Gaussian Process: Mathematical Formulation

1. Definition of Gaussian Processes

A Gaussian Process (GP) is defined as a distribution over functions:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

where:

- $m(x)$ is the **mean function**, defined as:

$$m(x) = \mathbb{E}[f(x)]$$

- $k(x, x')$ is the **covariance function** or **kernel**, defined as:

$$k(x, x') = \mathrm{Cov}(f(x), f(x'))$$

2. Prior Distribution

Given a set of input points $X = \{x_1, x_2, \ldots, x_n\}$ and corresponding function values $f(X)$, the prior distribution is:

$$f(X) \sim \mathcal{N}(m(X), K(X, X))$$

where:

- $m(X)$ is an n-dimensional vector of mean values.
- $K(X, X)$ is an $n \times n$ covariance matrix computed using the kernel function $k(x_i, x_j)$.

For example, for two points $X = [x_1, x_2]$, the prior distribution is:

$$f(X) \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ m(x_2) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{bmatrix}\right)$$

3. Posterior Distribution

After observing data points (X, y), we update our belief using Bayes' Theorem.

Given a new input x_*, the joint distribution of the observed function values y and the predicted value f_* is:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(X) \\ m(x_*) \end{bmatrix}, \begin{bmatrix} K(X,X) & K(X,x_*) \\ K(x_*,X) & K(x_*,x_*) \end{bmatrix}\right)$$

where:

- $K(X, X)$ is the covariance matrix for the observed points.
- $K(X, x_*)$ is the covariance vector between the observed points and the new point.
- $K(x_*, X) = K(X, x_*)^\top$.
- $K(x_*, x_*)$ is the variance at the new point.

4. Gaussian Process Regression (Derivation)

The goal of Gaussian Process Regression (GPR) is to find the **posterior distribution** over the function values at a new point x_* given the observed data (X, y).

Using the properties of multivariate Gaussian distributions, the **posterior mean** and **posterior variance** are derived as:

Posterior Mean

$$\mu_* = m(x_*) + K(x_*, X)K(X, X)^{-1}(y - m(X))$$

Assuming a zero mean function $m(x) = 0$, this simplifies to:

$$\mu_* = K(x_*, X)K(X, X)^{-1}y$$

Posterior Variance

$$\sigma_*^2 = K(x_*, x_*) - K(x_*, X)K(X, X)^{-1}K(X, x_*)$$

5. Squared Exponential Kernel

The **Squared Exponential (SE) kernel** is one of the most commonly used kernels:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$

where:

- σ_f^2 is the variance of the function.
- ℓ is the length scale, controlling how quickly the function values can change.