**CS 412 — Introduction to Machine Learning (UIC)** April 24, 2025

Lecture 26

*Instructor: Aadirupa Saha* *Scribe(s): Amith Bhat Hosadurga Anand*

[This draft is not fully proofread. Please email any typos/errors to the scribe/instructor or directly edit the file.]

## Overview

In the last lecture, we covered the following main topics:

1. CNN
2. Dropout regularization
3. Vanishing & Exploding Gradients (VE-Grads)
4. RNN-LSTM

This lecture focuses on:

1. Properties of (Multivariate) Gaussian Distribution
2. Gaussian Process: A Bayesian Regression Technique

---

# 1 Properties of (Multivariate) Gaussian Distribution

## 1.1 Preliminaries

**Definition 1** (Multivariate Gaussian)**.** *A random variable $X \in \mathbb{R}^d$ is said to follow $\mathcal{MVG}$ distribution with mean and covariance parameters $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ (where $\Sigma$ is positive semi-definite), if:*

$$P(X = x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

*where $|\Sigma|$ denotes the determinant of the covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$.*

**Remark 1.** *Special case: If $\Sigma$ is diagonal, then all the $d$ coordinates of the random variables are independent. Note that the diagonal entries **MUST** be positive, and $\Sigma$ has to be positive semi-definite (PSD).*

## 1.2 Notation

Before we begin with the properties, let us define some notation.
Assume the $d$-coordinates of $X$ can be partitioned into two sets $A$ and $B$, such that $A \cup B = [d]$.
Without loss of generality, let $A = \{1, 2, \ldots, r\}$, for some $1 \leq r < d$, and hence $B = \{r+1, \ldots, d\}$,

So for any realization '$x$' in $X$, we will denote

$$x = \begin{bmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}$$

where $\mu_A \in \mathbb{R}^r, \quad \mu_B \in \mathbb{R}^{d-r}$
$\Sigma_{AA} \in \mathbb{R}^{r \times r}, \quad \Sigma_{AB} \in \mathbb{R}^{r \times (d-r)}, \quad \Sigma_{BA} \in \mathbb{R}^{(d-r) \times r}$ and $\Sigma_{BB} \in \mathbb{R}^{(d-r) \times (d-r)}$

Another method of deriving the dimensions of the partitions of the covariance matrix is through the definition of the covariance matrix

$$\Sigma = \mathbb{E}_{X \sim \mathcal{P}(\cdot|\mu,\Sigma)} \left[ (X - \mu)(X - \mu)^\top \right]$$

Further,

$$\Sigma_{AA} = \mathbb{E}_{X \sim \mathcal{P}(\cdot|\mu,\Sigma)} \left[ (\mathbf{x}_A - \mu_A)(\mathbf{x}_A - \mu_A)^\top \right] = \Sigma_{AA}^\top \in \mathbb{R}^{r \times r}$$

Similarly,

$$\Sigma_{BB} = \mathbb{E}_{X \sim \mathcal{P}(\cdot|\mu,\Sigma)} \left[ (\mathbf{x}_B - \mu_B)(\mathbf{x}_B - \mu_B)^\top \right] = \Sigma_{BB}^\top \in \mathbb{R}^{(d-r) \times (d-r)}$$

and lastly:

$$\Sigma_{AB} = \mathbb{E}_{X \sim \mathcal{P}(\cdot|\mu,\Sigma)} \left[ (\mathbf{x}_A - \mu_A)(\mathbf{x}_B - \mu_B)^\top \right] \quad \in \mathbb{R}^{r \times (d-r)}$$

$$= \Sigma_{BA}^\top \in \mathbb{R}^{r \times (d-r)}$$

## 1.3 Properties

**Definition 2.** *Marginals*
*The marginal probability distributions of a subset of random variables are obtained by integrating out the remaining variables. Specifically, for a random vector $X$ with partitioned components $X_A$ and $X_B$, the marginal probability distribution of $X_A$ is given by:*

$$P(X_A = x_A) = \int_{x_B} P\left( X = \begin{bmatrix} x_A \\ x_B \end{bmatrix}; \mu, \Sigma \right) dx_B$$

*Similarly, the marginal probability distribution of $X_B$ is given by:*

$$P(X_B = x_B) = \int_{x_A} P\left( X = \begin{bmatrix} x_A \\ x_B \end{bmatrix}; \mu, \Sigma \right) dx_A$$

### 1.3.1 Property 1

**Theorem 1.1: Marginal Distributions**

$X_A \sim \mathcal{N}(\mu_A, \Sigma_{AA})$, which is a $\mathcal{MVG}$ with mean $\mu_A$ and covariance $\sum_{AA}$ and
$X_B \sim \mathcal{N}(\mu_B, \Sigma_{BB})$, which is a $\mathcal{MVG}$ with mean $\mu_B$ and covariance $\sum_{BB}$.

**Proof:**
**Note:** *The presented proof is not from class but is included here for completeness.*
The proof follows from computing the marginal integrals and integrating out the irrelevant variables but there is a less computation heavy proof, based on the theory of multivariate normal distributions and follows from the general principles of linear transformations in multivariate normal distributions. For exact details, see the proof in the reference: Marginal distributions of the multivariate normal distribution.

We are given that $X = \begin{bmatrix} X_A \\ X_B \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma)$, where: $\mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}$

Let us derive the marginal distribution of $X_A$. The marginal distribution of $X_B$ follows a similar structure

**Step 1: Define the Subset Matrix $S$:**
For the marginal distribution of $X_A$, define the subset matrix $S$, which extracts the elements corresponding to $X_A$ from the full vector $X$. In this case, we define $S$ as a $r \times d$ matrix such that:

$$S = \begin{bmatrix} I_r & 0 \end{bmatrix}$$

Where $I_r$ is the $r \times r$ identity matrix, and 0 is the zero matrix of size $r \times (d - r)$. This matrix $S$ selects the first $r$-dimensional vector $X_A$ from the full vector $X$, so:

$$X_A = SX$$

**Step 2: Apply the Linear Transformation Theorem:**
By the Linear Transformation Theorem, since $X \sim \mathcal{N}(\mu, \Sigma)$, we know that:

$$X_A \sim \mathcal{N}(S\mu, S\Sigma S^T)$$

Substitute $S = \begin{bmatrix} I_r & 0 \end{bmatrix}$ into the equation:

$$S\mu = \begin{bmatrix} I_r & 0 \end{bmatrix} \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix} = \mu_A$$

Thus, the mean of $X_A$ is $\mu_A$.
Now, compute the covariance matrix $S\Sigma S^T$:

$$S\Sigma S^T = \begin{bmatrix} I_r & 0 \end{bmatrix} \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \begin{bmatrix} I_r \\ 0 \end{bmatrix}$$

This simplifies to: $S\Sigma S^T = \Sigma_{AA}$
Thus, the marginal distribution of $X_A$ is: $X_A \sim \mathcal{N}(\mu_A, \Sigma_{AA})$
Similarly, the marginal distribution of $X_B$ is: $X_B \sim \mathcal{N}(\mu_B, \Sigma_{BB})$ (Use $S = \begin{bmatrix} 0 & I_{d-r} \end{bmatrix}$) □

**Definition 3.** *Conditional Distribution*
*We begin with the conditional distribution of $X_A = x_A$ given $X_B = x_B$ where $x_A$ and $x_B$ are the given realized values of $X_A$ and $X_B$:*

$$P(X_A = x_A \mid X_B = x_B) = \frac{P(X_A = x_A \cap X_B = x_B)}{P(X_B = x_B)}$$

$$= \frac{P\left(X = \begin{bmatrix} x_A \\ x_B \end{bmatrix}; \mu, \Sigma\right)}{P(X_B = x_B; \mu_B, \Sigma_{BB})} \quad \textit{( denominator follows from Property 1)}$$

*Similarly:*

$$P(X_B = x_B \mid X_A = x_A) = \frac{P\left(X = \begin{bmatrix} x_A \\ x_B \end{bmatrix}; \mu, \Sigma\right)}{P(X_A = x_A; \mu_A, \Sigma_{AA})} \quad \textit{(again, denominator follows from Property 1)}$$

### 1.3.2 Property 2

**Theorem 1.2: Conditional Distributions**

a) $P(X_A \mid X_B = x_B) \sim \mathcal{N}\left(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B),\ \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)$

b) $P(X_B \mid X_A = x_A) \sim \mathcal{N}\left(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A),\ \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}\right)$

**Proof:** We will prove Property 2a as 2b will follow a similar analysis.
Note, by definition of conditional probability, we have:

$$P(X_A = x_A \mid X_B = x_B) = \frac{P\left(X = \begin{bmatrix} x_A \\ x_B \end{bmatrix}; \mu, \Sigma\right)}{P(X_B = x_B)}$$

Substitute the terms and simplify:

$$P(X_A = x_A \mid X_B = x_B) = \frac{\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}{P(X_B = x_B)}$$

Since $P(X_B = x_B)$ is independent of $x_A$, we can group it, along with $\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}$ as a constant $Z_1$
This simplifies to:

$$P(X_A = x_A \mid X_B = x_B) = \frac{1}{Z_1} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} x_A \\ x_B \end{bmatrix} - \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}\right)^T \Sigma^{-1} \left(\begin{bmatrix} x_A \\ x_B \end{bmatrix} - \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}\right)\right)$$

We can write

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} = \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix}$$

Thus, the conditional distribution can be written as:

$$P(X_A = x_A \mid X_B = x_B) = \frac{1}{Z_1} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} x_A \\ x_B \end{bmatrix} - \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}\right)^T \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} \left(\begin{bmatrix} x_A \\ x_B \end{bmatrix} - \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}\right)\right)$$

Expanding the matrices we get:

$$\frac{1}{Z_1}\exp\left[-\frac{1}{2}\Big((x_A-\mu_A)^T V_{AA}(x_A-\mu_A)+(x_A-\mu_A)^T V_{AB}(x_B-\mu_B)\right.$$
$$\left.+(x_B-\mu_B)^T V_{BA}(x_A-\mu_A)+(x_B-\mu_B)^T V_{BB}(x_B-\mu_B)\Big)\right]$$

Let us separate the terms independent of $x_A$, since we want to understand the distribution of $X_A \mid X_B = x_B$:

$$=\frac{1}{Z_2}\exp\left[-\frac{1}{2}\left(x_A^T V_{AA}x_A-2x_A^T V_{AA}\mu_A+2x_A^T V_{AB}(x_B-\mu_B)\right)\right]$$

where

$$\frac{1}{Z_2}=\frac{1}{Z_1}\exp\left[-\frac{1}{2}\left(\mu_A^T V_{AA}\mu_A-2\mu_A^T V_{AB}(x_B-\mu_B)+(x_B-\mu_B)^T V_{BB}(x_B-\mu_B)\right)\right]$$

Which are all the terms independent of $x_A$

This simplifies to:

$$=\frac{1}{Z_3}\exp\left(-\frac{1}{2}\left[x_A-\mu'_A\right]^T V_{AA}\left[x_A-\mu'_A\right]\right) \qquad \star$$

where:

$$\mu'_A=\mu_A-V_{AA}^{-1}V_{AB}(x_B-\mu_B) \quad \in \mathbb{R}^r$$

and $Z_3$ is adjusted accordingly.

The dimension of $\mu'_A$ follows as $\mu_A \in \mathbb{R}^r, V_{AA}^{-1} \in \mathbb{R}^{r\times r}, V_{AB} \in \mathbb{R}^{r\times(d-r)}$ and $(x_B-\mu_B) \in \mathbb{R}^{(d-r)\times 1}$
Further noting that,

$$\begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} = \begin{bmatrix} =\left(\Sigma_{AA}-\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)^{-1} & -(\Sigma_{AA}-\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})^{-1}\Sigma_{AB}\Sigma_{BB}^{-1}) \\ -\Sigma_{BB}^{-1}\Sigma_{BA}\left(\Sigma_{AA}-\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)^{-1} & (\Sigma_{BB}-\Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB})^{-1} \end{bmatrix}$$

which can be proved by matrix inversion technique.
The generalized process of obtaining such inversions can be referred to in this article on Schur Complement.

The form of $\star$ above justifies:

$$P(X_A \mid X_B = x_B) \sim \mathcal{N}(\mu'_A, V_{AA}) = \mathcal{N}\left(\mu_A+\Sigma_{AB}\Sigma_{BB}^{-1}(x_B-\mu_B), \Sigma_{AA}-\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)$$

A similar sequence of steps can be followed to obtain 2b.

# An Alternative (Easier/More Intuitive) Proof of Property 2

Note the above proof starts with the expression of $P(X_A = x_A \mid X_B = x_B)$ using conditional probability and tries to rearrange the terms to obtain another Multivariate Gaussian distribution expression whose mean and covariance turn out to be $\mu'_A$ and $V_{AA}$, after some useful regrouping of the terms.

However, since $P(X_A) \sim \mathcal{N}(\mu_A, \Sigma_{AA})$ and $P(X_B) \sim \mathcal{N}(\mu_B, \Sigma_{BB})$,
and we know that the conditional distribution of $\mathcal{MVG}$ is also another $\mathcal{MVG}$, we know that:

$$P(X_A = x_A \mid X_B = x_B) \sim \mathcal{N}(\tilde{\mu}, \tilde{\Sigma}) \quad \text{for some} \quad \tilde{\mu} \in \mathbb{R}^r \quad \text{and} \quad \tilde{\Sigma} \in \mathbb{R}^{r \times r} \quad \text{(PSD)}.$$

Our goal is to find $\tilde{\mu}$ and $\tilde{\Sigma}$.
To find $\tilde{\mu}$, we note that we derived above that:

$$P(X_A = x_A \mid X_B = x_B) = \frac{1}{Z_2} \exp\left(-\frac{1}{2}\left[x_A^T V_{AA} x_A - 2x_A^T V_{AA}\mu_A + (x_B - \mu_B)^T V_{BB}(x_B - \mu_B)\right]\right) \quad (1)$$

Now since $P(X_A \mid X_B) \sim \mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$, with $\tilde{\mu}$ and $\tilde{\Sigma}$ unknown.
Now we know the maximum density of any $\mathcal{MVG}$ is achieved at its mean. Hence, to find $\tilde{\mu}$ (the mean of $\mathcal{MVG}$ $P(X_A \mid X_B)$), we maximize Equation 1 with respect to $x_A$

$$\arg\max_{x_A \in \mathbb{R}^r} P(X_A = x_A \mid X_B = x_B)$$

$$= \arg\max_{x_A \in \mathbb{R}^r} \left[\frac{1}{Z_2} \exp\left(-\frac{1}{2}\left[x_A^T V_{AA} x_A - 2x_A^T V_{AA}\mu_A + 2x_A^T V_{AB}(x_B - \mu_B)\right]\right)\right]$$

Which is equivalent to maximizing,

$$= \arg\max_{x_A \in \mathbb{R}^r} \left[x_A^T V_{AA} x_A - 2x_A^T V_{AA}\mu_A + 2x_A^T V_{AB}(x_B - \mu_B)\right]$$

Let $f(x_A) = x_A^T V_{AA} x_A - 2x_A^T V_{AA}\mu_A + 2x_A^T V_{AB}(x_B - \mu_B)$
Then,

$$\nabla f(x_A) = 2V_{AA}x_A - 2V_{AA}\mu_A + 2V_{AB}(x_B - \mu_B)$$

and

$$\nabla^2 f(x_A) = V_{AA} \succeq 0$$

Which is PSD since we know that $V_{AA}\left(\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)^{-1}$

$\implies$ The maximum is achieved at: $\nabla f(x_A^*) = 0$

$$\Rightarrow V_{AA} \cdot x_A^* = V_{AA} \cdot \mu_A - V_{AB}(x_B - \mu_B)$$

$$\Rightarrow x_A^* = \mu_A - V_{AA}^{-1} V_{AB}(x_B - \mu_B) \quad (2)$$

But noting that $V_{AA}^{-1} = (\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})$ and $V_{AB} = -(\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})\Sigma_{AB}\Sigma_{BB}^{-1}$, we obtain:

$$V_{AA}^{-1} V_{AB} = -\Sigma_{AB}\Sigma_{BB}^{-1} \qquad \text{(from Eq(2))}$$

Thus, we get:

$$x_A^* = \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B) \quad \text{as desired.}$$

So we have:

$$\tilde{\mu} = x_A^* = \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B)$$

To find $\Sigma$:
Again, since $P(X_A \mid X_B) \sim \mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$, we know that the Hessian of $\mathcal{MVG}$ density is $\tilde{\Sigma}^{-1}$.
But Equation (3) gives:

$$\nabla^2 f(x_A^*) = V_{AA}$$

Thus:

$$\tilde{\Sigma}^{-1} = V_{AA} \quad \Rightarrow \quad \tilde{\Sigma} = V_{AA}^{-1} \quad \Rightarrow \quad \tilde{\Sigma} = \left(\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)^{-1}$$

$$\tilde{\Sigma} = \left(\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)^{-1} \text{ as desired.}$$

### 1.3.3 Property 3

**Theorem 1.3: Independence**

If $X$ and $Y$ are two RVs following

$$X \sim \mathcal{N}(\mu_1, \Sigma_1) \quad \text{and} \quad Y \sim \mathcal{N}(\mu_2, \Sigma_2),$$

respectively, then:

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2) \quad \text{if } X \text{ and } Y \text{ are independent } X \perp Y.$$

**Proof:** The proof follows by analyzing the density of $X + Y$ and exploiting the fact that $X \perp Y$
**Note:** *The presented proof is not from class but is included here for completeness.*
Let $Z = X + Y$. The sum $Z$ is a linear combination of the random variables $X$ and $Y$, and due to the linearity of normal distributions, $Z$ must also follow a normal distribution.
The mean of $Z$ is:

$$E[Z] = E[X] + E[Y] = \mu_1 + \mu_2$$

The covariance of $Z$ is:

$$\text{Cov}(Z) = \text{Cov}(X + Y) = \text{Cov}(X) + \text{Cov}(Y) + 2 \cdot \text{Cov}(X, Y)$$

Since $X$ and $Y$ are independent, $\text{Cov}(X, Y) = 0$, so:

$$\text{Cov}(Z) = \Sigma_1 + \Sigma_2$$

Thus, the sum of two independent normally distributed random variables $X$ and $Y$ is normally distributed with mean $\mu_1 + \mu_2$ and covariance $\Sigma_1 + \Sigma_2$:

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$$

---

# 2 Gaussian Processes (GP): A Bayesian Regression Technique

## 2.1 Gaussian Process Inference

GP (Gaussian Processes) is a method for modeling probability distributions over functions, which can be used to infer function values at unknown datapoints (i.e essentially regression tasks) using the properties of $\mathcal{MVG}$ we studied above.
More specifically, we say that $f : \mathbb{R}^d \to \mathbb{R}$ follows a GP with mean function $m(\cdot)$ and covariance function $k(\cdot, \cdot)$ if for any $x \in \text{domain}(f) \subseteq \mathbb{R}^d$,

$$f(x) \sim \mathcal{N}\left(m(x), k(x, x)\right)$$

and for any sequence of datapoints $x_1, \ldots, x_n \in \mathbb{R}^d$,

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix}, \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}\right)$$

where:

- The vector $\begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix} \in \mathbb{R}^n$, represents the mean of the $\mathcal{MVG}$ for the random variables $\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$,
- The matrix $\begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix} \in \mathbb{R}^{n \times n}$ represents the covariance of the $\mathcal{MVG}$ for the random variables $\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$,

Further, since $k(\cdot, \cdot)$ is a valid kernel for $f(\cdot)$, by the property of kernel functions, for any $n$, we have

$$\begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$$

is always a PSD matrix, as desired for it to be a valid covariance matrix.

Some popular choices of covariance functions $k(\cdot, \cdot)$ could be:

- **Squared Exponential Kernel:**

$$K_{\text{SE}}(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|_2^2\right), \quad \text{for any } x, x' \in \mathbb{R}^d,$$

(Also known as RBF)

- **Linear Kernel:**
$$K_{\text{Lin}}(x, x') = x^T x'$$
- **Polynomial Kernel:**
$$K_{\text{Pol}}(x, x') = \left(x^T x' + 1\right)^d$$
- **Sigmoid Kernel:**
$$K_{\text{Sig}}(x, x') = \tanh(\alpha x^T x' + \beta), \quad \text{etc.}$$

## 2.2 (Regression) Inference using Gaussian Processes:

Suppose we are given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ of $n$ points, such that $y_i = f(x_i) + \varepsilon_i$, for $\quad i \in [n]$, where $f : \mathbb{R}^d \to \mathbb{R}$ is some UNKNOWN function and $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$ for known $\sigma$ (i.e. zero-mean Gaussian noise).
Further, suppose $\tilde{x}_1, \ldots, \tilde{x}_m \in \mathbb{R}^d$ are given test points, where our goal is to infer (regression task) $f(\tilde{x}_1), \ldots, f(\tilde{x}_m)$ using GP inference. *Note:* If we knew $f(\cdot)$, the problem is trivial.

The task here is to approximate $f(\cdot)$ using Gaussian Processes (GP).
We assume the prior mean $f^*$ as $m(\cdot) := 0$, i.e., $m(x) = 0$ for $\forall x \in \mathbb{R}^d$, and we can choose any suitable kernel $k(\cdot, \cdot)$.

So initially:
$$f(\cdot) \sim \text{GP}(m(\cdot); k(\cdot, \cdot)) \quad \Leftrightarrow \quad f(x) \sim \mathcal{N}(0, K(x, x'))$$

Then,
$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} f(x_1) + \varepsilon_1 \\ \vdots \\ f(x_n) + \varepsilon_n \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, K_\varepsilon\right) \qquad \star$$

The by property 3
$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, K\right) \quad \text{and} \quad \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 I_{n\times n}\right) \quad \text{where} \quad K = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix} \in \mathbb{R}^{n\times n}$$

$$\text{say } K = K(X, X) \quad \text{then } K_\varepsilon = \left(K + \sigma^2 I_{n\times n}\right).,$$

where $K_\varepsilon$ is the noise covariance matrix that represents the uncertainty in the observed data due to the noise term $\varepsilon$

**Exercise 2.1: Property 3**

Show how Property 3 proves the above.

*Hint*: The function values $f(x_i)$ and the noise terms $\varepsilon_i$ are assumed to be independent, the sum of these two terms will follow a $\mathcal{MVG}$ whose mean is the sum of the means and covariance is the sum of the covariances.

Further owing to the GP assumption, we know that:

$$\begin{pmatrix} f(\tilde{x}_1) \\ \vdots \\ f(\tilde{x}_m) \end{pmatrix} \sim \mathcal{N}\left( \mathbf{0}, \begin{pmatrix} k(\tilde{x}_1, \tilde{x}_1) & \cdots & k(\tilde{x}_1, \tilde{x}_m) \\ \vdots & \ddots & \vdots \\ k(\tilde{x}_m, \tilde{x}_1) & \cdots & k(\tilde{x}_m, \tilde{x}_m) \end{pmatrix} \right)$$

Where the mean vector is a zero vector $\in \mathbb{R}^m$ and where the covariance matrix is $\in \mathbb{R}^{m \times m}$.
Then, combining the above with $\star$, we have the joint distribution:

$$\begin{pmatrix} f(\tilde{x}_1) \\ \vdots \\ f(\tilde{x}_m) \\ y_1 \\ \vdots \\ y_n \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} k(\tilde{X}, \tilde{X}) & k(\tilde{X}, Y) \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ k(Y, \tilde{X}) & K(X, X) + \sigma^2 I_{n \times n} \end{pmatrix} \right)$$

Where $\mathbf{0} \in \mathbb{R}^{m+n}$, and the covariance matrix $\in \mathbb{R}^{(m+n) \times (m+n)}$ is block matrixed as:

$$\begin{pmatrix} k(\tilde{X}, \tilde{X}) & k(\tilde{X}, Y) \\ k(Y, \tilde{X}) & K(X, X) + \sigma^2 I_{n \times n} \end{pmatrix}$$

where $k(\tilde{X}, \tilde{X}) \in \mathbb{R}^{m \times m}$, $\quad k(\tilde{X}, Y) \in \mathbb{R}^{m \times n}$, $\quad k(Y, \tilde{X}) \in \mathbb{R}^{n \times m}$ $\quad$ and $\quad K(X, X) + \sigma^2 I_{n \times n} \in \mathbb{R}^{n \times m}$

Then, by MVG conditioning (a.k.a. posterior) of $\begin{pmatrix} f(\tilde{x}_1) \\ \vdots \\ f(\tilde{x}_m) \end{pmatrix}$ given the observed realization of $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$
using Property 2 of MVG, we know:

$$P\left( \begin{pmatrix} f(\tilde{x}_1) \\ \vdots \\ f(\tilde{x}_m) \end{pmatrix} \middle| \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right) \sim \mathcal{N}(\hat{m}, \hat{K}); \text{ where}$$

$$\hat{m} = \mu_A + \Sigma_{AB} \Sigma_{BB}^{-1} (\tilde{x}_B - \mu_B) \quad \text{[from Prop 2]}$$

$$= 0 + K(\tilde{X}, y) \left( K(X, X) + \sigma^2 I_{n \times n} \right)^{-1} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\text{and } \tilde{K} = \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA}$$

$$\implies \tilde{K} = K(\tilde{X}, X) - K(\tilde{X}, y) \left( K(X, X) + \sigma^2 I_{n \times n} \right)^{-1} K(Y, \tilde{X})$$

where
$\tilde{K} \in \mathbb{R}^{m \times m}$, $\quad K(\tilde{X}, X) \in \mathbb{R}^{m \times m}$, $\quad K(\tilde{X}, y) \in \mathbb{R}^{m \times n}$, $\quad \left( K(X, X) + \sigma^2 I_{n \times n} \right)^{-1} \in \mathbb{R}^{n \times n}$
and $K(Y, \tilde{X}) \in \mathbb{R}^{n \times m}$

Thus we can compute both $\tilde{m}$ and $\tilde{K}$, and then infer $\begin{pmatrix} f(\tilde{x}_1) \\ \vdots \\ f(\tilde{x}_m) \end{pmatrix}$ .We just need to draw a sample from $\mathcal{N}(\tilde{m}, \tilde{K})$.

This is how we make regression inferences using GP (which is inherently the posterior distribution of MVG random variables!).

## 2.3 Gaussian Process Visualization: Predictions and Confidence Intervals

In the below plots, we visualize the predictions made by a Gaussian Process (with RBF Kernel) regression model, along with its associated uncertainty. The red points represent the training data, while the blue line shows the predicted mean function. The shaded blue region represents the 95% confidence interval, which indicates the uncertainty in the GP's predictions. The wider the region, the higher the uncertainty, which is particularly noticeable in areas with fewer data points.

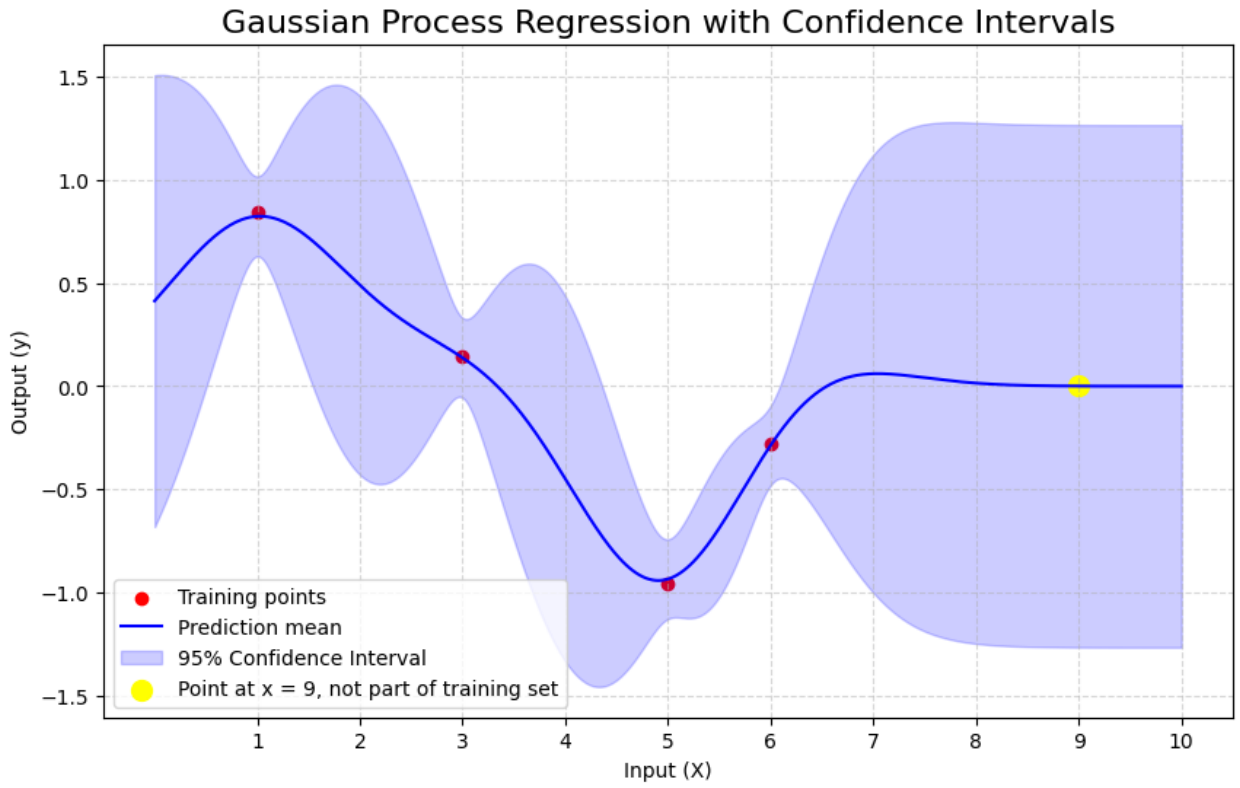


Figure 1: This plot shows the Gaussian Process regression prediction mean and the 95% confidence interval, with $x = 9$ being outside the training set. The wider confidence intervals at $x = 9$ indicate higher uncertainty in predictions for points farther from known data points.

Additionally, the point at x=9 is included to demonstrate how adding new data influences the GP's confidence. As we can see, the variance (confidence interval) near this new point is reduced, indicating that the GP has become more confident about the prediction.

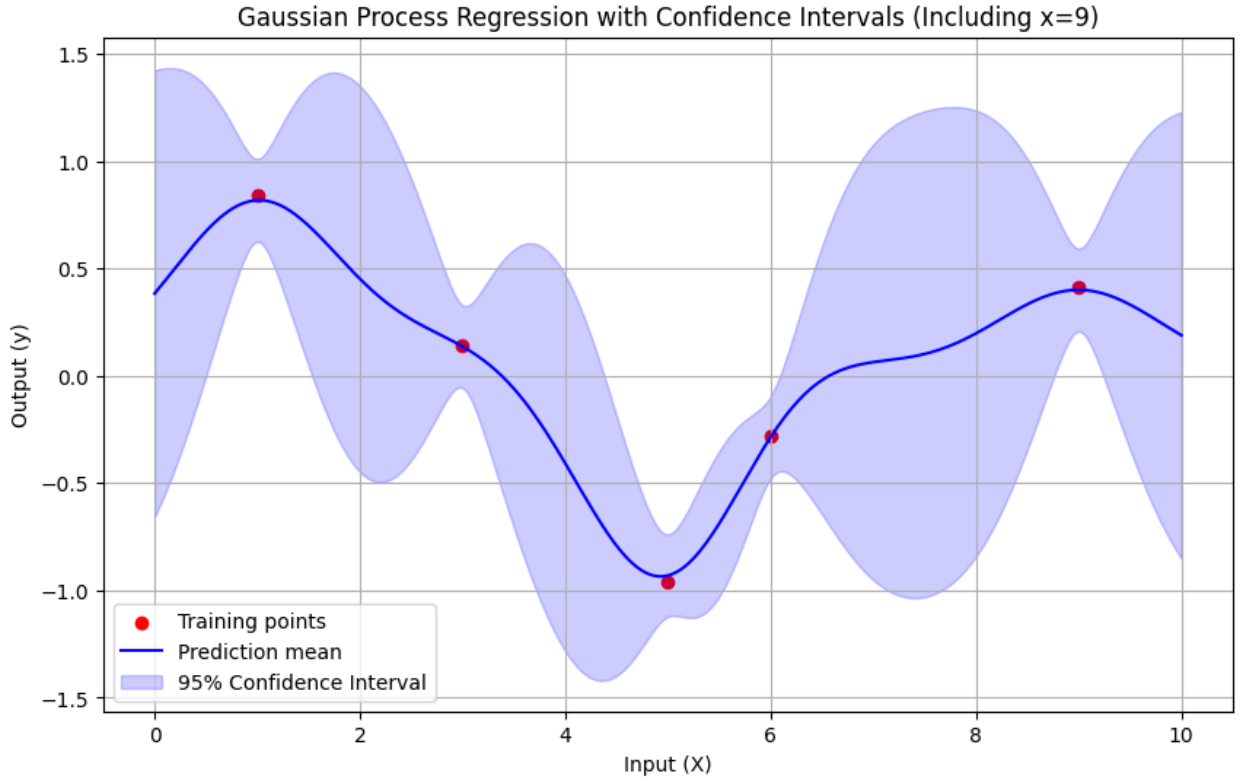


Figure 2: This plot shows the Gaussian Process regression prediction mean and the 95% confidence interval, with the new point $x = 9$ included as part of the training data. The variance decreases around $x = 8$, as the GP model has incorporated more data, providing tighter confidence bounds for predictions near known points

## Next Lecture

The next lecture will cover the following topics:

(i) Online (Sequential) Learning

(ii) Halving Algorithm (HA)

(iii Weighted Majority Algorithm (WMA)

## References:


1. Gaussian Processes, Lecture Notes from CS229: Machine Learning, 2008 Fall [Link]
2. Lecture 24, Lecture Notes from ESE 680-004: Learning and Control, 2019 Fall [Link]
3. Gaussian Processes visual explanation [Link1] and a more mathematical explanation [Link2]
4. Reference for Proofs involving Gaussian properties [Link]