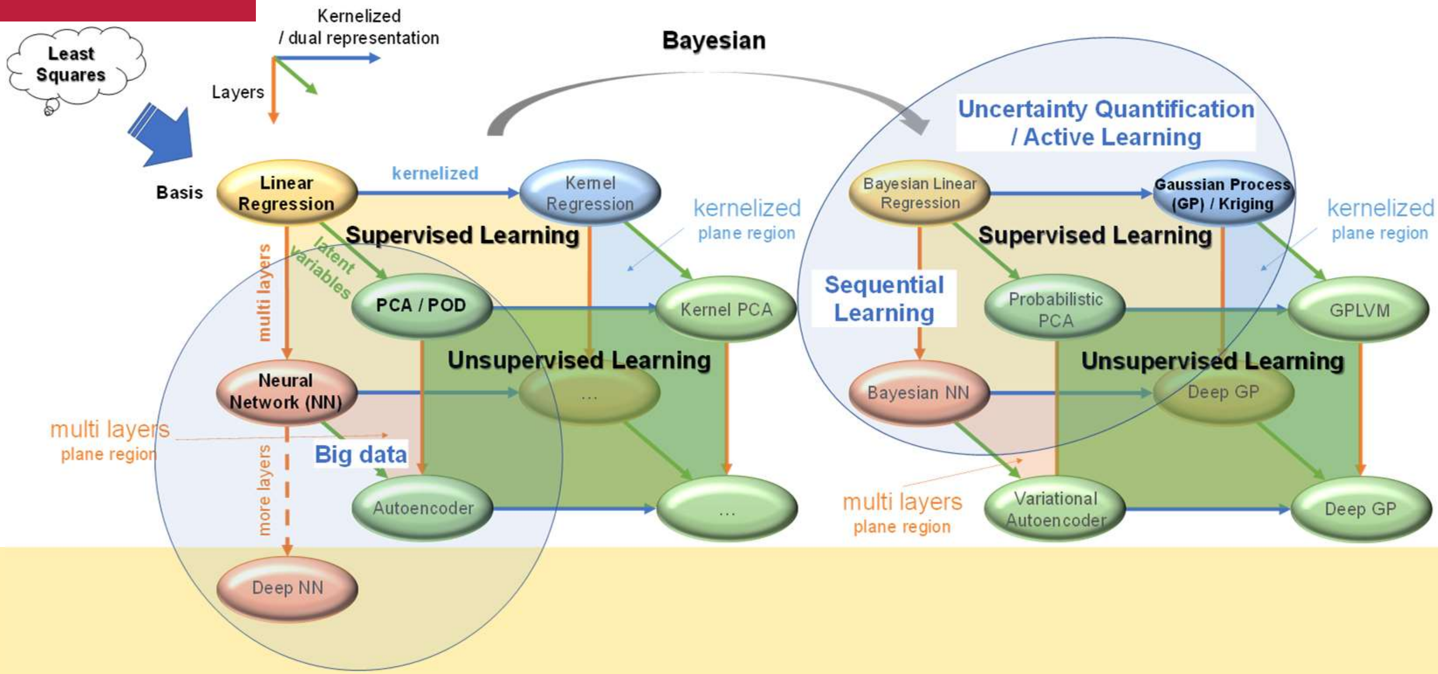
Technische
Universität
Braunschweig

IFL



Kernelized
/ dual representation
Layers
Least
Squares
Basis
Bayesian
Uncertainty Quantification
/ Active Learning
Linear
Regression
kernelized
Kernel
Regression
Supervised Learning
kernelized
plane region
latent
variables
multi layers
PCA / POD
Kernel PCA
Unsupervised Learning
Neural
Network (NN)
…
multi layers
plane region
Big data
more layers
Autoencoder
…
Deep NN
Bayesian Linear
Regression
Gaussian Process
(GP) / Kriging
Supervised Learning
kernelized
plane region
Sequential
Learning
Probabilistic
PCA
GPLVM
Unsupervised Learning
Bayesian NN
Deep GP
multi layers
plane region
Variational
Autoencoder
Deep GP

Scientific Machine Learning

Lecture 6: Bayesian Statistics (2/2) – Bayesian Linear Regression

Dr. Daigo Maruyama

Prof. Dr. Ali Elham

Where are we going now?

We are going to learn:

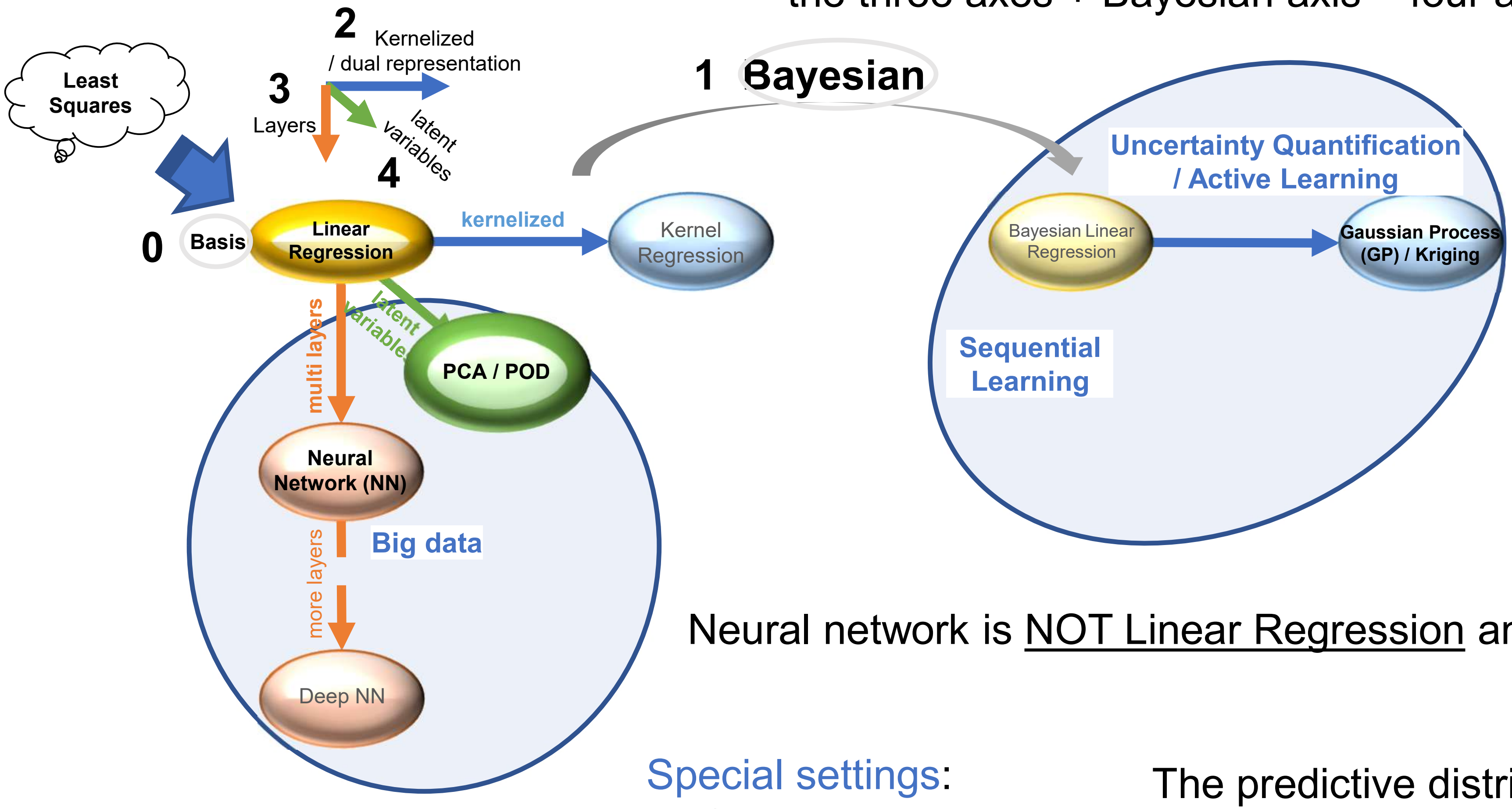
If one sentence is used to explain them:

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

by learning tools now.

Key Components

the three axes + Bayesian axis = four axes



Neural network is NOT Linear Regression anymore.

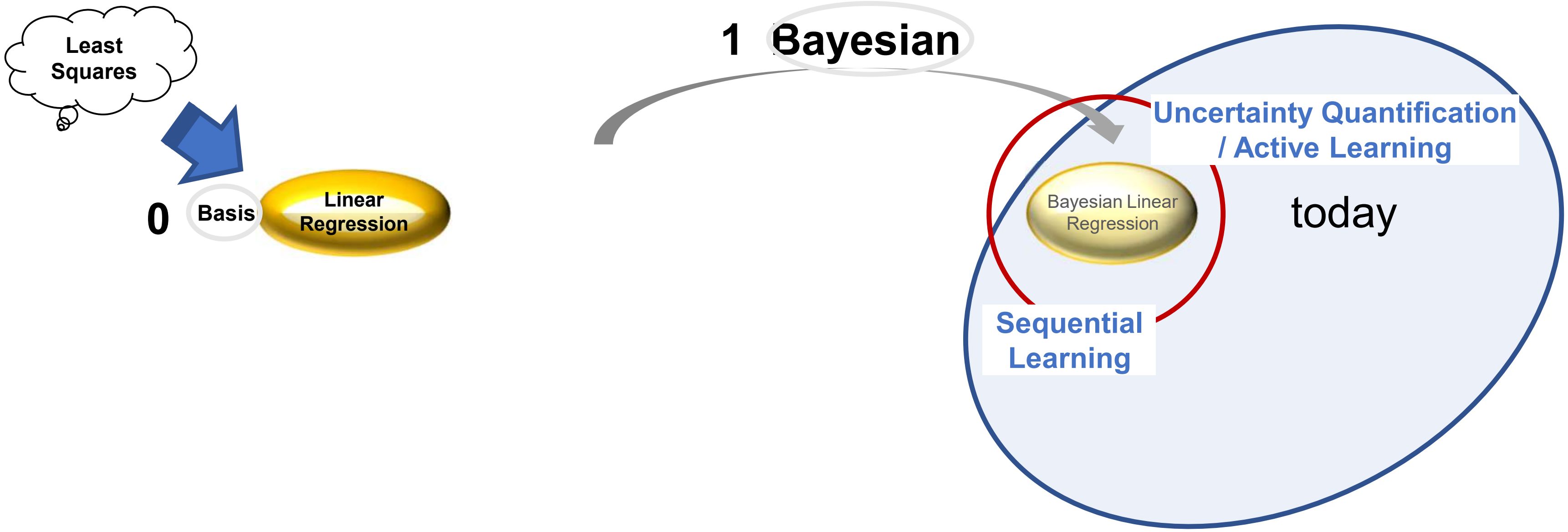
Special settings:

- Conjugate prior
- Linear regression

➡ The predictive distribution (the goal) can be analytically obtained.

This can be also a preparation of Gaussian Process.

Current Position



Special settings:
- Conjugate prior
- Linear regression

⇒ The predictive distribution (the goal) can be analytically obtained.

This can be also a preparation of Gaussian Process.

Lecture content

- Bayesian approach (review)
- Bayesian linear regression
  - Bayesian sequential learning
  - Uncertainty due to data (active learning)

The lecture of this time basically follows the Sections 3.3 and 2.3.6 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.

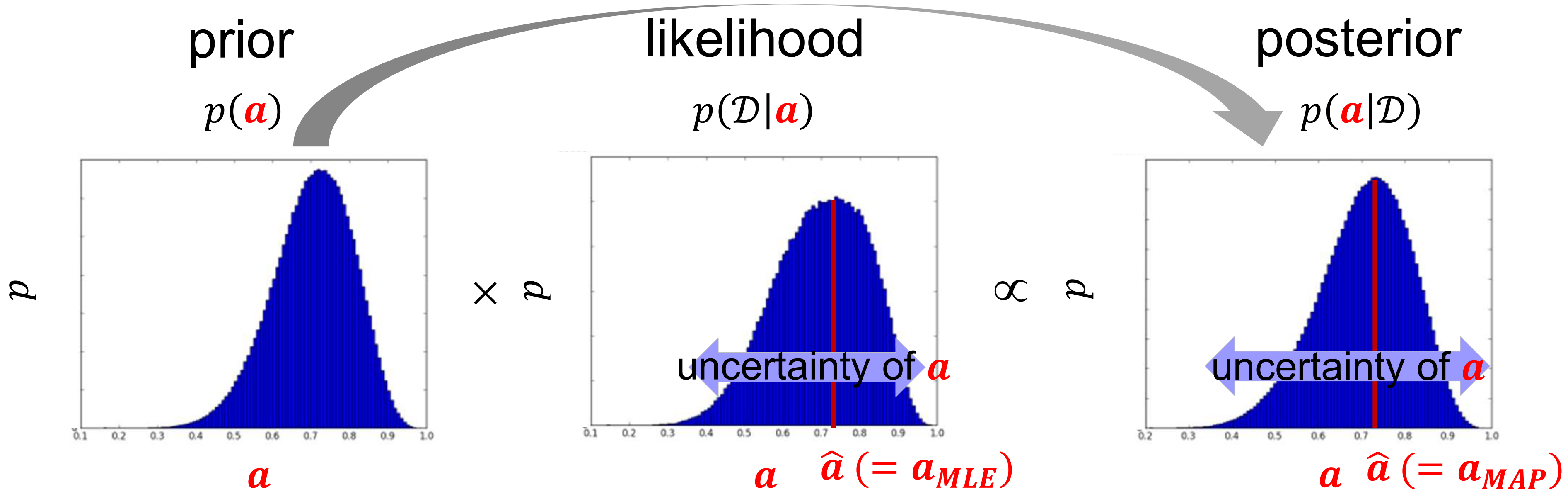
Lecture content

- Bayesian approach (review)

Bayesian Approach - Review

Generalized Process

1. Define a probabilistic model
2. Then, **compute the posterior**
    - (Define a prior distribution)
        - **Point estimate** (Deterministic)
        - **Probability distribution** (Stochastic) ← computations hard



prior $p(\boldsymbol{a})$ × likelihood $p(\mathcal{D}|\boldsymbol{a})$ ∝ posterior $p(\boldsymbol{a}|\mathcal{D})$

uncertainty of $\boldsymbol{a}$

uncertainty of $\boldsymbol{a}$

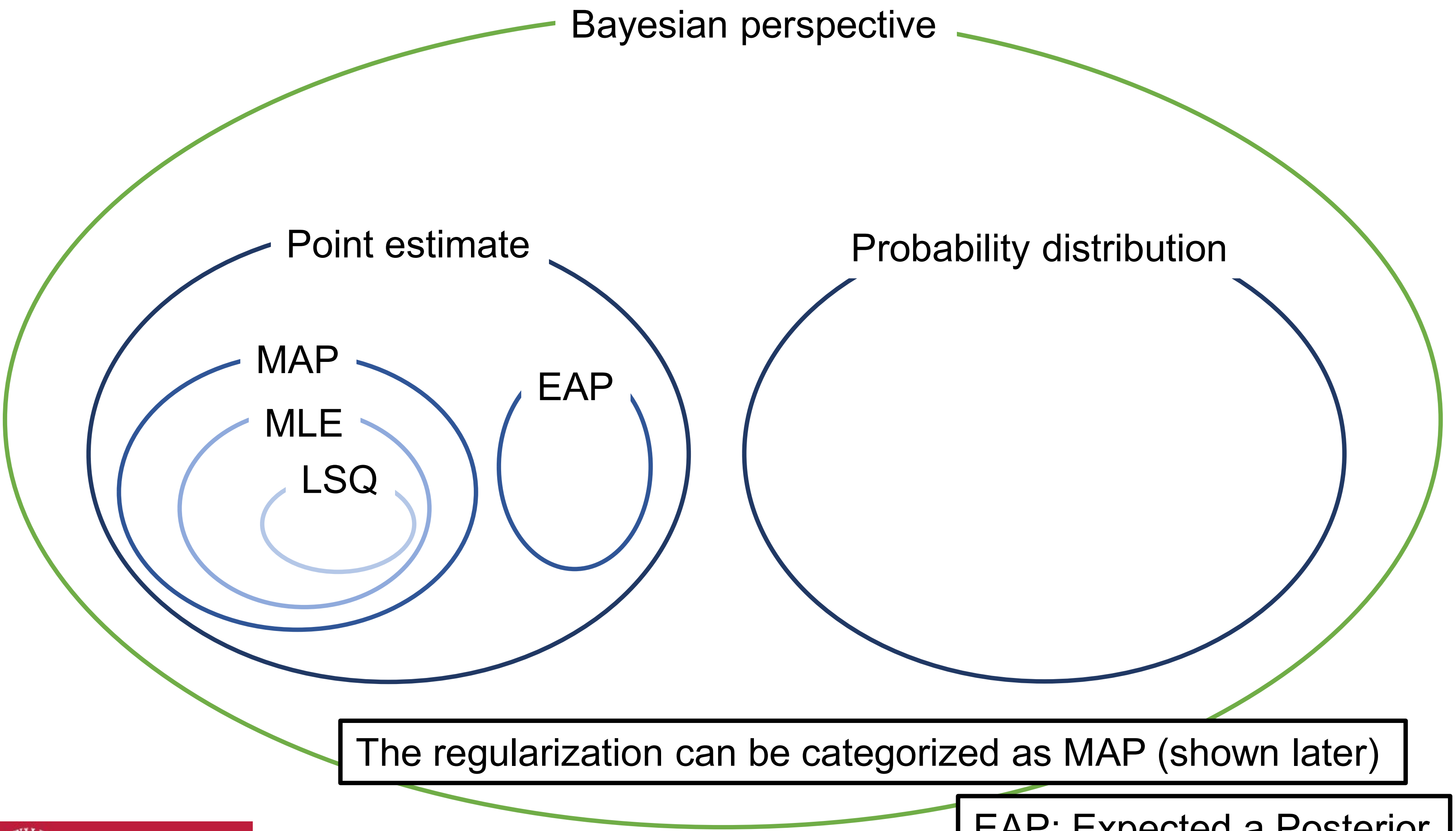
$\boldsymbol{a}$

$\boldsymbol{a}$ $\hat{\boldsymbol{a}}\ (= \boldsymbol{a}_{MLE})$

$\boldsymbol{a}$ $\hat{\boldsymbol{a}}\ (= \boldsymbol{a}_{MAP})$

$$prior \times likelihood \propto posterior$$

Bayesian Approach – Generalized Perspective



Bayesian perspective

Point estimate

MAP

MLE

LSQ

EAP

Probability distribution

The regularization can be categorized as MAP (shown later)

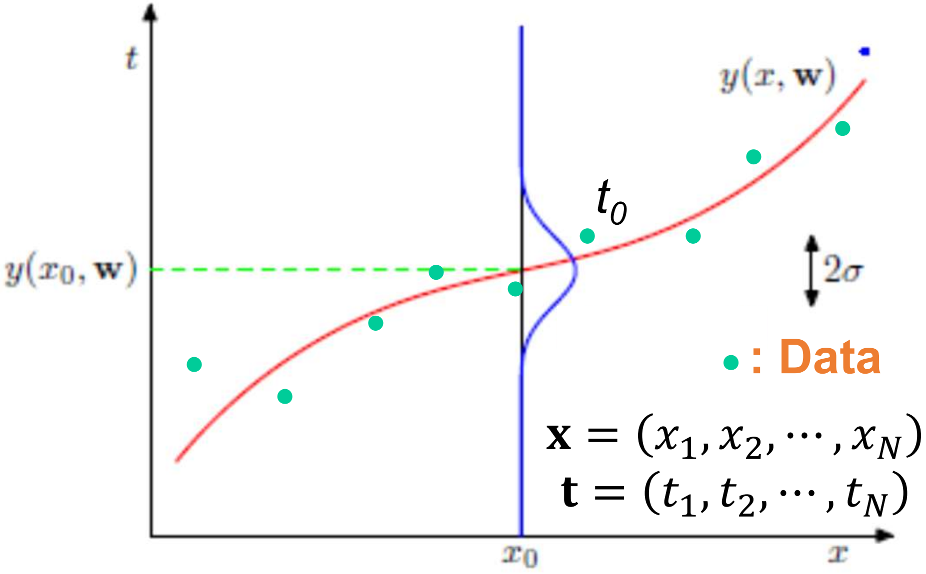
EAP: Expected a Posterior

Lecture content

- Bayesian linear regression

Curve Fitting Revisiting

The least square method and the regularization method are summarized in perspectives based in the **probability theory**.



● : **Data**

$$\mathbf{x} = (x_1, x_2, \cdots, x_N)$$
$$\mathbf{t} = (t_1, t_2, \cdots, t_N)$$

based on PRML, p. 29

Define **a Probabilistic model**

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

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

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

x: deterministic variable

t: random variable

$\boldsymbol{w}, \sigma$: random variable

e.g. $y(x, \boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x})$ is a polynomial function.

Curve Fitting Revisiting (MLE)

Likelihood

$$p(\mathbf{t}|\mathbf{x},\boldsymbol{w},\sigma)=\prod_{i=1}^{N}\mathcal{N}(t_i|y(x_i,\boldsymbol{w}),\sigma^2)$$

The likelihood function is uniquely determined by the data $\mathbf{x},\mathbf{t}$.

MLE $\quad \hat{\boldsymbol{w}},\hat{\sigma}=\underset{\boldsymbol{w},\sigma}{\operatorname{argmax}}\,p(\mathbf{t}|\mathbf{x},\boldsymbol{w},\sigma)$

$$\hat{\boldsymbol{w}}=\underset{\boldsymbol{w}}{\operatorname{argmin}}\sum_{i=1}^{N}\{t_i-y(x_i,\boldsymbol{w})\}^2 \qquad =\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$

$$\hat{\sigma}^2=\frac{1}{N}\sum_{i=1}^{N}\{t_i-y(x_i,\hat{\boldsymbol{w}})\}^2$$

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

(If **Linear regression**)

x: deterministic variable
t: random variable
$\boldsymbol{w},\sigma$: **deterministic variable**

$$\left(-\ln p(\mathbf{t}|\mathbf{x},\boldsymbol{w},\sigma)=\frac{1}{2\sigma^2}\sum_{i=1}^{N}\{t_i-y(x_i,\boldsymbol{w})\}^2+\frac{N}{2}\ln(2\pi\sigma^2)\right)$$

Curve Fitting Revisiting (Bayesian Perspective)

Prior

$$p(\boldsymbol{w},\sigma)$$

You can set it.

Likelihood

$$p(\mathbf{t}|\mathbf{x},\boldsymbol{w},\sigma)=\prod_{i=1}^{N}\mathcal{N}(t_i|y(x_i,\boldsymbol{w}),\sigma^2)$$

The likelihood function is uniquely determined by the data $\mathbf{x},\mathbf{t}$.

Posterior

$$p(\boldsymbol{w},\sigma|\mathbf{x},\mathbf{t})$$

by the Bayes' theorem

$$p(\boldsymbol{w},\sigma)p(\mathbf{t}|\mathbf{x},\boldsymbol{w},\sigma)\propto p(\boldsymbol{w},\sigma|\mathbf{x},\mathbf{t})$$

If the prior $p(\boldsymbol{w},\sigma)$ is a uniform distribution, $p(\mathbf{t}|\mathbf{x},\boldsymbol{w},\sigma)\propto p(\boldsymbol{w},\sigma|\mathbf{x},\mathbf{t})$

x: deterministic variable ← Please do not consider x

t: random variable

$\boldsymbol{w},\sigma$: **random variable**

$$\boldsymbol{a}\equiv\boldsymbol{w},\sigma$$

$$p(\boldsymbol{a})p(\mathbf{t}|\boldsymbol{a})\propto p(\boldsymbol{a}|\mathbf{t})$$

Bayesian Approach

Let's try to apply this concept to **the curve fitting problem**.

Probabilistic model

$$p(t|x,\boldsymbol{w},\sigma) = \mathcal{N}(t|y(x,\boldsymbol{w}),\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\{t-y(x,\boldsymbol{w})\}^2}{2\sigma^2}\right\}$$

Likelihood function

$$p(\mathbf{t}|\mathbf{x},\boldsymbol{w},\sigma) = \prod_{i=1}^{N}\mathcal{N}(t_i|y(x_i,\boldsymbol{w}),\sigma^2) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}exp\left[-\frac{\{t_i-y(x_i,\boldsymbol{w})\}^2}{2\sigma^2}\right]$$

➡ **Posterior distribution** $p(\boldsymbol{w},\sigma|\mathbf{x},\mathbf{t}) = \boldsymbol{complicated}$

This is a Gaussian distribution wrt t_i,
but NOT a Gaussian distribution wrt $\boldsymbol{w},\sigma$.

Predictive distribution

$$p(t|x,\mathbf{x},\mathbf{t}) = \iint p(t|x,\boldsymbol{w},\sigma)p(\boldsymbol{w},\sigma|\mathbf{x},\mathbf{t})\mathrm{d}\boldsymbol{w}\mathrm{d}\sigma = \boldsymbol{complicated}$$

probabilistic model × posterior

Bayesian Approach

There are so many difficulties in the Bayesian approach in general.

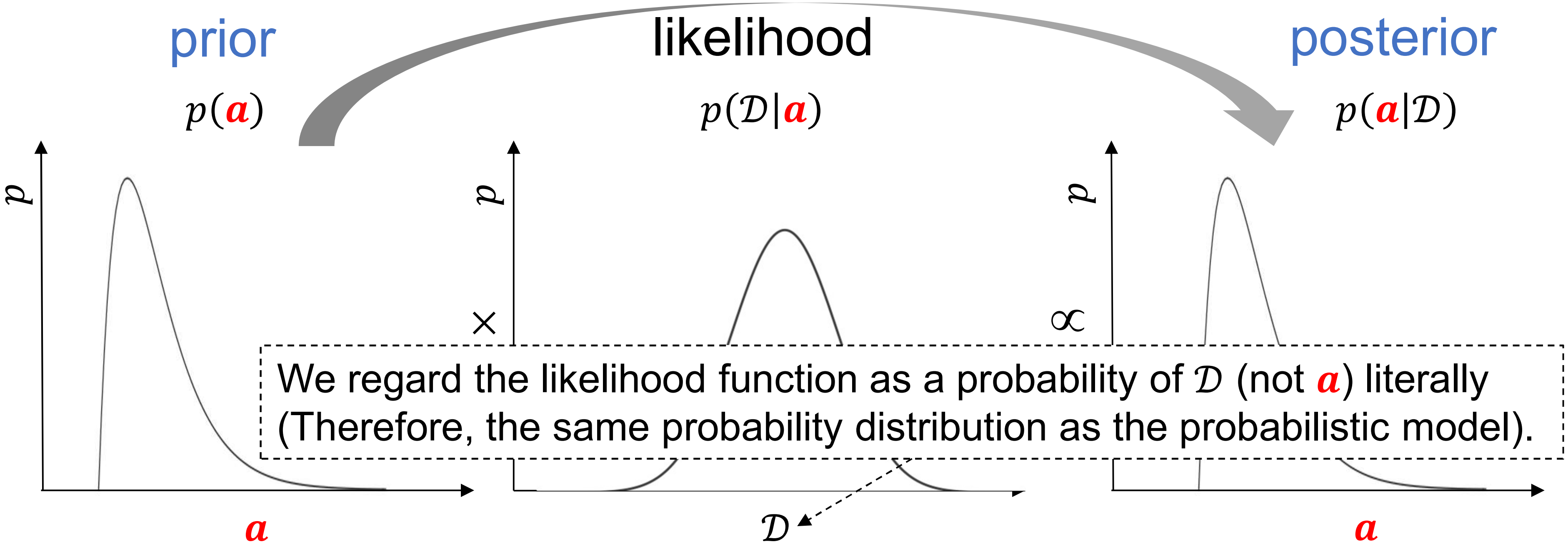
Difficulty 1: Computing the posterior distribution (the point estimate is fine)
Difficulty 2: Computing the predictive distribution (the goal)

Computations are hard.

It is rather rare if we can obtain analytical solutions of them.

Conjugate Prior

Special cases



We regard the likelihood function as a probability of $\mathcal{D}$ (not $\boldsymbol{a}$) literally (Therefore, the same probability distribution as the probabilistic model).

$$\textit{prior} \times \textit{likelihood} \propto \textit{posterior}$$

The posterior becomes the same type of probability distribution of **the prior**.

Conjugate Prior

Likelihood Function	Unknown Parameters	Conjugate Prior	Predictive Distribution
Bernoulli	μ	Beta	Bernoulli
Binomial	μ	Beta	Beta-Binomial
Categorical	$\boldsymbol{\mu}$	Dirichlet	Categorical
Multinomial	$\boldsymbol{\mu}$	Dirichlet	Dirichlet-Multinomial
Poisson	λ	Gamma	Negative Binomial
Gaussian	μ (σ^2 known)	Gaussian	Gaussian
Gaussian	σ^2 (μ known)	Gamma	Student-t
Gaussian	μ, σ^2	Gauss-Gamma	Student-t
Multivariate Gaussian	$\boldsymbol{\mu}$ ($\boldsymbol{\Sigma}$ known)	Multivariate Gaussian	Multivariate Gaussian
Multivariate Gaussian	$\boldsymbol{\Sigma}$ ($\boldsymbol{\mu}$ known)	Wishart	Multivariate Student-t
Multivariate Gaussian	$\boldsymbol{\mu}, \boldsymbol{\Sigma}$	Gaussian-Wishart	Multivariate Student-t

Conjugate Prior

Likelihood Function	Unknown Parameters	Conjugate Prior	Predictive Distribution
Bernoulli			Bernoulli
Binomial			-Binomial
Categorical			gorical
Multinomial			t-Multinomial
Poisson			egative Binomial
Gaussian		**aussian**	**Gaussian**
Gaussia		Gamma	Student-t
G	μ, σ^2	Gauss-Gamma	Student-t
Multivaria	$\boldsymbol{\mu}$ ($\boldsymbol{\Sigma}$ known)	Multivariate Gaussian	Multivariate Gaussian
Multivariate Gaussian	$\boldsymbol{\Sigma}$ ($\boldsymbol{\mu}$ known)	Wishart	Multivariate Student-t
Multivariate Gaussian	$\boldsymbol{\mu}, \boldsymbol{\Sigma}$	Gaussian-Wishart	Multivariate Student-t

When we can have cases like “σ^2 known”?

Please think about such cases by yours

The answers were alrea
points in the pre

No need to remember!

Bayesian Linear Regression

PRML, p.152-158

There is an important thing which is not emphasized in the book:

- σ is known and constant.

P. 152: For the moment, we shall treat the noise precision parameter β as a known constant.

✓ We have been doing MLE for μ, σ both of which are unknown.

➡ $\hat{\mu}, \hat{\sigma}$ obtained

✓ We will extend this to the Bayesian approach but only for μ as unknown, σ is known.

A possible idea: e.g. $\hat{\sigma}$ is used.

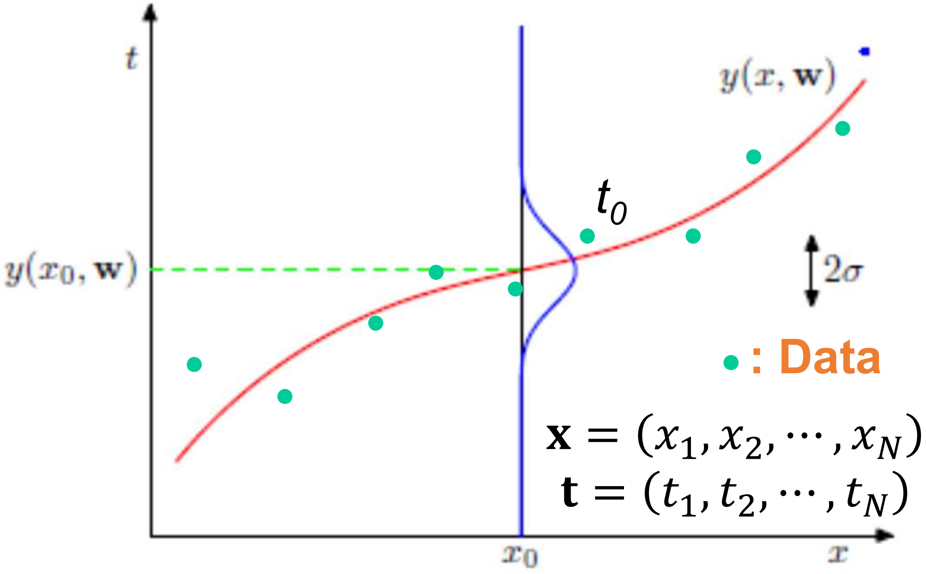
➡ Then, only μ is considered as probability as $p(\mu)$.

Conjugate Prior

Likelihood Function	Unknown Parameters	Conjugate Prior	Predictive Distribution
Bernoulli	μ	Beta	Bernoulli
Binomial	μ	Beta	Beta-Binomial
Categorical	$\boldsymbol{\mu}$	Dirichlet	Categorical
Multinomial	$\boldsymbol{\mu}$	Dirichlet	Dirichlet-Multinomial
Poisson	λ	Gamma	Negative Binomial
Gaussian	μ (σ^2 known)	**Gaussian**	**Gaussian**
Gaussian	σ^2 (μ known)	Gamma	Student-t
Gaussian	μ, σ^2	Gauss-Gamma	Student-t
Multivariate Gaussian	$\boldsymbol{\mu}$ ($\boldsymbol{\Sigma}$ known)	Multivariate Gaussian	Multivariate Gaussian
Multivariate Gaussian	$\boldsymbol{\Sigma}$ ($\boldsymbol{\mu}$ known)	Wishart	Multivariate Student-t
Multivariate Gaussian	$\boldsymbol{\mu}, \boldsymbol{\Sigma}$	Gaussian-Wishart	Multivariate Student-t

Bayesian Linear Regression

The least square method and the regularization method are summarized in perspectives based in the **probability theory**.



$$\mathbf{x} = (x_1, x_2, \cdots, x_N)$$
$$\mathbf{t} = (t_1, t_2, \cdots, t_N)$$

based on PRML, p. 29

$x, \hat{\sigma}$: deterministic variable
t: random variable
$\boldsymbol{w}$: <u>random variable</u>

Define **a Probabilistic model**

$$p(t|x, \mu) = \mathcal{N}(t|\mu(x), \hat{\sigma}^2)$$

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

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

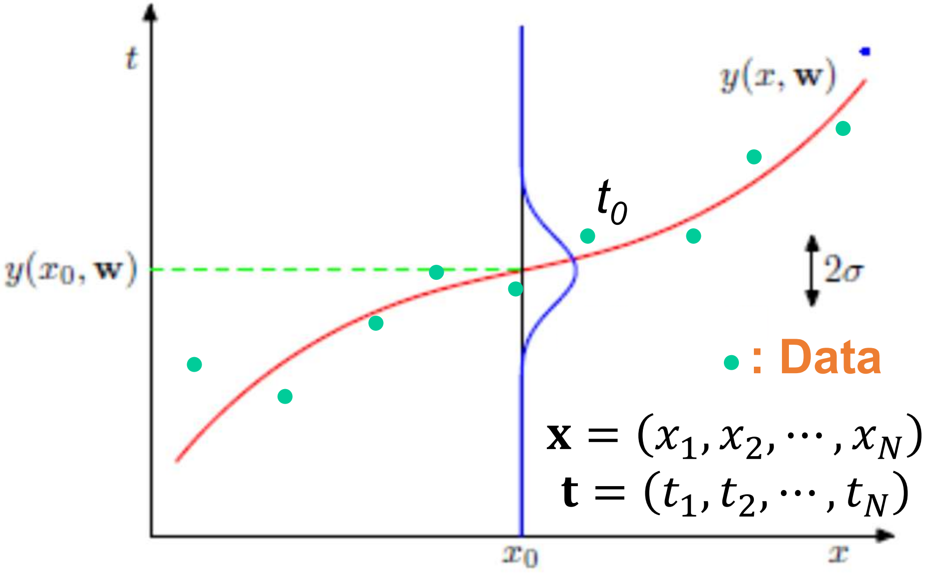
Likelihood can be then determined.

$$p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}) = \prod_{i=1}^{N} \mathcal{N}(t_i|y(x_i, \boldsymbol{w}), \hat{\sigma}^2)$$

<u>Gaussian distributions</u> wrt t_i

Bayesian Linear Regression

The least square method and the regularization method are summarized in perspectives based in the **probability theory**.



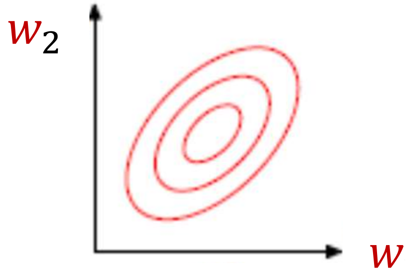
●: **Data**

$$\mathbf{x} = (x_1, x_2, \cdots, x_N)$$
$$\mathbf{t} = (t_1, t_2, \cdots, t_N)$$

based on PRML, p. 29

$x, \hat{\sigma}$: deterministic variable
t: random variable
$\boldsymbol{w}$: random variable

Define **a prior**



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

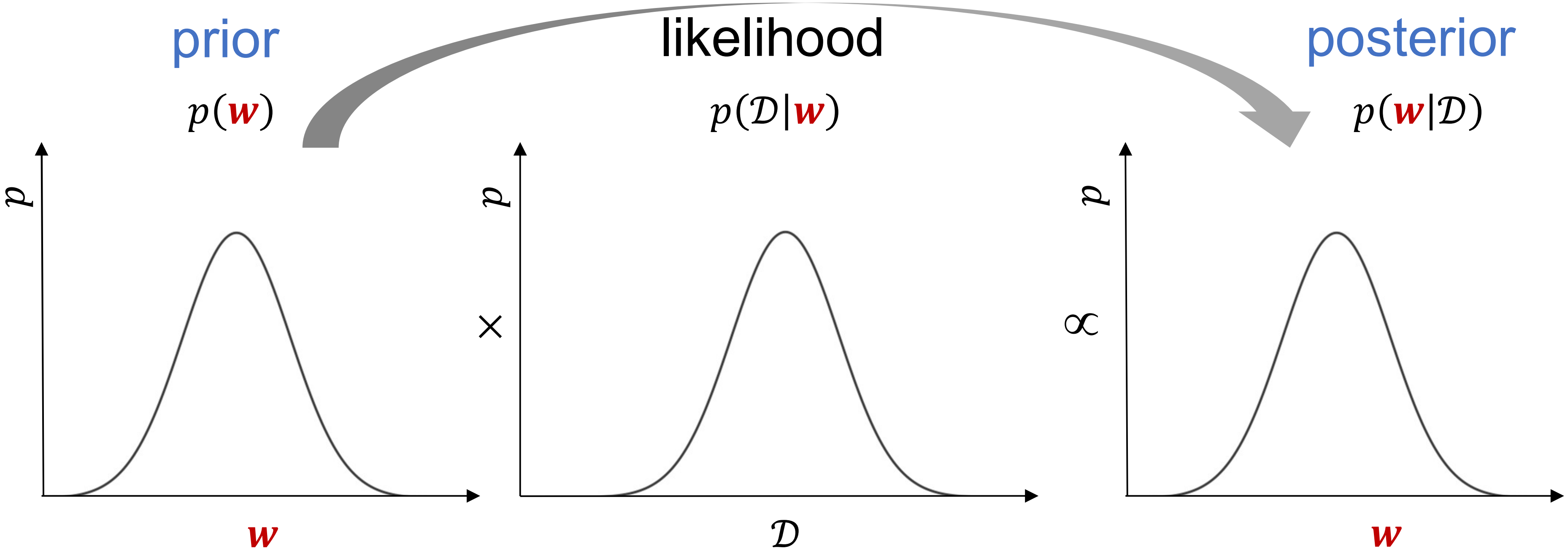
your setting

A Gaussian distribution in general

a posterior using the likelihood

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

Bayesian Linear Regression



prior

$p(\boldsymbol{w})$

likelihood

$p(\mathcal{D}|\boldsymbol{w})$

posterior

$p(\boldsymbol{w}|\mathcal{D})$

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

$$\boldsymbol{m}_0 = \mathbf{0}$$

$$\mathbf{S}_0 = {\sigma_0}^2\mathbf{I}$$

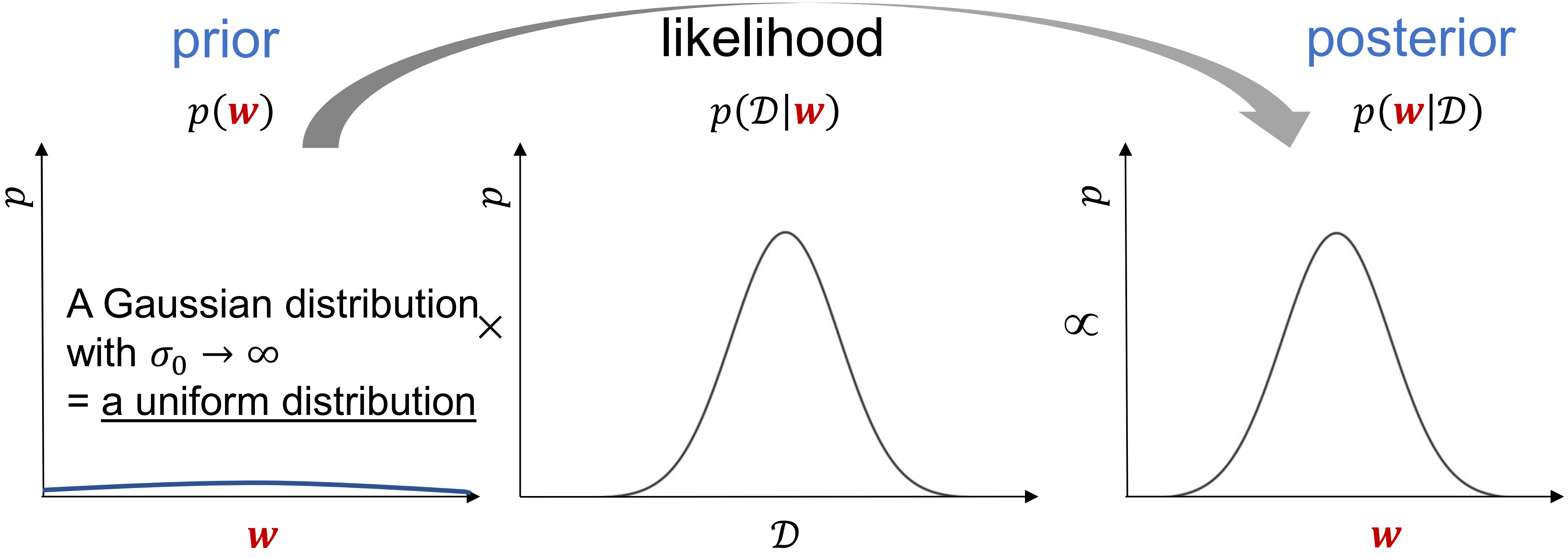
$$p(\boldsymbol{w}|\mathcal{D}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \mathbf{S}_N)$$

Conjugate prior
+ Linear regression

⇒

Analytical solutions of
- Posterior distribution
- Predictive distribution

Bayesian Linear Regression



prior

$p(\boldsymbol{w})$

likelihood

$p(\mathcal{D}|\boldsymbol{w})$

posterior

$p(\boldsymbol{w}|\mathcal{D})$

A Gaussian distribution
with $\sigma_0 \to \infty$
= a uniform distribution

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

$$\boldsymbol{m}_0 = \mathbf{0}$$

$$\mathbf{S}_0 = {\sigma_0}^2\mathbf{I}$$

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

Conjugate prior
+ Linear regression

Analytical solutions of

- Posterior distribution
- Predictive distribution

Bayesian Linear Regression

Details of the analytical solutions:
See PRML, p.152

Prior (your setting)

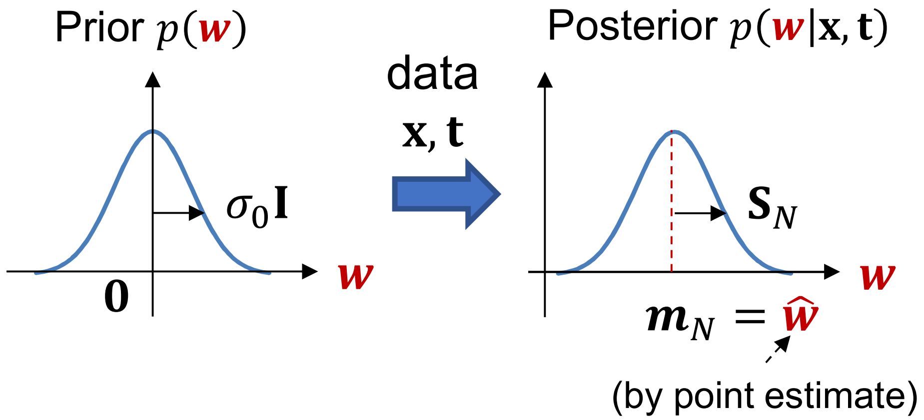
$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, {\sigma_0}^2\mathbf{I})$$

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**



By Bayes' theorem

$$p(\boldsymbol{w})p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}) \propto p(\boldsymbol{w}|\mathbf{x}, \mathbf{t})$$

where, $p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}) = \prod_{i=1}^{N} \mathcal{N}(t_i|y(x_i, \boldsymbol{w}), \hat{\sigma}^2)$

Next,

By the way,
How does the prior work in this setting?

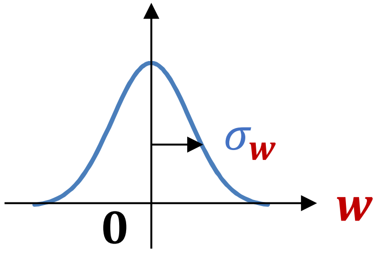
Relationship with
regularization

Curve Fitting Revisiting (Review from Lecture 3)

e.g. $p(\mathbf{w})$: a Gaussian distribution around $\mathbf{0}$

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

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



Prior distribution of $\mathbf{w}$

Bayes' theorem $p(\mathbf{w}|\mathcal{D}) \propto p(\mathcal{D}|\mathbf{w})p(\mathbf{w})$

$$-\ln p(\mathbf{w}|\mathcal{D}) \propto -\ln p(\mathcal{D}|\mathbf{w}) - \ln p(\mathbf{w})$$

$$= \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{N}\{t_i - y(x_i, \mathbf{w})\}^2 + \frac{N}{2}\ln(2\pi\hat{\sigma}^2) + \frac{1}{2{\sigma_{\mathbf{w}}}^2}\|\mathbf{w}\|^2$$

$$-\ln p(\mathbf{w}|\mathcal{D}) \propto \sum_{i=1}^{N}\{t_i - y(x_i, \mathbf{w})\}^2 + \left(\frac{\hat{\sigma}}{\sigma_{\mathbf{w}}}\right)^2\|\mathbf{w}\|^2 = \underbrace{E(\mathbf{w}) + \lambda\|\mathbf{w}\|^2}$$

where,

$$\lambda \equiv \left(\frac{\hat{\sigma}}{\sigma_{\mathbf{w}}}\right)^2$$

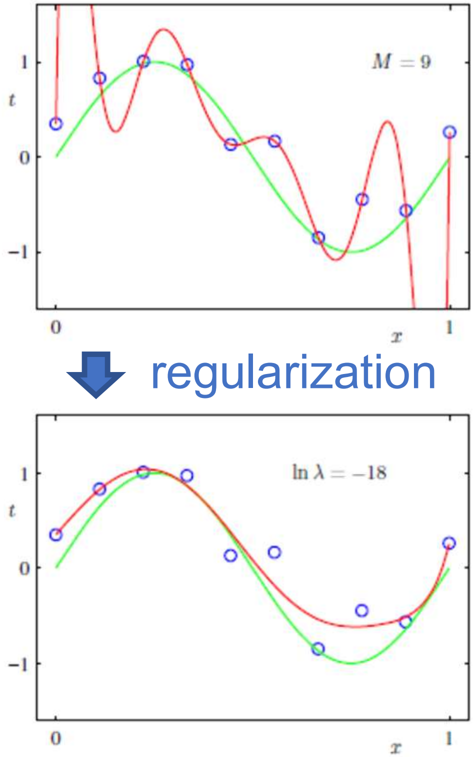
The same function as that for the regularization

Linear Regression (Review from Lecture 4)

If the regression $y(x_i, \boldsymbol{w})$ is a linear regression form $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x})$:

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

$$\widehat{\boldsymbol{w}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda\mathbf{I}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$



The regularization was (first introduced in Lecture 2):
The least square method with penalty term of $\boldsymbol{w}$

The regularization is now:
The point estimate with a prior distribution of $\boldsymbol{w}$

= Maximizing a posterior (MAP)

The prediction process is not decided only by the given data, but also with prior information.

Linear Regression (Review from Lecture 4)

Other regularization techniques

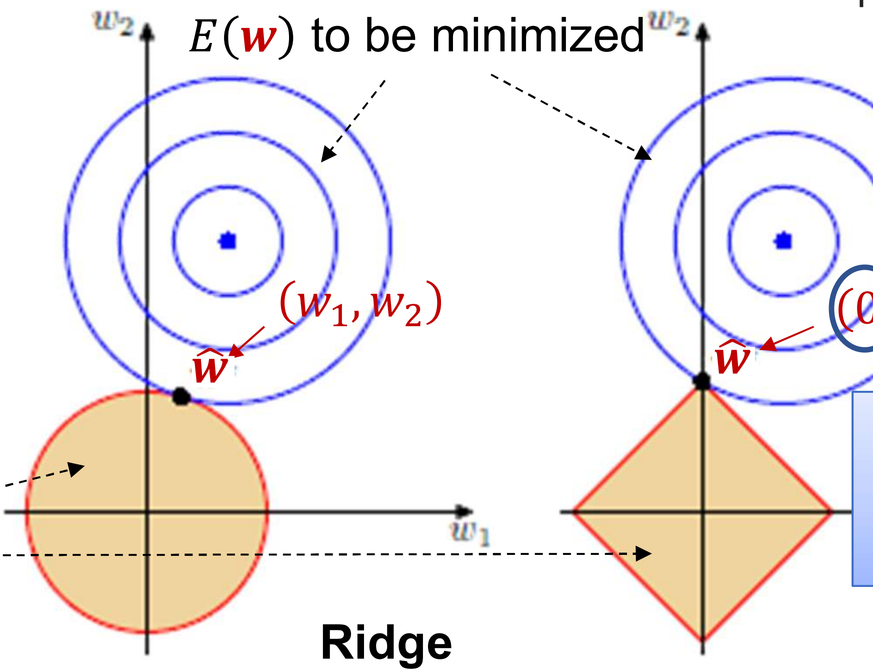
regularization term

$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, E_{reg}(\boldsymbol{w})$ where, $E_{reg}(\boldsymbol{w}) = E(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|^{q}$

parameter $\boldsymbol{w}$ space

PRML, p. 146

$E(\boldsymbol{w})$ to be minimized



A sparse model can be obtained!

But $\hat{\boldsymbol{w}}$ cannot be analytically obtained any more.

the domain $\boldsymbol{w}$ where can move

$\|\boldsymbol{w}\|^2 \leq \eta$

Ridge

when $q = 2$

Lasso

when $q = 1$

Regularization - LASSO Regression

Prior (your setting)

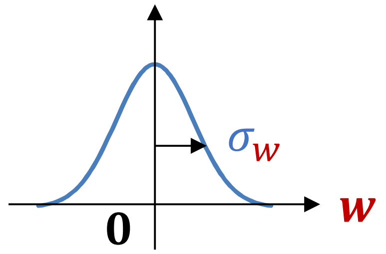
$$p(\mathbf{w}|\sigma_w) = \mathcal{N}(\mathbf{w}|0, \sigma_w{}^2\mathbf{I})$$

when the prior is a Gaussian distribution

Ridge

$$\hat{\mathbf{w}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} + \lambda\mathbf{I}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

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



Prior distribution of $\mathbf{w}$

Another Prior (your setting)

$$p(\mathbf{w}|b_w) = Laplace(\mathbf{w}|0, b_w)$$

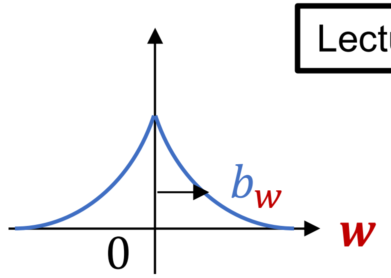
when is the prior is a Laplace distribution

Lasso

$\hat{\mathbf{w}}$: **no analytical solution**

$$p(\mathbf{w}|b_w) = \frac{1}{(2b_w)^{M+1}} \exp\left\{-\frac{\|\mathbf{w}\|}{b_w}\right\}$$

Lecture 4



Prior distribution of $\mathbf{w}$

Please confirm this by yourself

Bayesian Linear Regression (Note)

Details of the analytical solutions:
See PRML, p.152

Prior (your setting)

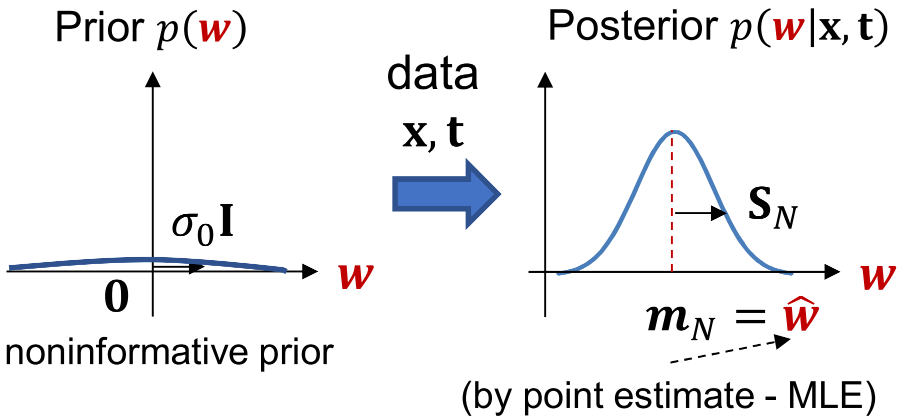
$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, {\sigma_0}^2\mathbf{I})$$

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

Posterior

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

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



$$\boldsymbol{m}_N = \hat{\sigma}^{-2}\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \xrightarrow{\sigma_0\to\infty} \hat{\sigma}^{-2}\left(\hat{\sigma}^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \quad = \hat{\boldsymbol{w}} \text{ by MLE}$$

$$\mathbf{S}_N{}^{-1} = {\sigma_0}^{-2}\mathbf{I} + \hat{\sigma}^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} \xrightarrow{\sigma_0\to\infty} \hat{\sigma}^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$$

$$p(\mathbf{t}|\mathbf{x},\boldsymbol{w}) \propto p(\boldsymbol{w}|\mathbf{x},\mathbf{t})$$

$$likelihood \propto posterior$$

Lecture content

- Bayesian sequential learning

Bayesian Sequential Learning

Bayes' theorem

$$p(\boldsymbol{w}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}) \qquad \text{wrt } \boldsymbol{w}$$

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})}{p(\mathcal{D})}$$

A dataset $\mathcal{D}$ divided into N datasets as $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_N$.

$$\underline{p(\boldsymbol{w}|\mathcal{D}_1)} \propto p(\mathcal{D}_1|\boldsymbol{w})p(\boldsymbol{w})$$

next prior

$$\underline{p(\boldsymbol{w}|\mathcal{D}_1, \mathcal{D}_2)} \propto p(\mathcal{D}_2|\boldsymbol{w})\underline{p(\boldsymbol{w}|\mathcal{D}_1)}$$

$$p(\boldsymbol{w}|\mathcal{D}_1, \mathcal{D}_2) \propto p(\mathcal{D}_2|\boldsymbol{w})p(\mathcal{D}_1|\boldsymbol{w})p(\boldsymbol{w})$$

$$p(\boldsymbol{w}|\mathcal{D}_1, \mathcal{D}_2) = \frac{p(\mathcal{D}_2|\boldsymbol{w})p(\boldsymbol{w}|\mathcal{D}_1)}{p(\mathcal{D}_2)}$$

next prior

...

$$p(\boldsymbol{w}|\mathcal{D}) \propto \prod_{i=1}^{N} p(\mathcal{D}_i|\boldsymbol{w})\, p(\boldsymbol{w})$$

Sequential view
Bayesian sequential learning

Bayesian Sequential Learning

The likelihood of data in batch itself

$$p(\boldsymbol{w}|\mathcal{D}) \propto \boxed{\prod_{i=1}^{N} p(\mathcal{D}_i|\boldsymbol{w})}\, p(\boldsymbol{w})$$

Sequential view
Bayesian sequential learning

This can be applied to any problems in which the observed data are assumed to be **i.i.d.**

Review of Lecture 2:

Independent and identically distributed (i.i.d.)

independent

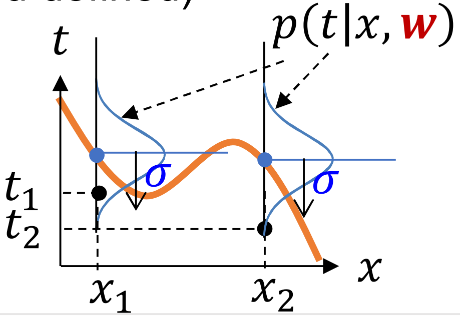
$$p(x_1, x_2) = p(x_1)p(x_2) = \prod_{i=1}^{2} p(x_i)$$

Each sample point is i.i.d.
independently generated from a Gaussian distribution
(the probabilistic model that you defined)

The likelihood function

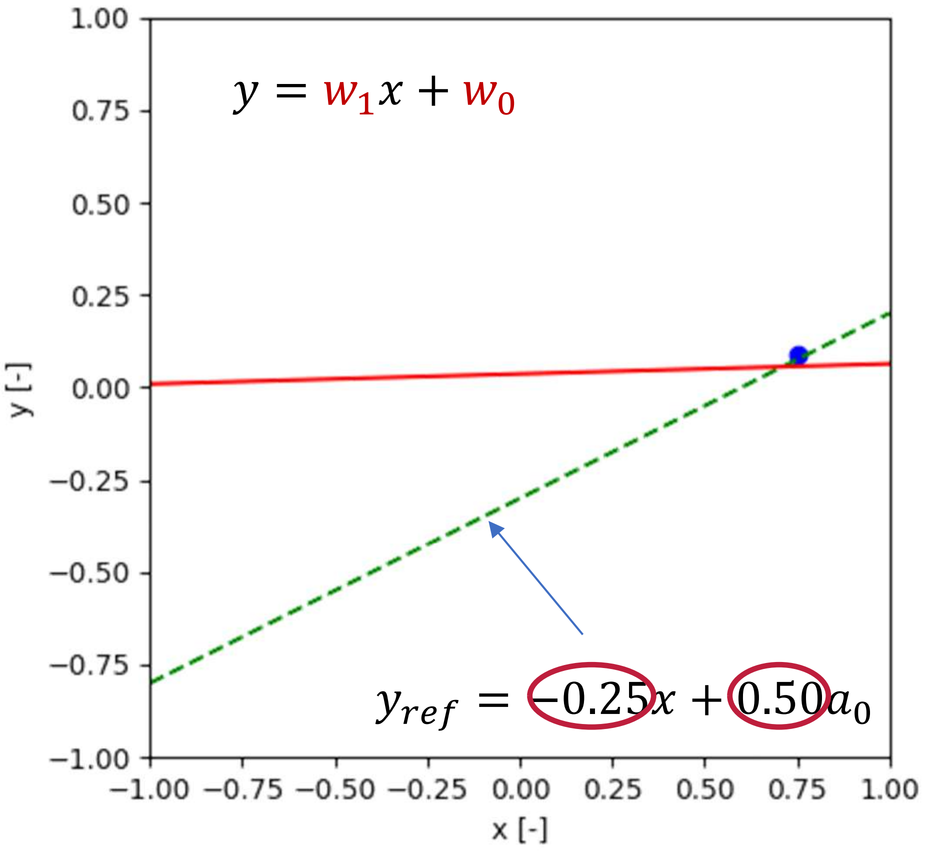
$$p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}) = \prod_{i=1}^{N} \mathcal{N}(t_i|y(x_i, \boldsymbol{w}), \hat{\sigma}^2)$$

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



Bayesian Sequential Learning

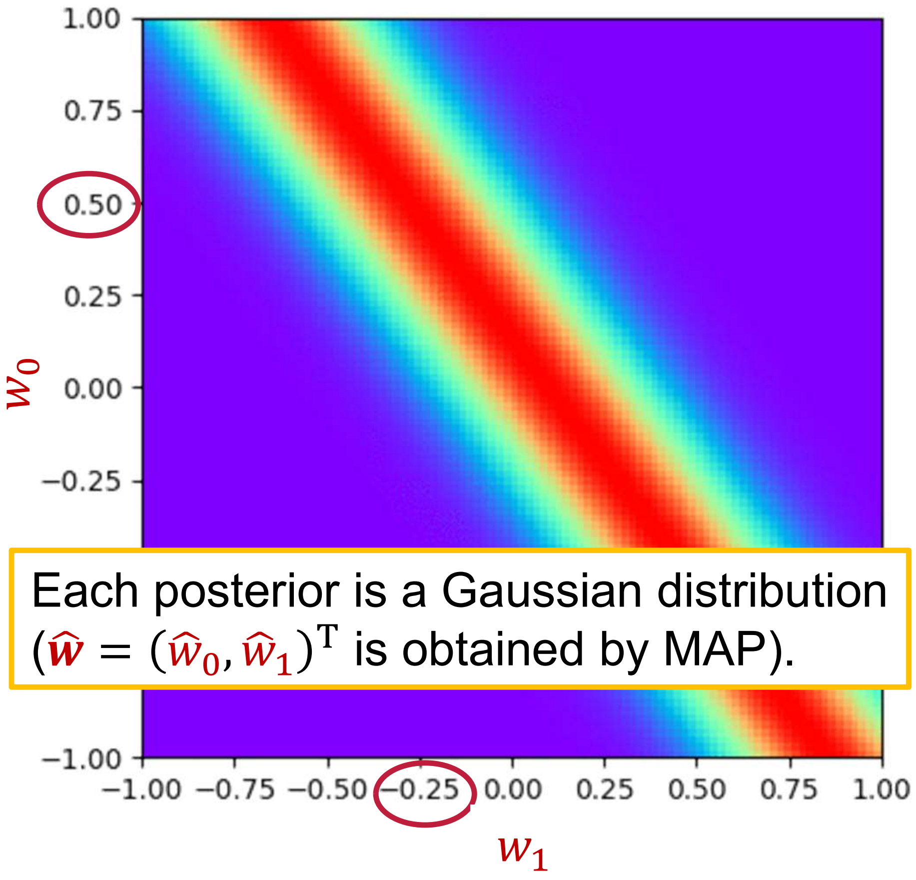
input - output space



$y = w_1 x + w_0$

$y_{ref} = -0.25x + 0.50a_0$

parameter space $p(\boldsymbol{w}|\mathcal{D})$



Each posterior is a Gaussian distribution ($\hat{\boldsymbol{w}} = (\hat{w}_0, \hat{w}_1)^{\mathrm{T}}$ is obtained by MAP).

The parameters w_0 and w_1 are being clarified to be around (-0.25, 0.50).

Lecture content

- Uncertainty due to data (active learning)

Bayesian Linear Regression

Details of the analytical solutions:
See PRML, p.152

Prior (your setting)

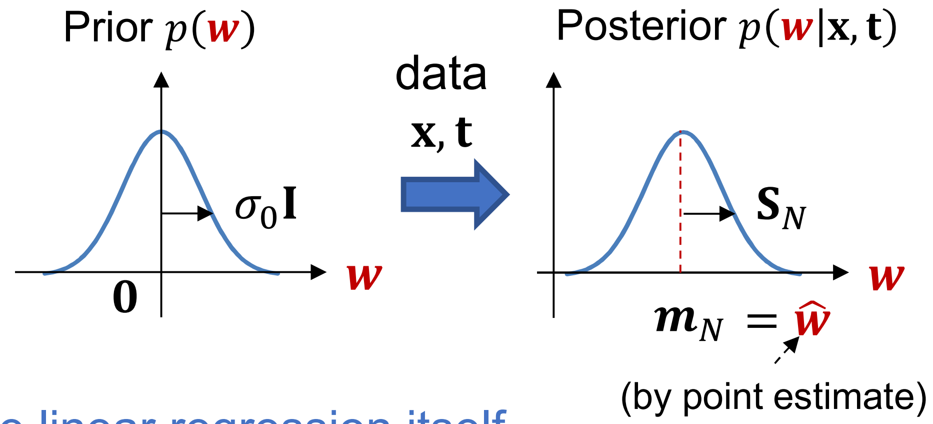
$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, {\sigma_0}^2\mathbf{I})$$

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

Posterior

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

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



Predictive distribution (the goal)

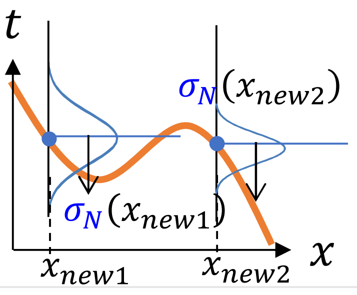
the linear regression itself

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

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

What is σ_N?

The predictive distribution result contains the result of “point estimate”.



Bayesian Linear Regression

Note:
The reason why the variance can be decomposed clearly like this is again due to the properties of Gaussian distributions in general.

$$\sigma_N{}^2(\boldsymbol{x}) = \hat{\sigma}^2 + \overset{N \to \infty}{\cancelto{0}{\sigma_N{}'^2(\boldsymbol{x})}}$$

= **noise** + **uncertainty associated with *w***

What does this mean?

where, $\sigma_N{}'^2(\boldsymbol{x}) = \phi(\boldsymbol{x})^{\mathrm{T}} \mathbf{S}_N \phi(\boldsymbol{x})$

The meanings of these two components are totally different from each other.

Bayesian Linear Regression

The probabilistic model

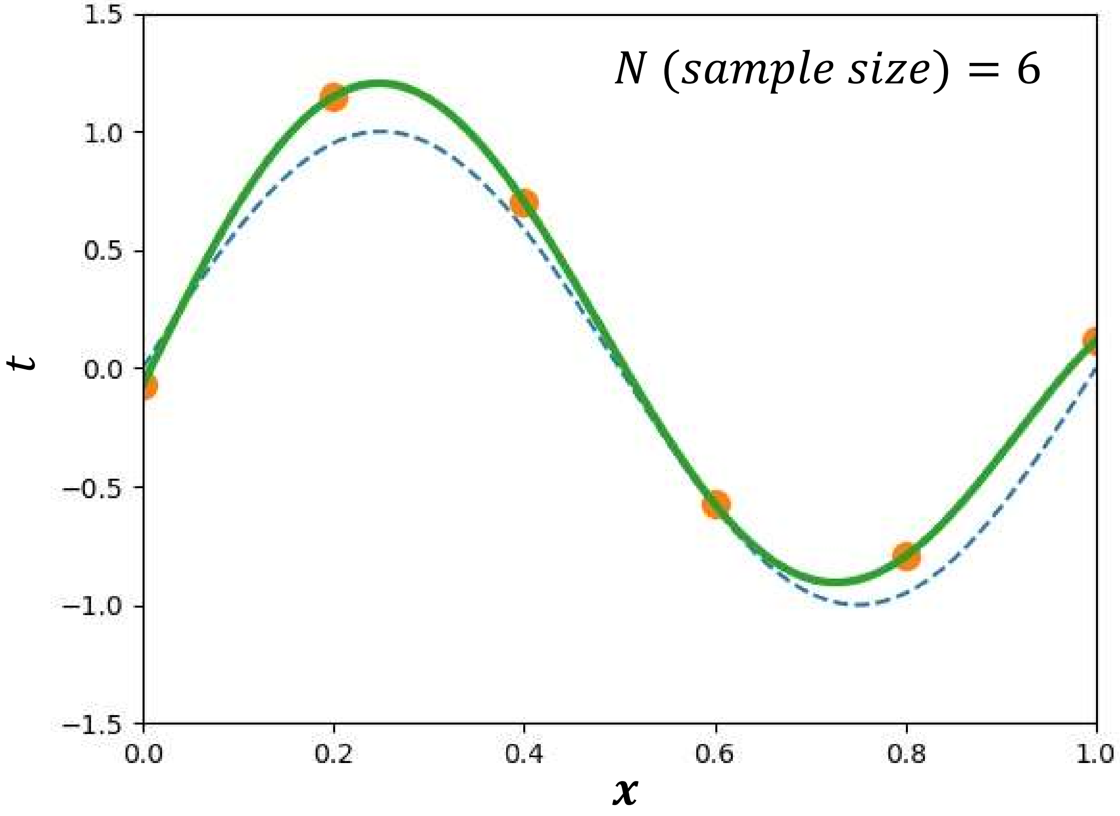
$p(t|x,\mathbf{w}) = \mathcal{N}\left(t\middle|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$

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

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

(MLE) Least square method

Examples



Bayesian Linear Regression

The probabilistic model

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

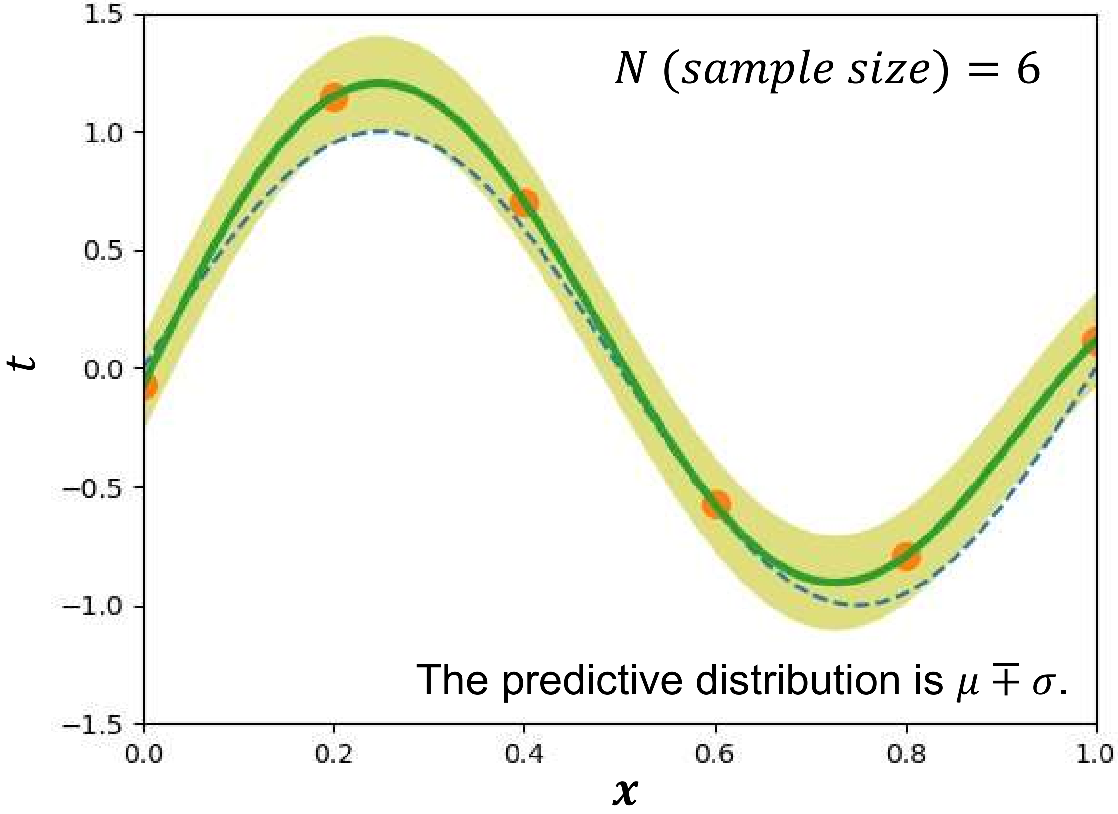
$\boldsymbol{\phi}(\boldsymbol{x})$ = polynomials $(M = 9)$

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

Examples

(MLE) Least square method

$\hat{\sigma}$



$\hat{\sigma}$: constant noise

Bayesian Linear Regression

The probabilistic model

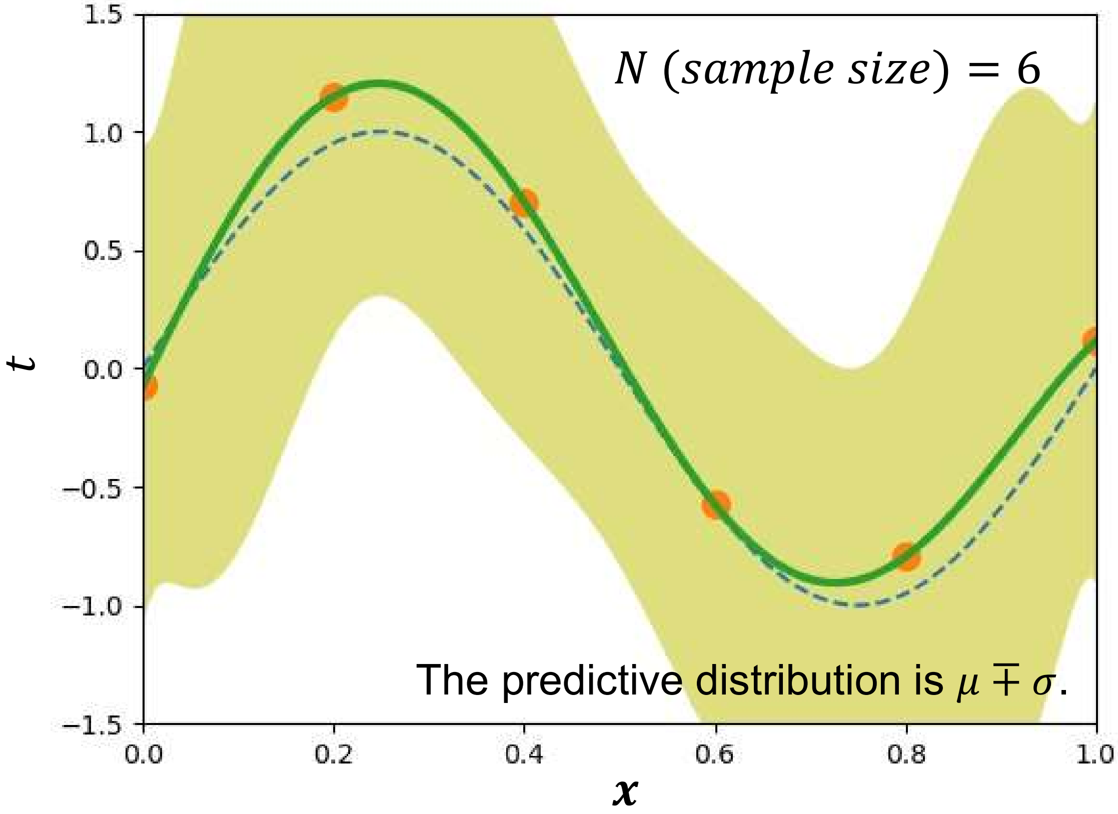
$p(t|x,\mathbf{w}) = \mathcal{N}\left(t\middle|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$

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

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

Examples

${\sigma_N}^2(\boldsymbol{x})$



${\sigma_N}^2(\boldsymbol{x})$: constant noise + uncertainty associated with $\mathbf{w}$

Bayesian Linear Regression

The probabilistic model

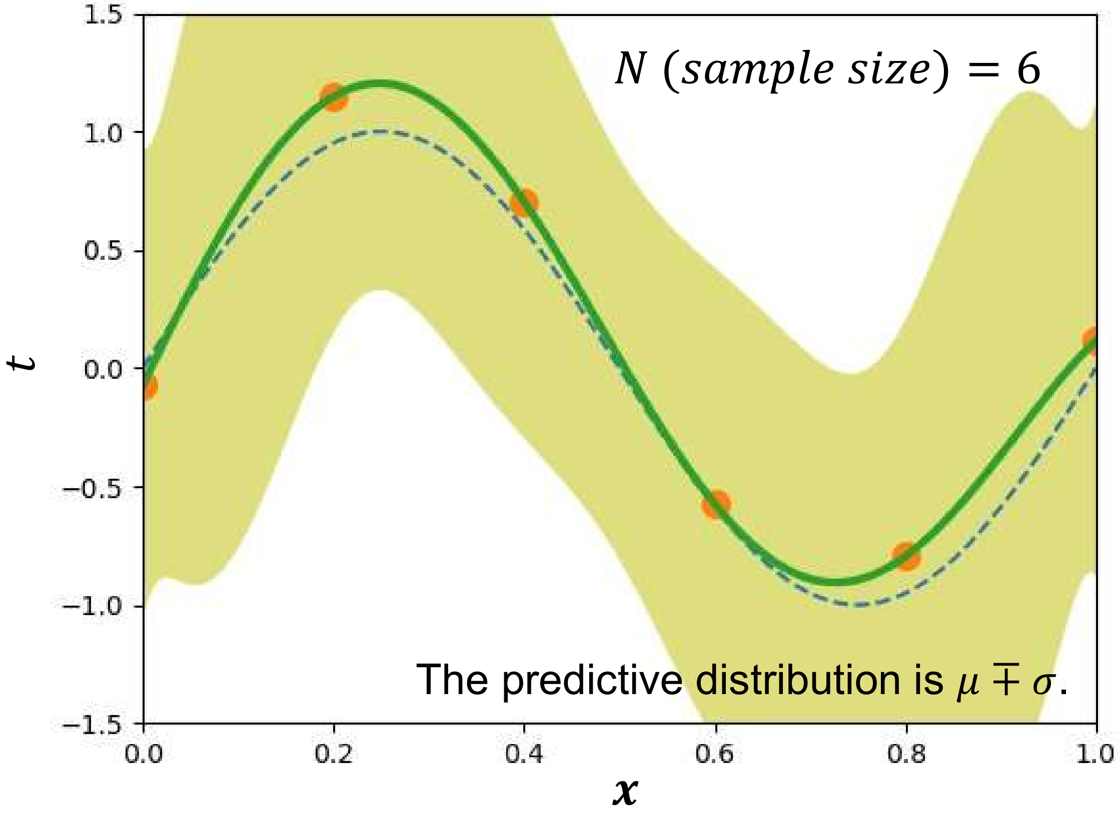
$p(t|x, \mathbf{w}) = \mathcal{N}\left(t \middle| \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$

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

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

Examples

$\sigma_N{'}^2(\boldsymbol{x})$



The predictive distribution is $\mu \mp \sigma$.

$\sigma_N{'}^2(\boldsymbol{x})$: uncertainty associated with $\mathbf{w}$

Bayesian Linear Regression

The probabilistic model

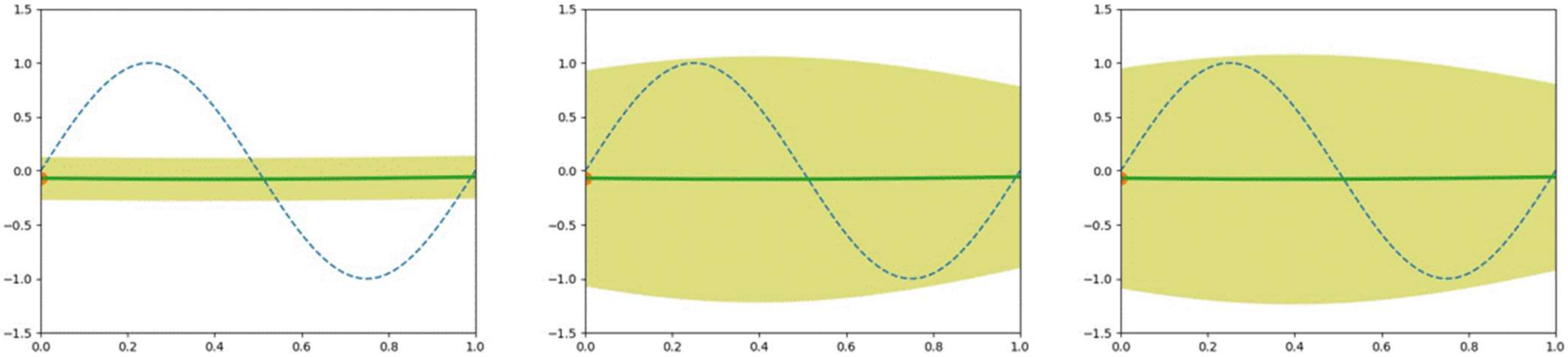
$p(t|x, \boldsymbol{w}) = \mathcal{N}\left(t \middle| \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$

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

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

Examples

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



$\hat{\sigma}^2$ $+$ ${\sigma_N}'^2(\boldsymbol{x})$ $=$ ${\sigma_N}^2(\boldsymbol{x})$

${\sigma_N}'^2(\boldsymbol{x})$: uncertainty associated with $\boldsymbol{w}$

Note: The predictive distribution is $\mu \mp \sigma$.

Bayesian Linear Regression

The probabilistic model

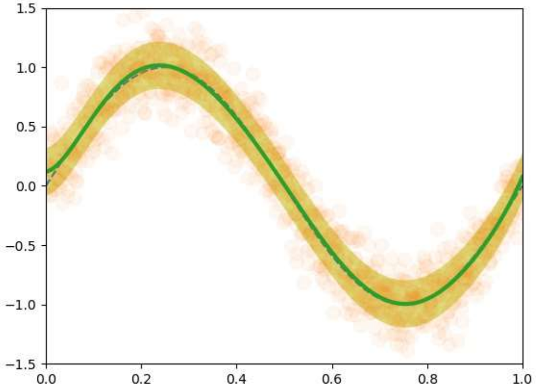
$$p(t|x,\boldsymbol{w}) = \mathcal{N}\left(t\middle|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}),\hat{\sigma}^2\right)$$

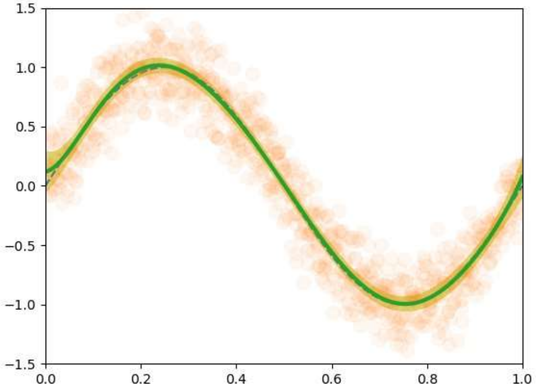
$\boldsymbol{\phi}(\boldsymbol{x}) =$ polynomials $(M = 9)$

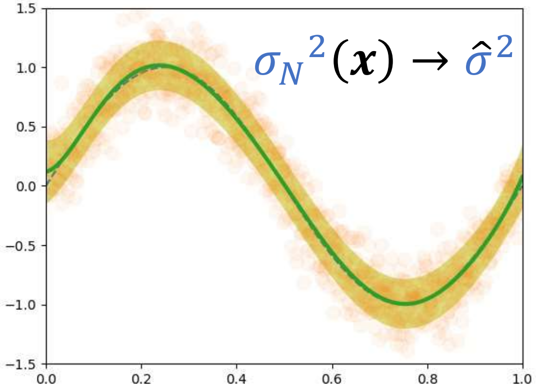
$\hat{\sigma} = 0.2$ (fixed)

Examples

$$N = 1000$$







$\sigma_N{}^2(\boldsymbol{x}) \rightarrow \hat{\sigma}^2$

$$\hat{\sigma}^2 \quad + \quad \cancelto{0}{\sigma_N{}'^2(\boldsymbol{x})} \quad = \quad \sigma_N{}^2(\boldsymbol{x})$$

$\sigma_N{}'^2(\boldsymbol{x})$: uncertainty associated with $\boldsymbol{w}$

Note: The predictive distribution is $\mu \mp \sigma$.

Bayesian Linear Regression

Question: How can we use this information?

Active learning

if evaluating the output t (for a given input $\boldsymbol{x}$) is **expensive**

Request new sample points on:

- where σ_N is large.
- where a criterion defined by σ_N and the optimum location among the current sample points is large.
- etc. (σ_N + other information)

Annotation in classification

Bayesian optimization

With a small amount of learning datasets, the output t is **efficiently** predicted.

→ **Gaussian Processes** (Lectures 7-9) are more often used
(since weak assumption on the probabilistic model can be used).

Bayesian Linear Regression

The probabilistic model

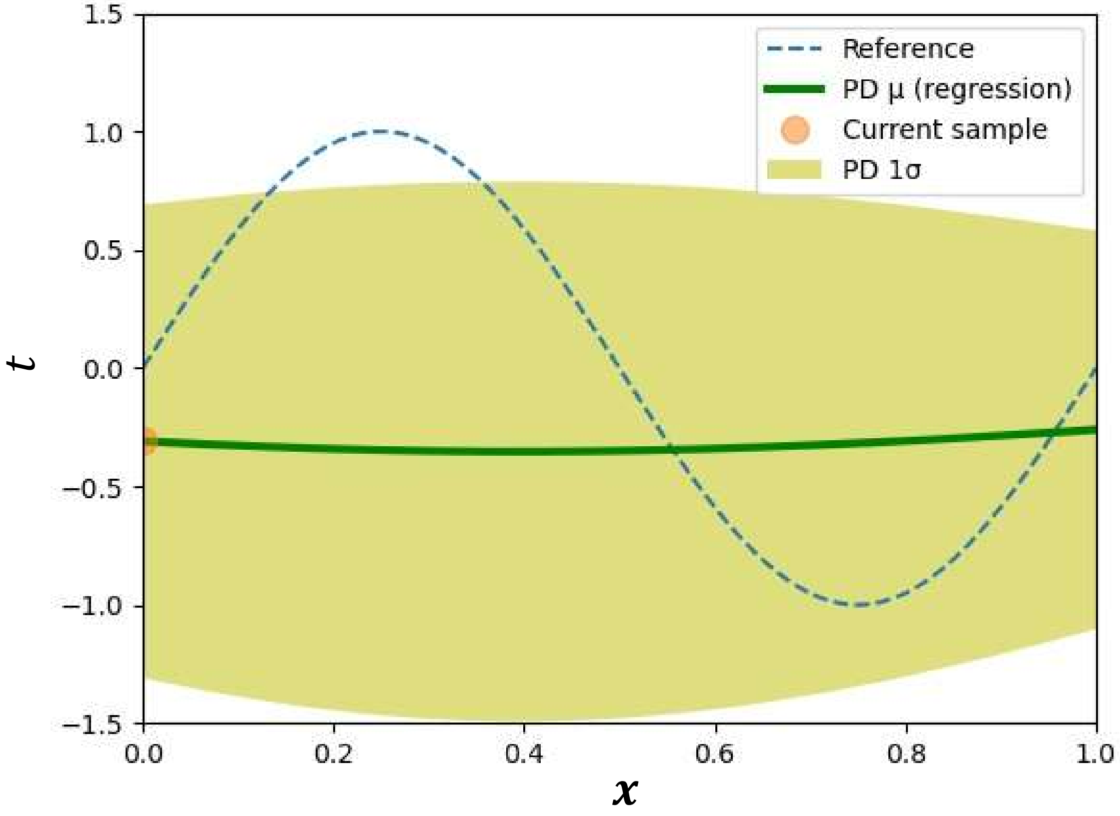
$p(t|x, \mathbf{w}) = \mathcal{N}\left(t \middle| \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$

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

Bayesian Linear Regression

The probabilistic model

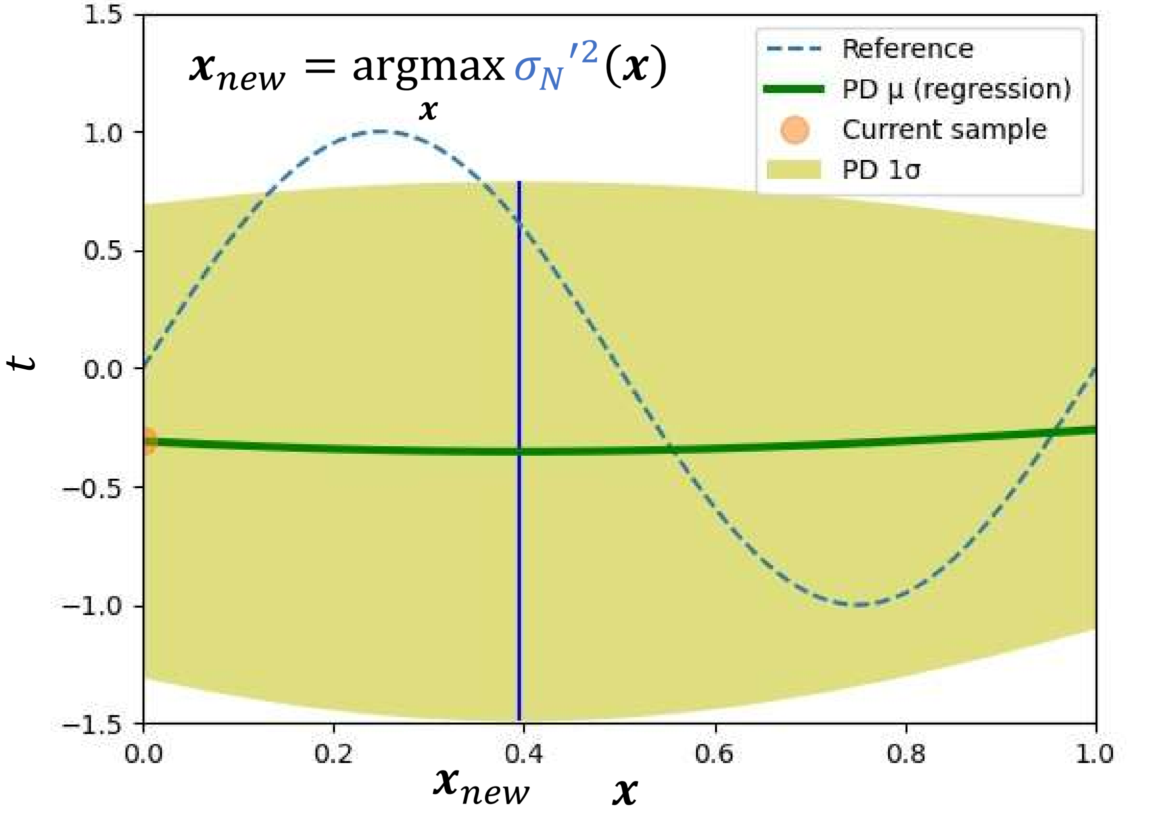
$$p(t|x,\boldsymbol{w}) = \mathcal{N}\left(t\middle|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\boldsymbol{x})$ = 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**



Ask to provide the output t for the location $\boldsymbol{x}_{new}$

Bayesian Linear Regression

The probabilistic model

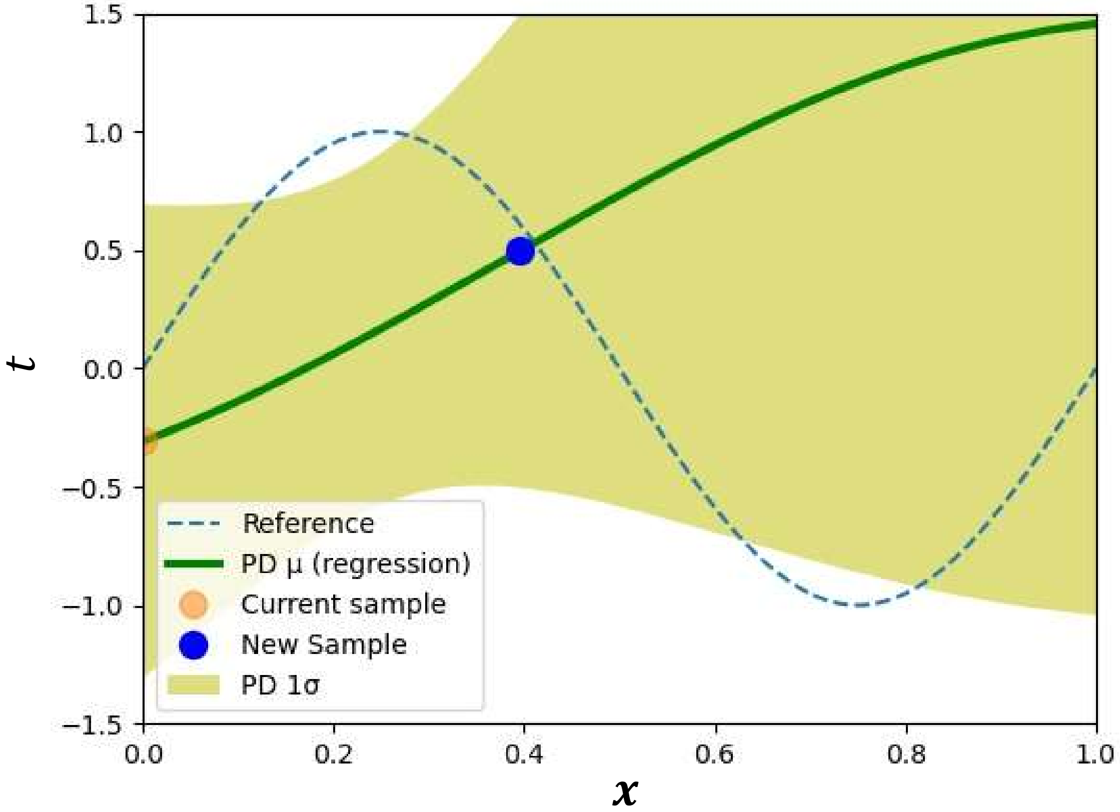
$$p(t|x,\mathbf{w}) = \mathcal{N}\left(t\middle|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\boldsymbol{x})$ = 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**



A new sample point has been added **actively/automatically**.

Bayesian Linear Regression

The probabilistic model

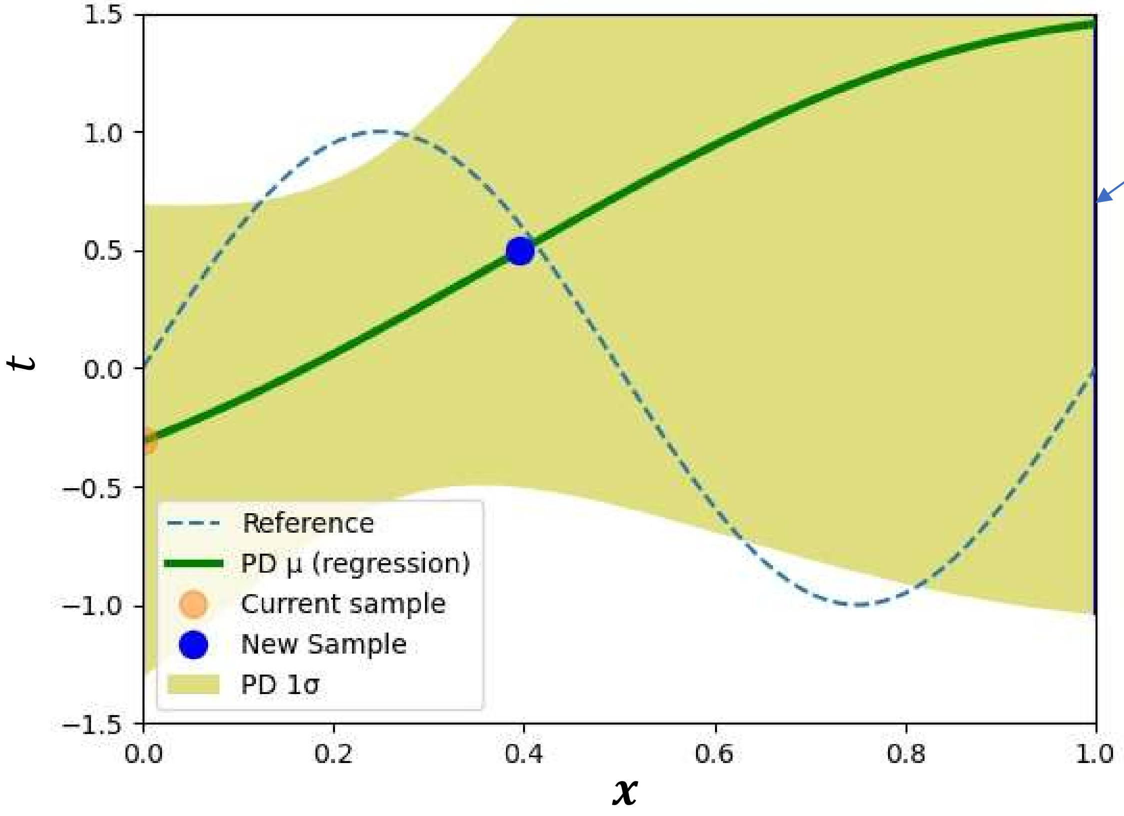
$p(t|x, \boldsymbol{w}) = \mathcal{N}\left(t \middle| \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$

$\boldsymbol{\phi}(\boldsymbol{x}) =$ 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**



difficult to see but here

Continue the iterative process…

Bayesian Linear Regression

The probabilistic model

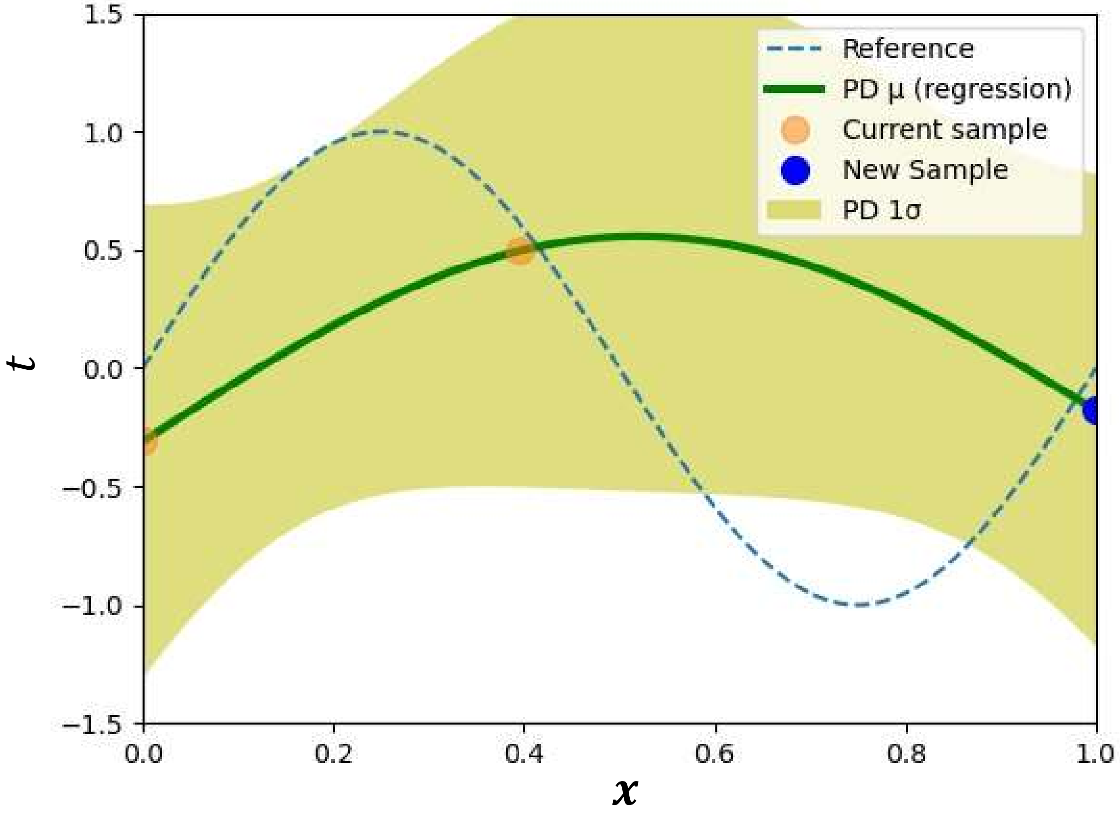
$$p(t|x, \boldsymbol{w}) = \mathcal{N}\left(t \middle| \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\boldsymbol{x})$ = 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**



Bayesian Linear Regression

The probabilistic model

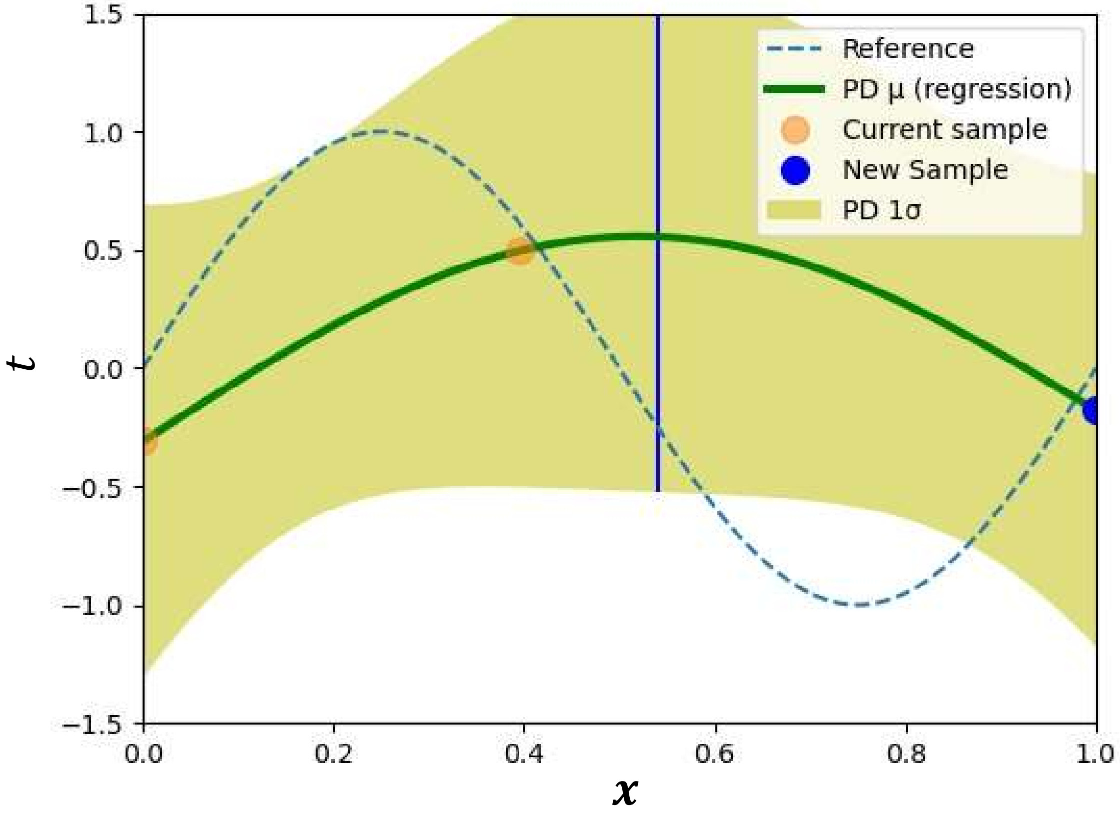
$$p(t|x, \mathbf{w}) = \mathcal{N}\left(t \middle| \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\boldsymbol{x})$ = 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**



Bayesian Linear Regression

The probabilistic model

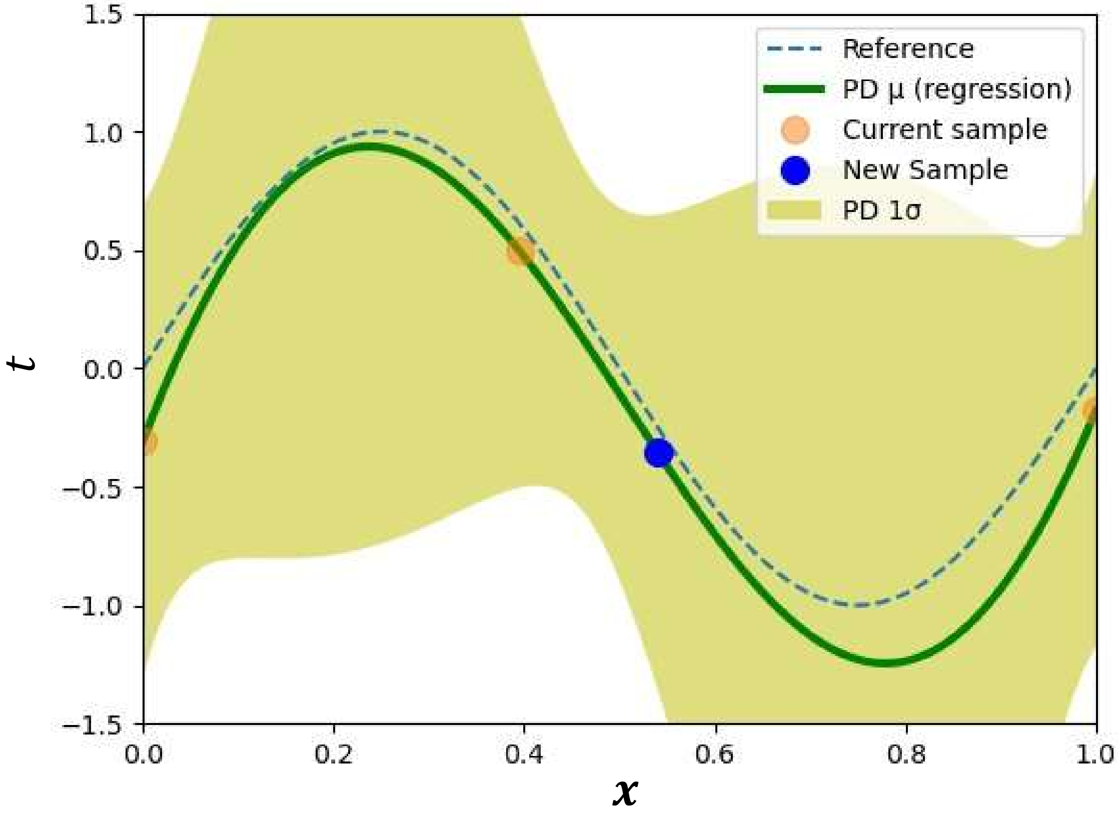
$$p(t|x, \mathbf{w}) = \mathcal{N}\left(t \middle| \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\mathbf{x})$ = 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**



Bayesian Linear Regression

The probabilistic model

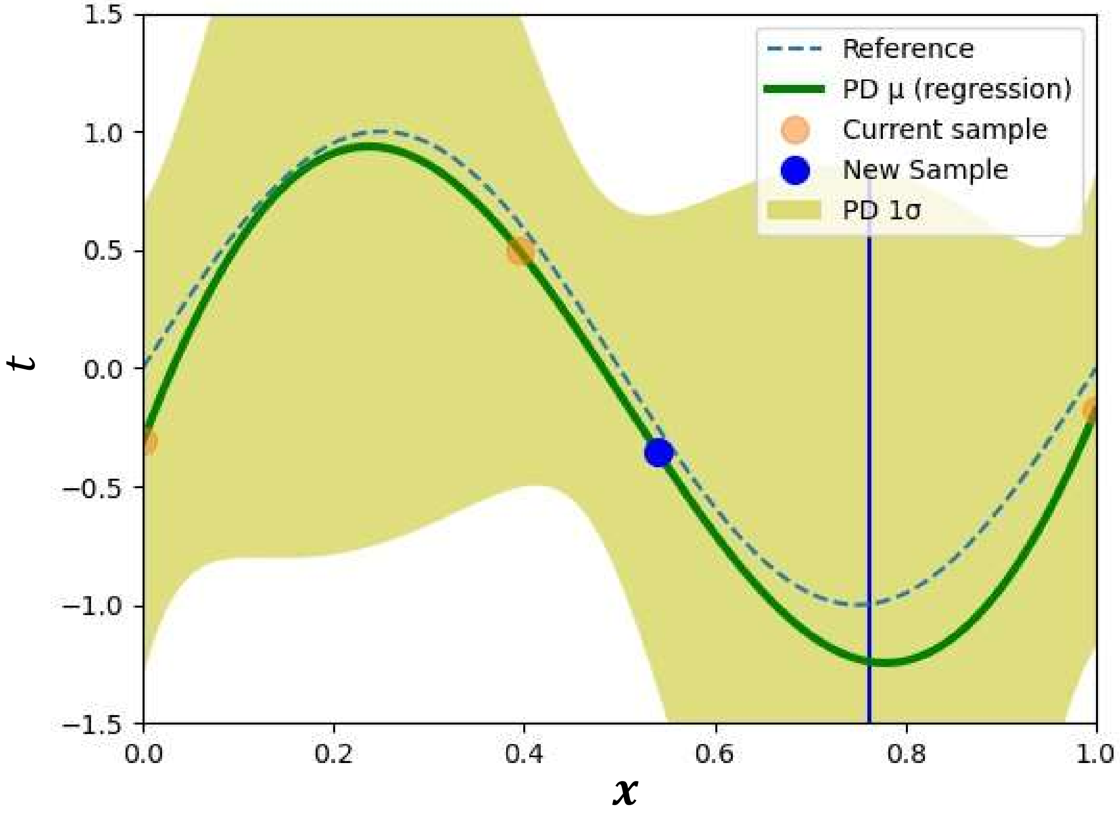
$$p(t|x,\mathbf{w}) = \mathcal{N}\left(t\middle|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\boldsymbol{x})$ = 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**



Bayesian Linear Regression

The probabilistic model

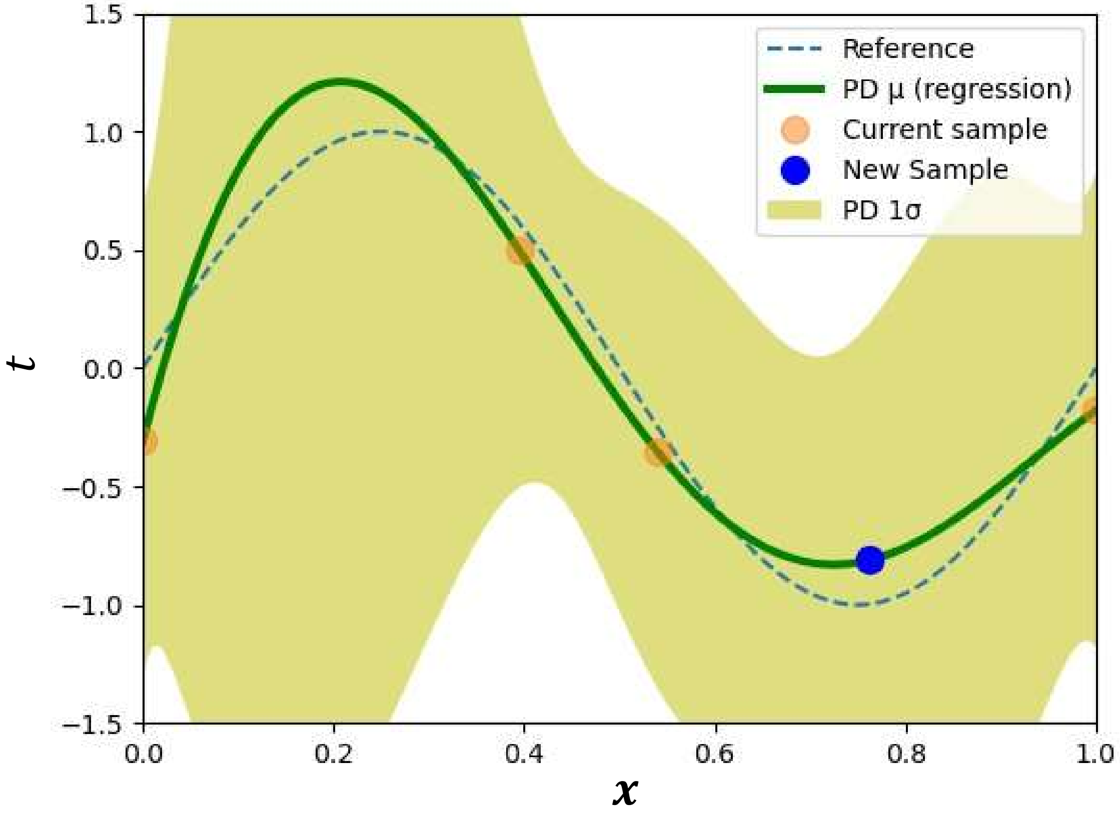
$$p(t|x, \mathbf{w}) = \mathcal{N}\left(t\middle|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\boldsymbol{x})$ = 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**



Bayesian Linear Regression

The probabilistic model

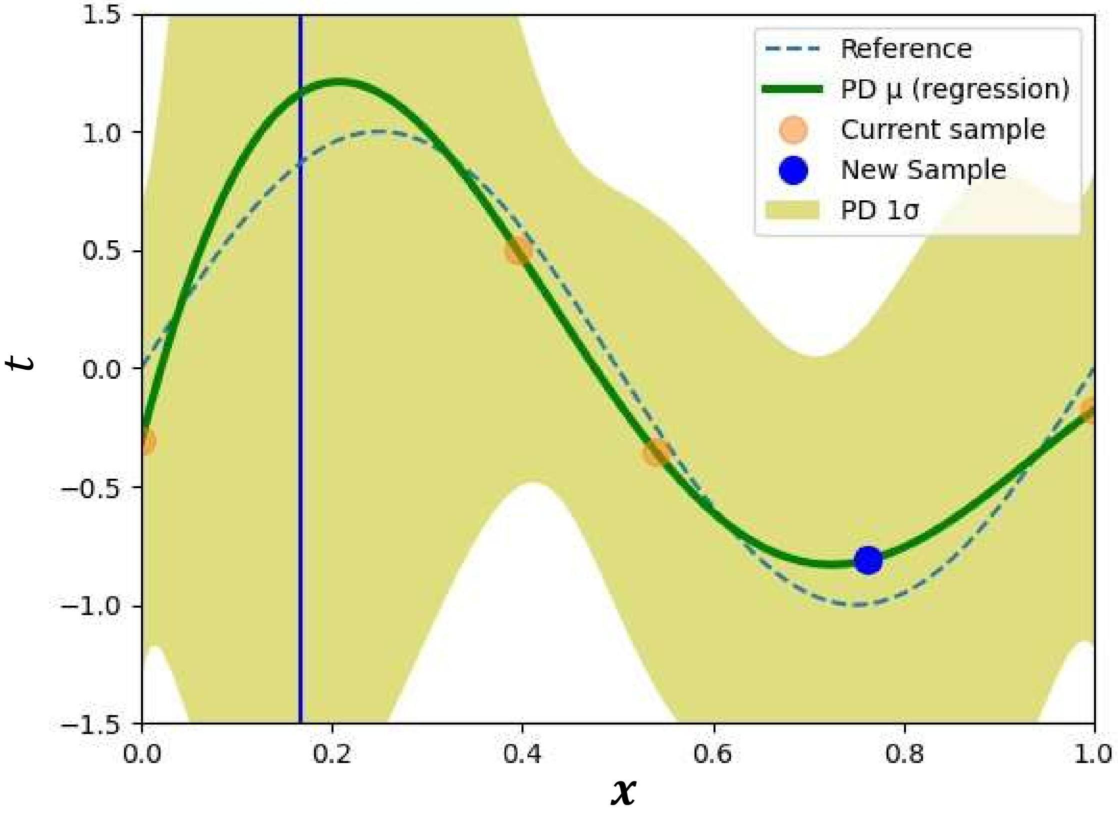
$$p(t|x,\mathbf{w}) = \mathcal{N}\left(t\middle|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\boldsymbol{x})$ = 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**



Bayesian Linear Regression

The probabilistic model

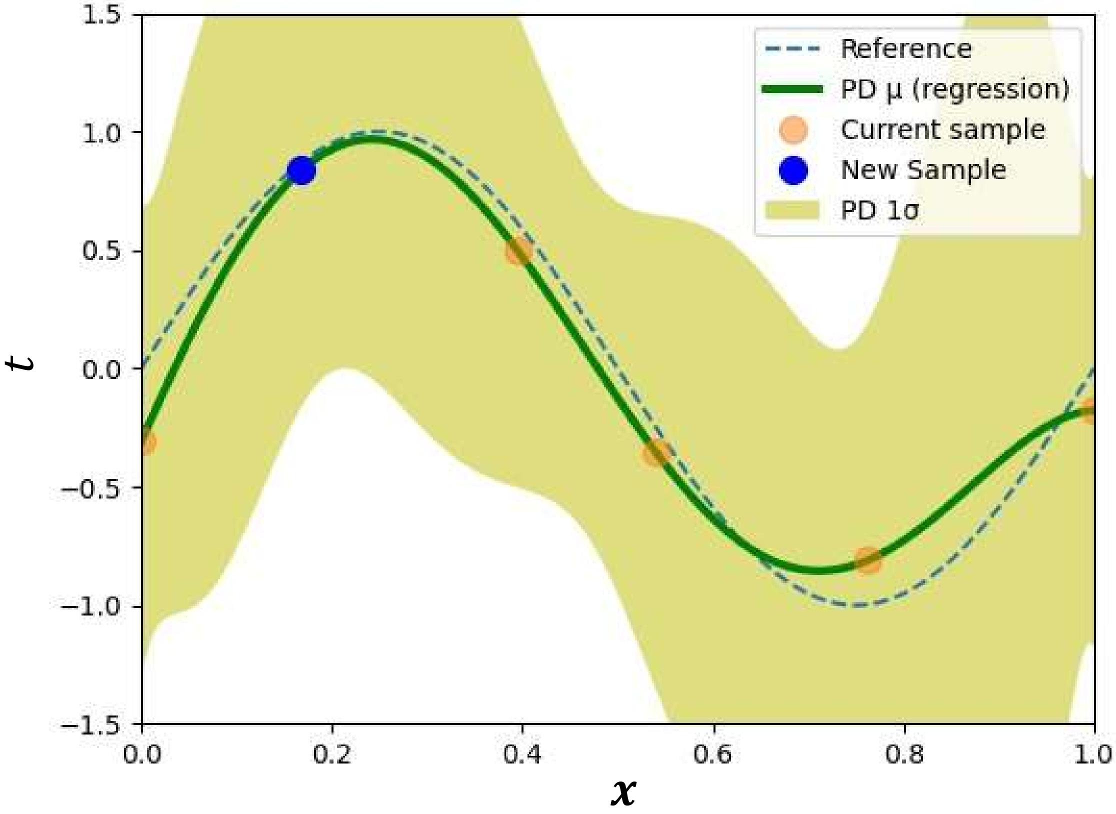
$$p(t|x,\boldsymbol{w}) = \mathcal{N}\left(t\middle|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\boldsymbol{x}) =$ 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**



Bayesian Linear Regression

The probabilistic model

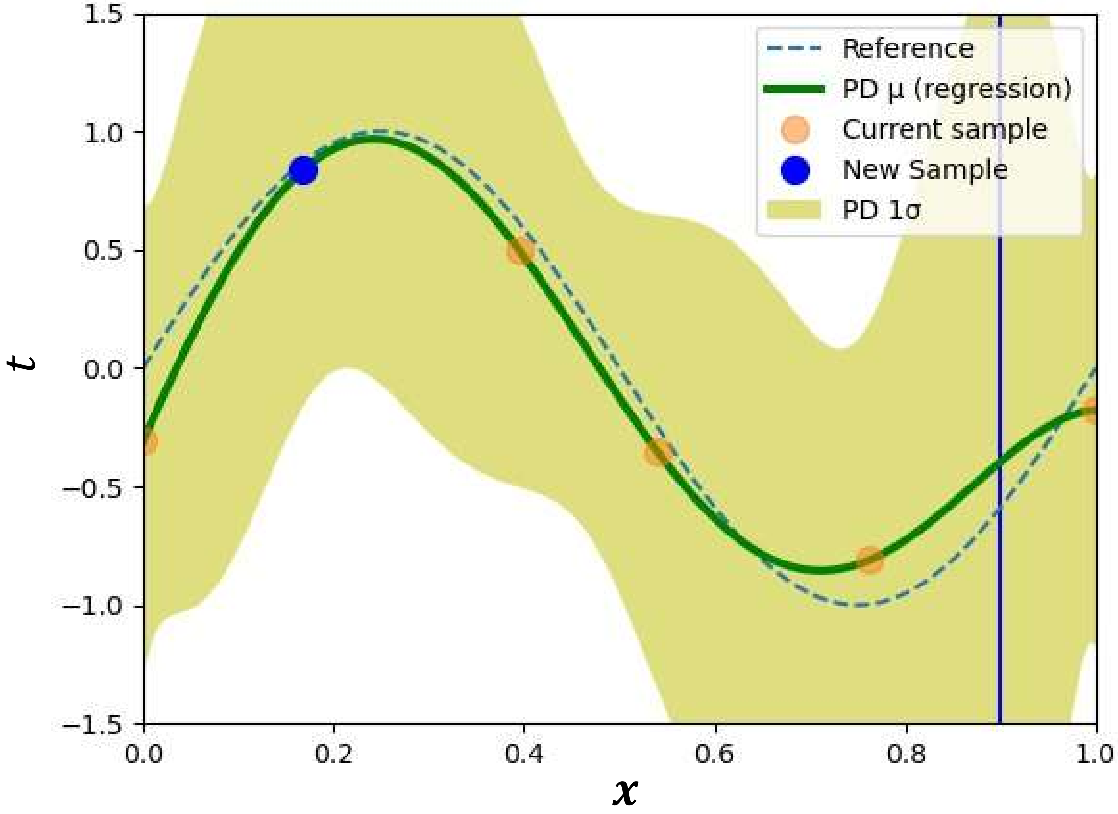
$$p(t|x, \boldsymbol{w}) = \mathcal{N}\left(t \middle| \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\boldsymbol{x})$ = 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**



Bayesian Linear Regression

The probabilistic model

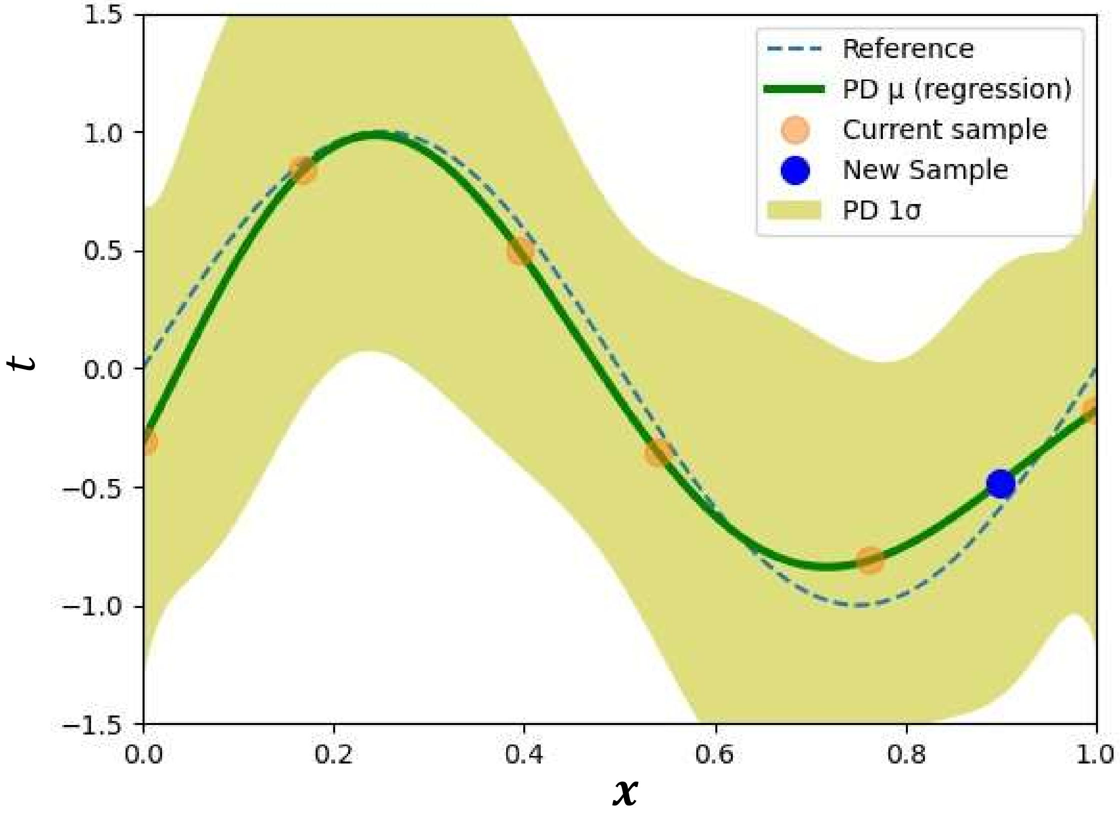
$$p(t|x, \boldsymbol{w}) = \mathcal{N}\left(t \middle| \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$$

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

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

Examples

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



Bayesian Linear Regression

The probabilistic model

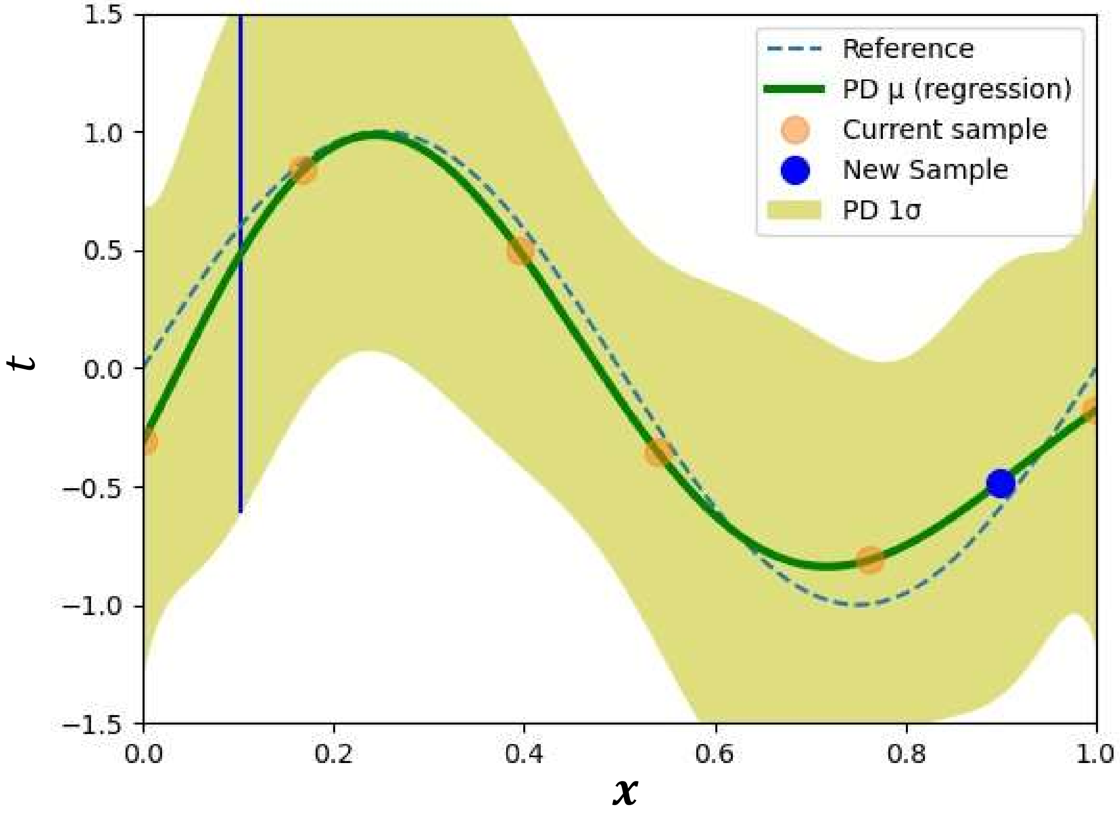
$p(t|x,\mathbf{w}) = \mathcal{N}\left(t\middle|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$

$\boldsymbol{\phi}(\boldsymbol{x}) =$ 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**



Bayesian Linear Regression

The probabilistic model

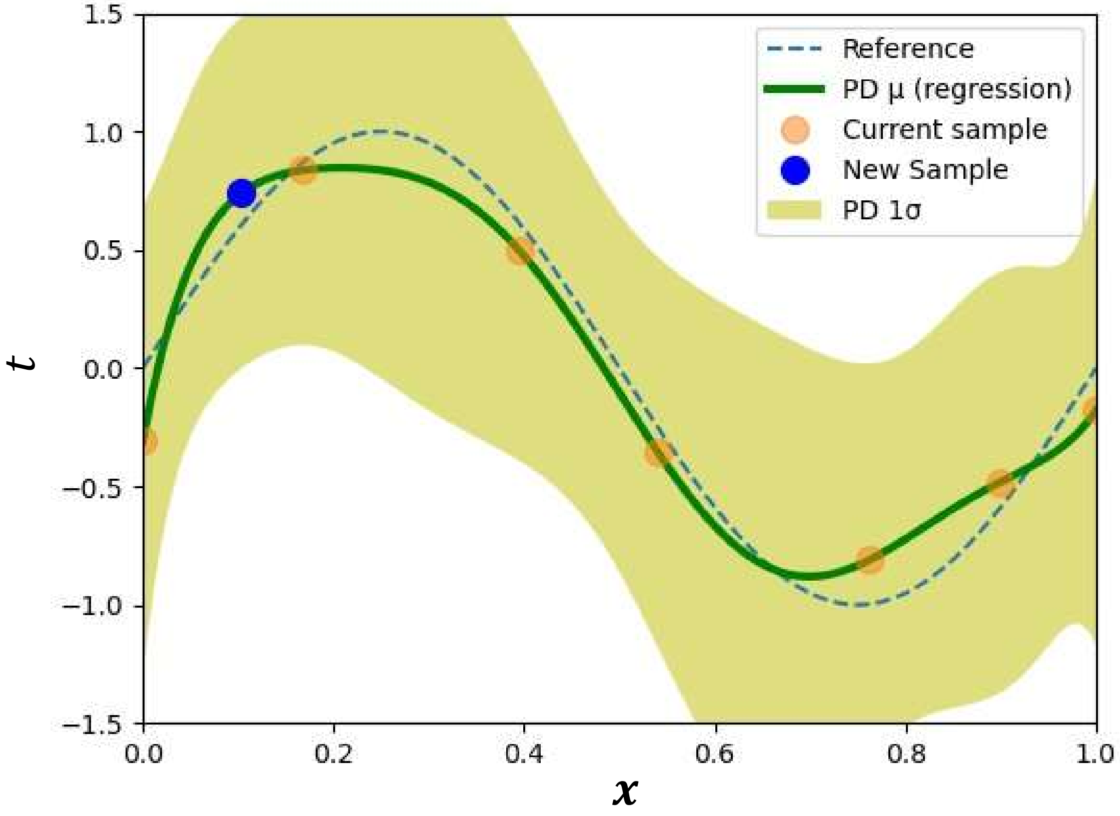
$$p(t|x,\mathbf{w}) = \mathcal{N}\left(t\middle|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}),\hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\boldsymbol{x})$ = 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**



Bayesian Linear Regression

The probabilistic model

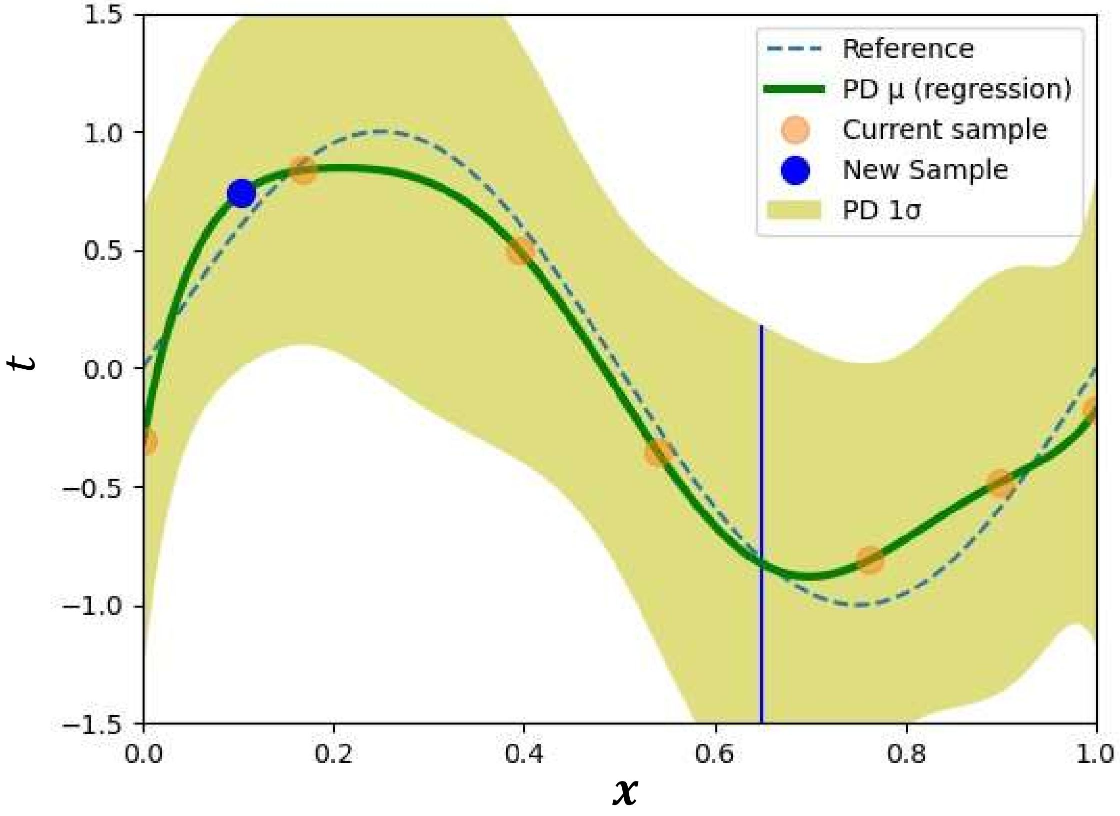
$$p(t|x,\boldsymbol{w}) = \mathcal{N}\left(t\middle|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}),\hat{\sigma}^2\right)$$

$\boldsymbol{\phi}(\boldsymbol{x})$ = 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**



Bayesian Linear Regression

The probabilistic model

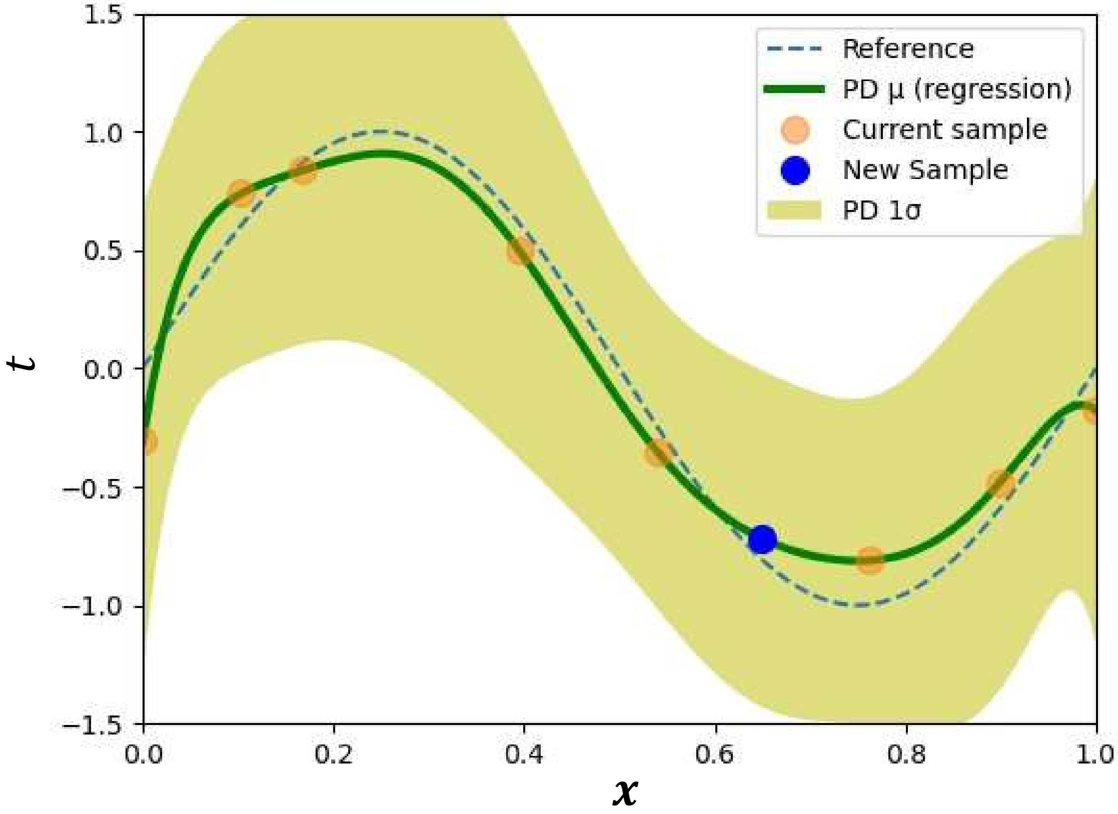
$p(t|x,\mathbf{w}) = \mathcal{N}\left(t\middle|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \hat{\sigma}^2\right)$

$\boldsymbol{\phi}(\boldsymbol{x})$ = 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**



Prior in General

Prior:

- Anything is fine.
  - Evidence in past
  - Regularization
  - Imaginary sample data
  - Your belief
  - etc.

Summary

Generalized Process

1. Define a probabilistic model
2. Then, **compute the posterior**
    - (Define a prior distribution)
        - **Point estimate** (Deterministic)
        - **Probability distribution** (Stochastic) ← computations hard

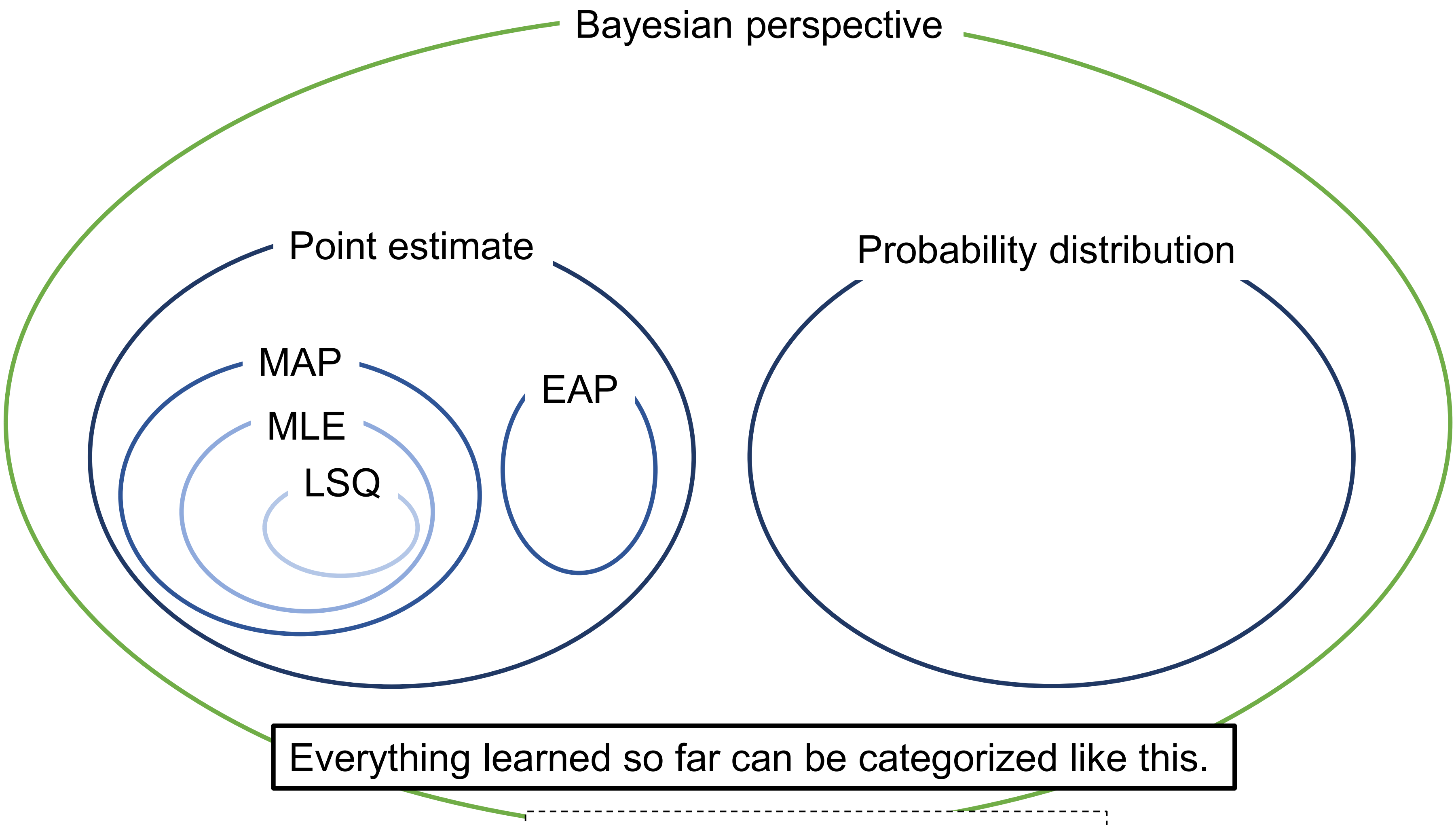
By using special cases,
We could see what we can do by the Bayesian perspectives

➡ We could know: Generalized perspectives + Various possibilities

There are numerical techniques – e.g. MCMC (Lecture 12) to compute the difficult parts:

- Posterior distribution
- Predictive distribution (as the goal)

Bayesian Approach – Generalized Perspective



Bayesian perspective

Point estimate

MAP

MLE

LSQ

EAP

Probability distribution

Everything learned so far can be categorized like this.

- Regularization
- Bayesian sequential learning
- …