

# 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  $\phi$  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

## Question

What is the *consistency* property in Kolmogorov's Extension Theorem?

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

Suppose we are given probability measures  $\mu_n$  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  $\mathcal{B}(\mathbb{R}^{\mathbf{N}}) = \{(\omega_1, \omega_2, \dots) : \omega_i \in \mathcal{B}(\mathbb{R})\}$

In [Durrett, 2019], theorem is not complete to use in our scenario because  $P$  defines on countable index set  $\mathbf{N}$ .

# Q&A: Countable vs. Uncountable Domains

## Question

What are the characteristics of modeling on countable vs uncountable domains?

## Definition: Sample path [Le Gall, 2016]

Let  $(X_t)_{t \in \mathcal{X}}$  be a random process with values in  $E$ . The *sample paths* of  $X$  are the mappings  $\mathcal{X} \ni t \mapsto X_t(\omega)$  obtained when fixing  $\omega \in \Omega$ . The sample paths of  $X$  thus form a collection of mappings from  $\mathcal{X}$  into  $E$  indexed by  $\omega \in \Omega$ .

## Countable infinite domain (e.g., $\mathcal{X} = \mathbb{Z}$ ):

- GP reduces to specifying consistent MVN distributions on all finite subsets
- Sample paths are sequences  $\{f(x_i)\}_{i=1}^{\infty}$
- No notion of “continuity” in the classical sense

# Q&A: Countable vs. Uncountable Domains

## Question

What are the characteristics of modeling on countable vs uncountable domains?

## Definition: Sample path [Le Gall, 2016]

Let  $(X_t)_{t \in \mathcal{X}}$  be a random process with values in  $E$ . The *sample paths* of  $X$  are the mappings  $\mathcal{X} \ni t \mapsto X_t(\omega)$  obtained when fixing  $\omega \in \Omega$ . The sample paths of  $X$  thus form a collection of mappings from  $\mathcal{X}$  into  $E$  indexed by  $\omega \in \Omega$ .

## Uncountable infinite domain (e.g., $\mathcal{X} = \mathbb{R}^d$ ):

- Sample paths are actual functions
- Continuity, differentiability become meaningful properties
- Requires careful treatment (measurability issues, sample path properties)

# Kolmogorov Extension Theorem: The Foundation

## Question

What is the *consistency* property in Kolmogorov's Extension Theorem?

Kolmogorov's Extension Theorem : Stochastic process ver. [Oksendal, 2003]

For all  $t_1, \dots, t_k \in T$ ,  $k \in \mathbb{N}$  let  $\nu_{t_1, \dots, t_k}$  be probability measures on  $\mathbb{R}^{nk}$  s.t.

$$\nu_{t_{\sigma(1)}, \dots, t_{\sigma(k)}}(F_1 \times \cdots \times F_k) = \nu_{t_1, \dots, t_k}(F_{\sigma^{-1}(1)} \times \cdots \times F_{\sigma^{-1}(k)}) \quad (\text{K1})$$

for all permutations  $\sigma$  on  $\{1, 2, \dots, k\}$  and

$$\nu_{t_1, \dots, t_k}(F_1 \times \cdots \times F_k) = \nu_{t_1, \dots, t_k, t_{k+1}, \dots, t_{k+m}}(F_1 \times \cdots \times F_k \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n) \quad (\text{K2})$$

for all  $m \in \mathbb{N}$ .

# Kolmogorov Extension Theorem: The Foundation

Kolmogorov's Extension Theorem : Stochastic process ver. (cont.) [Oksendal, 2003]

Then there exists a probability space  $(\Omega, \mathcal{F}, P)$  and a stochastic process  $\{X_t\}$  on  $\Omega$ ,  
 $X_t : \Omega \rightarrow \mathbb{R}^n$ , s.t.

$$\nu_{t_1, \dots, t_k}(F_1 \times \cdots \times F_k) = P[X_{t_1} \in F_1, \dots, X_{t_k} \in F_k],$$

for all  $t_i \in T$ ,  $k \in \mathbb{N}$  and all Borel sets  $F_i$ .

## Consistency Conditions:

1. **Permutation invariance** (K1): The joint distribution is unchanged by reordering indices
2. **Marginalization consistency** (K2): If  $x \subset x'$ , then marginalizing  $p(\phi' | x')$  over  $\phi' \setminus \phi$  yields  $p(\phi | x)$

For GPs, these are 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**.

# Q&A: Extension from Countable Dense Subsets

## Question

If  $f$  is a.s. continuous and  $D \subset \mathcal{X}$  is countable dense, can  $p(f | D)$  determine  $p(f)$ ?

## Proof.

Let  $f$  and  $g$  be two continuous functions on  $\mathcal{X}$ . Suppose  $f(x) = g(x)$  for  $\forall x \in D$ . For any  $x \in \mathcal{X}$ , there exists a sequence  $\{q_n\} \subset D$  such that  $q_n \rightarrow x$ . ( $\because D$  is dense in  $\mathcal{X}$ ) By continuity,

$$f(x) = f\left(\lim_{n \rightarrow \infty} q_n\right) = \lim_{n \rightarrow \infty} f(q_n) = \lim_{n \rightarrow \infty} g(q_n) = g\left(\lim_{n \rightarrow \infty} q_n\right) = g(x)$$

Thus,  $f(x) = g(x), \forall x \in \mathcal{X}$ . Since sample paths are a.s. continuous, the process is uniquely determined by its values on  $D$ .



# 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:

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

where

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

## 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

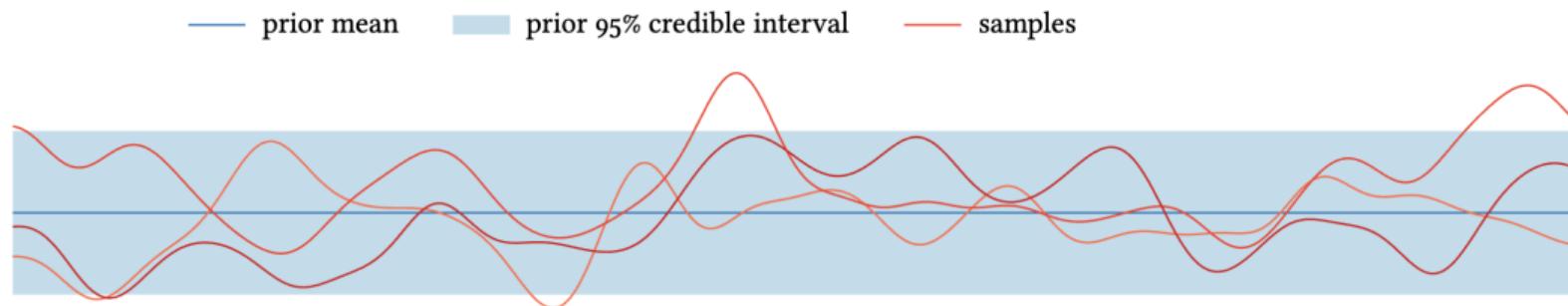


Figure 1: Example of Gaussian Process [Garnett, 2023]

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

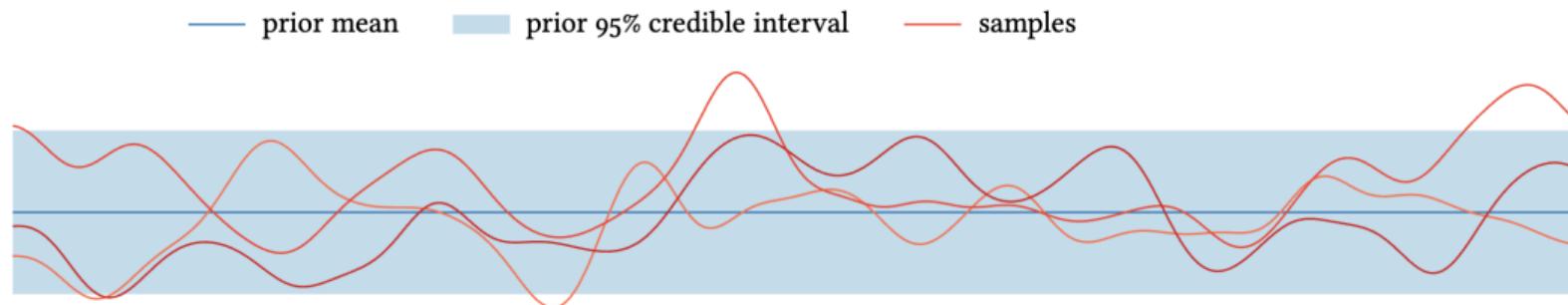


Figure 1: Example of Gaussian Process [Garnett, 2023]

## 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  $\mu$  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

# Q&A: What Observations $\mathbf{y}$ Are Jointly Gaussian with $f$ ?

## Question

Besides function values, what else can  $\mathbf{y}$  be?

Any of the following are jointly Gaussian with a GP:

- Function values:  $\mathbf{y} = f(\mathbf{x})$  (the basic case)
- Affine transformations:  $\mathbf{y} = \mathbf{A}f(\mathbf{x}) + \mathbf{b}$
- Limits of affine transformations:
  - Partial derivatives:  $\frac{\partial f}{\partial x_i}(x)$
  - Integrals/expectations:  $\int f(x)p(x)dx$
- Any of the above + independent Gaussian noise

# Q&A: What Observations $\mathbf{y}$ Are Jointly Gaussian with $f$ ?

## Question

Besides function values, what else can  $\mathbf{y}$  be?

**Other examples:** Based on linear (affine) operator, these are joint Gaussian distribution.

- Line/Path Integral Observations [Särkkä, 2011]

$$\mathbf{y} = \int_{\gamma} f(x) ds + \boldsymbol{\varepsilon}, \quad \kappa(x') = \int_{\gamma} K(x, x') ds$$

Used at Tomographic reconstruction (CT/MRI), Measure expected path of moving sensor

- Convolved/Smoothed Observations by smoothing kernel  $G$  [Alvarez et al., 2011]

$$\mathbf{y} = (f * G)(x_0) = \int f(x') G(x_0 - x') dx', \quad \kappa(x) = \int G(x_0 - x') K(x', x) dx'$$

Used at Sensor spatial averaging, image blur (PSF), multi-output GP via convolution

# 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  $\kappa$
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  $\phi = f(\mathbf{x})$ . We know that  $p(\phi|\mathbf{x}) = \mathcal{N}(\phi; \mu, \Sigma)$  so  $\kappa(x) = \text{cov}[\phi, \phi|\mathbf{x}, x]$ .

The posterior is  $p(f | \mathcal{D}) = \mathcal{GP}(f; \mu_{\mathcal{D}}, K_{\mathcal{D}})$  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.

# Inference with Noisy Function Evaluations

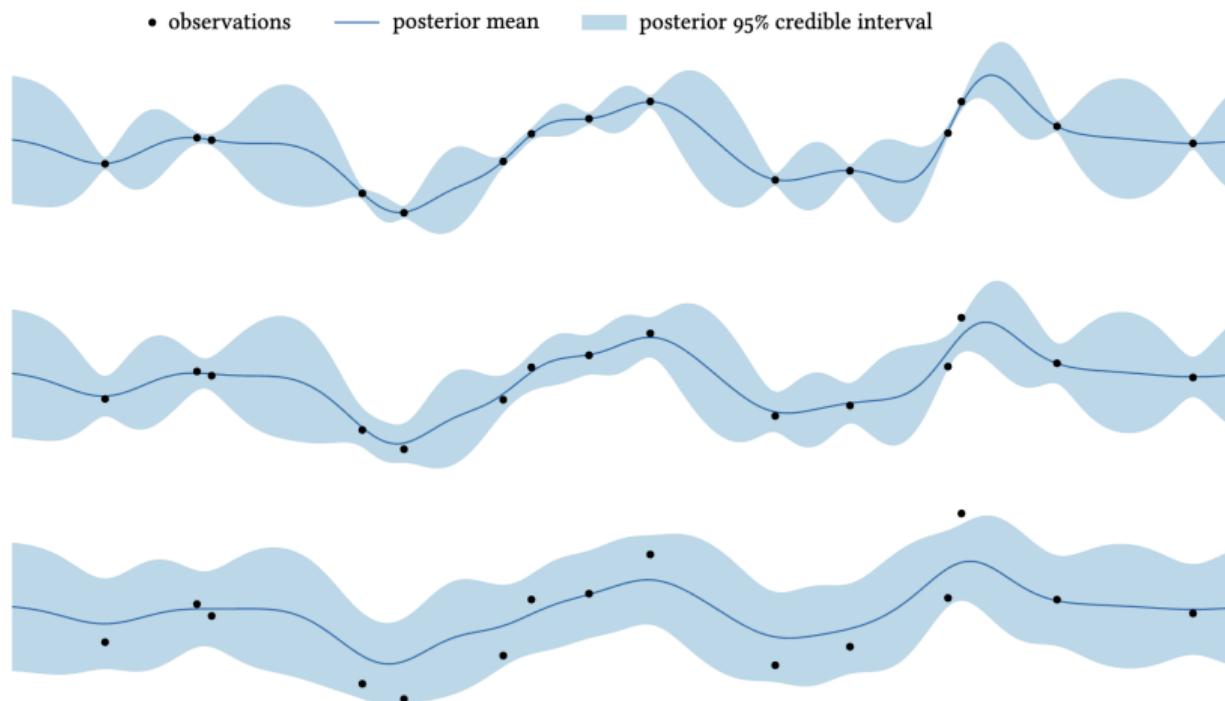


Figure 2: Conditioning GP on data corrupted by increasing levels of homoskedastic noise [Garnett, 2023]

# 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  $|\rho|$

# Interpretation of Posterior Moments: Vector Case

For vector-valued observations  $\mathbf{y}$ , factor the covariance as  $C = SPS$  where:

- $S$ : diagonal with  $S_{ii} = \sqrt{C_{ii}} = \text{std}[y_i]$
- $P = \text{corr}[\mathbf{y}]$ : observation correlation matrix

Define vectors of  $z$ -scores and cross-correlations:

$$z_i = \frac{y_i - m_i}{s_i}, \quad \rho_i = \frac{[\kappa(x)]_i}{\sigma s_i}$$

## Posterior Moments (Vector Case)

Posterior mean of  $\phi$  :  $\mu + \sigma \boldsymbol{\rho}^\top P^{-1} \mathbf{z}$

Posterior std of  $\phi$  :  $\sigma \sqrt{1 - \boldsymbol{\rho}^\top P^{-1} \boldsymbol{\rho}}$

# Q&A: Differential Entropy as Global Uncertainty

## Question

Why is differential entropy a measure of global uncertainty?

## Definition: Differential Entropy

For a random variable  $\omega$  with density  $p(\omega)$ , *differential entropy* is defined as:

$$H[\omega] = - \int p(\omega) \log p(\omega) d\omega$$

## For Gaussian distributions:

- Univariate:  $H[\mathcal{N}(\mu, \sigma^2)] = \frac{1}{2} \log(2\pi e \sigma^2)$
- Multivariate:  $H[\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})] = \frac{1}{2} \log |2\pi e \boldsymbol{\Sigma}|$

For GPs: The joint entropy of function values at locations  $\mathbf{x}$  is:

$$H[\phi \mid \mathbf{x}] = \frac{1}{2} \log |2\pi e K(\mathbf{x}, \mathbf{x})|$$

# Q&A: Why Differential Entropy Measures Global Uncertainty

**Key insight:** Entropy depends on the **determinant of the covariance matrix**.

**Why “Global”?**

- Marginal variance  $\sigma_i^2 = \Sigma_{ii}$ : uncertainty of **individual** variables
- Determinant  $|\Sigma|$ : **joint** uncertainty of **all** variables together
- Also reflects correlations: high correlation  $\Rightarrow$  smaller  $|\Sigma|$

# Q&A: Entropy Reduction in the Posterior

## Prior vs Posterior Covariance:

$$\text{Prior : } K(\mathbf{x}, \mathbf{x})$$

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

## Global Uncertainty Reduction

For positive semidefinite matrices  $A, B$ :  $|A| \leq |A + B|$ , hence:

$$|K_{\mathcal{D}}(\mathbf{x}, \mathbf{x})| \leq |K(\mathbf{x}, \mathbf{x})|$$

Therefore, posterior entropy is **always  $\leq$  prior entropy**.

## Interpretation:

- Independent observations ( $\rho = 0$ ): no entropy change
- More precise observations (larger  $C^{-1}$ ): greater entropy reduction  $\Rightarrow$  more information gained

# 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  $\sigma_n^2$ ):

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

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

# 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:

- $\mu_f, K_f$  and  $\mu_g, K_g$ : 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:

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

# Q&A: Designing Cross-Covariance Functions

## Question

How do we choose the cross-covariance function  $K_{fg}$ ?

### Common approaches:

1. **Scaled base kernel:**  $K_{fg}(x, x') = \rho \cdot K(x, x')$  where  $|\rho| < 1$ 
  - Simple, interpretable:  $\rho$  is the correlation at any point
2. **Linear Model of Coregionalization (LMC)** [Alvarez et al., 2011]:

$$K_{fg}(x, x') = \sum_{q=1}^Q a_q^{(f)} a_q^{(g)} K_q(x, x')$$

3. **Convolution processes:** Define via convolution with a smoothing kernel

**Constraint:** The full covariance matrix must remain positive semidefinite!

## Example: Correlated Functions

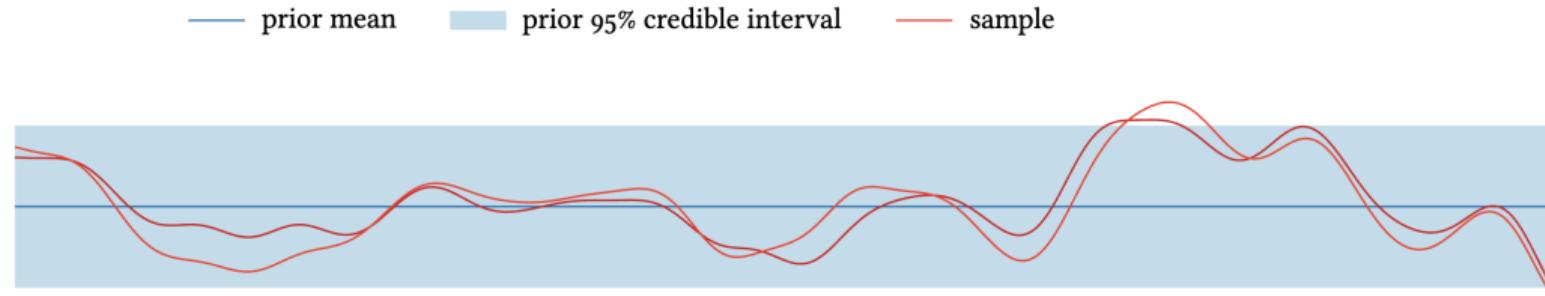


Figure 3: Example of Joint Gaussian Process [Garnett, 2023]

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

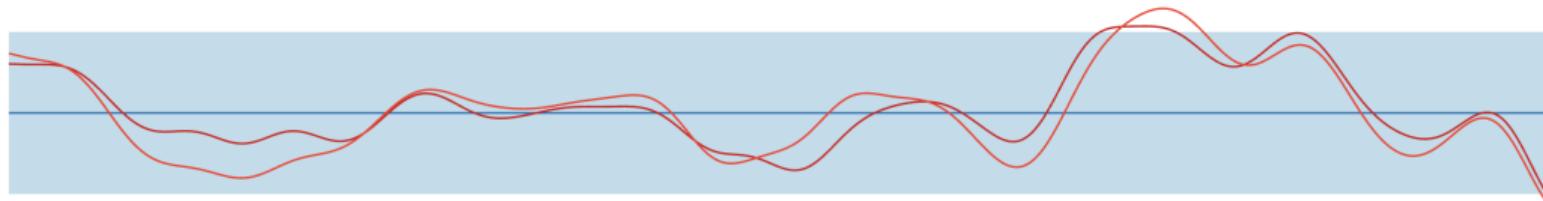


Figure 3: Example of Joint Gaussian Process [Garnett, 2023]

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.”

# Q&A: Why Disjoint Union for Joint GP Domains?

## Question

Is disjoint union the standard way to define joint GP domains?

Yes, mathematically this is the cleanest approach:

$$\bigsqcup f : \mathcal{X} = \bigsqcup_i \mathcal{X}_i \rightarrow \mathbb{R}, \quad \text{where } (\bigsqcup f)|_{\mathcal{X}_i} \equiv f_i$$

## Why disjoint union?

- Ensures each point  $(x, i)$  uniquely identifies *which* function
- In previous example, if  $\mathcal{X}_1 = \mathcal{X}_2 = [0, 30]$ , we distinguish  $f(10)$  from  $g(10)$ 
  - 엄밀하게는 multi-output GP
- The kernel can then define both within-function and cross-function covariance

# Q&A: Domain Size and Function Influence in Joint GPs

## Question

Can domain size differences affect function influence in joint GPs?

Yes, this is a real concern in practice!

- Function with more observations may dominate inference
- Different scales of domains may require different length scales
- Numerical conditioning can suffer from imbalanced data

## Mitigation strategies:

1. **Output scaling:** Normalize each function to similar variance
2. **Weighted likelihoods:** Down-weight abundant data sources
3. **Hierarchical priors:** Place priors on correlation parameters

# 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.

# 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  $\mu$  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 I

-  Alvarez, M. A., Rosasco, L., and Lawrence, N. D. (2011).  
Kernels for vector-valued functions: A review.
-  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.
-  Le Gall, J.-F. (2016).  
*Brownian motion, martingales, and stochastic calculus.*  
Graduate texts in mathematics. Springer International Publishing, 2016 edition.
-  Oksendal, B. (2003).  
*Stochastic differential equations: An introduction with applications.*  
Universitext. Springer.

## References II



Särkkä, S. (2011).

*Linear operators and stochastic partial differential equations in Gaussian process regression*, pages 151–158.

Lecture Notes in Computer Science. Springer Berlin Heidelberg.

# Thank You