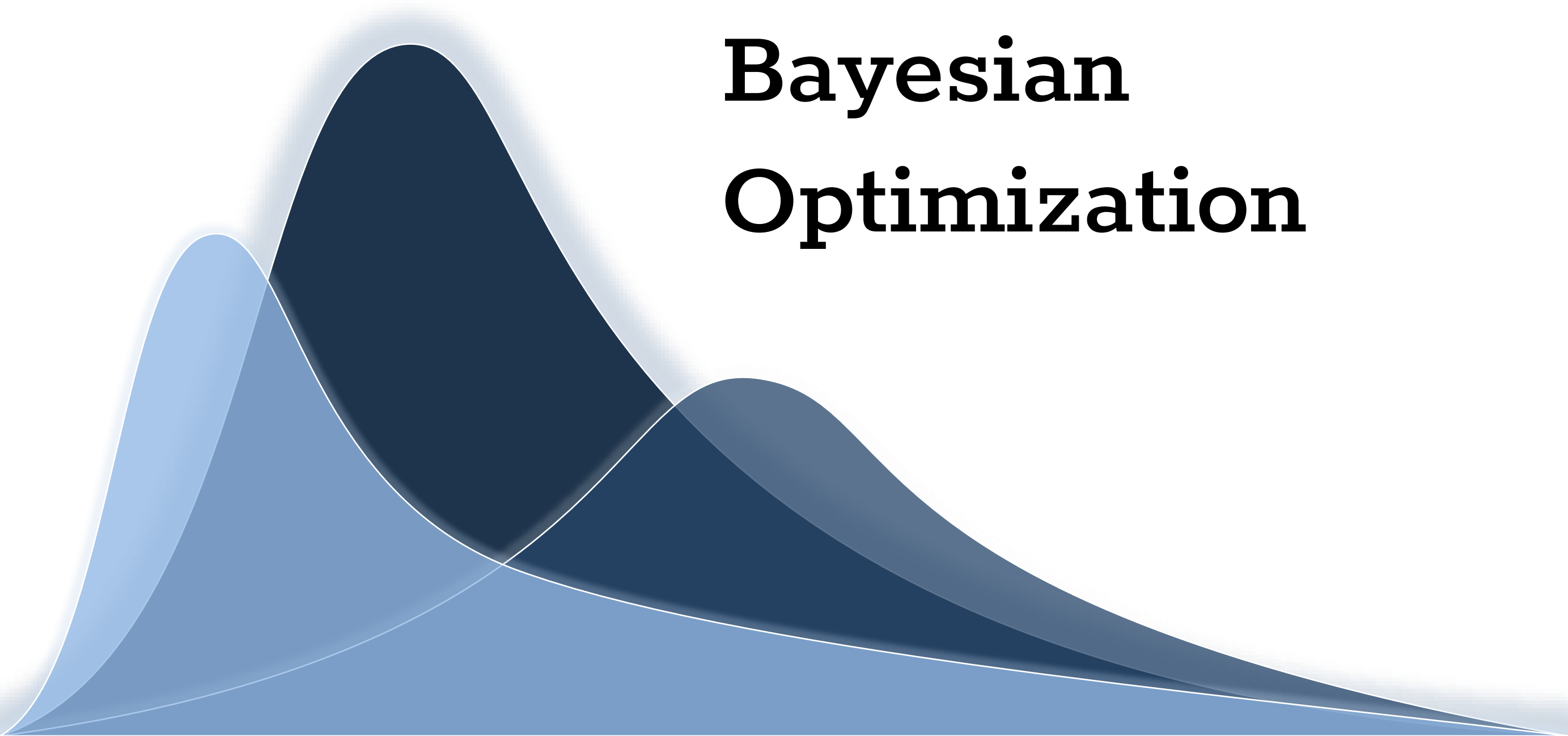
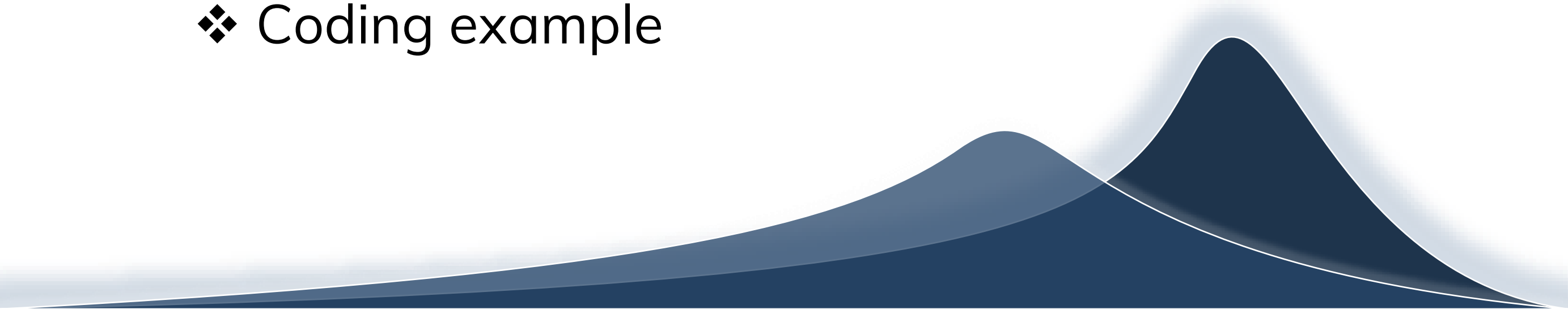


# Bayesian Optimization



# Content

---

- ❖ Concept
  - ❖ Structure
  - ❖ Process
  - ❖ Coding example
- 

# Concept



$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

*Bayes' theorem*

Bayesian inference

(Bayes' theorem) is used to update the probability for a hypothesis as more available information.

Bayesian optimization

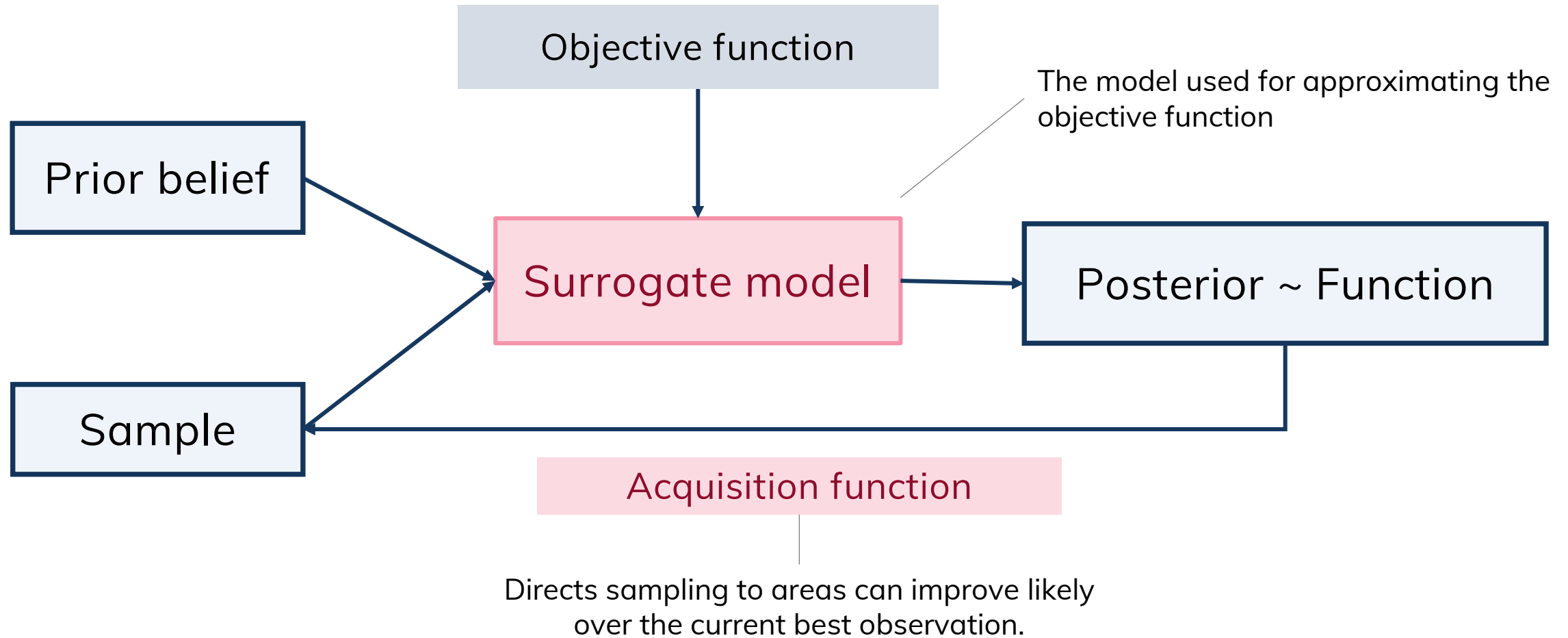
Sequential design algorithm for global optimization of black-box functions to find the global optimum in a minimum number of steps



Optimize expensive-to-evaluate functions.

Ex: Hyper-parameter tuning

# Structure



# Structure

## Surrogate model Infer a distribution over functions

Multivariate Gaussian distributions

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma)$$

Training data X, Training new data Y.

The key idea of Gaussian processes is to model the underlying distribution of X together with Y as a multivariate normal distribution

Mean vector  $\mu$  and covariance matrix  $\Sigma$

$$P_{X,Y} = \begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma) = \mathcal{N} \left( \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \right)$$

$$\underbrace{P(X|Y)}_{\text{posterior}} = \frac{P(Y|X) \times \underbrace{P(X)}_{\text{prior}}}{P(Y)}$$

## Gaussian processes

- Defines a prior over functions.
- Observed some function values
- Converted into a posterior over functions.

# Structure

## Surrogate model      Gaussian processes

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

$F(x)$  is function value at input  $X$  (training data)

$\boldsymbol{\mu}=(m(x_1),\dots,m(x_N))$  with  $m$  is the mean function and it is common to use  $m(x)=0$  as GPs are flexible enough to model the mean arbitrarily well.

$K$  is kernel function (covariance function), define distribution over function shape.

A GP prior can be converted into a GP posterior  $p(\mathbf{f}_*|\mathbf{X}_*,\mathbf{X},\mathbf{f})$

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix}\right)$$

GP posterior  $p(\mathbf{f}_*|\mathbf{X}_*,\mathbf{X},\mathbf{f})$  which can then be used to make predictions  $\mathbf{f}_*$  at new inputs  $\mathbf{X}_*$ .

$$\begin{aligned} p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{f}) &= \mathcal{N}(\mathbf{f}_*|\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \\ \boldsymbol{\mu}_* &= \mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{f} \\ \boldsymbol{\Sigma}_* &= \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{K}_* \end{aligned}$$

# Structure

## Surrogate model      Gaussian processes – Kernel function

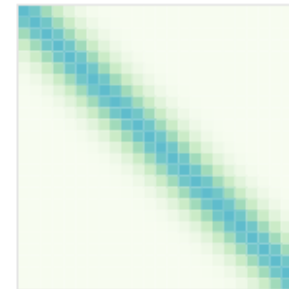
K is kernel function (covariance function), define distribution over function shape

Kernel function receives two points as input and returns a similarity measure between those points in the form of a scalar

$$k : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}, \quad \Sigma = \text{Cov}(X, X') = k(t, t')$$

RBK KERNEL

$$\sigma^2 \exp \left( -\frac{\|t-t'\|^2}{2l^2} \right)$$



Variance  $\sigma$  = 0.76

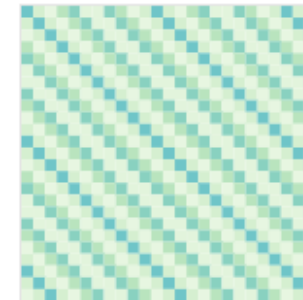


Length l = 0.91



PERIODIC

$$\sigma^2 \exp \left( -\frac{2 \sin^2(\pi|t-t'|/p)}{l^2} \right)$$



Variance  $\sigma$  = 0.73



Length l = 1.05

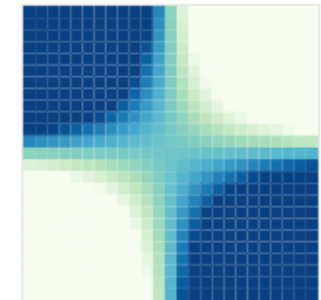


Periodicity p = 0.55



LINEAR

$$\sigma_b^2 + \sigma^2(t - c)(t' - c)$$



Variance  $\sigma$  = 0.41



Variance  $\sigma_b$  = 0.53



Offset c = -0.3



# Structure



Surrogate model      Gaussian processes

Training dataset with noisy function values  $y=f+\epsilon$  where noise  $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_y^2 \mathbf{I})$

$$\begin{aligned} p(\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{f}) &= \mathcal{N}(\mathbf{f}_* | \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \\ \boldsymbol{\mu}_* &= \mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{f} \\ \boldsymbol{\Sigma}_* &= \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{K}_* \end{aligned}$$



$$\begin{aligned} p(\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{y}) &= \mathcal{N}(\mathbf{f}_* | \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \\ \boldsymbol{\mu}_* &= \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{y} \\ \boldsymbol{\Sigma}_* &= \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{K}_* \end{aligned}$$

with

$$\mathbf{K}_y = \mathbf{K} + \sigma_y^2 \mathbf{I}.$$

$$p(y_* | \mathbf{X}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(y_* | \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_* + \sigma_y^2 \mathbf{I})$$



# Structure



## Acquisition function

Exploitation: sampling where the surrogate model predicts a high objective.

Exploration: sampling at locations where the prediction uncertainty is high.



Maximize the acquisition function to determine the next sampling point.

The objective function  $f$  will be sampled at

$$\mathbf{x}_t = \operatorname{argmax}_{\mathbf{x}} u(\mathbf{x} | \mathcal{D}_{1:t-1})$$

Where  $u$  is the acquisition function

$\mathcal{D}_{1:t-1} = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{t-1}, y_{t-1})$  are the  $t-1$  samples drawn from  $f$  so far

Popular acquisition functions: maximum probability of improvement (MPI), expected improvement (EI) and upper confidence bound (UCB)

# Structure

Acquisition function Expected improvement (EI)

$$\text{EI}(\mathbf{x}) = \mathbb{E} \max(f(\mathbf{x}) - f(\mathbf{x}^+), 0)$$

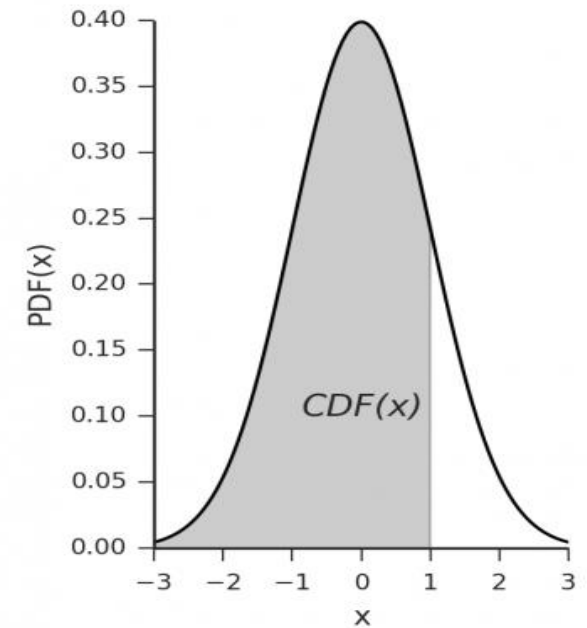
where  $f(\mathbf{x}^+)$  is the value of the best sample so far and  $\mathbf{x}^+ = \operatorname{argmax}_{\mathbf{x}_i \in \mathbf{x}_{1:t}} f(\mathbf{x}_i)$

EI can be evaluated analytically under the GP model.

$$\text{EI}(\mathbf{x}) = \begin{cases} \overbrace{(\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\Phi(Z)}^{\text{Exploitation}} + \overbrace{\sigma(\mathbf{x})\phi(Z)}^{\text{Exploration}} & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases}$$

where

$$Z = \begin{cases} \frac{\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi}{\sigma(\mathbf{x})} & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases}$$



$\mu(\mathbf{x})$  and  $\sigma(\mathbf{x})$  are the mean and the standard deviation of the GP posterior predictive at  $\mathbf{x}$ .  $\Phi$  and  $\phi$  are the CDF and PDF of the standard normal distribution.

$\xi$  determines the amount of exploration during optimization and higher  $\xi$  values lead to more exploration

# Process



The Bayesian optimization procedure:

For  $t=1,2,\dots$  repeat:

- Find the next sampling point  $\mathbf{x}(t)$  by optimizing the acquisition function over the GP:

$$\mathbf{x}_t = \operatorname{argmax}_{\mathbf{x}} u(\mathbf{x}|\mathcal{D}_{1:t-1})$$

- Obtain a possibly noisy sample  $y_t = f(\mathbf{x}_t) + \epsilon_t$  from the objective function  $f$ .
- Add the sample to previous samples  $\mathcal{D}_{1:t} = \mathcal{D}_{1:t-1}, (\mathbf{x}_t, y_t)$  and update the GP.

# Coding



Topic: Predicting debt group of ending month payment of the customer by some customer attribute.

(vintage, loan amt with ins, interest rate, tenure, product, age, gender, marital status, social status, education, pos province, province city current, fb members, personal income → debt group)

## **Encode**

Pos province → Price index by province

Province city current → Income average by province

Link coding:

<https://drive.google.com/file/d/1XqzH119DKlu6rBclRwBiNPOFhHwVZrui/view?usp=sharing>