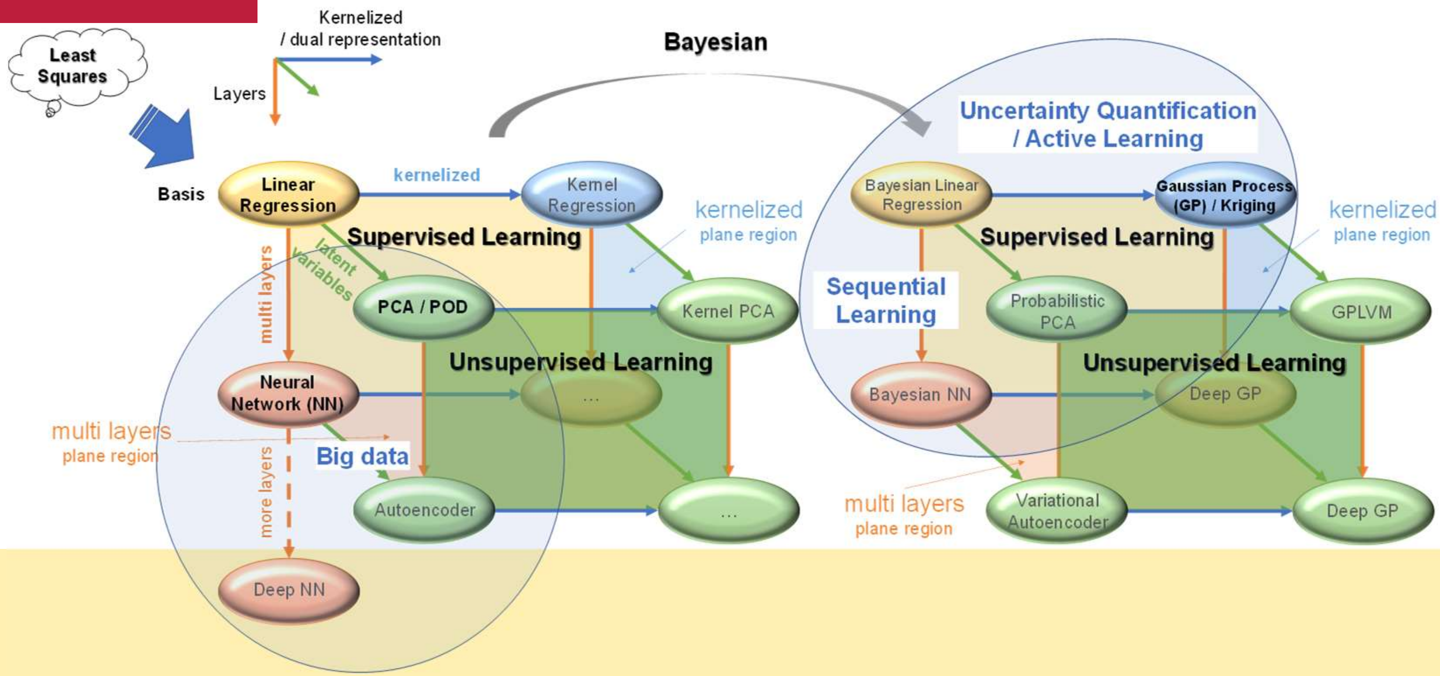


Scientific Machine Learning

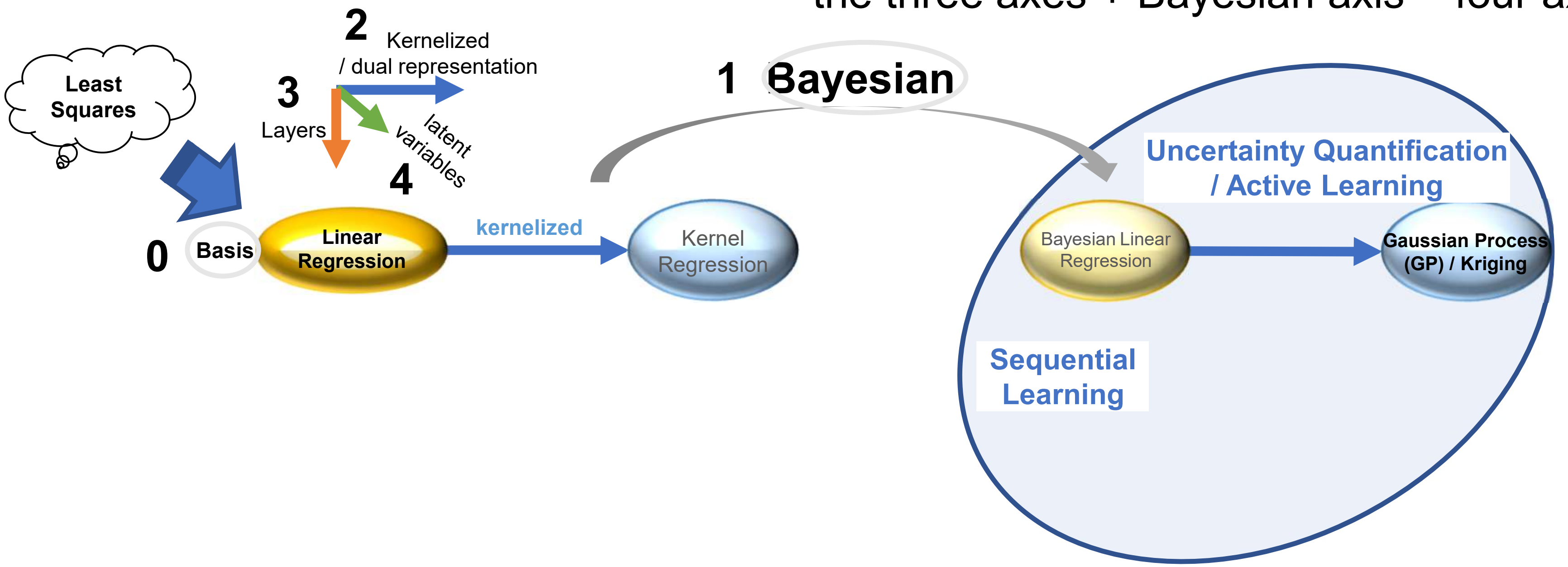
Lecture 9: Gaussian Process (2/2)

Dr. Daigo Maruyama

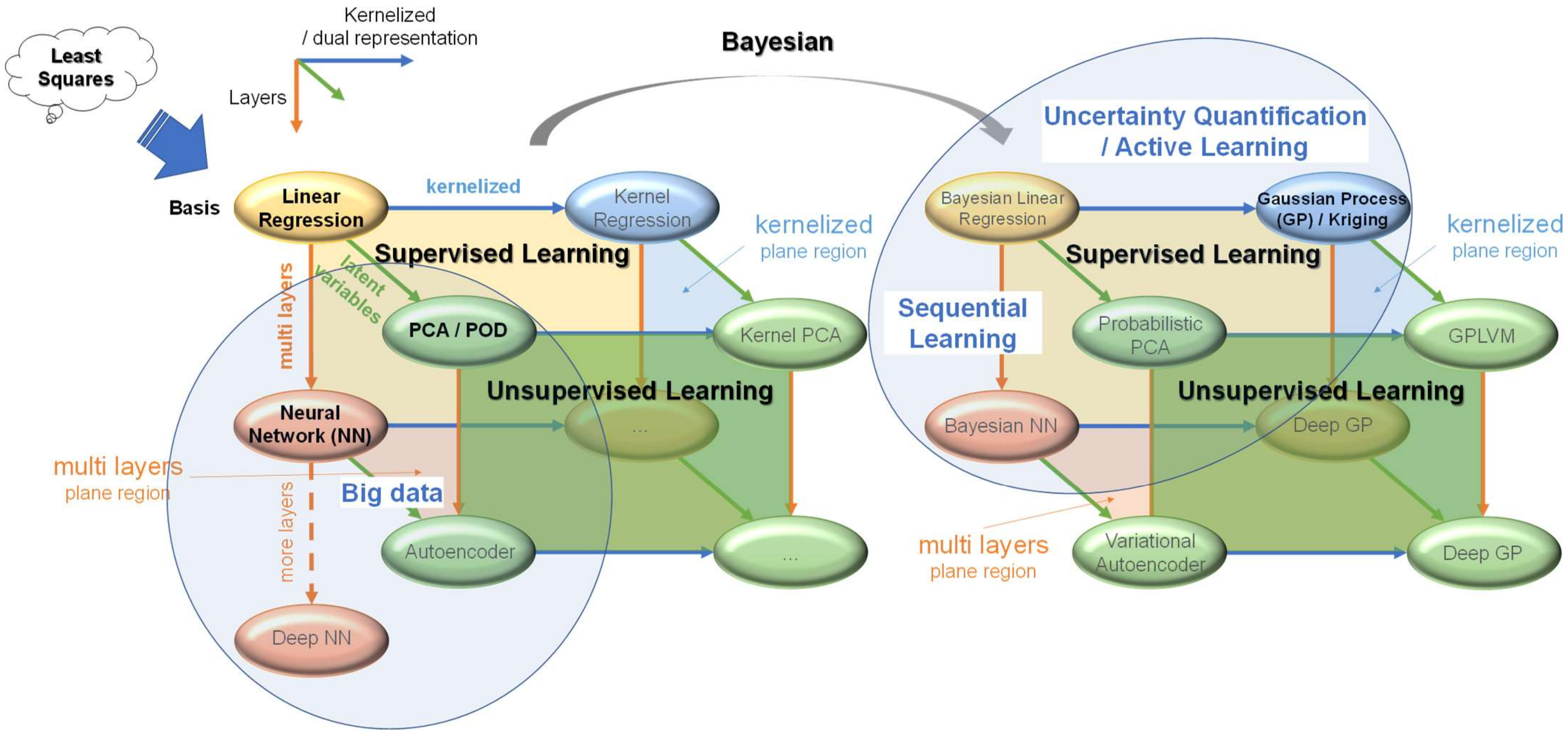
Prof. Dr. Ali Elham

Key Components

the three axes + Bayesian axis = four axes



Structures of the Techniques/Methods



Lecture content

- Further topics of Gaussian processes and introduction of Kriging
- Applications of Gaussian processes
- Advanced Methods/Topics

The lecture of this time partially follows the Section 6.4 of the book:
Christopher M. Bishop "Pattern Recognition And Machine Learning" Springer-Verlag (2006)
The name of this book is shown as “PRML” when it is referred in the slides.

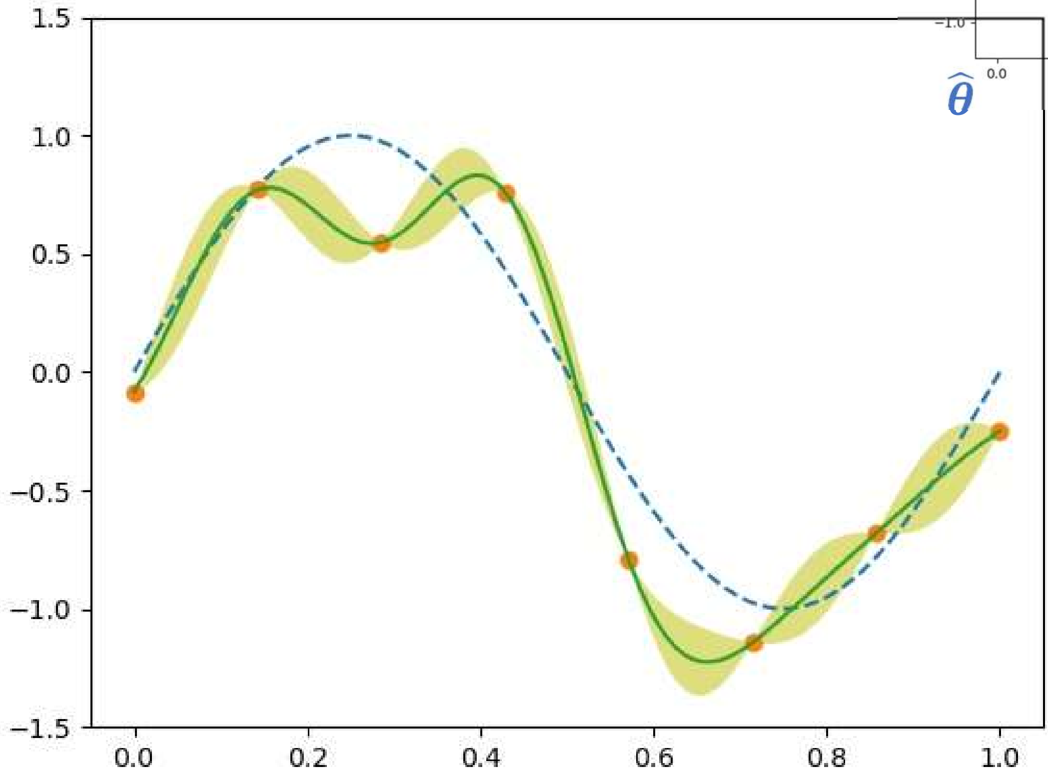
The lecture slides contains many original contents in the context apart from the above sections in the book.

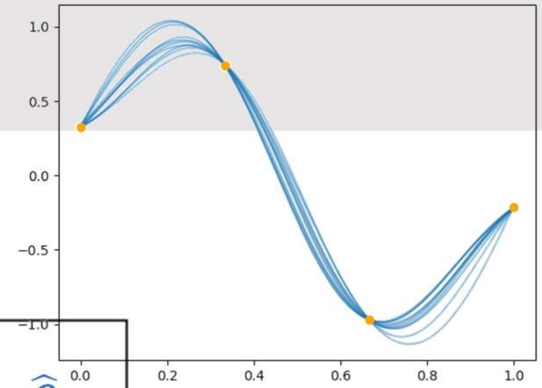
Lecture content

- Further topics of Gaussian processes and introduction of Kriging

Gaussian Processes

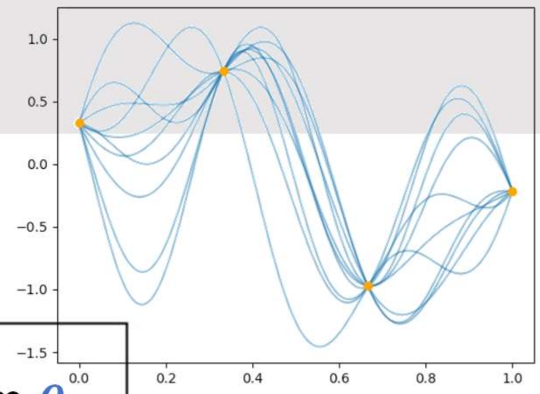
Example ($\hat{\boldsymbol{\theta}}$ by MLE)



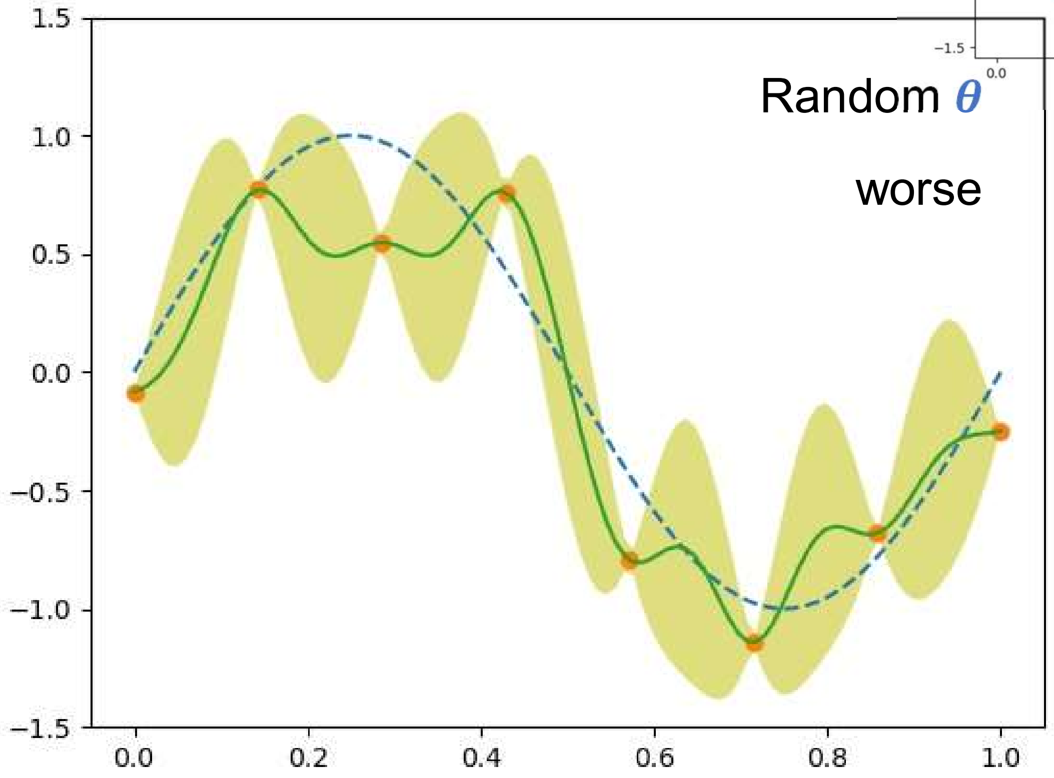


$\hat{\boldsymbol{\theta}}$

Gaussian Processes

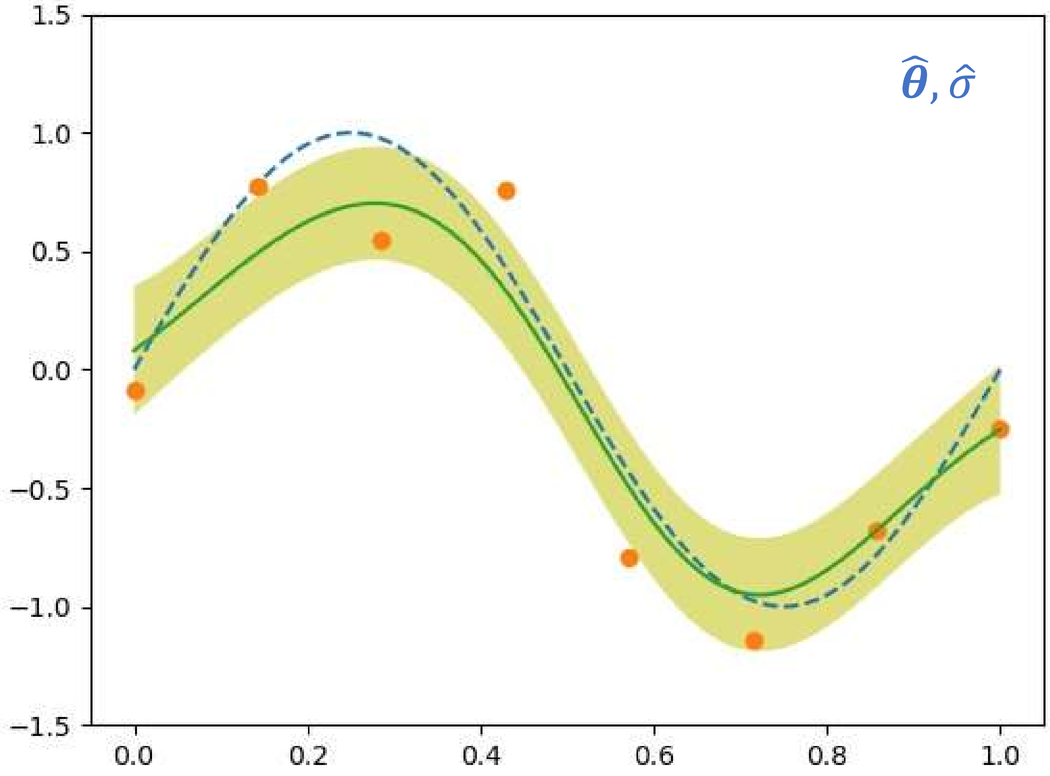


Examples (random $\boldsymbol{\theta}$ without learning from the data)



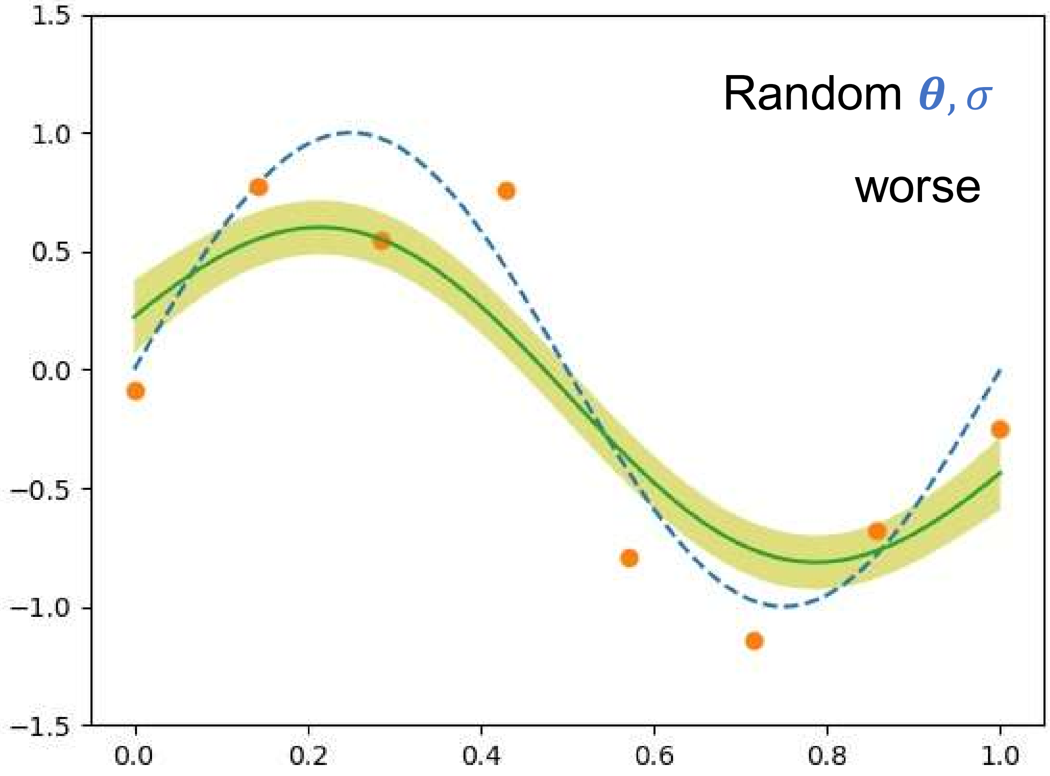
Gaussian Processes

Examples ($\hat{\boldsymbol{\theta}}, \hat{\sigma}$ by MLE)



Gaussian Processes

Examples (random $\boldsymbol{\theta}, \sigma$ without learning from the data)

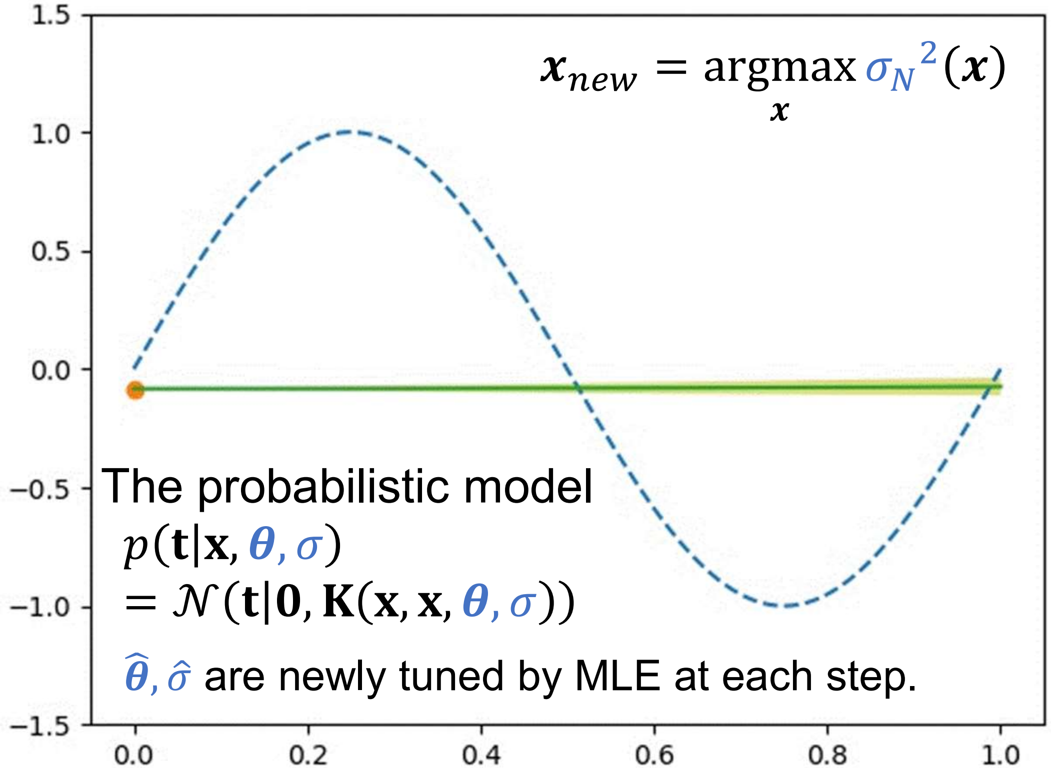


Gaussian Process + adaptive sampling

Examples

Adding a new point at the location $\boldsymbol{x}$ **where $\sigma_N{}^2(\boldsymbol{x})$ is max**

Exactly the same concept as slides 44-60 in Lecture 6

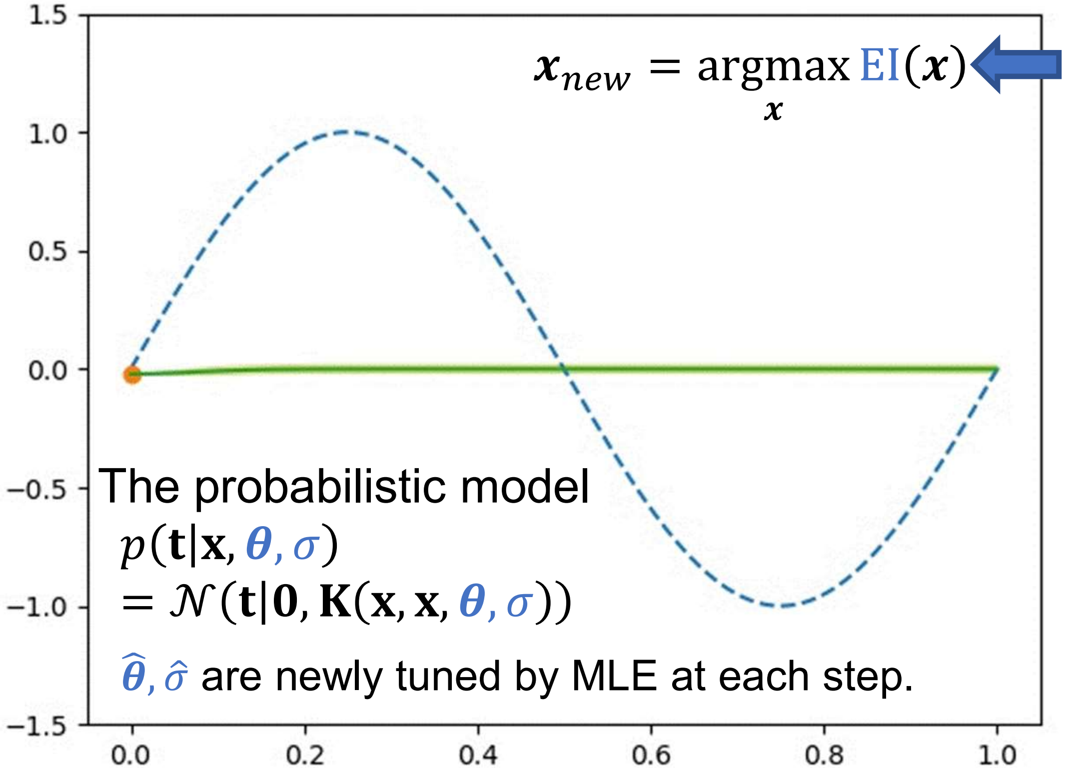


$$\sigma_N{}^2(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}, \widehat{\boldsymbol{\theta}}, \hat{\sigma}) - \boldsymbol{k}(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}, \hat{\sigma})^{\mathrm{T}} \mathbf{K}(\widehat{\boldsymbol{\theta}}, \hat{\sigma})^{-1} \boldsymbol{k}(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}, \hat{\sigma})$$

Gaussian Process + adaptive sampling (for Optimization)

Examples

Adding a new point at the location $\boldsymbol{x}$ **where** a function
$\mathrm{EI}(\boldsymbol{x}) = f(\mu(\boldsymbol{x}), {\sigma_N}^2(\boldsymbol{x}), current\ minimum\ sample\ point)$ **is max**



a suitable criterion
to find optimum

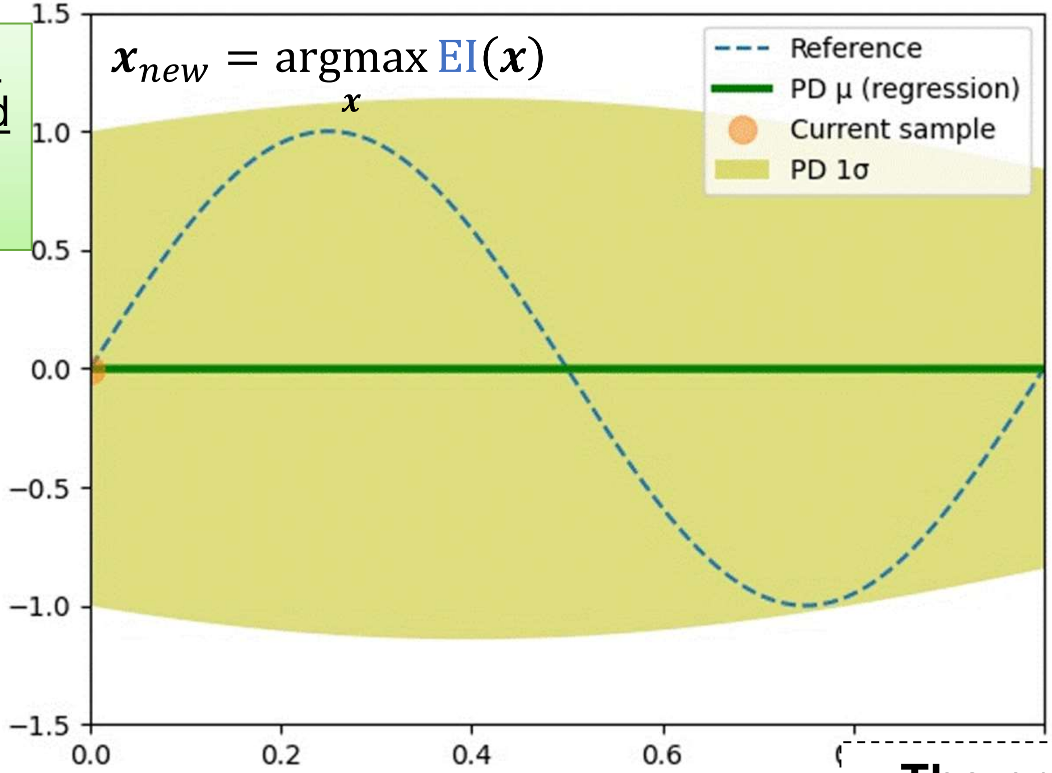
$\mathrm{EI}(\boldsymbol{x}) = f(\mu(\boldsymbol{x}), {\sigma_N}^2(\boldsymbol{x}), current\ minimum\ sample\ point)$ $\quad (\mu(\boldsymbol{x}) = \boldsymbol{k}(\mathbf{x})^{\mathrm{T}} \mathbf{K}(\widehat{\boldsymbol{\theta}}, \hat{\sigma})^{-1} \mathbf{T})$

Bayesian Linear Regression + adaptive sampling (for Opt.)

Examples

Adding a new point at the location $\boldsymbol{x}$ **where** a function
$\mathrm{EI}(\boldsymbol{x}) = f(\mu(\boldsymbol{x}), {\sigma_N}^2(\boldsymbol{x}), current\ minimum\ sample\ point)$ **is max**

Of course, <u>the same concept can be used</u> in the Bayesian linear regression.



$\mathrm{EI}(\boldsymbol{x}) = f(\mu(\boldsymbol{x}), {\sigma_N}^2(\boldsymbol{x}), current\ minimum\ sample\ point)$

The probabilistic model

$p(t|x, \boldsymbol{w}) = \mathcal{N}\left(t \middle| \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$

$\boldsymbol{\phi}(\boldsymbol{x})$ = polynomials ($M = 9$)

$\hat{\sigma} = 1e - 3$ (fixed)

Bayesian Linear Regression

The probabilistic model

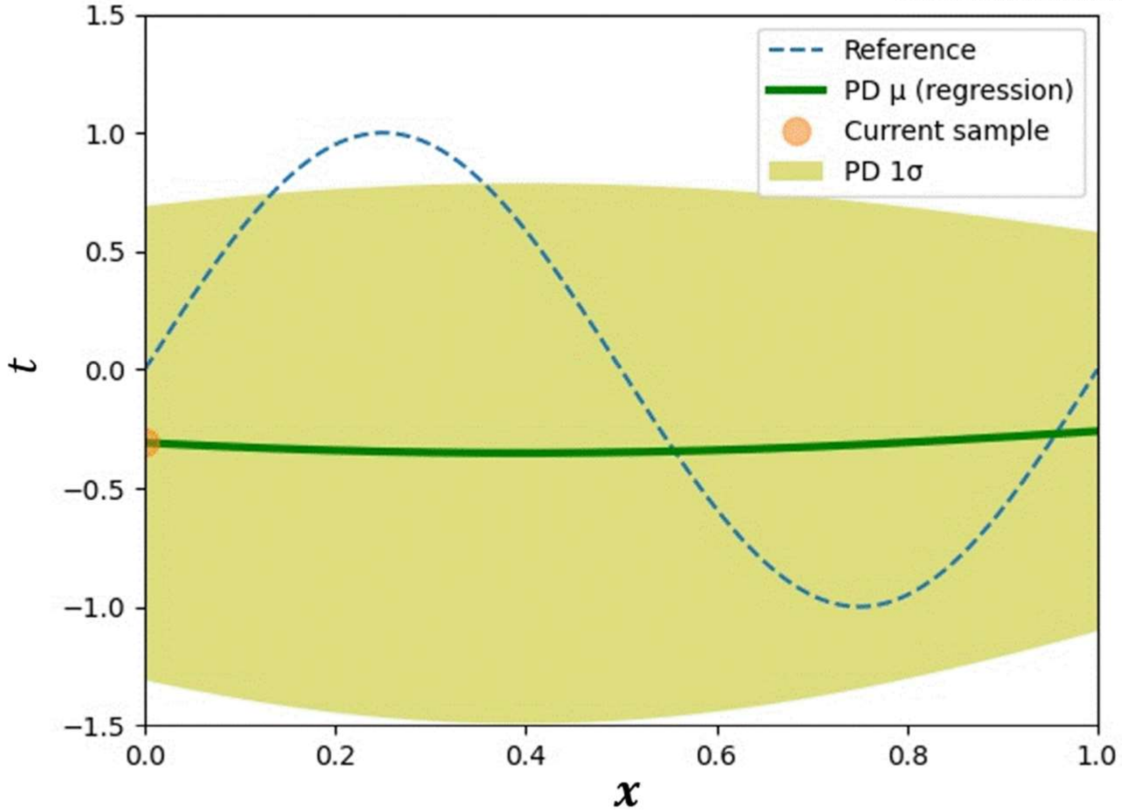
$p(t|x, \boldsymbol{w}) = \mathcal{N}\left(t \middle| \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$

$\boldsymbol{\phi}(\boldsymbol{x})$ = polynomials $(M = 9)$

$\hat{\sigma} = 0.2$ (fixed)

Examples

Adding a new point at the location $\boldsymbol{x}$ **where** ${\sigma_N}'^2(\boldsymbol{x})$ (or ${\sigma_N}^2(\boldsymbol{x})$) **is max**



starting from $N\ (sample\ size) = 1$

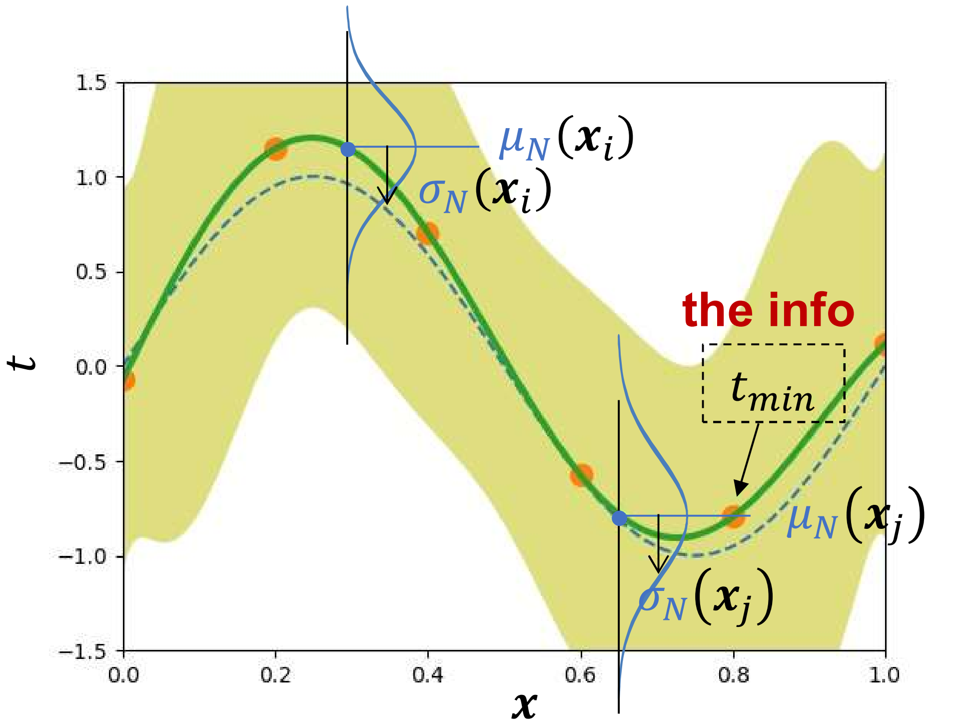
called acquisition functions in general

1. Variance (Entropy)

$$\boldsymbol{x}_{new} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, \sigma_N{}^2(\boldsymbol{x})$$

2. Expected Improvement (EI)

$$\boldsymbol{x}_{new} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, \mathrm{EI}(\boldsymbol{x})$$



The predictive distribution (a single input $\boldsymbol{x}$)

the info $\quad p(t|\boldsymbol{x}) = \mathcal{N}(t|\mu_N(\boldsymbol{x}), \sigma_N{}^2(\boldsymbol{x}))$

$$\mathrm{EI}(\boldsymbol{x}) = \left(t_{min} - \mu_N(\boldsymbol{x})\right)\mathrm{cdf}\left(p_{std}(\boldsymbol{x})\right) + \sigma_N(\boldsymbol{x})\mathrm{pdf}\left(p_{std}(\boldsymbol{x})\right)$$

$$p_{std}(\boldsymbol{x}) = \frac{t_{min} - \mu_N(\boldsymbol{x})}{\sigma_N(\boldsymbol{x})}$$

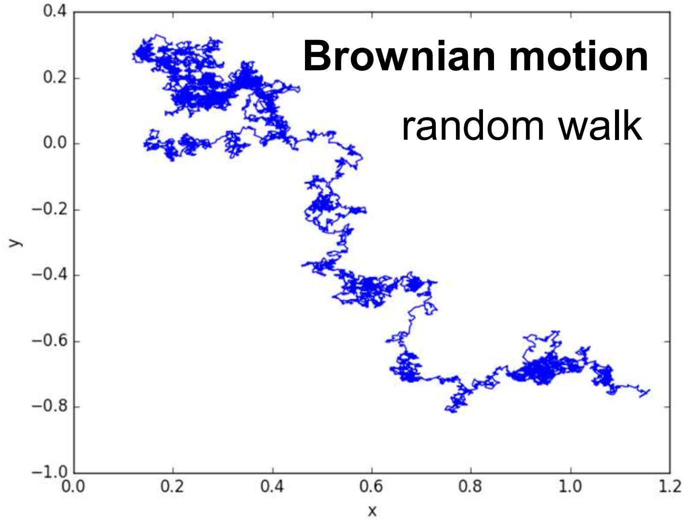
standardized

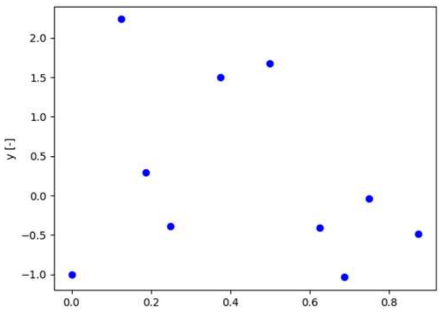
If you are interested, you can follow the meaning of the equation and can understand it as a probability where min location is expected to be updated.

History of Gaussian Process and Kriging

Gaussian Process

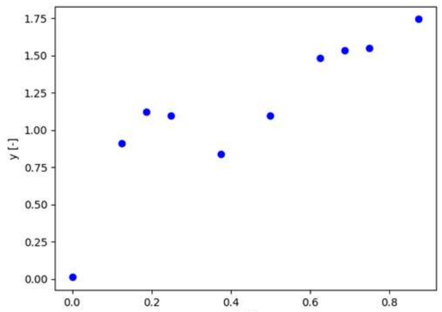
"Gaussian process" have been modeled and developed to describe random motions in physics.





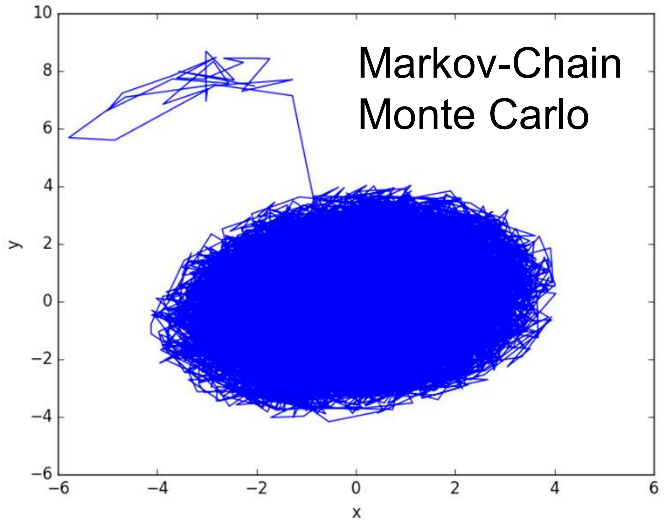
$t\ [time]$

$p(y) = \mathcal{N}(y|0, \sigma^2)$



$t\ [time]$

$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K})$



History of Gaussian Process and Kriging

Kriging

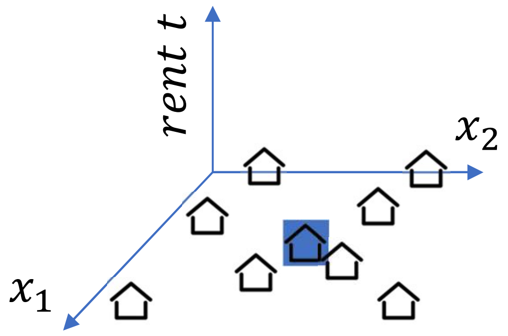
- Spatial statistics
- Geostatistics

to interpolate the value of a random field at an unobserved location from observations of its value at nearby locations

$$\tilde{\mathbf{t}}(\boldsymbol{x}) = \boldsymbol{w}(\boldsymbol{x})^{\mathrm{T}}\mathbf{T}$$

Example:
We want to predict the rent of an apartment.

using:
- The known rents of apartments around it
- Spatial data (in this case the location $\mathbf{X}$)



Spatial autocorrelation: "*closer to each other, closer the properties are*."

$$\hat{\boldsymbol{w}}(\boldsymbol{x}) = \min_{\boldsymbol{w}(\boldsymbol{x})} MSE \qquad MSE = E\left[\left(\mathbf{t}(\boldsymbol{x}) - \tilde{\mathbf{t}}(\boldsymbol{x})\right)^2\right] \qquad \Rightarrow \qquad \tilde{\mathbf{t}}(\boldsymbol{x}) = \hat{\boldsymbol{w}}(\boldsymbol{x})^{\mathrm{T}}\mathbf{T}$$

Prediction in Dual Representation

$$p(\mathbf{t}|\mathbf{x},\mathcal{D}) = \mathcal{N}\left(\mathbf{t}\middle|\underbrace{\boldsymbol{k}(\mathbf{x})^{\mathrm{T}}\mathbf{K}^{-1}}_{\chi(\boldsymbol{x})^{\mathrm{T}}}\mathbf{T}, \mathbf{K}(\mathbf{x},\mathbf{x}) - \underbrace{\boldsymbol{k}(\mathbf{x})^{\mathrm{T}}\mathbf{K}^{-1}}_{\chi(\boldsymbol{x})^{\mathrm{T}}}\boldsymbol{k}(\mathbf{x})\right)$$

$$\mathcal{D} = (\mathbf{X}, \mathbf{T})$$

Gaussian Process $\tilde{\mathbf{t}}(\mathbf{x}) = \underbrace{\boldsymbol{k}(\mathbf{x})^{\mathrm{T}}\mathbf{K}^{-1}}_{\chi(\boldsymbol{x})^{\mathrm{T}}}\mathbf{T}$

Kriging $\tilde{\mathbf{t}}(\mathbf{x}) = \chi(\boldsymbol{x})^{\mathrm{T}}\mathbf{T}$

(Bayesian) linear regression

$\tilde{\mathbf{t}}(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x})$

$\tilde{\mathbf{t}}(\mathbf{x}) = \boldsymbol{\Phi}(\mathbf{x})\boldsymbol{w}$

Prediction using
the training data $\mathcal{D} = (\mathbf{X}, \mathbf{T})$

Prediction using
the learned parameters $\boldsymbol{w}$

The training data $\mathcal{D} = (\mathbf{X}, \mathbf{T})$ can be discarded in the prediction process.

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

$$\boldsymbol{\mu} = \chi(\boldsymbol{x})^{\mathrm{T}}\mathbf{T}$$

$$\boldsymbol{\Sigma} = \mathbf{K}(\mathbf{x},\mathbf{x}) - \chi(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{k}(\mathbf{x})$$

Note: The covariance can be slightly different in Kriging.

Lecture content

- Applications of Gaussian processes

Optimization in Engineering Problems

Optimization formulation:

$$\min_{\boldsymbol{x}} y(\boldsymbol{x})$$

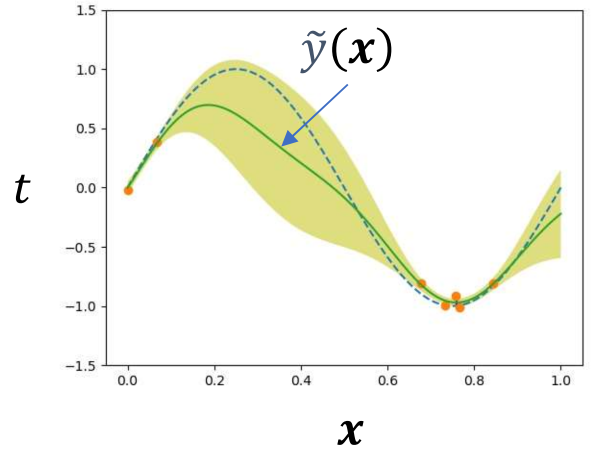
$\hat{\boldsymbol{x}}$ and $\hat{y} \equiv y(\hat{\boldsymbol{x}})$

$$\min_{\boldsymbol{x}} \tilde{y}(\boldsymbol{x})$$

$\tilde{y}$ is **approximation**.

$\boldsymbol{x}$: input

t: output



Conditions:
$y(\boldsymbol{x})$ for arbitrary $\boldsymbol{x}$ is <u>expensive</u> to obtain (e.g. experimental data).

Find $\hat{\boldsymbol{x}}$ accurately but using as small numbers of sample size N as possible

Surrogate-Based Optimization

Design of Experiments (DOE)

Evaluation of the expensive function

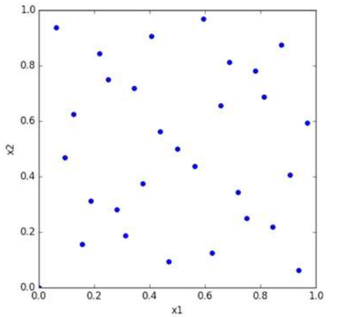
Gaussian Processes

Adaptive sampling

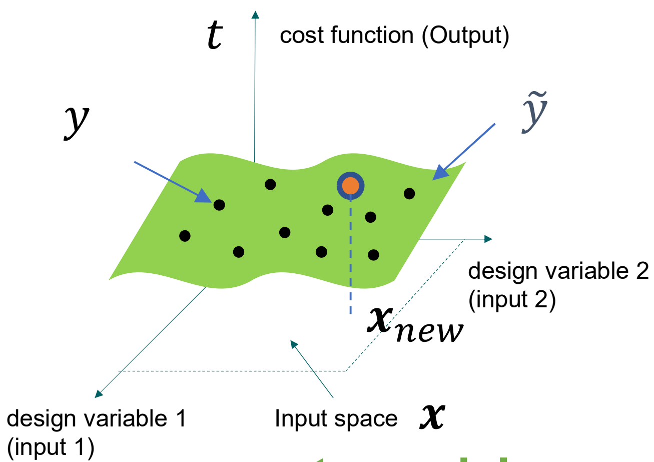
surrogate model

Optimization

DoE (see Lecture 4)



Sampling uniformly in the input space $\boldsymbol{x}$

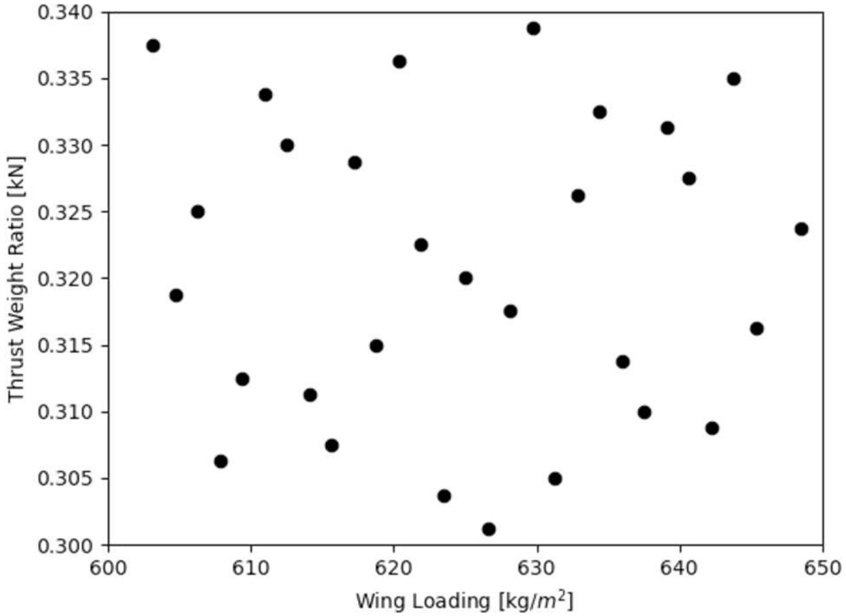


surrogate model

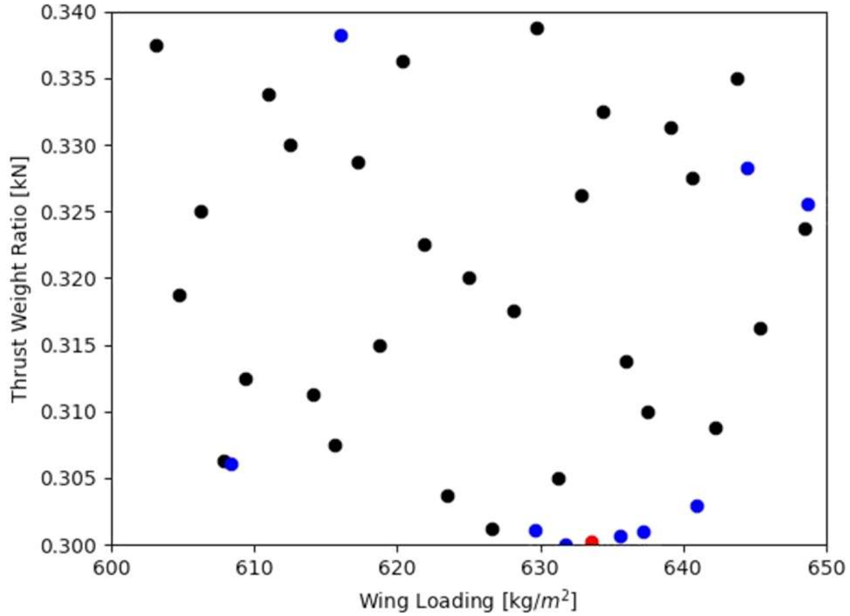
Surrogate-Based Optimization

DOE: Quasi Monte Carlo – Sobol sequence

initial DoE sample



initial DoE sample
with adaptive sample points



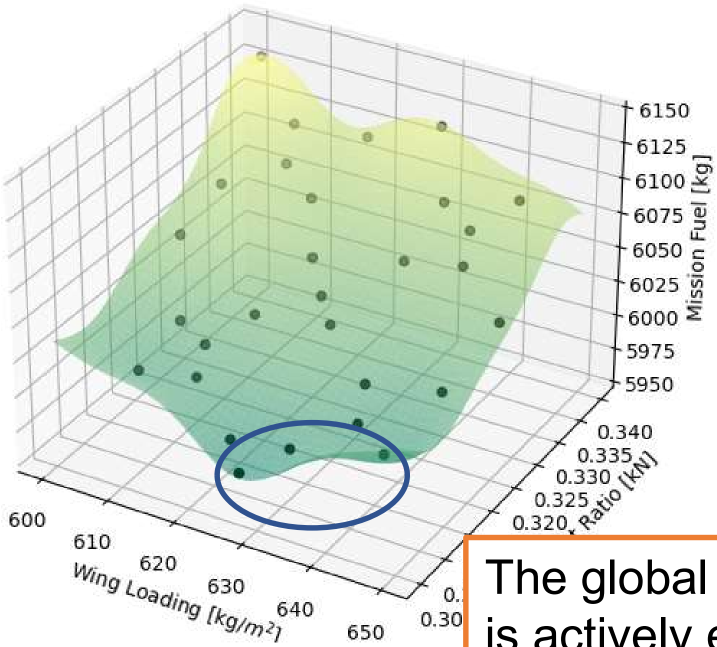
Sample points in **input design variables x space**

Wing loading $\in [600, 650]$
Thrust weight ratio $\in [0.3, 0.34]$

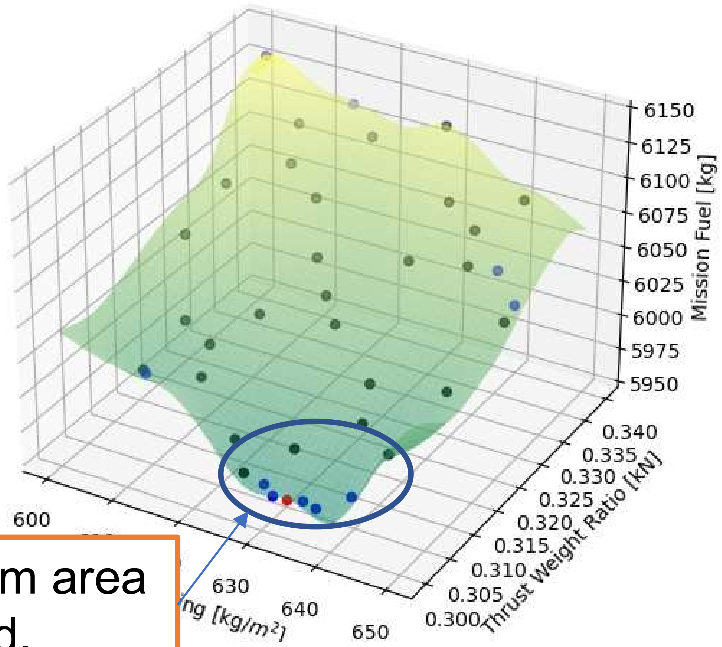
- initial DoE sample
- adaptive sample points
- the optimum point

Surrogate-Based Optimization

initial DoE sample



initial DoE sample
with adaptive sample points



The global minimum area is actively exploited.

The corresponding **cost function values** $y(\boldsymbol{x})$ on the sample points $\boldsymbol{x}$, and **surrogate models** $\tilde{y}(\boldsymbol{x})$ (by Gaussian Processes)

Technical Issues

A multivariate Gaussian distribution

$$p(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \underset{\text{covariance}}{\underset{\uparrow}{\boldsymbol{\Sigma}}})$$

$$\boldsymbol{\Sigma}(\mathbf{T},\mathbf{T}) = \begin{pmatrix} cov[t(\boldsymbol{x}_1), t(\boldsymbol{x}_1)] & \cdots & cov[t(\boldsymbol{x}_1), t(\boldsymbol{x}_N)] \\ \vdots & \ddots & \vdots \\ cov[t(\boldsymbol{x}_N), t(\boldsymbol{x}_1)] & \cdots & cov[t(\boldsymbol{x}_N), t(\boldsymbol{x}_N)] \end{pmatrix}$$

$$= {\sigma_{\mathrm{T}}}^2 \begin{pmatrix} cor[t(\boldsymbol{x}_1), t(\boldsymbol{x}_1)] & \cdots & cor[t(\boldsymbol{x}_1), t(\boldsymbol{x}_N)] \\ \vdots & 1 & \vdots \\ cor[t(\boldsymbol{x}_N), t(\boldsymbol{x}_1)] & \cdots & cor[t(\boldsymbol{x}_N), t(\boldsymbol{x}_N)] \end{pmatrix}$$

$$cov[T, T'] = \frac{cor[T, T']}{\sigma_{\mathrm{T}}\sigma_{\mathrm{T'}}}$$

$$0 \le cor[T, T'] \le 1$$

$$= {\sigma_{\mathrm{T}}}^2 \underbrace{\begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix}}_{\text{represented by } \boldsymbol{x}} = {\sigma_{\mathrm{T}}}^2 \mathbf{K}$$

Gram matrix $\mathbf{K}$: correlation matrix

represented by $\boldsymbol{x}$ ⇒ $\boldsymbol{x}$ should be scaled (standardized).

Technical Issues

A multivariate Gaussian distribution

$$p(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \boldsymbol{\Sigma})$$

covariance

$$\boldsymbol{\Sigma}(\mathbf{T},\mathbf{T}) = \begin{pmatrix} cov[t(\boldsymbol{x}_1), t(\boldsymbol{x}_1)] & \cdots & cov[t(\boldsymbol{x}_1), t(\boldsymbol{x}_N)] \\ \vdots & \ddots & \vdots \\ cov[t(\boldsymbol{x}_N), t(\boldsymbol{x}_1)] & \cdots & cov[t(\boldsymbol{x}_N), t(\boldsymbol{x}_N)] \end{pmatrix}$$

$$= {\sigma_\mathrm{T}}^2 \begin{pmatrix} cor[t(\boldsymbol{x}_1), t(\boldsymbol{x}_1)] & \cdots & cor[t(\boldsymbol{x}_1), t(\boldsymbol{x}_N)] \\ \vdots & 1 & \vdots \\ cor[t(\boldsymbol{x}_N), t(\boldsymbol{x}_1)] & \cdots & cor[t(\boldsymbol{x}_N), t(\boldsymbol{x}_N)] \end{pmatrix}$$

$$cov[T, T'] = \frac{cor[T, T']}{\sigma_\mathrm{T}\sigma_{\mathrm{T}'}}$$

$$0 \leq cor[T, T'] \leq 1$$

$$= {\sigma_\mathrm{T}}^2 \underbrace{\begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1, \boldsymbol{\theta}) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_N, \boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1, \boldsymbol{\theta}) & \cdots & k(\boldsymbol{x}_N, \boldsymbol{x}_N, \boldsymbol{\theta}) \end{pmatrix}}_{\text{represented by } \boldsymbol{x}} = {\sigma_\mathrm{T}}^2 \mathbf{K}(\boldsymbol{\theta})$$

Gram matrix $\mathbf{K}(\boldsymbol{\theta})$: correlation matrix

represented by $\boldsymbol{x}$ ⇒ $\boldsymbol{x}$ should be scaled (standardized).

Technical Issues

A multivariate Gaussian distribution

$$p(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \sigma^2\mathbf{K}(\boldsymbol{\theta}))$$

Then, the **likelihood function** when data $\mathcal{D} = (\mathbf{X}, \mathbf{T})$ is given,

$$p(\mathbf{T}|\mathbf{X}) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \sigma^2\mathbf{K}(\boldsymbol{\theta})) = \frac{1}{(2\pi\sigma)^{N/2}|\mathbf{K}(\boldsymbol{\theta})|^{1/2}} \exp\left(-\frac{\mathbf{T}^{\mathrm{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{T}}{2\sigma^2}\right)$$

negative log as usual → **Error function** E

$$E(\sigma, \boldsymbol{\theta}) = -\ln p(\mathbf{T}|\mathbf{X}) = \frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\sigma^2 + \frac{1}{2}\ln|\mathbf{K}(\boldsymbol{\theta})| + \frac{\mathbf{T}^{\mathrm{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{T}}{2\sigma^2}$$

$$\frac{\partial E}{\partial \sigma} = 0, \frac{\partial E}{\partial \boldsymbol{\theta}} = 0$$

Please derive $E(\boldsymbol{\theta})$ by yourself.

cautions in computation!

$$\widehat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} E(\boldsymbol{\theta}) \qquad E(\boldsymbol{\theta}) = \frac{N}{2}\ln\hat{\sigma}(\boldsymbol{\theta})^2 + \frac{1}{2}\ln\boxed{|\mathbf{K}(\boldsymbol{\theta})|} + \mathrm{C}$$

where,

$$\hat{\sigma}(\boldsymbol{\theta})^2 = \frac{\mathbf{T}^{\mathrm{T}}\boxed{\mathbf{K}(\boldsymbol{\theta})^{-1}}\mathbf{T}}{N}$$

Useful Kernel Functions

- Gaussian kernel (RBF kernel)

$$k(\boldsymbol{x}, \boldsymbol{x}', \theta) = \exp(-\theta \|\boldsymbol{x} - \boldsymbol{x}'\|^2) \quad \text{or} \quad k(\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{\theta}) = \exp\left(-\sum_{i=1}^{D} \theta_i \left\| x^{(i)} - x'^{(i)} \right\|^2\right)$$

- Linear kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}'$$

- Exponential kernel

$$k(\boldsymbol{x}, \boldsymbol{x}', \theta) = \exp(-\theta \|\boldsymbol{x} - \boldsymbol{x}'\|) \longrightarrow k(\boldsymbol{x}, \boldsymbol{x}', \theta, p) = \exp(-\theta \|\boldsymbol{x} - \boldsymbol{x}'\|^p)$$

- Matérn kernel

$$k_\nu(\boldsymbol{x}, \boldsymbol{x}', \theta) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\pi}r}{\theta}\right)^\nu K_\nu\left(\frac{\sqrt{2\pi}r}{\theta}\right)$$

$$r = \|\boldsymbol{x} - \boldsymbol{x}'\|$$

when $\nu = \frac{1}{2}$, Exponential kernel

when $\nu = \frac{3}{2}$, Matérn3

when $\nu = \frac{5}{2}$, Matérn5

when $\nu = \infty$, Gaussian kernel

Lecture content

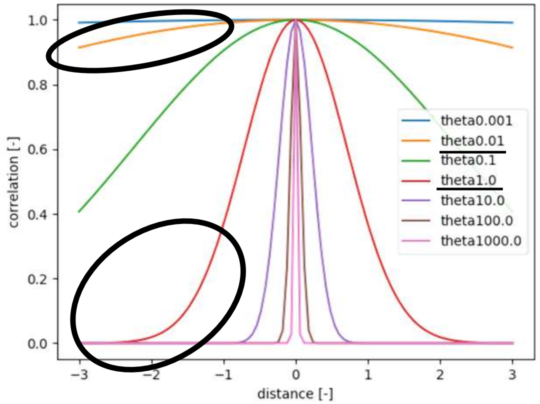
- Advanced Methods/Topics

Sensitivity Analysis (Automatic Relevance Determination - ARD)

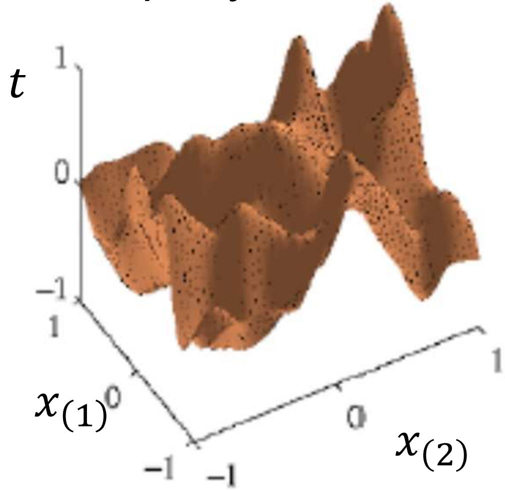
A kernel with hyperparameters at each component of the input

$$k(\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{\theta}) = \exp\left(-\sum_{i=1}^{2} \theta_i \left\| x^{(i)} - x'^{(i)} \right\|^2 \right)$$

e.g. 2D case
$\boldsymbol{x} = (x_{(1)}, x_{(2)})$

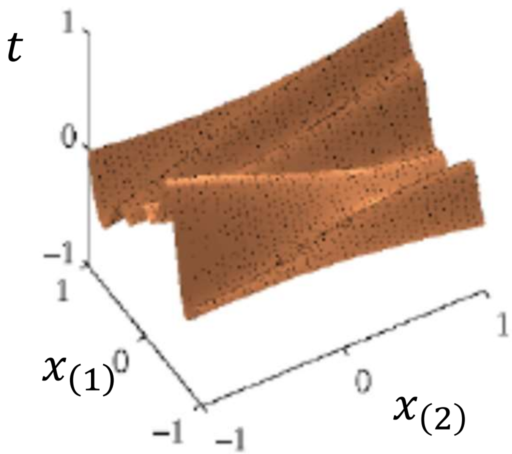


Equally sensitive



$\boldsymbol{\theta} = (\theta_1, \theta_2)$
$= (1, 1)$

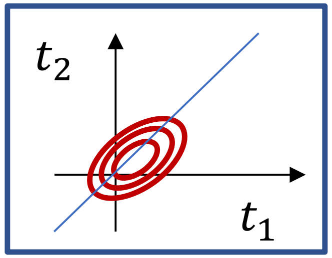
Not that sensitive in $x_{(2)}$



PRML, p.312

$\boldsymbol{\theta} = (\theta_1, \theta_2)$
$= (1, 0.01)$

t is correlated strongly in $x_{(2)}$ (even the distance $\|x_i - x_i'\|^2$ is large).



$t_1, t_2, t_3, \cdots$ are very similar in $x_{(2)}$.

After MLE with obtaining $\hat{\boldsymbol{\theta}}$, the info of the sensitivity analysis can be obtained.

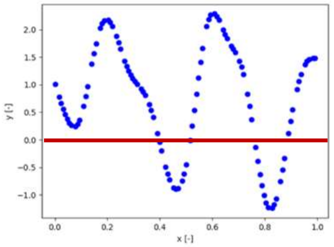
Other variations (from Kriging)

Note:
A little different explanation from that from Kriging in geostatistics

A multivariate Gaussian distribution

$$p(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}, \sigma^2\mathbf{K}(\boldsymbol{\theta}))$$

We can model also the trend.



the trend: $\boldsymbol{\mu}$

When $\boldsymbol{\mu}$ is:

1. $\boldsymbol{\mu} = \mathbf{0}$ (or $\bar{\mathbf{T}}$)
   Simple Kriging

2. $\boldsymbol{\mu} = \mathbf{1}\mu$: a constant value
   Ordinary Kriging
   $$\hat{\mu}(\boldsymbol{\theta}) = \left(\mathbf{1}^{\mathrm{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{1}\right)^{-1}\mathbf{1}^{\mathrm{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{T}$$

3. $\boldsymbol{\mu} = \boldsymbol{\Phi}\boldsymbol{w}$: linear regression model
   Universal Kriging
   $$\hat{\boldsymbol{w}}(\boldsymbol{\theta}) = \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{T}$$

$\hat{\boldsymbol{\mu}}$ (by MLE) in general has the closed-form (analytically expressed by $\boldsymbol{\theta}$) as **generalized least-squares (GLS)**.

$$\boxed{\hat{\boldsymbol{w}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{T}}$$

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{K}(\boldsymbol{\theta})$$

when,

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$$

(see Lecture 4)

$$\boxed{\hat{\boldsymbol{w}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{T}}$$

Least-squares in linear regression

ordinary least-squares (OLS).

Other variations (from Kriging)

A multivariate Gaussian distribution

$$p(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}, \sigma^2\mathbf{K}(\boldsymbol{\theta}))$$

Basically the same as Slide 25

Then, the **likelihood function** when data $\mathcal{D} = (\mathbf{X}, \mathbf{T})$ is given,

$$p(\mathbf{T}|\mathbf{X}) = \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}, \sigma^2\mathbf{K}(\boldsymbol{\theta})) = \frac{1}{(2\pi\sigma)^{N/2}|\mathbf{K}(\boldsymbol{\theta})|^{1/2}} \exp\left(-\frac{(\mathbf{T}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}(\mathbf{T}-\boldsymbol{\mu})}{2\sigma^2}\right)$$

negative log as usual → **Error function** E

$$E(\boldsymbol{\mu}, \sigma, \boldsymbol{\theta}) = -\ln p(\mathbf{T}|\mathbf{X}) = \frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\sigma^2 + \frac{1}{2}\ln|\mathbf{K}(\boldsymbol{\theta})| + \frac{(\mathbf{T}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}(\mathbf{T}-\boldsymbol{\mu})}{2\sigma^2}$$

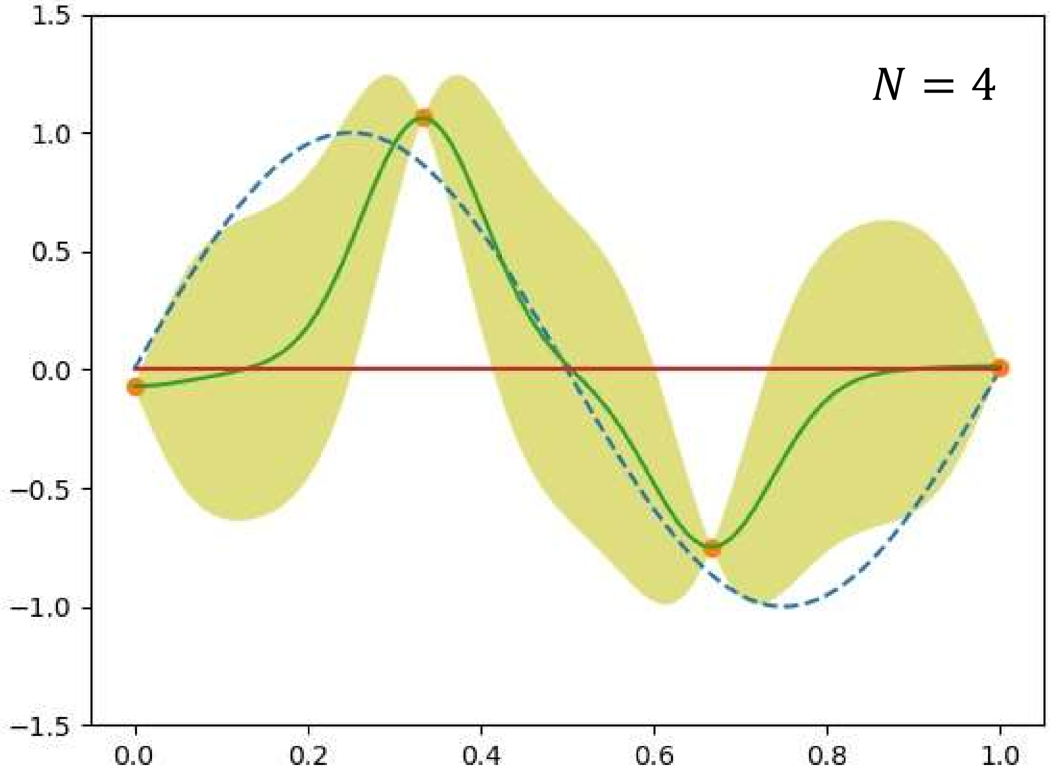
$$\frac{\partial E}{\partial \boldsymbol{\mu}} = 0, \frac{\partial E}{\partial \sigma} = 0, \frac{\partial E}{\partial \boldsymbol{\theta}} = 0$$

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} E(\boldsymbol{\theta}) \quad E(\boldsymbol{\theta}) = \frac{N}{2}\ln\hat{\sigma}(\boldsymbol{\theta})^2 + \frac{1}{2}\ln|\mathbf{K}(\boldsymbol{\theta})| + \mathrm{C}$$

$$\hat{\boldsymbol{\mu}} = \mathrm{GLS}$$

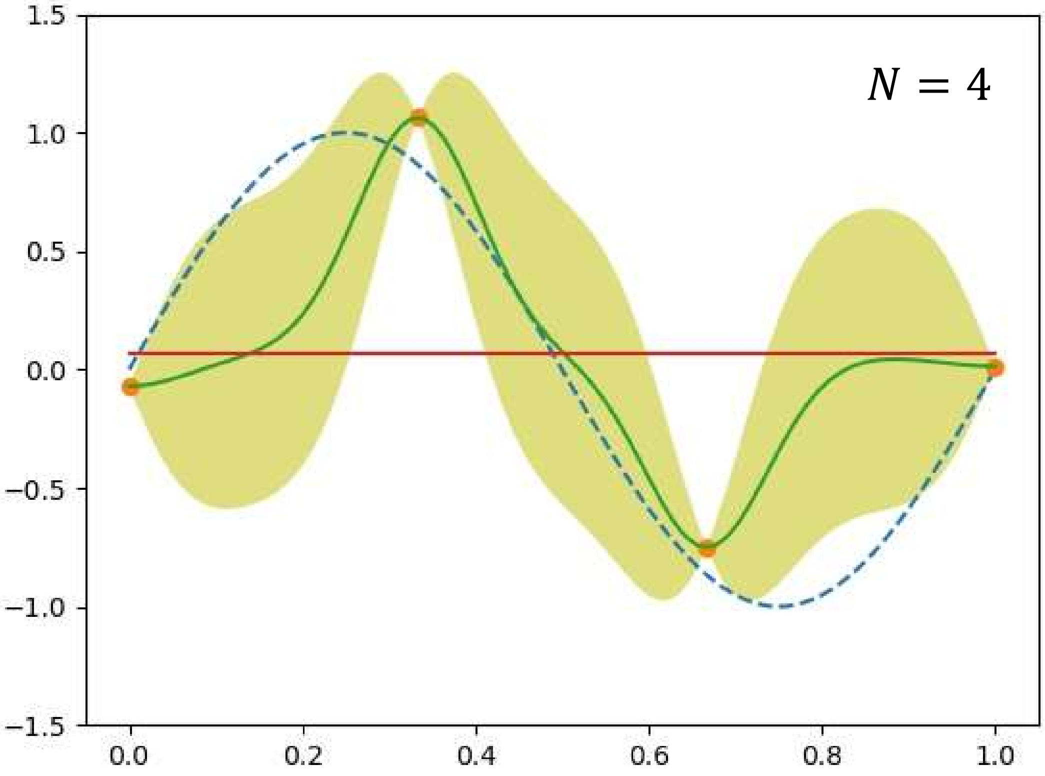
$$\hat{\sigma}(\boldsymbol{\theta})^2 = \frac{(\mathbf{T}-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}(\mathbf{T}-\hat{\boldsymbol{\mu}})}{N}$$

Simple Kriging (Normal Gaussian Process)



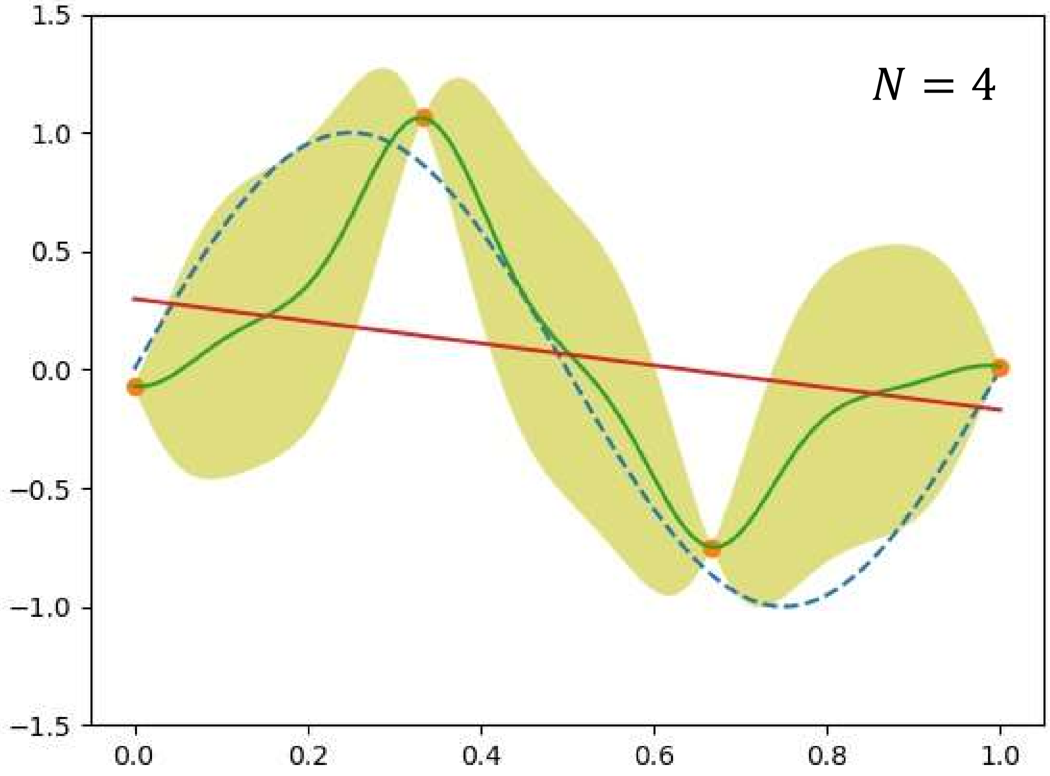
$\hat{\boldsymbol{\mu}} = \mathbf{0}$

Ordinary Kriging



$\hat{\boldsymbol{\mu}} = \mathbf{1}\hat{\mu}$

Universal Kriging

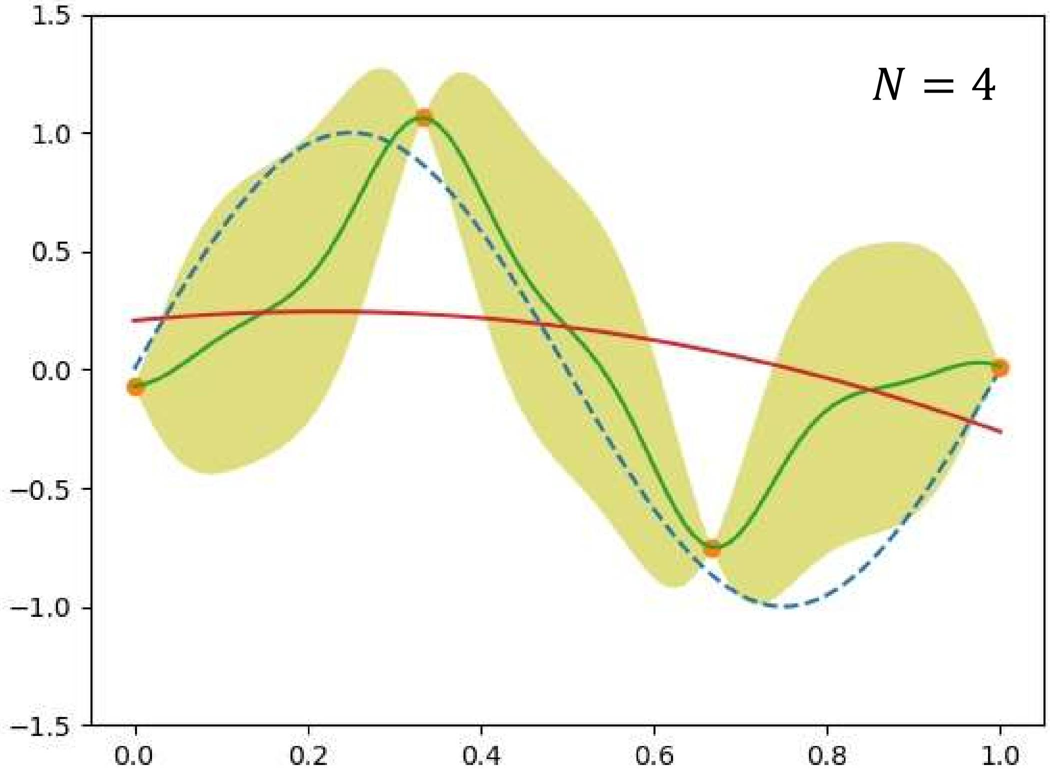


$\hat{\boldsymbol{\mu}} = \boldsymbol{\Phi}\hat{\boldsymbol{w}}$

$\boldsymbol{\phi}(x)$
$= (x^0, x^1)^{\mathrm{T}}$

Trend:
linear function

Universal Kriging



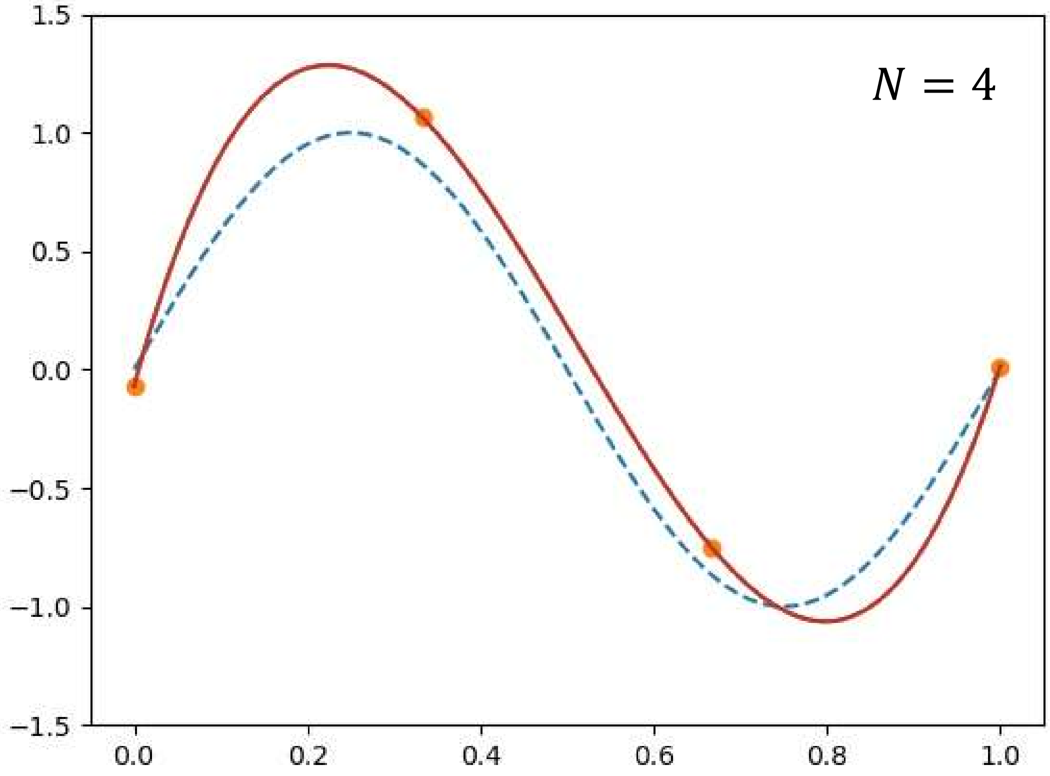
$\hat{\boldsymbol{\mu}} = \boldsymbol{\Phi}\hat{\boldsymbol{w}}$

$\boldsymbol{\phi}(x)$
$= (x^0, x^1, x^2)^{\mathrm{T}}$

Trend:
quadratic function

Universal Kriging

If you know the characteristic of the trend (it is a cubic function in this case), the universal Kriging works better (not always).

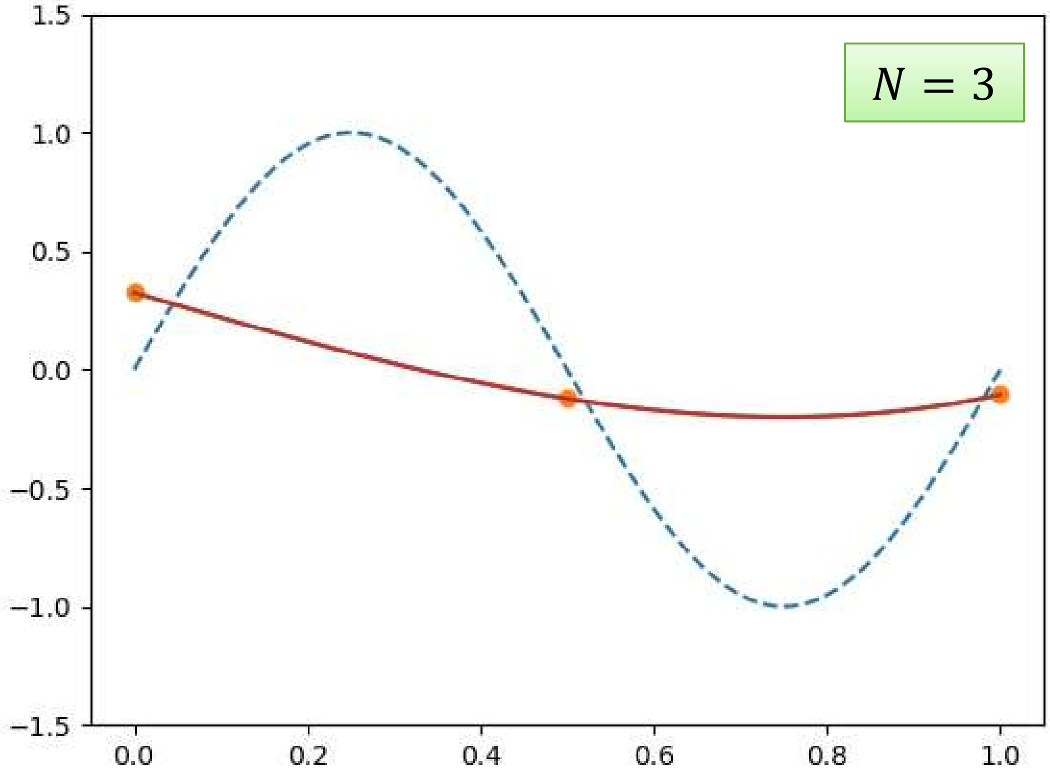


$\hat{\boldsymbol{\mu}} = \boldsymbol{\Phi}\hat{\boldsymbol{w}}$

$\boldsymbol{\phi}(x) = (x^0, x^1, x^2, x^3)^{\mathrm{T}}$

Trend:
cubic function

Universal Kriging (3/3)



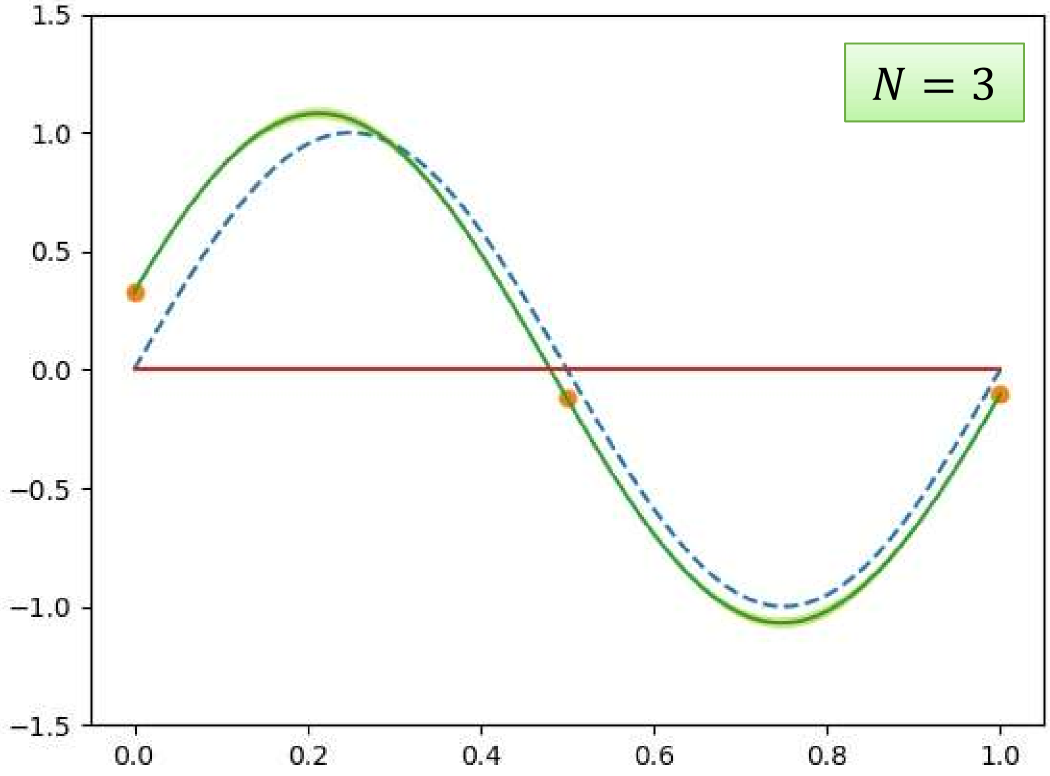
$\hat{\boldsymbol{\mu}} = \boldsymbol{\Phi}\hat{\boldsymbol{w}}$

$\boldsymbol{\phi}(x)$
$= (x^0, x^1, x^2, x^3)^{\mathrm{T}}$

Trend:
cubic function

Gradient Enhanced

Instead of adding new sample points, using **the gradient information** of the existing sample points



$\hat{\boldsymbol{\mu}} = \mathbf{0}$

(based on Simple Kriging in this case)

Gradient Enhanced

when the kernel function is differentiable w.r.t. $\boldsymbol{x}$

Instead of obtaining new sample points

e.g. $$k(\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{\theta}) = \exp\left(-\sum_{i=1}^{D} \theta_i \left\| x^{(i)} - x'^{(i)} \right\|^2\right)$$

$$\dot{\mathbf{K}} = \begin{pmatrix} \mathbf{K} & \dfrac{\partial \mathbf{K}}{\partial x^{(i)}} \\ \dfrac{\partial \mathbf{K}}{\partial x^{(j)}} & \dfrac{\partial^2 \mathbf{K}}{\partial x^{(i)} \partial x^{(j)}} \end{pmatrix} \quad i, j = 1, \cdots, D$$

$$\begin{array}{c|cc} & N & ND \\ \hline N & \mathbf{K} & \dfrac{\partial \mathbf{K}}{\partial x^{(i)}} \\ ND & \dfrac{\partial \mathbf{K}}{\partial x^{(j)}} & \dfrac{\partial^2 \mathbf{K}}{\partial x^{(i)} \partial x^{(j)}} \end{array}$$

the size of the matrix

$\mathbf{K}$: $N \times N$

$\dot{\mathbf{K}}$: $N(1 + D) \times N(1 + D)$

Computational time (during MLE):
cubic proportional to the size of the matrix

Gradient Enhanced

Practical applications:

- When the computations of the gradients are relatively cheap:
  - **Automatic differentiation** (the chain rule)
    - Forward mode
    - Reverse mode
      - e.g. single output t, multiple input $\boldsymbol{x}$

Topics of "calculus"

Note:
AD by the reverse mode is also used in neural networks for different purposes (Lecture 10).

Examples:
- adjoint solvers in computational fluid dynamics (CFD)

If there are tools to efficiently compute the gradients, the gradient-enhanced model is very useful.

Pros and Cons (Gaussian Process / Kriging in General)

Pros

- Weak assumption on using Bayesian approach
  - The model uncertainty can be analytically obtained.

Due to this property, the Gaussian processes are **powerful when data is expensive to obtain**.

In contrast to Big Data

Cons

- Computational cost in large datasets - $O(N^3)$
  - (topics of “linear algebra” and “computer science”)

within 1 sec. to days!

Especially in the cases of high-dimensional input $\boldsymbol{x}$, the sample size N tends to large.

Technical Issues on Implementation (1/2)

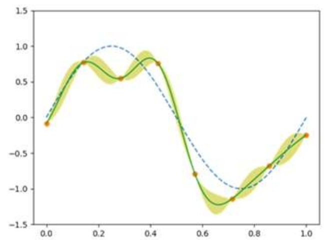
Computation of the likelihood function (error function)

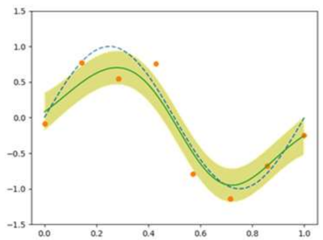
$$E(\boldsymbol{\theta}) = \frac{N}{2} \ln \hat{\sigma}(\boldsymbol{\theta})^2 + \frac{1}{2} \ln \boxed{|\mathbf{K}(\boldsymbol{\theta})|} + \mathrm{C} \qquad \hat{\sigma}(\boldsymbol{\theta})^2 = \frac{\mathbf{T}^{\mathrm{T}} \boxed{\mathbf{K}(\boldsymbol{\theta})^{-1}} \mathbf{T}}{N}$$

- Effectively use "LU decomposition" and it results
- Use double precision
- If numerically the above process is unstable, a fixed value can be added on the diagonal of $\mathbf{K}(\boldsymbol{\theta})$.

⇒ <u>Theoretically</u> this is equivalent to avoid overfitting for noisy data.

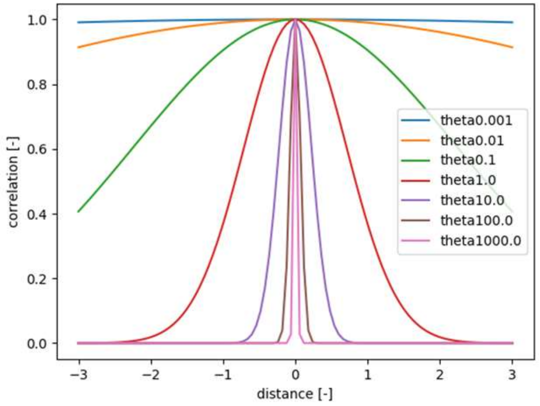
See Lecture 8, slides 30-34



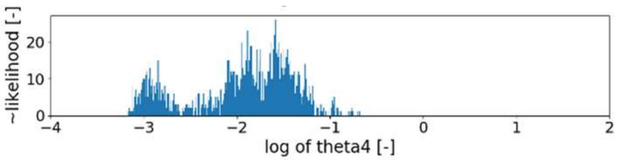


Technical Issues on Implementation (2/2)

- The hyperparameter $\boldsymbol{\theta}$ should be searched in the **log-scale**.



- The likelihood function (error function) $E(\boldsymbol{\theta})$ is multimodal.
  - $\frac{\partial E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ can be analytically obtained but **global optimization methods** is useful.



MCMC was used to make this fig.

- The input $\boldsymbol{x}$ should be standardized ($0 \leq \boldsymbol{x} \leq 1$).

A common kernel function can be used and ARD can be used.

Technical Issues on Implementation - Advanced

Advanced topics (**research level**):

- “Cholesky decomposition” (instead of LU decomposition) for further speed-up

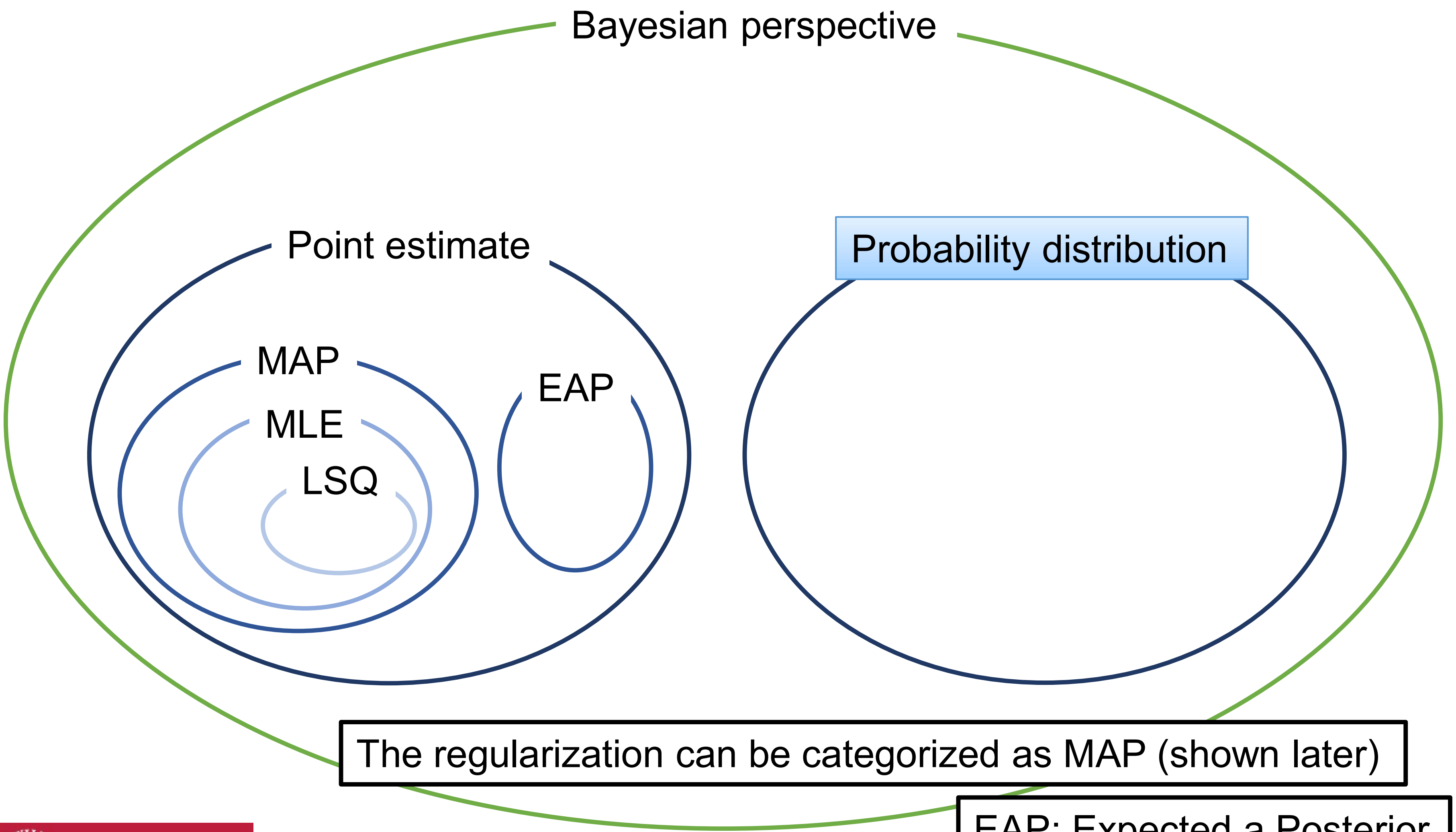
- There are various techniques to reduce the computational costs in large datasets.
  - Basically approximation

Other Advanced Methods

Other advanced topics that were not shown here:

- Cokriging (many variations in different communities)
  - Variable-fidelity methods
- Non-stationary kernel
- Regression Kriging (essentially the same as Universal Kriging)
- Deep Gaussian process

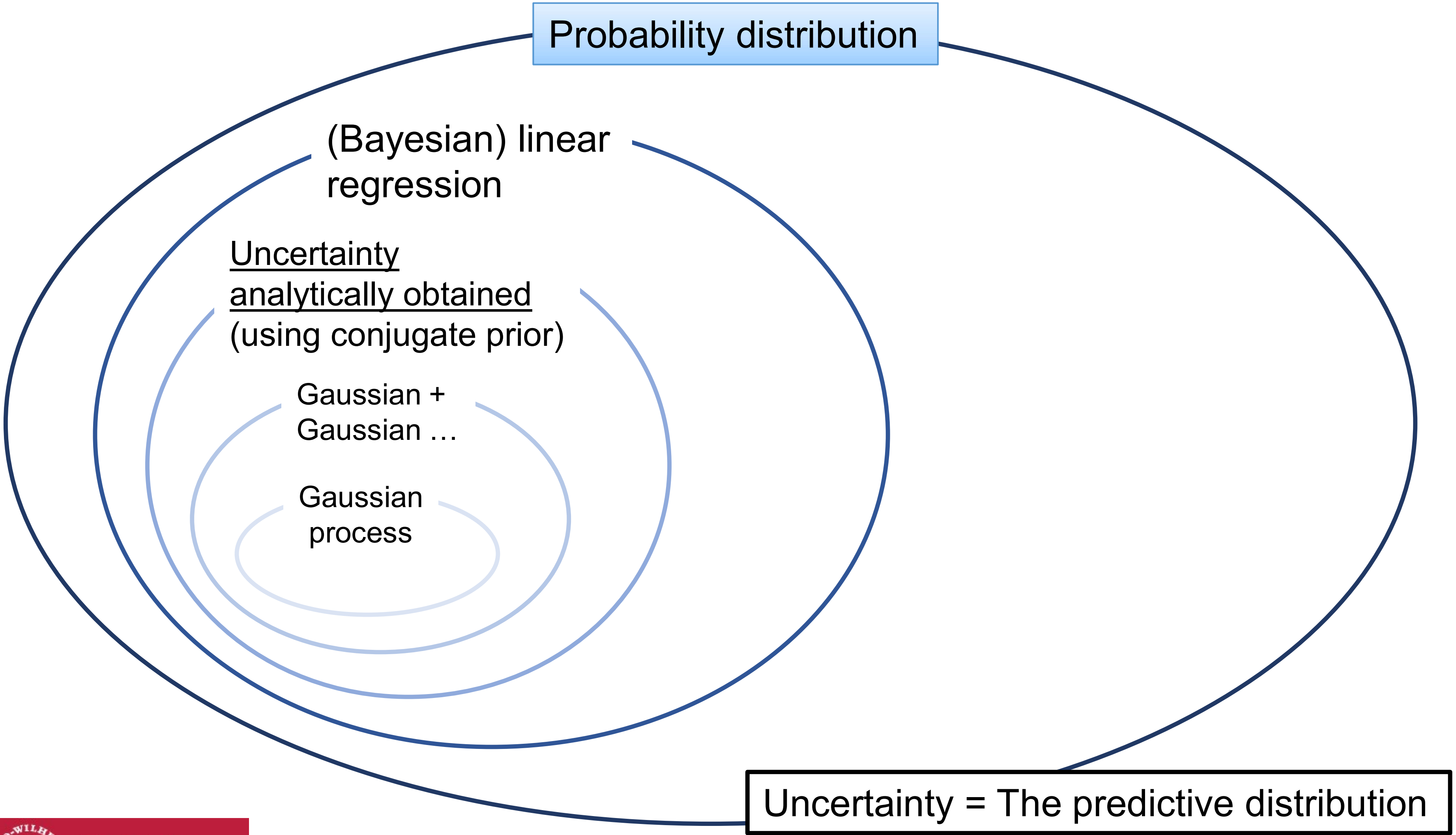
Bayesian Approach – Generalized Perspective (Lecture 6)



The regularization can be categorized as MAP (shown later)

EAP: Expected a Posterior

Uncertainty (to obtain the predictive distribution)



Uncertainty (to obtain the predictive distribution)