My solutions to

Deep Learning: Foundations and Concepts

Dario Miro Konopatzki

2 Probabilities

2.1

$$
\begin{aligned}
p(C=1|T=1) &= \frac{p(T=1|C=1)p(C=1)}{p(T=1)} \\
&= \frac{p(T=1|C=1)p(C=1)}{p(T=1,C=0)+p(T=1,C=1)} \\
&= \frac{p(T=1|C=1)p(C=1)}{p(T=1|C=0)p(C=0)+p(T=1|C=1)p(C=1)} \\
&= \frac{0.9\cdot 0.001}{0.03\cdot(1-0.001)+0.9\cdot 0.001} \\
&\approx 0.029
\end{aligned}
$$

2.2

Let Y denote the yellow die, B the blue die, G the green die and R the red die. We consider throws of pairs of independent dice, i.e. $p(D_1, D_2) = p(D_1)p(D_2)$. Each die takes on a unique value in a given throw, such that e.g. $(G = 5) := (G = 5, (B = 0 \text{ or } B = 4))$ and $(G = 1, B = 0)$ are mutually exclusive events, hence $p(G = 5 \text{ or } (G = 1, B = 0)) = P(G = 5) + P(G = 1, B = 0)$.

$$
\begin{aligned}
p(B>Y) &= p(B=4, Y=3) \\
&= p(B=4)p(Y=3) \\
&= \frac{4}{6}\cdot\frac{6}{6} \\
&= \frac{2}{3}
\end{aligned}
$$

$$\begin{aligned}
p(G > B) &= p(G = 5 \text{ or } (G = 1, B = 0)) \\
&= p(G = 5) + p(G = 1)p(B = 0) \\
&= \frac{3}{6} + \frac{3}{6} \cdot \frac{2}{6} \\
&= \frac{2}{3} \\
p(R > G) &= p(R = 6 \text{ or } (R = 2, G = 1)) \\
&= p(R = 6) + p(R = 2)p(G = 1) \\
&= \frac{2}{6} + \frac{4}{6} \cdot \frac{3}{6} \\
&= \frac{2}{3} \\
p(Y > R) &= p(Y = 3, R = 2) \\
&= p(Y = 3)p(R = 2) \\
&= \frac{6}{6} \cdot \frac{4}{6} \\
&= \frac{2}{3}
\end{aligned}$$

2.3

$$\begin{aligned}
\int_{\mathbb{R}} p_U(u)p_V(y-u)du &= \frac{d}{dy}\int_{-\infty}^{y}\left(\int_{\mathbb{R}} p_U(u)p_V(\tilde{y}-u)du\right)d\tilde{y} \\
&= \frac{d}{dy}\int_{\mathbb{R}}\left(\int_{\mathbb{R}} \mathbb{1}_{\tilde{y}\le y}(u,\tilde{y})p_U(u)p_V(\tilde{y}-u)du\right)d\tilde{y} \\
&= \frac{d}{dy}\int_{\mathbb{R}^2} \mathbb{1}_{\tilde{y}\le y}(u,\tilde{y})p_U(u)p_V(\tilde{y}-u)d(u,\tilde{y}) \qquad \text{Fubini} \\
&\overset{\star}{=} \frac{d}{dy}\int_{\mathbb{R}^2} \mathbb{1}_{u+v\le y}(u,v)p_U(u)p_V(v)\underbrace{\left|\det\begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}\right|}_{=1}d(u,v) \\
&= \frac{d}{dy}\int_{\mathbb{R}^2} \mathbb{1}_{u+v\le y}(u,v)p_{U,V}(u,v)d(u,v) \qquad U,V \text{ ind.} \\
&= \frac{d}{dy}P(U+V\le y) \\
&= \frac{d}{dy}P(Y\le y) \\
&= p_Y(y)
\end{aligned}$$

$\star$ Transformation $\tilde{y}(u,v) := (u, u+v)$.

2.4

$$
\begin{aligned}
\int_c^d p(x)dx &= \int_c^d \frac{1}{d-c}dx \\
&= \frac{1}{d-c}\int_c^d dx \\
&= \frac{1}{d-c}\left[x\right]_c^d \\
&= \frac{1}{d-c}(d-c) \\
&= 1 \\
\mathbb{E}_p[X] &= \int_c^d xp(x)dx \\
&= \int_c^d x\frac{1}{d-c}dx \\
&= \frac{1}{d-c}\int_c^d xdx \\
&= \frac{1}{d-c}\left[\frac{1}{2}x^2\right]_c^d \\
&= \frac{1}{2(d-c)}(d^2-c^2) \\
&= \frac{1}{2(d-c)}(d-c)(d+c) \\
&= \frac{d+c}{2} \\
\mathbb{E}_p[X^2] &= \int_c^d x^2p(x)dx \\
&= \int_c^d x^2\frac{1}{d-c}dx \\
&= \frac{1}{d-c}\int_c^d x^2dx \\
&= \frac{1}{d-c}\left[\frac{1}{3}x^3\right]_c^d \\
&= \frac{1}{3(d-c)}(d^3-c^3) \\
&= \frac{1}{3(d-c)}(d-c)(d^2+c^2+cd)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{3}(d^2 + c^2 + cd) \\
\mathbb{E}_p[X]^2 &= \left(\frac{d+c}{2}\right)^2 \\
&= \frac{d^2 + 2cd + c^2}{4} \\
\mathbb{V}_p[X] &= \mathbb{E}_p[X^2] - \mathbb{E}_p[X]^2 \\
&= \frac{1}{3}(d^2 + c^2 + cd) - \frac{(d+c)^2}{2^2} \\
&= \frac{1}{3}(d^2 + c^2 + cd) - \frac{d^2 + 2cd + c^2}{4} \\
&= \frac{1}{12}(4d^2 + 4c^2 + 4cd - 3d^2 - 6cd - 3c^2) \\
&= \frac{1}{12}(d^2 - 2cd + c^2) \\
&= \frac{1}{12}(d-c)^2
\end{aligned}
$$

2.5

Exponential distribution

$$
\begin{aligned}
\int p(x|\lambda)dx &= \int_0^\infty \lambda e^{-\lambda x} dx \\
&= \lambda \int_0^\infty e^{-\lambda x} dx \\
&= \lambda \left[-\frac{1}{\lambda} e^{-\lambda x}\right]_0^\infty \\
&= \lambda \left[0 - \left(-\frac{1}{\lambda} e^{-\lambda \cdot 0}\right)\right] \\
&= \lambda \cdot \frac{1}{\lambda} \\
&= 1
\end{aligned}
$$

Laplace distribution

$$
\begin{aligned}
\int p(x|\mu,\gamma) &= \int_{-\infty}^\infty \frac{1}{2\gamma} e^{-\frac{|x-\mu|}{\gamma}} dx \\
&= \frac{1}{2\gamma} \int_{-\infty}^\infty e^{-\frac{|x-\mu|}{\gamma}} dx
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2\gamma}\left(\int_{-\infty}^{\mu} e^{-\frac{|x-\mu|}{\gamma}}dx + \int_{\mu}^{\infty} \frac{1}{\gamma} e^{-\frac{|x-\mu|}{\gamma}}dx\right) \\
&= \frac{1}{2\gamma}\left(\int_{-\infty}^{\mu} e^{-\frac{\mu-x}{\gamma}}dx + \int_{\mu}^{\infty} e^{-\frac{x-\mu}{\gamma}}dx\right) \\
&= \frac{1}{2\gamma}\left(\int_{-\infty}^{\mu} e^{\frac{x-\mu}{\gamma}}dx + \int_{\mu}^{\infty} e^{\frac{\mu-x}{\gamma}}dx\right) \\
&= \frac{1}{2\gamma}\left(\left[\gamma e^{\frac{x-\mu}{\gamma}}\right]_{-\infty}^{\mu} + \left[-\gamma e^{\frac{\mu-x}{\gamma}}\right]_{\mu}^{\infty}\right) \\
&= \frac{1}{2\gamma}\left(\gamma\left[e^{0} - 0\right] - \gamma\left[0 - (e^{0})\right]\right) \\
&= \frac{\gamma}{2\gamma}(1-(-1)) \\
&= \frac{1}{2}\cdot 2 \\
&= 1
\end{aligned}
$$

2.6

$$
\begin{aligned}
\int p(x|\mathcal{D}) &= \int_{-\infty}^{\infty} \frac{1}{N}\sum_{n=1}^{N} \delta(x-x_n)dx \\
&= \frac{1}{N}\sum_{n=1}^{N}\int_{-\infty}^{\infty} \delta(x-x_n)dx \qquad \text{finite sum} \\
&= \frac{1}{N}\sum_{n=1}^{N} 1 \qquad \text{by def. of } \delta \\
&= \frac{1}{N}\cdot N \\
&= 1
\end{aligned}
$$

2.8

$$
\begin{aligned}
\mathbb{V}[f] &= \mathbb{E}\left[(f(X) - \mathbb{E}[f(X)])^2\right] \\
&= \mathbb{E}\left[f(X)^2 - 2f(X)\mathbb{E}[f(X)] + \mathbb{E}[f(X)]^2\right] \\
&= \mathbb{E}\left[f(X)^2\right] - 2\mathbb{E}[f(X)]\mathbb{E}[f(X)] + \mathbb{E}[f(X)]^2 \qquad \text{linearity of } \mathbb{E} \\
&= \mathbb{E}\left[f(X)^2\right] - 2\mathbb{E}[f(X)]^2 + \mathbb{E}[f(X)]^2 \\
&= \mathbb{E}\left[f(X)^2\right] - \mathbb{E}[f(X)]^2
\end{aligned}
$$

2.9

For independent random variables it holds that

$$\begin{aligned}
\mathbb{E}[XY] &:= \int xyp(x,y)d(x,y) \\
&= \int\int yxp(x)p(y)dxdy \qquad \text{ind.} \\
&= \int y\left(\int xp(x)dx\right)p(y)dy \\
&= \int xp(x)dx \int yp(y)dy \\
&= \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}$$

and consequently

$$\begin{aligned}
\text{cov}[X,Y] &:= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= 0
\end{aligned}$$

2.10

$$\begin{aligned}
\mathbb{E}[X+Z] &= \int (x+z)p(x,z)d(x,z) \\
&= \int\int (x+z)p(x)p(z)dxdz \qquad \text{ind.} \\
&= \int\int xp(x)p(z)dxdz + \int\int zp(x)p(z)dxdz \\
&= \int \underbrace{\left(\int xp(x)dx\right)}_{=\mathbb{E}[X]} p(z)dz + \int z \underbrace{\left(\int p(x)dx\right)}_{=1} p(z)dz \\
&= \mathbb{E}[X] \underbrace{\int p(z)dz}_{=1} + \int zp(z)dz \\
&= \mathbb{E}[X] + \mathbb{E}[Z]
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}[X+Z] &:= \mathbb{E}[(X+Z)^2] - \mathbb{E}[X+Z]^2 \\
&= \mathbb{E}[X^2 + 2XZ + Z^2] - \mathbb{E}[X+Z]^2 \qquad \text{linearity (cf. above)}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}[X^2] + 2\mathbb{E}[XZ] + \mathbb{E}[Z^2] - (\mathbb{E}[X] + \mathbb{E}[Z])^2 \\
&= \mathbb{E}[X^2] + 2\mathbb{E}[XZ] + \mathbb{E}[Z^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Z] - \mathbb{E}[Z]^2 \quad \text{ind.} \\
&= \mathbb{E}[X^2] \cancel{+ 2\mathbb{E}[X]\mathbb{E}[Z]} + \mathbb{E}[Z^2] - \mathbb{E}[X]^2 \cancel{- 2\mathbb{E}[X]\mathbb{E}[Z]} - \mathbb{E}[Z]^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \\
&= \mathbb{V}[X] + \mathbb{V}[Z]
\end{aligned}
$$

2.11

Expectation

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{E}[X|Y]\right] &= \int_{\mathbb{R}} \int_{\mathbb{R}} x p(x|y) dx \, p(y) dy \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} x \frac{p(x,y)}{p(y)} dx \, p(y) dy \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} x p(x,y) dx \, dy \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} x p(x,y) dy \, dx \qquad \text{Fubini } (\star) \\
&= \int_{\mathbb{R}} x \underbrace{\int_{\mathbb{R}} p(x,y) dy}_{=p(x)} \, dx \\
&= \mathbb{E}[X]
\end{aligned}
$$

($\star$) Assuming X integrable: $\infty > \mathbb{E}\left[|X|\right] = \int_{\mathbb{R}} |x| p(x) dx = \int_{\mathbb{R}} |x| \int_{\mathbb{R}} p(x,y) dy dx = \int_{\mathbb{R}} \int_{\mathbb{R}} |x| p(x,y) dy dx \underset{p \geq 0}{=} \int_{\mathbb{R}} \int_{\mathbb{R}} |x p(x,y)| dy dx$, i.e. $xp(x,y)$ integrable.

Variance

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{V}\left[X|Y\right]\right] + \mathbb{V}\left[\mathbb{E}\left[X|Y\right]\right] &= \mathbb{E}\left[\mathbb{E}\left[X^2|Y\right] - \mathbb{E}\left[X|Y\right]^2\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[X|Y\right]^2\right] - \mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right]^2 \\
&= \mathbb{E}\left[\mathbb{E}\left[X^2|Y\right]\right] \cancel{- \mathbb{E}\left[\mathbb{E}\left[X|Y\right]^2\right]} \\
&\quad \cancel{+ \mathbb{E}\left[\mathbb{E}\left[X|Y\right]^2\right]} - \mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right]^2 \qquad \mathbb{E} \text{ linear} \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 \qquad (\dagger) \\
&= \mathbb{V}\left[X\right]
\end{aligned}
$$

($\dagger$) Assuming X^2 integrable (and consequently X integrable e.g. via Cauchy-Schwarz), use result for expectation from above.

2.14

$$\begin{aligned}
\frac{\partial}{\partial x} p(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
&= p(x)\left(-\frac{\cancel{2}(x-\mu)}{\cancel{2}\sigma^2}\right) \\
&= -\sigma^{-2} p(x)(x-\mu)
\end{aligned}$$

By definition of univariate Gaussian $\sigma^2 > 0$, and consequently $\sigma^{-2} > 0$. Since also $p(x) > 0$ for all x, $\frac{\partial}{\partial x} p(x) = 0$ only if $x = \mu$.

$$\begin{aligned}
\frac{\partial^2}{\partial x^2} p(x) &= -\sigma^{-2}\left(-\sigma^{-2} p(x)(x-\mu)^2 + p(x)\right) \\
&= \sigma^{-2} p(x)\left(\sigma^{-2}(x-\mu)^2 - 1\right)
\end{aligned}$$

$\implies$ $\frac{\partial^2}{\partial x^2} p(\mu) = -\sigma^2 p(\mu) < 0$, so μ maximum of p which means μ mode of univariate Gaussian.

2.15

Mean

$$\begin{aligned}
0 &\overset{!}{=} \frac{\partial}{\partial \mu} \ln p(x|\mu,\sigma^2) \\
&= \frac{\partial}{\partial \mu}\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)\right) \\
&= -\frac{1}{2\sigma^2}\sum_{n=1}^{N} 2(x_n-\mu)(-1) \\
&= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) \\
&= \frac{1}{\sigma^2}\sum_{n=1}^{N} x_n - \frac{N\mu}{\sigma^2} \\
\overset{N>0}{\Longrightarrow} \quad \mu &= \frac{1}{N}\sum_{n=1}^{N} x_n := \tilde{\mu}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2}{\partial\mu^2}\ln p(x|\mu,\sigma^2) &= \frac{\partial}{\partial\mu}\left(\frac{1}{\sigma^2}\sum_{n=1}^{N}x_n - \frac{N\mu}{\sigma^2}\right)\\
&= -\frac{N}{\sigma^2}\\
\implies \quad \frac{\partial^2\ln p}{\partial\mu^2}(\tilde{\mu}) &= -\frac{N}{\sigma^2} \overset{N>0}{<} 0 \implies \tilde{\mu}\text{ maximum}
\end{aligned}$$

Variance

$$\begin{aligned}
0 &\overset{!}{=} \frac{\partial}{\partial\sigma^2}\ln p(x|\mu,\sigma^2)\\
&= \frac{\partial}{\partial\sigma^2}\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)\right)\\
&= -\frac{1}{2\sigma^4}(-1)\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\frac{1}{\sigma^2}\\
\implies \quad 0 &= \sum_{n=1}^{N}(x_n-\mu)^2 - N\sigma^2\\
\overset{N>0}{\implies} \quad \sigma^2 &= \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2 := \tilde{\sigma}^2
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2}{\partial\sigma^4}\ln p(x|\mu,\sigma^2) &= \frac{\partial}{\partial\sigma^2}\left(\frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2\sigma^2}\right)\\
&= \frac{1}{2\sigma^6}(-2)\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2\sigma^4}(-1)\\
&= -\frac{1}{\sigma^6}\sum_{n=1}^{N}(x_n-\mu)^2 + \frac{N}{2\sigma^4}\\
&= -\frac{1}{\sigma^4}\left(\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\right)\\
\implies \frac{\partial^2\ln p}{\partial\sigma^4}(\tilde{\sigma}^2) &= -\frac{1}{\tilde{\sigma}^4}\left(\frac{1}{\tilde{\sigma}^2}N\tilde{\sigma}^2 - \frac{N}{2}\right) = -\frac{N}{2\tilde{\sigma}^4} \overset{N>0}{<} 0 \implies \tilde{\sigma}^2\text{ max.}
\end{aligned}$$

2.16

If $n \neq m$, then by assumption $X_n \perp X_m$. Hence $\mathbb{E}\left[X_n X_m\right] = \mathbb{E}\left[X_n\right]\mathbb{E}\left[X_m\right] \overset{(2.52)}{=} \mu \cdot \mu = \mu^2$. If $n = m$, then $\mathbb{E}\left[X_n X_m\right] = \mathbb{E}\left[X_m^2\right] \overset{(2.53)}{=} \mu^2 + \sigma^2$. Taken together $\mathbb{E}\left[X_n X_m\right] = \mu^2 + \delta_{nm}\sigma^2$.

Mean

$$\begin{aligned}
\mathbb{E}\left[\mu_{ML}\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] && (2.57) \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[X_n] && \text{linearity of expectation} \\
&= \frac{1}{N}\sum_{n=1}^{N}\mu && \text{by assumption} \\
&= \frac{1}{N}(N\mu) \\
&= \mu
\end{aligned}$$

Variance

$$\begin{aligned}
\mathbb{E}\left[\sigma_{ML}^2\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left(X_n - \mu_{ML}\right)^2\right] && (2.58) \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left(X_n - \mu_{ML}\right)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left(X_n - \frac{1}{N}\sum_{m=1}^{N} X_m\right)^2\right] && (2.57) \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[X_n^2 - \frac{2}{N}\sum_{m=1}^{N} X_n X_m + \frac{1}{N^2}\sum_{l=1}^{N}\sum_{k=1}^{N} X_k X_l\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}\left[X_n^2\right] - \frac{2}{N}\sum_{m=1}^{N}\mathbb{E}\left[X_n X_m\right] + \frac{1}{N^2}\sum_{l=1}^{N}\sum_{k=1}^{N}\mathbb{E}\left[X_k X_l\right]\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\Bigg(\mu^2 + \sigma^2 - \frac{2}{N}(1 \cdot (\mu^2 + \sigma^2) + (N-1) \cdot \mu^2) \\
&\qquad + \frac{1}{N^2}(N \cdot (\mu^2 + \sigma^2) + (N^2 - N) \cdot \mu^2)\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\cancel{N}}\cancel{N}\left(\frac{1}{N}(N\mu^2 + N\sigma^2) - \frac{1}{N}(2\sigma^2 + 2N\mu^2) + \frac{1}{N}(\sigma^2 + N\mu^2)\right) \\
&= \left(\frac{N-1}{N}\right)\sigma^2
\end{aligned}
$$

2.17

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}^2\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[(X_n - \mu)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[X_n^2 - 2\mu X_n + \mu^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}\left[X_n^2\right] - 2\mu\mathbb{E}\left[X_n\right] + \mu^2\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2 + \sigma^2 - 2\mu\cdot\mu + \mu^2\right) \\
&= \frac{1}{N}N\sigma^2 \\
&= \sigma^2
\end{aligned}
$$

2.18

Analogous to variance in (2.15).

2.20

$$
\mathrm{J}_g = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 1 + 5\,\mathrm{sech}^2(5x_1) & 0 \\ x_1^2 & 1 + 5\,\mathrm{sech}^2(5x_2) \end{pmatrix}
$$

2.22

$$
f(p) := -\sum_{i=1}^{M} p_i \ln p_i
$$

$$\begin{aligned} g(p) &:= \sum_{i=1}^{M} p_i - 1 \\ L(p, \lambda) &:= f(p) + \lambda g(p) \\ &= -\sum_{i=1}^{M} p_i \ln p_i + \lambda \left(\sum_{i=1}^{M} p_i - 1 \right) \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial p_i} &= -1 \cdot \ln p_i - p_i \cdot \frac{1}{p_i} + \lambda \cdot (1 - 0) \\ &= -\ln p_i - 1 + \lambda \\ \frac{\partial L}{\partial \lambda} &= -\left(\sum_{i=1}^{M} p_i - 1 \right) \\ &= 1 - \sum_{i=1}^{M} p_i \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial p_i} &\overset{!}{=} 0 \\ \Leftrightarrow \quad -\ln p_i - 1 + \lambda &= 0 \\ \Leftrightarrow \quad -\ln p_i &= 1 - \lambda \\ \Leftrightarrow \quad \ln p_i &= \lambda - 1 \\ \Leftrightarrow \quad p_i &= e^{\lambda - 1} \quad (1) \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial \lambda} &\overset{!}{=} 0 \\ \Leftrightarrow \quad 1 - \sum_{i=1}^{M} p_i &= 0 \\ \overset{(1)}{\Rightarrow} \quad 1 - \sum_{i=1}^{M} e^{\lambda - 1} &= 0 \\ \Leftrightarrow \quad 1 - M \cdot e^{\lambda - 1} &= 0 \\ \Leftrightarrow \quad 1 &= M \cdot e^{\lambda - 1} \\ \Leftrightarrow \quad \frac{1}{M} &= e^{\lambda - 1} \end{aligned}$$

Hence for all $i \in \{1, ..., M\}$:

$$p_i = \frac{1}{M}$$

and consequently

$$\begin{aligned}
H[p] &= -\sum_{i=1}^{M} p_i \ln p_i \\
&= -\sum_{i=1}^{M} \frac{1}{M} \ln \frac{1}{M} \\
&= -M \cdot \frac{1}{M} \left(-\ln M\right) \\
&= \ln M
\end{aligned}$$

2.23

$$\begin{aligned}
-\mathrm{H}[X] &= -\sum_{m=1}^{M} p_m \ln\left(\frac{1}{p_m}\right) \\
&= \sum_{m=1}^{M} p_m \left(-\ln\left(\frac{1}{p_m}\right)\right) \\
&\geq -\ln\left(\sum_{m=1}^{M} p_m \frac{1}{p_m}\right) \qquad \text{ln concave} \implies -\ln \text{ convex; Jensen} \\
&= -\ln(M) \\
\implies \mathrm{H}[X] &\leq \ln(M)
\end{aligned}$$

2.25

$$\begin{aligned}
\mathrm{H}[X] &= -\int p(x) \ln p(x) dx \\
&= -\int p(x) \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx \\
&= -\int p(x) \left(\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \ln\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)\right) dx \\
&= -\int p(x) \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) dx - \int p(x) \ln\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx \\
&= -\left(-\frac{1}{2}\right) \ln\left(2\pi\sigma^2\right) \int p(x) dx - \left(-\frac{1}{2\sigma^2}\right) \int (x-\mu)^2 p(x) dx \\
&= \frac{1}{2} \ln\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\sigma^2 \qquad (2.93)\ \&\ (2.95)
\end{aligned}$$

$$= \frac{1}{2}\left(1 + \ln\left(2\pi\sigma^2\right)\right)$$

2.27

$$\begin{aligned}
\mathrm{KL}(p||q) &= \mathbb{E}_p\left[\ln\left(\frac{p(X)}{q(X)}\right)\right] \\
&= \mathbb{E}_p\left[\ln\left(\frac{\frac{1}{\sqrt{\cancel{2\pi}\sigma^2}}e^{-\frac{(X-\mu)^2}{2\sigma^2}}}{\frac{1}{\sqrt{\cancel{2\pi}s^2}}e^{-\frac{(X-m)^2}{2s^2}}}\right)\right] \\
&= \mathbb{E}_p\left[\ln\left(\frac{s}{\sigma}\cdot e^{-\frac{(X-\mu)^2}{2\sigma^2}-\left(-\frac{(X-m)^2}{2s^2}\right)}\right)\right] \\
&= \mathbb{E}_p\left[\ln\left(\frac{s}{\sigma}\right) + \frac{(X-m)^2}{2s^2} - \frac{(X-\mu)^2}{2\sigma^2}\right] \\
&= \ln\left(\frac{s}{\sigma}\right) + \frac{\mathbb{E}_p\left[X^2\right] - 2m\mathbb{E}_p[X] + m^2}{2s^2} - \frac{\mathbb{E}_p\left[(X-\mu)^2\right]}{2\sigma^2} \\
&= \ln\left(\frac{s}{\sigma}\right) + \frac{(\sigma^2+\mu^2) - 2m\mu + m^2}{2s^2} - \frac{\cancel{\sigma^2}}{2\cancel{\sigma^2}} \\
&= \ln\left(\frac{s}{\sigma}\right) + \frac{\sigma^2 + (\mu-m)^2}{2s^2} - \frac{1}{2}
\end{aligned}$$

2.30

$$\begin{aligned}
\mathrm{H}(Y) &= -\mathbb{E}\left[\ln p_Y(Y)\right] \\
&= -\mathbb{E}\left[\ln p_Y(AX)\right] \\
&\overset{\star}{=} -\mathbb{E}\left[\ln\left(p_X\left(A^{-1}AX\right)\left|\det A^{-1}\right|\right)\right] \\
&= -\mathbb{E}\left[\ln p_X\left(X\right) + \ln\left|\det A\right|^{-1}\right] \\
&= -\mathbb{E}\left[\ln p_X(X)\right] - \mathbb{E}\left[-\ln\left|\det A\right|\right] \\
&= \mathrm{H}(X) + \ln\left|\det A\right|
\end{aligned}$$

$\star$ By assumption, A invertible, i.e. A^{-1} exists. Transformation of densities via $g(x) := Ax$.