# Exercise 1

**Given:**

$$\boldsymbol{\theta} \sim \mathcal{N}_2(\boldsymbol{m_0}, P_0), \quad p(\boldsymbol{y}|\boldsymbol{\theta}) \sim \mathcal{N}_n(H\boldsymbol{\theta}, \sigma^2 I_N) \tag{1}$$

**To be Proved:**

$$\boldsymbol{\theta}|\boldsymbol{y} \sim N_2(\nu, C) \tag{2}$$

where

$$C^{-1} = H'\sigma^{-2}H + P_0^{-1}$$

and

$$\nu = C(H'\sigma^{-2}\boldsymbol{y} + P_0^{-1}\boldsymbol{m_0})$$

---

Using Bayes' Rule, we can write the posterior distribution from 2 as:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

We can rewrite the right-hand side of the above proportionality as a product of two Gaussians with mean and co-variance from 1:

$$\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{y}) &\propto \frac{1}{\sqrt{2\pi}|\sigma^2\boldsymbol{I}|^{1/2}} \exp(-\frac{1}{2}(\boldsymbol{y} - H\boldsymbol{\theta})'(\sigma^2\boldsymbol{I_n})^{-1}(\boldsymbol{y} - H\boldsymbol{\theta})) \\
&\quad \times \frac{1}{\sqrt{2\pi}|P_0|^{1/2}} \exp(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{m_0})'P_0^{-1}(\boldsymbol{\theta} - \boldsymbol{m_0})) \\
&\propto \exp(-\frac{1}{2\sigma^2}(\boldsymbol{y} - H\boldsymbol{\theta})'(\boldsymbol{y} - H\boldsymbol{\theta})) \times \exp(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{m_0})'P_0^{-1}(\boldsymbol{\theta} - \boldsymbol{m_0})) \\
&\propto \exp(-\frac{1}{2}(\frac{1}{\sigma^2}(\boldsymbol{y} - H\boldsymbol{\theta})'(\boldsymbol{y} - H\boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{m_0})'P_0^{-1}(\boldsymbol{\theta} - \boldsymbol{m_0}))) \\
&\propto \exp(-\frac{1}{2}(-\frac{2}{\sigma^2}\boldsymbol{y}'H\boldsymbol{\theta} + \frac{1}{\sigma^2}\boldsymbol{\theta}'H'H\boldsymbol{\theta} + \boldsymbol{\theta}'P_0^{-1}\boldsymbol{\theta} - 2\boldsymbol{m_0}'P_0^{-1}\boldsymbol{\theta})) \qquad (3)
\end{aligned}$$

In the above process, we simplify by removing terms that do not have $\boldsymbol{\theta}$. Comparing 3 to a generic normal distribution $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{C})$:

$$\begin{aligned}
\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{C}) &\propto \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\nu})'\boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{\nu})) \\
&\propto \exp(-\frac{1}{2}(\boldsymbol{x}'\boldsymbol{C}^{-1}\boldsymbol{x} - 2\boldsymbol{\nu}'\boldsymbol{C}^{-1}\boldsymbol{x})) \qquad (4)
\end{aligned}$$

We can rewrite equation 3 in the form of a generic normal distribution as given in equation 4. To do this, we equate the quadratic terms to get the mean and the linear terms to get the covariance.

Equating the quadratic terms:

$$\begin{aligned}
\boldsymbol{x}'\boldsymbol{C}^{-1}\boldsymbol{x} &= \frac{1}{\sigma^2}\boldsymbol{\theta}'H'H\boldsymbol{\theta} + \boldsymbol{\theta}'P_0^{-1}\boldsymbol{\theta} \\
&= \boldsymbol{\theta}'(\frac{1}{\sigma^2}H'H + P_0^{-1})\boldsymbol{\theta}
\end{aligned}$$

Hence, the covariance in the new distribution is:

$$\boldsymbol{C} = (\frac{1}{\sigma^2} H'H + P_0^{-1})^{-1} \tag{5}$$

Now, we equate the linear terms to get $\boldsymbol{\nu}$:

$$\begin{aligned}
-2\boldsymbol{\nu}'\boldsymbol{C}^{-1}\boldsymbol{x} &= -\frac{2}{\sigma^2}\boldsymbol{y}'H\boldsymbol{\theta} - 2\boldsymbol{m_0}'P_0^{-1}\boldsymbol{\theta} \\
\boldsymbol{\nu}'\boldsymbol{C}^{-1} &= \frac{1}{\sigma^2}\boldsymbol{y}'H + \boldsymbol{m_0}'P_0^{-1} \\
\boldsymbol{\nu}'\boldsymbol{C}^{-1}\boldsymbol{C} &= (\frac{1}{\sigma^2}\boldsymbol{y}'H + \boldsymbol{m_0}'P_0^{-1})\boldsymbol{C} \\
\boldsymbol{\nu}' &= (\frac{1}{\sigma^2}\boldsymbol{y}'H + \boldsymbol{m_0}'P_0^{-1})\boldsymbol{C} \\
\boldsymbol{\nu} &= \boldsymbol{C}(\frac{1}{\sigma^2}H'\boldsymbol{y} + P_0^{-1}\boldsymbol{m_0}) \qquad (6)
\end{aligned}$$

Hence, we can write the distribution of $\boldsymbol{\theta}|\boldsymbol{y}$ as a Gaussian with mean and covariance from 6 and 5:

$$\boldsymbol{\theta}|\boldsymbol{y} \sim \mathcal{N}_2(\boldsymbol{\nu}, \boldsymbol{C})$$

where

$$\boldsymbol{\nu} = \boldsymbol{C}(\frac{1}{\sigma^2}H'\boldsymbol{y} + P_0^{-1}\boldsymbol{m_0})$$

and

$$\boldsymbol{C^{-1}} = \frac{1}{\sigma^2}H'H + P_0^{-1}$$

which proves 2.

# Exercise 2

**Given:**

$$P_k^{-1} = H_k' \sigma^{-2} H_k + P_{k-1}^{-1} \tag{7}$$

**Woodbury Matrix Identity:**

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(c^{-1} + VA^{-1}U)^{-1}VA^{-1} \tag{8}$$

**To be Proved:**

$$P_k = P_{k-1} - P_{k-1}H_k'(\sigma^2 + H_k P_{k-1} H_k')^{-1} H_k P_{k-1} \tag{9}$$

---

Substituting $A = P_{k-1}^{-1}, C = \sigma^{-2}, U = H_k', V = H_k$ in 8, we get:

$$(H_k' \sigma^{-2} H_k + P_{k-1}^{-1})^{-1} = P_{k-1} - P_{k-1}H_k'(\sigma^2 + H_k P_{k-1} H_k')^{-1} H_k P_{k-1}$$

Comparing the above equation to 7, the LHS is just $P_k^{-1}$, so we get:

$$\begin{aligned}(P_k^{-1})^{-1} &= P_{k-1} - P_{k-1}H_k'(\sigma^2 + H_k P_{k-1} H_k')^{-1} H_k P_{k-1} \\ P_k &= P_{k-1} - P_{k-1}H_k'(\sigma^2 + H_k P_{k-1} H_k')^{-1} H_k P_{k-1}\end{aligned}$$

which proves 9.

---

Using Woodbury Matrix Identity is numerically advantageous to compute $P_k$ because the inverse term is a scalar, which is trivial to invert(compared to the inverse in the original equation for $P_k$). Let us verify this by checking the dimensions of each term of the inverse in equation 9:

$$H_k \in \mathbb{R}^{1\times k}, P_{k-1} \in \mathbb{R}^{k\times k}, \sigma^2 \in \mathbb{R}$$

$$\therefore (\sigma^2 + H_k P_{k-1} H_k') \in \mathbb{R}^1$$

Comparing this to the inverse term in the original equation for $P_k$, as described in 7:

$$H_k' \sigma^{-2} H_k + P_{k-1}^{-1} \in \mathbb{R}^{k\times k}$$

Computing this can be non-trivial for large values of $k$. Hence, it would be numerically advantageous to use the Woodbury Matrix Identity to compute the updated co-variance matrix.

# Exercise 3

1. The code that computes the posterior distribution is in the Appendix at the end of the report.

2. Posterior Means and Co-Variance Matrices:

   **After 10 observations:**

$$m_{10} = \begin{bmatrix} 40067.19 \\ 233.92 \\ 663.49 \\ 23.29 \end{bmatrix}, P_{10} = \begin{bmatrix} 98.95 & -2.11 & -4.00 & 0.31 \\ -2.11 & 94.71 & -12.06 & 0.05 \\ -4.00 & -12.06 & 67.48 & -7.80 \\ 0.31 & 0.05 & -7.80 & 1.02 \end{bmatrix}$$

   **After full dataset:**

$$m_{30} = \begin{bmatrix} 40072.17 \\ 355.67 \\ 1324.71 \\ -53.82 \end{bmatrix} P_{30} = \begin{bmatrix} 98.74 & -2.73 & -5.02 & 0.43 \\ -2.73 & 91.88 & -21.13 & 1.17 \\ -5.02 & -21.13 & 26.78 & -2.63 \\ 0.43 & 1.17 & -2.63 & 0.28 \end{bmatrix}$$

3. Figure 1 shows the fitted curve in red, superimposed on the observed data (shown as a scatter plot). As we can see, the curve fits the observed data quite well.

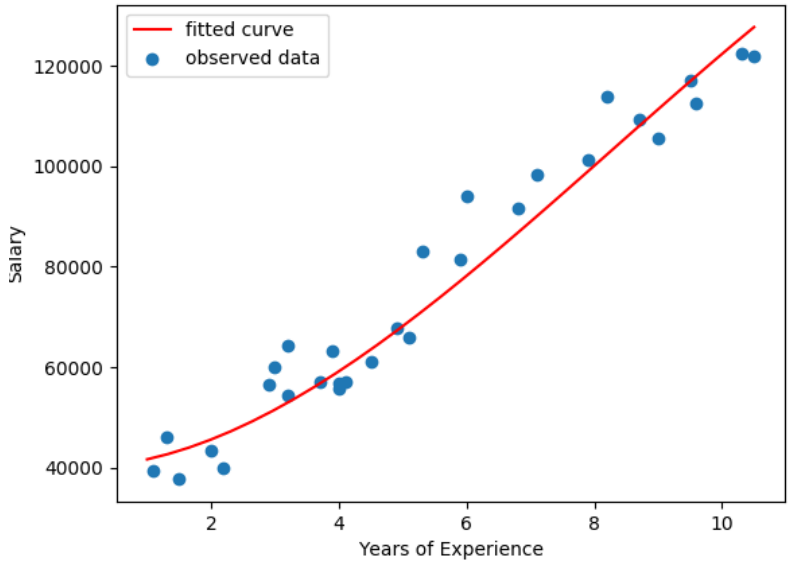


Figure 1: Fitted curve superimposed on the observed data

# Exercise 4

**Given:**

$$(\theta_{k-1}|y_{1:k-1}) \sim \mathcal{N}(m_{k-1}, P_{k-1}), \quad (\theta_k|\theta_{k-1}) \sim \mathcal{N}(A\theta_{k-1}, \theta) \tag{10}$$

and **Lemma 1:** If the random vectors $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathbb{R}^m$ satisfy

$$X \sim \mathcal{N}(m, P)$$
$$Y|X \sim \mathcal{N}(Hx + u, R)$$

then

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} m \\ Hm + u \end{bmatrix}, \begin{bmatrix} P & PH' \\ HP & HPH' + R \end{bmatrix}) \tag{11}$$

**To be Proved:**

$$\theta_k|y_{1:k-1} \sim \mathcal{N}(m_k^-, P_k^-)$$

where

$$P_k^- = AP_{k-1}A' + Q, \quad m_k^- = Am_{k-1} \tag{12}$$

---

The distribution $P(\theta_k|y_{1:k-1})$ can be written as a marginalization over $\theta_{k-1}|y_{1:k}$:

$$P(\theta_k|y_{1:k-1}) = \int P(\theta_k, \theta_{k-1}|y_{1:k-1}) d(\theta_{k-1}|y_{1:k-1})$$

Since $\theta_k$ is independent of $\theta_{k-1}$ given previous observations $y_{1:k-1}$:

$$P(\theta_k, \theta_{k-1}|y_{1:k-1}) = P(\theta_k|\theta_{k-1}, y_{1:k-1}) P(\theta_{k-1}|y_{1:k-1})$$

Since $\theta_k$ is independent of previous observations $y_{1:k-1}$ given $\theta_{k-1}$:

$$P(\theta_k, \theta_{k-1}|y_{1:k-1}) = P(\theta_k|\theta_{k-1}) P(\theta_{k-1}|y_{1:k-1})$$

Now, to represent the joint distribution of $(\theta_k, \theta_{k-1})|y_{1:k-1}$ as a Multivariate Gaussian, we use Lemma 1.

We can see clearly that $(\theta_{k-1}|y_{1:k-1})$ and $(\theta_k|\theta_{k-1})$ from 10 have a distribution of the same form as $X$ and $Y$ from Lemma 1.

Rewriting 11 by substituting from 10, we get:

$$\begin{bmatrix} \theta_{k-1}|y_{1:k-1} \\ \theta_k|\theta_{k-1} \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} m_{k-1} \\ Am_{k-1} \end{bmatrix}, \begin{bmatrix} P_{k-1} & P_{k-1}A' \\ AP_{k-1} & AP_{k-1}A' + Q \end{bmatrix})$$

Using the property that *The marginal of a Joint-Gaussian is a Gaussian*, we can marginalize the above joint Gaussian into its component distribution to get $\theta_k|\theta_{k-1}$:

$$\theta_k|\theta_{k-1} \sim \mathcal{N}(Am_{k-1}, AP_{k-1}A' + Q)$$

which proves 12.

# Appendix: Code for exercise 3

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt


def get_feature_matrix(x):
    """
    Transform the input into a feature matrix of order 3
    """
    return np.stack((x**0, x**1, x**2, x**3), axis=1)

def compute_var(var, x, sd):
    """
    update co-variance(Pk) using a single observation.
    """
    return np.linalg.inv((1/sd**2)*x.T.dot(x) + np.linalg.inv(var))

def compute_var_woodbury(var, x, sd):
    """
    Update co-variance(Pk) using a single observation.
    Implements the woodbury matrix identity to simplify the inverse.
    """
    return var - var.dot(x.T).dot((x.dot(var).dot(x.T) + sd**2)**-1).dot(x).dot(var)

def compute_mean(mean, x, sd, y, var_new, var_old):
    """
    Update the mean(mk) using a single observation.
    """
    return var_new.dot((1/sd**2) * x.T.dot(y) + np.linalg.inv(var_old).dot(mean))

def predict(x, model_params):
    """
    Predict the salary, given a new observation and updated model parameters (i.e. theta)
    """
    y = x.dot(model_params)
    return y

# Read the CSV into a pandas dataframe
df = pd.read_csv("./SalaryData.csv")

# Transform vector of "years of experience" into
# a feature matrix with 1st, 2nd and 3rd order terms.
X = get_feature_matrix(df["YearsExperience"])
# Get the salary data from dataframe
Y = df[["Salary"]].to_numpy()
```

```
46
47 # initialize standard deviation of measurement
48 sd_measurement = 250
49 # initialize prior mean and co-variance
50 m0 = np.array((40000, 0, 0, 0))
51 p0 = np.array((100, 100, 100, 100))*np.eye(4)
52
53 var_old = p0
54 mean_old = m0
55 # update posterior mean and co-variance, one observation at a time
56 for i in range(0, X.shape[0]):
57     # fetch the ith input and output observations
58     x = np.expand_dims(X[i], axis=0)
59     y = Y[i]
60
61     # compute the co-variance and mean using ith observation
62     var_new = compute_var_woodbury(var_old, x, sd_measurement)
63     mean_new = compute_mean(mean_old, x, sd_measurement, y, var_new,
    var_old)
64
65     # store current co-variance and mean for the next update
66     var_old = var_new
67     mean_old = mean_new
68
69 # set the final mean and co-variance as the model mean and co-variance
70 model_mean = mean_old
71 model_var = var_old
72
73 # scatter plot for the observed data inputs vs. outputs
74 plt.scatter(X[:,1], Y, marker="o", label="observed data")
75
76 # create a set of new inputs (Years of Working Experience)
77 # in the same domain as the observations
78 X_new = get_feature_matrix(np.linspace(1, 10.5, 30))
79 y_new = np.empty((X_new.shape[0], ), dtype=np.float32)
80
81 # For the set of new inputs, we predict
82 # the output(Salary) using the predict() method
83 for i in range(0, y_new.shape[0]):
84     # Sample from a multivariate gaussian with model mean and co-variance
85     model_params = np.random.multivariate_normal(model_mean, model_var)
86     # Predict salary for a new value of "years of experience"
87     y_new[i] = predict(X_new[i], model_params)
88
89 # Plot curve of the new observations vs. the model outputs
90 plt.plot(X_new[:,1], y_new, c="r", label="fitted curve")
91 plt.xlabel("Years of Experience")
92 plt.ylabel("Salary")
93 plt.legend()
94 plt.show()
```