Solution

Problem 1

1. **Idempotence:** Show that $P^2 = P$.

Solution:

To show $P^2 = P$, expand P^2:

$$P^2 = \left(\Phi(\Phi^T\Phi)^{-1}\Phi^T\right)^2.$$

Using the associative property of matrix multiplication, we rewrite:

$$P^2 = \Phi(\Phi^T\Phi)^{-1}\Phi^T\Phi(\Phi^T\Phi)^{-1}\Phi^T.$$

Since $\Phi^T\Phi(\Phi^T\Phi)^{-1} = I_n$ (the identity matrix), this simplifies to:

$$P^2 = \Phi(\Phi^T\Phi)^{-1}\Phi^T = P.$$

Thus, P is idempotent.

2. **Symmetry:** Show that $P^T = P$.

Solution:

The transpose of P is:

$$P^T = \left(\Phi(\Phi^T\Phi)^{-1}\Phi^T\right)^T.$$

Using the property $(AB)^T = B^TA^T$, we have:

$$P^T = (\Phi^T)^T\left((\Phi^T\Phi)^{-1}\right)^T\Phi^T.$$

Since $\Phi^T\Phi$ is symmetric ($(\Phi^T\Phi)^T = \Phi^T\Phi$), $(\Phi^T\Phi)^{-1}$ is also symmetric. Additionally, $(\Phi^T)^T = \Phi$. Thus:

$$P^T = \Phi(\Phi^T\Phi)^{-1}\Phi^T = P.$$

Therefore, P is symmetric.

3. **Projection Properties:** Verify $\Phi^T(v - Pv) = 0$.

Solution:

Compute $v - Pv$:

$$v - Pv = v - \Phi(\Phi^T\Phi)^{-1}\Phi^Tv.$$

Multiply by Φ^T to check orthogonality:

$$\Phi^T(v - Pv) = \Phi^Tv - \Phi^T\Phi(\Phi^T\Phi)^{-1}\Phi^Tv.$$

Using $\Phi^T\Phi(\Phi^T\Phi)^{-1} = I_n$, this simplifies to:

$$\Phi^T(v - Pv) = \Phi^Tv - \Phi^Tv = 0.$$

Thus, $v - Pv$ is orthogonal to the columns of Φ.

Problem 2

1. **Solution:**

(a) From the definition of $E[X|Y]$, we can express the expectation of X as:

$$\begin{aligned}
E[E[X|Y]] &= E\left[\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y)\, dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\, dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\, dxdy \\
&= \int_{-\infty}^{\infty} x f_X(x)\, dx \\
&= E[X]
\end{aligned}$$

(b) If X and Y are independent, then the joint density function $f_{X,Y}(x,y)$ is given by:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y),$$

and the conditional density function becomes:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = f_X(x).$$

Using this, the conditional expectation simplifies as follows:

$$E[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx = \int_{-\infty}^{\infty} x f_X(x)\, dx = E[X].$$

Since $E[X|Y]$ does not depend on Y, we conclude that:

$$E[X|Y] = E[X].$$

(c)

$$\begin{aligned}
E[E[X|Y,Z]|Y] &= \int_{-\infty}^{\infty} E[X|Y=y, Z=z] f_{Z|Y}(z|y)\, dz \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x \cdot f_{X|Y,Z}(x|y,z)\, dx\right) f_{Z|Y}(z|y)\, dz \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f_{X|Y,Z}(x|y,z) f_{Z|Y}(z|y)\, dx\, dz \\
&= \int_{-\infty}^{\infty} x \cdot \frac{\int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\, dz}{f_Y(y)}\, dx \\
&= \int_{-\infty}^{\infty} x \cdot \frac{f_{X,Y}(x,y)}{f_Y(y)}\, dx \\
&= \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)\, dx \\
&= E[X|Y=y].
\end{aligned}$$

2. The Kullback-Leibler (KL) Divergence $D_{KL}(P\|Q)$ between two distributions $P(x)$ and $Q(x)$ is defined as:

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} P(x) \log \frac{P(x)}{Q(x)}\, dx.$$

What's the KL Divergence between two Gaussian distribution $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$?

Solution for Univariate Gaussian:

For the two Gaussian distributions:

$$P(x) = \mathcal{N}(\mu_1, \sigma_1^2),$$
$$Q(x) = \mathcal{N}(\mu_2, \sigma_2^2),$$

their probability density functions are:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right),$$
$$Q(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right).$$

Substituting $P(x)$ and $Q(x)$ into the KL divergence definition:

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} P(x) \left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)}\right] dx.$$

Simplify the logarithm term:

$$\log \frac{P(x)}{Q(x)} = \log \frac{\sqrt{2\pi\sigma_2^2}}{\sqrt{2\pi\sigma_1^2}} - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}.$$

Separate the integral into three parts:

$$D_{KL}(P\|Q) = \underbrace{\int_{-\infty}^{\infty} P(x) \log \frac{\sigma_2}{\sigma_1} dx}_{(A)} + \underbrace{\int_{-\infty}^{\infty} P(x) \left(\frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}\right) dx}_{(B)}.$$

For (A), we know:

$$\int_{-\infty}^{\infty} P(x) \log \frac{\sigma_2}{\sigma_1} dx = \log \frac{\sigma_2}{\sigma_1},$$

since $\log \frac{\sigma_2}{\sigma_1}$ is independent of x and integrates to 1 over $P(x)$.

For (B), by the definition of variance, we know:

$$\mathbb{E}_{x\sim P(x)}[(x-\mu_1)^2] = \sigma_1^2.$$

Substitute and simplify:

$$\int_{-\infty}^{\infty} P(x) \frac{(x-\mu_1)^2}{2\sigma_1^2} dx = \frac{\sigma_1^2}{2\sigma_1^2} = \frac{1}{2}.$$

For the cross-term:

$$
\begin{aligned}
\int_{-\infty}^{\infty} P(x)(x-\mu_2)^2 dx &= \mathbb{E}[(x-\mu_2)^2] \\
&= \mathbb{E}[(x-\mathbb{E}[x]+\mathbb{E}[x]-\mu_2)^2] \\
&= \mathbb{E}[(x-\mathbb{E}[x])^2] + 2\mathbb{E}[(x-\mathbb{E}[x])(\mathbb{E}[x]-\mu_2)] + \mathbb{E}[(\mathbb{E}[x]-\mu_2)^2] \\
&= \mathrm{Var}(x) + (\mathbb{E}[x]-\mu_2)^2 \\
&= \sigma_1^2 + (\mu_1-\mu_2)^2,
\end{aligned}
$$

where the third equality follows from $\mathbb{E}[x-\mathbb{E}[x]] = 0$.

Combine all terms:

$$
D_{KL}(P\|Q) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.
$$

Solution for Multivariate Gaussian: Suppose the covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$.Following the definition of Multivariate Gaussian, we have:

$$
\ln\frac{P(\mathbf{x})}{Q(\mathbf{x})} = \ln\Bigg(\sqrt{\frac{\det(\Sigma_2)}{\det(\Sigma_1)}}\Bigg) - \frac{1}{2}(\mathbf{x}-\mu_1)^\top\Sigma_1^{-1}(\mathbf{x}-\mu_1) + \frac{1}{2}(\mathbf{x}-\mu_2)^\top\Sigma_2^{-1}(\mathbf{x}-\mu_2).
$$

Thus,

$$
\begin{aligned}
D_{\mathrm{KL}}(P\|Q) &= \int P(\mathbf{x})\left[\frac{1}{2}\ln\Big(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\Big) - \frac{1}{2}(\mathbf{x}-\mu_1)^\top\Sigma_1^{-1}(\mathbf{x}-\mu_1) + \frac{1}{2}(\mathbf{x}-\mu_2)^\top\Sigma_2^{-1}(\mathbf{x}-\mu_2)\right] d\mathbf{x} \\
&= \frac{1}{2}\ln\Big(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\Big) - \underbrace{\int P(\mathbf{x})\frac{1}{2}(\mathbf{x}-\mu_1)^\top\Sigma_1^{-1}(\mathbf{x}-\mu_1)d\mathbf{x}}_{(A)} + \underbrace{\int P(\mathbf{x})\frac{1}{2}(\mathbf{x}-\mu_2)^\top\Sigma_2^{-1}(\mathbf{x}-\mu_2)d\mathbf{x}}_{B}.
\end{aligned}
$$

For (A), note that

$$
\mathbb{E}[(\mathbf{x}-\mu_1)(\mathbf{x}-\mu_1)^\top] = \Sigma_1.
$$

Moreover, we know that

$$
\mathbb{E}[\mathrm{tr}(A)] = \mathrm{tr}(\mathbb{E}[A]).
$$

(Please refer to https://statproofbook.github.io/P/mean-tr.html for more detail.)

Therefore,

$$
\begin{aligned}
(A) &= \frac{1}{2}\mathbb{E}[\mathrm{tr}(\mathbf{x}-\mu_1)^\top\Sigma_1^{-1}(\mathbf{x}-\mu_1)] \\
&= \frac{1}{2}\mathbb{E}[\Sigma_1^{-1}\mathrm{tr}(\mathbf{x}-\mu_1)^\top(\mathbf{x}-\mu_1)] \\
&= \frac{1}{2}\mathrm{tr}(\mathbb{E}[\Sigma_1^{-1}(\mathbf{x}-\mu_1)^\top(\mathbf{x}-\mu_1)]) \\
&= \frac{1}{2}\mathrm{tr}(\mathbb{E}[\Sigma_1^{-1}\Sigma_1]) \\
&= \frac{1}{2}\mathrm{tr}(\mathbf{1}_d) = \frac{d}{2}.
\end{aligned}
$$

For (B), note that

$$\begin{aligned}\mathbb{E}[(\mathbf{x}-\mu_2)(\mathbf{x}-\mu_2)^\top] &= \mathbb{E}[(\mathbf{x}-\mu_1+\mu_1-\mu_2)(\mathbf{x}-\mu_1+\mu_1-\mu_2)^\top] \\ &= \mathbb{E}[(\mathbf{x}-\mu_1)(\mathbf{x}-\mu_1)^\top] + (\mu_1-\mu_2)(\mu_1-\mu_2)^\top \\ &= \Sigma_1 + (\mu_1-\mu_2)(\mu_1-\mu_2)^\top.\end{aligned}$$

Thus,

$$\begin{aligned}(B) &= \frac{1}{2}\mathrm{tr}(\mathbb{E}[\Sigma_2^{-1}(\mathbf{x}-\mu_2)^\top(\mathbf{x}-\mu_2)]) \\ &= \frac{1}{2}\mathrm{tr}([\Sigma_2^{-1}(\Sigma_1 + (\mu_1-\mu_2)(\mu_1-\mu_2)^\top)]) \\ &= \frac{1}{2}\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + \frac{1}{2}(\mu_1-\mu_2)^\top\Sigma_2^{-1}(\mu_1-\mu_2).\end{aligned}$$

Putting these pieces together, we get

$$\begin{aligned}D_{\mathrm{KL}}(P\|Q) &= \frac{1}{2}\ln\Big(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\Big) - \frac{1}{2}d + \frac{1}{2}\,\mathrm{tr}\big(\Sigma_2^{-1}\Sigma_1\big) + \frac{1}{2}\,(\mu_1-\mu_2)^\top\Sigma_2^{-1}(\mu_1-\mu_2) \\ &= \frac{1}{2}\Big[\mathrm{tr}\big(\Sigma_2^{-1}\Sigma_1\big) + (\mu_2-\mu_1)^\top\Sigma_2^{-1}(\mu_2-\mu_1) - d + \ln\Big(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\Big)\Big].\end{aligned}$$

Problem 3 We are tasked with minimizing the function:

$$F(x, y) = y + (y - x)^2.$$

1. **Compute the Gradient $\nabla F(x, y)$:**

The gradient $\nabla F(x, y) = \begin{bmatrix} \frac{\partial F}{\partial x} \\ \frac{\partial F}{\partial y} \end{bmatrix}$ is calculated as follows:

1. Partial derivative with respect to x:

$$\frac{\partial F}{\partial x} = \frac{\partial}{\partial x}\left(y + (y - x)^2\right) = \frac{\partial}{\partial x}(y) + \frac{\partial}{\partial x}\left((y - x)^2\right).$$

Since y is independent of x, its derivative is 0, and:

$$\frac{\partial}{\partial x}\left((y - x)^2\right) = 2(y - x)(-1) = -2(y - x).$$

Thus:

$$\frac{\partial F}{\partial x} = -2(y - x).$$

2. Partial derivative with respect to y:

$$\frac{\partial F}{\partial y} = \frac{\partial}{\partial y}\left(y + (y - x)^2\right) = \frac{\partial}{\partial y}(y) + \frac{\partial}{\partial y}\left((y - x)^2\right).$$

The derivative of y is 1, and:

$$\frac{\partial}{\partial y}\left((y - x)^2\right) = 2(y - x)(1) = 2(y - x).$$

Thus:

$$\frac{\partial F}{\partial y} = 1 + 2(y - x).$$

The gradient is:

$$\nabla F(x, y) = \begin{bmatrix} -2(y - x) \\ 1 + 2(y - x) \end{bmatrix}.$$

2. **Evaluate $\nabla F(x, y)$ at $(x_0, y_0) = (1, 1)$:**

Substituting $x = 1$ and $y = 1$:

$$\nabla F(1, 1) = \begin{bmatrix} -2(1 - 1) \\ 1 + 2(1 - 1) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

3. **Perform the Gradient Descent Update:**

The gradient descent update rule is:

$$(x_{k+1}, y_{k+1}) = (x_k, y_k) - s\nabla F(x_k, y_k).$$

Substituting $s = \frac{1}{2}$, $(x_0, y_0) = (1, 1)$, and $\nabla F(1, 1) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$:

$$(x_1, y_1) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Simplify:

$$(x_1, y_1) = \begin{bmatrix} 1 \\ 1 - \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix}.$$

4. **Final Answer:**

After one step of gradient descent, the updated point is:

$$(x_1, y_1) = (1, \frac{1}{2}).$$