Solutions to Pattern Recognition and Machine Learning

Joey Miller

October 14, 2017

Chapter 1

1.1 Differentiating (1.2) with respect to $\mathbf{w} = \{w_i\}$ yields

$$\begin{aligned}\frac{\partial}{\partial w_i}\left[\frac{1}{2}\sum_{n=1}^{N}\left(y(x_n,\mathbf{w})-t_n\right)^2\right] &= \sum_{n=1}^{N}\left(y(x_n,\mathbf{w})-t_n\right)\frac{\partial}{\partial w_i}y(x_n\mathbf{w}) \\ &= \sum_{n=1}^{N}\left(\sum_{j=0}^{M}(w_j x_n^j)-t_n\right)x_n^i.\end{aligned}$$

Setting it equal to 0 and rearranging gets us the following

$$\sum_{n=1}^{N}\sum_{j=1}^{M} w_j x_n^{j+i} = \sum_{n=1}^{N} t_n x_n^i$$

1.4 Differentiating (with respect to $g(y)$) the change of variable formula for a univariate PDF with transformation $x = g(y)$ and setting to zero to state the condition for a maximum, we get

$$0 = P_X'(g(y))|g'(y)| \pm g''(y)P_X(g(y))$$

where we are adding when $g'(y) > 0$ and subtracting when $g'(y) < 0$. Clearly, evaluating at $\hat{y}$, s.t. $P_X'(g(\hat{y})) = 0$ does not yield 0 for most functions g. In the case of a linear transformation, hoIver, $g''(y) = 0$ for all y, so we get our maximum at $\hat{y}$.

1.7 Beginning with

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2}(x^2+y^2)\right)dxdy$$

we change variables to polar coordinates resulting in

$$\begin{aligned}I^2 &= \int_{-\pi}^{\pi}\int_{0}^{\infty}\exp\left(-\frac{1}{2\sigma^2}r^2\right)rdrd\theta \\ &= \int_{-\pi}^{\pi}-\sigma^2\exp\left(-\frac{1}{2\sigma^2}r^2\right)\Big|_0^{\infty}d\theta \\ &= \int_{-\pi}^{\pi}\sigma^2 d\theta \\ &= 2\pi\sigma^2.\end{aligned}$$

Thus, $I = \sqrt{2\pi\sigma^2}$. Noting that I is the kernel of the $\mathcal{N}(x|\mu,\sigma^2)$,

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

1.8 Changing variables $z = x - \mu$ we integrate the expected value of the $\mathcal{N}(\mu, \sigma^2)$ variable,

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2} z^2\right)(z+\mu)dz \\
=&\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2} z^2\right) z dz + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2} z^2\right) \mu dz \\
=&0 + \mu
\end{aligned}$$

by the first term's odd kernel and the second's previously derived value.

Next, differentiating with respect to σ^2,

$$\begin{aligned}
&\frac{d}{d\sigma^2} \int_{-\infty}^{\infty} P(x|\mu, \sigma^2)dx = 0 \\
\Rightarrow &\frac{1}{2\sigma^4\sqrt{2\pi\sigma^2}} \int_{\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\left[(x-\mu)^2 - \sigma^2\right] dx = 0 \\
\Rightarrow &\int_{\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)(x-\mu)^2 dx = \sigma^2 \qquad (1) \\
\Rightarrow &\int_{\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\left(x^2 - 2x\mu + \mu^2\right) = \sigma^2 \\
\Rightarrow &\mathrm{E}\left[x^2\right] - 2\mu\mathrm{E}\left[x\right] + \mu^2 = \sigma^2 \\
\Rightarrow &\mathrm{E}\left[x^2\right] = \mu^2 + \sigma^2. \qquad (2)
\end{aligned}$$

Noting that equation (1) is $\mathrm{var}(x)$ and asserts that $\mathrm{var}(x) = \sigma^2$, it then follows that equation 1.51 from the text holds by substitution into (2): $\mathrm{var}(x) = \mathrm{E}[x^2] - \mathrm{E}[x]^2$

1.9 The mode of the Gaussian is μ and is shown by differentiating and solving at 0, then confirming that the second derivative is negative at $x = \mu$:

$$\begin{aligned}
\frac{d}{dx} P(x|\mu, \sigma^2) = \frac{d}{dx}\frac{1}{K} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) &= 0 \\
\Rightarrow -2(x-\mu)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) &= 0 \\
\Rightarrow x &= \mu.
\end{aligned}$$

$$\begin{aligned}
\frac{d^2}{dx^2} P(x|\mu, \sigma^2) &< 0 \\
\Rightarrow \frac{d}{dx} - 2(x-\mu)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) &< 0 \\
\Rightarrow -\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) + \frac{2(x-\mu)}{2\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)(-(x-\mu)) &< 0 \\
\Rightarrow -1 + 0 < 0 \text{ when evaluated at } x &= \mu.
\end{aligned}$$

For the multivariate Gaussian, we follow the same procedure:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{x}} P(\mathbf{x}|\boldsymbol{\mu}, \Sigma) \propto \frac{\partial}{\partial \mathbf{x}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}^T)\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) &= 0 \\
\Rightarrow -\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}^T)\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) &= 0 \qquad (3) \\
\Rightarrow \mathbf{x} &= \boldsymbol{\mu}.
\end{aligned}$$

For the second derivative test, it helps to work with the log-density to get directly at the quadratic form:

$$\begin{aligned}
\frac{\partial^2}{\partial \boldsymbol{x}^2} \log P(\mathbf{x}|\boldsymbol{\mu}, \Sigma) &= \frac{\partial}{\partial \boldsymbol{x}} - \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{x}) \\
&= -\Sigma^{-1}.
\end{aligned}$$

By the properties of the covariance matrix, the precision matrix has a positive determinant, and its negative is correspondingly negative.

1.10 If two variables x and z are independent, then

$$\mathrm{E}[x+z] = \int\int (x+z)P(x,z)dxdz = \int\int (x+z)P(x)P(z)dxdz = \int\int xP(x)dx + \int\int zP(z)dz = \mathrm{E}[x] + \mathrm{E}[z].$$

Similarly,

$$\begin{aligned}
\mathrm{var}[x+z] &= \int\int (x+z-\mu_x-\mu_z)^2 P(x,z)dxdz \\
&= \int\int (x+z)^2 P(x,z)dxdz - 2\int\int (x+z)(\mu_x+\mu_z)P(x,z)dxdz + (\mu_x+\mu_z)^2 \\
&= \int\int (x^2+2xz+z^2)P(x)P(z)dxdz - 2(\mu_x+\mu_z)^2 + (\mu_x+\mu_z)^2 \\
&= \int\int x^2 P(x)P(z)dxdz + 2\int xP(x)dx \int zP(z)dz + \int\int z^2 P(x)P(z)dxdz - (\mu_x+\mu_z)^2 \\
&= \mathrm{var}[x] + \mu_x^2 + 2\mu_x\mu_z + \mathrm{var}[z] + \mu_z^2 - (\mu_x+\mu_z)^2 \\
&= \mathrm{var}[x] + \mathrm{var}[z].
\end{aligned}$$

1.12 In the case that $n = m$, $\mathrm{E}[x_n x_m] = \mathrm{E}[x^2] = \mu^2 + \sigma^2$. If $n \neq m$, where by stipulation x are independent draws from a common distribution, $\mathrm{E}[x_n x_m] = \mu^2$ using the results implied by exercise (1.10).

Now, we verify that $\mathrm{E}\left[\mu_{\mathrm{ML}}\right] = \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^N x_n\right] = \frac{1}{N}N\mu_x = \mu_x$. Likewise, $\mathrm{E}\left[\sigma^2_{\mathrm{ML}}\right] = \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^N \left(x_n - \left(\frac{1}{N}\sum_{n=1}^N x_n\right)\right)^2\right]$ which after multiplying out yields

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^N \mathrm{E}[x_n^2] - \frac{2}{N^2}\sum_{n=1}^N\sum_{m=1}^N \mathrm{E}[x_n x_m] + \frac{1}{N^3}\sum_{n=1}^N\sum_{m=1}^N\sum_{p=1}^N \mathrm{E}[x_m x_p] \\
=&\mu^2 + \sigma^2 - \frac{1}{N^2}\sum_{n=1}^N\sum_{m=1}^N \mathrm{E}[x_n x_m] \\
=&\mu_2 + \sigma^2 - \frac{(N^2-N)\mu^2 + N(\mu^2+\sigma^2)}{N^2} \\
=&\sigma^2 + \frac{\sigma^2}{N} = \frac{N-1}{N}\sigma^2
\end{aligned}$$

1.15 Starting with expressing the M^{th} term of a polynomial in D dimensions as

$$\sum_{i_1=1}^D \sum_{i_2=1}^D \cdots \sum_{i_M=1}^D w_{i_1 i_2 \cdots i_M} x_{i_1} x_{i_2} \cdots x_{i_M},$$

we can simplify this expression by noting that many of the $x_{i_1} x_{i_2} \cdots x_{i_M}$ are identical when the indices all go from 1 to D. The corresponding number of independent parameters is thus many feIr than D^M. If we enforce the restriction that successive indices should be non-increasing, we eliminate all redundancy because any combination (corresponding to a product of values) needs only be written one way, without losing any terms (because every combination has a way of ordering the terms non-increasingly). Thus,

$$\sum_{i_1=1}^D \sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} w_{i_1 i_2 \cdots i_M} x_{i_1} x_{i_2} \cdots x_{i_M}.$$

Next, letting $n(D, M)$ be the number of independent parameters which appear at order M, we have

$$n(D,M) = \sum_{i=1}^D n(i, M-1) \tag{4}$$

which is easily seen by replacing $w_{i_1 i_2 \cdots i_M} x_{i_1} x_{i_2} \cdots x_{i_M}$ with 1 and substituting:

$$n(D,M) = \sum_{i_1=1}^D \sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} 1 = \sum_{i_1}^D n(i_1, M-1).$$

Now, I'll prove the following result by induction:

$$\sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!}. \tag{5}$$

For $D = 1$, we clearly have $\frac{(1+M-2)!}{0!(M-1)!} = \frac{(1+M-1!)}{0!M!}$ which is true for all M. Next, assuming the identity is true, I'll add the next term in the summation and do the following algebra:

$$\sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!(M-1)!} + \frac{(D+1+M-2)!}{(D+1-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!} + \frac{(D+1+M-2)!}{(D+1-1)!(M-1)!}$$

where the RHS is manipulated

$$\begin{aligned}
&= \frac{(D+M-1)!}{(D-1)!M!}\frac{D}{D} + \frac{(D+M-1)!}{(D)!(M-1)!}\frac{M}{M} \\
&= \frac{(D+M-1)!}{(D)!M!}\frac{D}{1} + \frac{(D+M-1)!}{(D)!(M)!}\frac{M}{1} \\
&= \frac{(D+M)!}{D!M!} \\
&= \frac{(D+1+M-1)!}{(D+1-1)!M!}
\end{aligned}$$

which proves the identity for $D+1$. Lastly, assuming that $n(D,M) = \frac{D+M-1}{(D-1)!M!}$ and verifying that for $M = 2$, the number of terms should simply be $\binom{D+1}{2}$, representing any combination of two numbers from 0 to D. Now, by making the following substitutions, we see that the assumption holds:

$$\begin{aligned}
n(D,M) &= \sum_{i=1}^{D} n(i, M-1) && \text{by (4)} \\
&= \sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!(M-1)!} && \text{by the assumption} \\
&= \frac{(D+M-1)!}{(D-1)!M!} && \text{by (5).}
\end{aligned}$$

1.17 We show the gamma function's recurrence relation through integration by parts with the following substitutions:

$$\begin{aligned}
dv = u^{x-1} \qquad & w = e^{-u} \\
v = \frac{u^x}{x} \qquad & dw = -e^{-u}
\end{aligned}$$

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

Evaluating $\Gamma(1) = \int_0^\infty e^0 e^{-u}du = -e^{-u}\Big|_0^\infty = 1$. Thus, $\Gamma(x+1) = x\Gamma(x) = x!$.

1.18 Using the given result (1.142) from the text, we can evaluate the LHS and RHS as the following:

$$\begin{aligned}
\text{LHS} &= \prod_{i=1}^{D} \sqrt{\pi} \\
\text{RHS} &= S_D \int_0^\infty e^{-r^2}(r^2)^{D/2-1} r dr.
\end{aligned} \tag{6}$$

Performing a change of variables $\sqrt{u} = r$, we can proceed on (6):

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

Uniting both sides now yields $\prod_{i=1}^{D} \sqrt{\pi} = S_D \frac{1}{2}\Gamma(\frac{D}{2})$ or $S_D = 2\pi^{\frac{D}{2}}/\Gamma(\frac{D}{2})$.

Using the formula for S_D, we can generalize it to a sphere of any size by simply scaling it to have the appropriate dimensions: $S_D(r) = S_D r^{D-1}$. Integrating this from $r = 0$ to $r = 1$ is done,

$$V_D = \int_0^1 S_D r^{D-1} dr = \frac{S_D}{D} r^D \Big|_0^1 = \frac{S_D}{D}.$$

Evaluating at $D = 2$ and $D = 3$ we get the familiar results $S_2 = 2\pi$, $V_2 = \pi$, $S_3 = 4\pi$, and $V_3 = \frac{4}{3}\pi$.

1.20 Starting with the D dimensional spherical Gaussian

$$P(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{||\mathbf{x}||^2}{2\sigma^2}\right) \tag{7}$$

we parameterize in terms of $r^2 = ||\mathbf{x}||^2$ and integrate around the surface of the D-sphere:

$$\int_{S_D} P(\mathbf{x}) = \int_{S_D} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) dr = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right).$$

Using the first-order approximation to the integral over r, we have

$$\int_{r_0-\epsilon/2}^{r_0+\epsilon/2} P(r) dr \approx P(r_0)\epsilon.$$

Next, we find the stationary point, $\hat{r}$, of $P(r)$ by differentiating

$$\begin{aligned}
\frac{d}{dr} P(r) &\propto (D-1) r^{D-2} - \frac{2r}{2\sigma^2} r^{D-1} = 0 \\
&\Rightarrow (D-1) = \frac{r^2}{\sigma^2} \\
&\Rightarrow r = \sqrt{D-1}\sigma \approx \sqrt{D}\sigma.
\end{aligned}$$

Considering $P(\hat{r} + \epsilon)$ for a small ϵ and large D, we can approximate $P(\hat{r} + \epsilon)$ by evaluating

$$\begin{aligned}
\frac{P(\hat{r}+\epsilon)}{P(\hat{r})} &= \left(1 + \frac{\epsilon}{\hat{r}}\right)^{D-1} \exp\left(-\frac{\epsilon^2 + 2\hat{r}\epsilon}{2\sigma^2}\right) \\
&\approx \left(\left((1 + \frac{\epsilon}{\hat{r}}\right)^{\hat{r}}\right)^{\frac{\hat{r}}{\sigma^2}} \exp\left(-\frac{\epsilon^2 + 2\hat{r}\epsilon}{2\sigma^2}\right) \\
&= \exp\left(\frac{\hat{r}\epsilon}{\sigma^2}\right) \exp\left(-\frac{\epsilon^2 + 2\hat{r}\epsilon}{2\sigma^2}\right) \\
&= \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)
\end{aligned}$$

implying that $P(\hat{r} + \epsilon) = P(\hat{r}) \exp\left(\frac{-\epsilon^2}{2\sigma^2}\right)$ (note that this differs from the textbook's assertion). Finally, we compare the density (7) at the origin and at a radius $\hat{r}$

$$\frac{P(\mathbf{0})}{P(\sqrt{D}\sigma)} = \exp\left(\frac{D}{2}\right).$$

1.22 Given a loss matrix with elements $L_{kj} = 1 - I_{kj}$, where I is the identity matrix, we can minimize the loss by choosing j such that the posterior probability for that class is highest. This is seen by noting that the risk is simply $1 - P(C_k|\mathbf{x})$, which minimizes at the highest probability. This matrix is interpreted as the 0-1 loss function.

1.24 With the loss matrix, L_{kj} and reject option with loss λ, the decision criterion for the minimum expected loss is choosing the option that yields a minimum of $\{\sum_k L_{kj} P(C_k|\mathbf{x}), \lambda\}$ where the decision is over j and the option to choose λ.

When the loss matrix is $L_{kj} = 1 - I_{kj}$, this simplifies to choosing a minimum of $\{1 - P(C_k|\mathbf{x}), \lambda\}$. Restating that as choosing the maximum of $\{P(C_k|\mathbf{x}), \theta\}$ where $\theta = 1 - \lambda$ achieves the result stated in 1.5.3.

1.25 Letting $\mathbf{t}$ be a vector of target variables won't change the derivation of this result. Considering

$$L(y, \epsilon) = \int \int (y(\mathbf{x}) + \epsilon\eta(\mathbf{x}) - \mathbf{t})^2 P(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}$$

we take its derivative with respect to ϵ, evaluate at $\epsilon = 0$, and enforcing the condition for all $\eta(\mathbf{x})$,

$$\left.\frac{\partial L(y, \epsilon)}{\partial \epsilon}\right|_{\epsilon=0} = \int \int 2(y(\mathbf{x}) - \mathbf{t})\eta(\mathbf{x}) P(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} = 0.$$

Reordering the integrals and considering $\eta(\mathbf{x})$ as a continuous function such that $\eta(\mathbf{x}) < 0$ iff $y(\mathbf{x}) < \mathbf{t}$, we can conclude that the kernel is 0. Now,

$$\left.\frac{\partial L(y, \epsilon)}{\partial \epsilon}\right|_{\epsilon=0} = \int 2(y(\mathbf{x}) - \mathbf{t})\eta(\mathbf{x}) P(\mathbf{x}, \mathbf{t}) d\mathbf{t} = 0.$$

Then, considering the case where $\eta(\mathbf{x}) = 1$, and rearranging

$$\begin{aligned}
\left.\frac{\partial L(y, \epsilon)}{\partial \epsilon}\right|_{\epsilon=0} &= \int 2(y(\mathbf{x}) - \mathbf{t}) P(\mathbf{x}, \mathbf{t}) d\mathbf{t} = 0 \\
y(\mathbf{x}) P(\mathbf{x}) &= \int \mathbf{t} P(\mathbf{t}|\mathbf{x}) P(\mathbf{x}) d\mathbf{t} \\
y(\mathbf{x}) &= \int \mathbf{t} P(\mathbf{t}|\mathbf{x}) d\mathbf{x} = \mathrm{E}[\mathbf{t}|\mathbf{x}].
\end{aligned}$$

1.27 The condition that $y(\mathbf{x})$ must satisfy to minimize $\mathrm{E}[L_q]$ is $\int \int |y(\mathbf{x}) - t|^q P(\mathbf{x}, t) d\mathbf{x} dt$. Assuming we are allowed arbitrary $y(\mathbf{x})$, we can write $y = y(\mathbf{x})$ so this can be simplified to $\int |y - t|^q P(t|\mathbf{x}) dt$ by integrating out $P(x)$.

First, I'll consider the case where $q \to 0$. The cost function is then a function of y that is 0 is a neighborhood around t and 1 elsewhere, corresponding to the continuous version of the 0-1 loss. Likewise, the value y that minimizes this integral for any $\mathbf{x}$ is the conditional mode of $P(t|\mathbf{x})$.

For the case where $q = 1$, we set the functional derivative to 0, noting the discontinuity:

$$\begin{aligned}
\frac{\partial L_1}{\partial \epsilon} &= \int_{-\infty}^{y} P(t|\mathbf{x}) dt + \int_{y}^{\infty} -P(t|\mathbf{x}) dt = 0 \\
\int_{-\infty}^{y} P(t|\mathbf{x}) dt &= \int_{-\infty}^{y} P(t|\mathbf{x}) dt
\end{aligned}$$

which occurs when y is the conditional median of $P(t|\mathbf{x})$.

1.29 An M-state discrete random variable x has entropy bounded by $\ln M$. This is shown by applying Jensen's inequality to the function $f(\cdot) = -\ln(\cdot)$, with $P(x_i)$, and $1/P(x_i)$ as follows:

$$\begin{aligned}
f(\sum_{i=1}^{M} \lambda_i x_i) &\leq \sum_{i=1}^{M} \lambda_i f(x_i) \\
\Rightarrow -\ln\left(\sum_{i=1}^{M} P(x_i) \frac{1}{P(x_i)}\right) &\leq \sum_{i=1}^{M} P(x_i) \left[-\ln\left(\frac{1}{P(x_i)}\right)\right] \\
\Rightarrow \ln(M) &\geq \sum_{i=1}^{M} P(x_i) \ln\left(\frac{1}{P(x_i)}\right) = \mathrm{H}[x].
\end{aligned}$$

1.31 Given two variables x and y with joint distribution $P(x, y)$, the differential entropy satisfies the property $\mathrm{H}[x, y] \leq \mathrm{H}[x] + \mathrm{H}[y]$ with equality when x and y are independent. This can be proven by showing that $\mathrm{H}[x] \geq \mathrm{H}[x|y]$ with the same equality condition. First, assuming independence,

$$\mathrm{H}[x|y] = -\int \int P(x, y) \ln(P(x|y)) dx dy = -\int \int P(x) P(y) \ln(P(x)) dx dy = \mathrm{H}[x] \int P(y) dy = \mathrm{H}[x].$$

And in general, $\mathrm{H}[x] - \mathrm{H}[x|y] = \mathrm{I}[x, y] \geq 0$, because the mutual information is simply a KL-divergence between $P(x, y)$ and $P(x)P(y)$.

1.34 Differentiating the functional (1.108) with respect to $P(x)$ with the Euler-Lagrange formula term by term and setting to 0 easily yields

$$\frac{\partial L}{\partial P(x)} = -(\ln P(x) + 1) + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2 = 0.$$

Solving for $P(x)$ implies (1.108). Simply noting that $P(x)$ is a quadratic form in x, the Gaussian needs only two independent parameters in the exponential, and that the mean of the distribution if assumed to be μ by (1.106), we know that $\lambda_2 = 0$, $-1 + \lambda_1 = \frac{1}{(2\pi\sigma^2)^{1/2}}$ by the normalization constraint, which implies $\lambda_3 = -\frac{1}{2\sigma^2}$.

$$P(x) = \exp(a) \exp\left(+\lambda_2 x + \lambda_3 (x - \mu)^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

1.35 The entropy of the univariate Gaussian is derived:

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

1.38 Assuming (1.114) holds, and that (1.115) holds for M terms, we can write the $M + 1^{\text{st}}$ as

$$f\left(\sum_{i=1}^{M} \lambda_i x_i + \lambda_{M+1} x_{M+1}\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i) + \lambda_{M+1} f(x_{M+1}).$$

Letting $\sum_{i=1}^{M} \lambda_i = \lambda$, $\sum_{i=1}^{M} x_i = a$, $\lambda_{M+1} = 1 - \lambda$, and $x_{M+1} = b$ allows us to apply (1.114).

1.41 The mutual information, $\mathrm{I}[x, y] = -\int\int P(x, y) \ln\left(\frac{P(x)P(y)}{P(x,y)}\right) dxdy$ and can be decomposed as follows:

$$\begin{aligned}
\mathrm{I}[x, y] &= -\int\int P(x, y) \ln(P(x))\, dxdy + \int\int P(x, y) \ln\left(\frac{P(x, y)}{P(y)}\right) dxdy \\
&= -\int P(x) \ln(P(x))\, dx + \int\int P(x, y) \ln(P(x|y))\, dxdy \\
&= \mathrm{H}[x] - \mathrm{H}[x|y] \text{ and by symmetry,} \\
&= \mathrm{H}[y] - \mathrm{H}[y|x].
\end{aligned}$$

Chapter 2

2.1 The Bernoulli distribution is normalized and is verified by

$$\sum_{x=0}^{1} P(x|\mu) = (1 - \mu) + \mu = 1.$$

Its mean is

$$\mathrm{E}[x] = \sum_{x=0}^{1} x P(x|\mu) = 0(1 - \mu) + 1\mu = \mu,$$

and variance is

$$\mathrm{var}[x] = \mathrm{E}[x^2] - \mathrm{E}[x]^2 = \sum_{x=0}^{1} x^2 P(x|\mu) - \mu^2 = \mu(1 - \mu).$$

Finally, the entropy is

$$\mathrm{H}[x] = -\sum_{x=0}^{1} P(x|\mu)\ln\left(P(x|\mu)\right) = -(1-\mu)\ln(1-\mu) - \mu\ln\mu.$$

2.3 The identity $\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}$ is easily seen by considering the the $N+1^{\text{st}}$ object, O_{N+1}, of the RHS (the number of possible groups of m objects out of a total of $N+1$ objects). There are $\binom{N}{m}$ possible groups that do not contain O_{N+1}, and $\binom{N}{m-1}$ possible groups that do contain O_{N+1}. Using this result, we can prove the binomial theorem by induction. For $N=1$, we have $(1+x)^1 = \sum_{m=0}^{1}\binom{1}{m}x^m = (1+x)$. Applying the theorem at (8),

$$\begin{aligned}
(1+x)^{N+1} &= (1+x)^N + (1+x)^N x \\
&= \sum_{m=0}^{N}\binom{N}{m}x^m + \sum_{m=0}^{N}\binom{N}{m}x^{m+1} \\
&= \sum_{m=0}^{N}\binom{N}{m+1}x^{m+1} + \sum_{m=0}^{N}\binom{N}{m}x^{m+1} + 1 \\
&= \sum_{m=0}^{N}\binom{N+1}{m+1}x^{m+1} + 1 \\
&= \sum_{m=1}^{N+1}\binom{N+1}{m}x^{m} + 1 \\
&= \sum_{m=0}^{N+1}\binom{N+1}{m}x^{m}
\end{aligned} \tag{8}$$

and the theorem is proved. Lastly, to prove that the binomial distribution is normalized, we apply the theorem

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

2.5 Beginning with

$$\Gamma(a)\Gamma(b) = \int_0^\infty \exp(-x)x^{a-1}dx\int_0^\infty \exp(-y)y^{b-1}dy$$

we bring in the integral over y resulting in

$$\int_0^\infty\int_0^\infty \exp(-(x+y))x^{a-1}y^{b-1}dydx$$

and make the substitution $y = t - x$

$$\int_0^\infty\int_x^\infty \exp(-t)x^{a-1}(t-x)^{b-1}dtdx.$$

Next, interchanging the order of integration (and correctly specifying the bounds),

$$\int_0^\infty\int_x^\infty \exp(-t)x^{a-1}(t-x)^{b-1}dxdt$$

and substituting $x = t\mu$ we have

$$\int_0^\infty\int_0^1 \exp(-t)(t\mu)^{a-1}(t-t\mu)^{b-1}td\mu dt.$$

Factoring the integral then using the definition of the Gamma function,

$$\int_0^\infty \exp(-t)t^{a+b-1}dt\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu = \Gamma(a+b)\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu.$$

2.9 To prove the normalization of the Dirichlet distribution, we consider

$$P_M(\mu_1,\ldots,\mu_{M-1}) = C_M \prod_{k=1}^{M-1} \mu_k^{\alpha_k-1}\left(1-\sum_{j=1}^{M-1}\mu_j\right)^{\alpha_M-1}.$$

Integrating over μ_{M-1},

$$\int_0^{1-\sum_{j=1}^{M-2}\mu_j} C_M \prod_{k=1}^{M-2}\mu_k^{\alpha_k-1}\mu_{M-1}^{\alpha_{M-1}}\left(\left[1-\sum_{j=1}^{M-2}\mu_j\right]-\mu_{M-1}\right)^{\alpha_M-1} d\mu_{M-1}$$

and changing variables by a scale factor, $\left(1-\sum_{j=1}^{M-2}\mu_j\right)v = \mu_{M-1}$ so the bounds on the integral in v go from 0 to 1 when μ_{M-1} varies from 0 to $1-\sum_{j=1}^{M-2}\mu_j$

$$\begin{aligned}
&= C_M \prod_{k=1}^{M-2}\mu_k^{\alpha_k-1}\int_0^1 \left(\left[1-\sum_{j=1}^{M-2}\mu_j\right]-\left[1-\sum_{j=1}^{M-2}\mu_j\right]x\right)^{\alpha_M-1}\left(\left[1-\sum_{j=1}^{M-2}\mu_j\right]x\right)^{\alpha_{M-1}-1}\left[1-\sum_{j=1}^{M-2}\mu_j\right]^{-1} dx\\
&= C_M \prod_{k=1}^{M-2}\mu_k^{\alpha_k-1}\left[1-\sum_{j=1}^{M-2}\mu_j\right]^{\alpha_M+\alpha_{M-1}-1}\int_0^1 (1-x)^{\alpha_M-1}x^{\alpha_{M-1}-1}dx\\
&= C_M \prod_{k=1}^{M-2}\mu_k^{\alpha_k-1}\left[1-\sum_{j=1}^{M-2}\mu_j\right]^{\alpha_M+\alpha_{M-1}-1}\frac{\Gamma(\alpha_M)\Gamma(\alpha_{M-1})}{\Gamma(\alpha_M+\alpha_{M-1})}.
\end{aligned}$$

Note that this process results in a Dirichlet with the jth component taking on the sum of the parameters associated with the jth through Mth components. In the end, after integrating over all components we have

$$\begin{aligned}
1 &= C_M \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\sum_{j=1}^M\alpha_j)}\\
C_M &= \frac{\Gamma(\sum_{j=1}^M\alpha_j)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}.
\end{aligned}$$

2.11 The expectation of the logarithm of an element μ_j from a Dirichlet distribution can be found as follows:

$$\begin{aligned}
\mathrm{E}[\ln\mu_j] &\propto \prod_{k\neq j}\int_{\mu_k}\mu_k^{\alpha_k-1}d\mu_k\int_{\mu_j} ln(\mu_j)\mu_j^{\alpha_j-1}d\mu_j\\
&= \frac{\partial}{\partial\alpha_j}\prod_k\int_{\mu_k}\mu_k^{\alpha_k-1}d\mu_k\\
&= \frac{\partial}{\partial\alpha_j}\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0)}.
\end{aligned}$$

Now stating the proportionality constant explicitly, we have

$$\begin{aligned}
\mathrm{E}[\ln\mu_j] &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)\cdots\Gamma(\alpha_K)}\frac{\partial}{\partial\alpha_j}\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0)}\\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)}\frac{\partial}{\partial\alpha_j}\frac{\Gamma(\alpha_j)}{\Gamma(\alpha_0)}\\
&= \frac{\partial}{\partial\alpha_j}\ln\left(\frac{\Gamma(\alpha_j)}{\Gamma(\alpha_0)}\right)\\
&= \frac{\partial}{\partial\alpha_j}\ln(\Gamma(\alpha_j)) - \frac{\partial}{\partial\alpha_j}\ln(\Gamma(\alpha_0)).
\end{aligned}$$

Finally, noting that $\partial\alpha_0 = \partial\alpha_j$, the last line can be written as $\frac{\partial}{\partial\alpha_j}\ln(\Gamma(\alpha_j)) - \frac{\partial}{\partial\alpha_0}\ln(\Gamma(\alpha_0)) = \psi(\alpha_j) - \psi(\alpha_0)$.

2.14 Using an analogous technique to exercise (1.34), we assign the constraints Lagrange multipliers and find stationary points of the functional

$$L(\mathbf{x}) = -\int P(\mathbf{x})\ln P(\mathbf{x})d\mathbf{x} + \lambda_1\left(\int P(\mathbf{x})d\mathbf{x}-1\right) + \lambda_2\left(\int \mathbf{x}P(\mathbf{x})\mathbf{x}d\mathbf{x}-\boldsymbol{\mu}\right) + \lambda_3\left(\int P(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T d\mathbf{x}-\Sigma\right)$$

Differentiating with respect to $P(\mathbf{x})$ and setting equal to zero yields the relationship

$$\frac{\partial L}{\partial P(\mathbf{x})} = \exp\left(-1 + \lambda_1 + \lambda_2 \mathbf{x} + \lambda_3 (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\right.$$

which is a quadratic form in $\mathbf{x}$ and is thus a Gaussian.

2.16 Using the following factorization, we have

$$\begin{aligned} P(x) &= \int_{-\infty}^{\infty} P(x|x_2) P(x_2) dx_2 \\ &\propto \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\tau_1 (x - (\mu_1 + x_2))^2\right] \exp\left[-\frac{1}{2}\tau_2 (x_2 - \mu_2)^2\right] dx_2 \end{aligned}$$

where $P(x|x_2) = P(x_1 + x_2|x_2)$, $x_1 + x_2 \sim \mathcal{N}(\mu_1 + x_2, \tau_1^{-1})$ and $x_2 \sim \mathcal{N}(\mu_2, \tau_2^{-1})$. Next, expanding terms retaining only those with either x or x_2 yields the next line which is correct up to a (to be determined) proportionality constant.

$$P(x) \propto \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\tau_1 \left[x^2 - 2x(\mu_1 + x_2) + (\mu_1 + x_2)^2\right]\right)\right] \exp\left[-\frac{1}{2}\tau_2 (x_2^2 - 2x_2\mu_2)\right] dx_2.$$

Considering only the exponent with terms containing x_2, we complete its square by adding the term

$$\frac{-2(\tau_1(\mu_1 - x) - \tau_2\mu_2)^2}{\tau_1 + \tau_2}.$$

This results in the exponent of a quadratic form in x_2, with x being treated as a constant as part of the mean term, and disappearing from the exponent after the integral is applied. Since the final form will be normalized (over x), only the remaining terms containing x^2 will need to be considered, which includes the prior fraction obtained by completing the square. This leaves

$$P(x) \propto \exp\left[-\frac{1}{2}\frac{\tau_1\tau_2}{\tau_1 + \tau_2}(x - K)^2\right] \tag{1}$$

implying that the precision is $\frac{\tau_1\tau_2}{\tau_1+\tau_2}$, which is all that is needed for the entropy. Thus, $H(x) = \frac{1}{2}\ln\left[2\pi\frac{\tau_1\tau_2}{\tau_1+\tau_2}\right]$.

2.17 Rewriting a square matrix A as

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

where the equality is obvious, we can verify that the first term is symmetric because each element $a_{ij} = \frac{1}{2}a_{ij} + \frac{1}{2}a_{ji} = \frac{1}{2}a_{ji} + \frac{1}{2}a_{ij} = a_{ji}$. The second term likewise, is anti-symmetric: $a_{ij} = \frac{1}{2}a_{ij} - \frac{1}{2}a_{ji} = -(\frac{1}{2}a_{ji} - \frac{1}{2}a_{ij}) = -a_{ji}$. Letting A be the precision matrix,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T\left(A^{\text{symm}} + A^{\text{anti}}\right)(\mathbf{x} - \boldsymbol{\mu})\right].$$

Considering an arbitrary off-diagonal term of the quadratic form associated with A_{anti} (the diagonals are 0, but the following still holds), we have $(x_i - \mu_i)a_{ij}(x_j - \mu_j) + (x_j - \mu_j)a_{ji}(x_i - \mu_i) = 0$. Thus, only A^{symm} need be considered.

2.20 By the spectral decomposition $\Sigma = \sum_i \lambda_i u_i u^T$ and considering any $\mathbf{a} = \sum_i a_i$ as a projection in each direction of u_i, we see that $\mathbf{a}^T\Sigma\mathbf{a} = \sum_i a_i^T \lambda_i u_i u^T a_i$. If for all $\mathbf{a}$, this is positive, then the λ_i must clearly all be positive. If all of the eigenvalues λ_i are positive, then for any $\mathbf{a}$, the result is positive.

2.22 Any symmetric matrix X can be decomposed as $U^T D U$, where D is a diagonal matrix and U is orthogonal. Thus, $(U^T D U)(U^T D^{-1} U) = I$ where D^{-1} is the diagonal matrix with reciprocal nonzero elements. Thus, the matrix $U^T D^{-1} U = X^{-1}$ is the (unique) inverse of X. To see that X^{-1} is symmetric, consider u_i as the i^{th} column of U^T. Then, X^{-1} is $\sum_i d_i u_i u_i^T$, a sum of symmetric matrices (hence, symmetric), where d_i is the i^{th} diagonal element of D.

2.24 Noting that $M = (A - BD^{-1}C)^{-1}$ and multiplying on the left (multiplying on the right results in combinations that aren't helpful) by $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ yields the relationship

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix}.$$

The LHS is clearly I, and there is no need to make use of the relationships implied between the products of blocks. Considering only the right, it remains to show that the resulting partitioned matrix is the identity. First, the top-left, replacing $(A - BD^{-1}C)^{-1} = M$:

$$A(A - BD^{-1}C)^{-1} + B(-D^{-1}C(A - BD^{-1}C)^{-1}) = (A - BD^{-1}C)(A - BD^{-1}C)^{-1} = I.$$

Now, the top right:

$$\begin{aligned} & A(-MBD^{-1}) + BD^{-1} + BD^{-1}CMBD^{-1} \\ =&(-A + BD^{-1}C)(MBD^{-1}) + BD^{-1} \\ =&- M^{-1}MBD^{-1} + BD^{-1} \\ =&0. \end{aligned}$$

Bottom left:

$$CM + D(-D^{-1}CM) = CM - CM = 0.$$

Bottom right:

$$C(-MBD^{-1}) + DD^{-1} + DD^{-1}CMBD^{-1} = -CMBD^{-1} + I + CMBD^{-1} = I.$$

2.28 Given a joint distribution over $\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$ with mean $\begin{pmatrix} \boldsymbol{\mu} \\ A\boldsymbol{\mu} + \boldsymbol{b} \end{pmatrix}$ and covariance $\begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^T \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T \end{pmatrix}$ we use the block partitioning results (2.92) and (2.93) and easily find that the random variable $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$.

Next, using (2.81) and (2.82) from the text, $P(\mathbf{y}|\mathbf{x})$ is seen to have parameters

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}} &= \mathbf{A}\boldsymbol{\mu} + \boldsymbol{b} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{A}\mathbf{x} + \boldsymbol{b} \\ \boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}} &= \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T - \mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T = \mathbf{L}^{-1} \end{aligned}$$

and thus the distribution $\mathcal{N}(\mathbf{A}\mathbf{x} + \boldsymbol{b}, \mathbf{L}^{-1})$.

2.32 To-write (the algebra is very messy)

2.34 Maximizing the log-likelihood of (2.118) is straightforwardly accomplished by differentiating with respect to Σ^{-1} and setting equal to zero:

$$\begin{aligned} \frac{\partial P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}} &= \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \left(-\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln(1/|\boldsymbol{\Sigma}^{-1}|) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})^T \right) \\ &= \frac{N}{2}\boldsymbol{\Sigma}^T - \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}}\frac{1}{2}\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}\mathrm{Tr}\left[(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\right] \\ &= \frac{N}{2}\boldsymbol{\Sigma}^T - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T = 0 \\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T. \end{aligned}$$

2.36 Sequential ML estimate for the variance of a Normal distribution using the Robbins-Monro formula:

$$\begin{aligned} \sigma^2_{(N)} &= \sigma^2_{(N-1)} + \alpha_{(N-1)}\frac{\partial}{\partial \sigma^2_{(N-1)}}\ln P(x_N|\theta_{(N-1)}) \\ &= \sigma^2_{(N-1)} + \alpha_{(N-1)}\left[-\frac{1}{2\sigma^2_{(N-1)}} + \frac{1}{2\sigma^2_{(N-1)}}(x_N - \mu)^2 \right] \\ &= \sigma^2_{(N-1)} + \frac{\alpha_{(N-1)}}{2\sigma^4_{(N-1)}}\left[-\sigma^2_{(N-1)} + (x_N - \mu)^2 \right]. \end{aligned}$$

Comparing to

$$\begin{aligned}
\sigma^2_{(N)} &= \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2 \\
&= \frac{1}{N}\left((x_N-\mu)^2+\sum_{n=1}^{N}(x_n-\mu)^2\right) \\
&= \frac{1}{N}(x_N-\mu)^2+\frac{N-1}{N}\left[\frac{1}{N-1}\sum_{n=1}^{N-1}(x_n-\mu)^2\right] \\
&= \frac{1}{N}(x_N-\mu)^2+\frac{N-1}{N}\sigma^2_{(N-1)} \\
&= \frac{1}{N}(x_N-\mu)^2+\sigma^2_{(N-1)}-\frac{\sigma^2_{(N-1)}}{N} \\
&= \sigma^2_{(N-1)}+\frac{1}{N}\left[-\sigma^2_{(N-1)}+(x_N-\mu)^2\right]
\end{aligned}$$

implying that $\frac{1}{N}=\frac{\alpha_{(N-1)}}{2\sigma^4_{(N-1)}}$ so $\alpha_{(N-1)}=\frac{2\sigma^4_{(N-1)}}{N}$.

2.40 Posterior inference for a multivariate Normal with unknown mean and known variance proceeds as follows:

$$\begin{aligned}
P(\boldsymbol{\mu}|\mathbf{x},\boldsymbol{\Sigma}) &\propto P(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})P(\boldsymbol{\mu}|\boldsymbol{\Sigma}) \\
&\propto \exp\left(\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\right)\exp\left((\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right)
\end{aligned}$$

and keeping only terms attached to $\boldsymbol{\mu}$

$$\propto \exp\left(\boldsymbol{\mu}^T(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})\boldsymbol{\mu}-2\boldsymbol{\mu}^T(N\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu_0})\right)$$

finally, completing the square yields

$$\propto \exp\left(\boldsymbol{\mu}-(N\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu_0})(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})^{-1}\right)^T(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})\left(\boldsymbol{\mu}-(N\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu_0})(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})^{-1}\right)$$

implying that the posterior, $\boldsymbol{\mu}\sim\mathcal{N}(N\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu_0},(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})^{-1})$.

2.46 Using a Gamma(a,b) prior on the precision, the predictive distribution for a $\mathcal{N}(\mu,\tau^{-1})$ is the t-distribution and is found by integrating

$$\begin{aligned}
P(x|\mu,a,b) &= \int_0^{\inf}\frac{b^ae^{-b\tau}}{\Gamma(a)}\left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}}\exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\}d\tau \\
&= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}}\int_0^{\infty}e^{-b\tau}\tau^{a-1}\tau^{\frac{1}{2}}\exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\}d\tau \\
&= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}}\int_0^{\infty}\tau^{a-\frac{1}{2}}\exp\left\{-b\tau-\frac{\tau}{2}(x-\mu)^2\right\}d\tau
\end{aligned}$$

and changing variables $z=\tau(b+\frac{(x-\mu)^2}{2})$ yields

$$\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}}\frac{1}{\left(b+\frac{(x-\mu)^2}{2}\right)^{a+\frac{1}{2}}}\int_0^{\infty}z^{a-\frac{1}{2}}\exp(-z)dz=\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}}\frac{1}{\left(b+\frac{(x-\mu)^2}{2}\right)^{a+\frac{1}{2}}}\Gamma\left(a+\frac{1}{2}\right).$$

Rewriting with $a=\frac{v}{2}$ and $b=\frac{v}{2\lambda}$

$$\begin{aligned}
&= \frac{\left(\frac{v}{2\lambda}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)\sqrt{2\pi}}\frac{\Gamma\left(\frac{v}{2}+\frac{1}{2}\right)}{\left(\frac{v}{2\lambda}+\frac{(x-\mu)^2}{2}\right)^{\frac{v}{2}+\frac{1}{2}}} \\
&= \frac{\Gamma\left(\frac{v}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)}\left(\frac{v\pi}{\lambda}\right)^{-\frac{1}{2}}\left(1+\frac{\lambda(x-\mu)^2}{v}\right)^{-\frac{v}{2}-\frac{1}{2}}.
\end{aligned}$$

2.47 The t-distribution has the following factor with x

$$\left[1+\frac{\lambda(x-\mu)^2}{v}\right]^{-\frac{v-1}{2}} = \left(\left[1+\frac{\lambda(x-\mu)^2}{v}\right]^{v+1}\right)^{-\frac{1}{2}}.$$

Letting $v \to \infty$ and recognizing the power as an exponential leaves us with

$$\left(\left[1+\frac{\lambda(x-\mu)^2}{v}\right]^{v+1}\right)^{-\frac{1}{2}} \approx \left(\left[1+\frac{\lambda(x-\mu)^2}{v}\right]^{v}\right)^{-\frac{1}{2}} = \left(\exp\left[\lambda(x-\mu)^2\right]\right)^{-\frac{1}{2}} = \exp\left[-\frac{\lambda(x-\mu)^2}{2}\right].$$

Thus, the distribution is normal with mean μ and $\sigma^2 = 1/\lambda$.

2.51 Given that $\exp(iA) = \cos A + i\sin A$ and considering that $\exp(iA)\exp(-iA) = 1$, (2.177), $1 = \sin^2 A + \cos^2 A$ is proved by

$$\begin{aligned}
1 &= \exp(iA)\exp(-iA) \\
&= (\cos A + i\sin A)(\cos(-A) + i\sin(-A)) \\
&= (\cos A + i\sin A)(\cos(A) + -i\sin(A)) \\
&= \cos^2 A + sin^2 A.
\end{aligned}$$

Next, $\cos A\cos B + \sin A\sin B = \cos(A-B)$ (2.178) is proved using $\cos(A-B) = \Re\left[\exp(i(A-B))\right]$:

$$\begin{aligned}
\Re\left[\exp(i(A-B))\right] &= \Re\left[\exp(iA)\exp(-iB))\right] \\
&= \Re\left[(\cos A + i\sin A)(\cos(-B) + i\sin(-B)\right] \\
&= \Re\left[\cos A\cos B + \sin A\sin B\right] \\
&= \cos A\cos B + \sin A\sin B.
\end{aligned}$$

Lastly, noting that $\sin(A-B) = \Im\left[\exp(i(A-B))\right]$, we have

$$\begin{aligned}
\sin(A-B) = \Im\left[\exp(i(A-B))\right] &= \Im\left[(\cos A + i\sin A)(\cos(-B) + i\sin(-B))\right] \\
&= \Im\left[i\sin A\cos B - i\sin B\cos A\right] \\
&= \sin A\cos B - \sin B\cos A
\end{aligned}$$

which is (2.183).

2.56 Noting the general form and notation of the exponential family distributions:

$$P(\mathbf{x}|a,b) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left[(\boldsymbol{\eta})^T u(\mathbf{x})\right]$$

we find the natural parameters of the Beta, Gamma, and Von-Mises distributions.

Beta:

$$\begin{aligned}
P(\mu|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\exp\left[(a-1)\ln\mu + (b-1)\ln(1-\mu)\right]
\end{aligned}$$

so $h(\mathbf{x}) = 1$, $g(\boldsymbol{\eta}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$, and $u(\mathbf{x}) = (\mu, 1-\mu)$ where $\mathbf{x} = \mu$ and $\boldsymbol{\eta}^T = (a-1, b-1)$.

Gamma:

$$\begin{aligned}
P(\lambda|a,b) &= \lambda^{a-1}\frac{b^a}{\Gamma(a)}\exp(-b\lambda) \\
&= \exp(\ln(\lambda^{a-1}))\frac{b^a}{\Gamma(a)}\exp(-b\lambda) \\
&= \frac{b^a}{\Gamma(a)}\exp((a-1)\ln\lambda - b\lambda)
\end{aligned}$$

so again $h(\mathbf{x}) = 1$, $g(\boldsymbol{\eta}) = b^a/\Gamma(a)$, $u(\mathbf{x}) = (\ln\lambda, -\lambda)^T$ where $\boldsymbol{\eta}^T = (a-1, b)$ and $\mathbf{x} = \lambda$.

Von-Mises: First, note that $\cos(\theta - \theta_0) = \cos\theta\cos\theta_0 + \sin\theta\sin\theta_0$. Then,

$$\begin{aligned} P(\theta|\theta_0, m) &= \frac{1}{2\pi I_0(m)} \exp(m\cos(\theta - \theta_0)) \\ &= \frac{1}{2\pi I_0(m)} \exp\left(m\cos\theta\cos\theta_0 + m\sin\theta\sin\theta_0\right). \end{aligned}$$

So $h(\mathbf{x}) = 1$, $g(\boldsymbol{\eta}) = 1/(2\pi I_1(m))$, $\boldsymbol{\eta}^T = (m\cos\theta_0, m\sin\theta_0)$, $u(\mathbf{x}) = \mathbf{x}$ and $\mathbf{x} = (\cos\theta, \sin\theta)^T$.

2.60 Using the log-likelihood of this histogram-density model, $P(\mathbf{x}|\mathbf{h}, \boldsymbol{\Delta}) = \prod_i h_i^{n_i}$, and the normalization constraint $\sum_i h_i\Delta_i = 1$, the Lagrangian is

$$L(\mathbf{h}, \lambda) = \sum_i n_i \log h_i + \lambda\left(\sum_i (h_i\Delta_i) - 1\right).$$

Differentiating with respect to each h_j and λ and setting to zero yields

$$\begin{aligned} \frac{\partial L}{\partial h_j} &= \frac{n_i}{h_j} + \lambda\Delta_j = 0 \\ \frac{\partial L}{\partial \lambda} &= \sum_i (h_i\Delta_i) - 1 = 0. \end{aligned}$$

Consequently,

$$\begin{aligned} \lambda &= -N \\ h_j &= \frac{n_j}{N\Delta_j}. \end{aligned}$$

Thus, the density is proportional to the count of observations that fall into region j, divided by the volume of region j, as expected.

Chapter 3

3.1 Given $\tanh(a) = \frac{1-e^{-2a}}{1+e^{-2a}}$,

$$\tanh(a) = \frac{1-e^{-2a}}{1+e^{-2a}} = \sigma(2a) - \frac{e^{-2a}}{1+e^{-2a}} = \sigma(2a) - 1 + 1 - \frac{e^{-2a}}{1+e^{-2a}} = \sigma(2a) - 1 + \frac{1+e^{-2a}}{1+e^{-2a}} - \frac{e^{-2a}}{1+e^{-2a}} = 2\sigma(2a) - 1.$$

A (scaled) linear combination of the basis functions $\sigma(\bullet)$ and $\tanh(\bullet)$ are related by

$$\begin{aligned} y(x, \mathbf{u}) &= u_0 + \sum_{j=1}^{M} u_j \tanh\left(\frac{x-\mu_j}{s}\right) \\ &= u_0 + \sum_{j=1}^{M} 2u_j\sigma\left(\frac{x-\mu_j}{s/2}\right) - u_j \\ &= u_0 - \sum_{j=1}^{M}\mu_j + \sum_{j=1}^{M} 2u_j\sigma\left(\frac{x-\mu_j}{s/2}\right). \end{aligned}$$

It is noted that the above can be put into the form

$$w_0 + \sum_{j=1}^{M} w_j\sigma\left(\frac{x-\mu_j}{t}\right)$$

and the parameters easily related.

3.2 Decompose $v = v_1 + v_2$, where v_1 is in $\text{col}(\boldsymbol{\Phi})$ and v_2 is in $\text{col}(\boldsymbol{\Phi})^\perp$. Noting that $v_1 = \boldsymbol{\Phi}x$ for some x, and $\boldsymbol{\Phi}^T v_2 = 0$ (because v_2 lies in the left nullspace of $\boldsymbol{\Phi}$). Then

$$\begin{aligned} \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T v &= \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T v_1 + \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T v_2 \\ &= \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi}x + \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}0 \\ &= \boldsymbol{\Phi}x \\ &= v_1. \end{aligned}$$

Now, the least squares solution $\mathbf{w} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}\mathbf{t}$ multiplied on the left by $\mathbf{\Phi}$ yields the estimate $\hat{\mathbf{t}} = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}\mathbf{t}$ which is easily seen as the projection of $\mathbf{t}$ onto $\mathcal{S}$, or col($\mathbf{\Phi}$).

3.4 Considering a linear model of the form

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

and the sum of squares error function, one shows that adding independent error random variables, ϵ, with mean 0 and variance σ^2 to the $\mathbf{x}$, and averaging over the errors, results in

$$\begin{aligned}
\int E_D(\mathbf{w}, \epsilon) d\epsilon &= \int \frac{1}{2} \sum_{n=1}^{N} \left(y(\mathbf{x}_n + \epsilon_n, \mathbf{w}) - t_n\right)^2 d\epsilon \\
&= \int \frac{1}{2} \sum_{n=1}^{N} \left(y(\mathbf{x}_n, \mathbf{w}) + \sum_{i}^{D} w_i \epsilon_{n,i} - t_n \right)^2 d\epsilon \\
&= \int \frac{1}{2} \sum_{n=1}^{N} \left(\left[y(\mathbf{x}_n - t_n)\right]^2 + 2 \left[\sum_{i=1}^{D} w_i \epsilon_{n,i}\right] \left[y(\mathbf{x}_n, \mathbf{w})\right] + \left[\sum_{i=1}^{D} w_i \epsilon_{n,i}\right]^2 \right) d\epsilon \\
&= \frac{1}{2} \sum_{n=1}^{N} \left(\left[y(\mathbf{x}_n - t_n)\right]^2 + 0 + \sum_{i=1}^{D} w_i^2 \sigma^2 \right).
\end{aligned}$$

The result implies the equivalence to a mean-squared-plus-L2 regularization on the non-intercept weights with penalty σ^2.

3.5 It is easily seen that minimizing

$$\frac{1}{2} \sum_{n=1}^{N} (t_n - \mathbf{w}^T \phi(\mathbf{x}))^2$$

subject to $\sum_{j=1}^{M} |w_i|^q \leq \eta$ for some η is equivalent to minimizing the Lagrangian

$$\begin{aligned}
L(\mathbf{w}, \lambda/2) &= \frac{1}{2} \sum_{n=1}^{N} (t_n - \mathbf{w}^T \phi(\mathbf{x}))^2 - \frac{\lambda}{2} \left(\eta - \sum_{j=1}^{M} |w_i|^q \right) \\
&= \frac{1}{2} \sum_{n=1}^{N} (t_n - \mathbf{w}^T \phi(\mathbf{x}))^2 + \frac{\lambda}{2} \sum_{j=1}^{M} |w_i|^q - \frac{\lambda}{2} \eta
\end{aligned}$$

with respect to $\frac{\lambda}{2}$ and $\mathbf{w}$ subject to the same condition, $\frac{\lambda}{2} \geq 0$, and $\frac{\lambda}{2} \left(\eta - \sum_{j=1}^{M} |w_i|^q \right) = 0$. Thus, if the constraint is active, then $\lambda > 0$ and $\eta = \sum_{j=1}^{M} |w_i|^q$ for the corresponding optimal values of $\mathbf{w}$ and λ. The Lagrangian, with the conditions, is identical to the form (3.29), the regularized error function.

3.6 In the linear basis function regression model for a multivariate target variable $\mathbf{t}$ having a distribution $\mathcal{N}(\mathbf{y}(\mathbf{x}, \mathbf{W}), \mathbf{\Sigma})$ where $\mathbf{y}(\mathbf{x}) = \mathbf{W}^T \boldsymbol{\phi}(\mathbf{x})$, it is shown that the ML solution $\mathbf{W}_{\text{ML}}$ has, for each column, a solution given by expression (3.15). Maximizing the log-likelihood with respect to $\mathbf{W}$ yields

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{W}} \log L(\mathbf{W}) &= 0 \\
\sum_{n=1}^{N} \mathbf{\Sigma}^{-1} (t_n - \mathbf{W}^T \boldsymbol{\phi}(\mathbf{x})) \boldsymbol{\phi}(\mathbf{x})^T &= 0 \\
\sum_{n=1}^{N} (t_n - \mathbf{W}^T \boldsymbol{\phi}(\mathbf{x})) \boldsymbol{\phi}(\mathbf{x})^T &= 0 \\
\sum_{n=1}^{N} t_n \phi(\mathbf{x})^T &= \mathbf{W}^T \phi(\mathbf{x}) \phi(\mathbf{x})^T
\end{aligned}$$

Keeping in mind that the above is a vector equality and taking the equation element by element, we see that this implies that each column of $\mathbf{W}$ is given by something like (3.15).
Following the exact same procedure as in exercise (2.34), we easily see the maximum-likelihood solution is given by

$$\hat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n))(\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n))^T$$

where the ML estimate of $\mathbf{W}$ is likewise used.

3.9 Consider the linear basis function model in Section 3.1, and suppose that we have already observed N data points, so that the posterior distribution over w is given by (3.49). This posterior can be regarded as the prior for the next observation. By considering an additional data point (x_{N+1}, t_{N+1}), and by completing the square in the exponential, show that the resulting posterior distribution is again given by (3.49) but with S_N replaced by S_{N+1} and m_N replaced by m_{N+1}. Repeat the previous exercise but instead of completing the square by hand, make use of the general result for linear-Gaussian models given by (2.116).
Writing $\Pr(\mathbf{w}|\mathbf{t}_N) = \Pr(\mathbf{w})$, (suppressing dependence on the earlier data and treating the posterior of the weight vector as its prior) we can apply Bayes Rule as follows:

$$\Pr(\mathbf{w}|\mathbf{t}_{N+1}) = \frac{\Pr(\mathbf{w})\Pr(\mathbf{t}_{N+1}|\mathbf{w})}{\Pr(\mathbf{t}_{N+1})}$$

where $\mathbf{w} \sim \mathcal{N}(\mathbf{m}_N, \mathbf{S}_N)$, and $\mathbf{t}_{N+1} \sim \mathcal{N}(\boldsymbol{\phi}\mathbf{w}, \boldsymbol{\beta}^{-1})$, where $\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\beta}\boldsymbol{\phi}^T\mathbf{t})$. Applying (2.116) [CONTINUE THIS...]

3.13 Show that the predictive distribution $p(t|\mathbf{x}, \mathbf{t})$ for the model discussed in Ex-ercise 3.12 is given by a Students t-distribution of the form

$$p(t|\mathbf{x}, \mathbf{t}) = St(t|\mu, \lambda, \nu)(3.114)$$

and obtain expressions for μ, λ, and ν.

3.16 Derive the result (3.86) for the log evidence function $p(\mathbf{t}|\alpha, \beta)$ of the linear regression model by making use of (2.115) to evaluate the integral (3.77) directly.

3.21

3.24

Chapter 4