Exercise of Supervised Learning: Gaussian Processes 1

Yawei Li

yawei.li@stat.uni-muenchen.de

January 28, 2025

Exercise 1: Bayesian Linear Model

In the Bayesian linear model, we assume that the data follows the following law:

$$y = f(\mathbf{x}) + \epsilon = \boldsymbol{\theta}^\top \mathbf{x} + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and independent of $\mathbf{x}$. On the data-level this corresponds to

$$y^{(i)} = f\left(\mathbf{x}^{(i)}\right) + \epsilon^{(i)} = \boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \epsilon^{(i)}, \quad \text{for } i \in [n],$$

where $\epsilon^{(i)} \in \mathcal{N}(0, \sigma^2)$ are i.i.d. and all independent of $\mathbf{x}^{(i)}$'s. In the Bayesian perspective it is assumed that the parameter vector $\boldsymbol{\theta}$ is stochastic and follows a distribution. Assume we are interested in the so-called maximum a posteriori estimate of $\boldsymbol{\theta}$, which is defined by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} | \mathbf{X}, \mathbf{y}).$$

(a) Show that if we choose a **uniform distribution** over the parameter vector $\boldsymbol{\theta}$ as the prior belief, i.e., $q(\boldsymbol{\theta}) \propto 1$, then the maximum a posteriori estimate coincides with the **empirical risk minimizer for the L2-loss** (over linear models).

1 (a): Construct Posterior from Bayes' Rule

$$\underbrace{p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})}_{\text{posterior}} = \frac{p(\boldsymbol{\theta}, \mathbf{y}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})}$$

- For a linear model, $\mathbf{y}|\mathbf{X}, \theta \sim \mathcal{N}(\mathbf{X}\theta, \sigma^2 \boldsymbol{I})$. Will be computed on the next slide.
- In 1(a), we choose a **uniform prior**, indicating $q(\theta) \propto 1$.
- $p(\mathbf{y}|\mathbf{X})$ does **not** depend on θ. $\rightsquigarrow$ Treated as constant when maxing the posterior of θ.

1 (a): Construct Posterior from Bayes' Rule

$$\underbrace{p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})}_{\text{posterior}} = \frac{p(\boldsymbol{\theta},\mathbf{y}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})}$$

$$= \frac{\overbrace{p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})}^{\text{likelihood}}\ \overbrace{q(\boldsymbol{\theta})}^{\textit{prior}}}{\underbrace{p(\mathbf{y}|\mathbf{X})}_{\text{marginal}}}$$

- For a linear model, $\mathbf{y}|\mathbf{X},\theta \sim \mathcal{N}(\mathbf{X}\theta, \sigma^2 \boldsymbol{I})$. Will be computed on the next slide.
- In 1(a), we choose a **uniform prior**, indicating $q(\theta) \propto 1$.
- $p(\mathbf{y}|\mathbf{X})$ does **not** depend on θ. $\rightsquigarrow$ Treated as constant when maxing the posterior of θ.

1 (a): Construct Posterior from Bayes' Rule

$$\underbrace{p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})}_{\text{posterior}} = \frac{p(\boldsymbol{\theta},\mathbf{y}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})}$$

$$= \frac{\overbrace{p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})}^{\text{likelihood}}\,\overbrace{q(\boldsymbol{\theta})}^{\textit{prior}}}{\underbrace{p(\mathbf{y}|\mathbf{X})}_{\text{marginal}}}$$

- For a linear model, $\mathbf{y}|\mathbf{X},\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta},\sigma^2\boldsymbol{I})$. Will be computed on the next slide.
- In 1(a), we choose a **uniform prior**, indicating $q(\theta) \propto 1$.
- $p(\mathbf{y}|\mathbf{X})$ does **not** depend on θ. $\rightsquigarrow$ Treated as constant when maxing the posterior of θ.

1 (a): Construct Posterior from Bayes' Rule

$$\underbrace{p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})}_{\text{posterior}} = \frac{p(\boldsymbol{\theta},\mathbf{y}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})}$$

$$= \frac{\overbrace{p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})}^{\text{likelihood}}\ \overbrace{q(\boldsymbol{\theta})}^{\textit{prior}}}{\underbrace{p(\mathbf{y}|\mathbf{X})}_{\text{marginal}}}$$

- For a linear model, $\mathbf{y}|\mathbf{X},\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta},\sigma^2\boldsymbol{I})$. Will be computed on the next slide.
- In 1(a), we choose a **uniform prior**, indicating $q(\boldsymbol{\theta}) \propto 1$.
- $p(\mathbf{y}|\mathbf{X})$ does **not** depend on θ. $\rightsquigarrow$ Treated as constant when maxing the posterior of θ.

1 (a): Construct Posterior from Bayes' Rule

$$\underbrace{p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})}_{\text{posterior}} = \frac{p(\boldsymbol{\theta},\mathbf{y}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})}$$

$$= \frac{\overbrace{p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})}^{\text{likelihood}}\ \overbrace{q(\boldsymbol{\theta})}^{\textit{prior}}}{\underbrace{p(\mathbf{y}|\mathbf{X})}_{\text{marginal}}}$$

- For a linear model, $\mathbf{y}|\mathbf{X},\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$. Will be computed on the next slide.
- In 1(a), we choose a **uniform prior**, indicating $q(\boldsymbol{\theta}) \propto 1$.
- $p(\mathbf{y}|\mathbf{X})$ does **not** depend on $\boldsymbol{\theta}$. $\rightsquigarrow$ Treated as constant when maxing the posterior of $\boldsymbol{\theta}$.

1 (a): The Likelihood $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\right]$$
$$= \exp\left[-\frac{||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||^2}{2\sigma^2}\right]$$
$$= \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2}\right].$$

In addition, recall that the prior $q(\theta) \propto 1$ and we don't care the marginal $p(\mathbf{y}|\mathbf{X})$.
Now, we plug these information into $p(\theta|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X},\theta)q(\theta)}{p(\mathbf{y}|\mathbf{X})} \propto p(\mathbf{y}|\mathbf{X}, \theta)q(\theta)$.

1 (a): The Likelihood $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) &\propto \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\right] \\
&= \exp\left[-\frac{||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||^2}{2\sigma^2}\right] \\
&= \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2}\right].
\end{aligned}
$$

In addition, recall that the prior $q(\boldsymbol{\theta}) \propto 1$ and we don't care the marginal $p(\mathbf{y}|\mathbf{X})$.
Now, we plug these information into $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})q(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{X})} \propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})q(\boldsymbol{\theta})$.

1 (a): MAP Estimate

$$p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) \propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) \cdot q(\boldsymbol{\theta}) \propto \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2}\right].$$

Now we compute the maximum a posterior estimate as:

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\boldsymbol{\theta}} p(\theta|\mathbf{X},\mathbf{y}) \\
&= \arg\max_{\boldsymbol{\theta}} \log(p(\theta|\mathbf{X},\mathbf{y})) \qquad \text{(log is a monotone increasing func.)} \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} \\
&= \arg\min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} \\
&= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2.
\end{aligned}$$

Therefore, in 1 (a), maximum a posteriori estimate $\Leftrightarrow$ ERM for the L2-loss.

1 (a): MAP Estimate

$$p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) \propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) \cdot q(\boldsymbol{\theta}) \propto \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2}\right].$$

Now we compute the maximum a posterior estimate as:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) \\
&= \arg\max_{\boldsymbol{\theta}} \log(p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})) \qquad \text{(log is a monotone increasing func.)} \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} \\
&= \arg\min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} \\
&= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2.
\end{aligned}$$

Therefore, in 1 (a), maximum a posteriori estimate $\Leftrightarrow$ ERM for the L2-loss.

Exercise 1(b)

Show that if we choose a **Gaussian distribution** over the parameter vectors $\boldsymbol{\theta}$ as the prior belief, i.e.,

$$q(\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2\tau^2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right], \quad \tau > 0,$$

then the maximum a posteriori estimate coincides for a specific choice of τ with the **regularized** empirical risk miminizer for the L2-loss with L2 penalty (over the linear models), i.e., the Ridge regression.

1 (b): Posterior with A Gaussian Prior

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) &\propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) q(\boldsymbol{\theta}) \\
&\propto \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{1}{2\tau^2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right] \\
&\propto \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{||\boldsymbol{\theta}||_2^2}{2\tau^2}\right].
\end{aligned}
$$

Next, we compute $\arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$. That is, $\arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$.

1 (b): MAP Estimate

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{||\boldsymbol{\theta}||_2^2}{2\tau^2} \\
&= \arg\min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} + \frac{||\boldsymbol{\theta}||_2^2}{2\tau^2} \\
&= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 + \frac{\sigma^2}{\tau^2}||\boldsymbol{\theta}||_2^2. \qquad \text{(Because } \times\sigma^2 \text{ doesn't change the argmin)}
\end{aligned}
$$

We define $\lambda = \frac{\sigma^2}{\tau^2}$, then the maximum a posteriori $\Leftrightarrow$ L2-loss with L2 penalty.

Exercise 1(c)

Show that if we choose a **Laplace distribution** over the parameter vectors $\boldsymbol{\theta}$ as the prior belief, i.e.,

$$q(\boldsymbol{\theta}) \propto \exp\left[-\frac{\sum_i^p |\boldsymbol{\theta}_i|}{\tau}\right], \quad \tau > 0,$$

then the maximum a posteriori estimate coincides for a specific choice of τ with the regularized empirical risk minimizer for the L2-loss with L1 penalty (over the linear models), i.e., the Lasso regression.

1 (c): Posterior with A Laplace Prior

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})q(\boldsymbol{\theta}) \\
&\propto \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{||\boldsymbol{\theta}||_1}{\tau}\right].
\end{aligned}
$$

1 (c): MAP Estimate

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{||\boldsymbol{\theta}||_1}{\tau} \\
&= \arg\min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} + \frac{||\boldsymbol{\theta}||_1}{\tau} \\
&= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 + \frac{2\sigma^2}{\tau}||\boldsymbol{\theta}||_1 \qquad \text{(Because } \times\sigma^2 \text{ doesn't change the argmin)}
\end{aligned}
$$

We define $\lambda = \frac{2\sigma^2}{\tau}$, and then the MAP estimate $\Leftrightarrow$ L2-loss with L1 penalty.

Exercise 2: Covariance Functions

Consider the commonly used covariance functions mentioned in the lecture slides: constant, linear, polynomial, squared exponential, Matern, exponential covariance functions.

(a) Show that they are valid covariance functions. (**Proofs for Matern and exp. cov. functions are out of scope and omitted.**) You may use the following composition rules. In these rules we assume that $k_0(\cdot,\cdot)$ and $k_1(\cdot,\cdot)$ are valid covariance functions.

1. $k(\mathbf{x},\mathbf{x}') = \mathbf{x}^\top\mathbf{x}'$ is a valid covariance function;
2. $k(\mathbf{x},\mathbf{x}') = c \cdot k_0(\mathbf{x},\mathbf{x}')$ is a valid covariance function if $c \geq 0$ is constant.
3. $k(\mathbf{x},\mathbf{x}') = k_0(\mathbf{x},\mathbf{x}') + k_1(\mathbf{x},\mathbf{x}')$ is a valid covariance function;
4. $k(\mathbf{x},\mathbf{x}') = k_0(\mathbf{x},\mathbf{x}') \cdot k_1(\mathbf{x},\mathbf{x}')$ is a valid covariance function;
5. $k(\mathbf{x},\mathbf{x}') = g(k_0(\mathbf{x},\mathbf{x}'))$ is a valid cov. func. if g is a polynomial function with **pos.** coefficients;
6. $k(\mathbf{x},\mathbf{x}') = t(\mathbf{x}) \cdot k_0(\mathbf{x},\mathbf{x}') \cdot t(\mathbf{x}')$ is a valid cov. function, where t is any **continuous** function;
7. $k(\mathbf{x},\mathbf{x}') = \exp(k_1(\mathbf{x},\mathbf{x}'))$ is a valid covariance function;
8. $k(\mathbf{x},\mathbf{x}') = \mathbf{x}^\top\mathbf{A}\mathbf{x}'$ is a valid covariance function if $\mathbf{A} \succeq 0$.

2 (a): Proof via Kernel Matrix

Construct of the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ from $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$.
Each element $K_{i,j} = k(\mathbf{x}, \mathbf{x}')$. In the current case $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2$, we have $K_{i,j} = \sigma_0^2$. In other words,

$$\mathbf{K} = \begin{pmatrix} \sigma_0^2 & \ldots & \sigma_0^2 \\ \vdots & \ddots & \vdots \\ \sigma_0^2 & \ldots & \sigma_0^2 \end{pmatrix}$$

Note: kernel matrix $\mathbf{K}$ is NOT kernel function $k(\cdot, \cdot)$. Don't claim "$k(\cdot, \cdot)$ is P.S.D." in the exam.
Now, We need to prove that $\mathbf{K}$ is P.S.D.

1. **Prove K is symmetric.**
2. **Prove that $\forall\, \boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{v}^\top \mathbf{K} \boldsymbol{v} \geq 0$.**

2 (a): How to Prove That A Matrix Is P.S.D.

1. Since $K_{i,j} = \sigma_0^2$ for all i, j, we have $\mathbf{K}^\top = \mathbf{K}$, thus $\mathbf{K}$ is symmetric.
2. For any $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^\top$, we need to prove $\boldsymbol{v}^\top \mathbf{K} \boldsymbol{v} \geq 0$.
   2.1 Naive way: First compute $(\mathbf{K}\boldsymbol{v})$ and then $\boldsymbol{v}^\top(\mathbf{K}\boldsymbol{v})$.

$$\mathbf{K}\boldsymbol{v} = \sigma_0^2\Big(\sum_i v_i, \sum_i v_i, \ldots, \sum_i v_i\Big)^\top$$

$$\boldsymbol{v}^\top \mathbf{K}\boldsymbol{v} = \sigma_0^2\Big[v_1\Big(\sum_i v_i\Big) + v_2\Big(\sum_i v_i\Big) + \ldots + v_n\Big(\sum_i v_i\Big)\Big] = \sigma_0^2(v_1 + v_2 + \ldots + v_n)\Big(\sum_i v_i\Big)$$

$$= \sigma_0^2(v_1 + \ldots + v_n)\Big(\sum_i v_i\Big) = \sigma_0^2\Big(\sum_i v_i\Big)\Big(\sum_i v_i\Big) = \sigma_0^2\Big(\sum_i v_i\Big)^2 \geq 0.$$

   2.2 Faster way: $\mathbf{K} = \sigma_0^2 \boldsymbol{l}\boldsymbol{l}^\top$, where $\boldsymbol{l} = (1, 1, \ldots, 1)^\top$. So, $\boldsymbol{v}^\top \mathbf{K}\boldsymbol{v} = \sigma_0^2 \boldsymbol{v}^\top \boldsymbol{l}\boldsymbol{l}^\top \boldsymbol{v} = \sigma_0^2(\boldsymbol{l}^\top \boldsymbol{v})^2 \geq 0$.

2 (a): How to Prove That A Matrix Is P.S.D.

1. Since $K_{i,j} = \sigma_0^2$ for all i, j, we have $\mathbf{K}^\top = \mathbf{K}$, thus $\mathbf{K}$ is symmetric.
2. For any $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^\top$, we need to prove $\boldsymbol{v}^\top \mathbf{K} \boldsymbol{v} \geq 0$.

   2.1 Naive way: First compute $(\mathbf{K}\boldsymbol{v})$ and then $\boldsymbol{v}^\top(\mathbf{K}\boldsymbol{v})$.

   $$\mathbf{K}\boldsymbol{v} = \sigma_0^2\left(\sum_i v_i, \sum_i v_i, \ldots, \sum_i v_i\right)^\top$$

   $$\begin{aligned}\boldsymbol{v}^\top\mathbf{K}\boldsymbol{v} &= \sigma_0^2\left[v_1\left(\sum_i v_i\right) + v_2\left(\sum_i v_i\right) + \ldots + v_n\left(\sum_i v_i\right)\right] = \sigma_0^2(v_1 + v_2 + \ldots + v_n)\left(\sum_i v_i\right) \\ &= \sigma_0^2(v_1 + \ldots + v_n)\left(\sum_i v_i\right) = \sigma_0^2\left(\sum_i v_i\right)\left(\sum_i v_i\right) = \sigma_0^2\left(\sum_i v_i\right)^2 \geq 0.\end{aligned}$$

   2.2 Faster way: $\mathbf{K} = \sigma_0^2 \boldsymbol{l}\boldsymbol{l}^\top$, where $\boldsymbol{l} = (1, 1, \ldots, 1)^\top$. So, $\boldsymbol{v}^\top\mathbf{K}\boldsymbol{v} = \sigma_0^2 \boldsymbol{v}^\top \boldsymbol{l}\boldsymbol{l}^\top \boldsymbol{v} = \sigma_0^2(\boldsymbol{l}^\top \boldsymbol{v})^2 \geq 0$.

2 (a): How to Prove That A Matrix Is P.S.D.

1. Since $K_{i,j} = \sigma_0^2$ for all i, j, we have $\mathbf{K}^\top = \mathbf{K}$, thus $\mathbf{K}$ is symmetric.
2. For any $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^\top$, we need to prove $\boldsymbol{v}^\top \mathbf{K} \boldsymbol{v} \geq 0$.
   2.1 Naive way: First compute $(\mathbf{K}\boldsymbol{v})$ and then $\boldsymbol{v}^\top(\mathbf{K}\boldsymbol{v})$.

$$
\begin{aligned}
\mathbf{K}\boldsymbol{v} &= \sigma_0^2 (\sum_i v_i, \sum_i v_i, \ldots, \sum_i v_i)^\top \\
\boldsymbol{v}^\top \mathbf{K}\boldsymbol{v} &= \sigma_0^2 [v_1(\sum_i v_i) + v_2(\sum_i v_i) + \ldots + v_n(\sum_i v_i)] = \sigma_0^2 (v_1 + v_2 + \ldots + v_n)(\sum_i v_i) \\
&= \sigma_0^2 (v_1 + \ldots + v_n)(\sum_i v_i) = \sigma_0^2 (\sum_i v_i)(\sum_i v_i) = \sigma_0^2 (\sum_i v_i)^2 \geq 0.
\end{aligned}
$$

   2.2 Faster way: $\mathbf{K} = \sigma_0^2 \boldsymbol{I}\boldsymbol{I}^\top$, where $\boldsymbol{I} = (1, 1, \ldots, 1)^\top$. So, $\boldsymbol{v}^\top \mathbf{K}\boldsymbol{v} = \sigma_0^2 \boldsymbol{v}^\top \boldsymbol{I}\boldsymbol{I}^\top \boldsymbol{v} = \sigma_0^2 (\boldsymbol{I}^\top \boldsymbol{v})^2 \geq 0$.

2 (a): How to Prove That A Matrix Is P.S.D.

1. Since $K_{i,j} = \sigma_0^2$ for all i, j, we have $\mathbf{K}^\top = \mathbf{K}$, thus $\mathbf{K}$ is symmetric.
2. For any $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^\top$, we need to prove $\boldsymbol{v}^\top \mathbf{K} \boldsymbol{v} \geq 0$.

   2.1 Naive way: First compute $(\mathbf{K}\boldsymbol{v})$ and then $\boldsymbol{v}^\top(\mathbf{K}\boldsymbol{v})$.

$$\mathbf{K}\boldsymbol{v} = \sigma_0^2(\sum_i v_i, \sum_i v_i, \ldots, \sum_i v_i)^\top$$

$$\begin{aligned}\boldsymbol{v}^\top \mathbf{K}\boldsymbol{v} &= \sigma_0^2[v_1(\sum_i v_i) + v_2(\sum_i v_i) + \ldots + v_n(\sum_i v_i)] = \sigma_0^2(v_1 + v_2 + \ldots + v_n)(\sum_i v_i) \\ &= \sigma_0^2(v_1 + \ldots + v_n)(\sum_i v_i) = \sigma_0^2(\sum_i v_i)(\sum_i v_i) = \sigma_0^2(\sum_i v_i)^2 \geq 0.\end{aligned}$$

   2.2 Faster way: $\mathbf{K} = \sigma_0^2 \boldsymbol{l}\boldsymbol{l}^\top$, where $\boldsymbol{l} = (1, 1, \ldots, 1)^\top$. So, $\boldsymbol{v}^\top \mathbf{K}\boldsymbol{v} = \sigma_0^2 \boldsymbol{v}^\top \boldsymbol{l}\boldsymbol{l}^\top \boldsymbol{v} = \sigma_0^2(\boldsymbol{l}^\top \boldsymbol{v})^2 \geq 0$.

2 (a): Proof via Transformed Feature Map

Alternatively, we can prove $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2$ is a valid cov. func. via writing $k(\mathbf{x}, \mathbf{x}')$ as **an inner product of two transformed feature maps**.
This requires to explicitly construct the feature map $\phi(\mathbf{x}) \in \mathbb{R}^d$ for some d.
In the current case, we can write

$$\phi(\mathbf{x}) = \sigma_0.$$

So that

$$\begin{aligned} k\left(\mathbf{x}, \mathbf{x}'\right) &= \sigma_0^2 \\ &= \langle \sigma_0, \sigma_0 \rangle \\ &= \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \end{aligned}$$

2 (a): Proof of $k\left(\mathbf{x}, \mathbf{x}'\right) = \sigma_0^2 + \mathbf{x}^\top \mathbf{x}'$

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \underbrace{\sigma_0^2}_{:=k_0(\mathbf{x}, \mathbf{x}')} + \underbrace{\mathbf{x}^\top \mathbf{x}'}_{:=k_1(\mathbf{x}, \mathbf{x}')}$$

1. We have shown that k_0 is a valid cov. func.
2. k_1 is a inner product $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$, where $\phi(\mathbf{x}) := \mathbf{x}$. So, k_1 is a valid cov. func.
3. Their sum $k_0 + k_1$ is also a cov. func.

2 (a): Proof of $k(\mathbf{x}, \mathbf{x}') = (\sigma_0^2 + \mathbf{x}^\top \mathbf{x})^p$

We define $k_2(\mathbf{x}, \mathbf{x}') = \sigma_0^2 + \mathbf{x}^\top \mathbf{x}$, so $k_3 = (\sigma_0^2 + \mathbf{x}^\top \mathbf{x})^p = k_2^p$.

1. We have shown that k_2 is a cov. func.
2. $k_3 = k_2^p$ is a polynomial of k_2 with only one p-order item k_2^p, and **the polynomial coefficient** 1 **is positive**. So $k_3 = k_2^p$ is a cov. func.

2 (a): Proof of $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||_2^2}{2\ell^2}\right)$

We can write

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||_2^2}{2\ell^2}\right) \\ &= \exp\left(-\frac{\mathbf{x}^\top\mathbf{x} - 2\mathbf{x}^\top\mathbf{x}' + \mathbf{x}'^\top\mathbf{x}'}{2\ell^2}\right) \\ &= \underbrace{\exp\left(-\frac{\mathbf{x}^\top\mathbf{x}}{2\ell^2}\right)}_{:=t(\mathbf{x})} \cdot \underbrace{\exp\left(\frac{\mathbf{x}^\top\mathbf{x}'}{\ell^2}\right)}_{:=k_4(\mathbf{x},\mathbf{x}')} \cdot \underbrace{\exp\left(-\frac{\mathbf{x}'^\top\mathbf{x}'}{2\ell^2}\right)}_{:=t(\mathbf{x}')}. \end{aligned}$$

where we defined a function $t(\cdot)$.
Furthermore, $\mathbf{x}^\top\mathbf{x}'$ is cov. func., so $\frac{\mathbf{x}^\top\mathbf{x}'}{\ell^2}$ is a kernel, so $\exp(\frac{\mathbf{x}^\top\mathbf{x}'}{\ell^2})$ is a cov. func.
Therefore, $k(\mathbf{x}, \mathbf{x}') = t(\mathbf{x}) \cdot k_4(\mathbf{x}, \mathbf{x}') \cdot t(\mathbf{x}')$ is a cov. func.

Exercise 2 (b)

(b): Are these covariance functions stationary or isotropic? Justify your answer.

2 (b): Stationary and Isotropic

1. $k(\cdot,\cdot)$ is stationary if $k(\mathbf{x}, \mathbf{x} + \boldsymbol{d}) = k(\mathbf{0}, \boldsymbol{d})$.
2. $k(\cdot,\cdot)$ is isotropic if it is a function of $||\mathbf{x} - \mathbf{x}'||$.

2 (b): Constant functions

1. $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2$ is stationary since $k(\mathbf{x}, \mathbf{x} + \boldsymbol{d}) = \sigma_0^2 = k(\mathbf{0}, \boldsymbol{d})$.
2. $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2$ is isotropic since $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 ||\mathbf{x} - \mathbf{x}'||^0$.

2 (b): Constant functions

1. $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2$ is stationary since $k(\mathbf{x}, \mathbf{x} + \boldsymbol{d}) = \sigma_0^2 = k(\mathbf{0}, \boldsymbol{d})$.
2. $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2$ is isotropic since $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 ||\mathbf{x} - \mathbf{x}'||^0$.

2 (b): $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 + \mathbf{x}^\top \mathbf{x}'$

1. $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 + \mathbf{x}^\top \mathbf{x}'$ is NOT stationary, since

$$k(\mathbf{x}, \mathbf{x} + \boldsymbol{d}) = \sigma_0^2 + \mathbf{x}^\top \mathbf{x} + \mathbf{x}^\top \boldsymbol{d}$$
$$k(\mathbf{0}, \boldsymbol{d}) = \sigma_0^2.$$

2. It is NOT isotropic, since it cannot be written as a func. of $||\mathbf{x} - \mathbf{x}'||$.

2 (b): $k\left(\mathbf{x}, \mathbf{x}'\right) = \sigma_0^2 + \mathbf{x}^\top \mathbf{x}'$

1. $k\left(\mathbf{x}, \mathbf{x}'\right) = \sigma_0^2 + \mathbf{x}^\top \mathbf{x}'$ is NOT stationary, since

$$k(\mathbf{x}, \mathbf{x} + \boldsymbol{d}) = \sigma_0^2 + \mathbf{x}^\top \mathbf{x} + \mathbf{x}^\top \boldsymbol{d}$$
$$k(\mathbf{0}, \boldsymbol{d}) = \sigma_0^2.$$

2. It is NOT isotropic, since it cannot be written as a func. of $||\mathbf{x} - \mathbf{x}'||$.

2 (b): Polynomial Cov. Func.

Similar to linear covariance functions, the polynomial covariance function is NOT stationary and NOT isotropic. (Prove this on your own.)

2 (b): $k\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||_2^2}{2\ell^2}\right)$

1. $k\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||_2^2}{2\ell^2}\right)$ is stationary, since $k(\mathbf{x}, \mathbf{x}+\boldsymbol{d}) = \exp\left(-\frac{||\boldsymbol{d}||_2^2}{2\ell^2}\right) = k(\mathbf{0}, \boldsymbol{d})$.
2. It is isotropic, since it is a function of $||\mathbf{x}-\mathbf{x}'||$.

2 (b): $k\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||_2^2}{2\ell^2}\right)$

1. $k\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||_2^2}{2\ell^2}\right)$ is stationary, since $k(\mathbf{x}, \mathbf{x} + \boldsymbol{d}) = \exp\left(-\frac{||\boldsymbol{d}||_2^2}{2\ell^2}\right) = k(\mathbf{0}, \boldsymbol{d})$.
2. It is isotropic, since it is a function of $||\mathbf{x} - \mathbf{x}'||$.

2 (b): Matern and Exponential Cov. Func.

Similar to the argument of squared exponential conv. func. $k\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||_2^2}{2\ell^2}\right)$.

1. Matern cov. func. is stationary and isotropic.

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \frac{1}{2^\nu}\left(\frac{\sqrt{2\nu}}{\ell}||\mathbf{x}-\mathbf{x}'||\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}}{\ell}||\mathbf{x}-\mathbf{x}'||\right)$$

2. Exponential cov. func. is stationary and isotropic.

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||}{\ell}\right)$$