

# IEE575 - Applied Stochastic Operations Research Models

## Lab 1: Gaussian Process Regressor

Tanmay Khandait  
tkhandai@asu.edu

April 16, 2025

# Hello, World!

- Hi! I'm **Tanmay Khandait**, your TA for the lab sessions.
- I'm a PhD student working on machine learning, especially probabilistic models.
- Ask questions anytime—I'm here to help!
- Labs can be chill but insightful—let's make GPs less mysterious together.
- You can reach me via , or catch me during office hours.  
I will let you all know the time in the announcements :).

- April 07 (Monday): Introduction to Python, Foundations of Gaussian Process (GP) Regressors
- April 14 (Monday): Deep Dive into GP Regressors, Starting Implementation in Python
- April 16 (Wednesday): Coding Gaussian Process Regressors
- April 21 (Monday): Bayesian Optimization (BO) Concepts, Integrating GP within the BO Framework
- April 23 (Wednesday): Coding Bayesian Optimization, Practical Applications of BO
- April 28 (Monday): Practical Application of BO, Lab Exam Prep
- April 30 (Wednesday): Lab Exam Day

- Bring Your Laptop + Charger to Every Session
- Set Up Python/Google Colab
  - Install Python (with NumPy, Matplotlib, scikit-learn) **or**
  - Use Google Colab: [colab.research.google.com](https://colab.research.google.com)
- We will be following this tutorial for the Gaussian Process Regressors.
  - Gaussian Process Tutorial — [arxiv.org/abs/2009.10862](https://arxiv.org/abs/2009.10862)
- **Mandatory Attendance**
  - Engage in discussions, ask questions, and collaborate!
  - Labs build toward the exam — attendance is 10% of your module grade.

## In the last class...

We covered a bunch of stuff

- Marginalization
- Conditioning

Today:

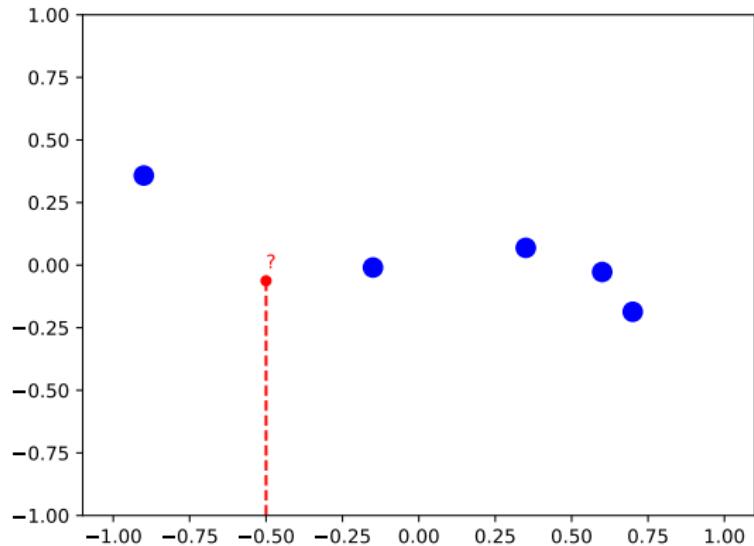
- We will implement GPR

Logistics

- Office Hours - Friday 9:00am - 11:00 am
- Exercises will be posted today
- We can discuss them in the office hours and on Monday

# Motivation

Consider the following regression problem:  
What would be the output be at the point in red?



## Motivation

Consider the following regression problem:

What would be the output be at the point in red?

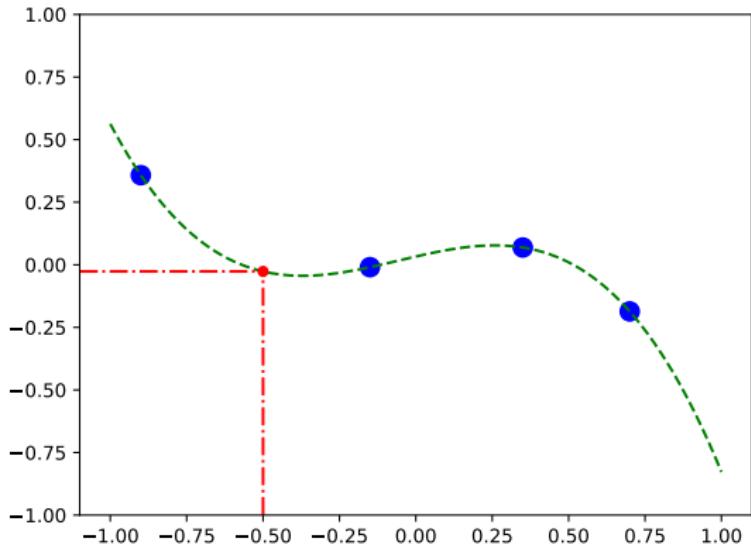


Figure: Here is an illustration of using non-linear regression.

## Motivation

Consider the following regression problem:

What would be the output be at the point in red?

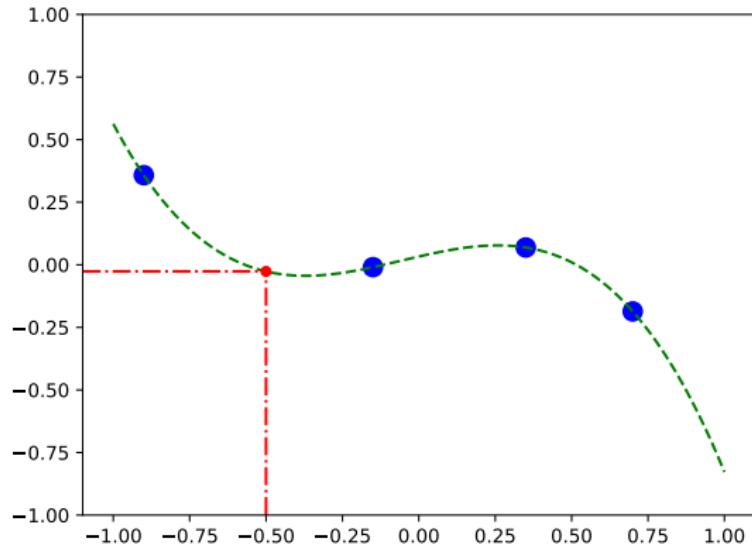


Figure: Here is an illustration of using non-linear regression.

However, in addition to this, we also need uncertainty estimates.

# Motivation

Can we do this using what we know about Gaussian Distributions?

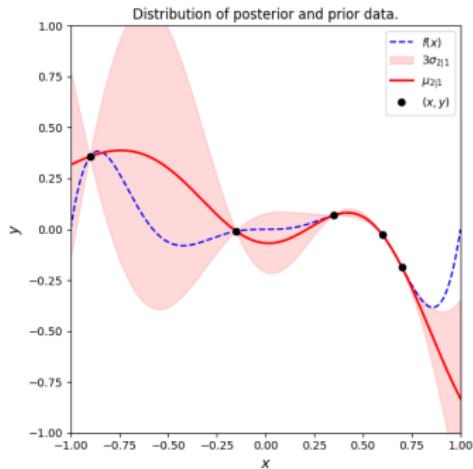


Figure: Here is an illustration of using non-linear regression.

# Univariate Gaussian Distribution

- A random variable  $X$  is normally distributed with mean  $\mu$  and variance  $\sigma^2$  if it has the probability density function of  $X$  as:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- In this expression, you see the squared difference between the variable  $x$  and its mean,  $\mu$ .
- This value will be minimized when  $x$  is equal to  $\mu$ .
- The quantity  $-\frac{x-\mu^2}{\sigma^2}$  will take its largest value when  $x$  is equal to  $\mu$  or likewise since the exponential function is a monotone function, the normal density takes a maximum value when  $x$  is equal to  $\mu$ .
- The variance  $\sigma^2$  defines the spread of the distribution about that maximum. If it is large, then the spread is going to be large, otherwise, if the value is small, then the spread will be small.
- If  $X$  is random variable that follows a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , then we will denote it as  $X \sim \mathcal{N}(\mu, \sigma^2)$ .

If you are not familiar, play around with this link:

<https://demonstrations.wolfram.com/TheNormalDistribution/>

# Multivariate Gaussian Distribution

- The multivariate normal distribution of a k-dimensional random vector  $\mathbf{X} = (X_1, X_2, \dots, X_k)^T$  is written as  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ .
- The probability density function is given as follows:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

where,

- $\boldsymbol{\mu} = E[\mathbf{X}] = (E[X_1], E[X_2], \dots, E[X_k])^T$
- $\Sigma_{i,j} = E[(X_i - \mu_i)(X_j - \mu_j)] = \text{Cov}[X_i, X_j]$

Fun fact: correctly constructed covariance matrices are always symmetric and positive semi-definite. And thus invertible.

But let's look at the covariance matrix in detail.

# Covariance Matrix I

Let us look at the covariance matrix in detail:  $\Sigma_{i,j} = E[(X_i - \mu_i)(X_j - \mu_j)] = Cov[X_i, X_j]$

$$\Sigma_{k \times k} = \begin{bmatrix} \sigma_1^2 & cov(x_1, x_2)^2 & \dots & cov(x_1, x_{k-1}) & cov(x_1, x_k) \\ cov(x_2, x_1)^2 & \sigma_2^2 & \dots & cov(x_2, x_{k-1}) & cov(x_2, x_k) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ cov(x_{k-1}, x_1)^2 & cov(x_{k-1}, x_2)^2 & \dots & \sigma_{k-1}^2 & cov(x_{k-1}, x_k)^2 \\ cov(x_k, x_1)^2 & cov(x_k, x_2)^2 & \dots & cov(x_k, x_{k-1})^2 & \sigma_k^2 \end{bmatrix}$$

- The diagonal elements of the matrix contain the variances of the variables.
- The off-diagonal elements contain the covariance between all possible pairs of variables.

**Question:** What happens when the off-diagonal elements are 0? What does it mean?

Let us look at a bivariate Gaussian Distribution for different values of the covariance matrix.

## Covariance Matrix II

$$\boldsymbol{\mu}_{2 \times 1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma}_{2 \times 2} = \begin{bmatrix} 1 & 0.0 \\ 0.0 & 1 \end{bmatrix}$$

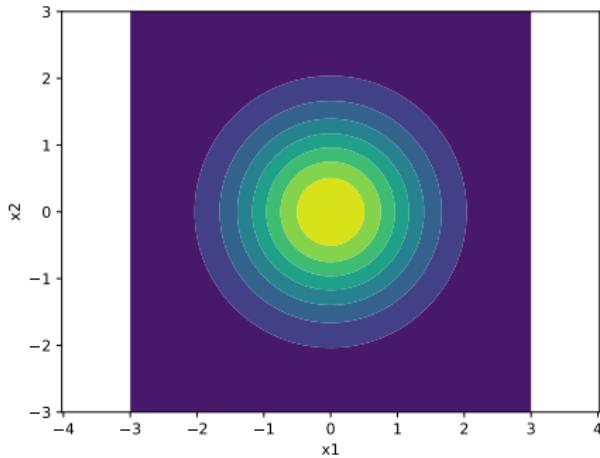


Figure: Distribution of Points

## Covariance Matrix III

$$\boldsymbol{\mu}_{2 \times 1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma}_{2 \times 2} = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$$

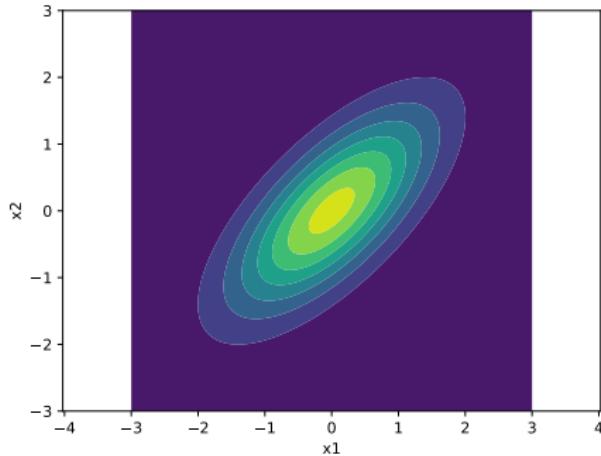


Figure: Distribution of Points

## Covariance Matrix IV

$$\mu_{2 \times 1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma_{2 \times 2} = \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}$$

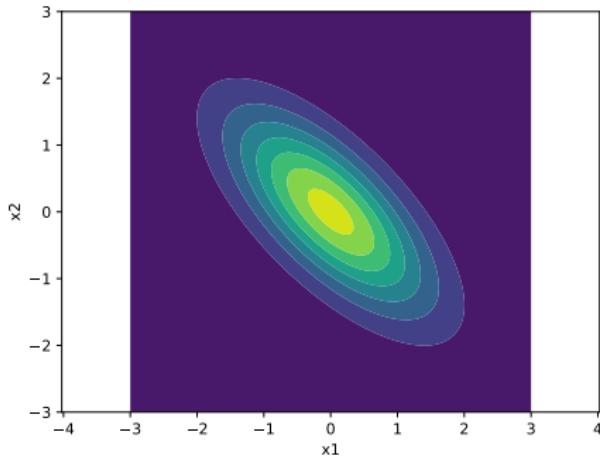


Figure: Distribution of Points

# Marginalization

- Given a multivariate distribution, can we compute the pdf of a single variable? - Yes  
 $f(X_1) = \int f(X_1, X_2) dX_2$
- Every random variables  $X_i \in \mathbf{X}$  has the following distribution:  $X_i \sim \mathcal{N}(\mu_i, \Sigma_{i,i})$

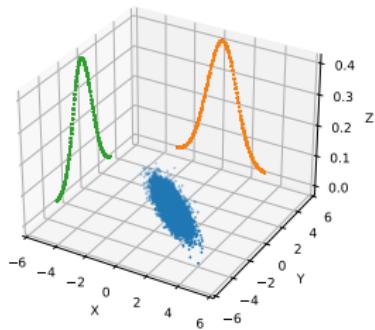


Figure: Covariance Matrix

$$\Sigma_{2 \times 2} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$

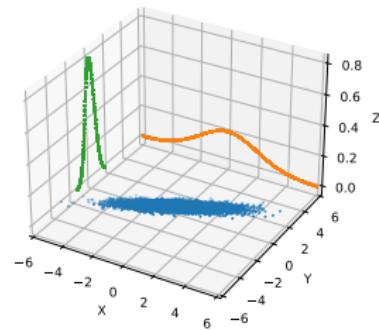


Figure: Covariance Matrix

$$\Sigma_{2 \times 2} = \begin{bmatrix} 2 & 0.9 \\ 0.9 & 0.5 \end{bmatrix}$$

# Conditioning

If some random variables from the set were fixed, what is the PDF?  
Or formally,  $f(X_i|X_j = x) = ?$

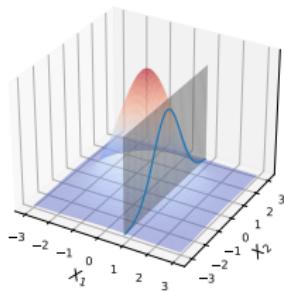


Figure: Covariance Matrix

$$\Sigma_{2 \times 2} = \begin{bmatrix} 1 & 0.0 \\ 0.0 & 1 \end{bmatrix}$$

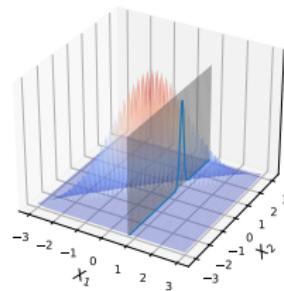


Figure: Covariance Matrix

$$\Sigma_{2 \times 2} = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$$

- In fact, it is a Gaussian. Or formally,  $X_i|(X_j = x) \sim \mathcal{N}(\mu_*, \Sigma_*)$ .
- In a bivariate case,  $f(X_2|X_1 = x_1) \propto \exp(-\frac{1}{2}(x_2 - \mu_*)\Sigma_*^{-1}(x_2 - \mu_*))$ .
- We will see what  $\mu_*$  and  $\Sigma_*$  looks like in a few slides.

# Conditioning

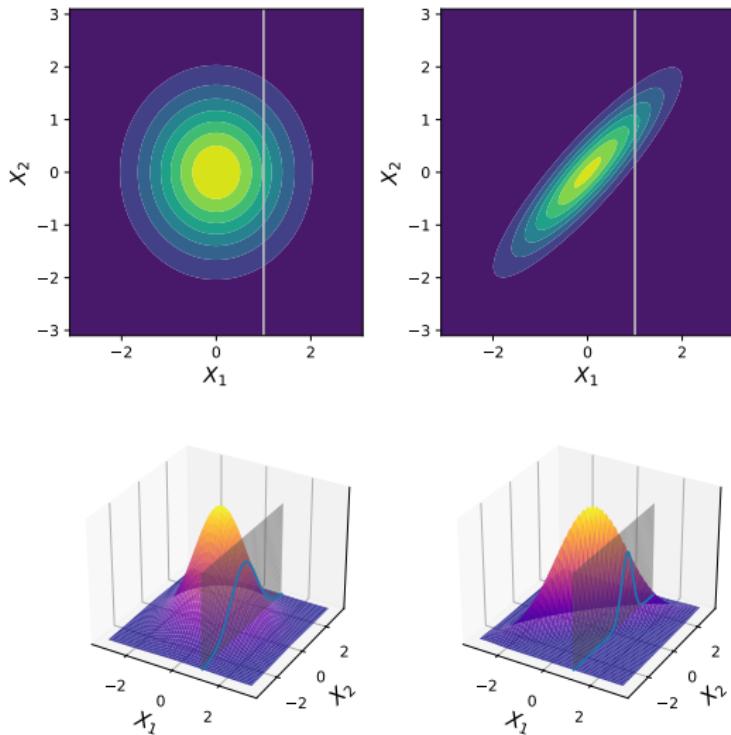
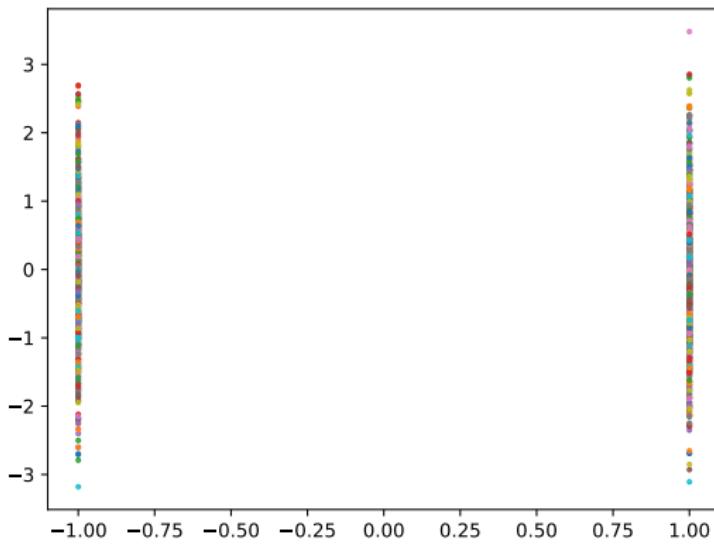


Figure: Detailed View of Conditioning

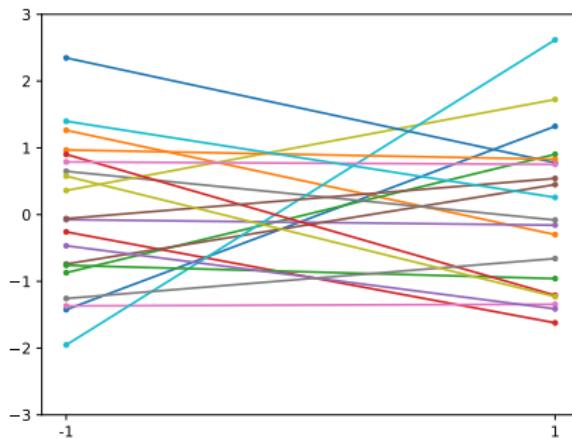
# Intuition

- Let us sample two random variables  $X_1 \sim \mathcal{N}(0, 1)$  and  $X_2 \sim \mathcal{N}(0, 1)$ .
- Next, we can plot multiple independent Gaussian in the coordinates. For example, put vector  $X_1$  at  $x = -1$  and another vector  $X_2$  at  $x = 1$ .



# Intuition

- Let's connect points of  $X_1$  and  $X_2$  by lines. For now, we only generate 20 random points, and then join them up as 10 lines. Keep in mind, that these randomly generated 10 points are Gaussian.



Going back to think about regression. These lines look like functions for each pair of points. On the other hand, the plot also looks like we are sampling the region with 20 linear functions even though there are only two points on each line. In the sampling perspective, the domain is our region of interest, i.e. the specific region we do our regression.

# Intuition

- Let's connect points of  $X_1$ ,  $X_2$ , and  $X_3$  by lines. For now, we only generate 10 random points, and then join them up as 10 lines. Keep in mind, that these randomly generated 10 points are Gaussian.

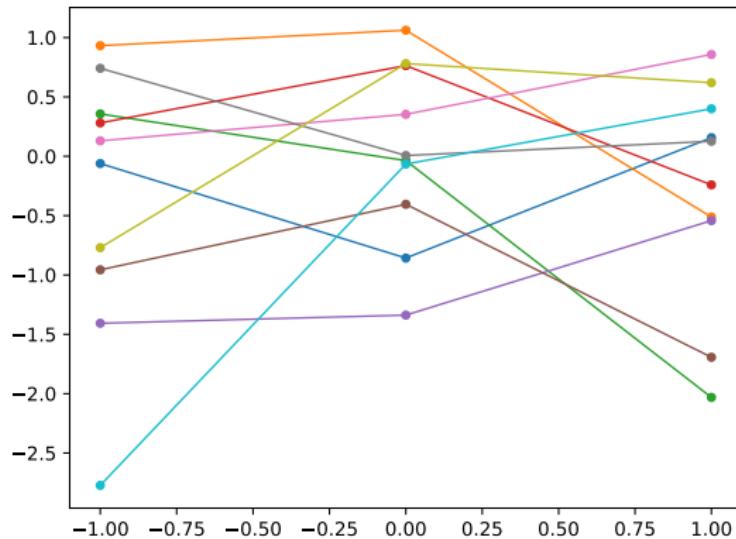


Figure:  $X_1$  will be at  $-1$ ,  $X_2$  will be at  $0$ , and  $X_3$  will be at  $1$

- Let's connect points of  $X_1, X_2, X_3, X_4$ , and  $X_5$  by lines.

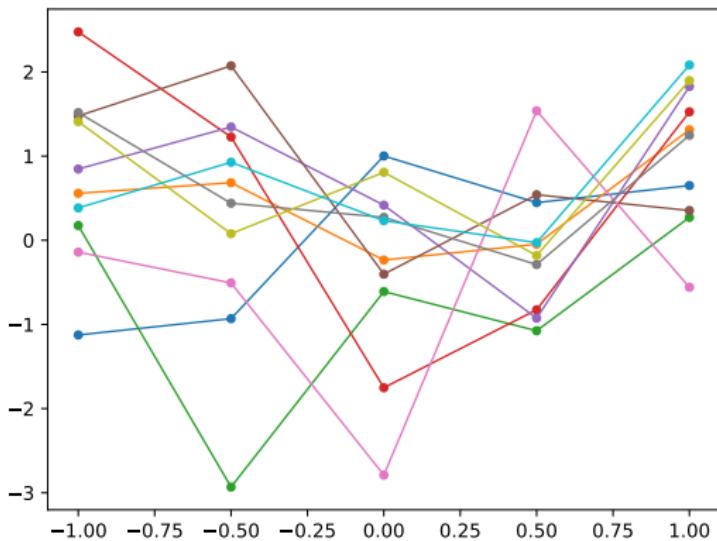


Figure:  $X_1$  will be at  $-1$ ,  $X_2$  will be at  $-0.5$ ,  $X_3$  will be at  $0$ ,  $X_4$  will be at  $0.5$ , and  $X_5$  will be at  $1$

# Intuition

This sampling looks even more clear if we generate more independent Gaussian and connecting points in order by lines.

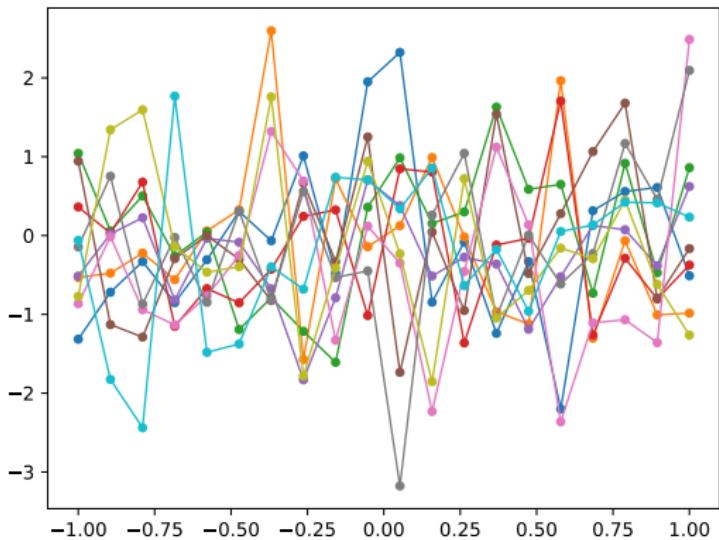


Figure:  $X_1$  will be at  $-1$ ,  $X_2$  will be at  $-0.89$ ,  $X_3$  will be at  $-0.78$ , . . . ,  $X_{19}$  will be at  $0.89$ , and  $X_{20}$  will be at  $1$

# What are the problems?

**Question:** We don't have a way of correlating these independent distributions. How do we solve this?

Use Multivariate Gaussian Distributions.

**Question:** How do we consider the points that we already have as prior?

Use Conditioning

Let's solve these two problems first!

## Revisiting Conditioning - Mathematically

So what exactly is the mean and variance when there is a conditional multivariate distribution?

$$\mathbf{X} = (X_1, X_2) \sim \mathcal{N} \left( \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \right)$$

$$X_1 | X_2 = \frac{X_1, X_2}{X_2}$$

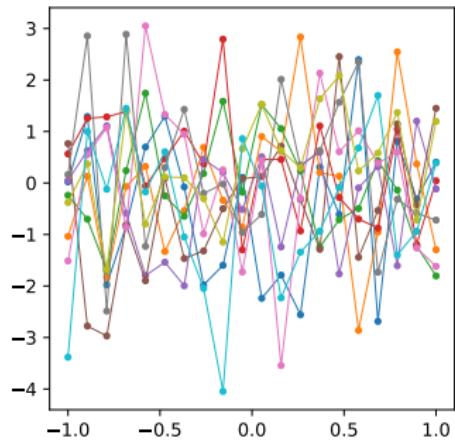
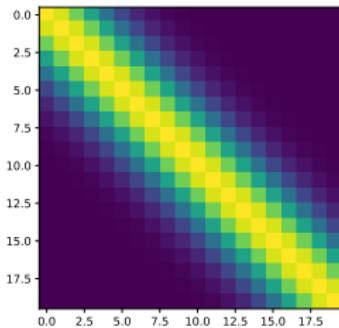
$$X_1 | (X_2 = \mathbf{x}) \sim \mathcal{N}(\mathbf{a} + \mathbf{B}\mathbf{C}^{-1}(\mathbf{x} - \mathbf{b}), \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T)$$

Do you notice the predictive mean is linear in data?

Do you notice that the predictive uncertainty is prior uncertainty minus the reduction in uncertainty?

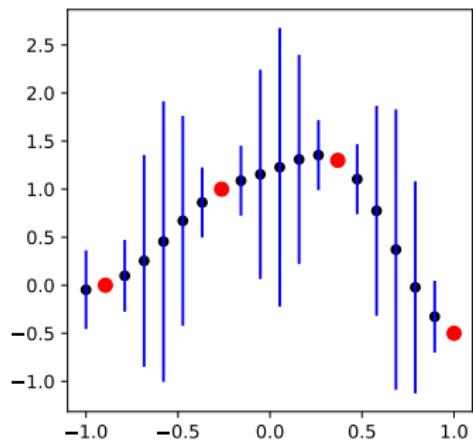
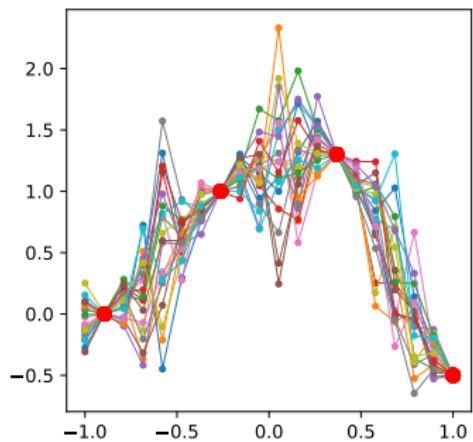
# Visualizing Multivariate Distributions with Conditioning

We have  $\mathbf{X}$  which is drawn from a Multivariate Distribution with 20 components. In this case,  $X_1$  will be at  $-1$ ,  $X_2$  will be at  $0.89$ ,  $X_3$  will be at  $-0.78$ ,  $\dots$ ,  $X_{19}$  will be at  $0.89$ , and  $X_{20}$  will be at  $1$ .



# Visualizing Multivariate Distributions with Conditioning

Let us fix  $X_1(x = -0.89)$ ,  $X_7(x = -0.26)$ ,  $X_{13}(x = 0.37)$ ,  $X_{19}(x = 1.0)$ .



## What are the problems?

**Question:** But, how do I make my regression smooth to predict at any value on the real line?

**Question:** Also, shouldn't similar points give similar results? How do I address that?

Let's see this equation again:

$$\mathbf{x} = (X_1, X_2) \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right)$$

$$X_1 | (X_2 = \mathbf{x}) \sim \mathcal{N}(\mathbf{a} + BC^{-1}(\mathbf{x} - \mathbf{b}), A - BC^{-1}B^T)$$

**Question:** Do you see that it has something to do with the covariance matrix?

# Covariance Function - Our Final Ingredient

- A kernel (also called a **covariance function**, kernel function, or covariance kernel), is a positive-definite function of two inputs.
- The covariance function  $\kappa(x_i, x_j)$  models the joint variability of the Gaussian process random variables. It returns the modeled covariance between each pair of points.
- Let's look at a covariance function called the squared exponential kernel.

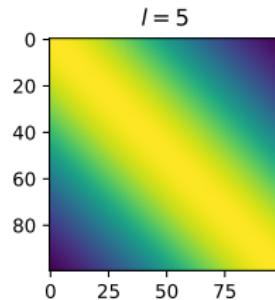
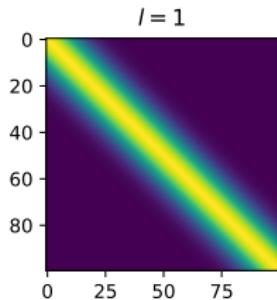
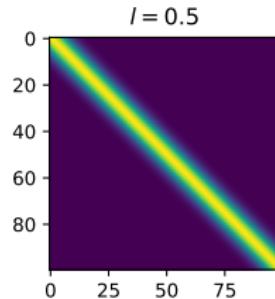
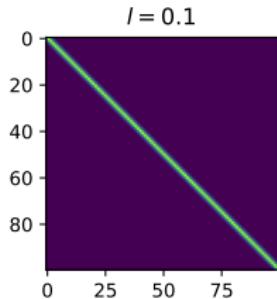
$$\kappa(x_i, x_j) = \sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{2l^2}\right)$$

We have two parameters here

- The scale parameter ( $\sigma$ ): The output variance  $\sigma^2$  determines the average distance of your function away from its mean. Every kernel has this parameter out in front; it's just a scale factor.
- The length parameter ( $l$ ): This affects the covariance matrix. Let's see what happens when we vary this.

# The Squared Exponential Kernel

$$\kappa(x_i, x_j) = \sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{2l^2}\right)$$



# Gaussian Processes

Generalization of the multivariate Gaussian Distribution to infinitely many variables.

## Definition

a Gaussian Process is a collection of random variables, any finite number of which have (consistent) Gaussian Distributions.

A Gaussian distribution is fully specified by a mean vector  $\mu$  and covariance matrix  $\Sigma$ :

$$\mathbf{f} = (f_1, \dots, f_n) \sim \mathcal{N}(\mu, \Sigma) \quad \text{where } i \text{ are the indices from } 1, \dots, n$$

A Gaussian Process is fully specified by a mean function  $m(x)$  and covariance function  $\kappa(x, x')$ :

$$f(x) \sim \mathcal{GP}(m(x), \kappa(x, x^*))$$

The function is modeled by a multivariable Gaussian as

$$\mathbf{f} | \mathbf{X} \sim \mathcal{N}(\mathbf{f} | \boldsymbol{\mu}, \mathbf{K})$$

where  $\mathbf{f} = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))$ ,  $\boldsymbol{\mu} = (m(\mathbf{x}_1), \dots, m(\mathbf{x}_n))$  and  $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ .

- $m$  is the mean function and it is common to use  $m(\mathbf{x}) = 0$  as GPs are flexible enough to model the mean arbitrarily well.
- $\kappa$  is a positive definite \*kernel function\* or \*covariance function\*.
- Thus, a Gaussian process is a distribution over functions whose shape (smoothness, ...) is defined by  $\mathbf{K}$ .
- If points  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are considered to be similar by the kernel the function values at these points,  $f(\mathbf{x}_i)$  and  $f(\mathbf{x}_j)$ , can be expected to be similar too.

The joint distribution of  $\mathbf{f}$  and  $\mathbf{f}_*$  can be modeled as:

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} m(\mathbf{X}) \\ m(\mathbf{X}_*) \end{pmatrix}, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix} \right)$$

where  $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$ ,  $\mathbf{K}_* = \kappa(\mathbf{X}, \mathbf{X}_*)$  and  $\mathbf{K}_{**} = \kappa(\mathbf{X}_*, \mathbf{X}_*)$ . And  $\begin{pmatrix} m(\mathbf{X}) \\ m(\mathbf{X}_*) \end{pmatrix} = \mathbf{0}$

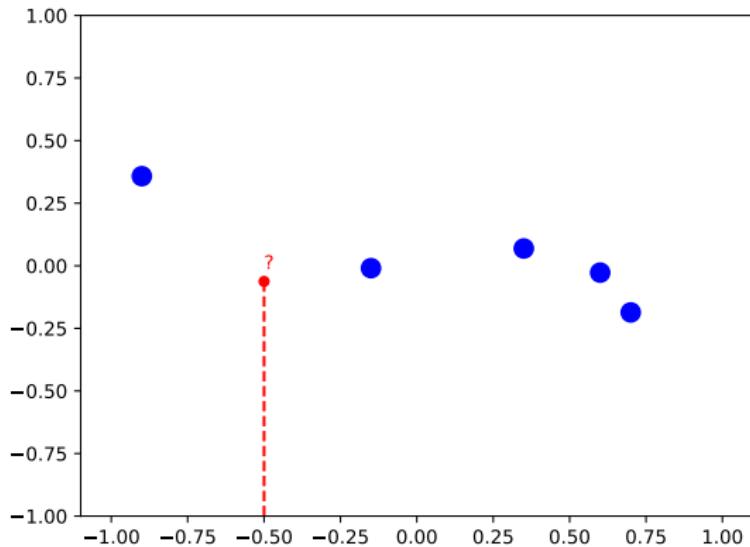
This is modeling a joint distribution  $p(\mathbf{f}, \mathbf{f}_* | \mathbf{X}, \mathbf{X}_*)$ , but we want the conditional distribution over  $\mathbf{f}_*$  only, which is  $p(\mathbf{f}_* | \mathbf{f}, \mathbf{X}, \mathbf{X}_*)$ .

- A standard result

$$\mathbf{f}_* | \mathbf{f}, \mathbf{X}, \mathbf{X}_* \sim \mathcal{N}(\mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{f}, \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{K}_*)$$

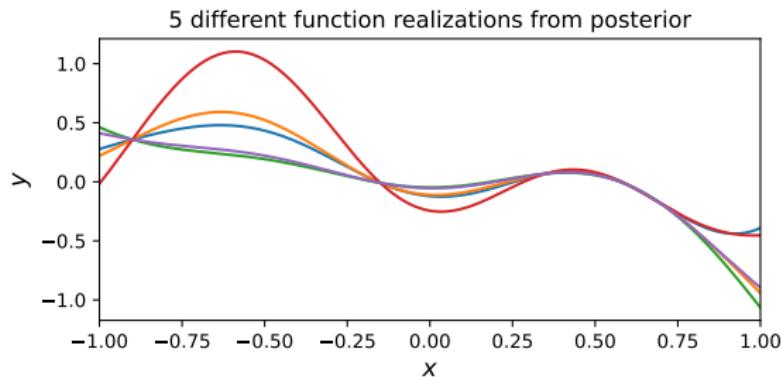
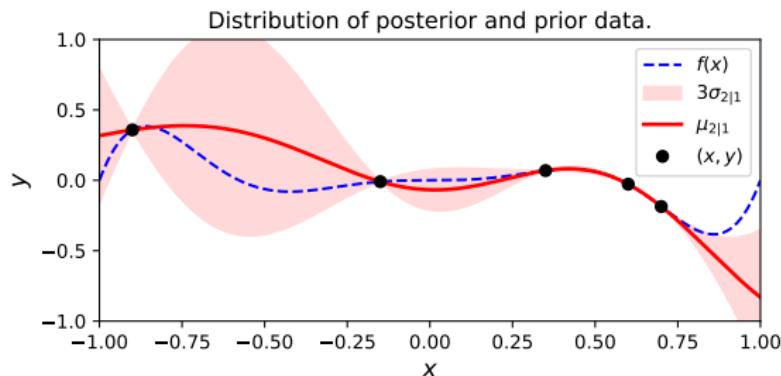
## Back to Motivation Problem

What would be the output be at the point in red?



## Back to Motivation Problem

Let us choose the RBF kernel with parameters  $\sigma = 1.5$  and  $l = 0.6$



Do we have everything ready?

**Question:** How do we choose parameters of the covariance function?

- We maximize the marginal likelihood of the Gaussian process distribution based on the observed data
- Let the (hyper-)parameters be called  $\Theta$ , inputs  $\mathbf{X}$  and their corresponding function values  $\mathbf{y} = f(\mathbf{X})$ .

$$\Theta^* = \arg \max_{\Theta} \log p(\mathbf{y} | \mathbf{X}, \Theta)$$

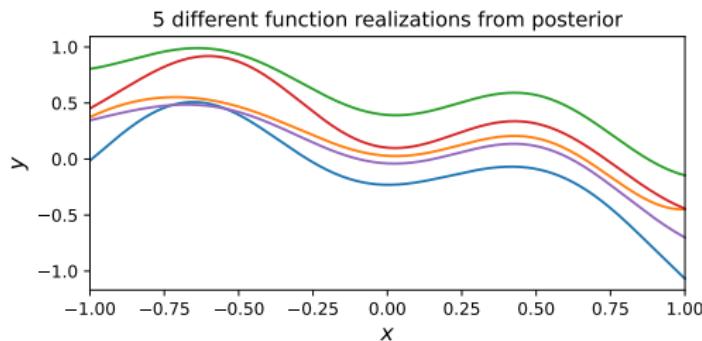
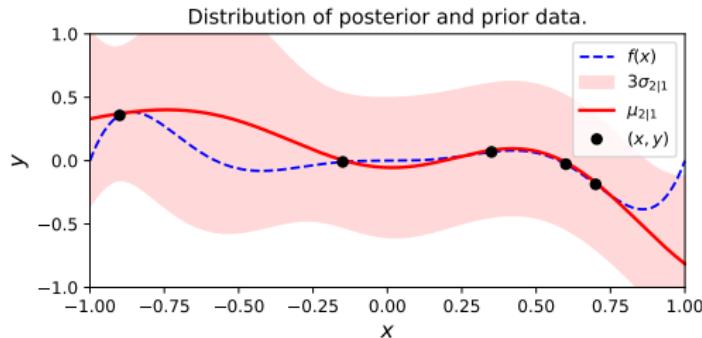
Observe that  $\mathbf{y} | \mathbf{X}, \Theta \sim \mathcal{N}(m(\mathbf{X}), K_\Theta)$ . Then,

$$\log p(\mathbf{y} | \mathbf{X}, \Theta) = -\frac{1}{2} (\mathbf{y} - \mathbf{m}_x)^T K_{\Theta_{xx}}^{-1} (\mathbf{y} - \mathbf{m}_x) - \frac{1}{2} \log(K_{\Theta_{xx}}) - \frac{n}{2} \log(2\pi)$$

## Back to Motivation Problem

We use the RBF kernel. Let's set  $\sigma = 1$ .

By maximizing the log marginal likelihood, we get  $l = 0.49$  and noise variance is 0.009



# Multiple Covariance Functions

A comprehensive overview and intuition is available at  
<https://www.cs.toronto.edu/~duvenaud/cookbook/>.  
Can you all please open the page?

**Question:** Can you code up the covariance functions and switch them up in the code we discussed?