Bayesian Deep Learning Tutorial

- Session 1: Bayesian methods for machine learning [Pengyu (Ben) Yuan]
  - Introduction & basic Bayesian rule
  - Non-Bayesian machine learning method
  - Bayesian machine learning method (Variational Inference)
  - Non-parametric machine learning method (Gaussian Process)
- Session 2: Bayesian deep learning [Dan Nguyen]
  - Uncertainty in model predictions
  - Bayesian deep learning with dropout
  - Some practical application examples
- Session 3: Bayesian deep learning demos [Pengyu (Ben) Yuan]
  - DropConnect Is Effective in Modeling Uncertainty of Bayesian Deep Networks
  - Skin lesion classification demos
  - Organ segmentation demos

MAIA Laboratory

H U L A Lab

UNIVERSITY of HOUSTON | ECE

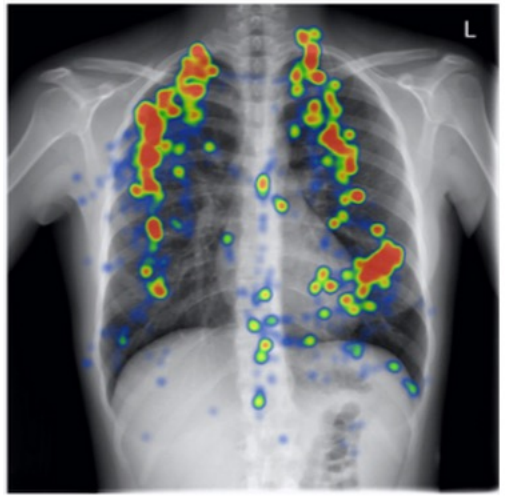
Bayesian methods for machine learning

HoUston Learning Algorithms (HULA) Lab
Presented by Pengyu (Ben) Yuan

About our lab

HoUston Learning Algorithms (HULA) Lab

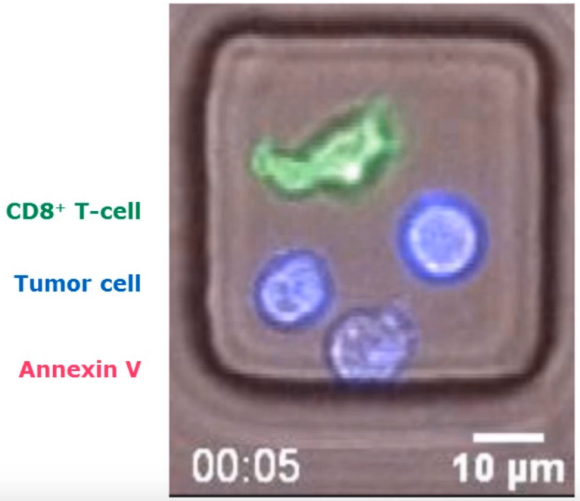
Funded by



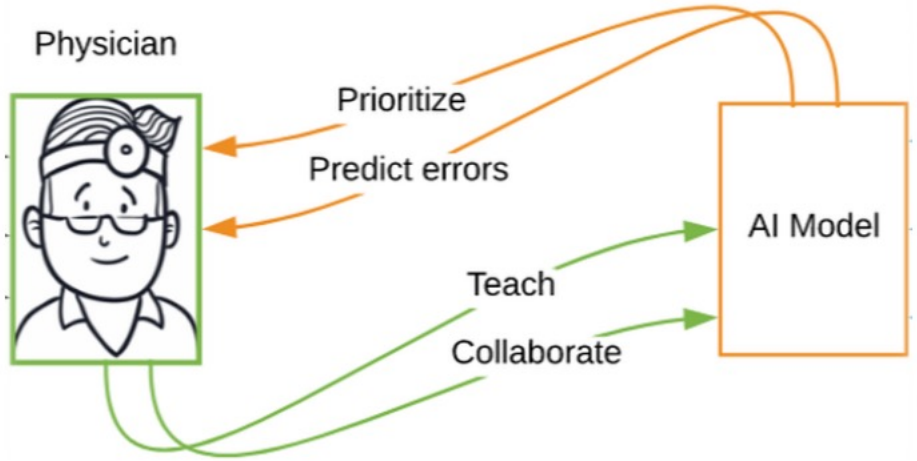
Understanding diagnostic errors

Domain generalization

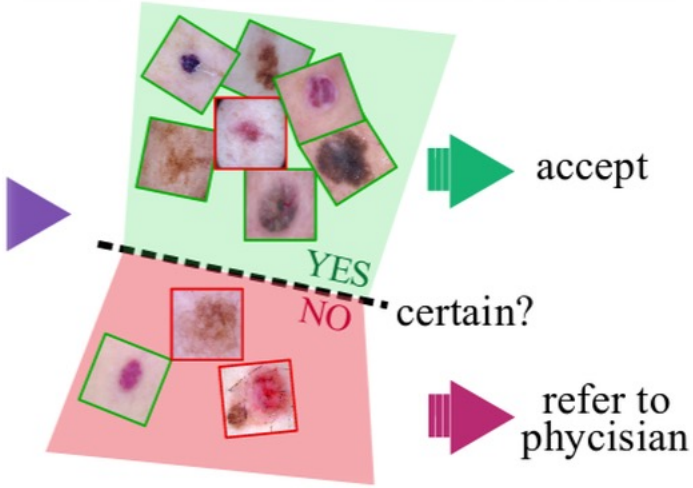
CAR+CD8+ T cells participate in efficient multi-killing



Analyzing cellular activities

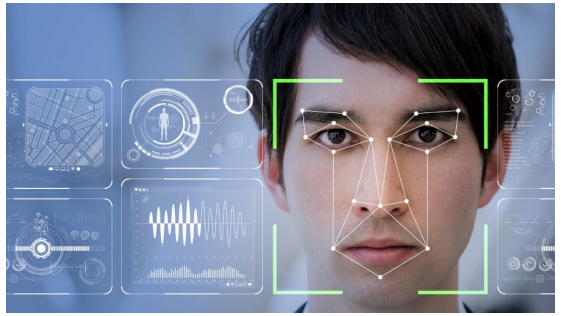


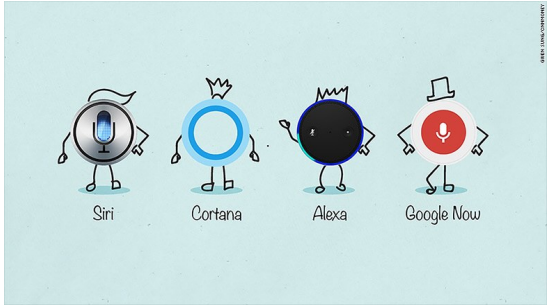
Physician-friendly AI

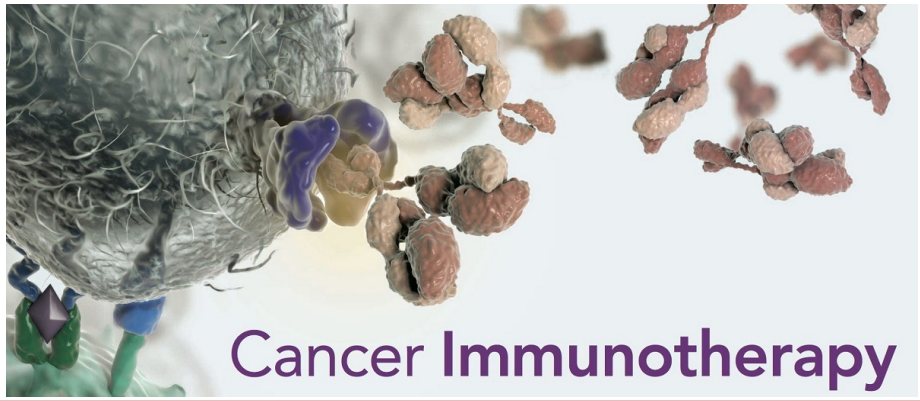


Risk-Aware Deep Learning

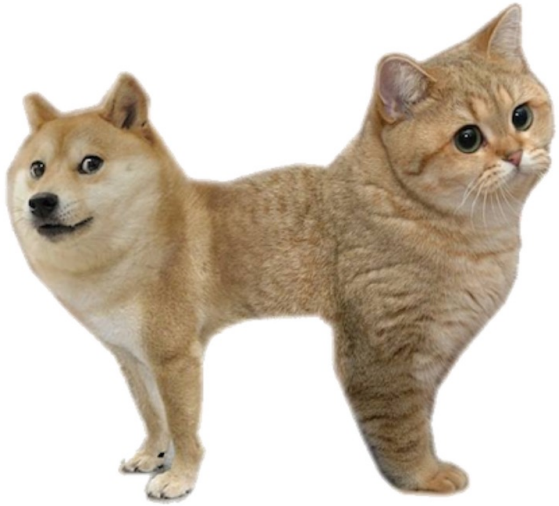
Machine learning is impacting our life



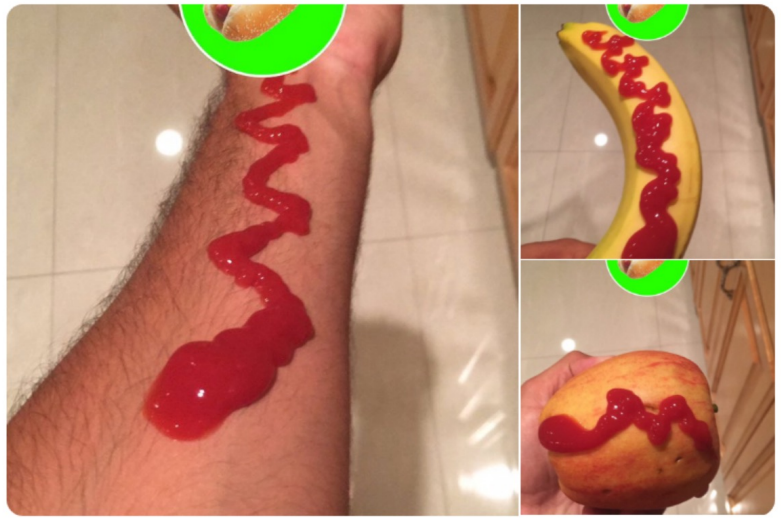




Is this a cat or dog?

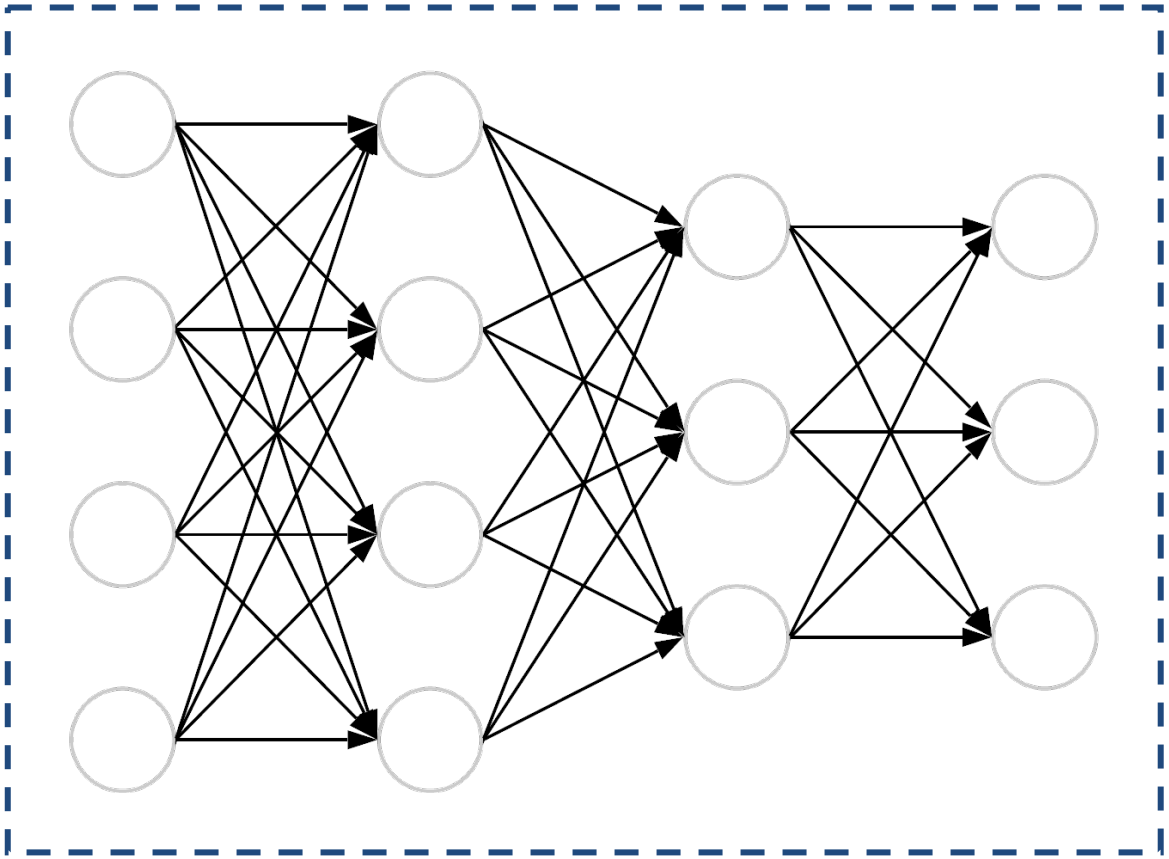


Are you sure this is a hot dog?



Input **x** → **Model** → Output **y**

x

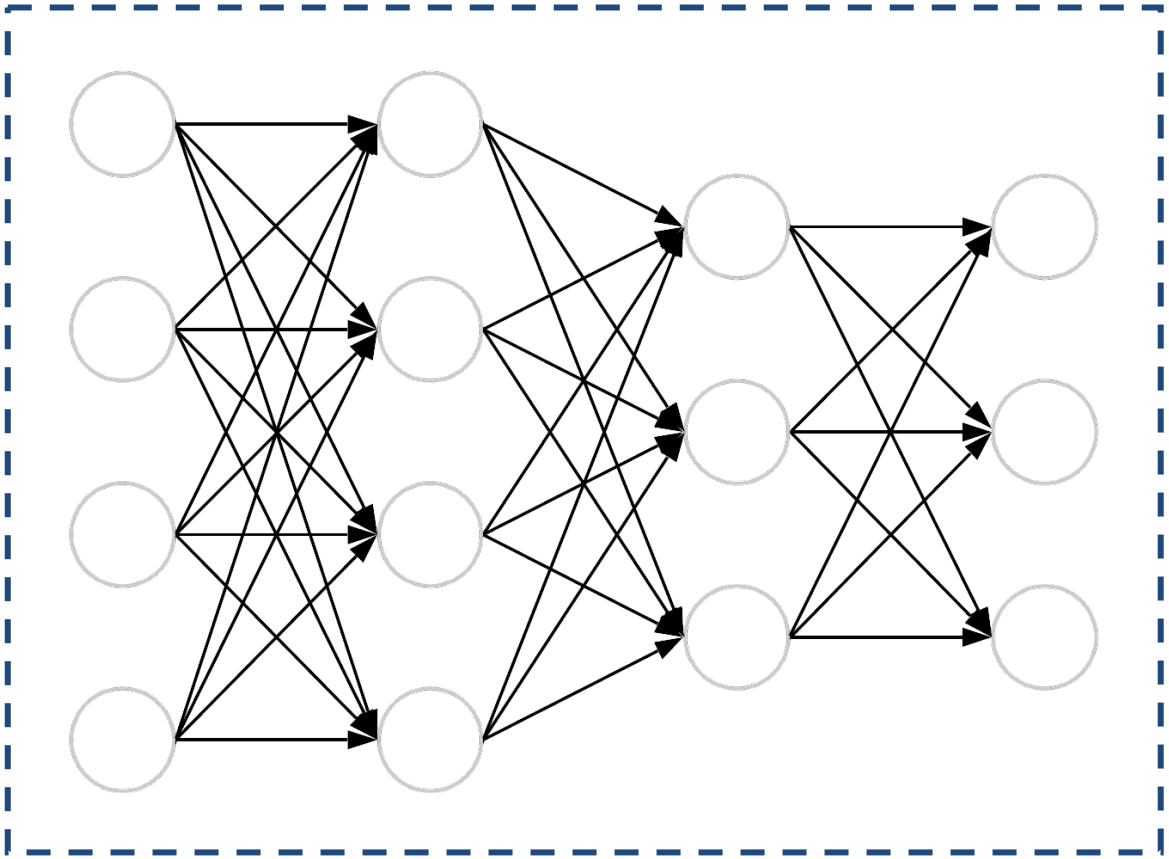


y

Dog	0.7
Cat	0.2
Bird	0.1

Input **x** → **Model** → Output **y**

x

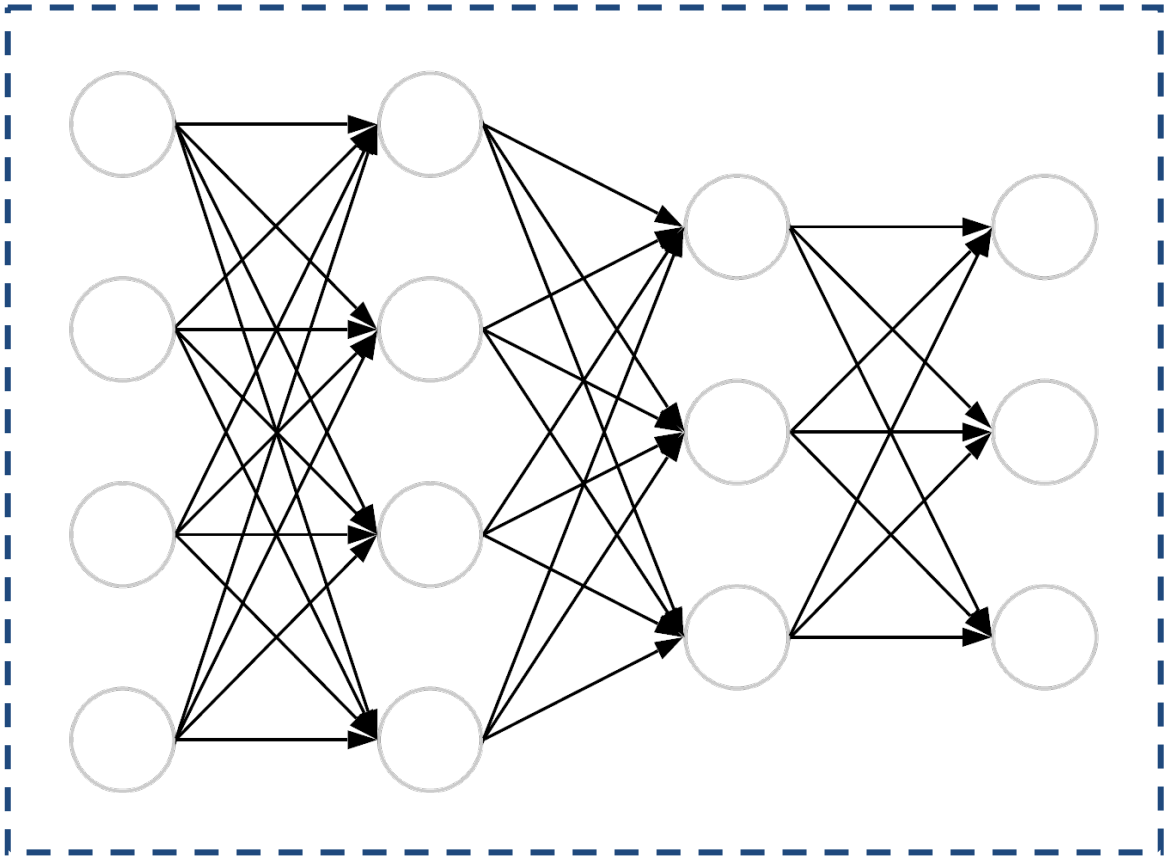


y

Dog	0.5
Cat	0.2
Bird	0.3

Input **x** → **Model** → Output **y**

x



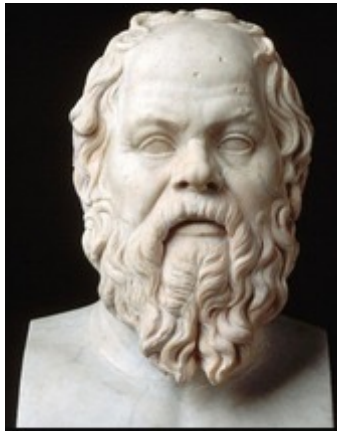
y

Dog	0.7
Cat	0.2
Bird	0.1

We cannot afford the cost from wrong predictions.

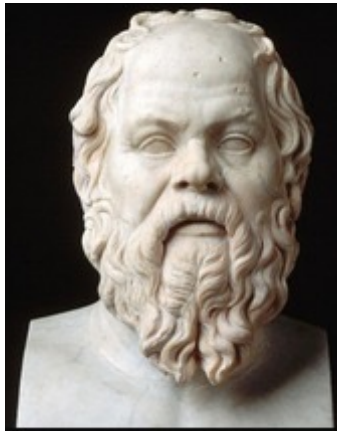
How to safely deal with uncertainty?

- Wisdom is knowing what you don’t know



~ Socrates

- Wisdom is knowing what you don't know



~ Socrates

Thomas Bayes ~

Bayesian methods

Outline

Introduction	Non-Bayesian method	Bayesian method	Non-parametric Bayesian
• Why we need uncertainty? • Basic Bayesian rules	• Linear Regression	• Bayesian Regression • Variational Inference	• Non-parametric method (GP) • Gaussian Process Regression

Outline

Introduction	Non-Bayesian method	Bayesian method	Non-parametric Bayesian
• Why we need uncertainty? • Basic Bayesian rules	• Linear Regression	• Bayesian Regression • Variational Inference	• Non-parametric method (GP) • Gaussian Process Regression

Basic Bayesian rules

- Bayesian rule/Inference

w – parameters

y – observations

Likelihood

Prior

Posterior

$$P(w|y) = \frac{P(y,w)}{P(y)} = \frac{P(y|w)P(w)}{P(y)}$$

Evidence

Basic Bayesian rules

- Bayesian rule/Inference

w – parameters

y – observations (labels)

X – Inputs

$$P(w|y) = \frac{P(y,w)}{P(y)} = \frac{P(y|w)P(w)}{P(y)}$$

Likelihood

Prior

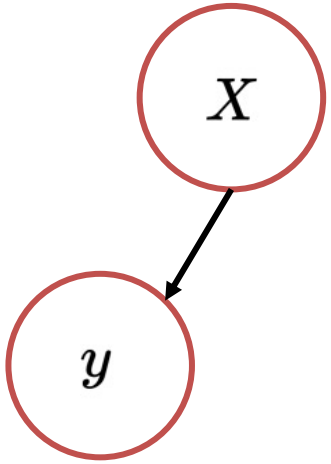
Posterior

$$P(w|y,X) = \frac{P(w,y|X)}{P(y|X)} = \frac{P(y|w,X)P(w)}{P(y|X)}$$

Evidence

Basic Bayesian rules

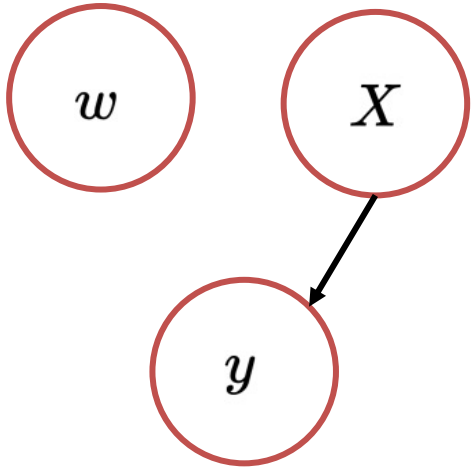
Basic machine learning problem:



where X, y is training data

Basic Bayesian rules

Basic machine learning problem:



Assume model:

$$y = f(w, X)$$

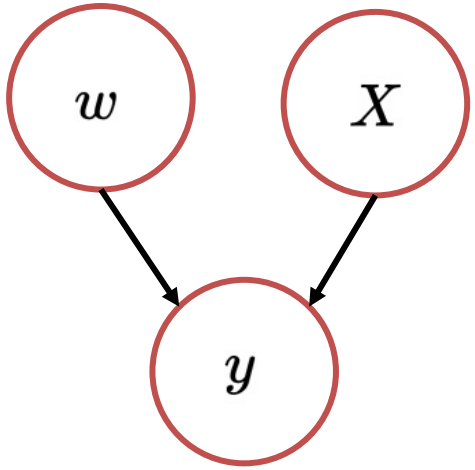
Prior knowledge about w:

$$P(w)$$

where w is model parameter

Basic Bayesian rules

Basic machine learning problem:

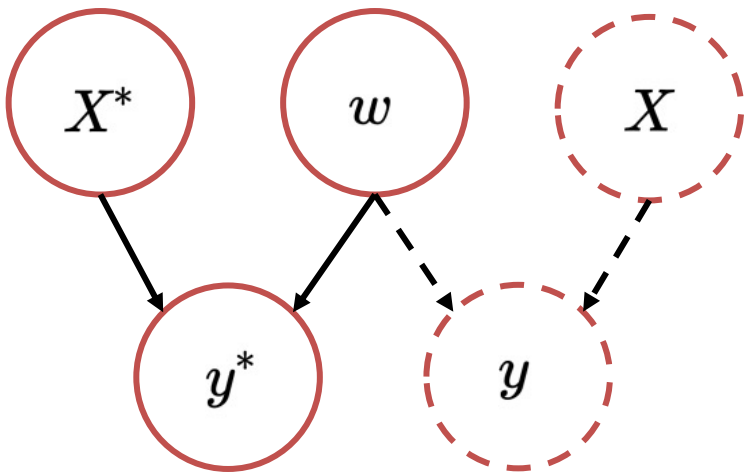


Use Bayesian rule to get posterior distribution of w

$$P(w|y,X) = \frac{P(w,y|X)}{P(y|X)} = \frac{P(y|w,X)P(w)}{P(y|X)}$$

Basic Bayesian rules

Basic machine learning problem:



Use posterior distribution of w for prediction

$$P(y^*|X^*, X, y) = \int P(y^*|X^*, w)P(w|X, y)dw$$

where X*, y* is test data

Outline

Introduction	Non-Bayesian method	Bayesian method	Non-parametric Bayesian
• Why we need uncertainty? • Basic Bayesian rules	• Linear Regression	• Bayesian Regression • Variational Inference	• Non-parametric method (GP) • Gaussian Process Regression

Linear regression

True model:

$$f(x) = 0.5 + \sin(2\pi x)$$

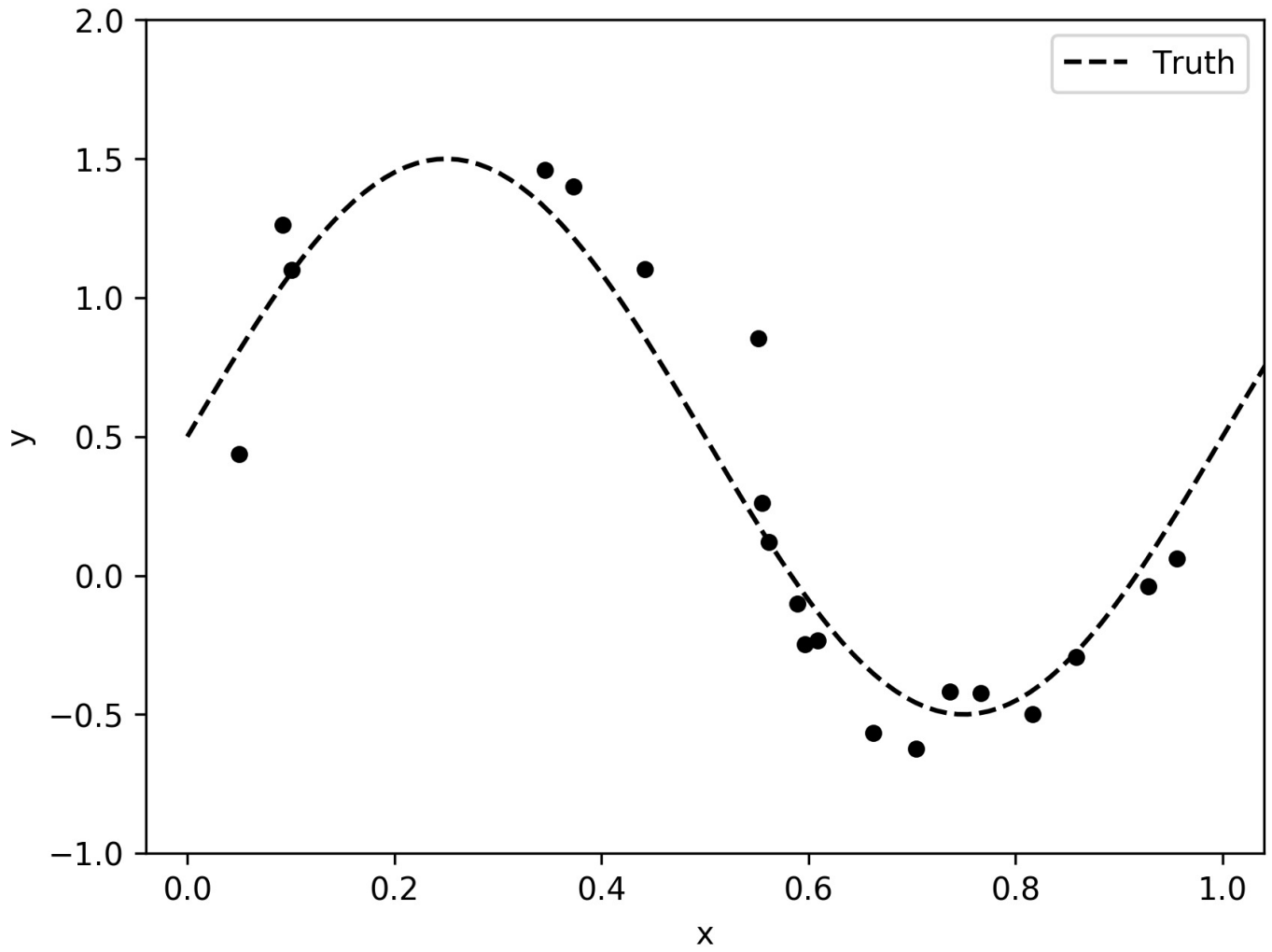
Observation:

$$y = 0.5 + \sin(2\pi x) + \epsilon$$

where

$$\epsilon \sim \mathcal{N}(0, \sigma_y^2)$$

N = 20 samples



Linear regression

Linear model

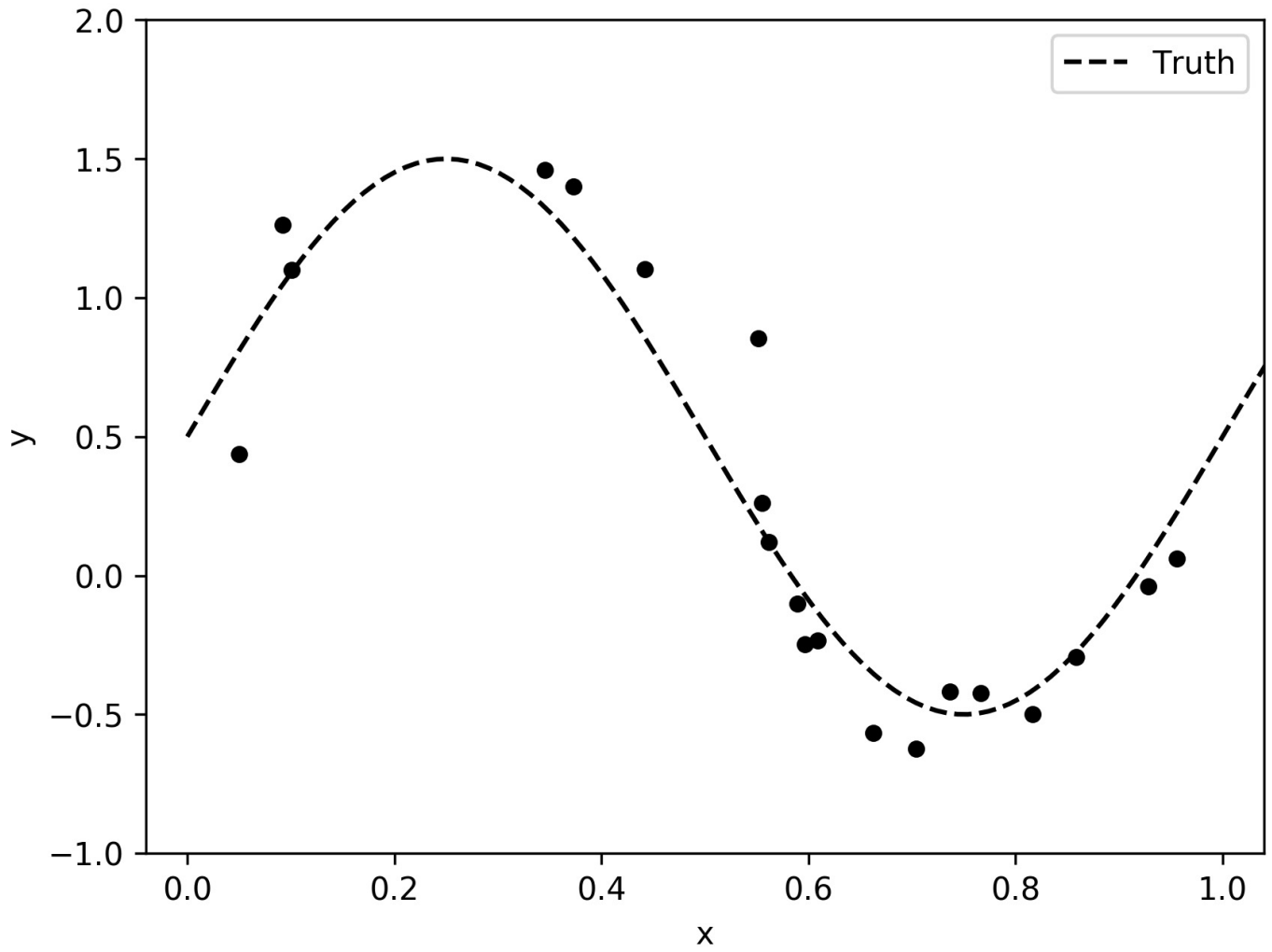
$$f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

where $\mathbf{w} = (w_0, \ldots, w_{M-1})^{\mathrm{T}}$

and $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_{M-1})^{\mathrm{T}}$

Observation

$$y = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}) + \epsilon$$



Basis functions $\phi(\mathbf{x})$

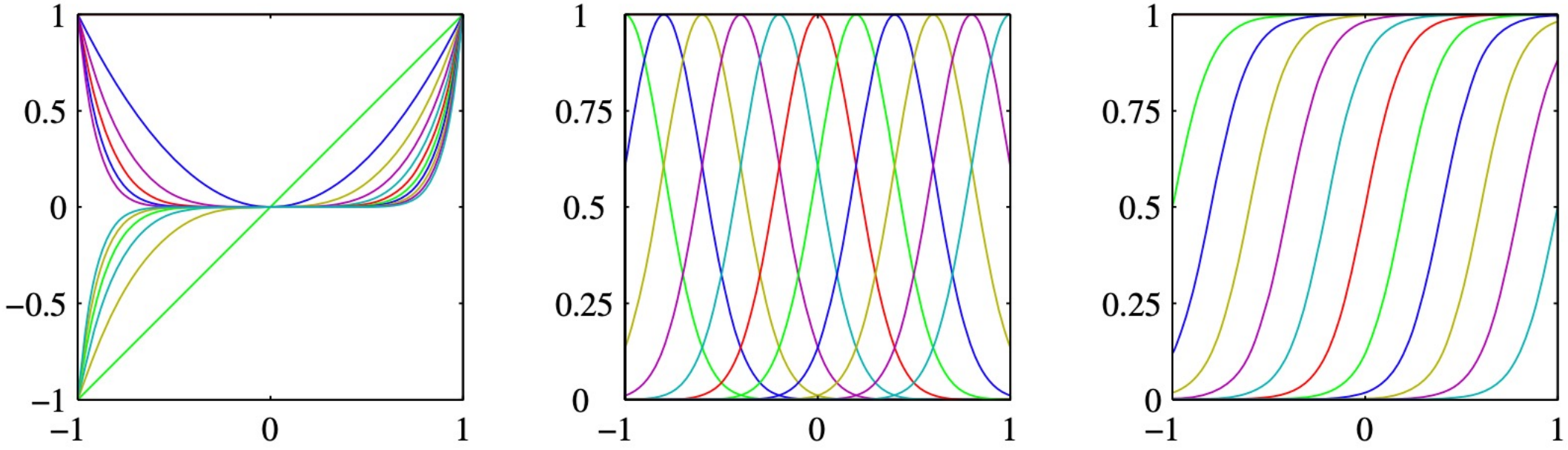


Figure 3.1 Examples of basis functions, showing polynomials on the left, Gaussians of the form (3.4) in the centre, and sigmoidal of the form (3.5) on the right.

Basis functions $\phi(\mathbf{x})$

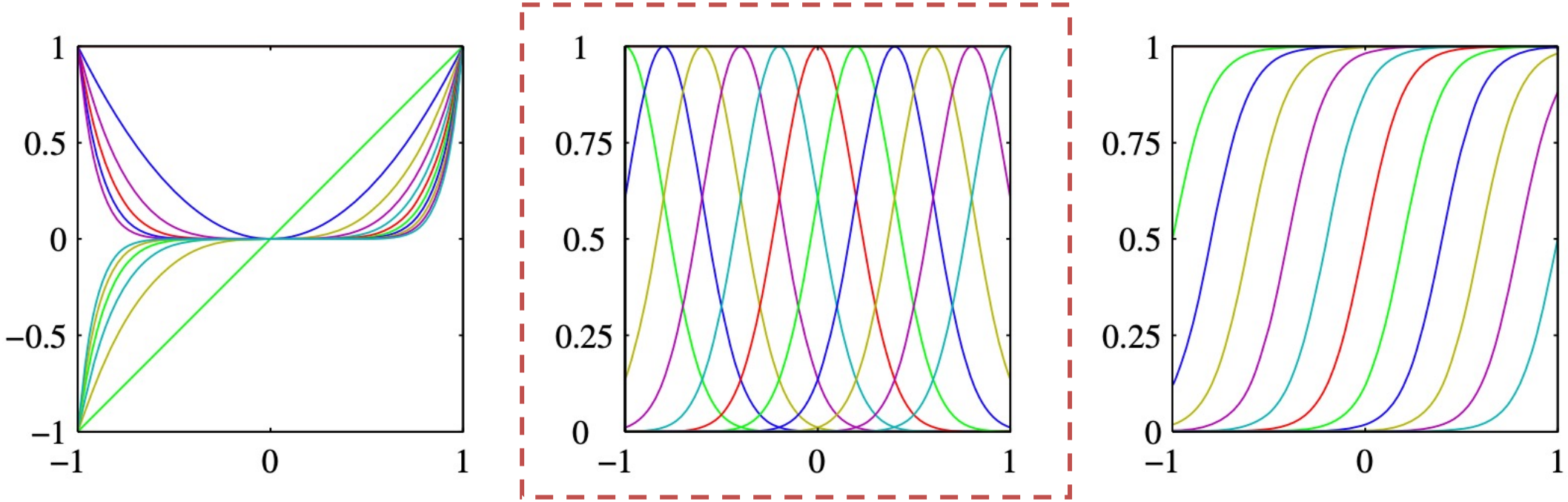


Figure 3.1 Examples of basis functions, showing polynomials on the left, Gaussians of the form (3.4) in the centre, and sigmoidal of the form (3.5) on the right.

Linear regression

Maximum likelihood estimator (MLE)

For MLE, maximize the likelihood :

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \ \ln P(\mathbf{y}|\mathbf{w}, \mathbf{X})$$

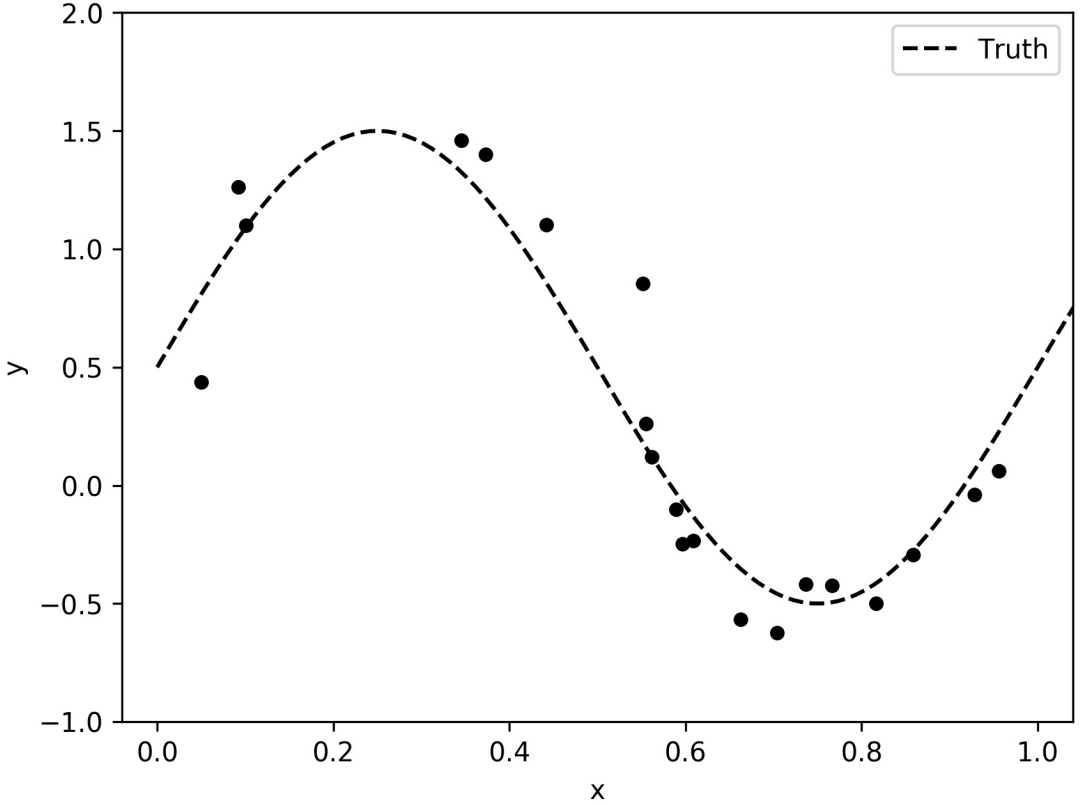
Because

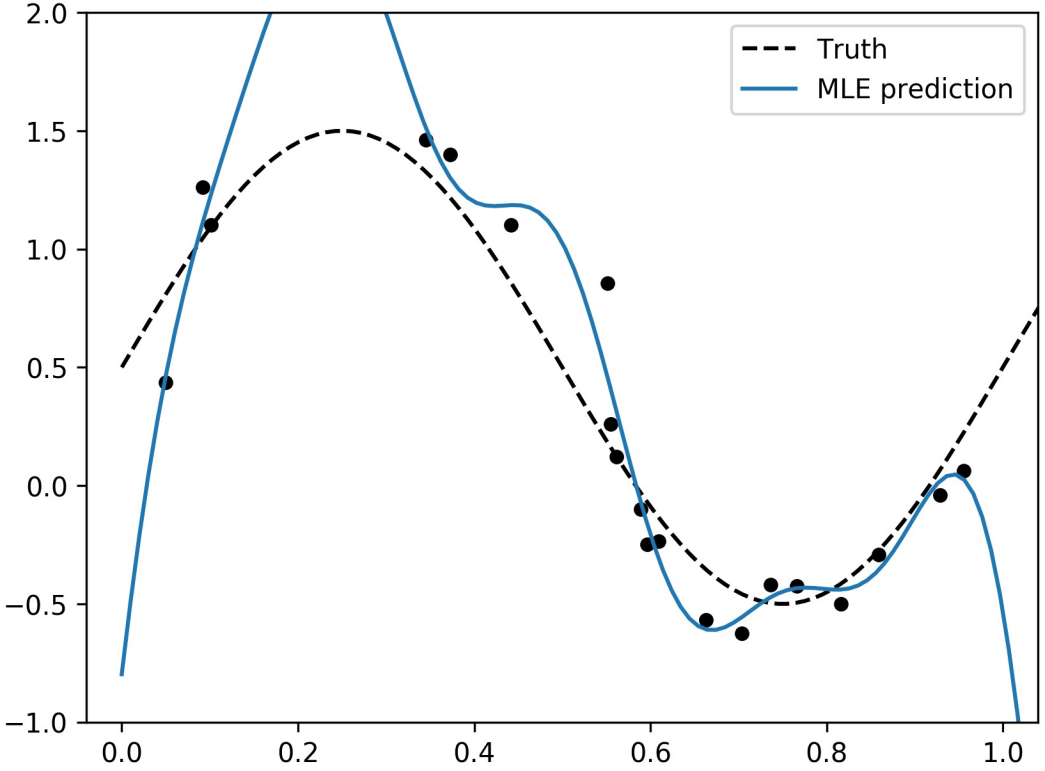
$$\ln p(\mathbf{y}|\mathbf{w}, \mathbf{X}) \propto - \| \mathbf{y} - \mathbf{\Phi}\mathbf{w} \|^2$$

Maximize likelihood is equivalent to **least squares** (if Gaussian)

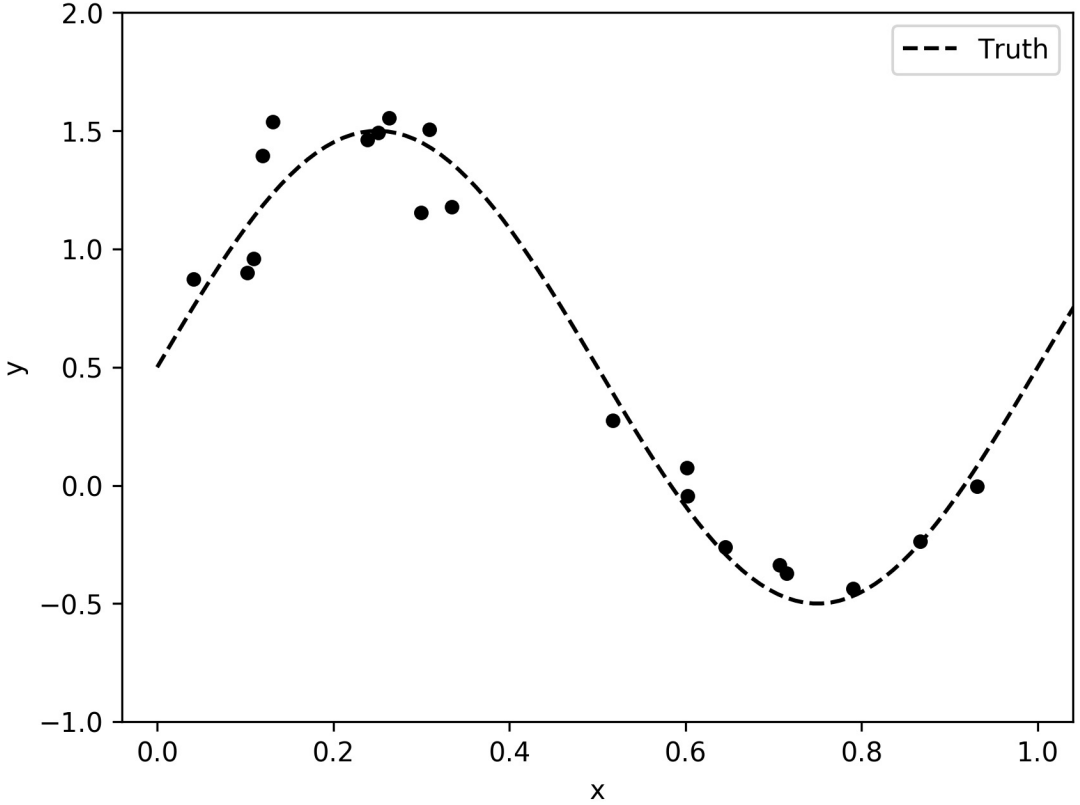
Linear regression

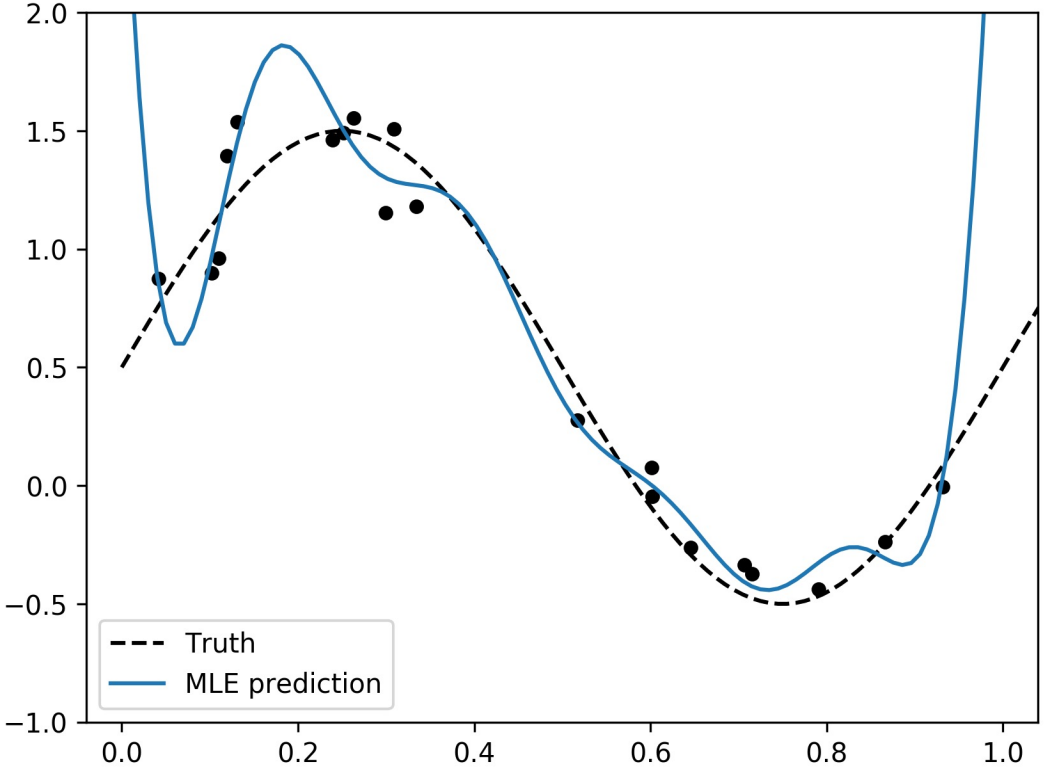
Maximum likelihood estimator (MLE)





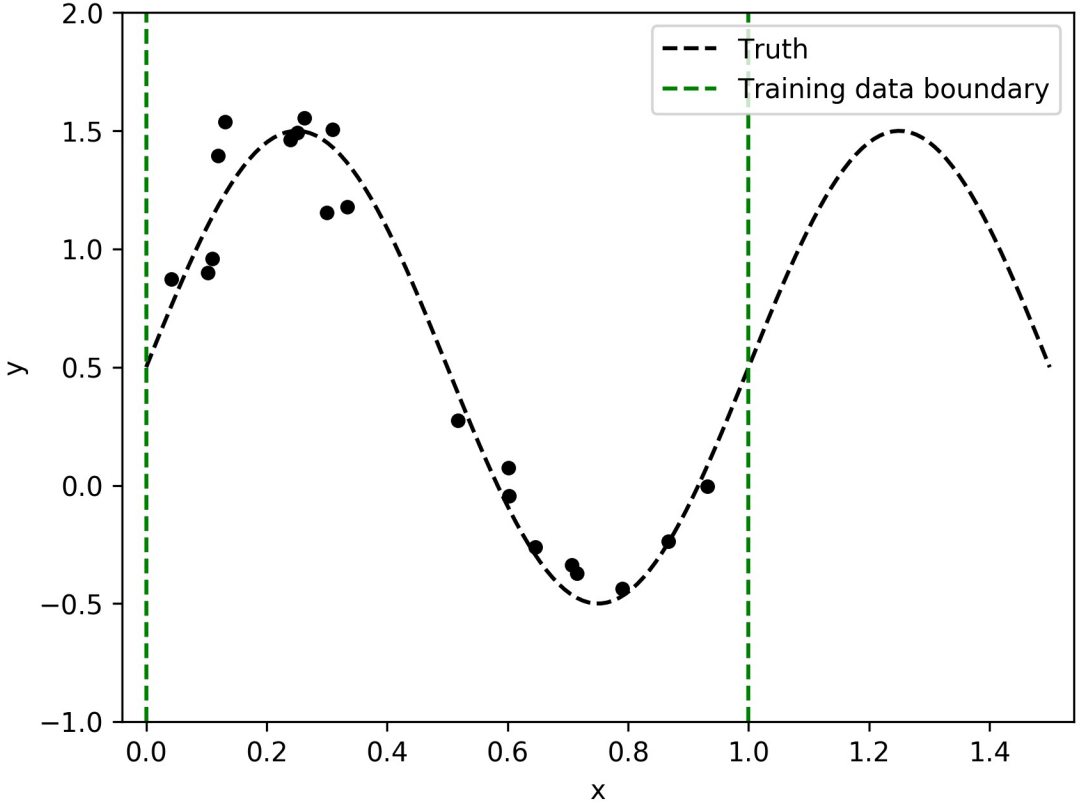
Maximum likelihood estimator (MLE) -- suffer from overfitting

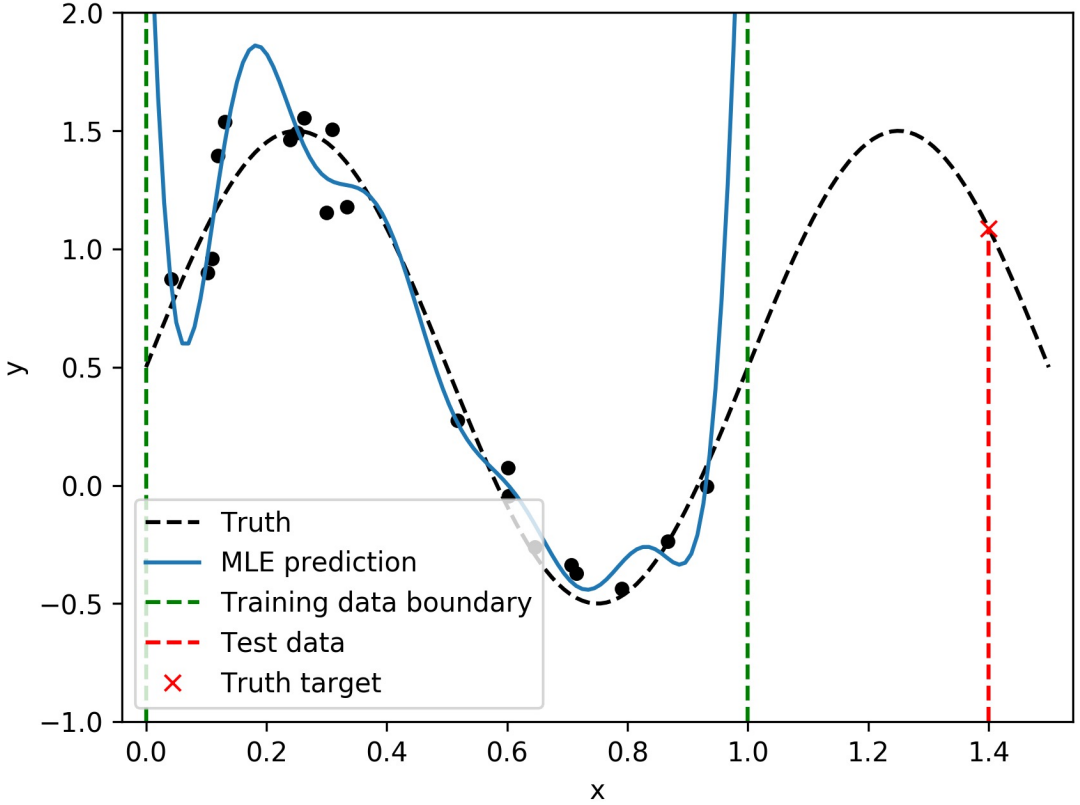




Linear regression

Maximum likelihood estimator (MLE) -- doesn't give prediction uncertainty





When $x^* = 1.4$, we have $y^*_{truth} = 0.5 + \sin(2\pi x^*) = 1.09$ $\quad y^* = \mathbf{w}_{ML}^{\mathrm{T}} \boldsymbol{\phi}(x^*) = 29.25$

Outline

Introduction	Non-Bayesian method	Bayesian method	Non-parametric Bayesian
• Why we need uncertainty? • Basic Bayesian rules	• Linear Regression	• Bayesian Regression • Variational Inference	• Non-parametric method (GP) • Gaussian Process Regression

Bayesian linear regression

Bayesian inference estimator (BIE)

Instead of maximize the likelihood:

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \ \ln P(\mathbf{w}, \mathbf{y}|\mathbf{X})$$

For BIE, we maximize the posterior :

$$\mathbf{w}_{B} = \arg\max_{\mathbf{w}} \ \ln P(\mathbf{w}|\mathbf{y}, \mathbf{X})$$

Bayesian equation

$$P(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{P(\mathbf{y}|\mathbf{w}, \mathbf{X})P(\mathbf{w})}{P(\mathbf{y}|\mathbf{X})}$$

Gaussian prior:

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

Bayesian linear regression

Bayesian inference estimator (BIE)

$$\mathbf{w}_B = \arg\max_{\mathbf{w}} \ln P(\mathbf{w}|\mathbf{y}, \mathbf{X})$$

Calculate log posterior:

$$\ln p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \propto - \| \mathbf{y} - \mathbf{\Phi}\mathbf{w} \|^2 - \lambda \| \mathbf{w} \|^2$$

Maximize posterior is equivalent to **regularized least squares** (if Gaussian)

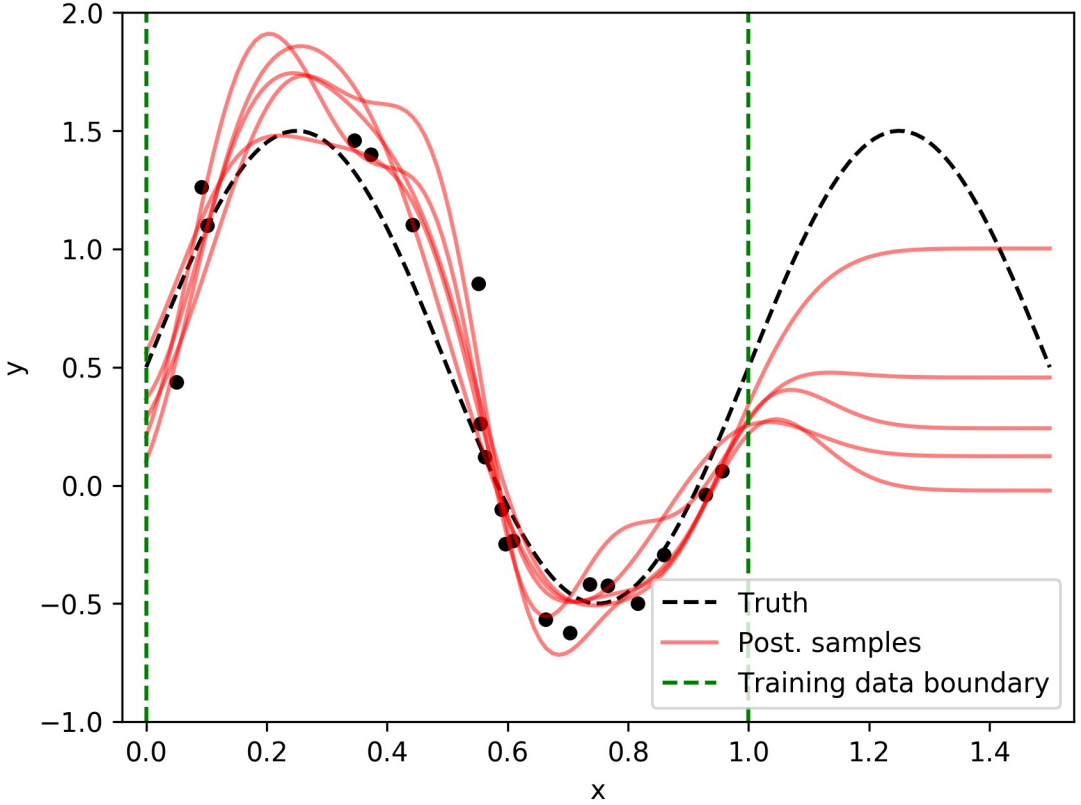
More importantly, the prediction can have a distribution:

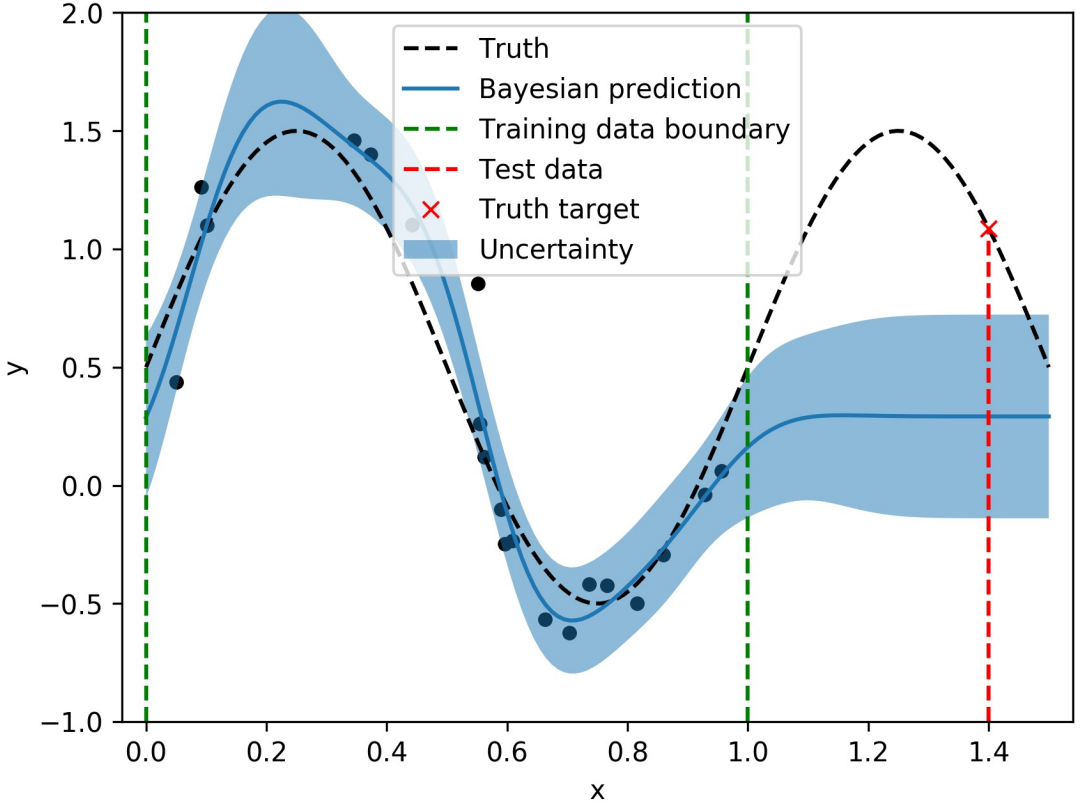
$$P(y^*|\mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(y^*|\mu^*, \mathbf{\Sigma}^*)$$

Compared with MLE: $y^* = \mathbf{w}_{ML}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}^*)$

Bayesian linear regression

Bayesian inference estimator (BIE)

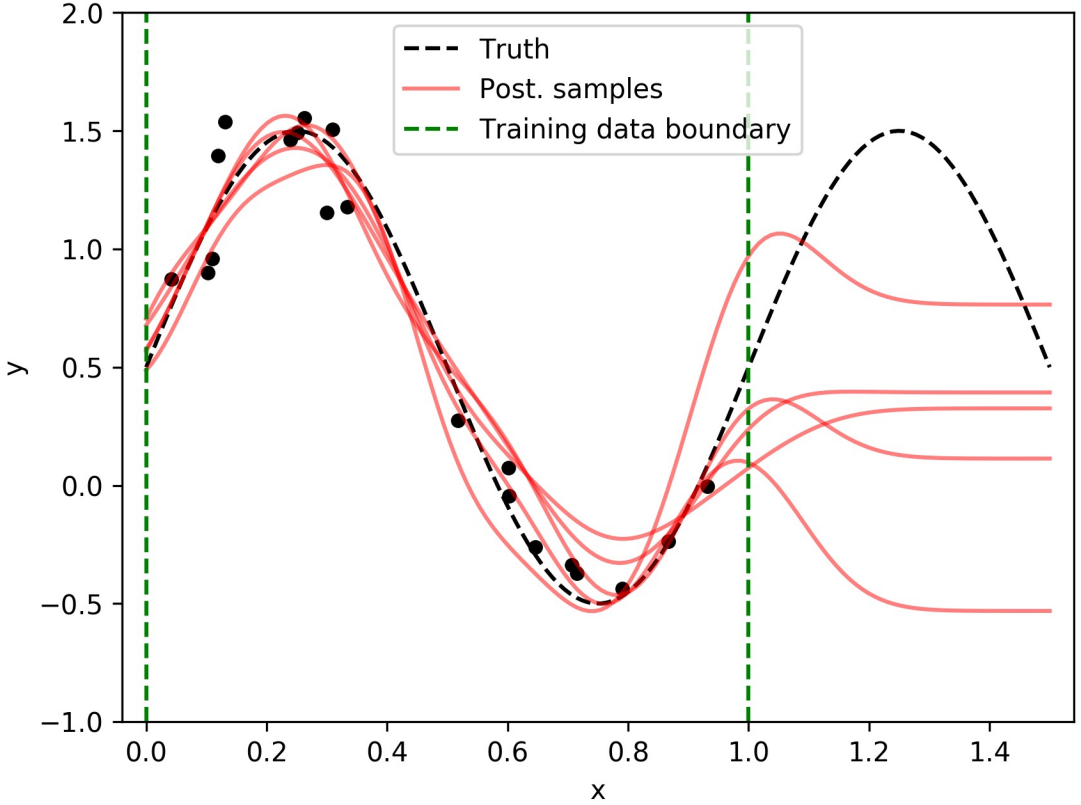


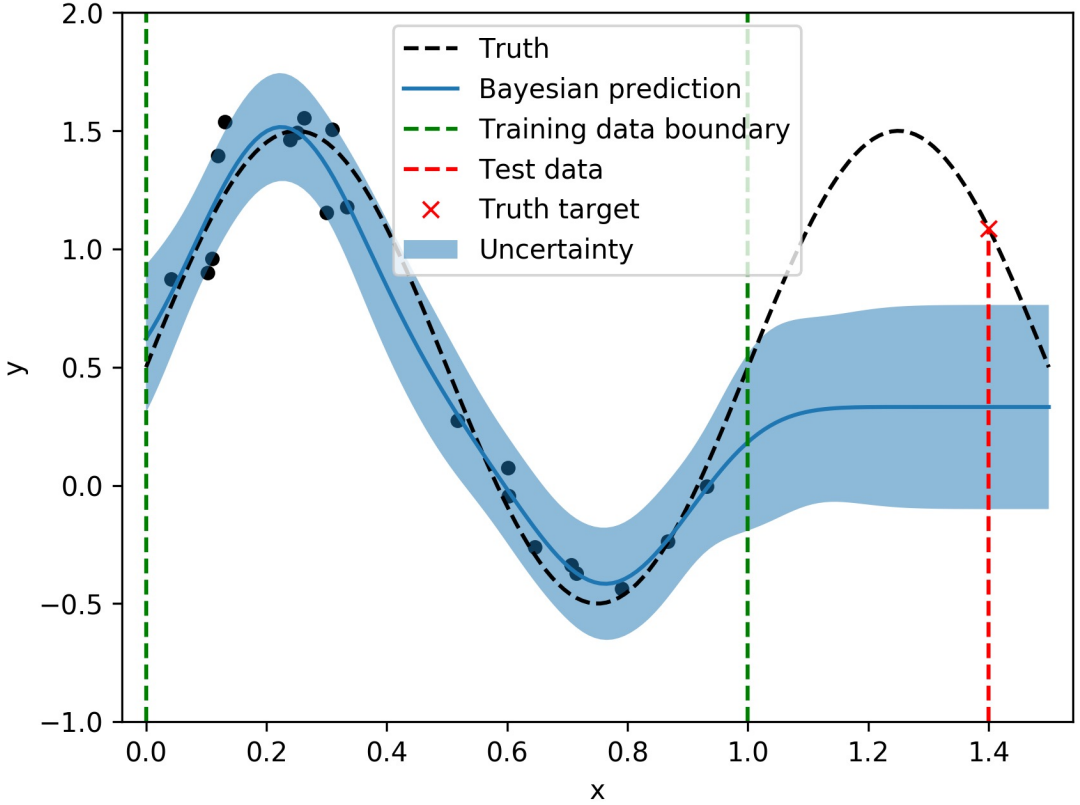


When $x^* = 1.4$, we have $y^*_{\text{truth}} = 1.09$, $E[y^*] = \mathbf{w}_B^{\mathrm{T}}\phi(\mathbf{x}^*) = 0.29$, $\sigma[y^*] = 0.43$

Bayesian linear regression

Bayesian inference estimator (BIE)

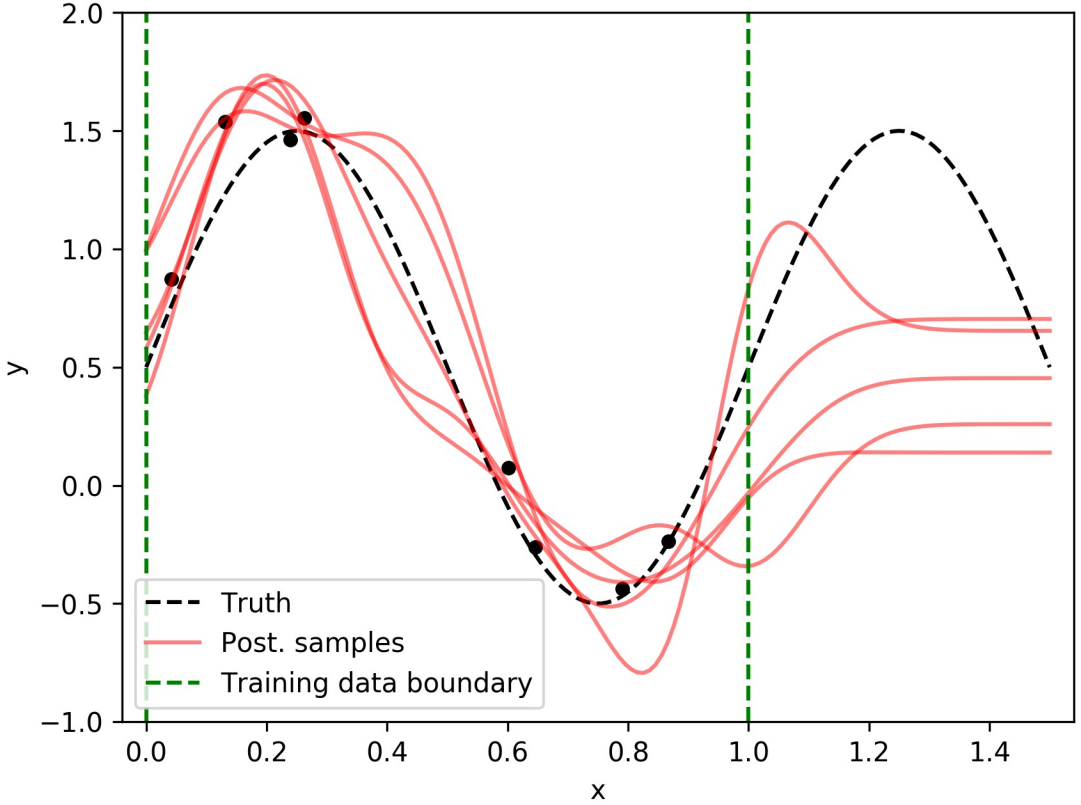


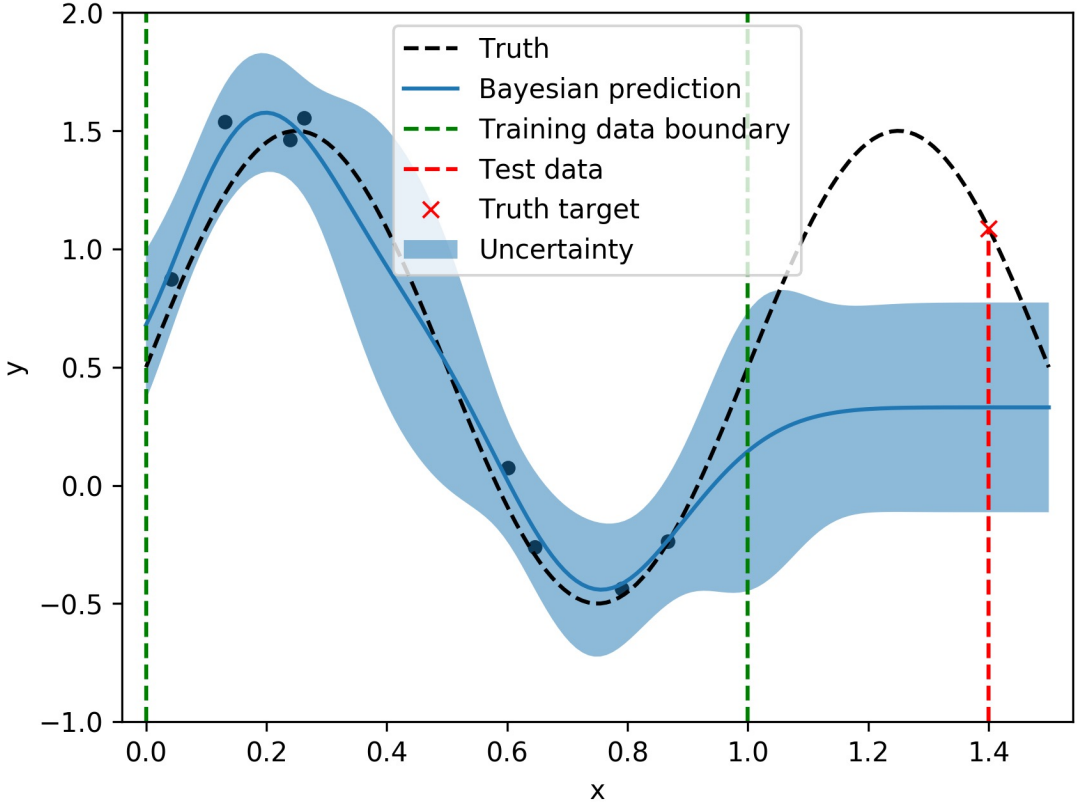


When $x^* = 1.4$, we have $y^*_{\text{truth}} = 1.09$, $E[y^*] = \mathbf{w}_B^{\mathrm{T}}\phi(\mathbf{x}^*) = 0.33$, $\sigma[y^*] = 0.43$

Bayesian linear regression

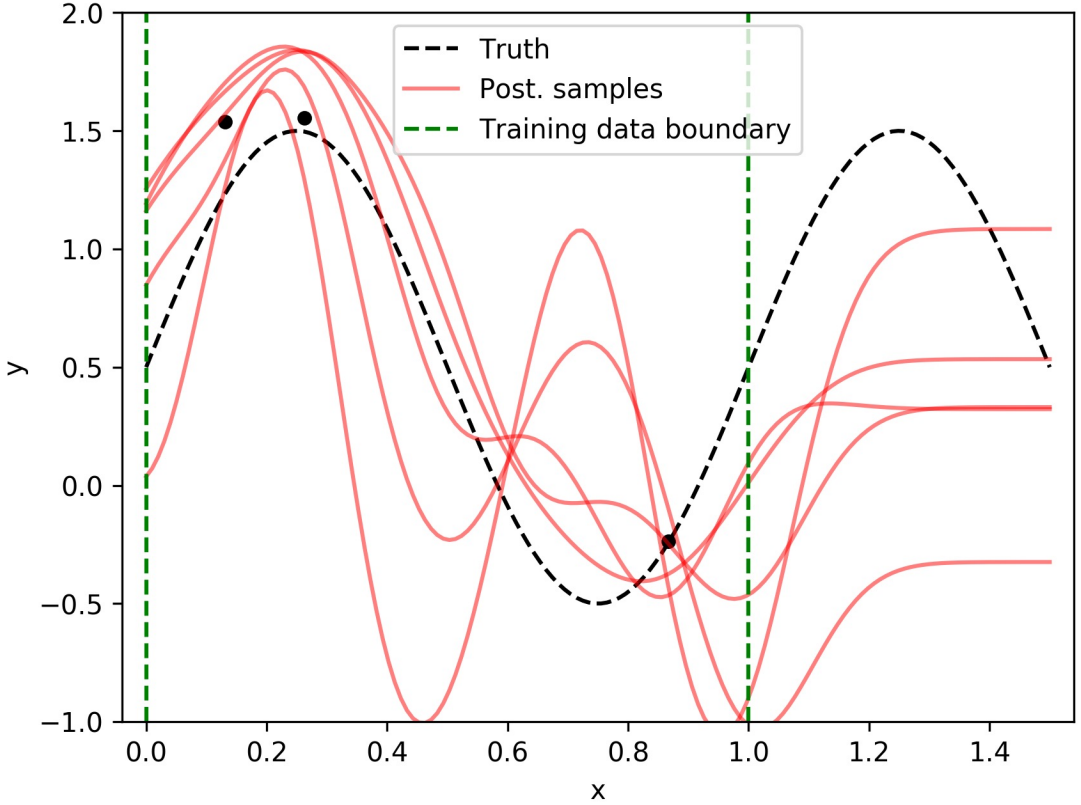
Reduce the number of samples (N = 8)

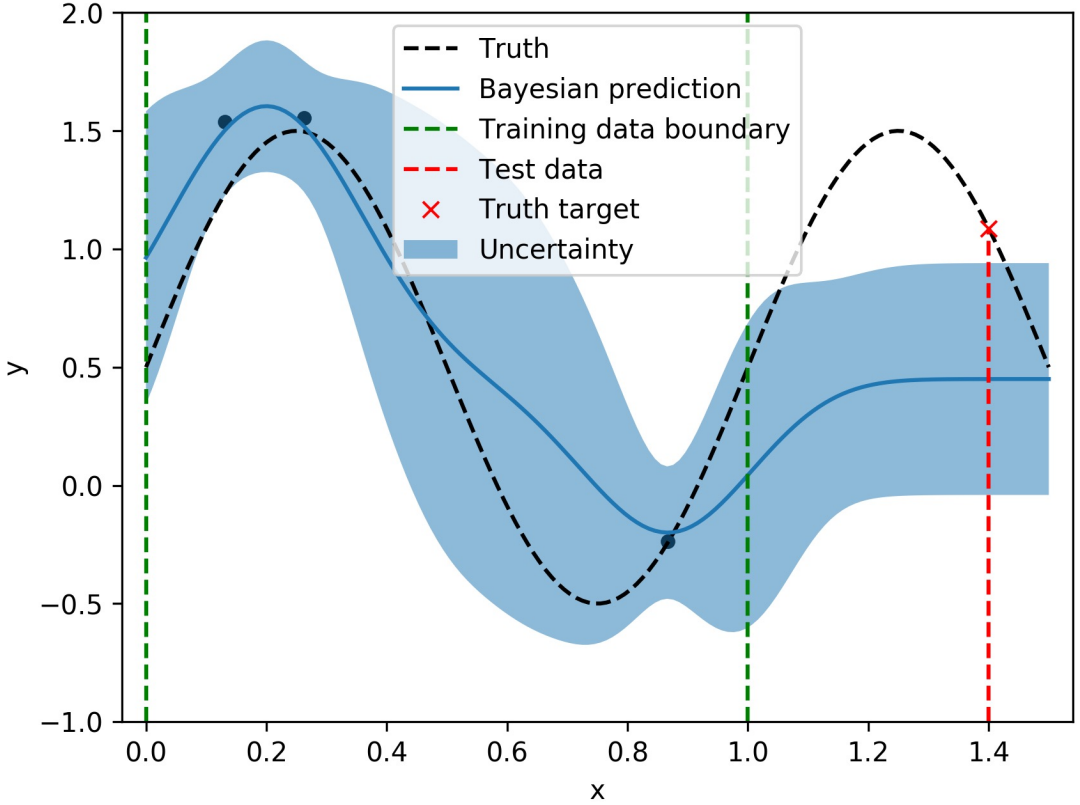




Bayesian linear regression

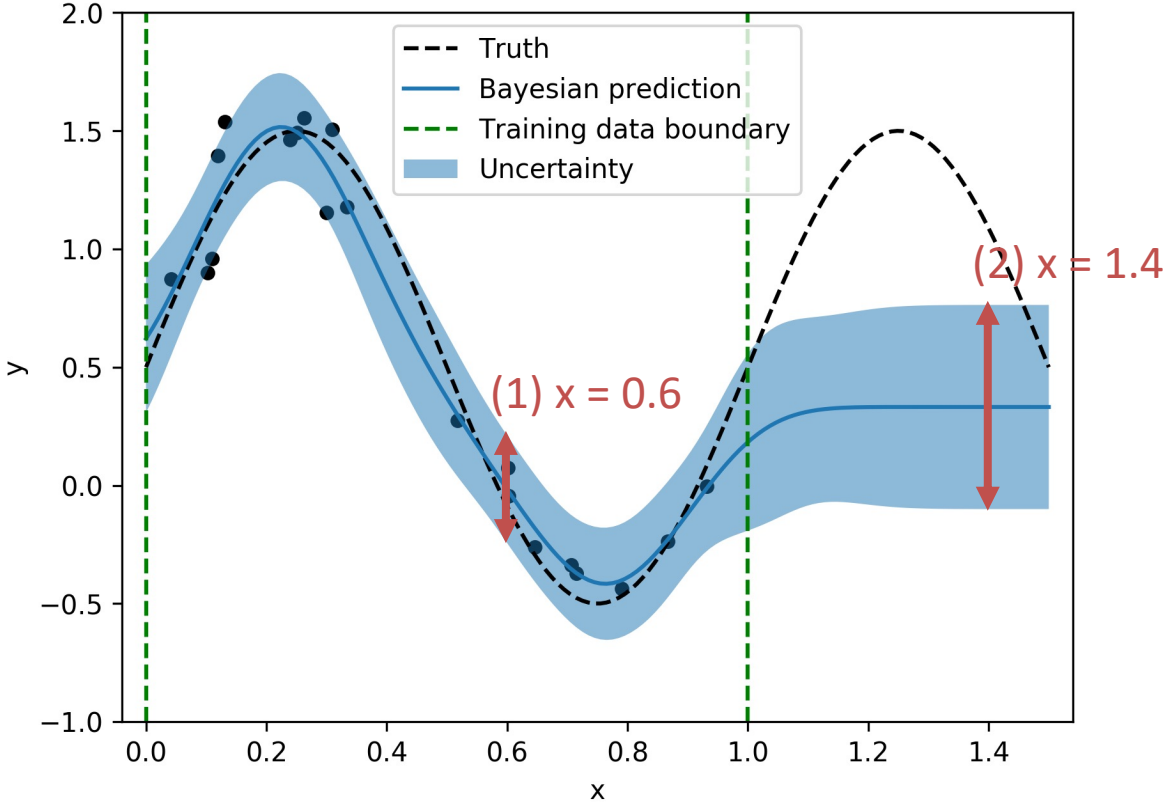
Reduce the number of samples (N = 3)





Bayesian linear regression

Uncertainty analysis



1) x = 0.6 is in the training distribution

2) x = 1.4 is out of training distribution

$$Var(x_1) < Var(x_2)$$

Bayesian linear regression

On-line learning

w – parameters

y – observations

Likelihood

Prior

New Prior

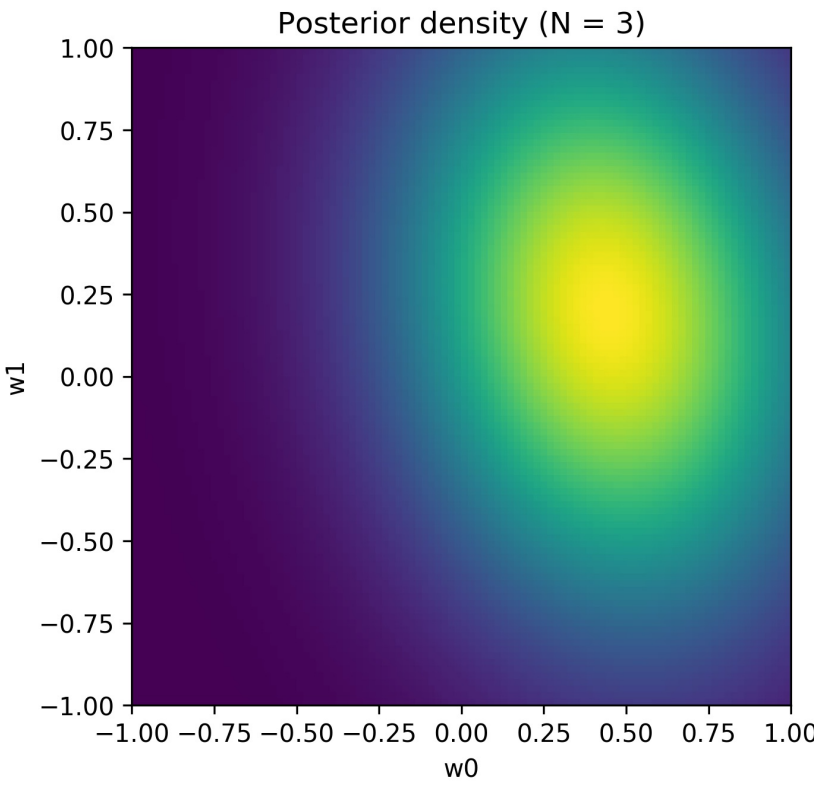
$$P_k(w) = P(w|y_k) = \frac{P(y_k|w)P_{k-1}(w)}{P(y_k)}$$

Posterior

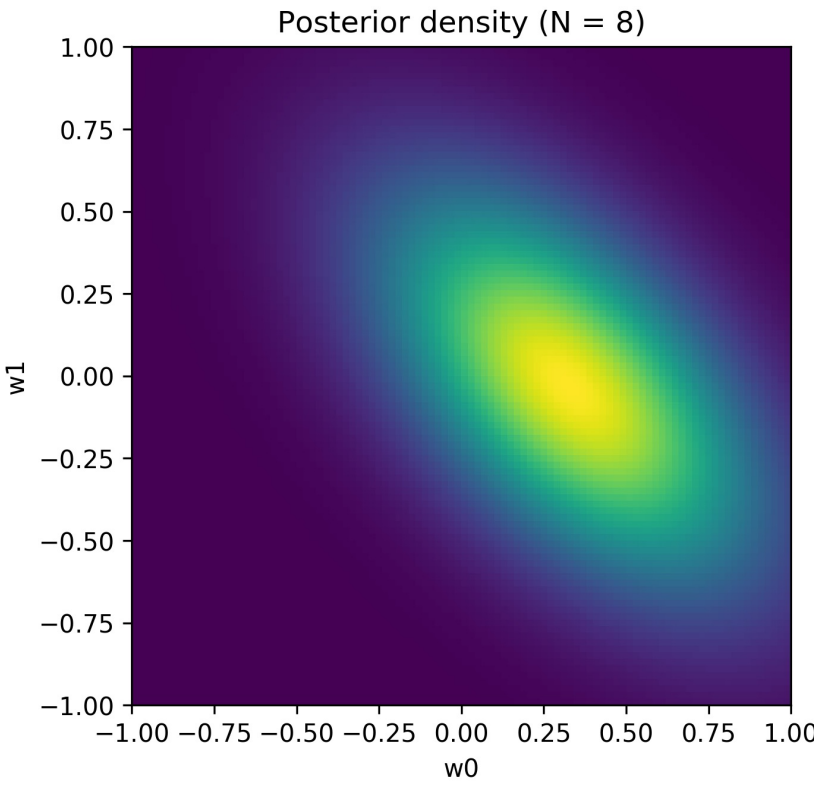
Evidence

Bayesian linear regression

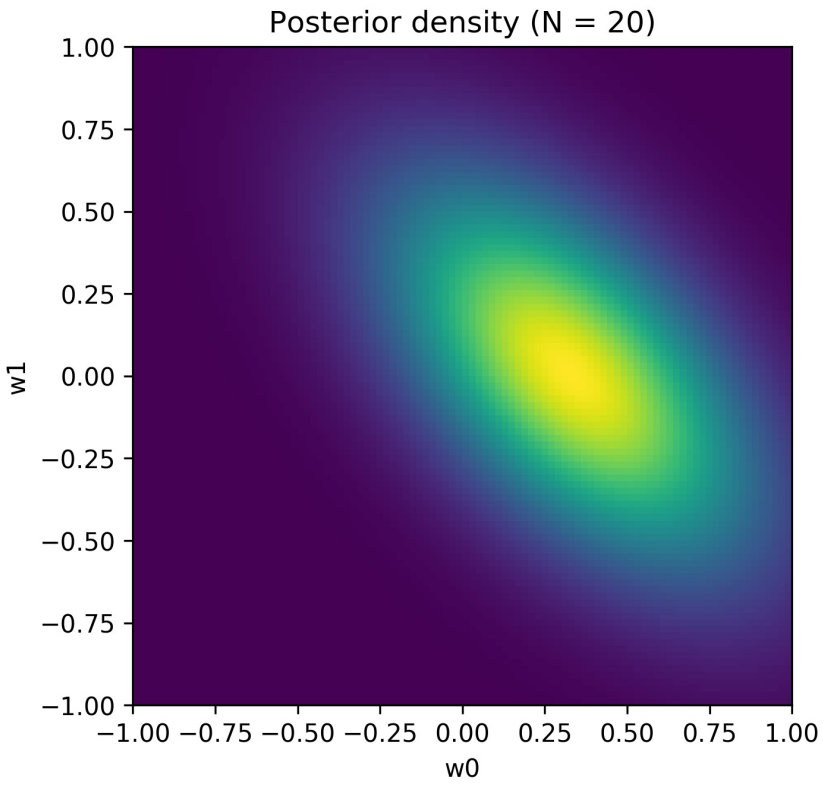
Evolving of posterior of the weights $p(\mathbf{w}|\mathbf{y},\mathbf{X})$ -- only show w_0, w_1 here



N = 3



N = 8



N = 20

Classical ML vs. Bayesian ML

Classical Machine Learning

1. **Starting point:** Set of possible models (w)
2. **Optimization:** Maximize the likelihood ($p(y|w, X)$)
3. **Result:** (usually) is a deterministic function
4. **Prediction:** (usually) by deterministic function call

Bayesian Machine Learning

1. **Starting point:** Distribution of possible models ($p(w)$)
2. **Optimization:** Maximize the posterior ($p(w|y, X)$)
3. **Result:** (usually) is an updated model distribution
4. **Prediction:** (usually) averaging model distribution

Classical ML vs. Bayesian ML

Classical Machine Learning

5. **Variance:** (may) suffer from overfitting
6. **Learning:** Offline learning only (Include all data for learning)

Bayesian Machine Learning

5. **Variance:** (usually) smaller, because of regularization term in optimization
6. **Learning:** can be used for online learning (only use new data for learning)

Summary

Pros:

1. Give us a distribution of prediction
2. Prevent overfitting problem
3. Can use on-line learning to update model gradually

Cons:

1. Computational intensive (especially when we can not avoid calculating the evidence, and sometimes the posterior is intractable)

Why we use variational inference?

Why the posterior is difficult to compute?

Fixed by model

Our own choice

$$p(w|y) = \frac{p(y|w)p(w)}{p(y)} = \frac{p(y|w)p(w)}{\int p(y|w)p(w)dw}$$

Fixed by data

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

Neural networks

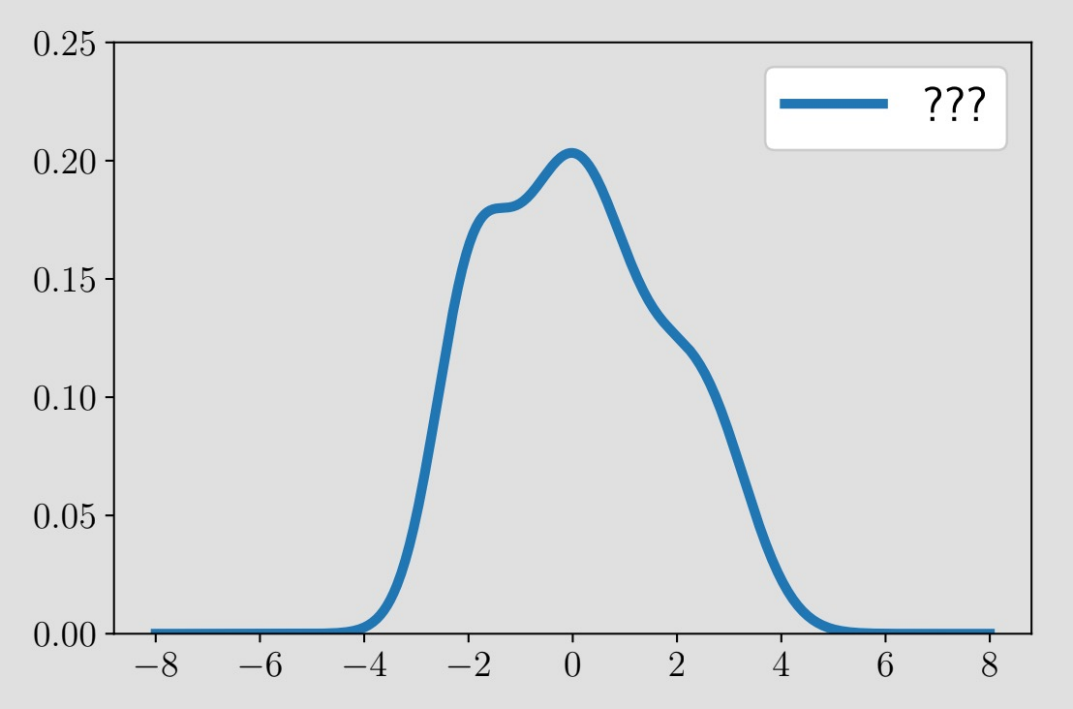
- It is intractable to compute $p(y)$ because it is impossible to consider all configurations w of the neural network.

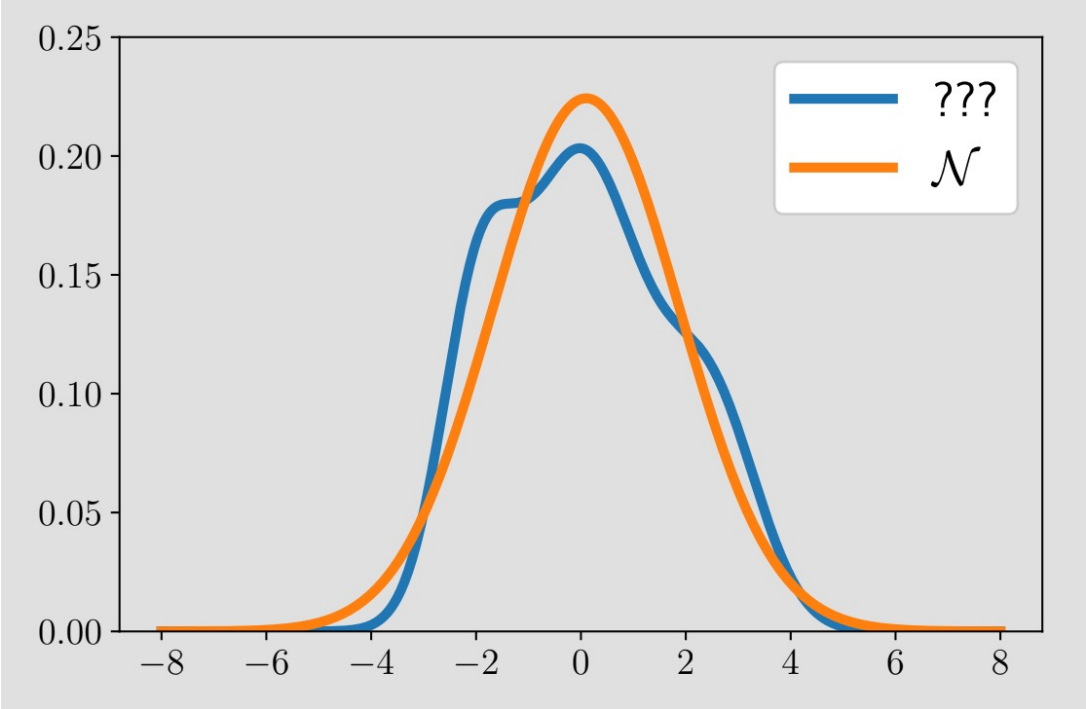
Outline

Introduction	Non-Bayesian method	Bayesian method	Non-parametric Bayesian
• Why we need uncertainty? • Basic Bayesian rules	• Linear Regression	• Bayesian Regression • Variational Inference	• Non-parametric method (GP) • Gaussian Process Regression

Variational inference

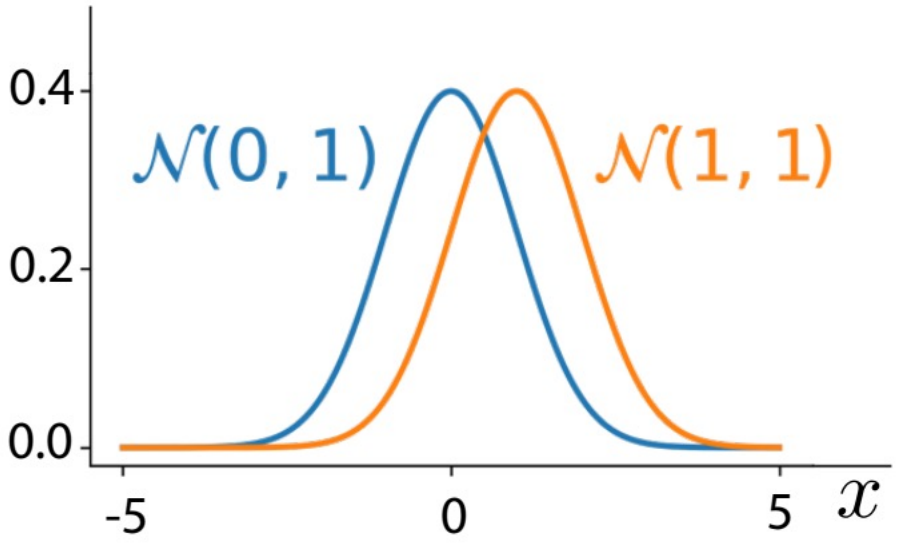
When the posterior is difficult to compute, why we just use an approximation to speed up the process? $q(w) \rightarrow p^*(w|y)$, $q(w) \in Q$



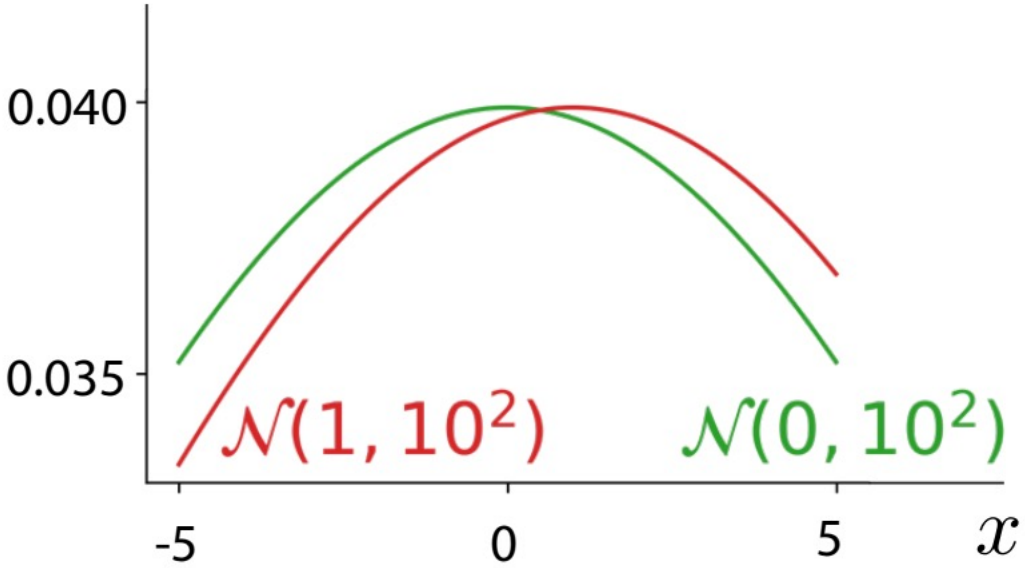


Kullback-Leibler (KL) divergence

Parameters difference: 1



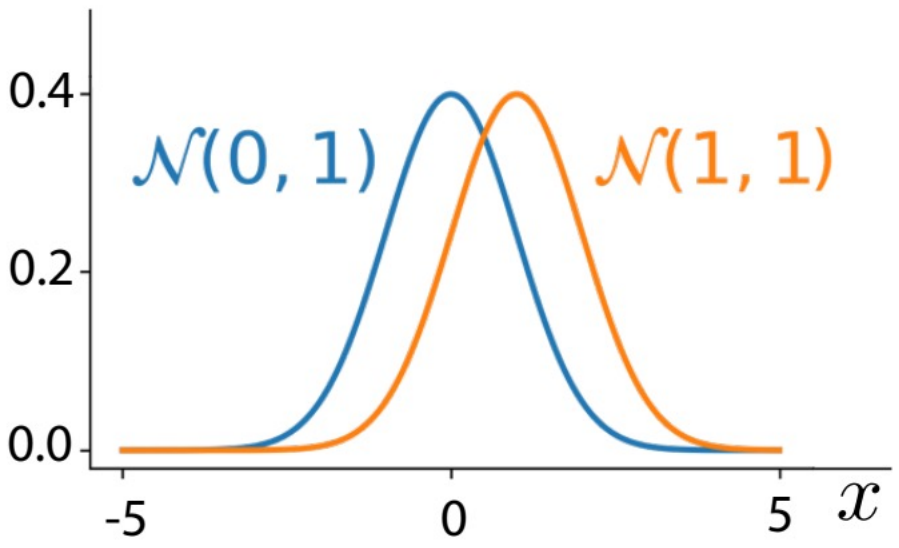
Parameters difference: 1



Kullback-Leibler (KL) divergence

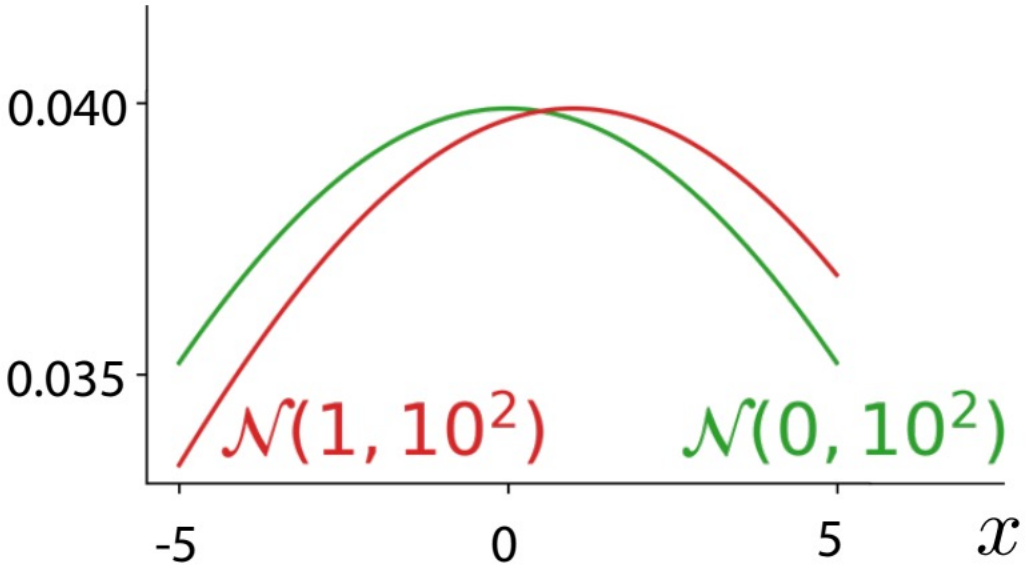
Parameters difference: 1

$$\mathcal{KL}(q_1 \parallel p_1) = 0.5$$



Parameters difference: 1

$$\mathcal{KL}(q_2 \parallel p_2) = 0.005$$



$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx \geqslant 0$$

Kullback-Leibler (KL) divergence

Numerical example:

$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx \geqslant 0$$

If $q(x) = p(x)$, $\mathcal{KL}(q \parallel p) = 0$

For Bernoulli distribution:

If $p(x) = \begin{cases} 0.5 & , x = 1 \\ 0.5 & , x = 0 \end{cases}$, when $q(x) = \begin{cases} 0.4 & , x = 1 \\ 0.6 & , x = 0 \end{cases}$; when $q(x) = \begin{cases} 0.2 & , x = 1 \\ 0.8 & , x = 0 \end{cases}$

$$\mathcal{KL}(q \parallel p) = 0.020 \qquad \mathcal{KL}(q \parallel p) = 0.193$$

Evidence Lower Bound (ELBO)

Log evidence

$$\ln p(\mathbf{y}) = ELBO + KL(q(\mathbf{w}) \parallel p(\mathbf{w}|\mathbf{y})) \geqslant ELBO$$

where

$$\boxed{ELBO = \int_{\mathbf{w}} q(\mathbf{w}) \ln \frac{p(\mathbf{y}, \mathbf{w})}{q(\mathbf{w})} d\mathbf{w}}$$

Evidence is fixed by data, thus $\min_q KL(q(\mathbf{w}) \parallel p(\mathbf{w}|\mathbf{y})) \rightarrow \max_q ELBO$

Evidence Lower Bound (ELBO)

Maximize ELBO

$$
\begin{aligned}
ELBO &= \int_{\mathbf{w}} q(\mathbf{w}) \ln \frac{p(\mathbf{y}, \mathbf{w})}{q(\mathbf{w})} d\mathbf{w} \\
&= \int_{\mathbf{w}} q(\mathbf{w}) \ln p(\mathbf{y}|\mathbf{w}) - q(\mathbf{w}) \ln \frac{q(\mathbf{w})}{p(\mathbf{w})} d\mathbf{w} \\
&= E_q[\ln p(\mathbf{y}|\mathbf{w})] - KL(q(\mathbf{w}) \parallel p(\mathbf{w}))
\end{aligned}
$$

q(w) – variational distribution

p(w|y) – true posterior distribution

p(y|w) – likelihood

p(w). – prior distribution

Evidence Lower Bound (ELBO)

Maximize ELBO

$$
\begin{aligned}
ELBO &= \int_{\mathbf{w}} q(\mathbf{w}) \ln \frac{p(\mathbf{y}, \mathbf{w})}{q(\mathbf{w})} d\mathbf{w} \\
&= \int_{\mathbf{w}} q(\mathbf{w}) \ln p(\mathbf{y}|\mathbf{w}) - q(\mathbf{w}) \ln \frac{q(\mathbf{w})}{p(\mathbf{w})} d\mathbf{w} \\
&= E_q[\ln p(\mathbf{y}|\mathbf{w})] - KL(q(\mathbf{w}) \parallel p(\mathbf{w}))
\end{aligned}
$$

Samples from $q(w)$ to perform original tasks

Regularization term

Evidence Lower Bound (ELBO)

Basic ideas: maximize ELBO to use $q(w)$ to approximate $p^*(w|y)$,

- During training:

  Train samples from $q(w)$ to perform original tasks and penalize $q(w)$ for differing from prior distribution $p(w)$

- During prediction:

  Sample w from $q(w)$ to do the prediction (use it as true posterior distribution)

Mean field variational inference

As **w** usually is high dimensional, mean field variational inference is a further simplification of variational distributions $q(\mathbf{w})$, by considering each dimension independently:

$$q(\mathbf{w}) = \prod_j q_j(w_j)$$

Example: $\mathcal{N}(\mu, \begin{pmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ 0 & & \ddots & \\ & & & \sigma_d^2 \end{pmatrix})$

Mean field variational inference

Example:

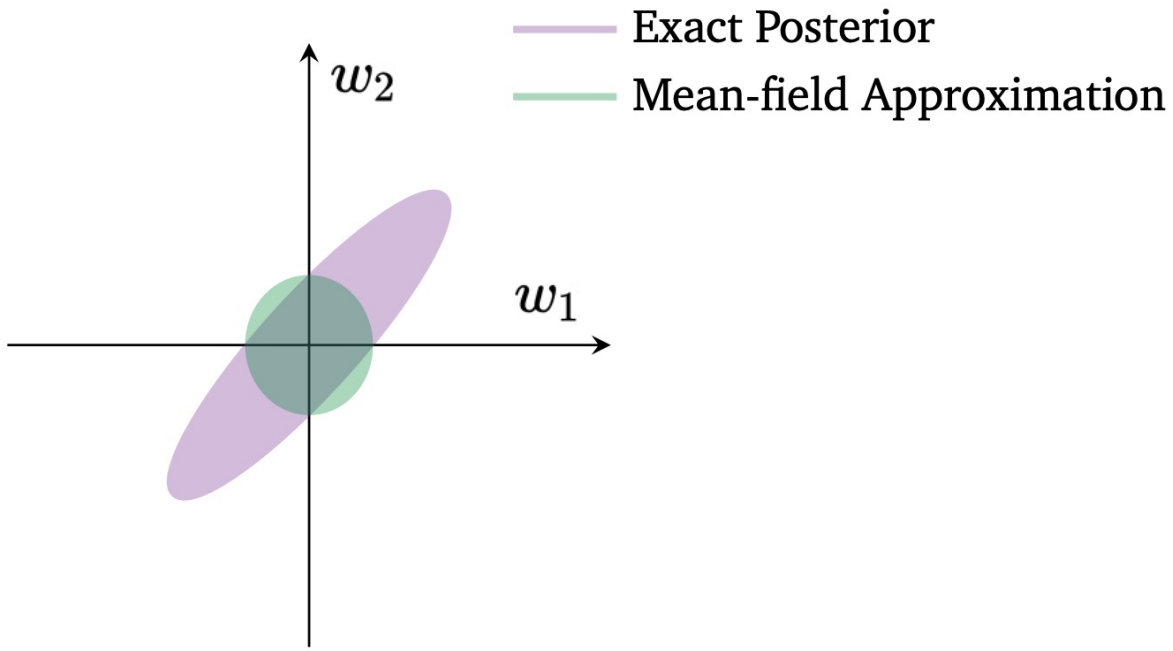


Figure 1: Visualizing the mean-field approximation to a two-dimensional Gaussian posterior. The ellipses show the effect of mean-field factorization. (The ellipses are 2σ contours of the Gaussian distributions.)

Problem of parametric method

The problem of linear regression:

Which basis functions to choose? How many parameters should we have?

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

Parameters → Model complexity

Basis function → Function form

If the assumption is wrong, more training data is not going to help at the end.

Outline

Introduction	Non-Bayesian method	Bayesian method	Non-parametric Bayesian
• Why we need uncertainty? • Basic Bayesian rules	• Linear Regression	• Bayesian Regression • Variational Inference	• Non-parametric method (GP) • Gaussian Process Regression

Parametric vs. Non-parametric

Parametric Method:

- Directly simplify the mapping function to a known form.
- Number of parameters is fixed.

Non-parametric Method:

- Do not make strong assumptions about the form of the mapping function, but about the correlations between different input.
- Model complexity grows with the size of training data.

Gaussian process

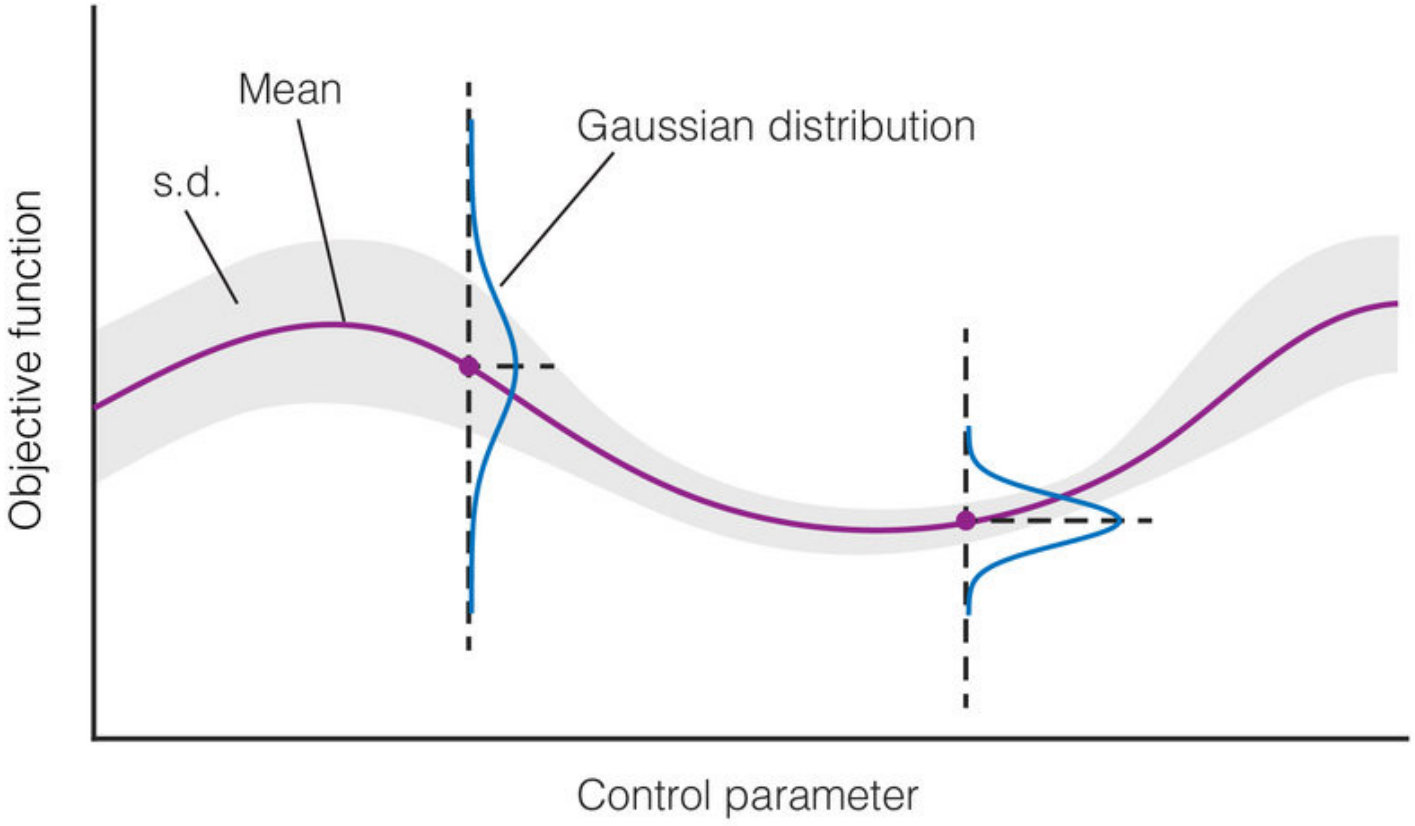
What is Gaussian process?

It can be seen as an infinite-dimensional generalization of multivariate normal distributions (any finite subset of which are Gaussian distributed).

Gaussian distribution	Gaussian Process
$x \sim \mathcal{N}(m(X), Var(X)),$ variable X	$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), \mathbf{K}(\mathbf{x},\mathbf{x}')\big),$ indices $\mathbf{x}$

Gaussian process

Function view



Control and Optimization of Soft Exosuit to Improve the efficiency of Human Walking - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Illustration-of-1-D-Gaussian-process-A-Gaussian-process-is-a-statistical-model-that_fig44_325386879

Outline

Introduction	Non-Bayesian method	Bayesian method	Non-parametric Bayesian
• Why we need uncertainty? • Basic Bayesian rules	• Linear Regression	• Bayesian Regression • Variational Inference	• Non-parametric method (GP) • Gaussian Process Regression

Linear regression vs. GP regression

Function view

$$y = f(\mathbf{X}) + \epsilon \quad , \quad \epsilon \sim \mathcal{N}(0, \sigma_y^2)$$

(Parametric)
Bayesian linear regression

$$f(\mathbf{X}) = \mathbf{\Phi w}$$

$$p(\mathbf{y}|\mathbf{w}, \mathbf{X}) = \mathcal{N}(\mathbf{y}|\mathbf{\Phi w}, \sigma_y^2 \mathbf{I})$$

$$p(\mathbf{y}|\mathbf{X}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma_\mathbf{w}^2 \mathbf{\Phi\Phi}^\mathbf{T} + \sigma_y^2 \mathbf{I})$$

(Non-parametric)
GP regression

$$f(\mathbf{X}) \sim \mathcal{N}(f|\mathbf{0}, \mathbf{K}(\mathbf{x},\mathbf{x}'))$$

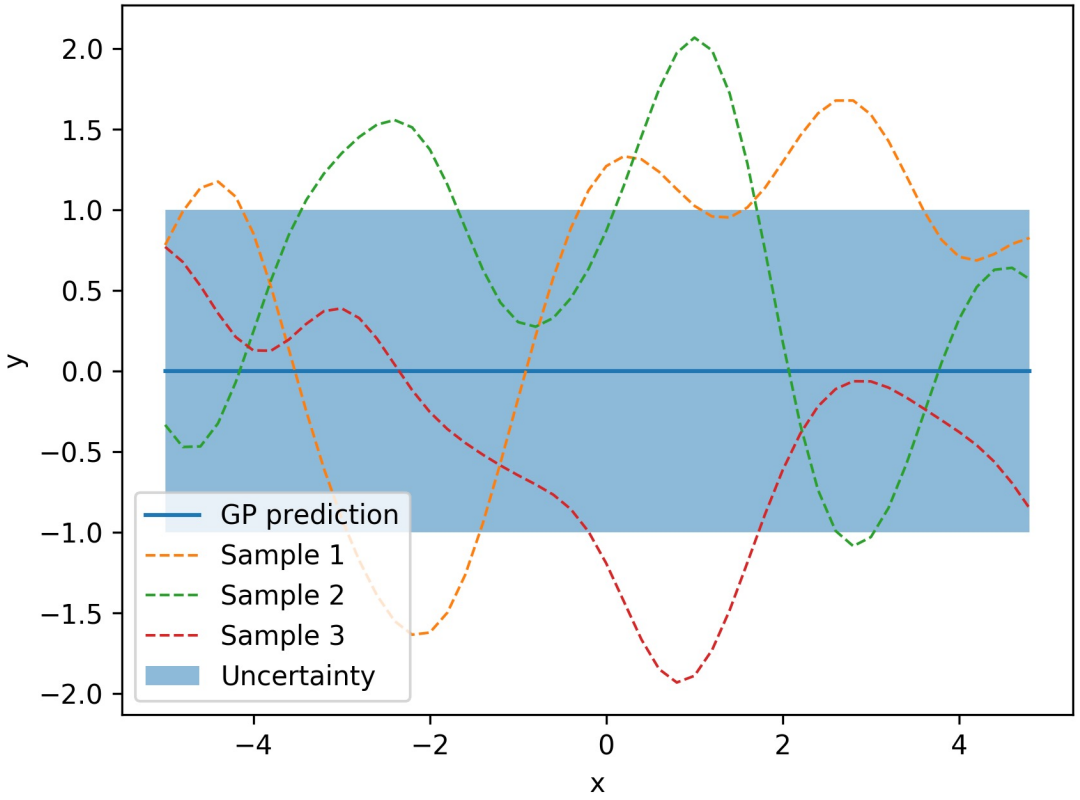
$$p(\mathbf{y}|\mathbf{f}, \mathbf{X}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}(\mathbf{x},\mathbf{x}') + \sigma_y^2 \mathbf{I})$$

$$\mathbf{K}(\mathbf{x},\mathbf{x}') = \sigma_\mathbf{w}^2 \mathbf{\Phi\Phi}^\mathbf{T}$$

$$\mathbf{\Phi} : (\infty \times M)$$

GP regression

Prior distribution on function



Prior:

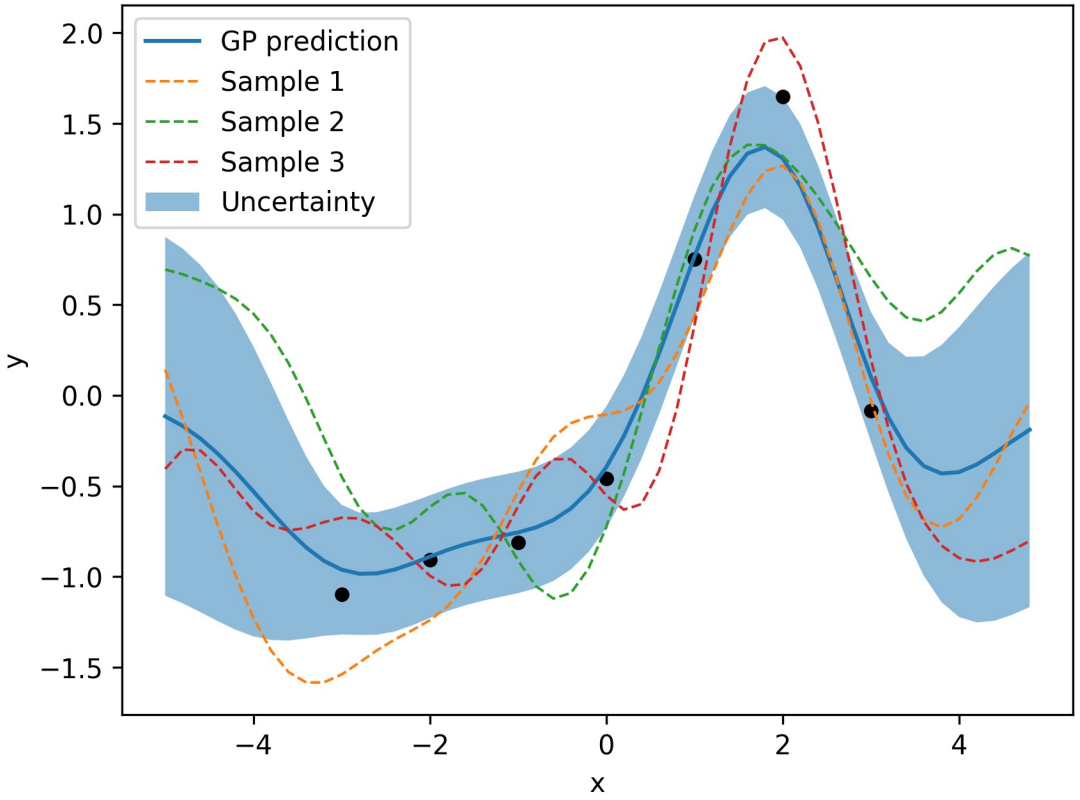
$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}\left(\mathbf{f}|\mathbf{0}, \mathbf{K}\left(\mathbf{x},\mathbf{x}'\right)\right)$$

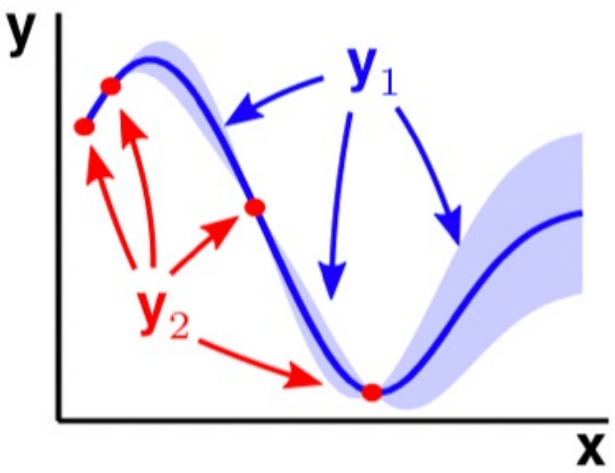
Gaussian Kernel:

$$\kappa(\mathbf{x}_i,\mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(\mathbf{x}_i-\mathbf{x}_j)^T(\mathbf{x}_i-\mathbf{x}_j)\right)$$

GP regression

GP prediction





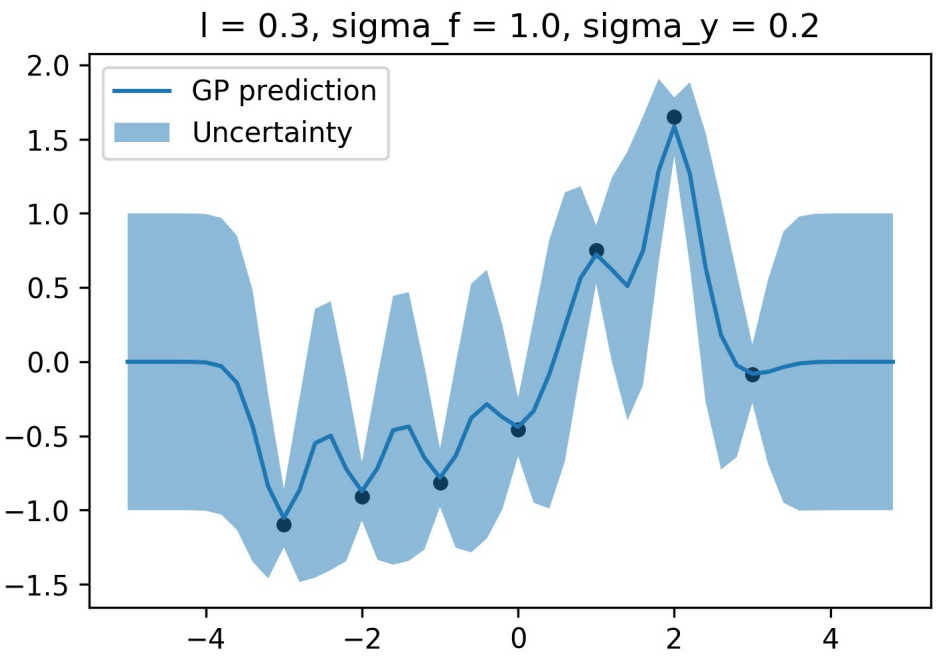
Use observed data to predict:

$$p(\mathbf{y}^*|\mathbf{X}^*,\mathbf{X},\mathbf{y}) = \int p(\mathbf{y}^*|\mathbf{X}^*,\mathbf{f})p(\mathbf{f}|\mathbf{X},\mathbf{y})d\mathbf{f}$$
$$= \mathcal{N}(\mathbf{y}^*|\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)$$

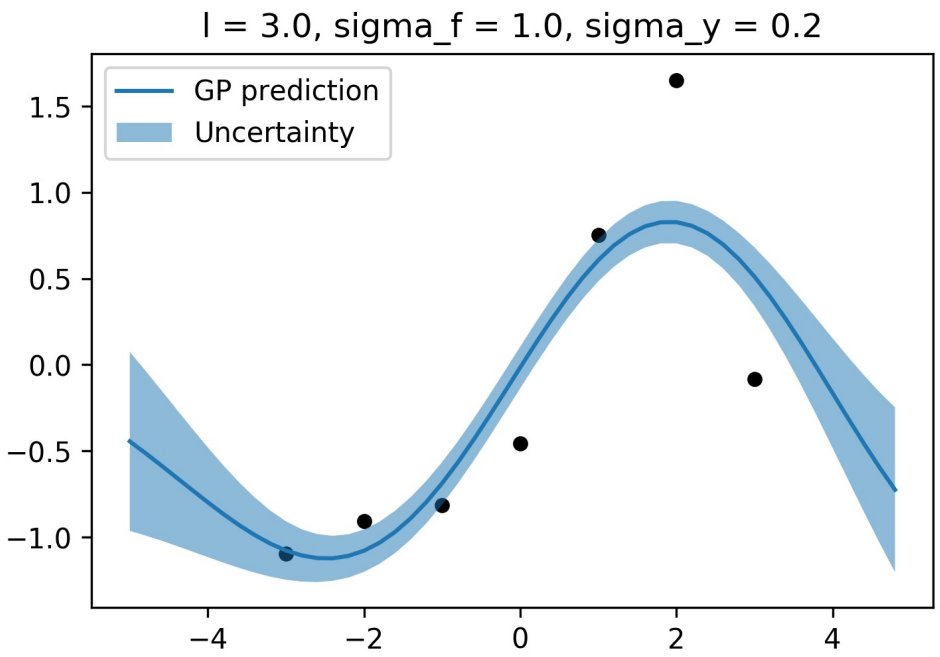
Different hyper-parameters

horizontal-scale

$$K(x_1, x_2) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$



small l

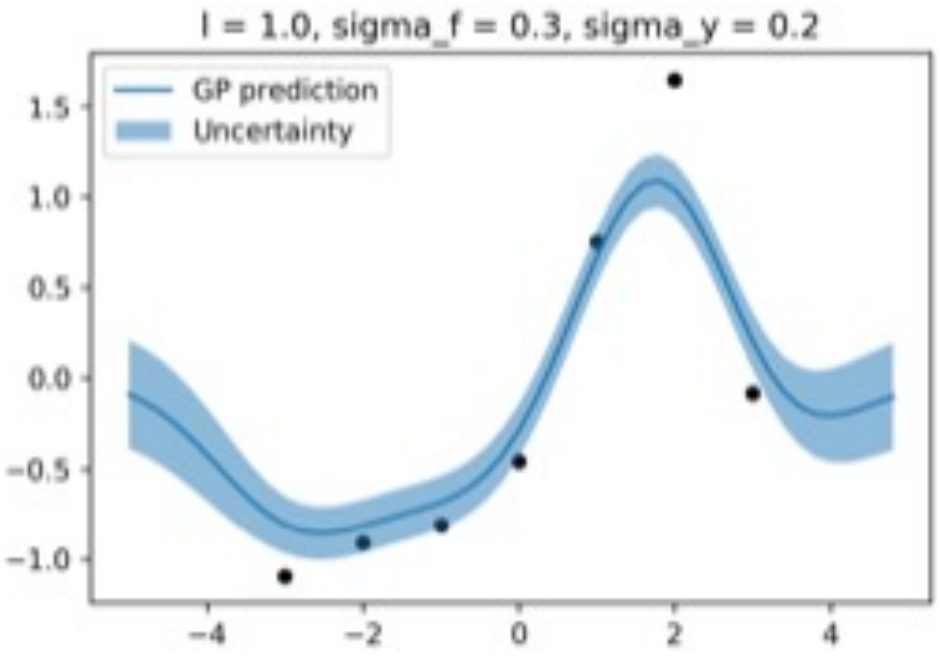


large l

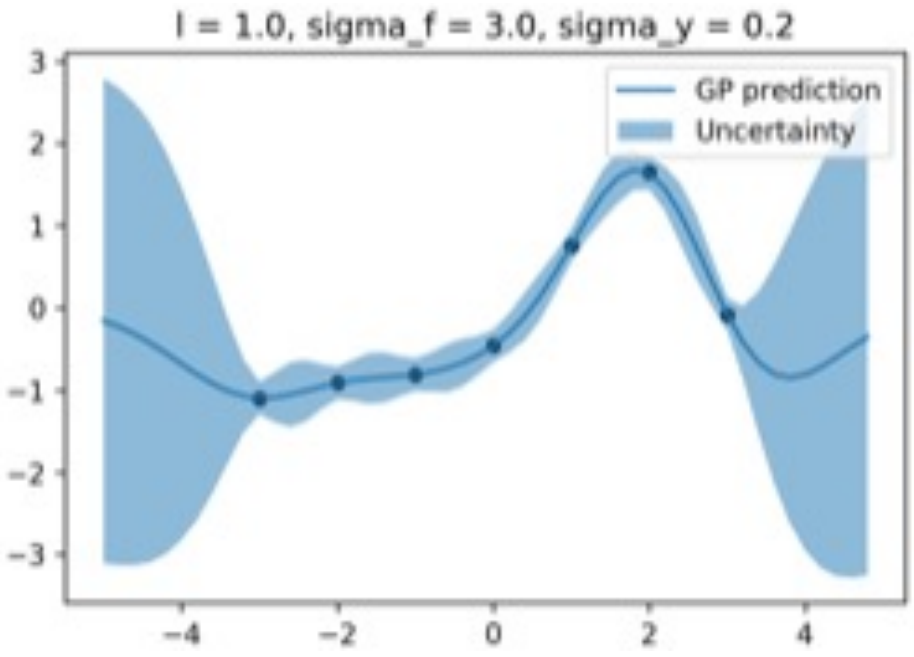
Different hyper-parameters

vertical-scale

$$K(x_1, x_2) = \sigma_f^2 \exp(-\frac{1}{2l^2}(x_1 - x_2)^2)$$



small σ_f

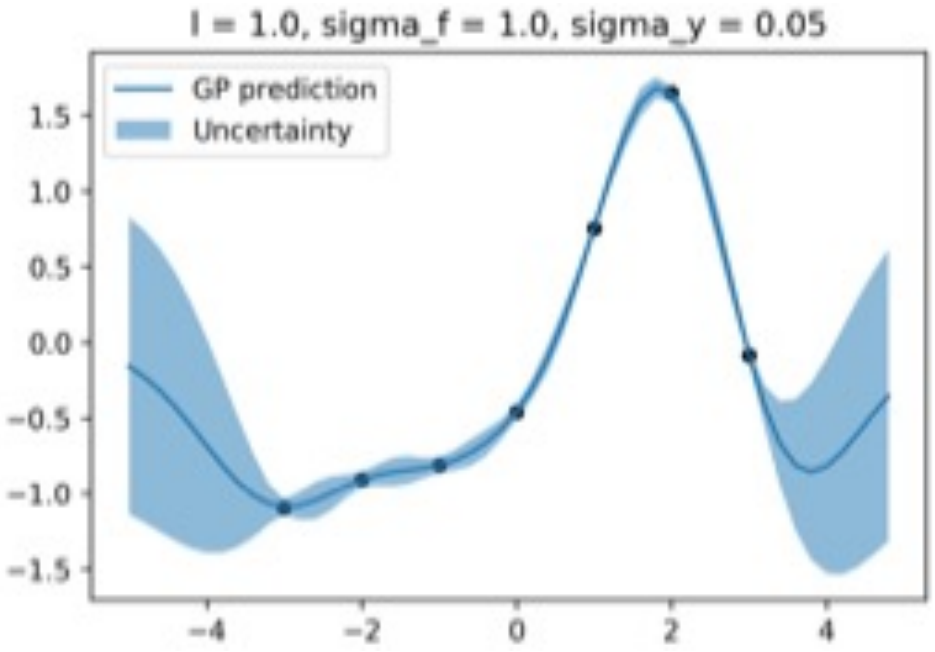


large σ_f

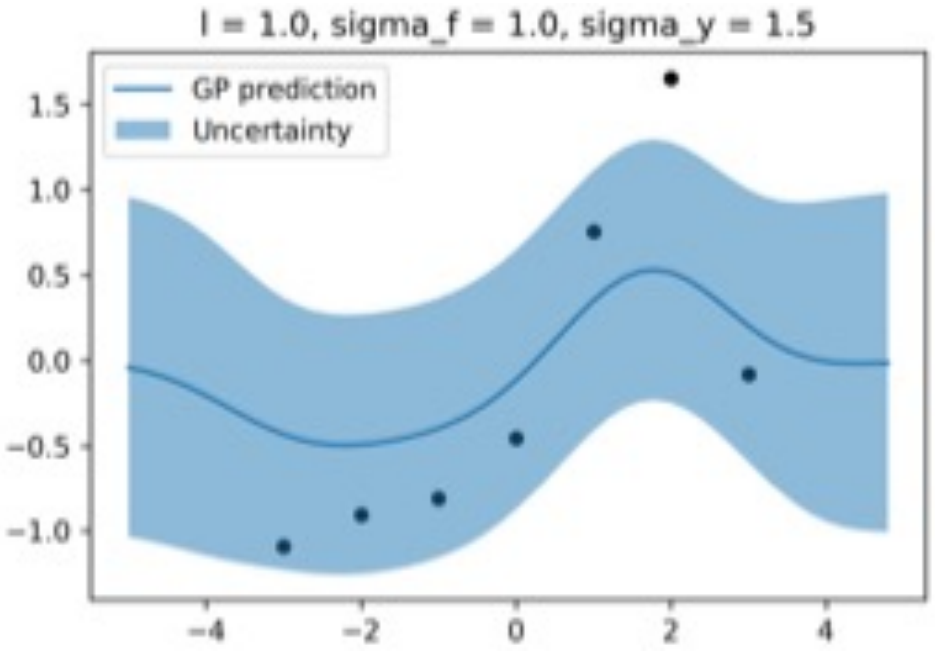
Different hyper-parameters

noise level

$$p\,(\mathbf{y}|\mathbf{f},\mathbf{X}) = N(\mathbf{y}|\mathbf{0}, \mathrm{K}(\mathbf{x},\mathbf{x}') + \sigma_y^2\mathbf{I})$$



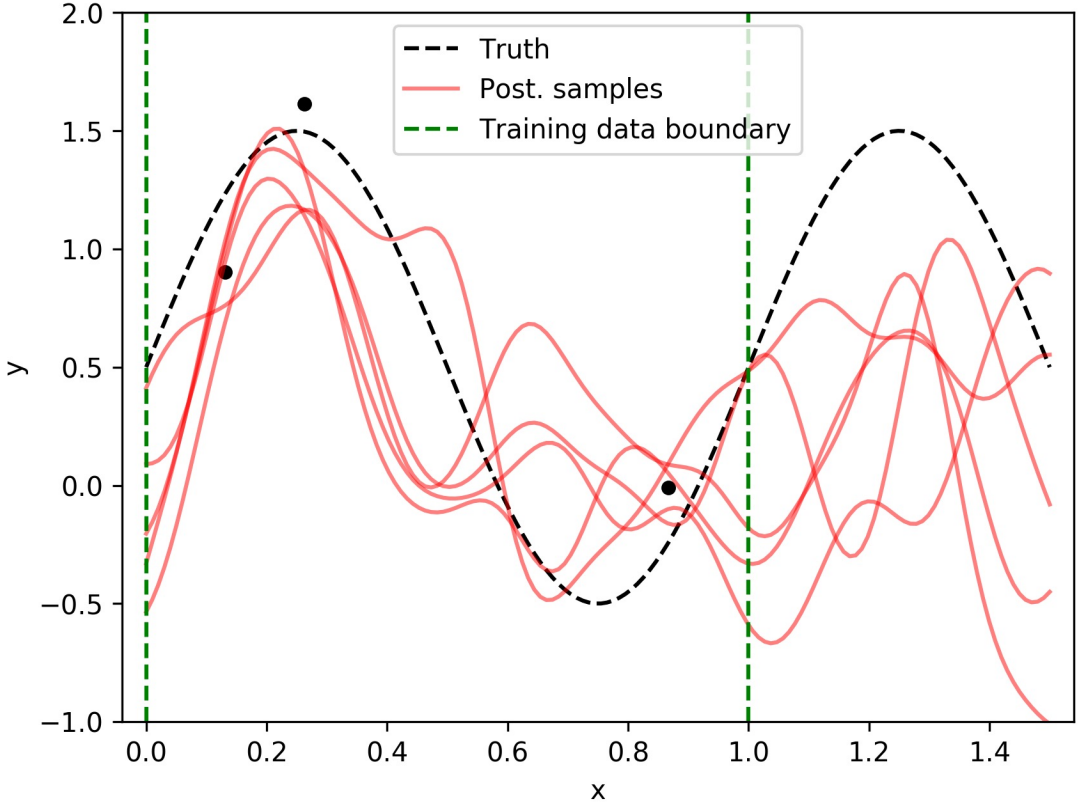
small σ_y

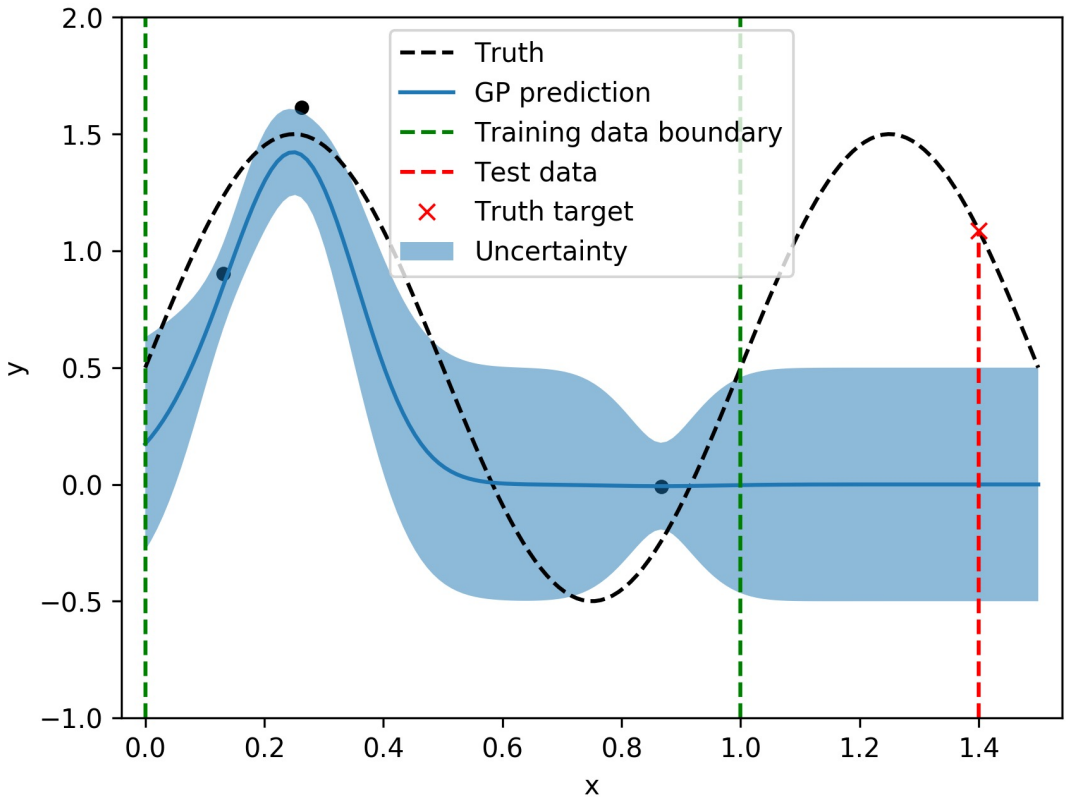


large σ_y

GP regression

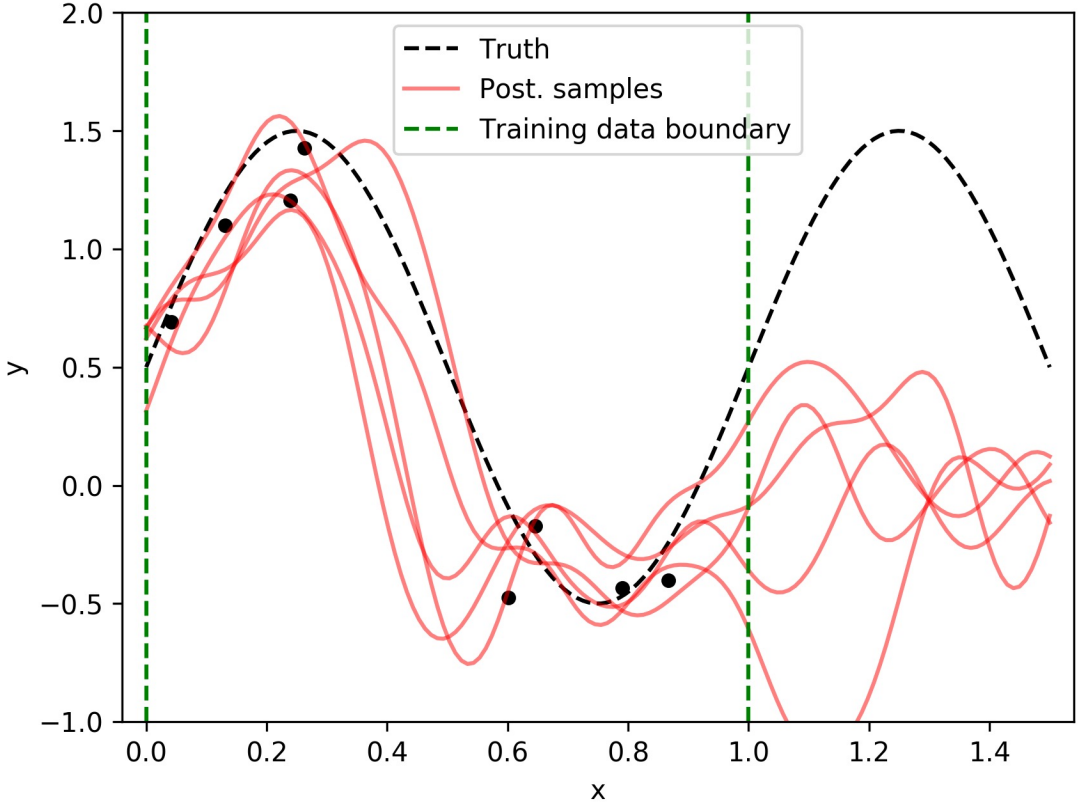
3 samples

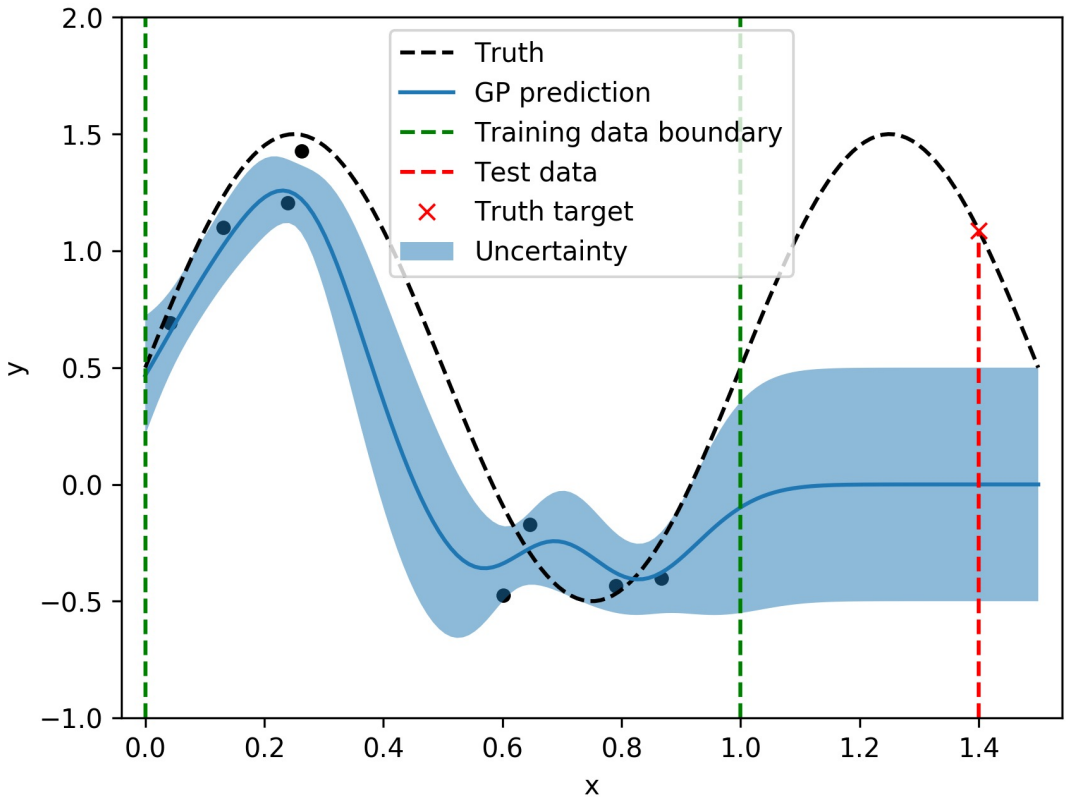




GP regression

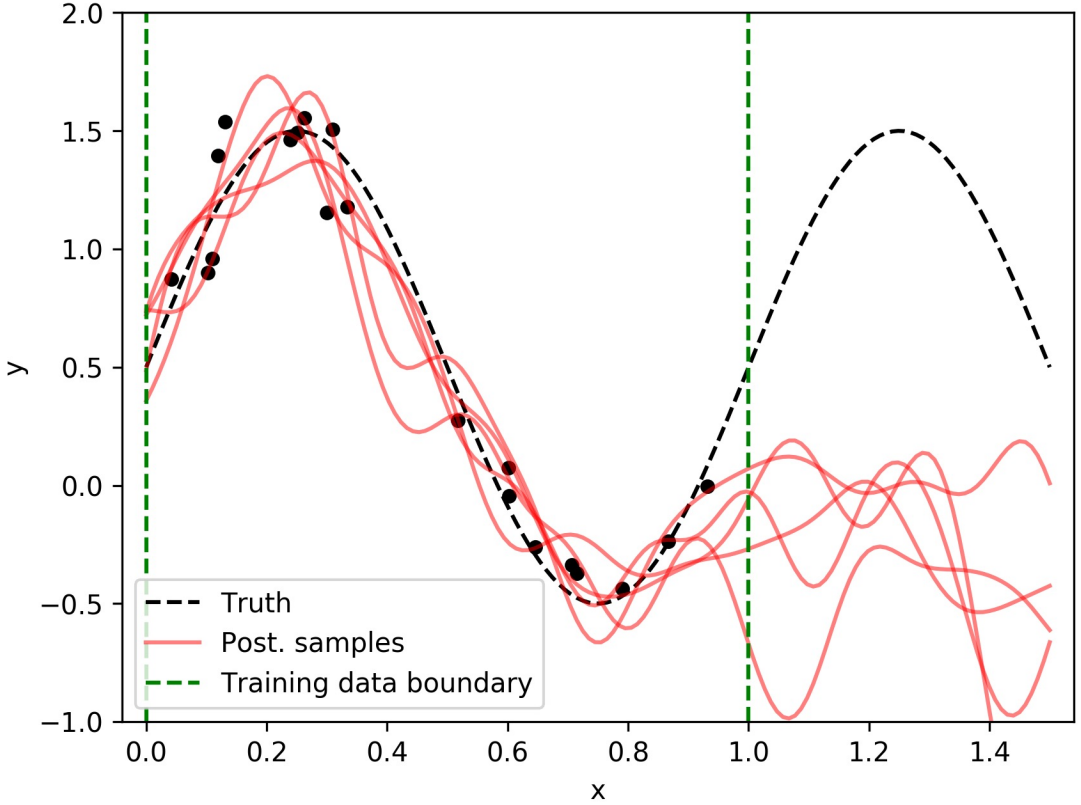
8 samples

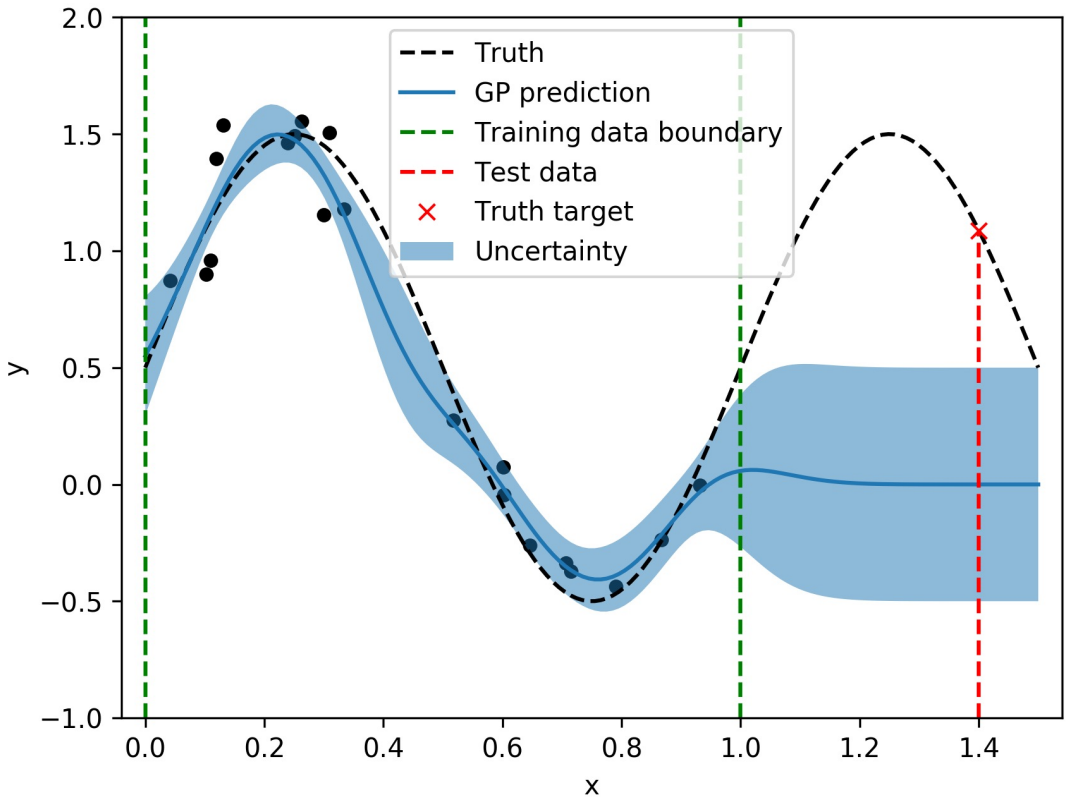




GP regression

20 samples

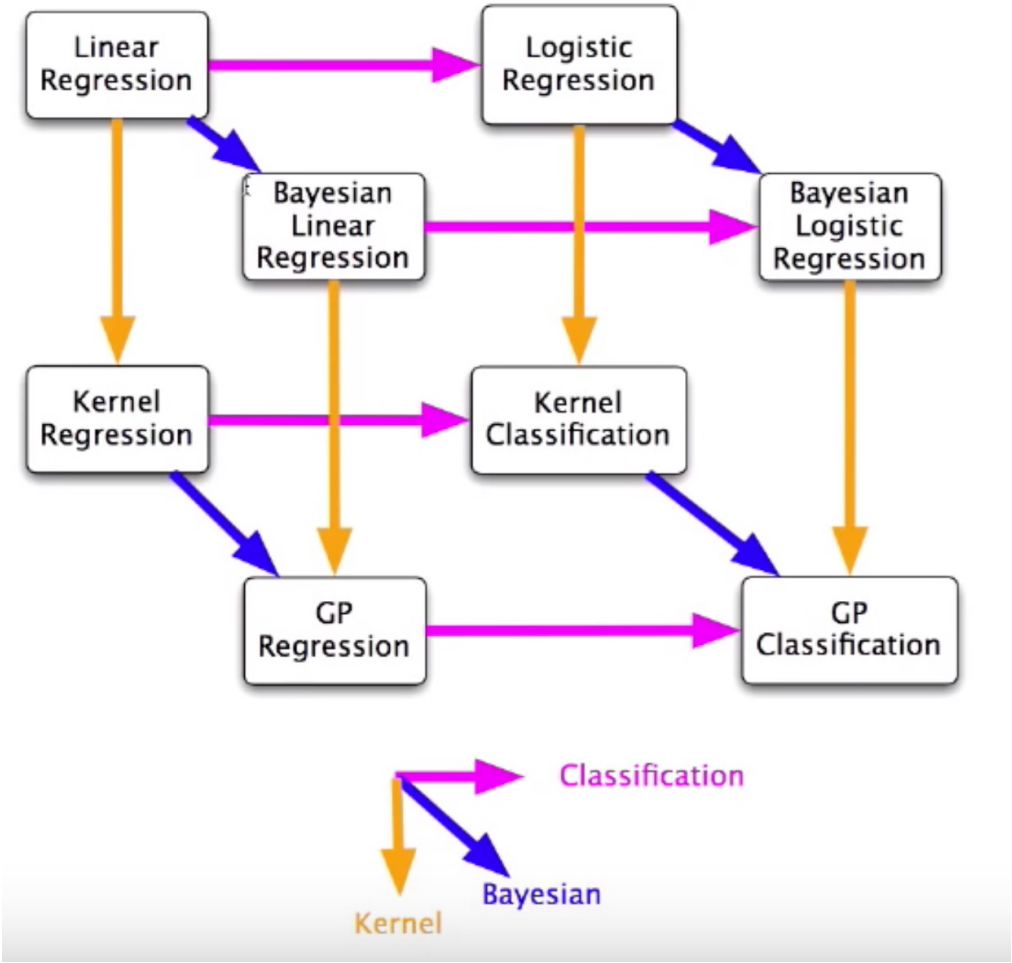




Comparison

The relationship between

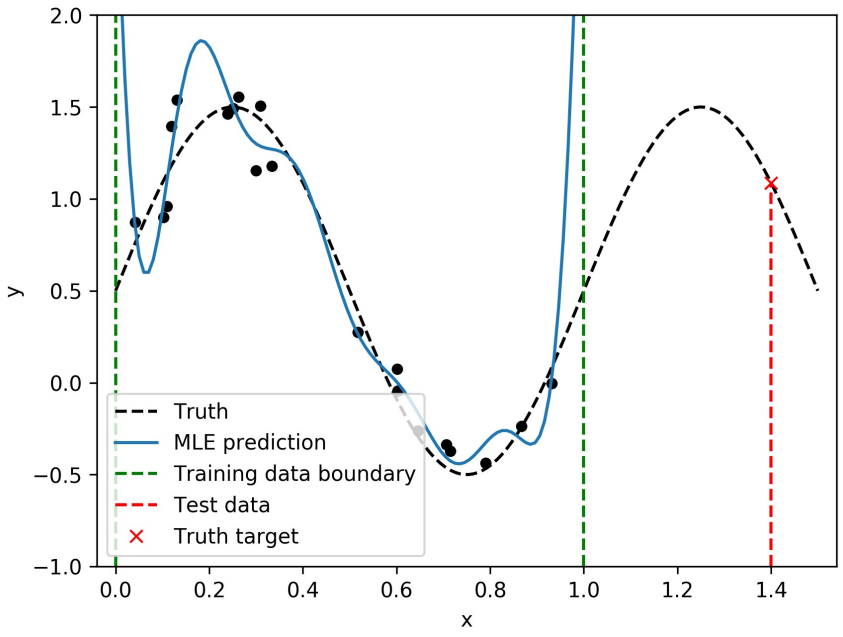
- Linear regression
- Bayesian linear regression
- GP regression

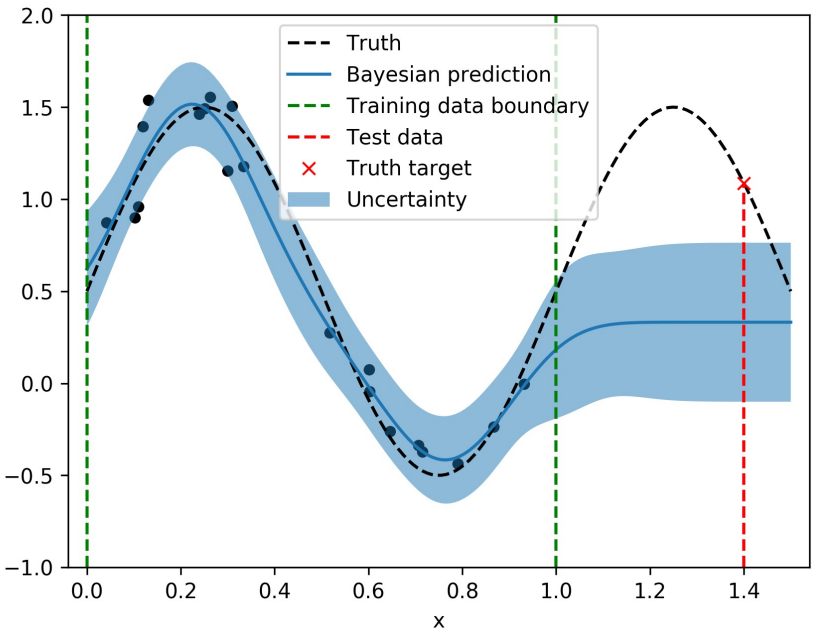


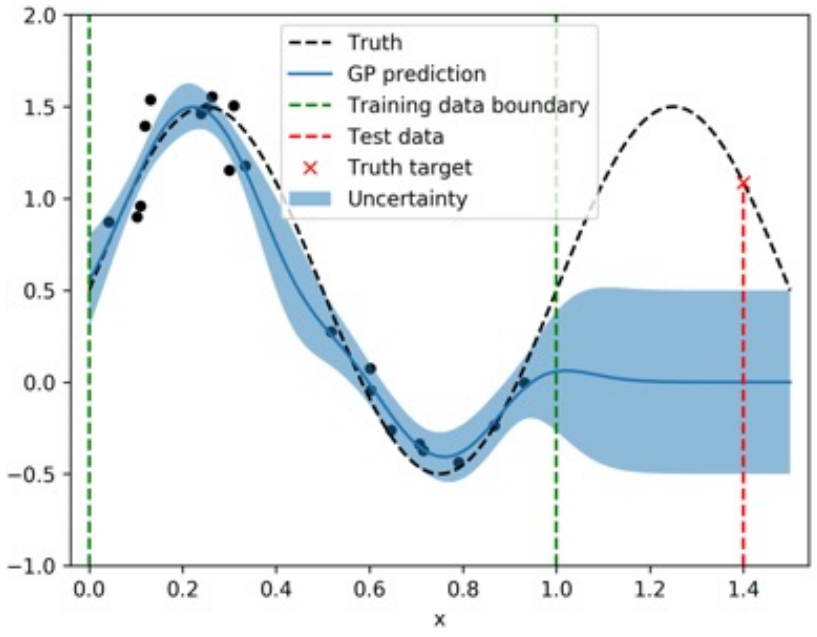
History of Bayesian Neural Networks (Keynote talk), NIPS 16
Available from: https://www.youtube.com/watch?v=FD8l2vPU5FY

Comparison

Linear regression → Bayesian → Linear Bayesian regression → Non-parametric → GP regression







Take away

- Bayesian method can give us a prediction distribution which can give us a sense of uncertainty of the prediction
- Instead of explicitly assuming the form of mapping function, non-parametric method predict based on the correlation of each samples
- For Bayesian method, when the posterior is intractable, we can use variational inference to approximate it with a familiar distribution model

References

- Bishop, Christopher M. *Pattern recognition and machine learning*. springer, 2006.(http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf)
- http://krasserm.github.io/2019/02/23/bayesian-linear-regression/
- https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1702.pdf
- https://www.coursera.org/learn/bayesian-methods-in-machine-learning/lecture/8e5un/why-approximate-inference
- http://krasserm.github.io/2018/03/19/gaussian-processes/
- https://arxiv.org/pdf/1601.00670.pdf (Variational Inference: A Review for Statisticians)
- http://gpss.cc/gpss13/assets/Sheffield-GPSS2013-Turner.pdf (GP)


рахмат
danke
謝謝
ngiyabonga
spas
teşekkür ederim
Баярлалаа
mersi
barka
welalin
tack
mahalo
спасибо
faafetai lava
vinaka
kia ora
dank je
misaotra
paldies
grazzi
tapadh leat
thank you
gracias
спасибі
blagodaram
matondo
хвала
kiitos
dankie
nanni
nandri
dhanyavad
asante
manana
obrigada
murakoze
tenki
bedankt
mauruuru
köszönöm
akun
dankon
aciû
hvala
gràcie
bayarlalaa
enkosi
djiere dieuf
tau
mochchakkeram
mamnun
дякую
chokrane
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
dēkuji
sagolun
sukriya
kop khun krap
obrigado
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
arigatô
takk
dakujem
trugarez
kam sah hamnida
rahmat
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
merçé
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci