Asymptotic Bayesian Generalization Error when Training and Test Distributions are Different

Keisuke Yamazaki [1)]
Motoaki Kawanabe [2)]
Sumio Watanabe [1)]
Masashi Sugiyama [1)]
Klaus-Robert Müller [2), 3)]

1) Tokyo Institute of Technology
2) Fraunhofer FIRST, IDA
3) Technical University of Berlin

Summary of This Talk

- Our target situation is non-regular models under the covariate shift.

	regular	non-regular
standard	statistics	algebraic geometry
covariate shift	importance weight	

- Non-regular model is a class of practical parametric models such as Gaussian mixtures, neural networks, hidden Markov models, etc.
- The covariate shift is the setting, where the training and test input distributions are different.

Summary of Our Theoretical Results

- Analytic expression of generalization error in large sample cases
  - Small order terms, which can be ignored in the absence of covariate shift, play an important role.
  - Small order terms are difficult to analyze in practice.
- Upper bound of generalization error in small sample cases
  - Our bound is computable for any sample size.
  - The worst case generalization error is elucidated.

Contents

1. Explanation of the table

	regular	non-regular
standard		
covariate shift		

2. Our results

First, I'll explain the table,
then, show our results.

Contents

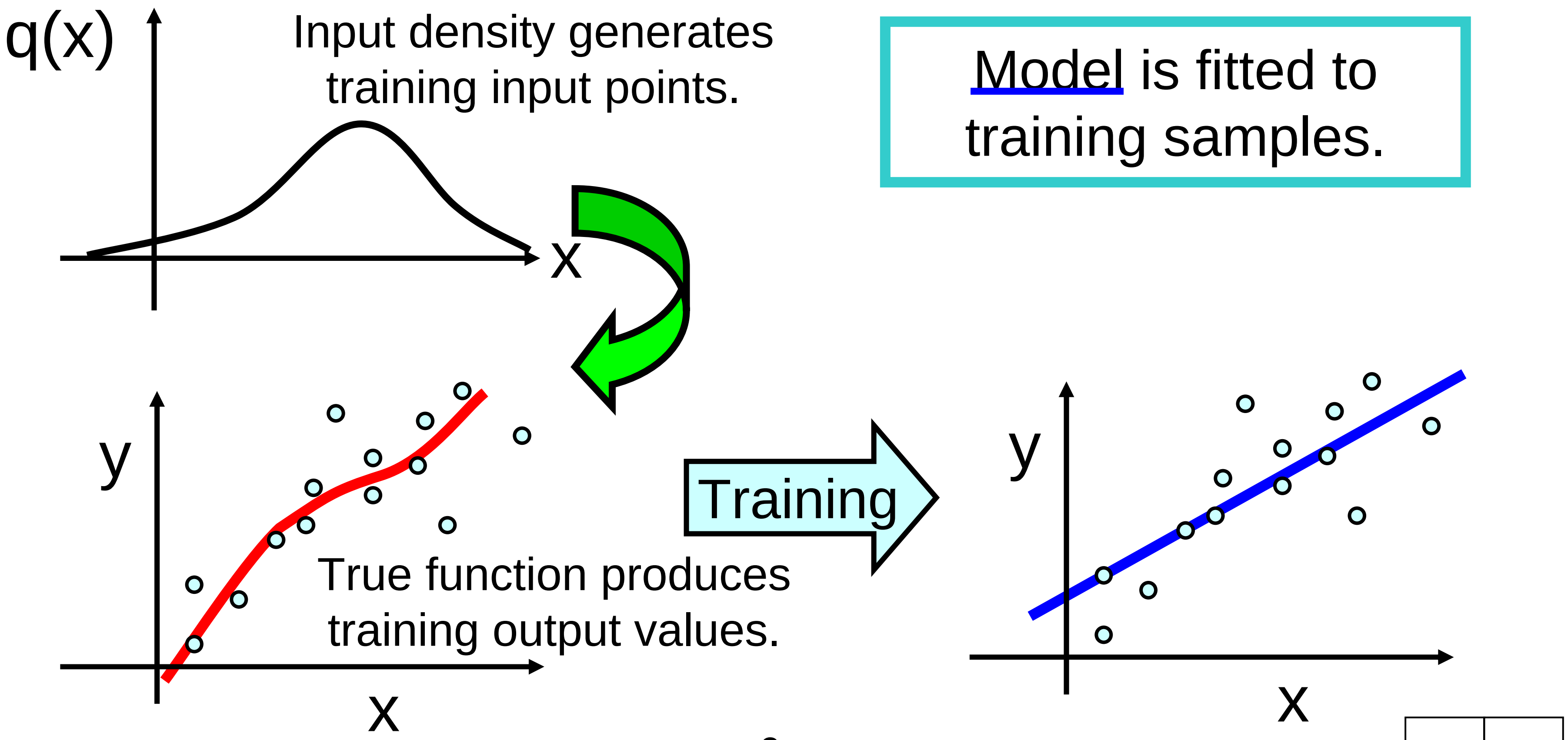
1. Explanation of the table

	regular	non-regular
standard		
covariate shift		

2. Our results

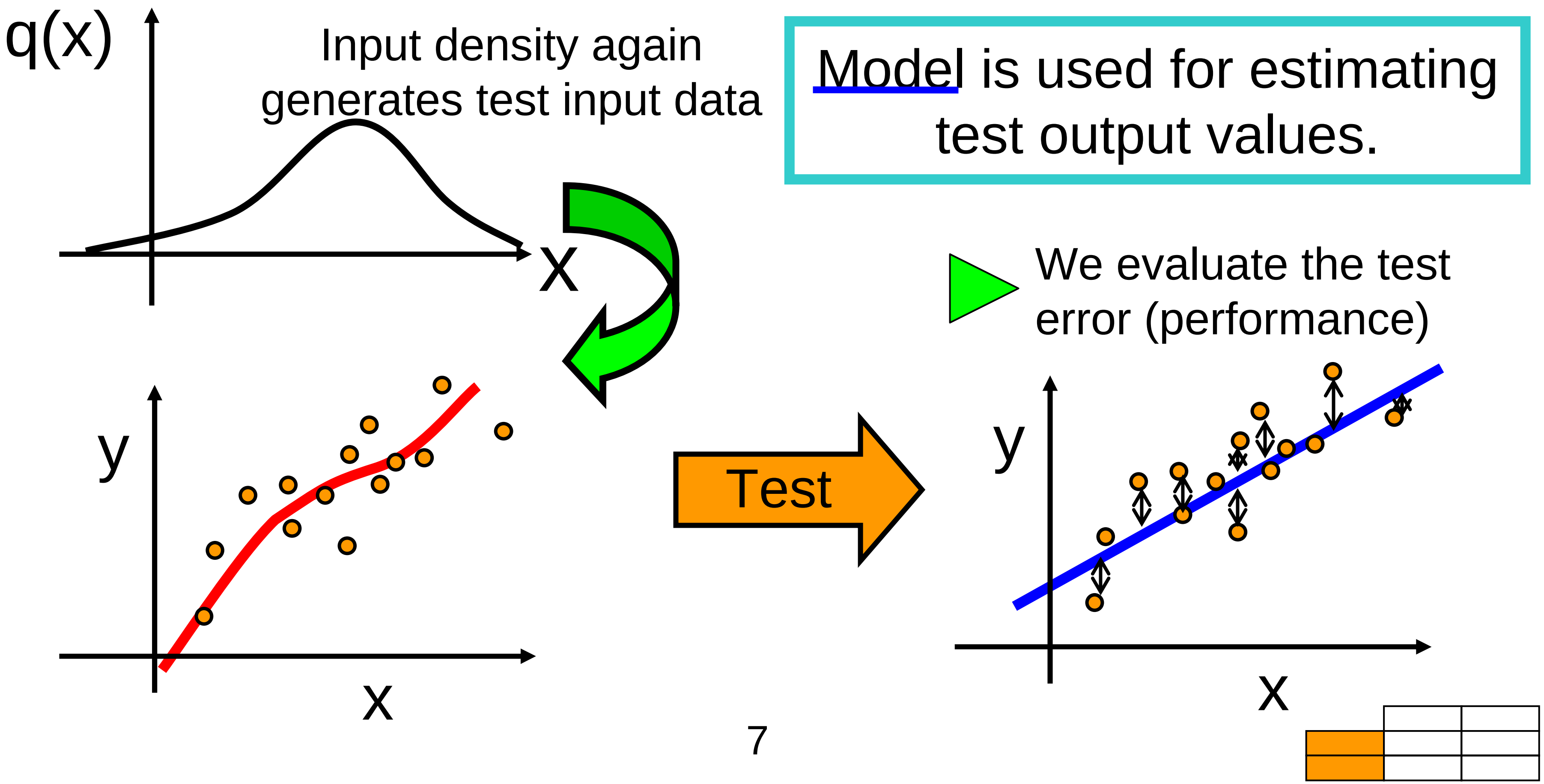
Regression Problem

- Training phase: learn input-output relation from training samples $r(y\,|\,x)$



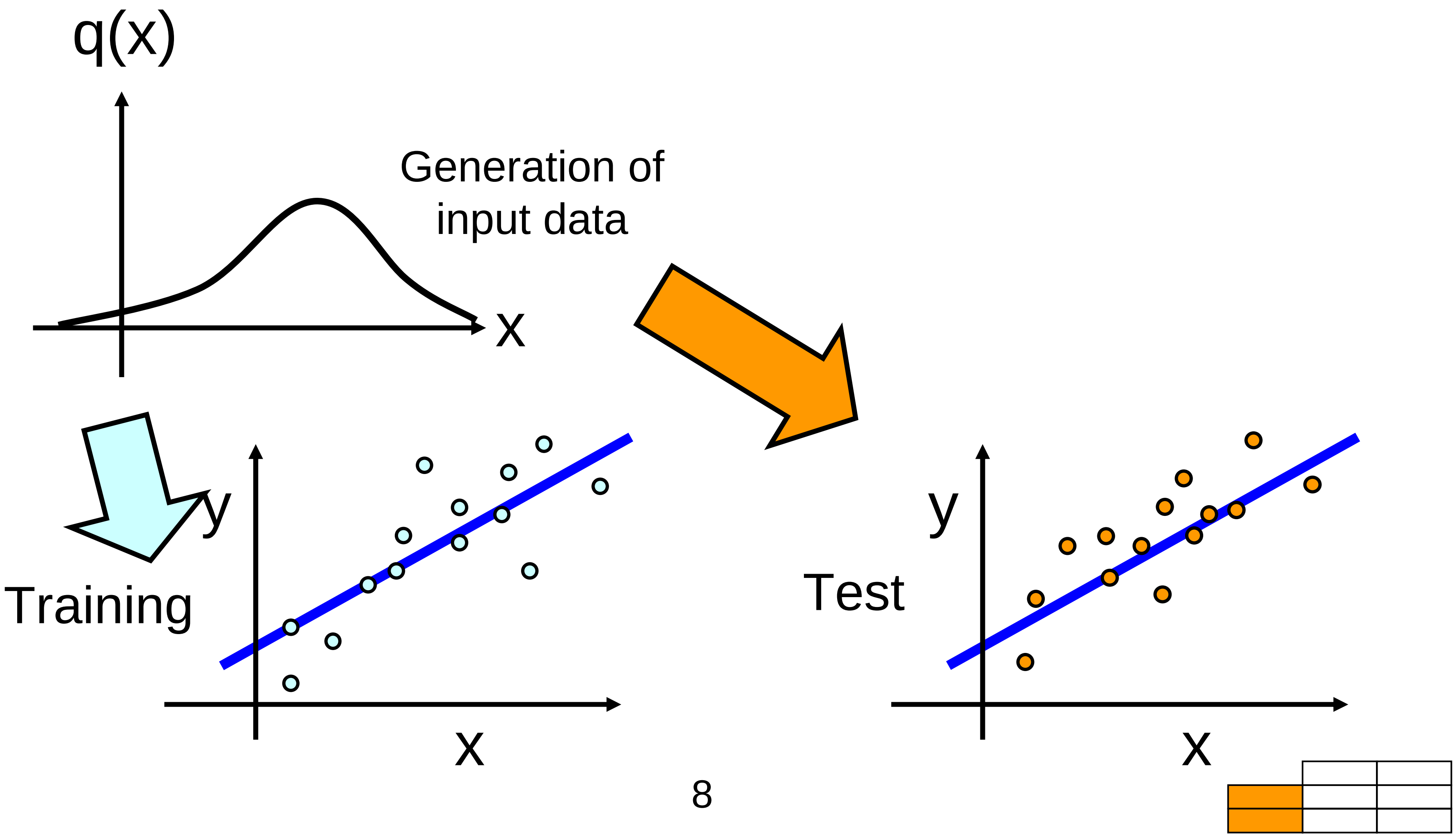
Regression Problem

- Test phase: predict test output values at given test input points



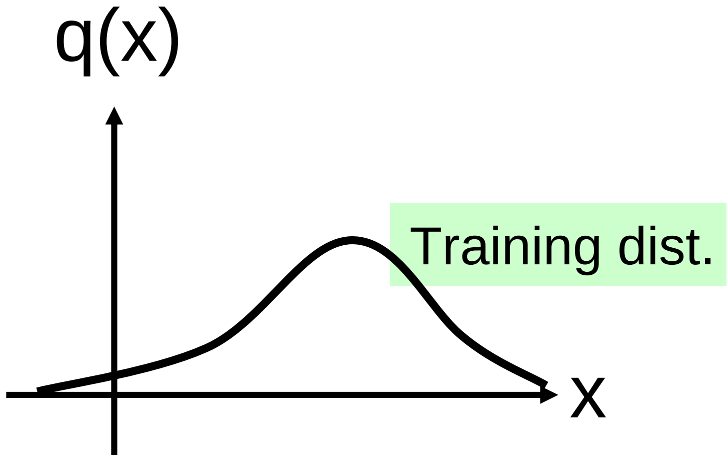
Input Distribution in Standard Setting

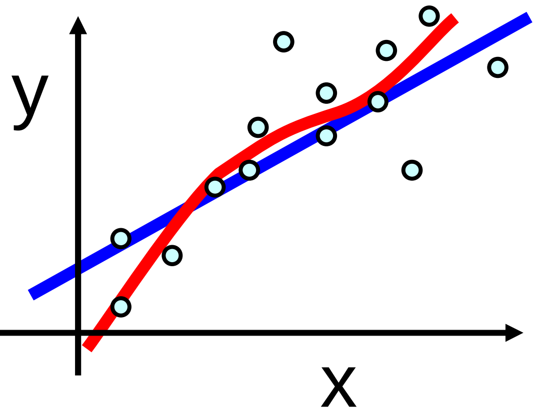
- The training and test distributions are same



Input Distribution in Practical Situations

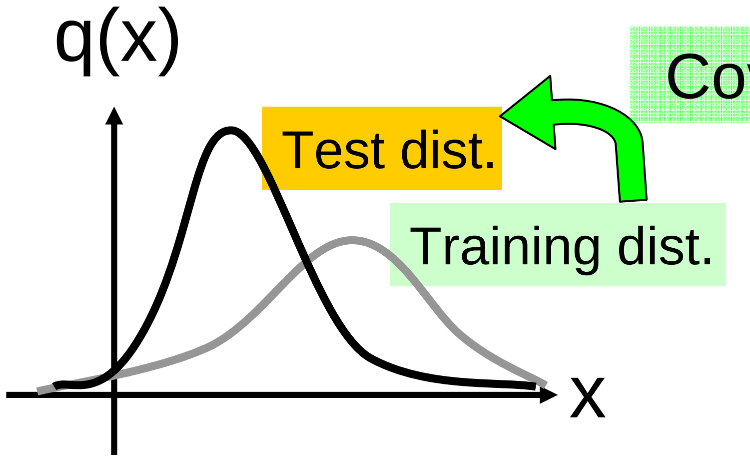
- The training and test input distributions are …





Input Distribution in Practical Situations

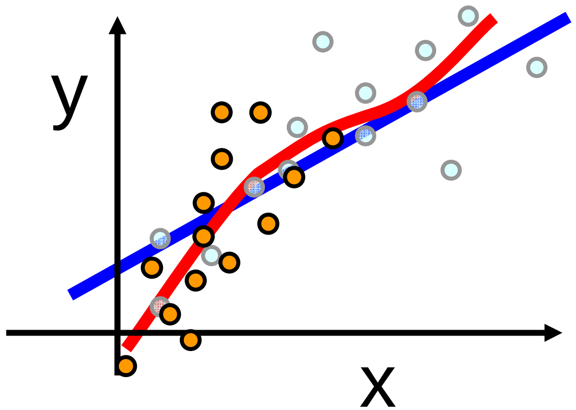
- The training and test input distributions are different!!!



Covariate shift

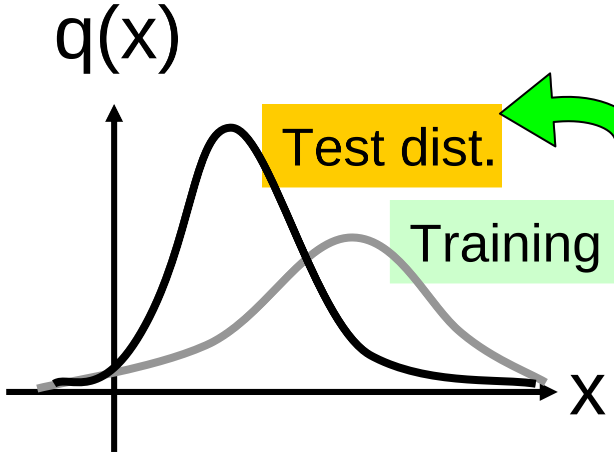
- Bioinformatics
  - [Baldi et al., 1998]
- Econometrics
  - [Heckman, 1979]
- Brain-computer interface
  - [Wolpaw et al., 2002]

etc.



Input Distribution in Practical Situations

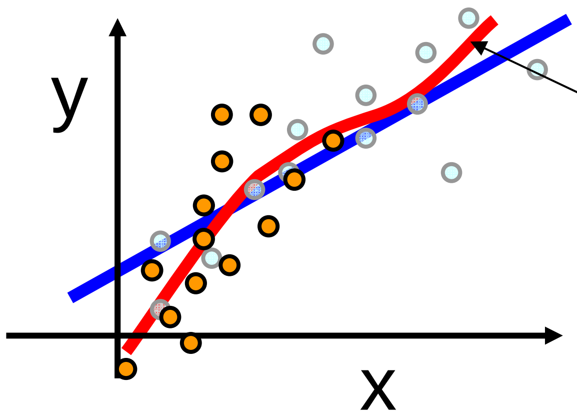
- The training and test input distributions are different!!!



Covariate shift

- Bioinformatics
  - [Baldi et al., 1998]
- Econometrics
  - [Heckman, 1979]
- Brain-computer interface
  - [Wolpaw et al., 2002]

etc.

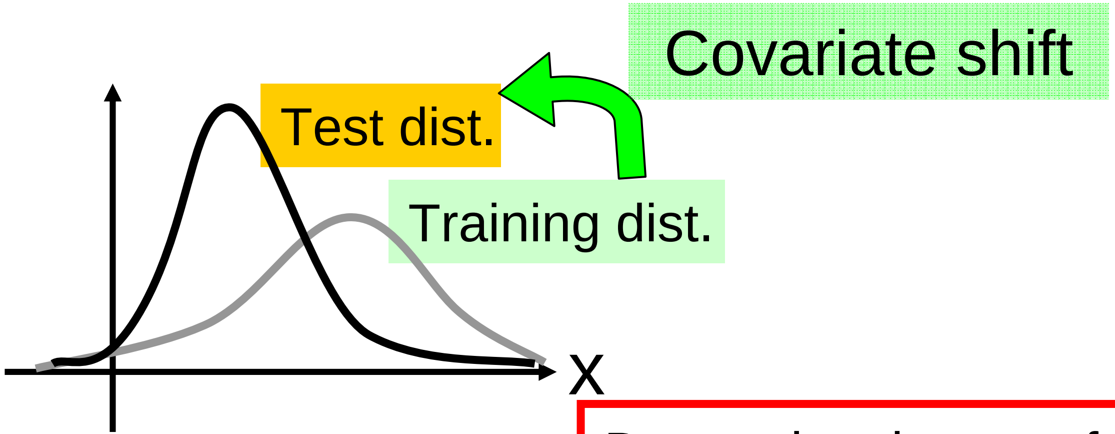


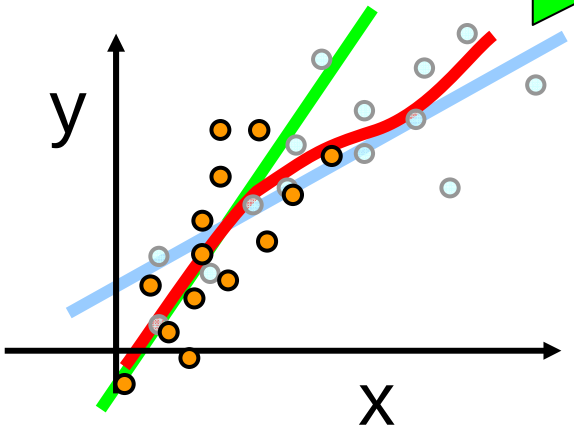
$r(y\,|\,x)$ **doesn't change.**

$q_0(x) \Rightarrow q_1(x)$

Input Distribution in Practical Situations

- The training and test input distributions are different!!!





Due to the change of data region,
the performance also changes.

A standard technique does NOT work.

Contents

1. Explanation of the table

	regular	non-regular
standard		
covariate shift		

2. Our results

Classes of Learning Models

- Non-/ Semi-parametric models
  - SVM

  etc.

- Parametric models
  - Regular
    - Polynomial regression
    - Linear model

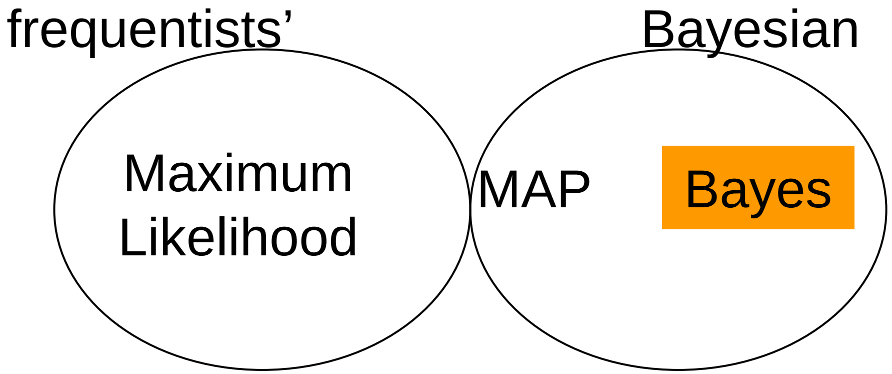
    etc.
  - Non-regular
    - Neural network
    - Gaussian mixture
    - Hidden Markov model
    - Bayesian network
    - Stochastic CFG

    etc.

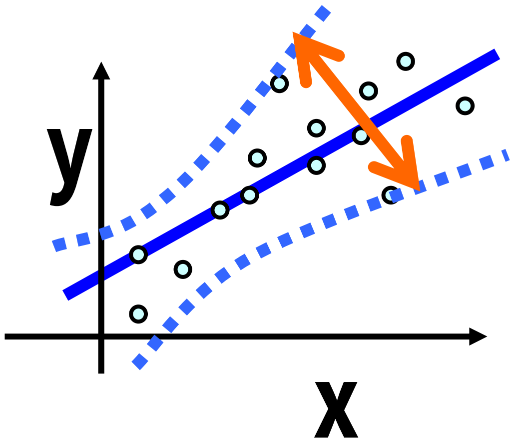
Non-regular models have hierarchical structure or hidden variables.

It is important to analyze non-regular models.

Our Learning Method is Bayesian



- Bayesian Learning [Parametric]



The Bayesian learning constructs the predictive distribution as the average of models.

Contents

1. Parametric Bayesian framework

	regular	non-regular
standard		
covariate shift		

2. Our results

Here, the interest is the generalization performance in each setting.

Before looking at each case, let us define how to measure the generalization performance.

How to Measure Generalization Performance

- Kullback divergence (or log-loss)

$$D(p_1 \| p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx$$

It shows the distance between densities.

$$p_1(x) = p_2(x) \Leftrightarrow D(p_1 \| p_2) = 0$$

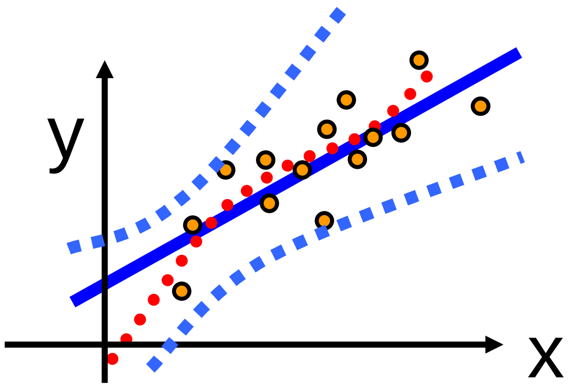
$$p_1(x) \neq p_2(x) \Leftrightarrow D(p_1 \| p_2) > 0$$

D(true function || predictive distribution)

▶ Kullback divergence from the true distribution to the predictive distribution.

Expected Kullback Divergence Is Our Generalization Error

$$G^0(n) = E^0_{X^n,Y^n}\left[\int r(y|x)q_0(x)\log\frac{r(y|x)}{p(y|x,X^n,Y^n)}dxdy\right]$$



We take the expectation over all training samples.

It is the function w.r.t. the training sample size.

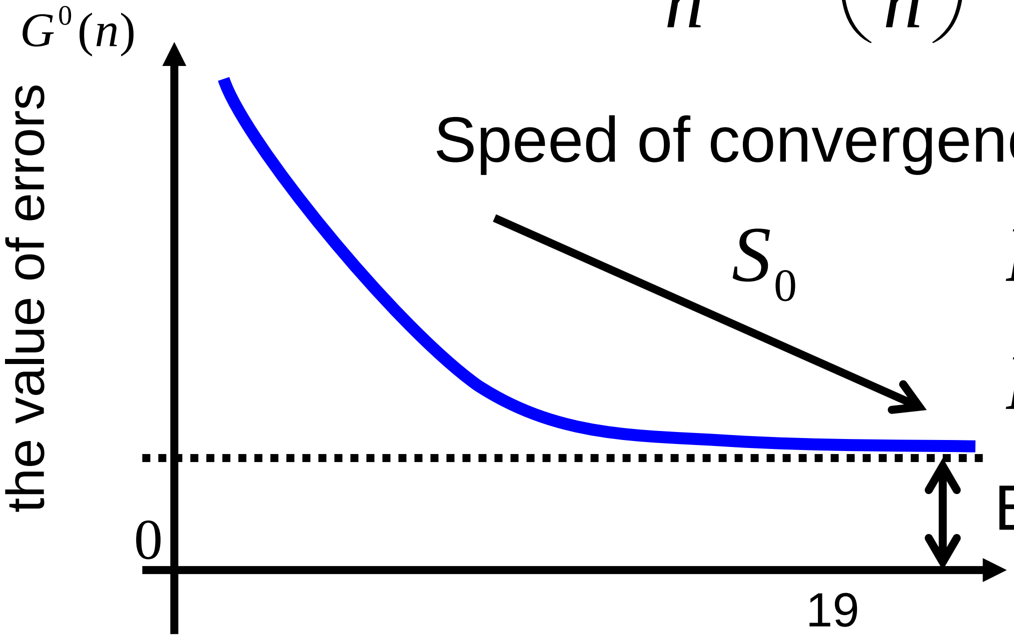
n

What Do We Want to Know?

- Learning curve: generalization error as a function of sample size

When the sample size n is sufficiently large,

$$G^0(n) = B_0 + \frac{S_0}{n} + o\left(\frac{1}{n}\right)$$



$B_0 > 0$ $\quad r(y|x) \neq p(y|x,w)$

$B_0 = 0$ $\quad r(y|x) = p(y|x,\hat{w})$

n :sample size

Contents

1. Parametric Bayesian framework

	regular	non-regular
standard		
covariate shift		

2. Our results

Now, we take a careful look at each case separately.

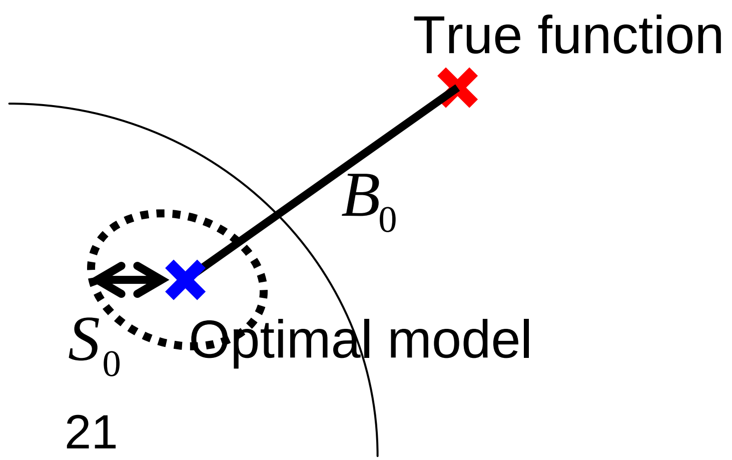
Regular Models in the Standard Input Dist.

- In statistics, the analysis has a long history.
  - Learning curve is well studied.

$$G^0(n) = B_0 + \frac{S_0}{n} + o\left(\frac{1}{n}\right)$$

B_0 Distance from the true function to the optimal model

S_0 (Dimension of parameter space)/2



Contents

1. Parametric Bayesian framework

	regular	non-regular
standard	statistics	
covariate shift		

2. Our results

Regular Models under Covariate Shift

Importance Weight $= \dfrac{q_1(x)}{q_0(x)}$

$$\int q_0(x) \times IW \times Loss(x)dx = \int q_0(x) \frac{q_1(x)}{q_0(x)} \times Loss(x)dx$$
$$= \int q_1(x) Loss(x)dx$$

- The importance weight improves the generalization error. [Shimodaira, 2000]

B_0 Distance to the optimal model following the test data.

S_0 Original speed + A factor from the importance weight.

$$G^0(n) = B_0 + \frac{S_0}{n} + o\left(\frac{1}{n}\right)$$

Contents

1. Parametric Bayesian framework

	regular	non-regular
standard	statistics	
covariate shift	importance weight	

2. Our results

Non-Regular Models without Covariate Shift

Stochastic Complexity: the average of marginal likelihood

$$U^0(n) = E^0_{X^n,Y^n}\left[-\log \int \prod_{i=1}^{n+1} \underbrace{p(Y_i|X_i,w)}_{\text{model}}\underbrace{\varphi(w)}_{\text{a prior}}dw\right]$$

Marginal likelihood is used for the model selection or the optimization of the prior.

n : the training data size

An asymptotic form of the stochastic complexity is

$$U^0(n) = a_0 n + b_0 \log n + o(\log n)$$

Analysis of Generalization Error in the Absence of Covariate Shift

According to the definition,

$$G^0(n) = U^0(n+1) - U^0(n)$$

$$\begin{aligned} U^0(n+1) &= a_0(n+1) + b_0 \log(n+1) + o(\log n) \\ - \quad U^0(n) &= a_0 n + b_0 \log n + o(\log n) \\ \hline G^0(n) &= a_0 + \frac{b_0}{n} + o\left(\frac{1}{n}\right) \end{aligned}$$

by very simple subtraction.

Generalization Error

$$G^0(n) = a_0 + \frac{b_0}{n} + o\left(\frac{1}{n}\right),$$ which includes regular cases.

Contents

1. Parametric Bayesian framework

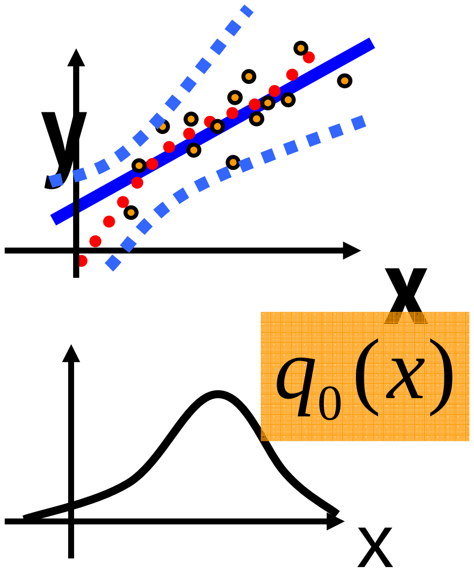
	regular	non-regular
standard	statistics	stochastic complexity
covariate shift	importance weight	

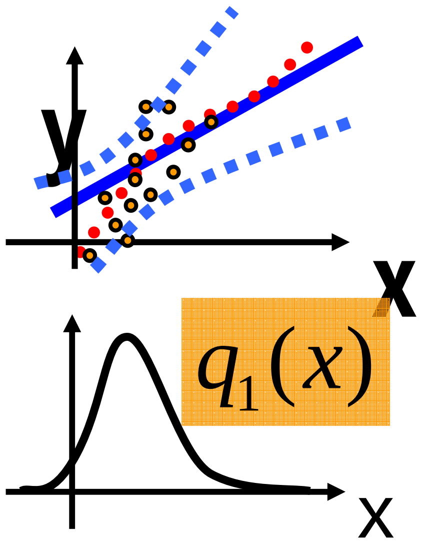
2. Our results
   - A) Large sample cases
   - B) Finite sample cases

The analysis for non-regular models under the covariate shift is still open !!!

Kullback Divergence w.r.t. Test Distribution

$$G^{i}(n) = E^{0}_{X^n,Y^n}\left[\int r(y|x)\,q_i(x)\log\frac{r(y|x)}{p(y|x,X^n,Y^n)}\,dxdy\right]$$





$G^0(n)$: standard case $G^1(n)$: covariate shift

Stochastic Complexity under Covariate Shift

We define shifted stochastic complexity:

$$U^i(n+1) = E^i_{X_{n+1},Y_{n+1}} E^0_{X^n,Y^n}\left[-\log\int\prod_{i=1}^{n+1} p(Y_i \mid X_i, w)\varphi(w)dw\right]$$

The expectation of test data is different.

The previous definition:

$$U^0(n) = E^0_{X^n,Y^n}\left[-\log\int\prod_{i=1}^{n+1} p(Y_i \mid X_i, w)\varphi(w)dw\right]$$

n : the training data size

Following the previous study,

Assumption: An asymptotic form of the stochastic complexity is

$$U^i(n) \cong a_i n + b_i \log n + \cdots + c_i + d_i / n + \cdots$$

The previous assumption:

$$U^0(n) = a_0 n + b_0 \log n + o(\log n)$$

n : the training data size

We Obtain Analytic expression of Generalization Error by subtraction

According to the definition,

$$G^i(n) = U^i(n+1) - U^0(n)$$

$$\begin{array}{rl} & U^1(n+1) = a_1(n+1) + b_1 \log(n+1) + \cdots + c_1 + d_1/(n+1) + o(1/n) \\ - & U^0(n) = a_0 n + b_0 \log n + \cdots + c_0 + d_0/n + o(1/n) \\ \hline & G^1(n) = (a_1 - a_0)n + (b_1 - b_0)\log n + a_1 + c_1 - c_0 + (b_1 + d_1 - d_0)/n + o(1/n) \end{array}$$

⇧

Based on a property of the learning curve, the expression can be simplified.

Small Order Terms Cannot be Ignored

- Theorem 1

$$G^1(n) = a_0 + (c_1 - c_0) + \frac{b_0 + (d_1 - d_0)}{n} + o\left(\frac{1}{n}\right)$$

$$(a_1 = a_0, b_1 = b_0)$$

$$G^0(n) \cong a_0 + \frac{b_0}{n}$$

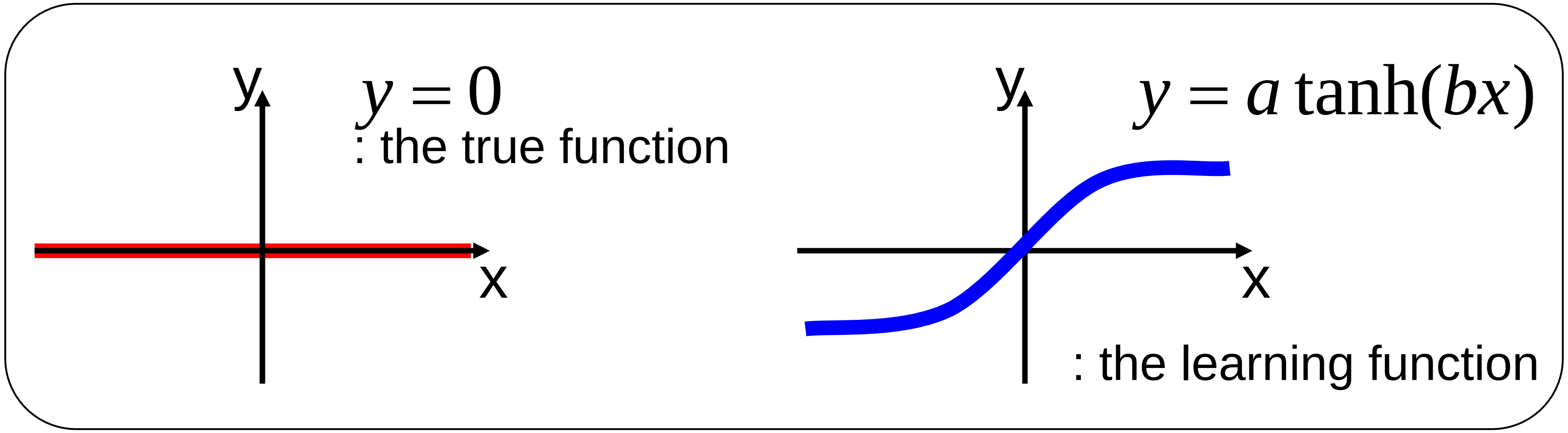
Small order terms are ignored in the standard asymptotic analysis.

$$U^i(n) \cong a_i n + b_i \log n + \cdots + c_i + d_i / n + \cdots$$

n : the training data size

Evaluation of Small Order Terms is Difficult!!!!

- Simple Neural Network



Evaluating small order terms is very hard
even in very simple settings.

Contents

1. Parametric Bayesian framework

	regular	non-regular
standard	statistics	stochastic complexity
covariate shift	importance weight	

2. Our results
   A) Large sample cases
   B) Finite sample cases

We Obtain an Finite-Sample Upper Bound

- Theorem 2

$$G^1(n) \leq M G^0(n)$$

Maximum ratio of input densities:

$$M = \max_{x \sim q_0(x)} \left[\frac{q_1(x)}{q_0(x)} \right] < \infty$$

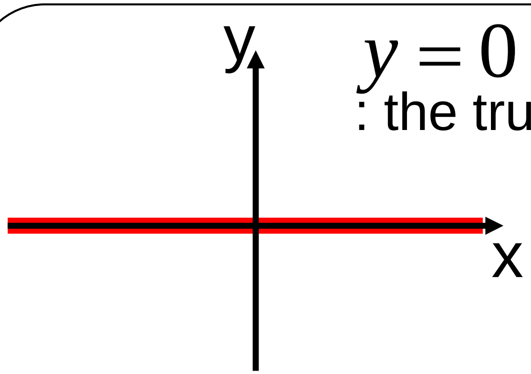
The upper bound can be easily computed!!!

We can overcome the difficulty in the previous theorem.

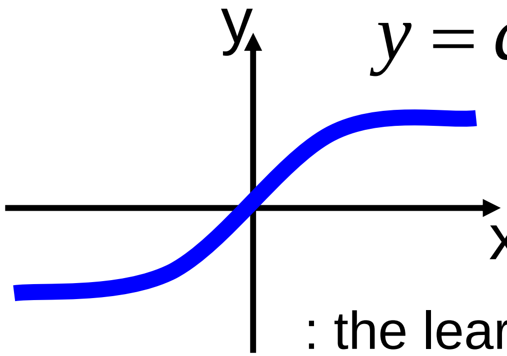
We Can Obtain Worst-Case Learning Curve

- Previous example

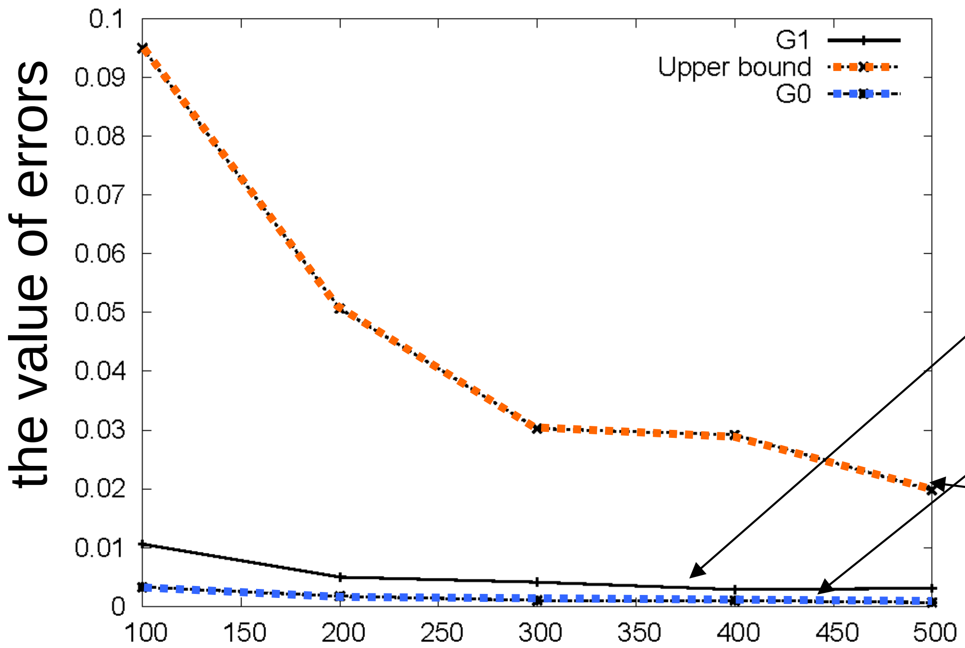
$y = 0$
: the true function



$y = a\tanh(bx)$



: the learning function



: sample size

$G^1(n)$: error under covariate shift

$G^0(n)$: error without shift

$MG^0(n)$: upper bound

Conclusions

- We analyzed Bayesian generalization error
  - of non-regular models: GM, HMM, NN etc.
  - under covariate shift: Input distribution change
- We proved that small order terms of stochastic complexity, which can be usually ignored, play important roles.
  - Directly evaluating generalization error is very hard.
- We derived a computable finite-sample upper bound
  - Worst-case generalization error is elucidated.