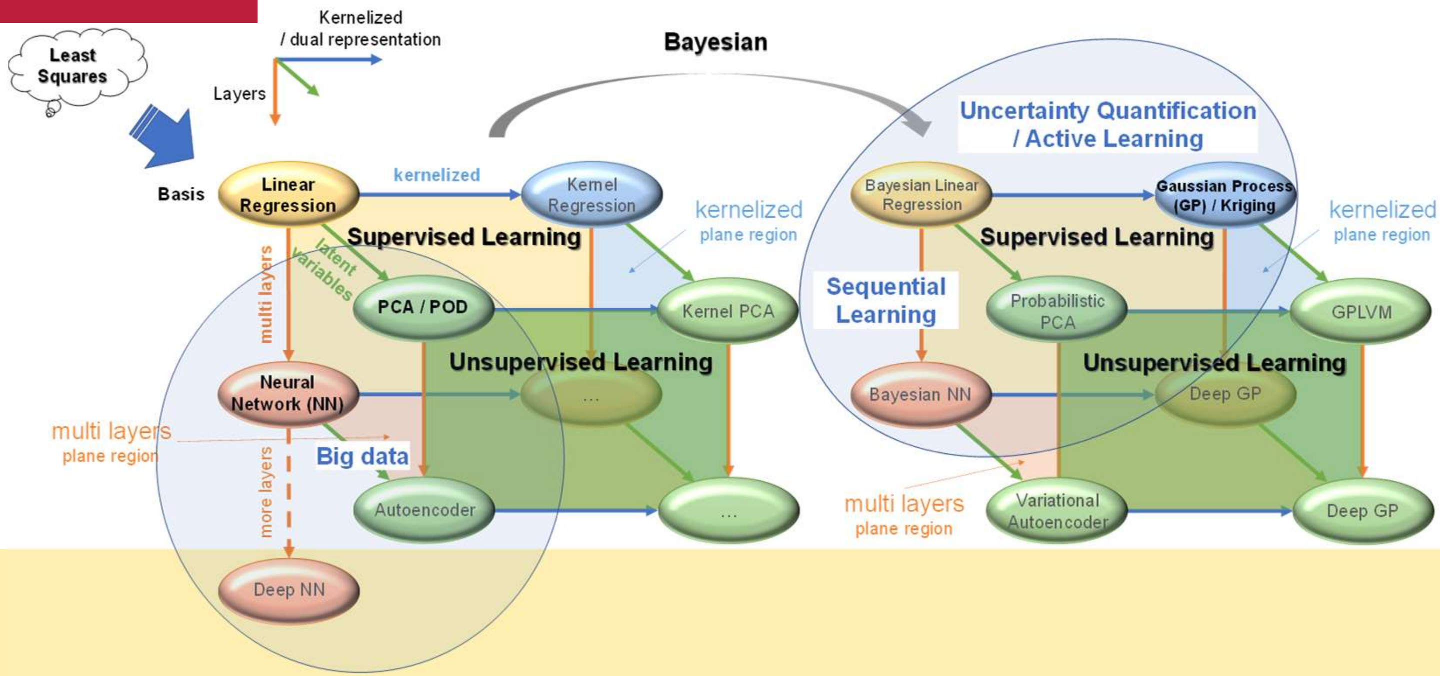


Scientific Machine Learning

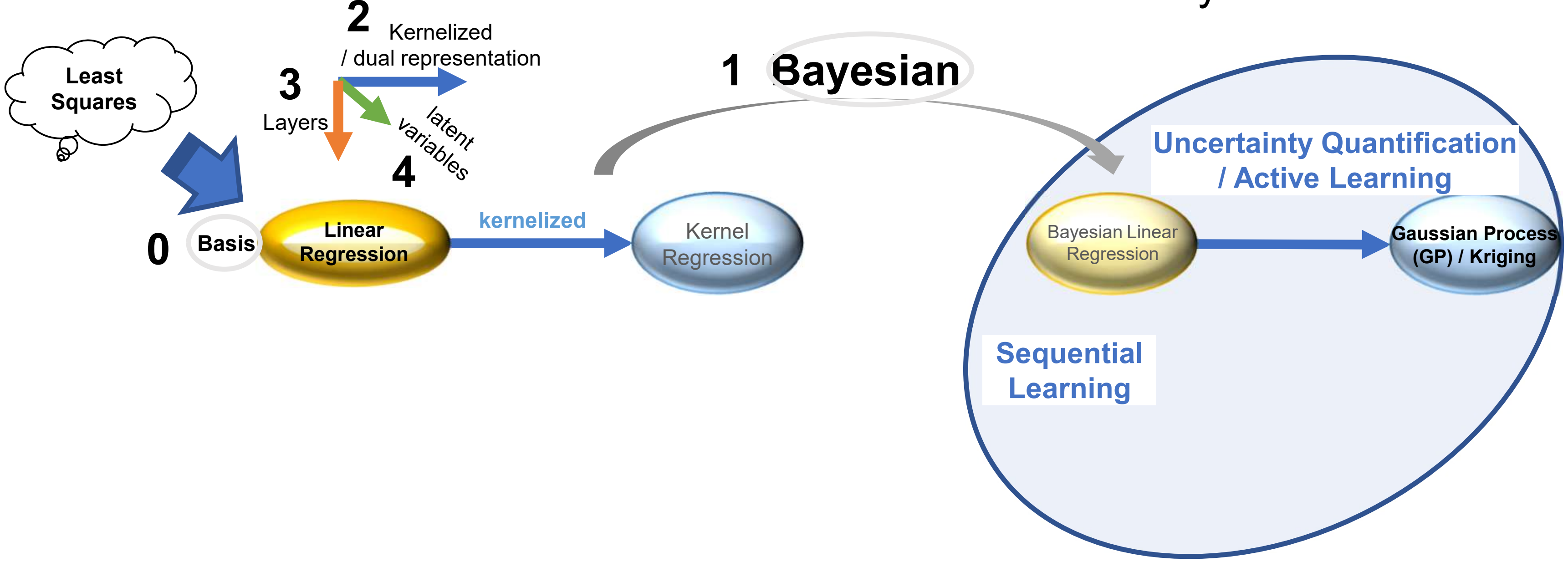
Lecture 8: Gaussian Process (1/2)

Dr. Daigo Maruyama

Prof. Dr. Ali Elham

Key Components

the three axes + Bayesian axis = four axes



Lecture content

- Introduction of Gaussian Processes (1)
- Introduction of Gaussian Processes (2)
- Learning hyperparameters in kernel functions
- Examples (analogy with Bayesian linear regression)

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.

Where are we going now?

We are going to learn:

If one sentence is used to explain them:

- **Gaussian Processes** → The probabilistic model is a multivariate Gaussian distribution.
- **Neural Networks** → Nonlinear regression

by learning tools now.

Gaussian Processes

Gaussian Processes (GPs)

In engineering (application) viewpoints:

- The regression model passes through all the sample points.
  - The regularization techniques also can be used.

- Uncertainty information can be used:
  - To show error bounds,
  - For new sample points.

Gaussian Processes

Gaussian Processes (GPs)

In theoretical (systematic) viewpoints:

- Bayesian linear regression in another expression (dual representation using kernel).
  - It is natural to have the uncertainty information
  - BUT, the model needs only **weak assumption**!

Bayesian linear regression
Specify the nonlinear function $\boldsymbol{\phi}(\boldsymbol{x})$ and the dimensionality of the parameter $\boldsymbol{w}$

$$\boldsymbol{\phi}(\boldsymbol{x}) = (1, x, x^2, x^3, \cdots)$$

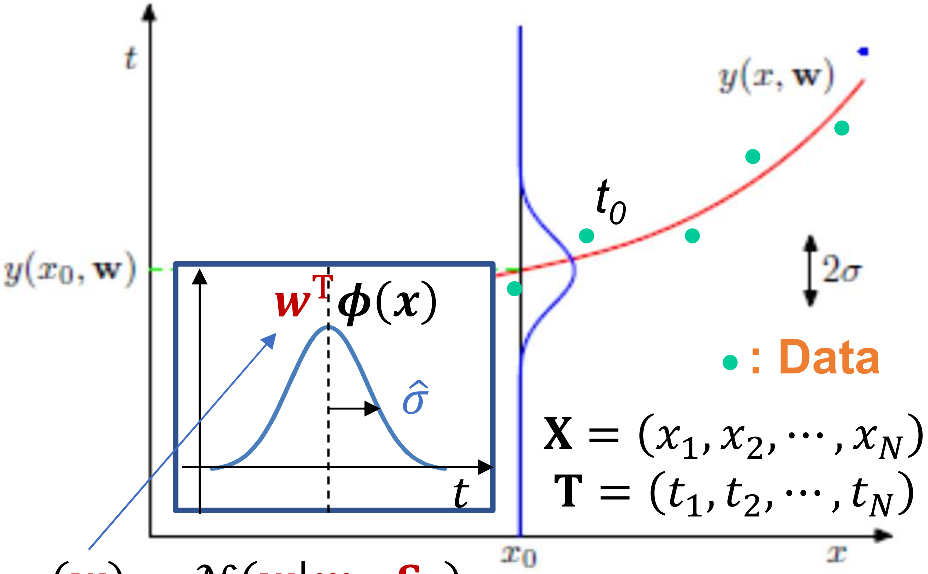
- How many degrees do we set?
- Which value do we choose as the regularization parameter λ?

Gaussian Process
Only one kernel function $k(\boldsymbol{x}, \boldsymbol{x}', \theta)$

Dual Representation (REVIEW)

Probabilistic model:
An isotropic Gaussian distribution

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

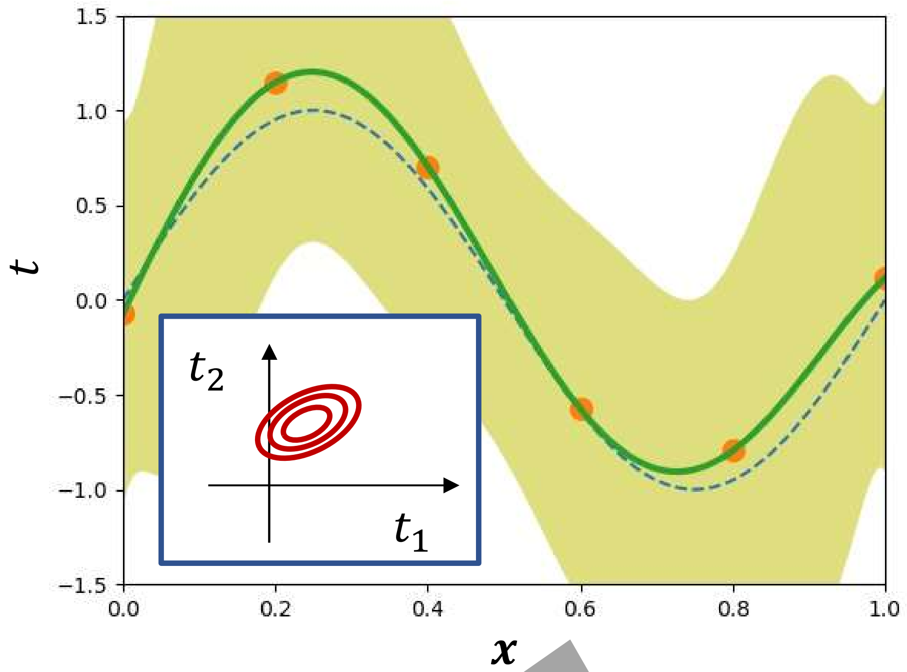


$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \mathbf{S}_0)$

Prior of $\boldsymbol{w}$

Predictive distribution:
A multivariate Gaussian distribution

$$p(\mathbf{t}|\mathbf{x},\mathbf{X},\mathbf{T}) = \mathcal{N}\left(\mathbf{t}\middle|\boldsymbol{\Phi}\boldsymbol{m}_N,\hat{\sigma}^2\mathbf{I}+\boldsymbol{\Phi}\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\right)$$



Start (traditional)

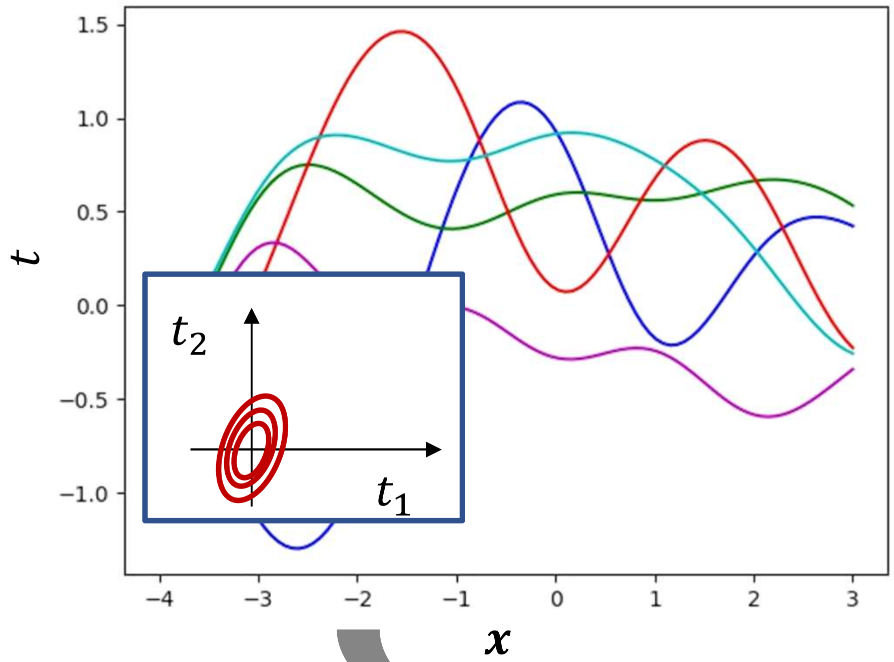
Arbitrary data: $\mathbf{X}, \mathbf{T}$

Goal (traditional)

Dual Representation (REVIEW)

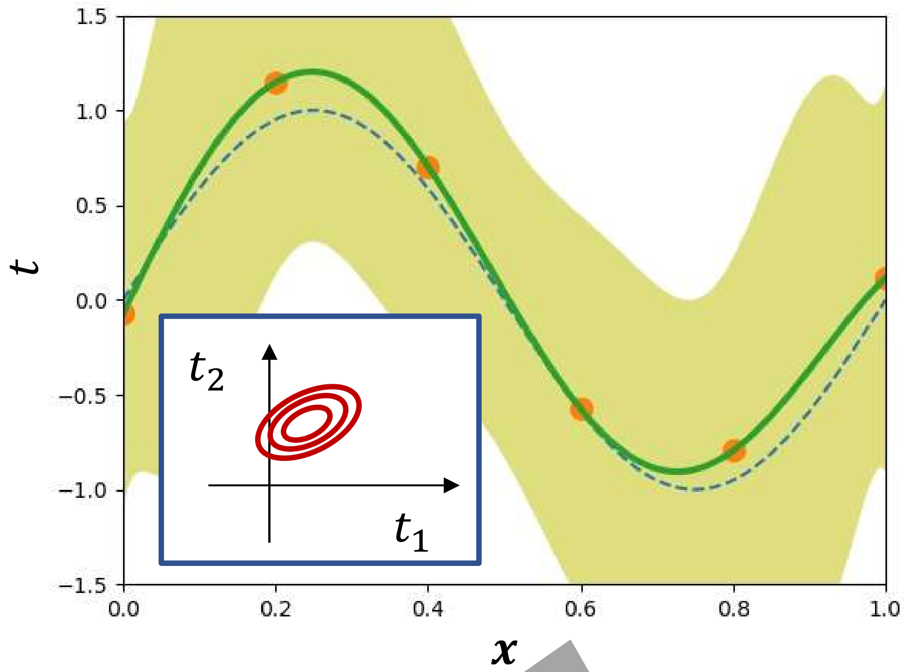
Probabilistic model:
An isotropic Gaussian distribution

$$p(\mathbf{t}|\mathbf{x}) = \mathcal{N}\left(\mathbf{t}\middle|\boldsymbol{\Phi}\boldsymbol{m}_0, \hat{\sigma}^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{S}_0\boldsymbol{\Phi}^{\mathrm{T}}\right)$$



Predictive distribution :
A multivariate Gaussian distribution

$$p(\mathbf{t}|\mathbf{x}, \mathbf{X}, \mathbf{T}) = \mathcal{N}\left(\mathbf{t}\middle|\boldsymbol{\Phi}\boldsymbol{m}_N, \hat{\sigma}^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\right)$$



Start (Dual)

Arbitrary data: $\mathbf{X}, \mathbf{T}$

Goal (Dual)

Lecture content

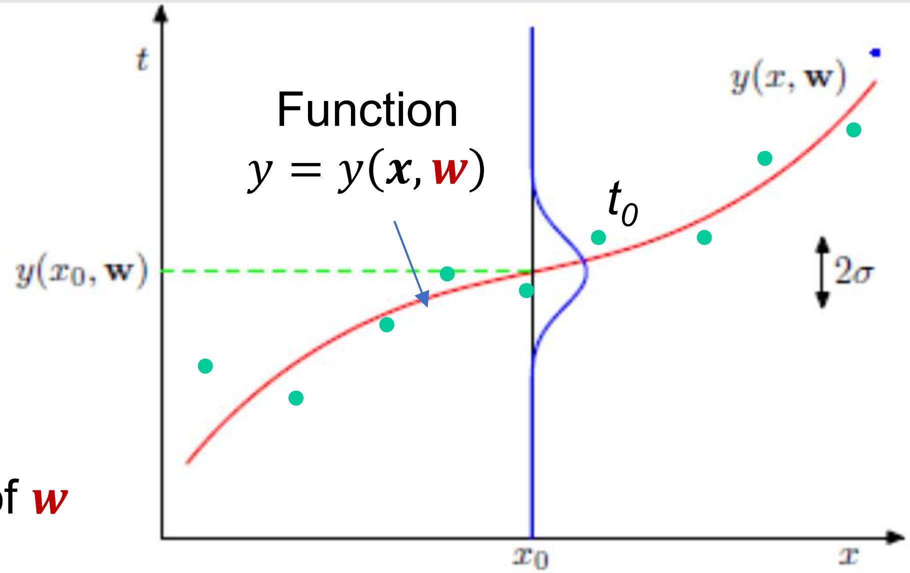
- Introduction of Gaussian Processes (1)

Gaussian Processes

Starting from a Linear Regression

$$p(t|\boldsymbol{x}, \boldsymbol{w}, \sigma) = \mathcal{N}(t|y(\boldsymbol{x}, \boldsymbol{w}), \sigma^2)$$

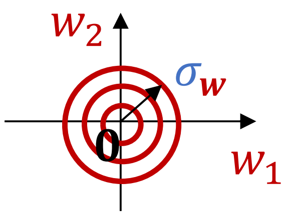
$$y(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x})$$



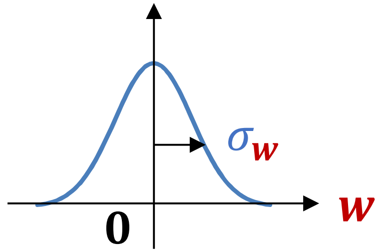
Next, we consider a Prior Distribution of $\boldsymbol{w}$

$p(\boldsymbol{w})$: an isotropic Gaussian distribution around $\mathbf{0}$

$$p(\boldsymbol{w}|\sigma_{\boldsymbol{w}}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, {\sigma_{\boldsymbol{w}}}^2 \mathbf{I})$$



$$p(\boldsymbol{w}|\sigma_{\boldsymbol{w}}) = \frac{1}{\left(\sqrt{2\pi {\sigma_{\boldsymbol{w}}}^2}\right)^{M+1}} exp\left[-\frac{\|\boldsymbol{w}\|^2}{2{\sigma_{\boldsymbol{w}}}^2}\right]$$



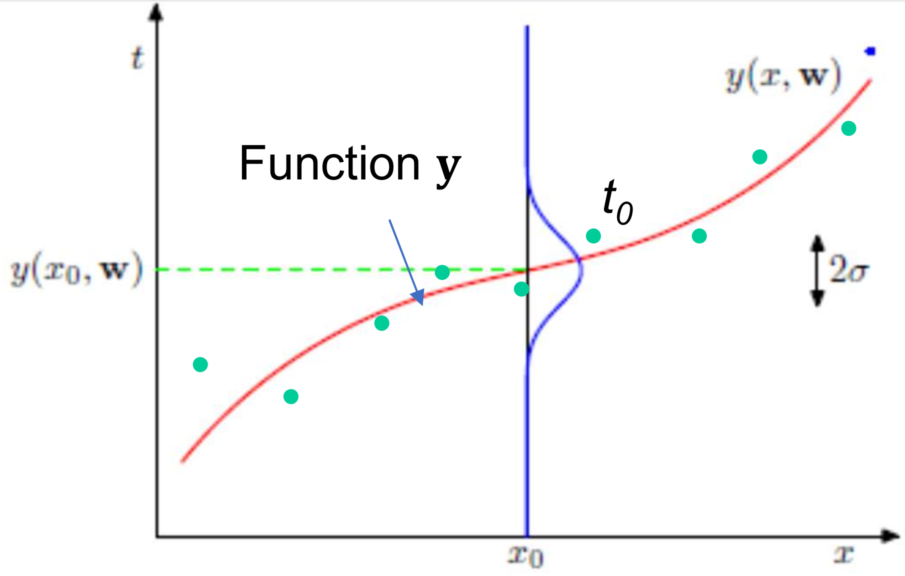
Prior distribution of $\boldsymbol{w}$

Gaussian Processes

Starting from a Linear Regression

$$p(t|\boldsymbol{x}, \boldsymbol{w}, \sigma) = \mathcal{N}(t|y(\boldsymbol{x}, \boldsymbol{w}), \sigma^2)$$

$$y(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x})$$



$$\boldsymbol{w} = (w_1, w_2, \cdots, w_M)^{\mathrm{T}}$$

$$\boldsymbol{\phi}(\boldsymbol{x}) = \left(\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \cdots, \phi_M(\boldsymbol{x})\right)^{\mathrm{T}}$$

$$\boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}) = w_1 \phi_1(\boldsymbol{x}) + w_2 \phi_2(\boldsymbol{x}) + \cdots + w_M \phi_M(\boldsymbol{x})$$

See Lecture 4

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \cdots & \phi_M(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_N) & \cdots & \phi_M(\boldsymbol{x}_N) \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_M \end{pmatrix}$$

$$\therefore \mathbf{y} = \boldsymbol{\Phi} \boldsymbol{w}$$

$N \times M$ matrix

See PRML, section 6.4.1

Gaussian Processes

$$\mathbf{y} = \underbrace{\boldsymbol{\Phi}\boldsymbol{w}}$$

Linear transformation
(linear algebra)

Linear transformation of a Gaussian distribution = a Gaussian distribution

$$\mathrm{E}[\mathbf{y}] = \boldsymbol{\Phi}\mathrm{E}[\boldsymbol{w}] = \mathbf{0}$$

$$\mathrm{cov}[\mathbf{y}] = \mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] - \underbrace{\mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^{\mathrm{T}}}_{\mathbf{0}} = \mathrm{E}[(\boldsymbol{\Phi}\boldsymbol{w})(\boldsymbol{\Phi}\boldsymbol{w})^{\mathrm{T}}] = \boldsymbol{\Phi}\underbrace{\mathrm{E}[\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}]}_{\sigma_{\boldsymbol{w}}{}^{2}\mathbf{I}}\boldsymbol{\Phi}^{\mathrm{T}} = \sigma_{\boldsymbol{w}}{}^{2}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}$$

See Lecture 4, slide 24

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathrm{E}[\mathbf{y}], \mathrm{cov}[\mathbf{y}]) = \mathcal{N}\left(\mathbf{y}\middle|\mathbf{0}, \sigma_{\boldsymbol{w}}{}^{2}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\right)$$

Gaussian Processes

$$\sigma_w{}^2\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} = \sigma_w{}^2 \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \cdots & \phi_M(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_N) & \cdots & \phi_M(\boldsymbol{x}_N) \end{pmatrix} \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \cdots & \phi_1(\boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ \phi_M(\boldsymbol{x}_1) & \cdots & \phi_M(\boldsymbol{x}_N) \end{pmatrix}$$

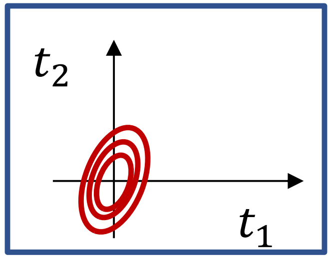
$$= \begin{pmatrix} \sigma_w{}^2\boldsymbol{\phi}(\boldsymbol{x}_1)^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_1) & \cdots & \sigma_w{}^2\boldsymbol{\phi}(\boldsymbol{x}_1)^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ \sigma_w{}^2\boldsymbol{\phi}(\boldsymbol{x}_N)^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_1) & \cdots & \sigma_w{}^2\boldsymbol{\phi}(\boldsymbol{x}_N)^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_N) \end{pmatrix}$$

$$= \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix}$$

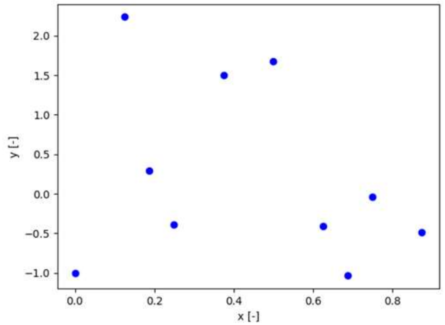
$$= \mathbf{K}$$

The kernel $k(\boldsymbol{x}, \boldsymbol{x}')$ is naturally derived.

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

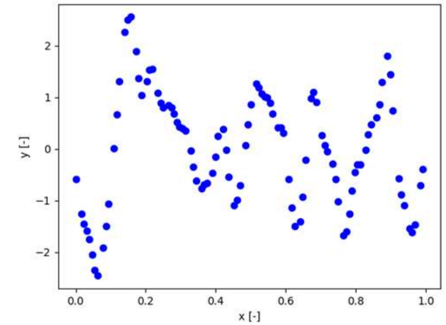


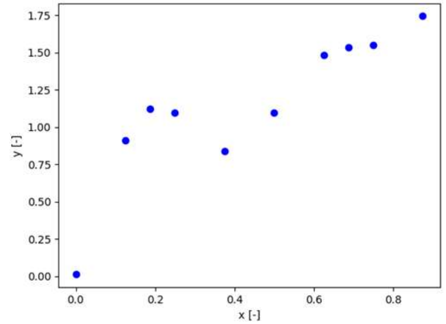
Gaussian Processes



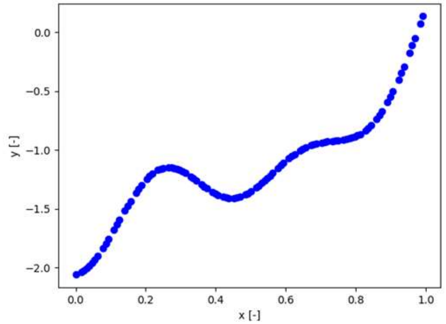
generated by
$p(y) = \mathcal{N}(y|0, \sigma^2)$?
Like **random**?

more sample
points generated





generated by ?



<u>One</u> multivariate Gaussian distribution (of infinite dimension)

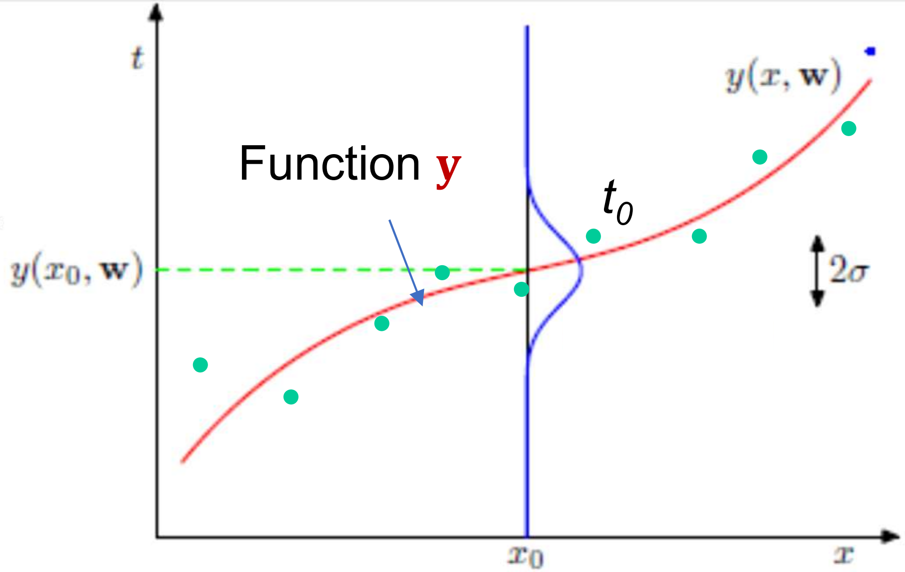
$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K})$$

Gaussian Processes

Gaussian process regression

$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K})$ $\mathbf{x}$ (as input) is omitted.

$p(\mathbf{t}|\mathbf{y}) = \mathcal{N}(\mathbf{t}|\mathbf{y}, \hat{\sigma}^2\mathbf{I})$



Our objective is $p(\mathbf{t})$.

$$p(\mathbf{t}) = \int p(\mathbf{t}, \mathbf{y})\mathrm{d}\mathbf{y} = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})\mathrm{d}\mathbf{y}$$

The concept: Lecture 5, slide 35
(will be explained and summarized more in Lecture 13)

$$= \int \mathcal{N}(\mathbf{t}|\mathbf{y}, \hat{\sigma}^2\mathbf{I})\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K})\mathrm{d}\mathbf{y}$$
$$= \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}')$$

$$\boxed{\mathbf{K}' = \mathbf{K} + \hat{\sigma}^2\mathbf{I}}$$

See PRML, section 6.4.2

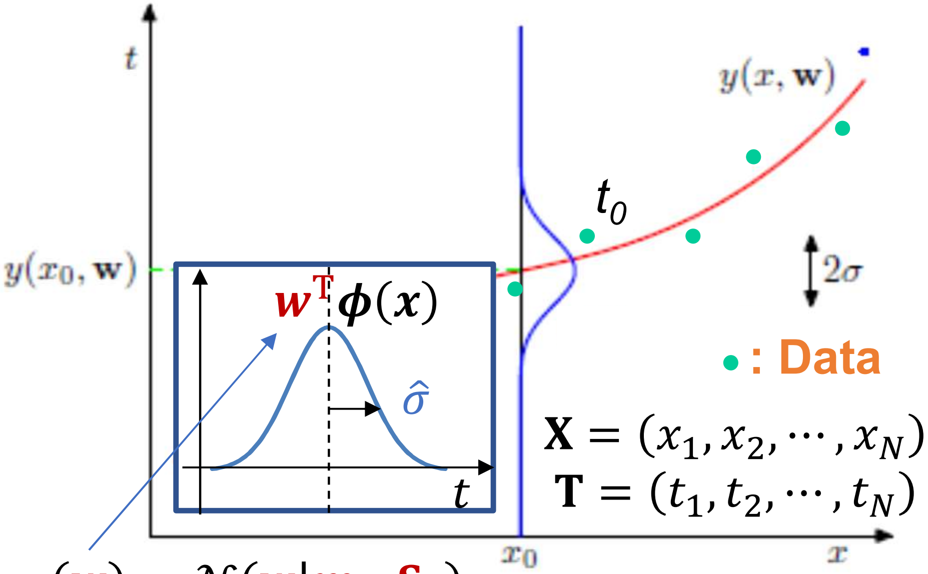
Lecture content

- Gaussian Processes (supplementary - generic perspective)

Dual Representation (REVIEW)

Probabilistic model:
An isotropic Gaussian distribution

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

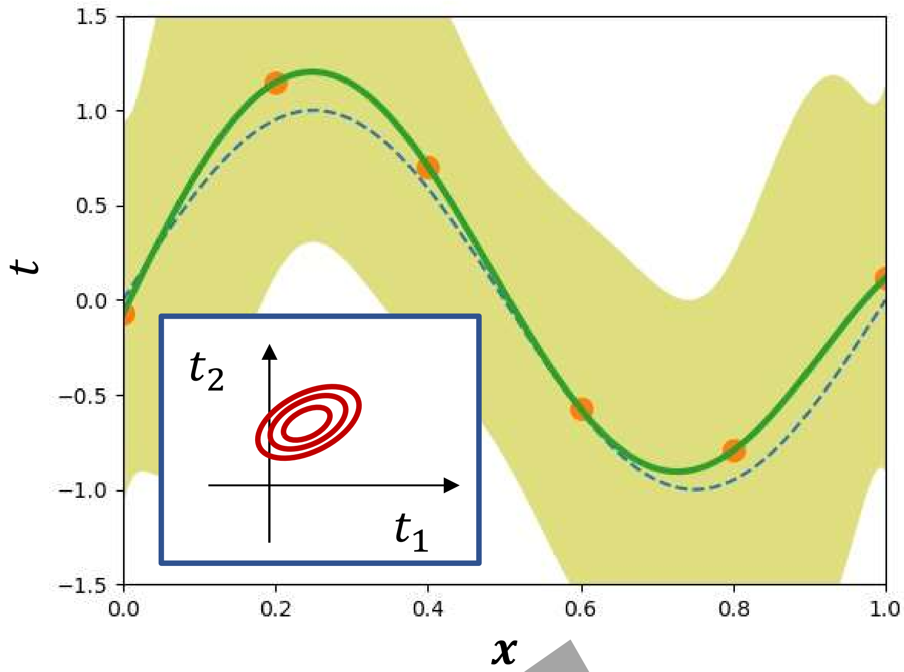


$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \mathbf{S}_0)$

Prior of $\boldsymbol{w}$

Predictive distribution:
A multivariate Gaussian distribution

$$p(\mathbf{t}|\mathbf{x}, \mathbf{X}, \mathbf{T}) = \mathcal{N}\left(\mathbf{t} \middle| \boldsymbol{\Phi}\boldsymbol{m}_N, \hat{\sigma}^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\right)$$



Start (traditional)

Arbitrary data: $\mathbf{X}, \mathbf{T}$

Goal (traditional)

Bayesian Linear Regression (REVIEW from Lecture 6)

Prior (your setting)

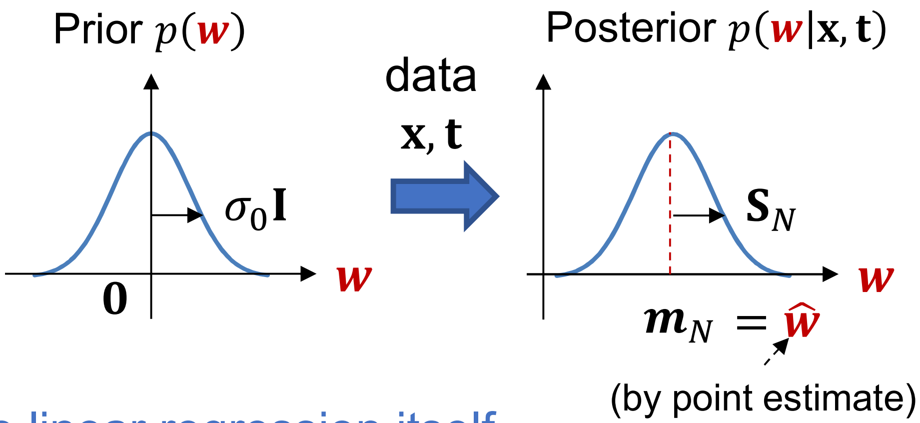
$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, {\sigma_0}^2\mathbf{I})$$

Special settings:
- Conjugate prior
  - and all Gaussian distributions
- Linear regression

Posterior

$$p(\boldsymbol{w}|\mathbf{x},\mathbf{t}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N)$$

$\boldsymbol{m}_N, \boldsymbol{S}_N$: analytically obtained



Predictive distribution (the goal)

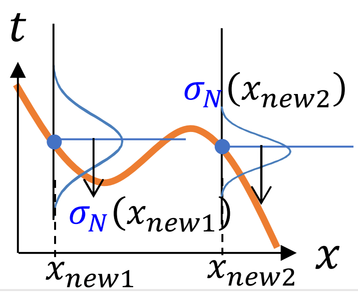
the linear regression itself

$$p(t|x,\mathbf{x},\mathbf{t}) = \int p(t|x,\boldsymbol{w})p(\boldsymbol{w}|\mathbf{x},\mathbf{t})\mathrm{d}\boldsymbol{w} = \mathcal{N}\left(t\,\middle|\,\boxed{{\boldsymbol{m}_N}^{\mathrm{T}}\phi(\boldsymbol{x})}, \underbrace{{\sigma_N}^2(\boldsymbol{x})}\right)$$

$\hat{\boldsymbol{w}}^{\mathrm{T}}\phi(\boldsymbol{x})$

What is σ_N?

The predictive distribution result contains the result of "point estimate".



please read 2.3.3 in PRML to follow the formulations.
But the concept described here is important.

Bayesian Linear Regression (REVIEW from Lecture 7)

1. We have (defined):
   - A probabilistic model (of $\mathbf{t}$)

     $$p(\mathbf{t}|\boldsymbol{w}) = \mathcal{N}\left(\mathbf{t}\middle|\boldsymbol{\Phi}\boldsymbol{w}, \hat{\Sigma}\right)$$

   - A prior distribution (of $\boldsymbol{w}$)

     $$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)$$

2. Then we obtain:
   - A posterior distribution (of $\boldsymbol{w}$)

     $$p(\boldsymbol{w}|\mathbf{t}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N)$$

   - A predictive distribution (of $\mathbf{t}$)

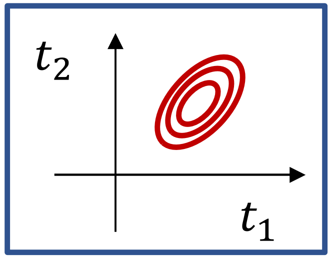
     $$p(\mathbf{t}) = \mathcal{N}\left(\mathbf{t}\middle|\boldsymbol{\Phi}\boldsymbol{m}, \hat{\Sigma} + \boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{\Phi}^{\mathrm{T}}\right)$$

$$p(\mathbf{t}|\boldsymbol{w}) = \mathcal{N}\left(\mathbf{t}\middle|\boldsymbol{\Phi}\boldsymbol{w}, \hat{\Sigma}\right)$$

$$p(\mathbf{t}) \quad = \mathcal{N}\left(\mathbf{t}\middle|\boldsymbol{\Phi}\boldsymbol{m}, \hat{\Sigma} + \boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{\Phi}^{\mathrm{T}}\right)$$

$\boldsymbol{m} = \boldsymbol{m}_0$ or $\boldsymbol{m}_N$

$\boldsymbol{S} = \boldsymbol{S}_0$ or $\boldsymbol{S}_N$



A multivariate Gaussian distribution in general

please read 2.3.3 in PRML to follow the formulations.
But the concept described here is important.

Bayesian Linear Regression (REVIEW from Lecture 7)

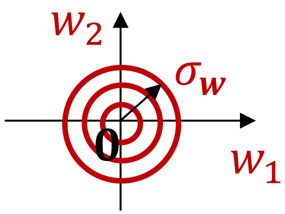
1. We have (defined):
   - A probabilistic model (of $\mathbf{t}$)

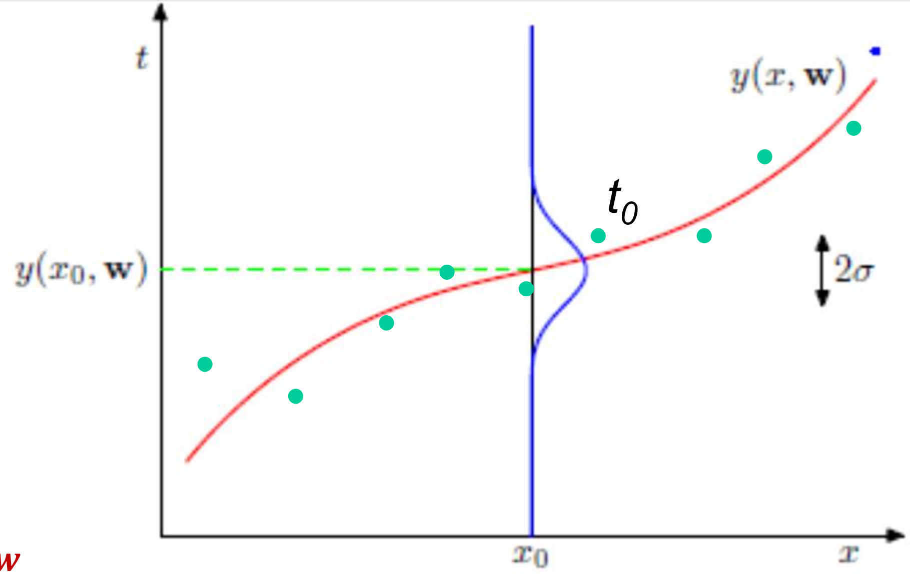
$$p(\mathbf{t}|\boldsymbol{w}) = \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\boldsymbol{w}, \hat{\sigma}^2\mathbf{I})$$

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

   - A prior distribution (of $\boldsymbol{w}$)

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, {\sigma_w}^2\mathbf{I})$$





2. Then we obtain:

   - A predictive distribution (of $\mathbf{t}$)

$$p(\mathbf{t}) = \mathcal{N}\left(\mathbf{t}\middle|\mathbf{0}, \hat{\sigma}^2\mathbf{I} + \underbrace{{\sigma_w}^2\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}}_{\mathbf{K}}\right) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \hat{\sigma}^2\mathbf{I} + \mathbf{K}) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{K}')$$

Gaussian Processes (Prediction)

using any Gram matrices $\mathbf{K}$ in general

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

Formula of conditional Gaussian distributions (Lecture 4)

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

$\mathbf{X}$ (as data) is omitted.

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

$$\mathbf{K}(\mathbf{X},\mathbf{X}) = \begin{pmatrix} k(\boldsymbol{x}_1,\boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1,\boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N,\boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_N,\boldsymbol{x}_N) \end{pmatrix}$$

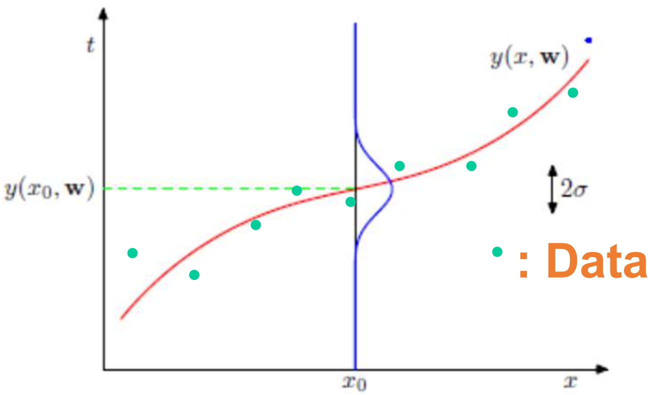
$$p\left(\begin{matrix}\mathbf{T} \\ \mathbf{t}\end{matrix}\middle|\begin{matrix}\mathbf{X} \\ \mathbf{x}\end{matrix}\right) = \mathcal{N}\left(\begin{pmatrix}\mathbf{T} \\ \mathbf{t}\end{pmatrix}\middle|\begin{pmatrix}\mathbf{0} \\ \mathbf{0}\end{pmatrix}, \underbrace{\begin{pmatrix}\mathbf{K}(\mathbf{X},\mathbf{X}) & \boldsymbol{k}(\mathbf{x}) \\ \boldsymbol{k}(\mathbf{x})^{\mathrm{T}} & \boldsymbol{k}(\mathbf{x},\mathbf{x})\end{pmatrix}}_{\text{The same kernel for all the elements}}\right)$$

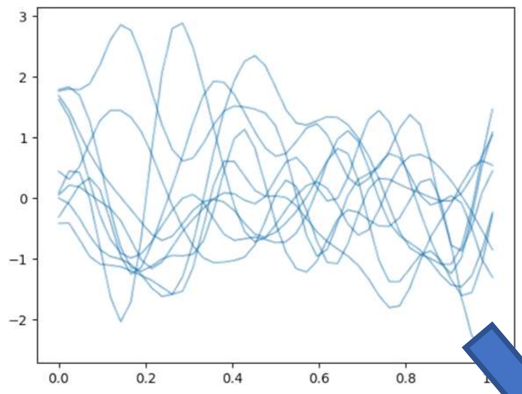
$\mathbf{x},\mathbf{t}$: new prediction

$\mathbf{X},\mathbf{T}$: data

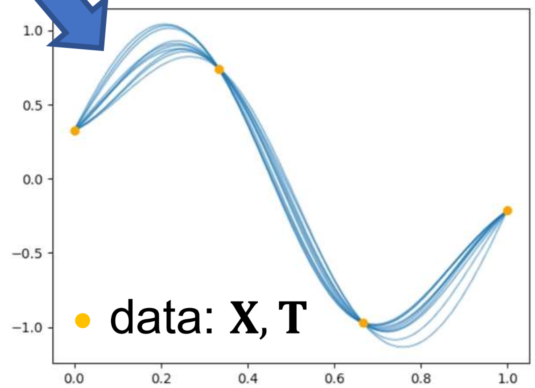
Bayesian Linear Regression

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





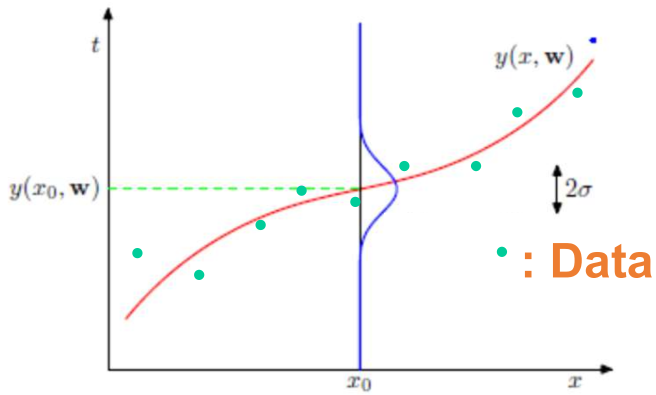
e.g. $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0},{\sigma_w}^2\mathbf{I})$



Bayesian linear regression

$$p(t|\boldsymbol{x},\mathbf{X},\mathbf{T}) = \mathcal{N}\left(t\middle|{\boldsymbol{m}_N}^{\mathrm{T}}\phi(\boldsymbol{x}),{\sigma_N}^2(\boldsymbol{x})\right)$$

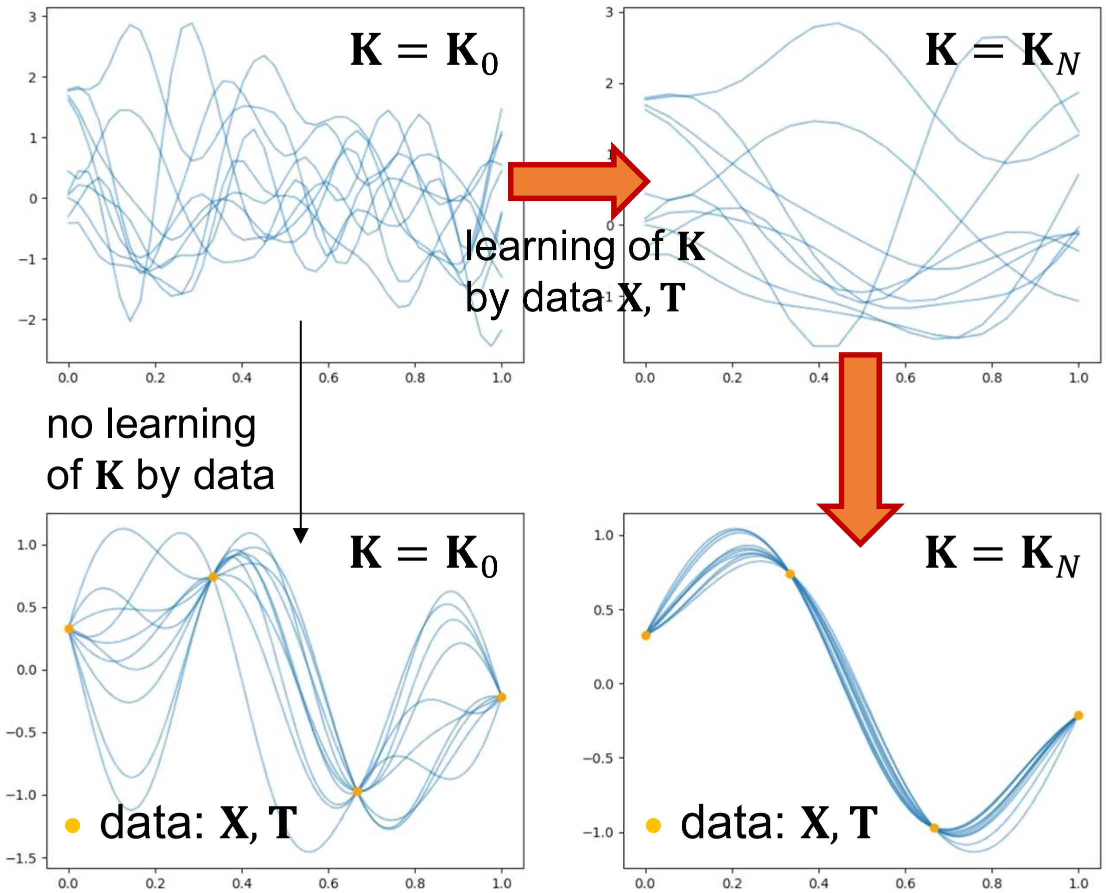
Gaussian Processes



Gaussian process

$$p(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{K}_0)$$

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



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

Lecture content

- Learning hyperparameters in kernel functions

Gaussian Processes

Update of Gram matrix

Learning the hyperparameters

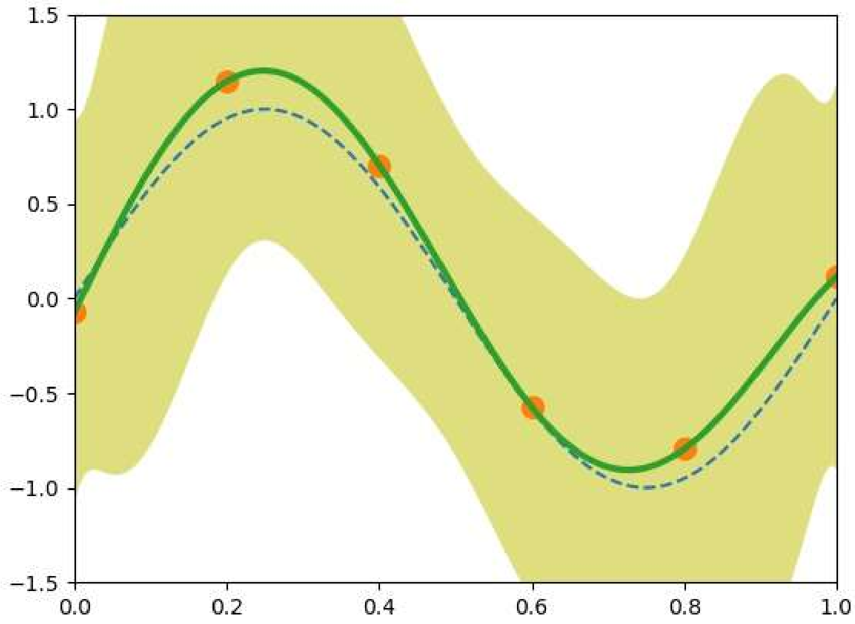
prior posterior

$$\mathbf{K}_0 \longrightarrow \mathbf{K}_N$$

$$\mathbf{K}(\hat{\boldsymbol{\theta}}_0) \longrightarrow \mathbf{K}(\hat{\boldsymbol{\theta}}_N)$$

MLE is normally used to determine $\hat{\boldsymbol{\theta}}_N$.

Why **MLE** in the Bayesian approach?



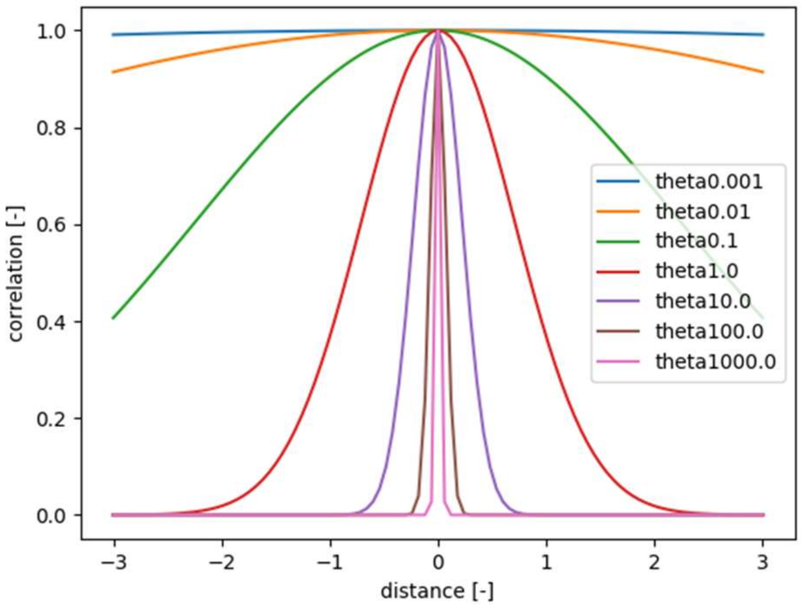
The deterministic $\hat{\boldsymbol{\theta}}$ controls the predictive distribution itself.

Note: No need to specify $\hat{\boldsymbol{\theta}}_0$ as we did not specify anything for σ in the curve fitting problem by MLE

Gaussian Processes

Gaussian kernel (see Lecture 7)

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right)$$



Example 1

$$k(\boldsymbol{x}, \boldsymbol{x}', \theta) = \exp(-\theta\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$$

Example 2

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

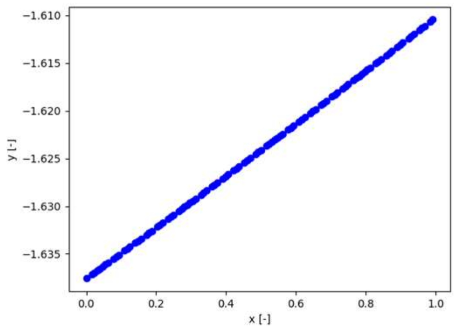
Note: Each component of the input $\boldsymbol{x}$ should be normalized (between 0 and 1).

Characteristics of the kernel parameterized by θ

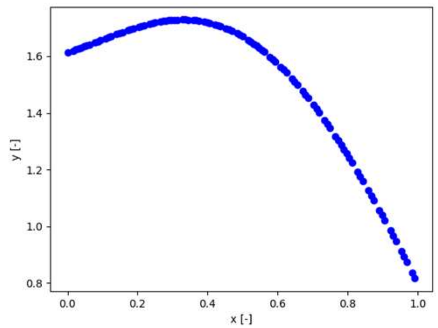
"One random sample" for each θ is generated by the probability below.

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

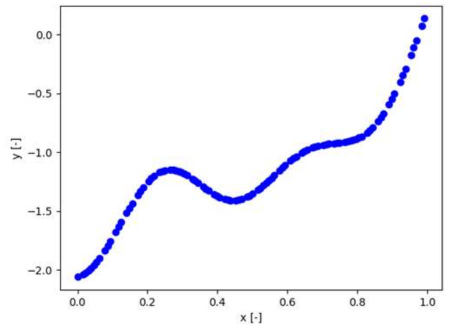
$$\mathbf{K}(\mathbf{X}, \mathbf{X}, \theta) = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1, \theta) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_N, \theta) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1, \theta) & \cdots & k(\boldsymbol{x}_N, \boldsymbol{x}_N, \theta) \end{pmatrix}$$



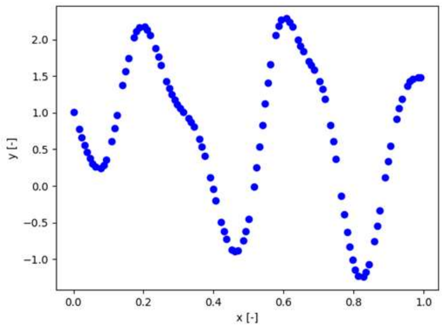
θ = 1e-3



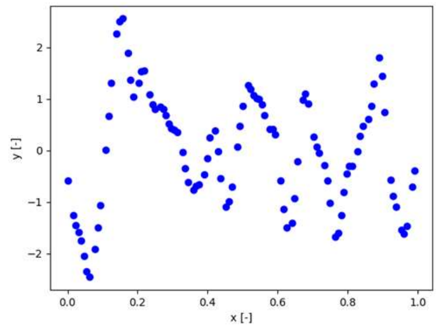
θ = 1e0



θ = 1e1



θ = 1e2



θ = 1e3

Gaussian Processes

Probabilistic model (originally a predictive distribution)

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

$\mathbf{x},\mathbf{t}$: prediction

$\mathbf{X},\mathbf{T}$: data

Likelihood function

$$p(\mathbf{T}|\mathbf{X},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{T}|\mathbf{0},\mathbf{K}(\mathbf{X},\mathbf{X},\boldsymbol{\theta})) \quad = -\frac{1}{2}\ln|\mathbf{K}| - \frac{1}{2}\mathbf{T}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{T} + \mathrm{C} \qquad (\mathbf{K} = \mathbf{K}(\mathbf{X},\mathbf{X},\boldsymbol{\theta}))$$

Optimization algorithm is required.

MLE

$$\boxed{\widehat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} p(\mathbf{T}|\mathbf{X},\boldsymbol{\theta})}$$

MLE to obtain $\widehat{\boldsymbol{\theta}}$

Predictive distribution

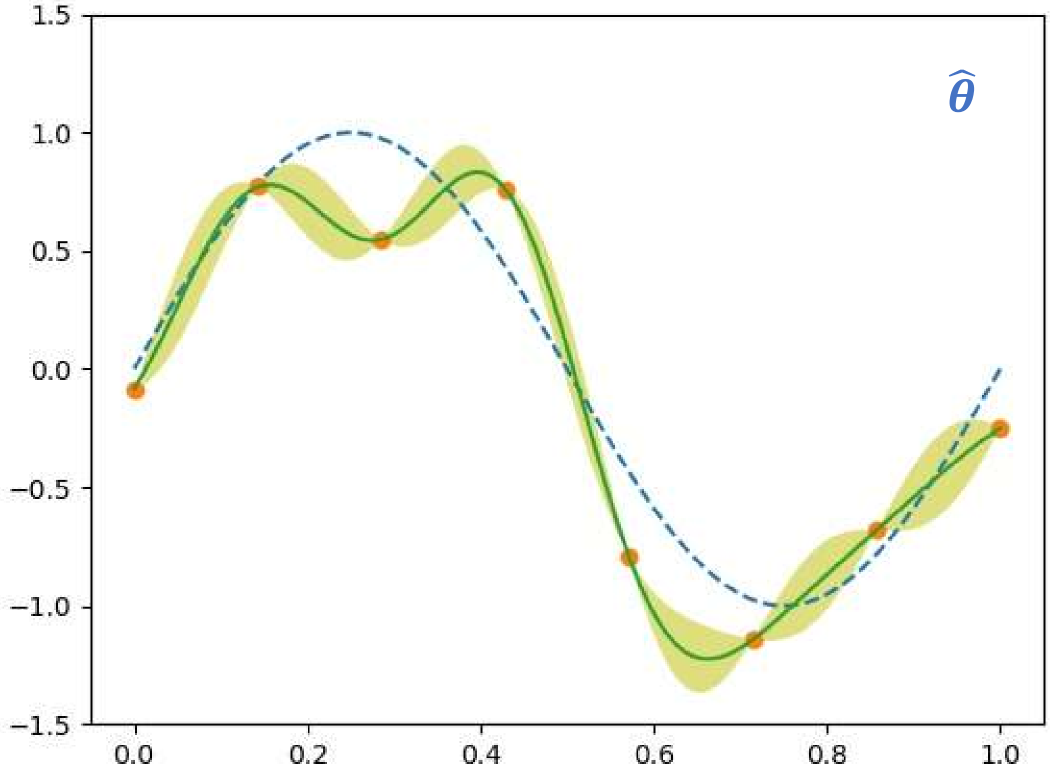
$$p(\mathbf{t}|\mathbf{x},\mathbf{X},\mathbf{T}) = \mathcal{N}\left(\mathbf{t}\middle|\boldsymbol{k}(\mathbf{x},\mathbf{X},\widehat{\boldsymbol{\theta}})^{\mathrm{T}}\mathbf{K}(\mathbf{X},\mathbf{X},\widehat{\boldsymbol{\theta}})^{-1}\mathbf{T}, \boldsymbol{k}(\mathbf{x},\mathbf{x},\widehat{\boldsymbol{\theta}}) - \boldsymbol{k}(\mathbf{x},\mathbf{X},\widehat{\boldsymbol{\theta}})^{\mathrm{T}}\mathbf{K}(\mathbf{X},\mathbf{X},\widehat{\boldsymbol{\theta}})^{-1}\boldsymbol{k}(\mathbf{x},\mathbf{X},\widehat{\boldsymbol{\theta}})\right)$$

Lecture content

- Examples (analogy with Bayesian linear regression)

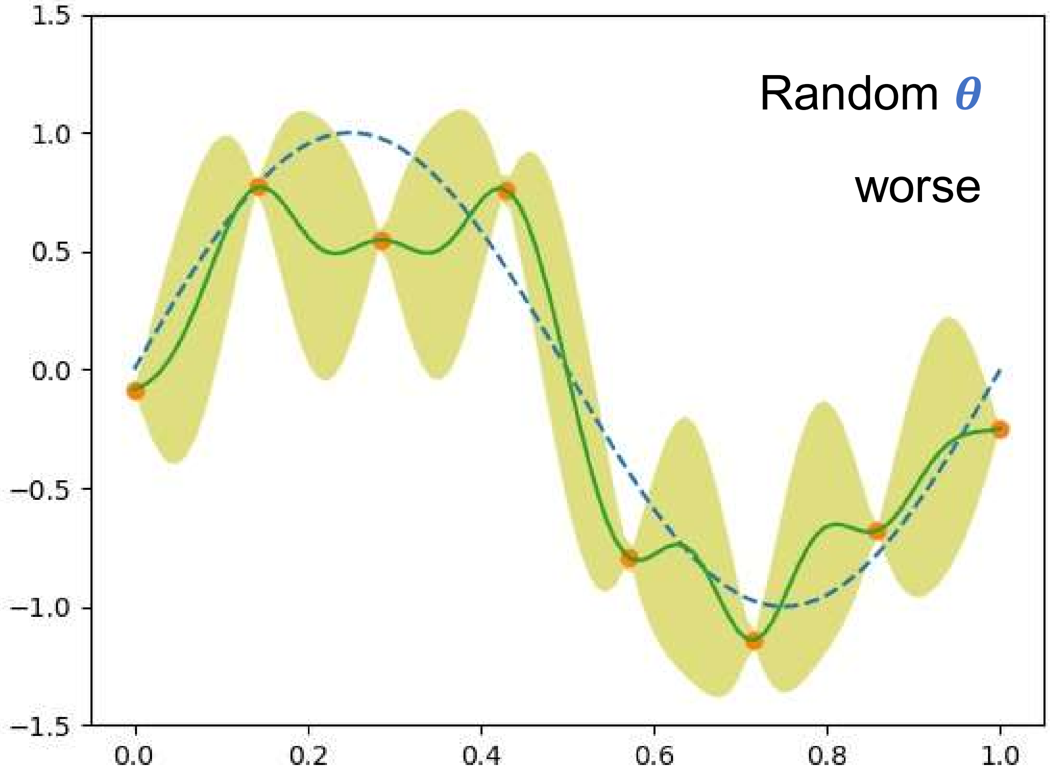
Gaussian Processes

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



Gaussian Processes

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



Gaussian Processes

Predictive distribution in general

$$p(\mathbf{t}) = \mathcal{N}\left(\mathbf{t}\middle|\boldsymbol{\Phi}\boldsymbol{m}, \quad \widehat{\boldsymbol{\Sigma}} \quad + \boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^{\mathrm{T}} \quad \right)$$

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

Prior of $\boldsymbol{w}$: $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, {\sigma_w}^2\mathbf{I})$

No Noise: $\sigma = 0$

$$\begin{aligned} p(\mathbf{t}) &= \mathcal{N}\left(\mathbf{t}\middle|\mathbf{0} \quad , \sigma^2\mathbf{I} + {\sigma_w}^2\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\right) \\ &= \mathcal{N}(\mathbf{t}|\mathbf{0} \quad , \sigma^2\mathbf{I} + \mathbf{K}(\boldsymbol{\theta}) \quad) \\ &= \mathcal{N}(\mathbf{t}|\mathbf{0} \quad , \quad \mathbf{K}(\boldsymbol{\theta}, \sigma) \ ) \end{aligned}$$

$$\begin{aligned} p(\mathbf{t}) &= \mathcal{N}\left(\mathbf{t}\middle|\mathbf{0} \quad , 0^2\mathbf{I} + {\sigma_w}^2\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\right) \\ &= \mathcal{N}(\mathbf{t}|\mathbf{0} \quad , \quad \mathbf{K}(\boldsymbol{\theta}) \quad) \end{aligned}$$

As far as σ is not treated as Bayesian, all can be Gaussians (conjugate prior – Lecture 6).

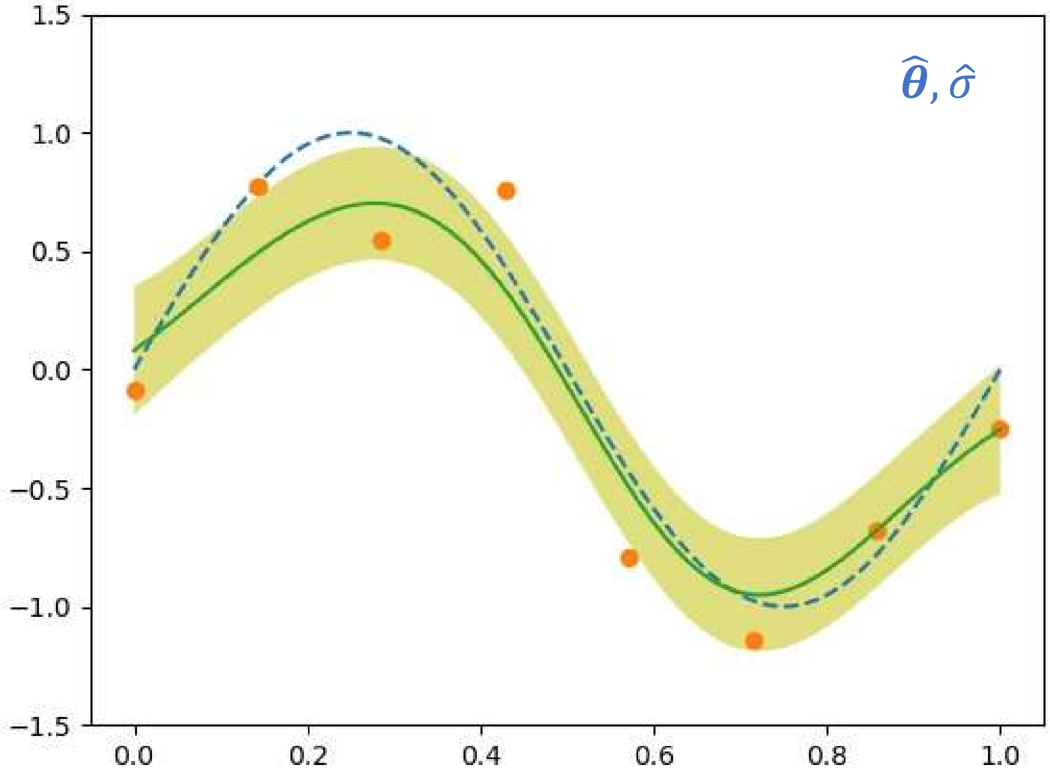
The situation when $\sigma = 0$ was already explained in the quiz in Lecture 6 slide 17.

$$\mathbf{K}(\boldsymbol{\theta}, \sigma) = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1, \boldsymbol{\theta}) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_N, \boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1, \boldsymbol{\theta}) & \cdots & k(\boldsymbol{x}_N, \boldsymbol{x}_N, \boldsymbol{\theta}) \end{pmatrix} + \sigma^2 \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \begin{cases} 1 \ if \ \boldsymbol{x} = \boldsymbol{x}' \\ 0 \ else \end{cases}$$

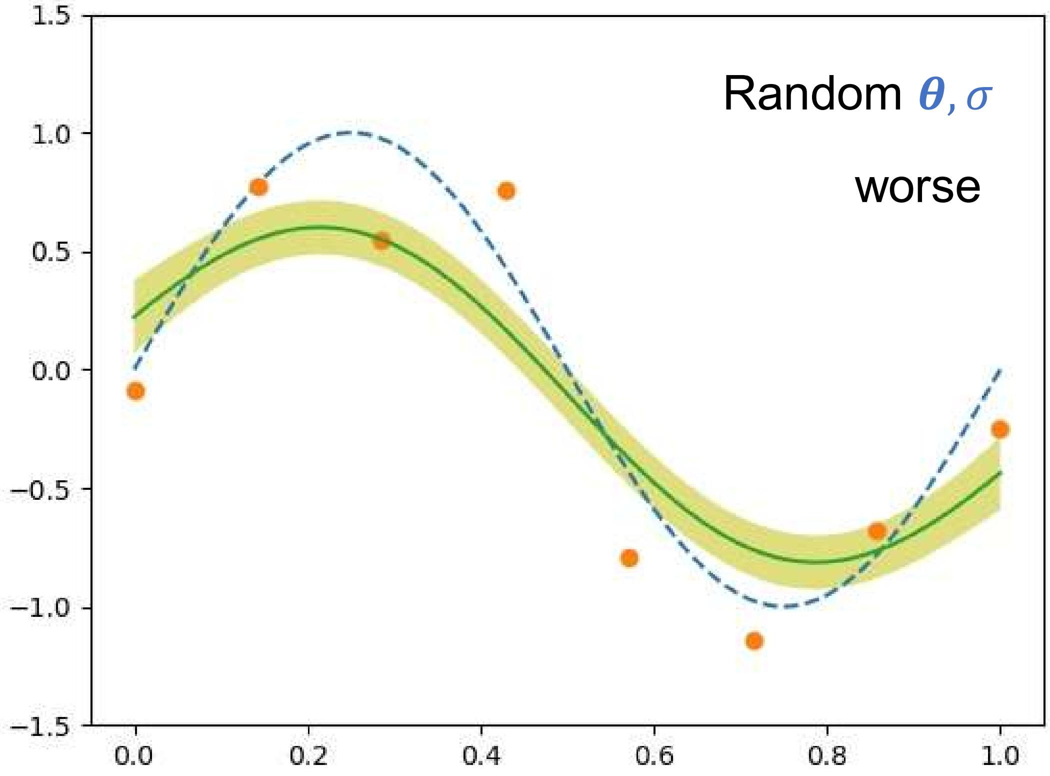
Gaussian Processes

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



Gaussian Processes

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

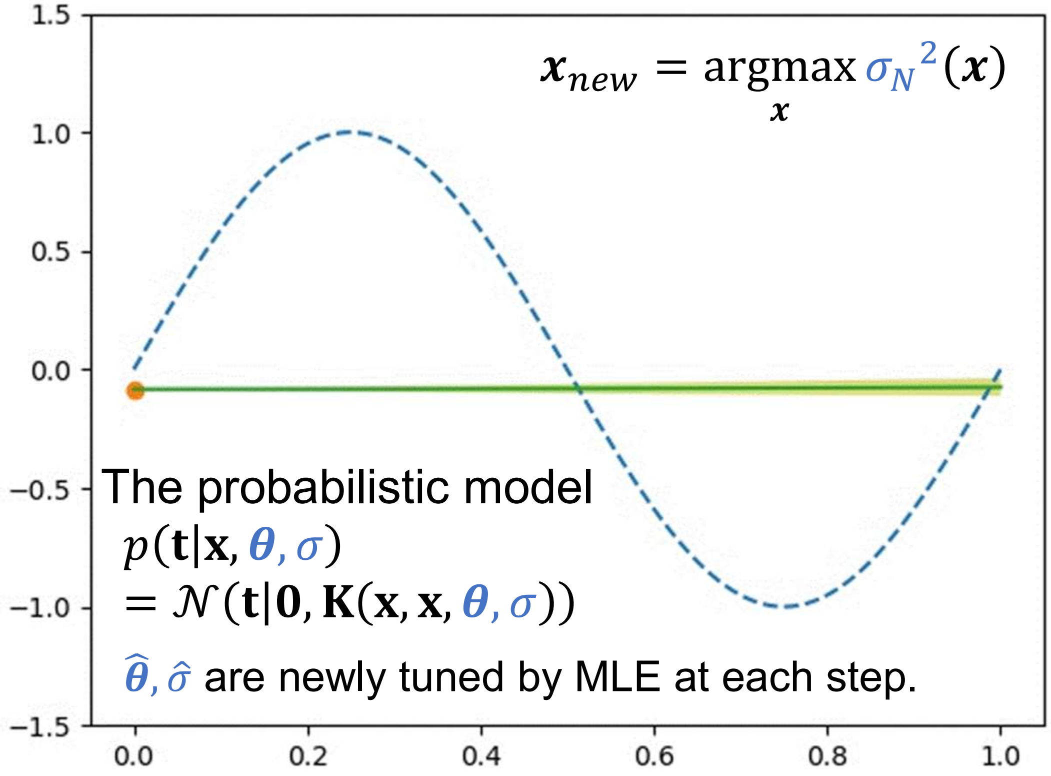


Gaussian Process + adaptive sampling

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

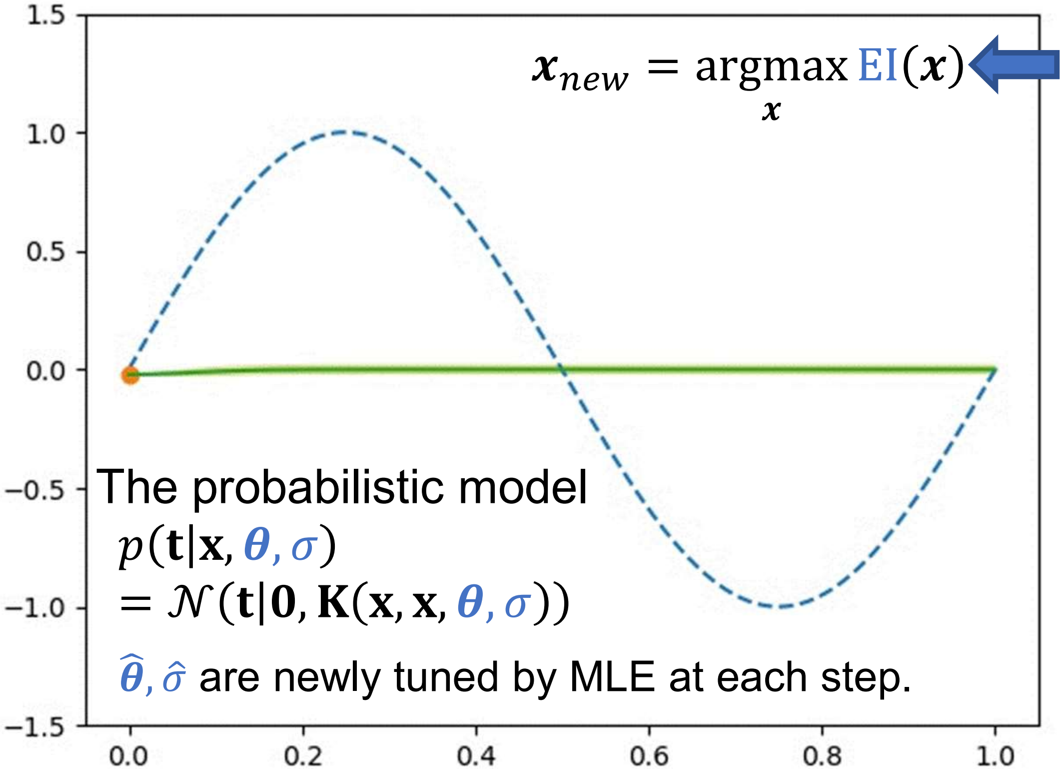


$$\sigma_N{}^2(\boldsymbol{x}) = \boldsymbol{k}\left(\boldsymbol{x}, \boldsymbol{x}, \widehat{\boldsymbol{\theta}}, \hat{\sigma}\right) - \boldsymbol{k}\left(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}, \hat{\sigma}\right)^{\mathrm{T}} \mathbf{K}\left(\widehat{\boldsymbol{\theta}}, \hat{\sigma}\right)^{-1}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

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**

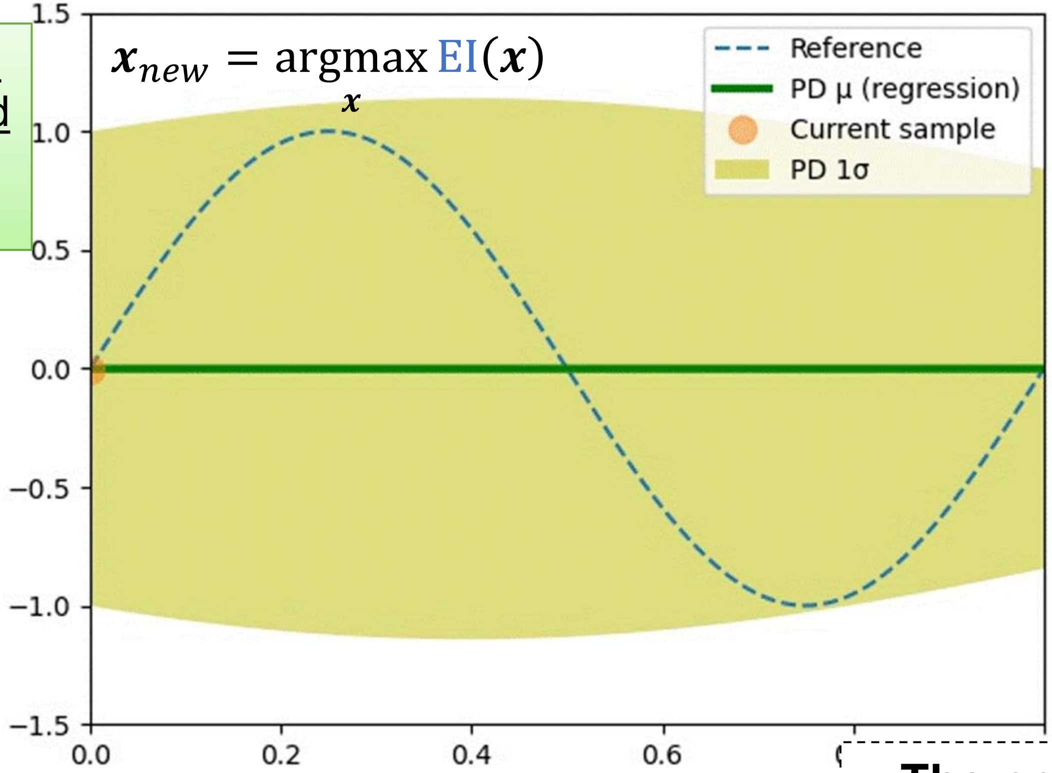
$$\mathrm{EI}(\boldsymbol{x}) = f(\mu(\boldsymbol{x}), {\sigma_N}^2(\boldsymbol{x}), current\ minimum\ sample\ point) \qquad \left(\mu(\boldsymbol{x}) = \boldsymbol{k}(\mathbf{x})^{\mathrm{T}} \mathbf{K}\left(\widehat{\boldsymbol{\theta}}, \hat{\sigma}\right)^{-1} \mathbf{T}\right)$$

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, the same concept can be used 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

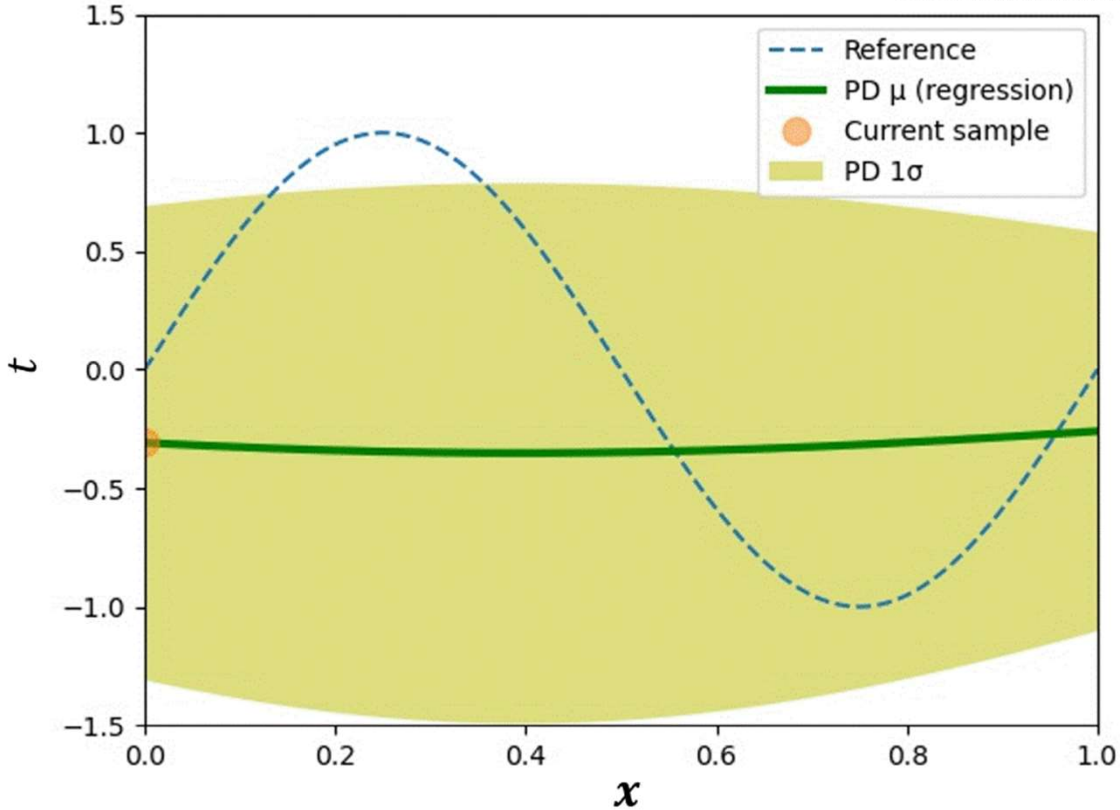
$$p(t|x,\boldsymbol{w}) = \mathcal{N}\left(t \middle| \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\boldsymbol{x})$ = 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 $N\ (sample\ size) = 1$

Brief Summary

Gaussian Processes

The probabilistic model is a multivariate Gaussian distribution.

1. Define a probabilistic model (with defining a kernel k):

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

2. Then, MLE (to find $\widehat{\boldsymbol{\theta}}$)

$\boldsymbol{\theta}$ is the representative of all the hyperparameters (e.g. $\boldsymbol{\theta}$ and σ)