

Ch2. Gaussian Processes (Part 1)

Bayesian Optimization Seminar

Sungwoo Park

Uncertainty Quantification Lab
Seoul National University

January 22, 2026

Overview

1. Review of Chapter 1: Introduction
2. Definition and Basic Properties
3. Inference with Exact and Noisy Observations
4. Joint Gaussian Processes
5. Summary

Contents

1. Review of Chapter 1: Introduction
2. Definition and Basic Properties
3. Inference with Exact and Noisy Observations
4. Joint Gaussian Processes
5. Summary

The Optimization Problem

Goal

Find the global optimum of an objective function $f : \mathcal{X} \rightarrow \mathbb{R}$:

$$x^* = \arg \max_{x \in \mathcal{X}} f(x); \quad f^* = \max_{x \in \mathcal{X}} f(x)$$

Setting:

- Objective function f is **expensive** to evaluate (time, cost, resources)
- We can only access f through **sequential observations**
- Observations may be **noisy**: $y = f(x) + \varepsilon$

Challenge: How do we decide where to observe next, given limited budget?

Observation Model

Observations are realized by a stochastic mechanism:

$$p(y | x, \phi), \quad \text{where } \phi = f(x)$$

Common model — Additive Gaussian noise:

$$y = \phi + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_n^2)$$

$$\Rightarrow p(y | x, \phi, \sigma_n) = \mathcal{N}(y; \phi, \sigma_n^2)$$

Assumption: Multiple observations are **conditionally independent** given the objective function values:

$$p(\mathbf{y} | \mathbf{x}, \phi) = \prod_i p(y_i | x_i, \phi_i)$$

The Bayesian Approach: Key Idea

Core Principle

Treat the unknown objective function f as a **random variable** and use **Bayesian inference** to reason about it.

Bayesian Inference Refresher:

1. Start with a **prior** $p(\phi | x)$ encoding initial beliefs
2. Observe data $\mathcal{D} = (x, y)$ via the **likelihood** $p(y | x, \phi)$
3. Update to the **posterior** via Bayes' rule:

$$p(\phi | \mathcal{D}) = \frac{p(y | x, \phi) p(\phi | x)}{p(y | x)}$$

The posterior captures what we now believe about ϕ after seeing the data.

Inference of the Objective Function

To reason about the *entire* objective function $f : \mathcal{X} \rightarrow \mathbb{R}$, we need a **stochastic process** — a probability distribution over functions.

Specifying a Stochastic Process

We specify the distribution of function values $\phi = f(\mathbf{x})$ for any finite set of locations $\mathbf{x} \subset \mathcal{X}$:

$$p(\phi | \mathbf{x})$$

Gaussian Processes: The family where all such finite-dimensional distributions are **multivariate Gaussian** — mathematically convenient and widely used in Bayesian optimization.

Posterior Predictive Distribution

After observing data \mathcal{D} , we can predict the outcome of a new observation at x :

Posterior Predictive Distribution

$$p(y' | x, \mathcal{D}) = \int p(y' | x, \phi) p(\phi | x, \mathcal{D}) d\phi$$

Interpretation:

- Integrates over all possible values of $\phi = f(x)$
- Weights by their plausibility under the posterior
- Naturally accounts for **uncertainty** in the objective function

This distribution is *instrumental* for making informed decisions about where to observe next.

Contents

1. Review of Chapter 1: Introduction
2. Definition and Basic Properties
3. Inference with Exact and Noisy Observations
4. Joint Gaussian Processes
5. Summary

What is a Gaussian Process?

A **Gaussian process (GP)** extends the multivariate normal distribution to model functions on infinite domains.

Key Idea

We model an objective function $f : \mathcal{X} \rightarrow \mathbb{R}$ as an infinite collection of random variables, one for each point in the domain. The **Kolmogorov extension theorem** allows us to specify this distribution through finite-dimensional marginals.

GPs inherit convenient mathematical properties of the multivariate normal distribution while remaining computationally tractable.

Recall: Kolmogorov's Extension Theorem

Theorem: Kolmogorov's Extension Theorem [Durrett, 2019]

Suppose we are given probability measures μ_n on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ that are *consistent*, that is,

$$\mu_{n+1}((a_1, b_1] \times \cdots \times (a_{n+1}, b_{n+1}] \times \mathbf{R}) = \mu_n((a_1, b_1] \times \cdots \times (a_n, b_n])$$

Then there is a unique probability measure P on $(\mathbb{R}^{\mathbf{N}}, \mathcal{B}(\mathbb{R}^{\mathbf{N}}))$ with

$$P(\omega : \omega_i \in (a_i, b_i], 1 \geq i \geq n) = \mu_n((a_1, b_1] \times \cdots \times (a_n, b_n])$$

where $\mathbf{N} = \{1, 2, \dots\}$ and $\mathcal{B}(\mathbb{R}^{\mathbf{N}}) = \{(\omega_1, \omega_2, \dots) : \omega_i \in \mathcal{B}(\mathbb{R})\}$

For GPs, consistency is automatically satisfied because the MVN satisfies them (the marginal of a MVN is MVN with the corresponding submatrix of the covariance).

GP Specification: Mean and Covariance Functions

A GP on f is specified by:

$$p(f) = \mathcal{GP}(f; \mu, K)$$

- **Mean function** $\mu : \mathcal{X} \rightarrow \mathbb{R}$: determines the expected function value

$$\mu(x) = \mathbb{E}[\phi | x], \quad \text{where } \phi = f(x)$$

- **Covariance function (kernel)** $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$: encodes the correlation structure

$$K(x, x') = \text{cov}[\phi, \phi' | x, x'], \quad \text{where } \phi' = f(x')$$

The covariance function must be **symmetric** and **positive semidefinite**.

Finite-Dimensional Marginals

For any finite set of points $\mathbf{x} \subset \mathcal{X}$, the corresponding function values $\phi = f(\mathbf{x})$ follow a multivariate normal distribution:

$$p(\phi \mid \mathbf{x}) = \mathcal{N}(\phi; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\mu} = \mathbb{E}[\phi \mid \mathbf{x}] = \mu(\mathbf{x}); \quad \boldsymbol{\Sigma} = \text{cov}[\phi \mid \mathbf{x}] = K(\mathbf{x}, \mathbf{x})$$

Gram Matrix

$K(\mathbf{x}, \mathbf{x})$ is the matrix formed by evaluating K for each pair of points:

$$\Sigma_{ij} = K(x_i, x_j)$$

Example: Squared Exponential Covariance

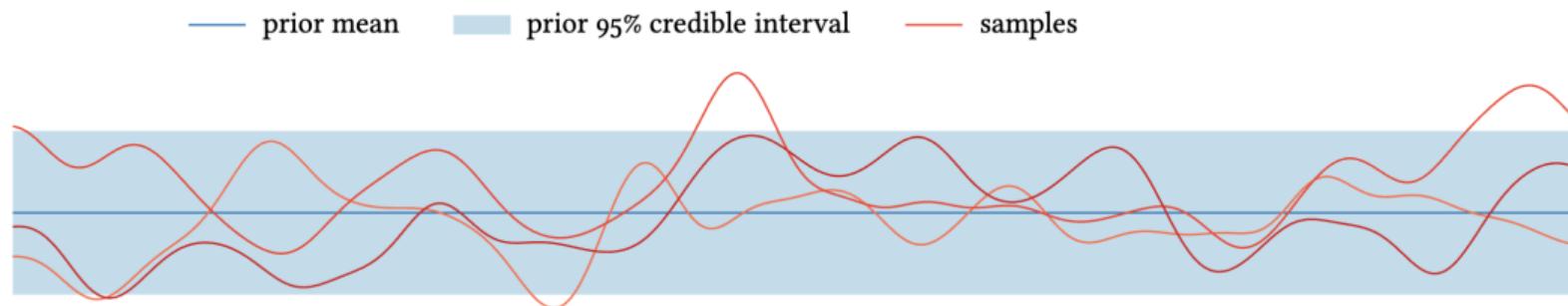


Figure 1: Example of Gaussian Process [Garnett, 2023]

Consider $\mathcal{X} = [0, 30]$ with:

- Mean function: $\mu \equiv 0$ (constant central tendency)
- Covariance function (squared exponential):

$$K(x, x') = \exp\left(-\frac{1}{2}|x - x'|^2\right)$$

Example: Squared Exponential Covariance

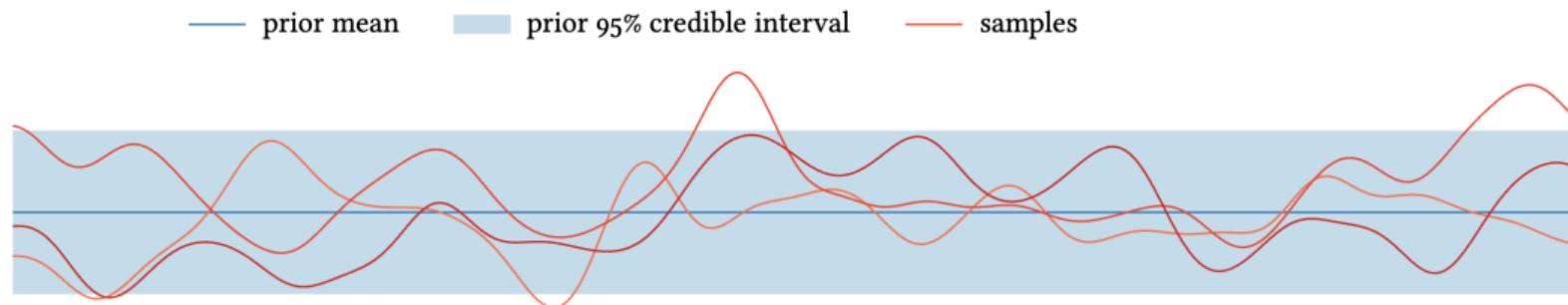


Figure 1: Example of Gaussian Process [Garnett, 2023]

Properties:

- $\text{var}[\phi | x] = K(x, x) = 1$ at every point
- Correlation decreases with distance: nearby values are highly correlated, distant values are nearly independent
- This encodes a statistical notion of **continuity**

Sampling from a Gaussian Process (Appendix A.2)

To sample from a GP with mean μ and covariance K :

1. Choose a finite grid of points $\mathbf{x} = (x_1, \dots, x_n)$
2. Compute $\boldsymbol{\mu} = \mu(\mathbf{x})$ and $\boldsymbol{\Sigma} = K(\mathbf{x}, \mathbf{x})$
3. Factor: $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$ (Cholesky decomposition)
4. Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5. Compute $\boldsymbol{\phi} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z}$

The resulting sample $\boldsymbol{\phi}$ represents function values at the chosen grid points, respecting the correlation structure encoded by K .

Contents

1. Review of Chapter 1: Introduction
2. Definition and Basic Properties
- 3. Inference with Exact and Noisy Observations**
4. Joint Gaussian Processes
5. Summary

General Framework: Jointly Gaussian Observations

We can condition a GP $p(f) = \mathcal{GP}(f; \mu, K)$ on any vector \mathbf{y} sharing a joint Gaussian distribution with f :

$$p(f, \mathbf{y}) = \mathcal{GP} \left(\begin{bmatrix} f \\ \mathbf{y} \end{bmatrix}; \begin{bmatrix} \mu \\ \mathbf{m} \end{bmatrix}, \begin{bmatrix} K & \boldsymbol{\kappa}^\top \\ \boldsymbol{\kappa} & \mathbf{C} \end{bmatrix} \right)$$

where:

- $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mathbf{m}, \mathbf{C})$: marginal distribution of observations
- $\boldsymbol{\kappa}(x) = \text{cov}[\mathbf{y}, \phi | x]$: cross-covariance function

Posterior Gaussian Process

Conditioning on observations $\mathcal{D} = \mathbf{y}$ yields a GP posterior:

$$p(f \mid \mathcal{D}) = \mathcal{GP}(f; \mu_{\mathcal{D}}, K_{\mathcal{D}})$$

Posterior Mean and Covariance

$$\mu_{\mathcal{D}}(x) = \mu(x) + \kappa(x)^{\top} \mathbf{C}^{-1} (\mathbf{y} - \mathbf{m})$$

$$K_{\mathcal{D}}(x, x') = K(x, x') - \kappa(x)^{\top} \mathbf{C}^{-1} \kappa(x')$$

Inference procedure:

1. Compute marginal distribution of \mathbf{y}
2. Derive cross-covariance function κ
3. Apply the posterior formulas

Handling Additive Gaussian Noise

Suppose we observe $\mathbf{z} = \mathbf{y} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{N})$ is independent noise.

Then:

$$p(\mathbf{z} \mid \mathbf{N}) = \mathcal{N}(\mathbf{z}; \mathbf{m}, \mathbf{C} + \mathbf{N}); \quad \text{cov}[\mathbf{z}, \phi \mid \mathbf{x}] = \kappa(\mathbf{x})$$

Key Result

Simply replace \mathbf{C} with $\mathbf{C} + \mathbf{N}$ in the posterior formulas!

As $\mathbf{N} \rightarrow \mathbf{0}$, the posterior converges to that from direct observation of \mathbf{y} .

Inference with Exact Function Evaluations

Suppose we observe f at locations \mathbf{x} , revealing $\phi = f(\mathbf{x})$.

The posterior is $p(f \mid \mathcal{D}) = \mathcal{GP}(f; \mu_{\mathcal{D}}, K_{\mathcal{D}})$ with:

$$\mu_{\mathcal{D}}(x) = \mu(x) + K(x, \mathbf{x})\Sigma^{-1}(\phi - \mu)$$

$$K_{\mathcal{D}}(x, x') = K(x, x') - K(x, \mathbf{x})\Sigma^{-1}K(\mathbf{x}, x')$$

where $\Sigma = K(\mathbf{x}, \mathbf{x})$ and $\mu = \mu(\mathbf{x})$.

Key properties:

- Posterior mean **interpolates** through observed points
- Posterior variance **vanishes** at observed locations
- Uncertainty remains unchanged far from observations

Inference with Noisy Function Evaluations

Suppose observations are corrupted: $\mathbf{y} = \phi + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{N})$.

Common noise models:

- **Homoskedastic:** $\mathbf{N} = \sigma_n^2 \mathbf{I}$ (constant noise)
- **Heteroskedastic:** $\mathbf{N} = \text{diag}(\sigma_n^2(\mathbf{x}))$ (location-dependent)

The posterior formulas become:

$$\begin{aligned}\mu_{\mathcal{D}}(x) &= \mu(x) + K(x, \mathbf{x})(\Sigma + \mathbf{N})^{-1}(\mathbf{y} - \mu) \\ K_{\mathcal{D}}(x, x') &= K(x, x') - K(x, \mathbf{x})(\Sigma + \mathbf{N})^{-1}K(\mathbf{x}, x')\end{aligned}$$

The posterior mean no longer interpolates exactly; extreme values may be “explained away” as noise.

Interpretation of Posterior Moments

Consider a single observation y with distribution $\mathcal{N}(y; m, s^2)$ and define:

- z-score: $z = \frac{y-m}{s}$
- Correlation: $\rho = \text{corr}[y, \phi \mid x] = \frac{\kappa(x)}{\sigma s}$

Posterior Moments (Scalar Case)

$$\text{Posterior mean of } \phi : \mu + \sigma \rho z$$

$$\text{Posterior std of } \phi : \sigma \sqrt{1 - \rho^2}$$

Intuition:

- Mean shifts proportionally to z-score and correlation strength
- Variance reduction depends only on correlation $|\rho|$

Posterior Predictive Distribution

For the latent function value $\phi = f(x)$:

$$p(\phi \mid x, \mathcal{D}) = \mathcal{N}(\phi; \mu_{\mathcal{D}}(x), K_{\mathcal{D}}(x, x))$$

For a **noisy observation** y at location x (with noise variance σ_n^2):

$$p(y \mid x, \mathcal{D}, \sigma_n) = \mathcal{N}(y; \mu_{\mathcal{D}}(x), K_{\mathcal{D}}(x, x) + \sigma_n^2)$$

The predictive credible intervals for noisy measurements are inflated compared to the latent function, reflecting observation uncertainty.

Contents

1. Review of Chapter 1: Introduction
2. Definition and Basic Properties
3. Inference with Exact and Noisy Observations
- 4. Joint Gaussian Processes**
5. Summary

Topics Covered in Remaining Sections

The remainder of Chapter 2 covers more specialized topics:

- **§2.4 Joint Gaussian Processes:** Modeling multiple correlated functions
- **§2.5 Continuity:** Conditions for continuous sample paths
- **§2.6 Differentiability:** Conditions for differentiable sample paths; derivative observations
- **§2.7 Existence/Uniqueness of Global Maxima:** Theoretical guarantees
- **§2.8 Non-Gaussian Observations:** Approximate inference methods

Key takeaway: Sections 2.1–2.2 provide sufficient foundation for most practical Bayesian optimization applications!

Motivation: Modeling Multiple Functions

In some settings, we need to jointly reason about **multiple related functions**:

- An objective function and its gradient
- An expensive objective and cheaper surrogates (multifidelity)
- Multiple objectives (multiobjective optimization)

Key Idea

“Paste together” multiple functions into a single function on a larger domain, then construct a standard GP on this combined function.

Definition of Joint Gaussian Process

Consider functions $\{f_i : \mathcal{X}_i \rightarrow \mathbb{R}\}$. Define the **disjoint union**:

$$\bigsqcup f : \mathcal{X} \rightarrow \mathbb{R}, \quad \mathcal{X} = \bigsqcup \mathcal{X}_i$$

such that $\bigsqcup f|_{\mathcal{X}_i} \equiv f_i$.

A **joint Gaussian process** is a GP on $\bigsqcup f$:

$$p(\bigsqcup f) = \mathcal{GP}(\bigsqcup f; \mu, K)$$

The mean and covariance functions on \mathcal{X} encode both:

- Marginal behavior of each function
- Cross-correlations between functions

Decomposed Notation

For two functions $f : \mathcal{F} \rightarrow \mathbb{R}$ and $g : \mathcal{G} \rightarrow \mathbb{R}$:

$$p(f, g) = \mathcal{GP}\left(\begin{bmatrix} f \\ g \end{bmatrix}; \begin{bmatrix} \mu_f \\ \mu_g \end{bmatrix}, \begin{bmatrix} K_f & K_{fg} \\ K_{gf} & K_g \end{bmatrix}\right)$$

Components:

- μ_f, K_f and μ_g, K_g : marginal GP parameters
- $K_{fg}(x, x') = \text{cov}[\phi, \gamma \mid x, x']$: cross-covariance
- $K_{gf} = K_{fg}^\top$

Marginal property: Each function has a marginal GP distribution:

$$p(f) = \mathcal{GP}(f; \mu_f, K_f); \quad p(g) = \mathcal{GP}(g; \mu_g, K_g)$$

Example: Correlated Functions

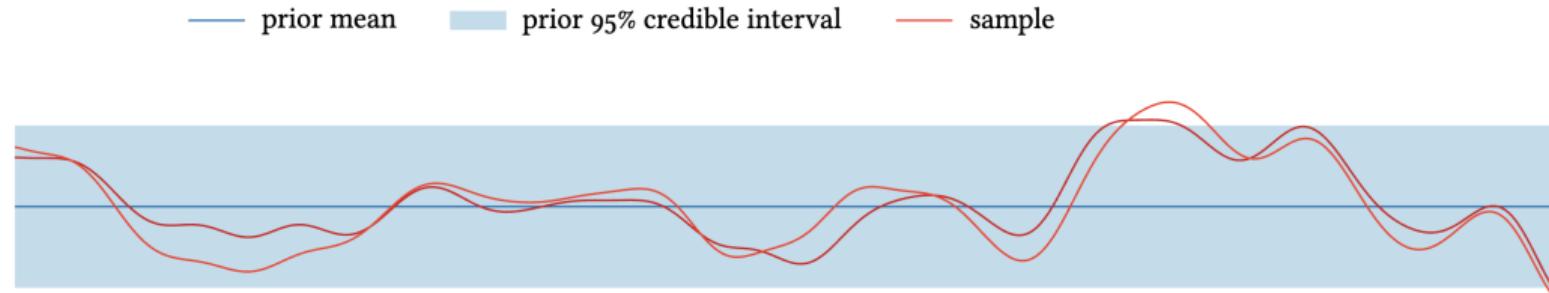


Figure 2: Example of Joint Gaussian Process [Garnett, 2023]

Consider $f, g : [0, 30] \rightarrow \mathbb{R}$ with:

- Same marginal: $\mu \equiv 0$, squared exponential covariance K
- Cross-covariance: $K_{fg}(x, x') = 0.9 \cdot K(x, x')$

Example: Correlated Functions

— prior mean ■ prior 95% credible interval — sample

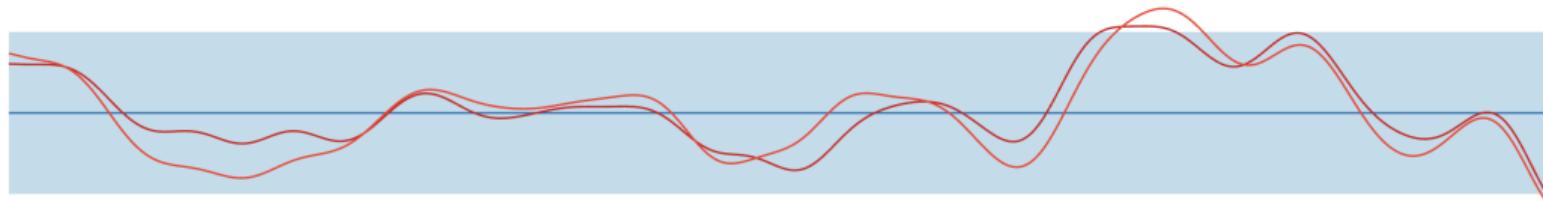


Figure 2: Example of Joint Gaussian Process [Garnett, 2023]

For any point x , the correlation between $\phi = f(x)$ and $\gamma = g(x)$ is:

$$\text{corr}[\phi, \gamma | x] = 0.9$$

Consequence: Samples from the joint distribution show strong coupling, the functions “move together.”

Inference for Joint GPs

The joint GP construction allows us to condition on observations of **any** of the functions using the standard inference procedure.

Examples

Given observations of f on the left side of the domain and observations of g on the right side:

- Observations of f inform our belief about g (and vice versa)
- Information propagates through the cross-covariance structure
- Strong correlation \Rightarrow strong information transfer

This is particularly useful for **multifidelity optimization**: cheap surrogate evaluations inform our belief about the expensive objective.

Extension to Vector-Valued Functions

A GP on a vector-valued function $\mathbf{f} : \mathcal{X} \rightarrow \mathbb{R}^d$ is defined by a joint GP on its coordinate functions $\{f_i\} : \mathcal{X} \rightarrow \mathbb{R}$.

Notation: $\mathcal{GP}(\mathbf{f}; \mu, K)$ where:

- $\mu : \mathcal{X} \rightarrow \mathbb{R}^d$ (vector-valued mean)
- $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{d \times d}$ (matrix-valued covariance)

Applications:

- Joint distribution of f and ∇f (gradient)
- Multiobjective optimization with correlated objectives
- Modeling spatial vector fields

Contents

1. Review of Chapter 1: Introduction
2. Definition and Basic Properties
3. Inference with Exact and Noisy Observations
4. Joint Gaussian Processes
5. Summary

Summary of Key Ideas

1. **GP Definition:** Specified by mean μ and covariance K functions; finite marginals are multivariate Gaussian
2. **Exact Inference:** Conditioning on jointly Gaussian observations yields a GP posterior with closed-form mean and covariance
3. **Noisy Inference:** Replace \mathbf{C} with $\mathbf{C} + \mathbf{N}$ to handle additive Gaussian noise
4. **Posterior Interpretation:** Mean update \propto (correlation \times z-score); variance reduction depends on correlation strength
5. **Joint GPs:** Model multiple correlated functions; enable information sharing across related tasks

References

-  Durrett, R. (2019).
Probability: Theory and examples.
Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
-  Garnett, R. (2023).
Bayesian optimization.
Cambridge University Press.

Thank You