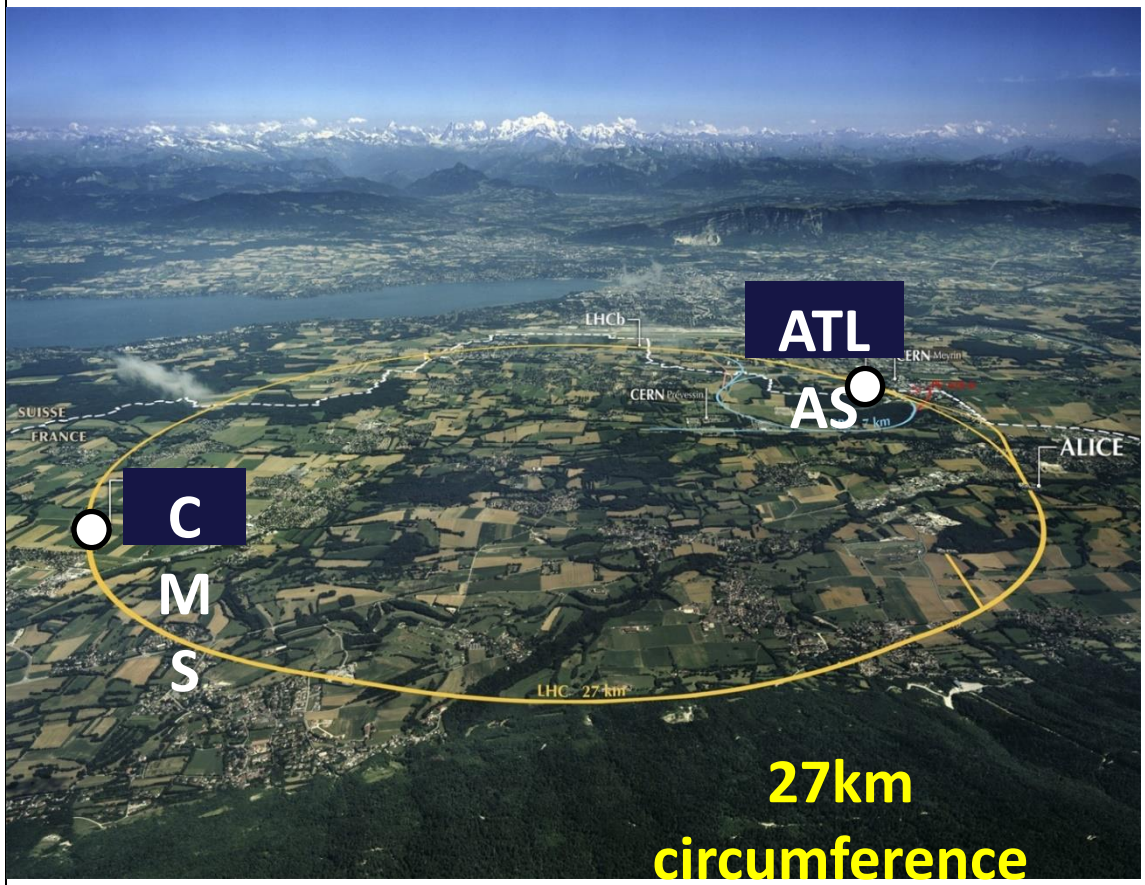
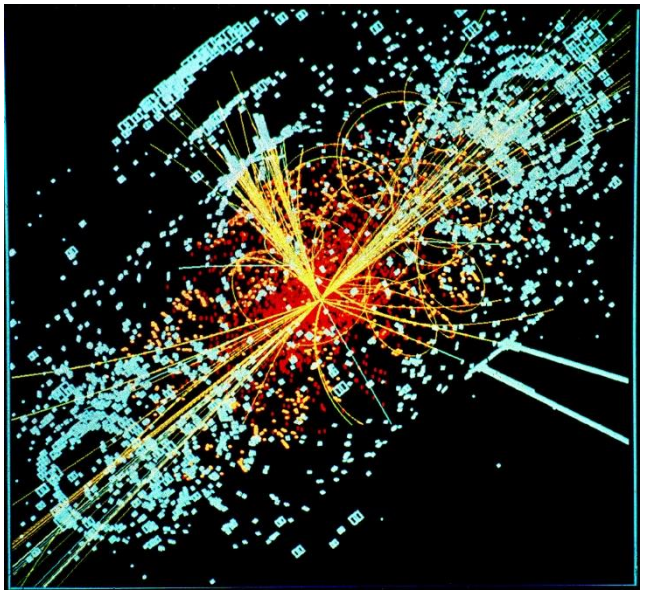
Bayesian Methods and Information Theory

CAC 고등과학원 여름학교
2023. 6. 26 - 6. 30. (Mon.-Fri.)

Yung-Kyun Noh (노영균)
Hanyang University &
Korea Institute for Advanced Study

CERN's Large Hadron Collider (LHC) for New Physics





Simulation of Higgs event at CMS detector

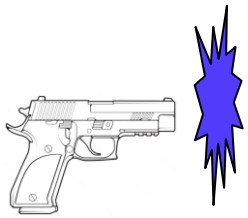
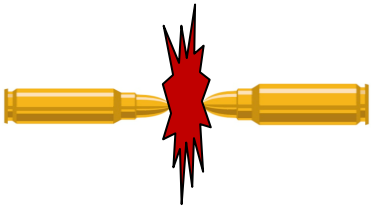
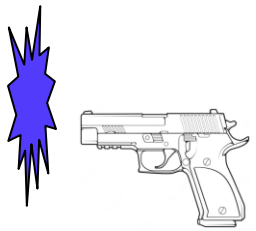
- LHC generates **50,000 TB** of data **per year** that may reveal new theories

Georgia Karagiorgi et al. (2022) Machine learning in the search for new fundamental physics, *Nature Review Physics, 4(6):399-412*

Slide credit: Dr. Cheongjae Jang

How Many Collisions for Data Production?

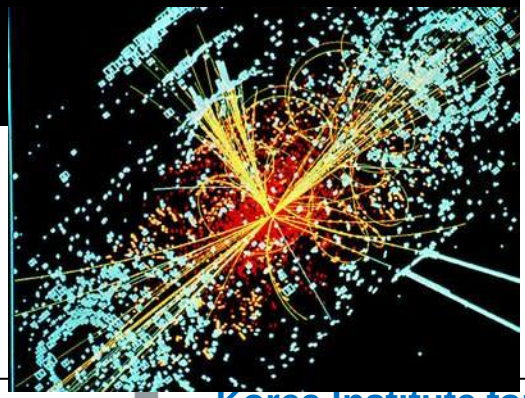
- How many collisions per second take place at the LHC?

총알 이미지: https://www.logoyogo.com/downloads/%EC%B4%9D%EC%95%8C-%EC%95%84%EC%9D%B4%EC%BD%98-%EB%A1%9C%EA%B3%A0-%EC%9D%BC%EB%9F%AC%EC%8A%A4%ED%8A%B8-ai-%EB%8B%A4%EC%9A%B4%EB%A1%9C%EB%93%9C/
총 이미지: https://kr.freepik.com/premium-vector/firearms-line-art-style-shooting-gun-weapon-illustration-vector-line-gun-illustration_32203402.htm

How Many Collisions for Data Production?

https://sg.finance.yahoo.com/news/happens-two-proton-beams-collide-204600660.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAGBHulCOwA468cineAcqonTroRccEurBqD6WEtyYRcTpNZuCrNflK8syrOHO1ean_9SPJFfrvVylzWW6ZwmTf-qMGaZT7Mduko21FYNn031BeD9jvVI3Jwp-tjO7YDK53jagzpSkG7cNXmGV1p4sAMTZE_ryga4Ar1KXni1NM5mv



How Many Collisions for Data Production?

https://www.freepik.com/free-vector/realistic-two-lights-collision-effect_4921513.htm#from_view=detail_alsolike

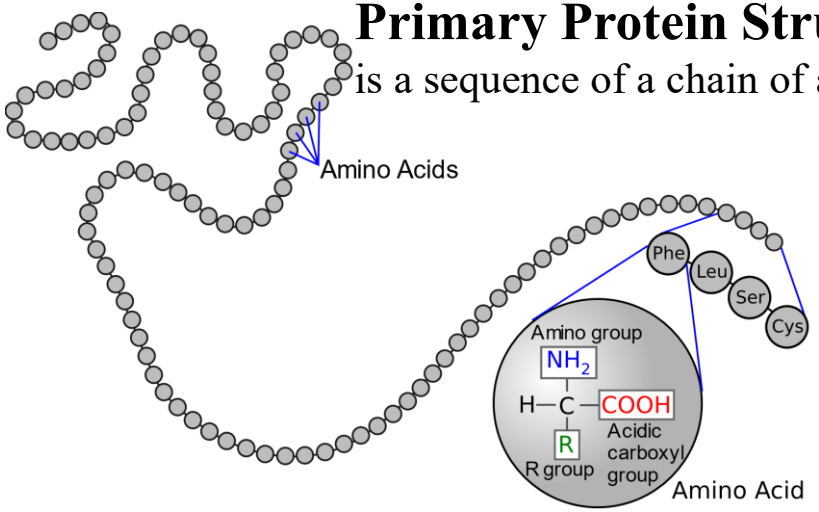
- Each beam consists of nearly 3000 bunches of particles and each bunch contains as many as 100 billion particles.
- The particles are so tiny that the chance of any two colliding is very small. When the bunches cross, there are up to 40 collisions between 200 billion particles.
- Bunches cross on average about 30 million times per second, so the LHC generates about 1 billion particle collisions per second.

Central Dogma and Amino Acid

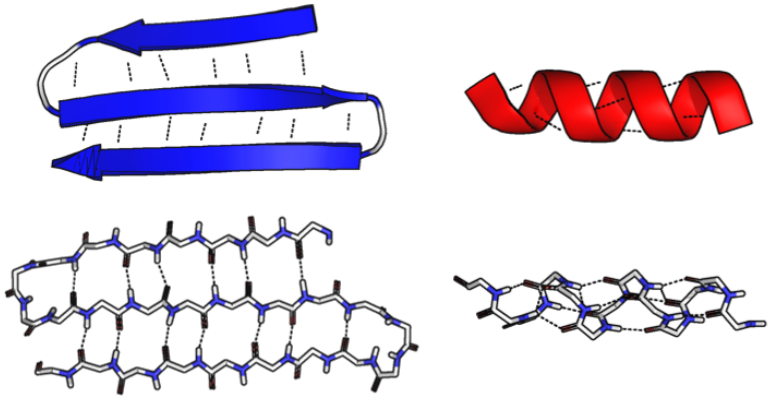
Ribosome and amino acid synthesis

Primary Protein Structure

is a sequence of a chain of amino acids



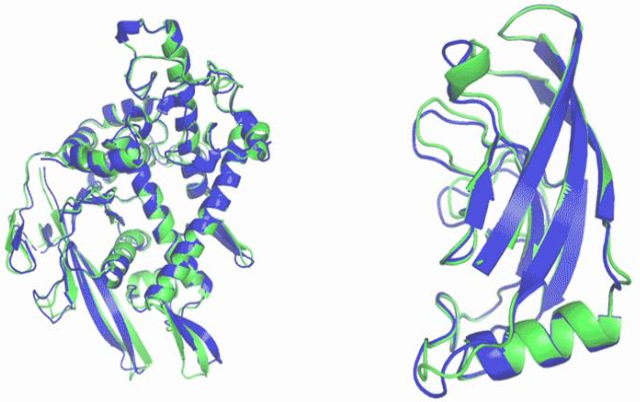
Secondary Structure



β-Sheet (3 strands)

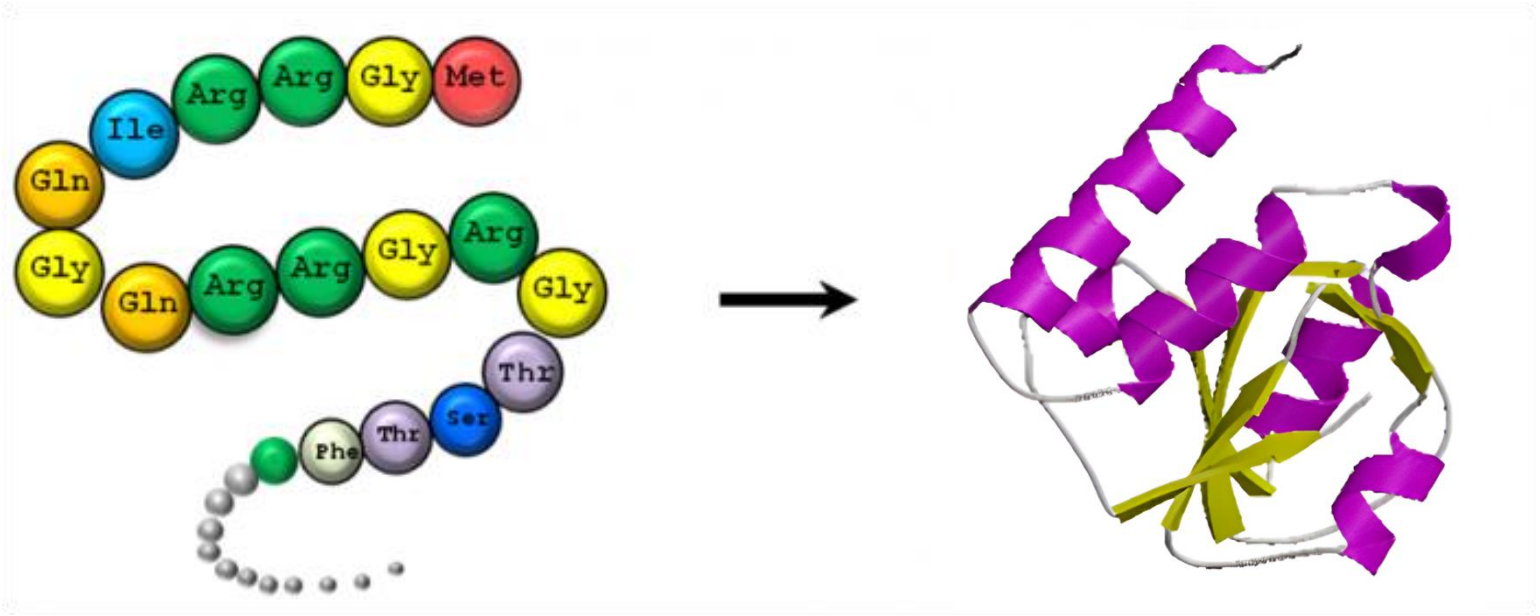
α-helix

Tertiary Structure



Three-dimensional structure

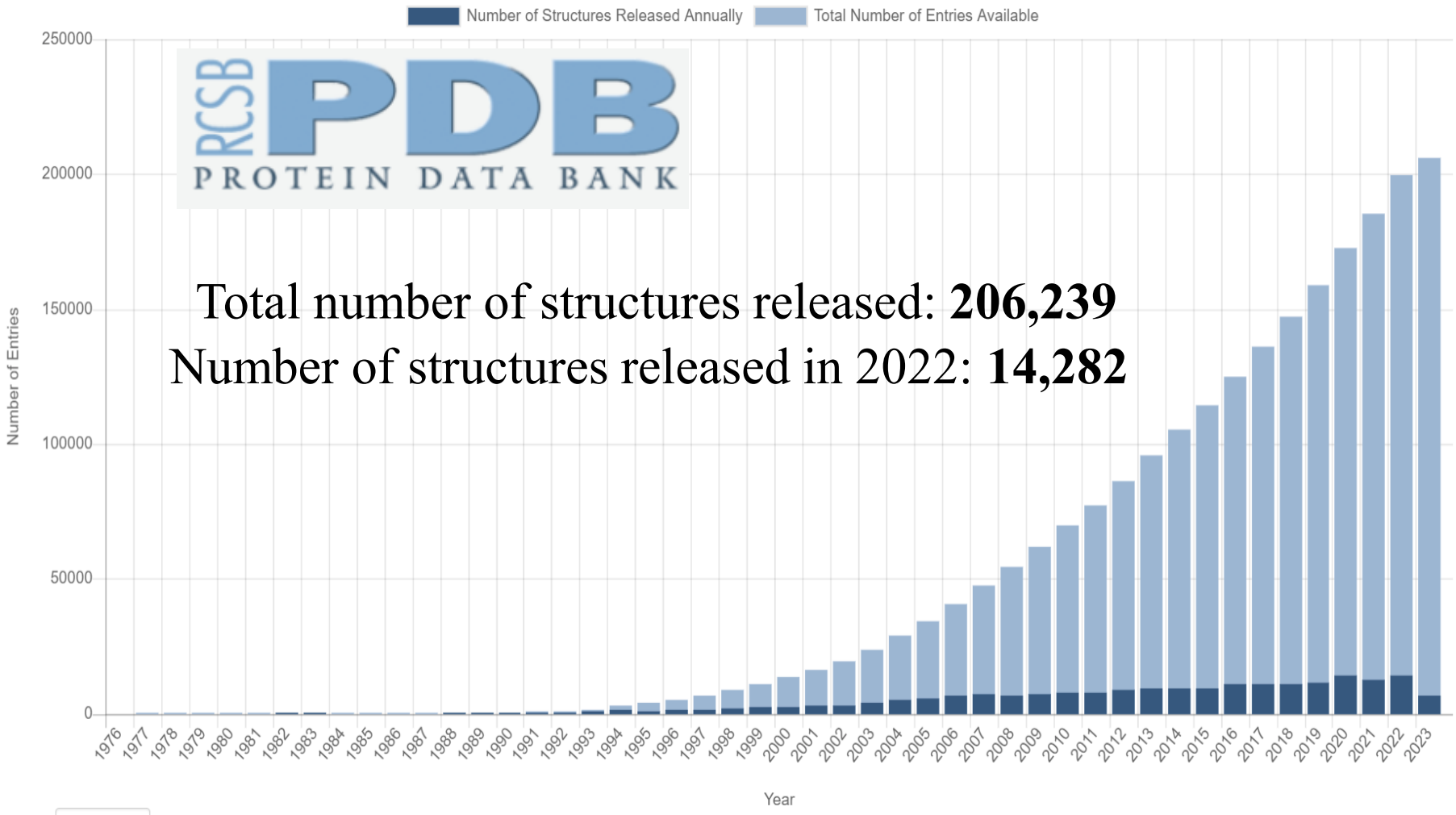
Protein Structure Prediction



Amino acid sequence **Backbone of Tertiary Structure**

https://www.ebi.ac.uk/training/online/courses/protein-classification-intro-ebi-resources/protein-classification/

Protein Data Bank

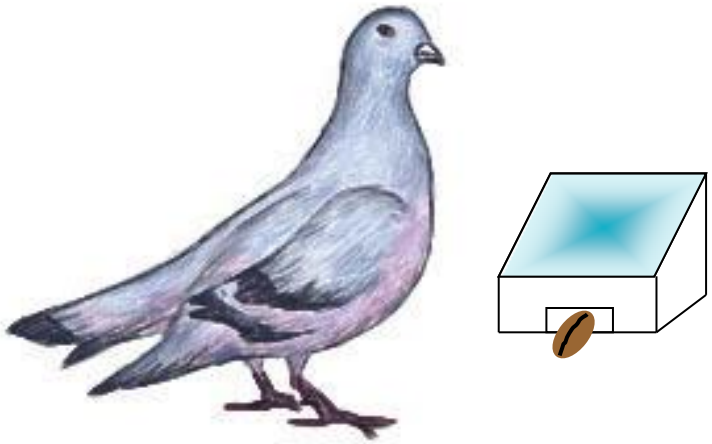


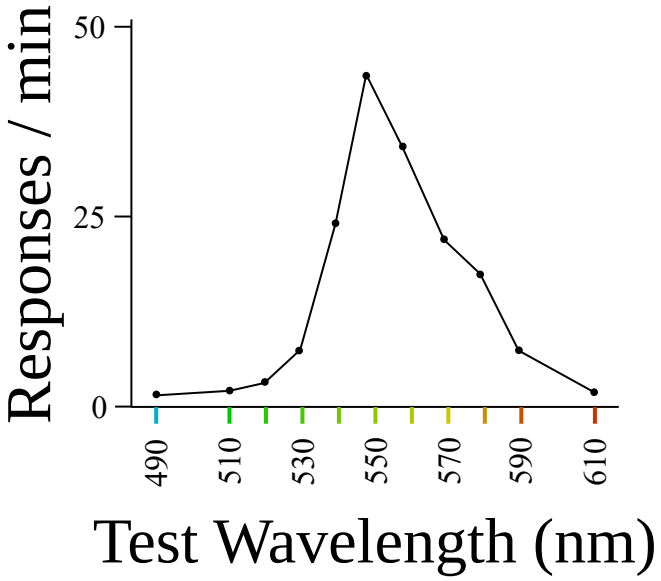
Contents

- Learning and generalization
- Bayesian approach for generalization
- Basic Bayesian formulations

Generalization of the Stimuli

- Classic research based on training & transfer with animals
- Guttman & Kalish (1956)

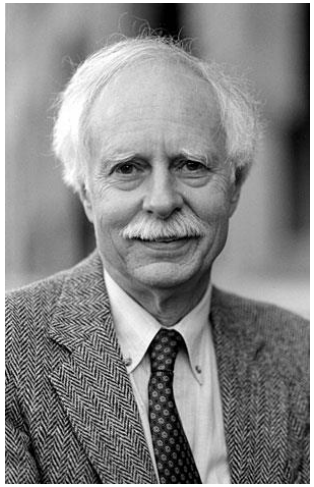




- Generalization strength depends on perceived similarity between stimuli (Shepard, 1957, 1987)

Slide credit: Matt Jones

Universal Law of Generalization

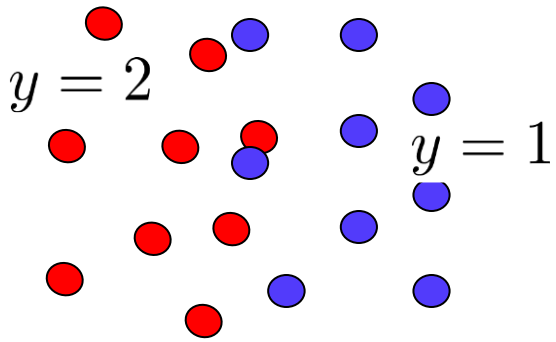


- Toward a Universal Law of Generalization for Psychological Science (Shepard, Science 1987)
  - The tercentenary of the publication, in 1687, of Newton's Principia prompts the question of whether **psychological science** has any hope of achieving a law that is comparable in generality to Newton's universal law of gravitation. Exploring the direction that currently seems most favorable for an affirmative answer, I outline empirical evidence and a theoretical rationale in support of a tentative candidate for a universal law of generalization.
  - Psychology's first general law should, I suggest, be a law of generalization.

Math for Learning

- Data

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N} \sim P \quad \text{(Regularity)}$$

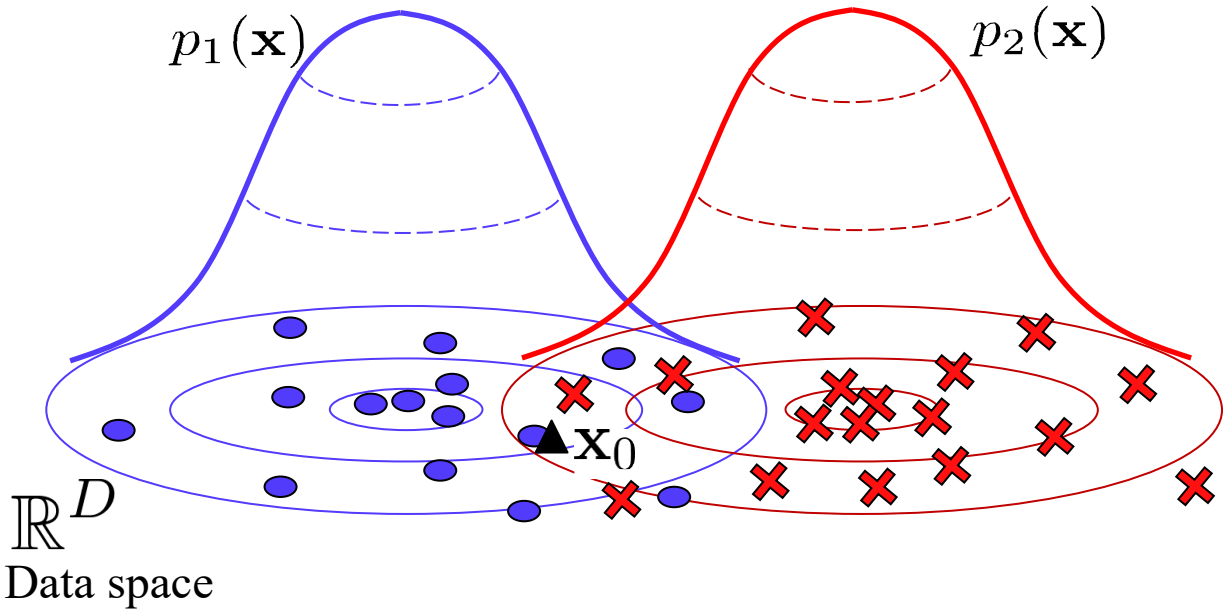


- Prediction

$$\mathbf{x} \in \mathbb{R}^D \xrightarrow{y = f(\mathbf{x})} \begin{array}{l} y \in \{1, 2, \ldots, C\} \\ y \in \mathbb{R} \end{array}$$

- Learning
  - ➢ Learn prediction function $f(\mathbf{x}) \in \mathcal{H}$ from data $\mathcal{D}$
    ($\mathcal{H}$: Hypothesis set/Candidate set)

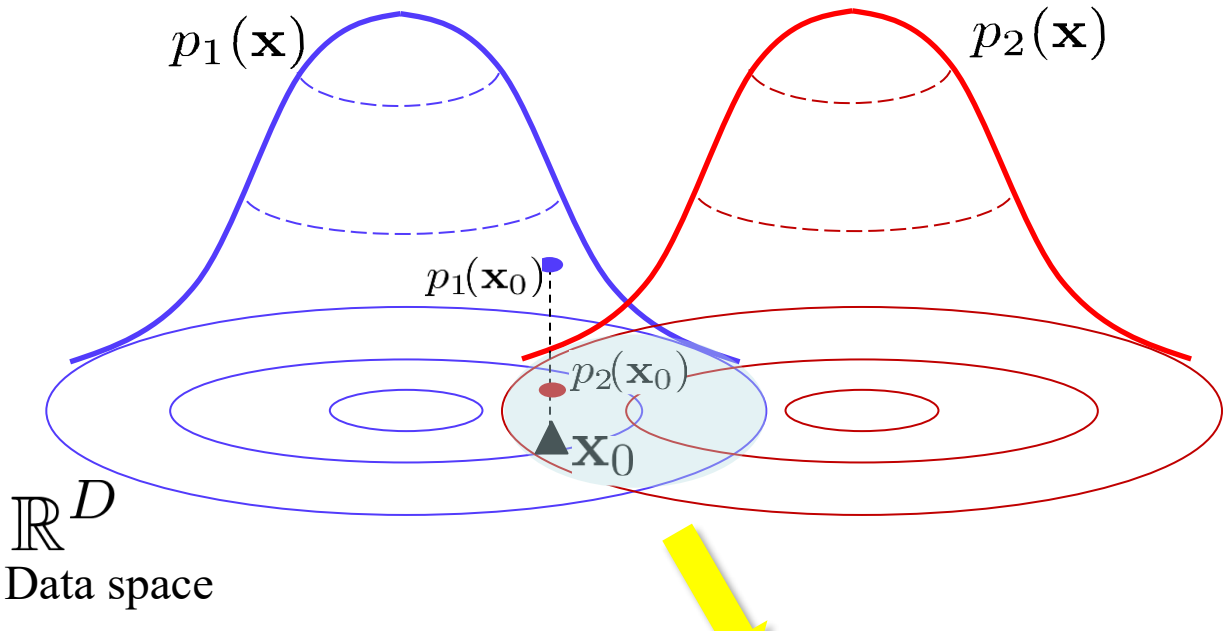
Assumption: Probabilistic REGULARITY



$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N} \sim p_1(\mathbf{x}), p_2(\mathbf{x})$$

Assumption: Probabilistic REGULARITY

- Bayes Error



$$E_{Bayes} = \frac{1}{2}\int \min[p_1, p_2]d\mathbf{x}$$

Model for Generalization

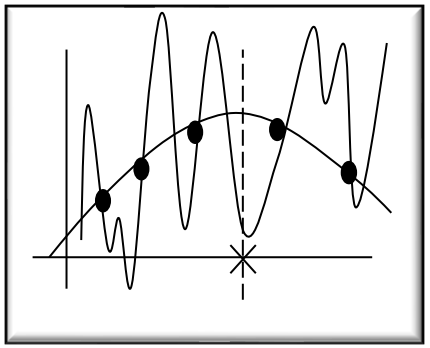
- Set of candidate functions

$$\mathcal{H} = \{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_{N_\mathcal{H}}(\mathbf{x})\}$$

In general, $N_\mathcal{H}$ is infinite.

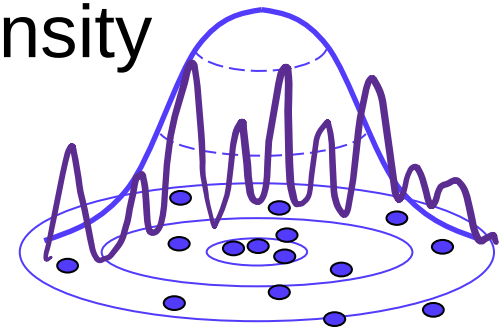
– $h_i(\mathbf{x})$ can be a prediction function

$$h_i(\mathbf{x}) \rightarrow y$$



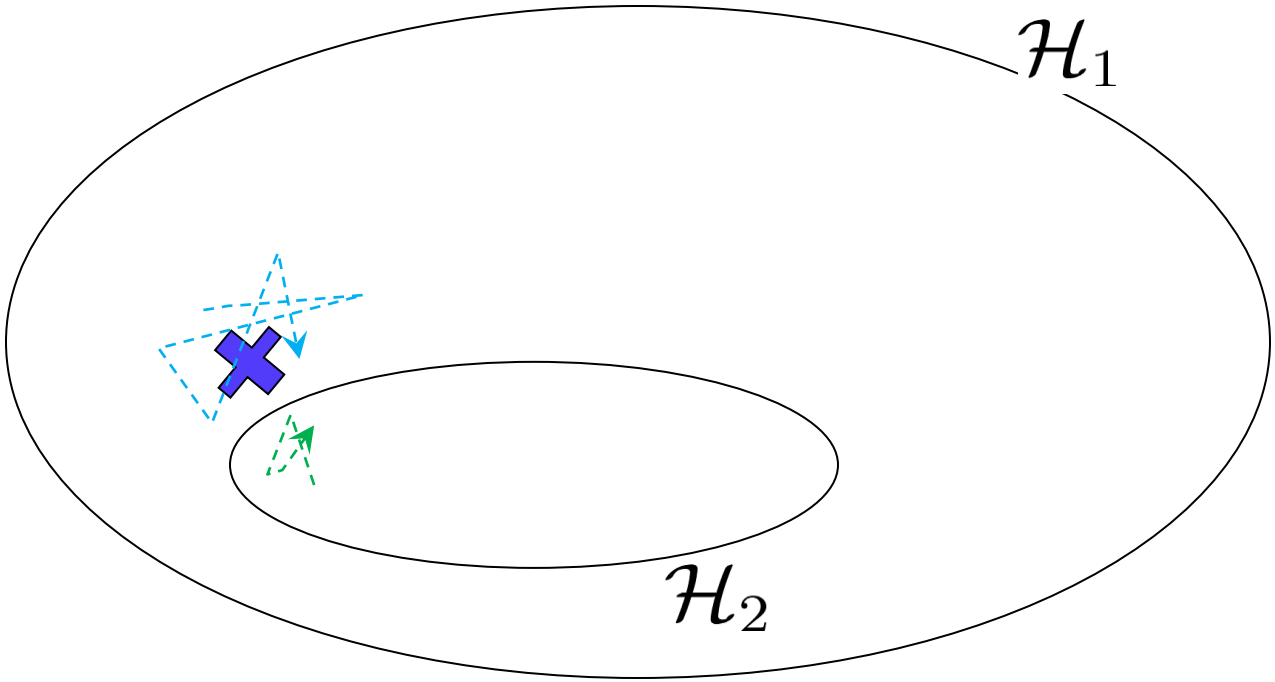
– $h_i(\mathbf{x})$ can be for probability density

$$h_i(\mathbf{x}) \rightarrow p(\mathbf{x})$$



Model – Confine Hypothesis Space $\mathcal{H}$

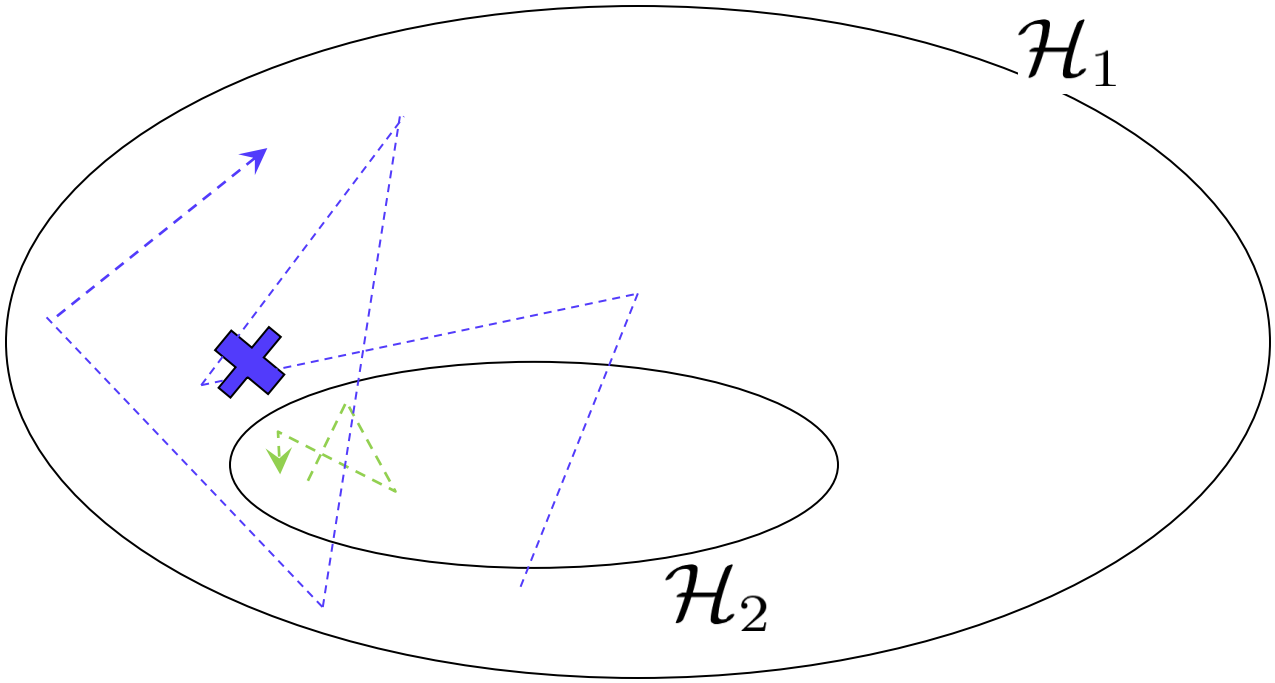
- Estimation with large number of data



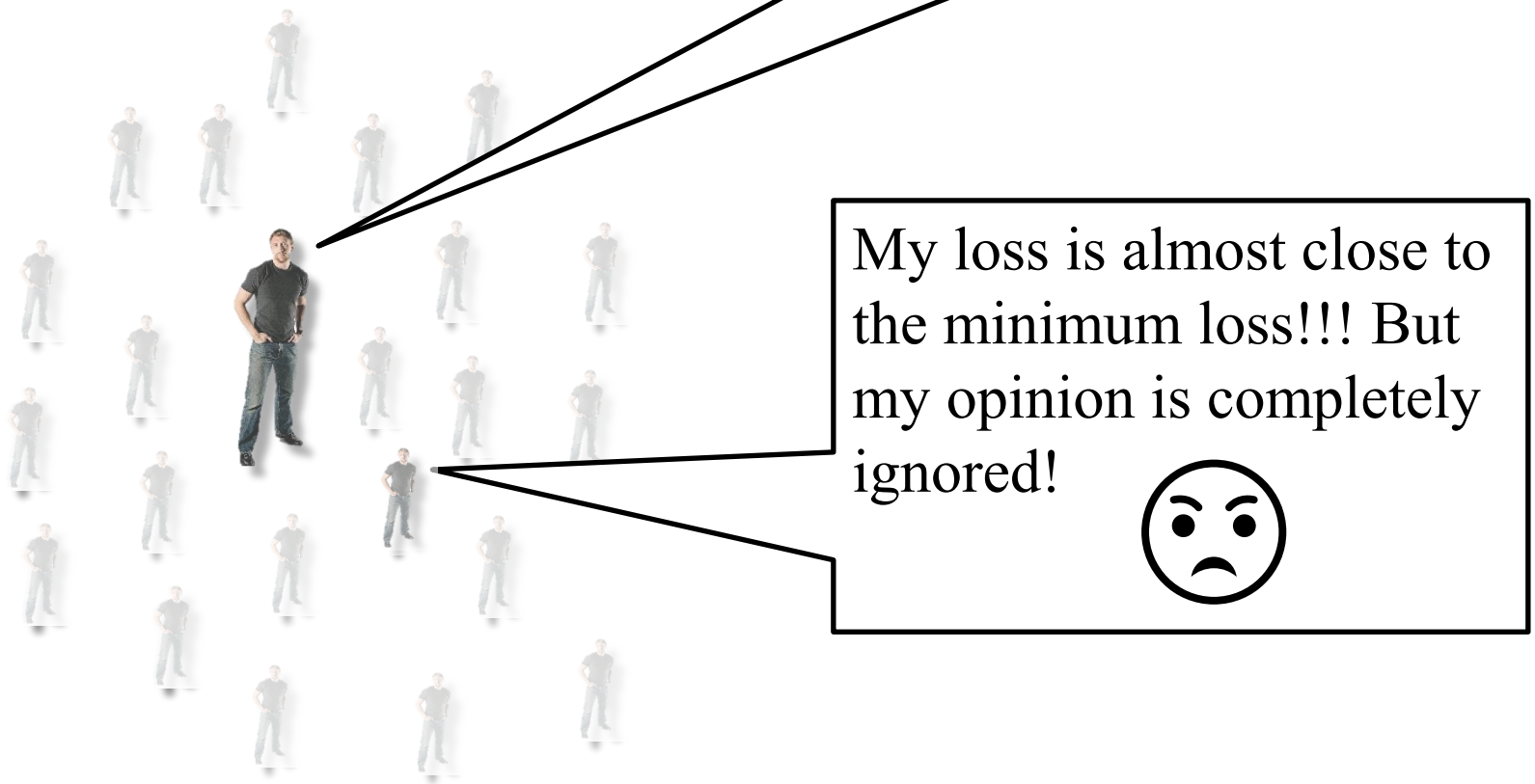
Optimal solution (✖) and the selected solutions of different realizations

Model – Confine Hypothesis Space $\mathcal{H}$

- Estimation with small number of data



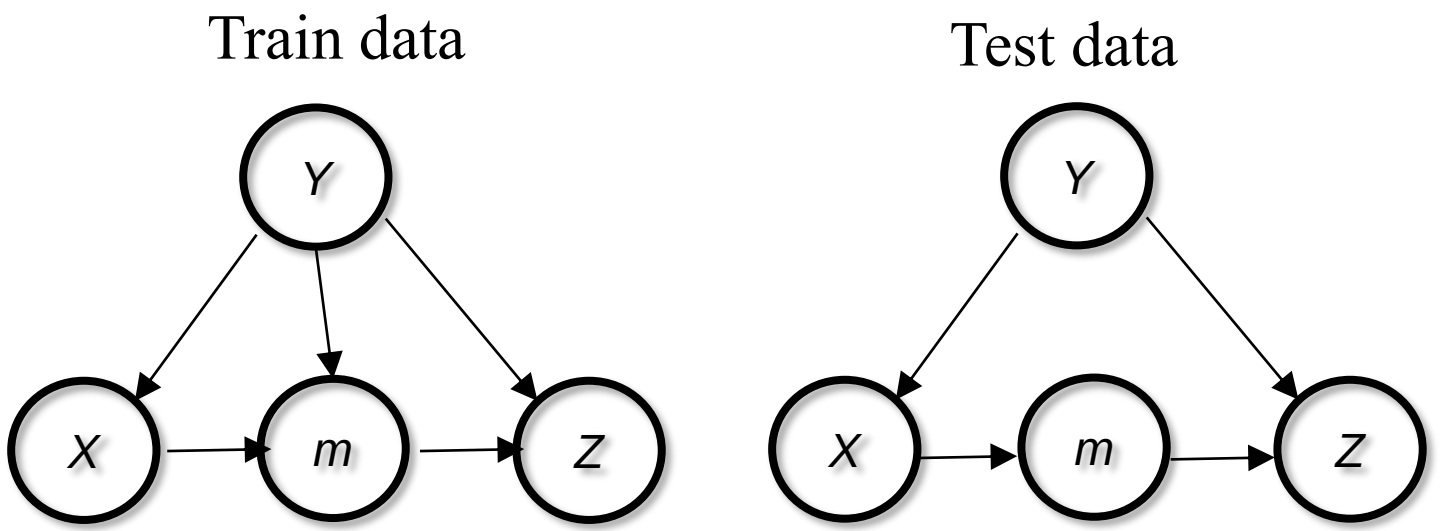
Optimal solution (✖) and the selected solutions of different realizations


I have the minimum loss.
My loss is almost close to the minimum loss!!! But my opinion is completely ignored!

We all have small losses, and our opinions are combined to make decisions!

Learning from Real World

- Traditional machine learning algorithm is considered as glorified curve fitting (by Judea Pearl).
- We need mathematical tools to learn from a complex nature.



Computations for Machine Learning

For given model $\mathcal{H} = \left\{ f(\mathbf{x};\theta) \;\middle|\; \theta \in \Theta \right\}$

- Optimization (Frequentist)
  - Find $f(\mathbf{x};\theta^*)$ s.t.

$$\theta^* = \arg\min_{\theta\in\Theta} \sum_{i=1}^{N} \left(f(\mathbf{x}_i;\theta) - y_i \right)^2 + \lambda\, \Omega(\theta)$$

- Integration (Bayesian)
  - Obtain y from

$$\widehat{y} = \arg\max_{y} P(y|\mathbf{x},\mathcal{D}) = \arg\max_{y} \int P(y|\mathbf{x};\theta)\, p(\theta|\mathcal{D})\, d\theta$$

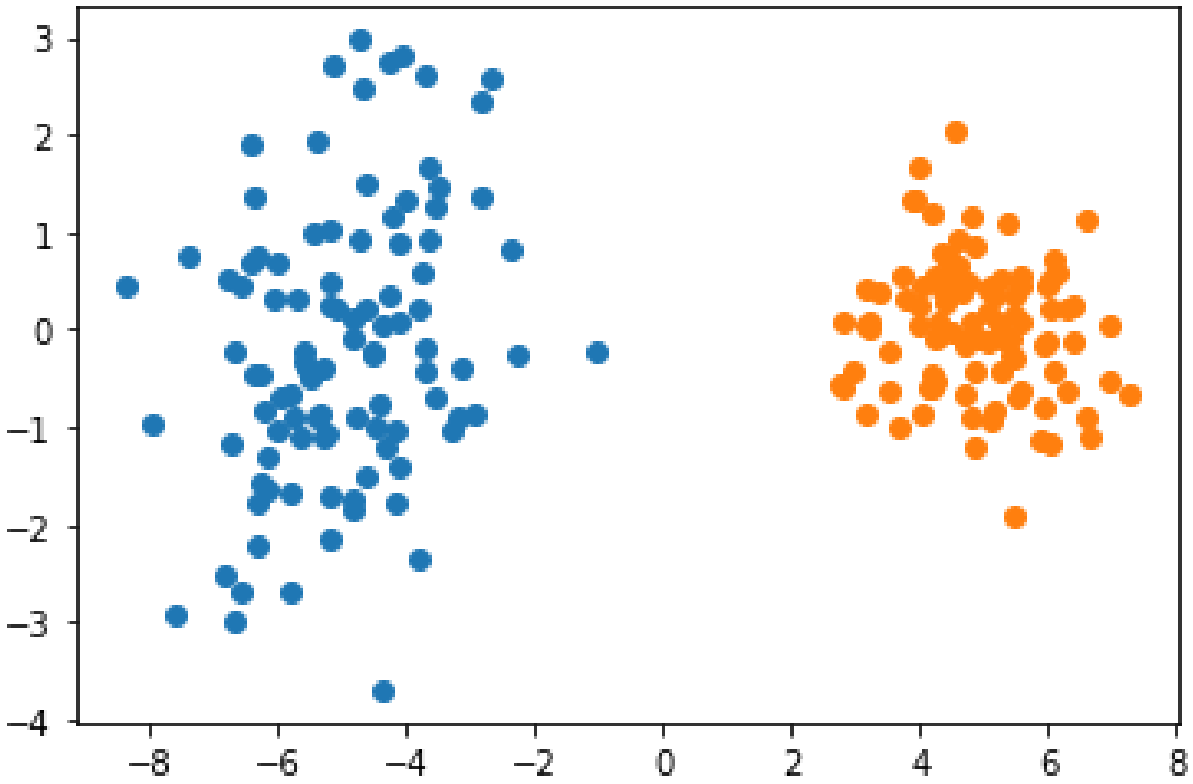
Classification

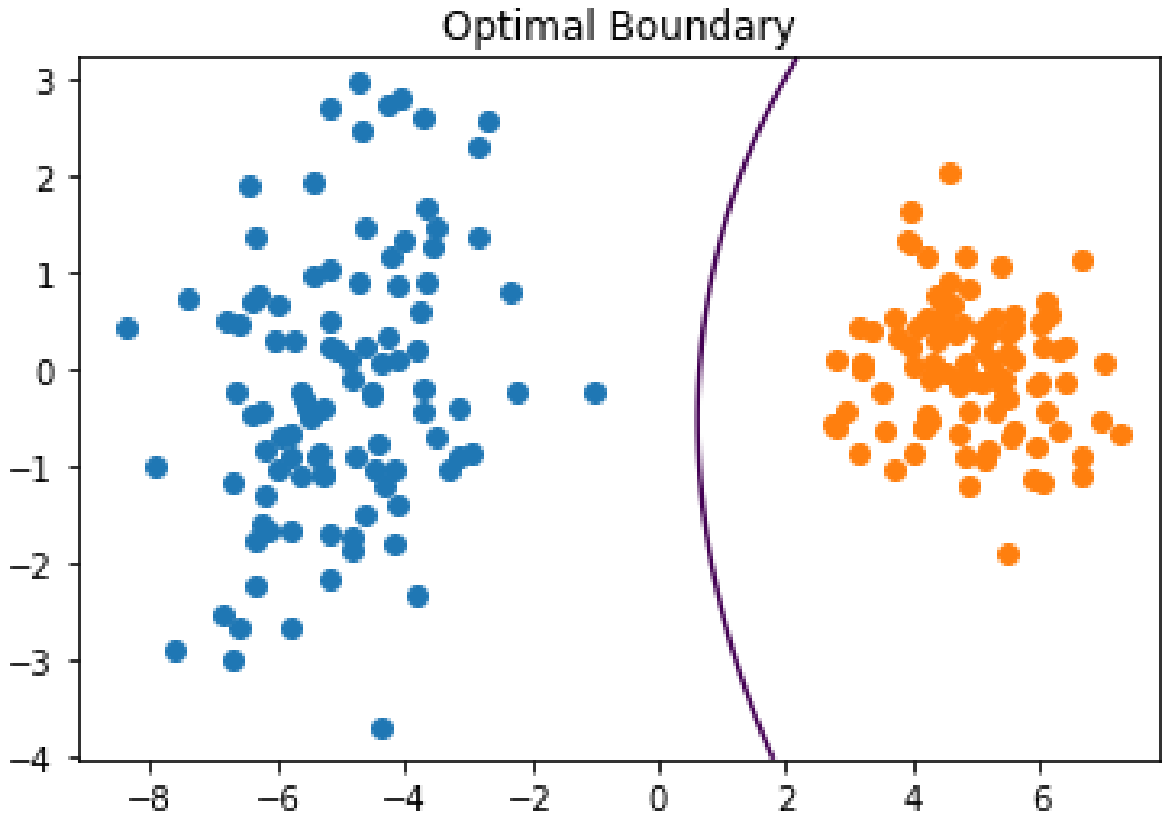
$$\widehat{y} = \int y\; p(y|\mathbf{x},\mathcal{D})\, dy = \iint y\; p(y|\mathbf{x},\theta)\, p(\theta|\mathcal{D})\, d\theta\, dy$$

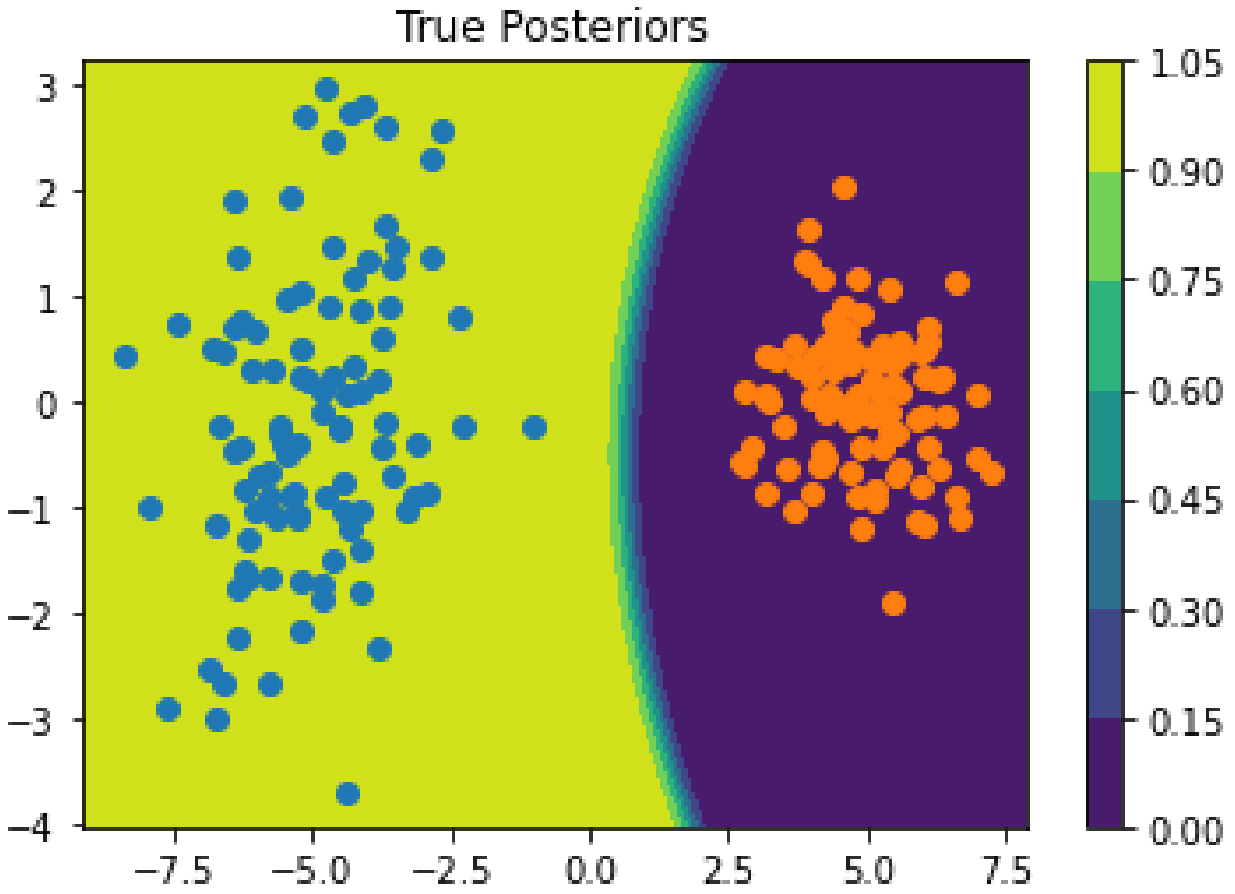
Regression

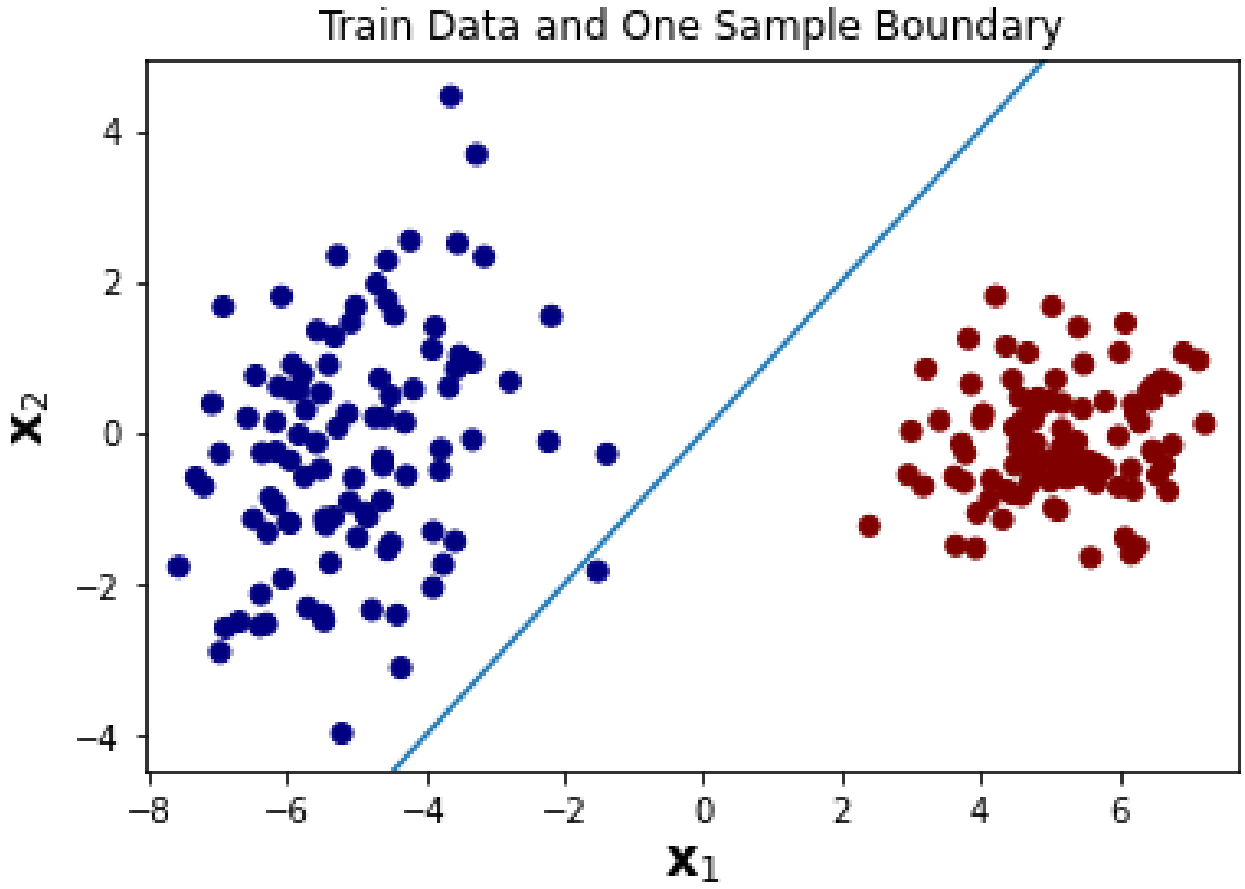
$f(\mathbf{x};\theta)$

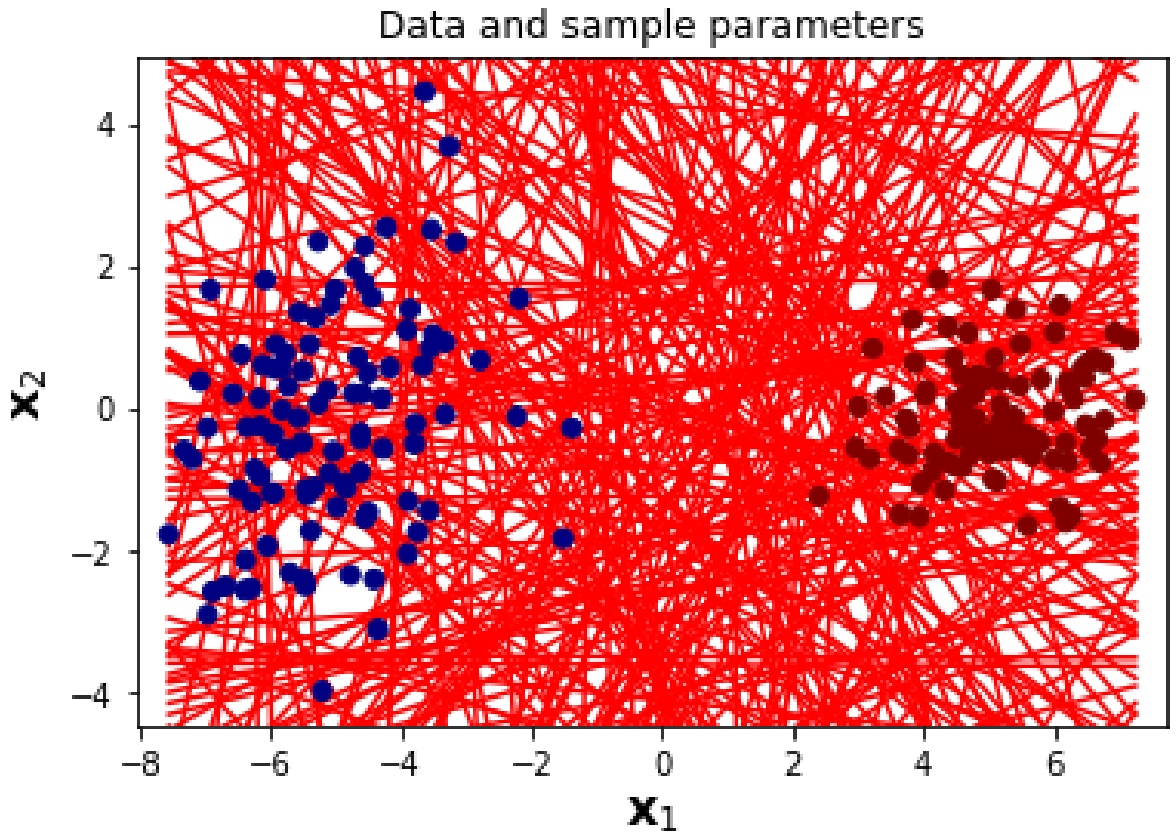
Who is more responsible?

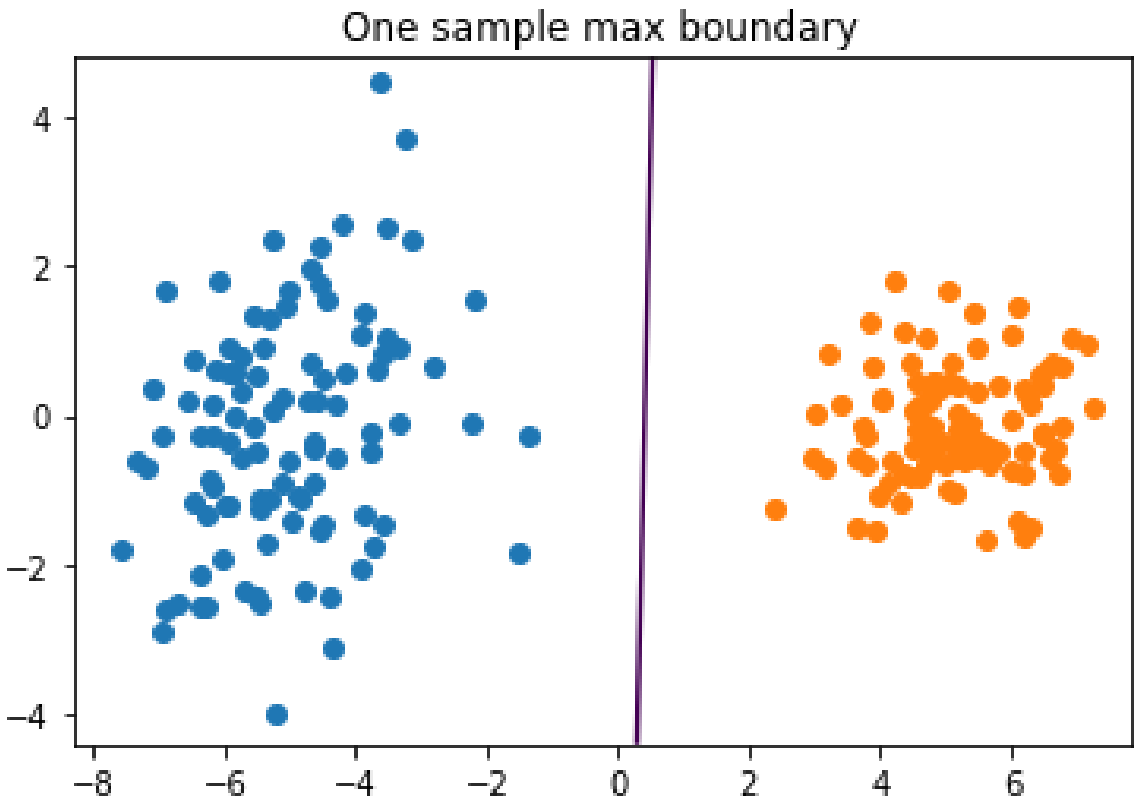

3
2
1
0
−1
−2
−3
−4
−8
−6
−4
−2
0
2
4
6
8

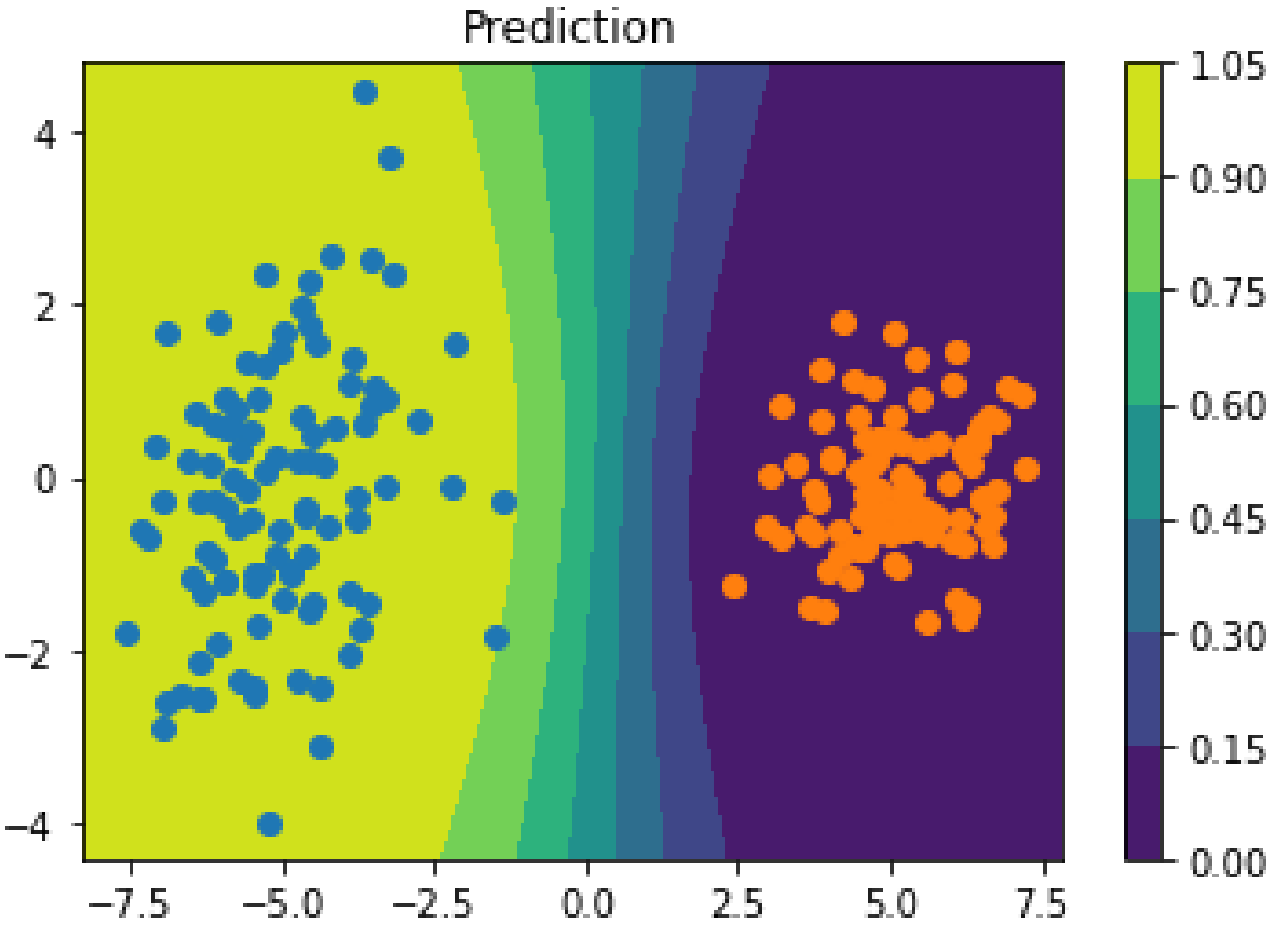

Optimal Boundary
3
2
1
0
−1
−2
−3
−4
−8
−6
−4
−2
0
2
4
6

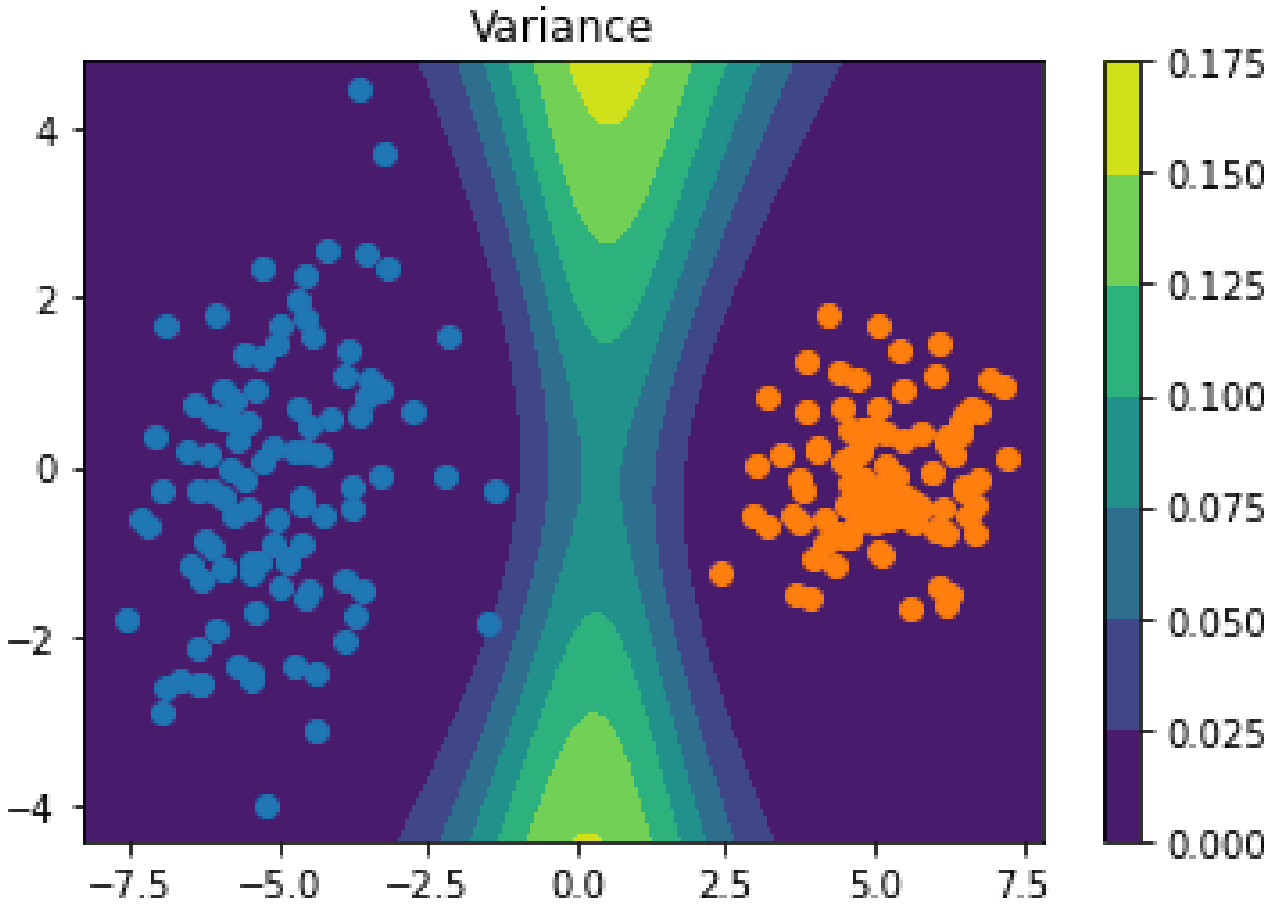

True Posteriors
3
2
1
0
−1
−2
−3
−4
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
1.05
0.90
0.75
0.60
0.45
0.30
0.15
0.00


Train Data and One Sample Boundary
4
2
0
−2
−4
−8
−6
−4
−2
0
2
4
6
x_2
x_1

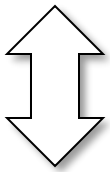

Data and sample parameters
4
2
0
−2
−4
−8
−6
−4
−2
0
2
4
6
x_2
x_1


One sample max boundary
4
2
0
−2
−4
−8
−6
−4
−2
0
2
4
6

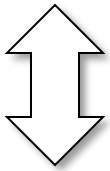

Prediction
4
2
0
−2
−4
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
1.05
0.90
0.75
0.60
0.45
0.30
0.15
0.00


Variance
4
2
0
−2
−4
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000

- Maximum likelihood solution



- Maximum A Posteriori solution (Regularized maximum likelihood solution)



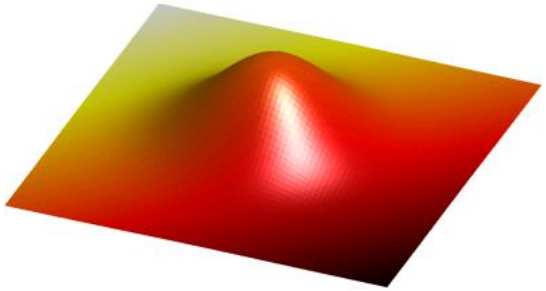
- Bayesian inference solution

Gaussian Densities

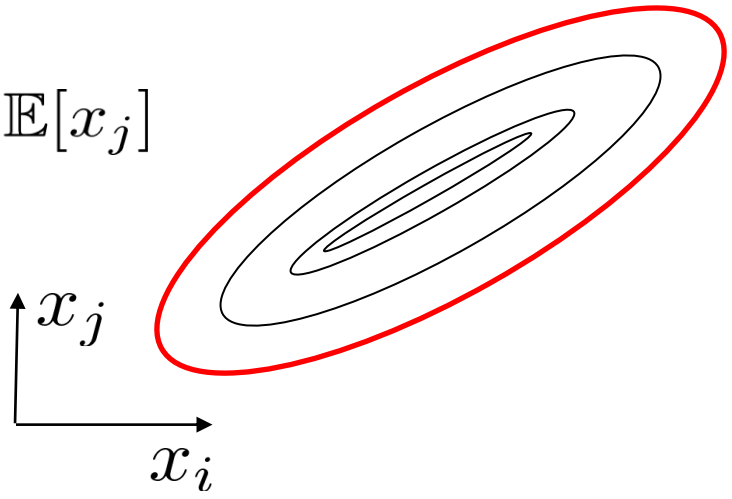
Gaussian Random Variable

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^{D}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix}$$



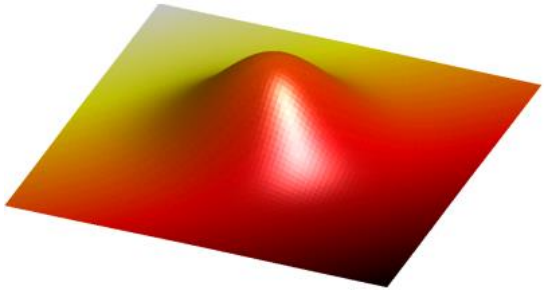
$$[\Sigma]_{ij} = \mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j]$$



Gaussian Random Variable

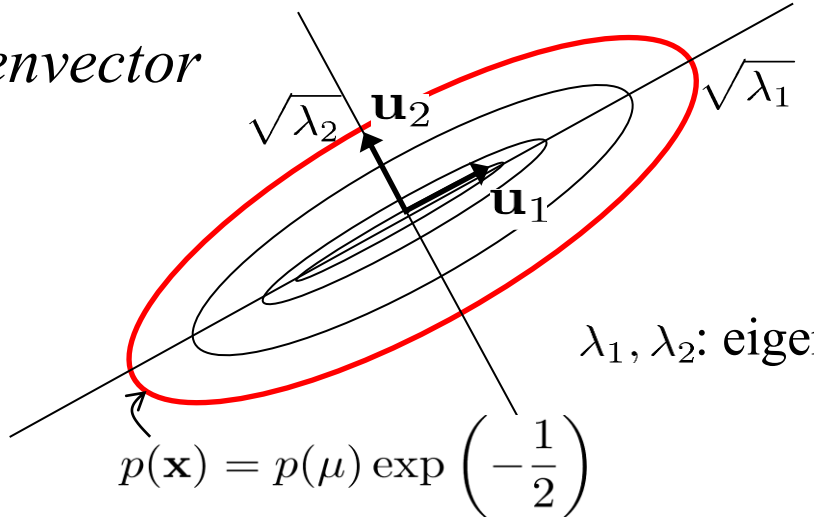
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^{D}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix}$$

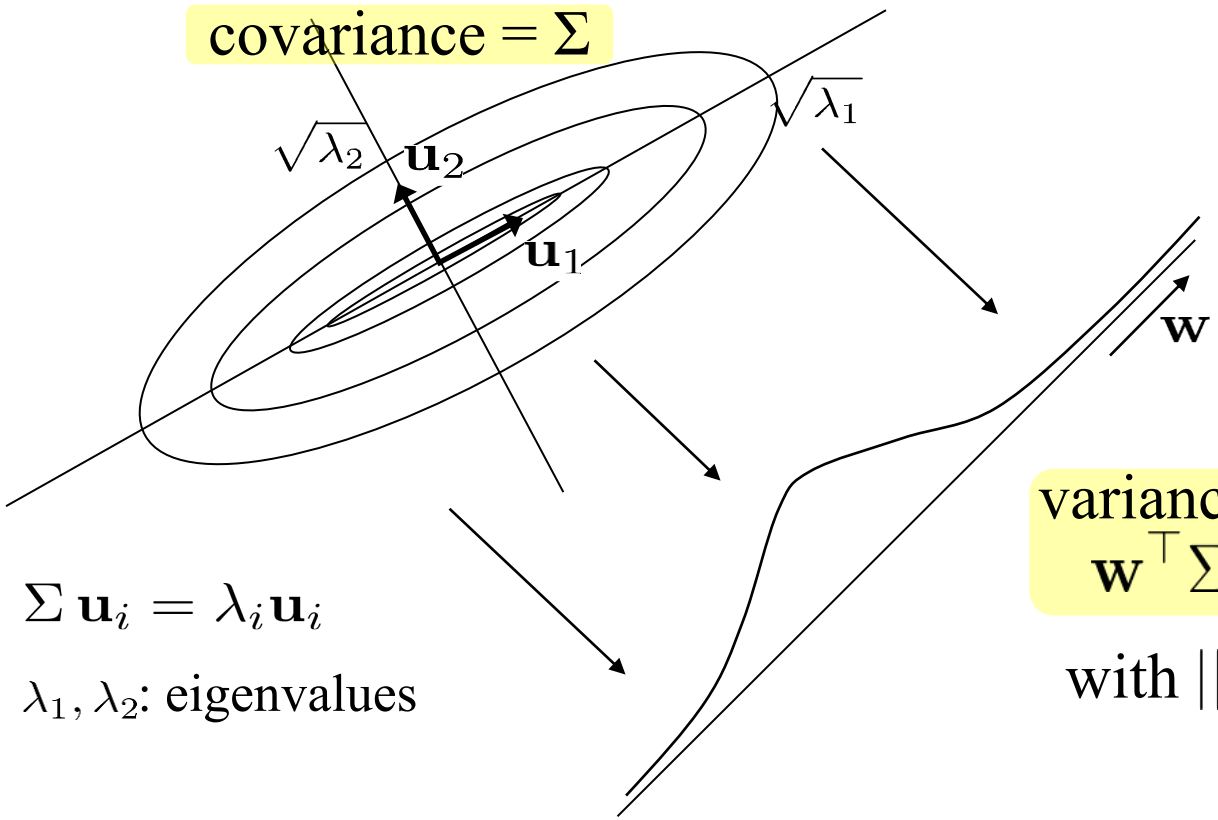


Principal axes are the eigenvector directions of Σ

$$\Sigma\,\mathbf{u}_i = \lambda_i \mathbf{u}_i$$



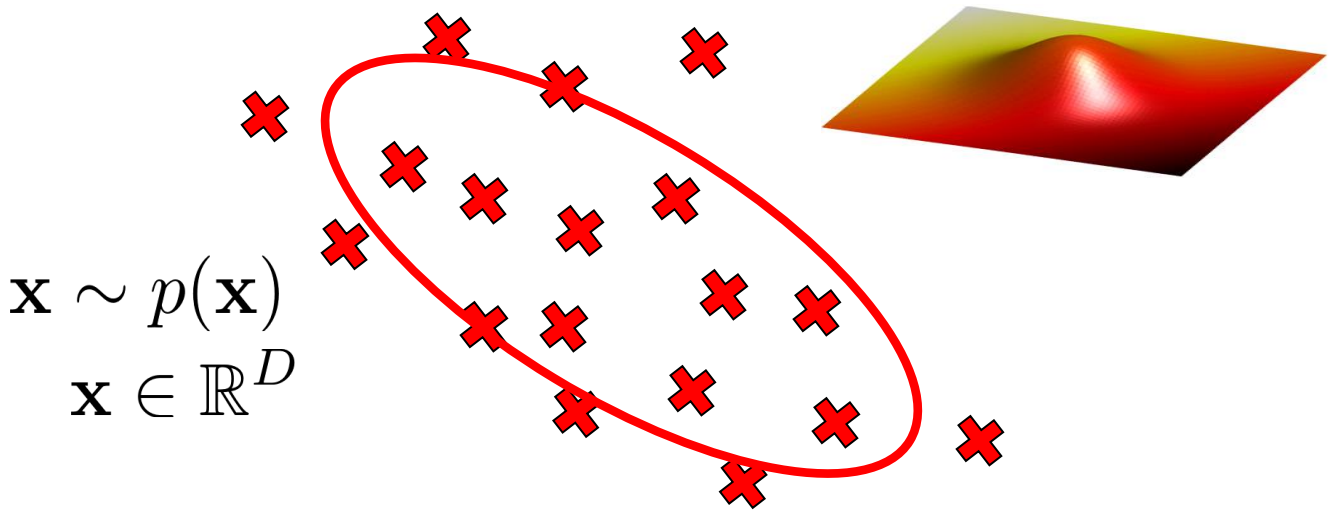
Covariance Matrix and Projection



PARAMETER ESTIMATION

Motivation – Parameter Estimation

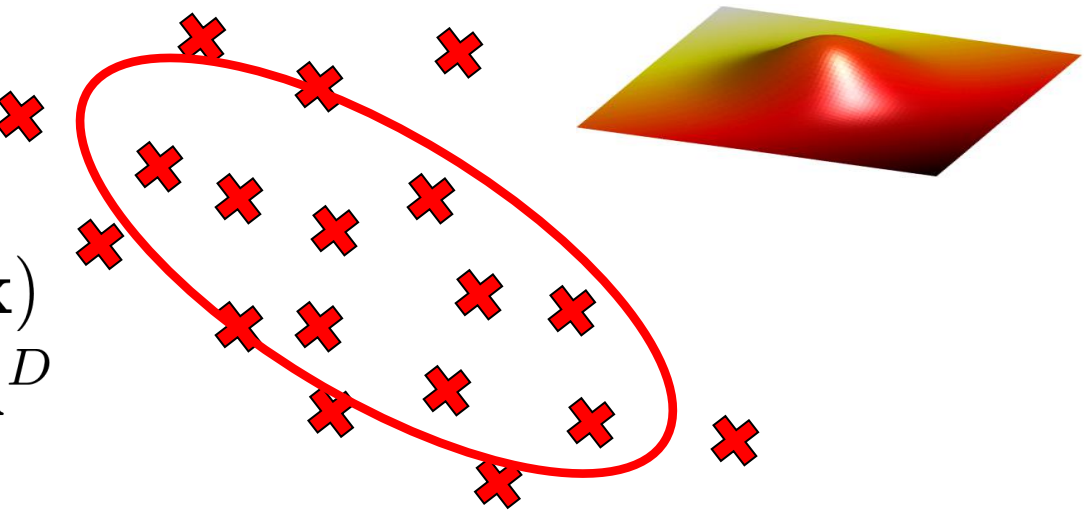
- Parameter estimation is an optimization problem



$\widehat{p}(\mathbf{x})$: estimated probability density function,
in other words, density function that fits data the most

Maximum Likelihood Estimation

- Parameter estimation is an optimization problem



$$\mathbf{x} \sim p(\mathbf{x})$$
$$\mathbf{x} \in \mathbb{R}^D$$

$$\widehat{p}(\mathbf{x}) = p(\mathbf{x}|\widehat{\mu}, \widehat{\Sigma})$$
$$\widehat{\mu}, \widehat{\Sigma} = \arg\max_{\mu,\Sigma} p(\mathbf{x}|\mu, \Sigma)$$

Maximum Likelihood for Gaussian

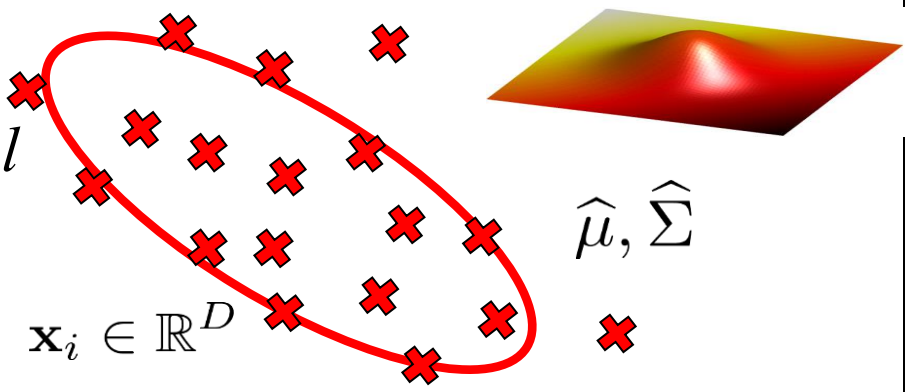
$$p(\mathbf{x}|\mu,\Sigma) = \frac{1}{\sqrt{2\pi}^{D}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

- With optimal parameters satisfying

$$\widehat{\mu},\widehat{\Sigma} = \arg\max_{\mu,\Sigma} p(X|\mu,\Sigma) = \arg\max_{\mu,\Sigma}\prod_{i=1}^{N} p(\mathbf{x}_i|\mu,\Sigma)$$

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \qquad \widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\widehat{\mu})(\mathbf{x}_i-\widehat{\mu})^{\top}$$

Empirical mean and empirical covariance are the maximum likelihood solutions.



Maximum Likelihood for Gaussian

$$p(\mathbf{x}|\mu,\Sigma) = \frac{1}{\sqrt{2\pi}^{D}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$\nabla_{\theta}\ln p(X|\theta) = \vec{0} \qquad \theta = \mu, \Sigma$$

$$\frac{\partial \ln p(X|\mu,\Sigma)}{\partial \mu} = 0 \Longrightarrow \widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

$$\frac{\partial \ln p(X|\mu,\Sigma)}{\partial \Sigma} = 0 \Longrightarrow \widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\widehat{\mu})(\mathbf{x}_i-\widehat{\mu})^{\top}$$

Maximum A Posteriori (MAP) Estimation

- MAP estimation

$$\theta^* = \arg\max_{\theta} p(\theta|X) \qquad \text{cf) } \theta^* = \arg\max_{\theta} p(X|\theta)$$

- Likelihood (Model): $p(\mathbf{x}|\theta)$
- Prior: $p(\theta)$
- Bayes rule:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

Maximum A Posteriori (MAP) Estimation for Gaussian

$$p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

$$\widehat{\mu} = \arg\max_{\mu} p(\mu|X) = \arg\max_{\mu} \prod_{i=1}^{N} p(\mu|x_i)$$

- Let the prior

$$p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)$$

- The posterior can be calculated using

$$p(\mu|X) \propto p(X|\mu)p(\mu) = \prod_{i=1}^{N} p(x_i|\mu)p(\mu) \sim \mathcal{N}(\mu_n, \sigma_n^2)$$

Maximum A Posteriori (MAP) Estimation for Gaussian

$$\prod_{i=1}^{N} p(x_i|\mu)p(\mu) = \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)\right] \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)$$

$$\propto \quad \exp\left(-\frac{1}{2}\left(\sum \frac{(x_i-\mu)^2}{\sigma^2} + \frac{\mu-\mu_0}{\sigma_0^2}\right)\right)$$

$$\propto \quad \exp\left(-\frac{1}{2}\left(\mu^2\left[\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right] - 2\mu\left[\frac{1}{\sigma^2}\sum x_i + \frac{\mu_0}{\sigma_0}\right]\right)\right)$$

$$\propto \exp(-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2)$$

Maximum A Posteriori (MAP) Estimation for Gaussian

- Posterior density

$$\propto \quad \exp\left(-\frac{1}{2}\left(\mu^2\left[\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right]-2\mu\left[\frac{1}{\sigma^2}\underbrace{\sum x_i}_{=N\cdot\widehat{\mu}_{ML}}+\frac{\mu_0}{\sigma_0}\right]\right)\right)$$

  - Caution: Posterior of μ, not the density function of x

- MAP of μ = Mean of μ = μ_n

$$\mu_n = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\widehat{\mu}_{ML} + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$$

MLE vs. MAP

- For Gaussian
  - When N is just a few (say N = 5),

$$\sigma_0^2 = 5, \sigma^2 = 3$$

$$\mu_n = \underbrace{\frac{25}{5 \cdot 5 + 3} \widehat{\mu}_{ML}}_{\text{Dominant}} + \frac{3}{5 \cdot 5 + 3} \mu_0$$

$$\sigma_n = \frac{5 \cdot 3}{25 + 3} \fallingdotseq 0.54$$

MLE vs. MAP

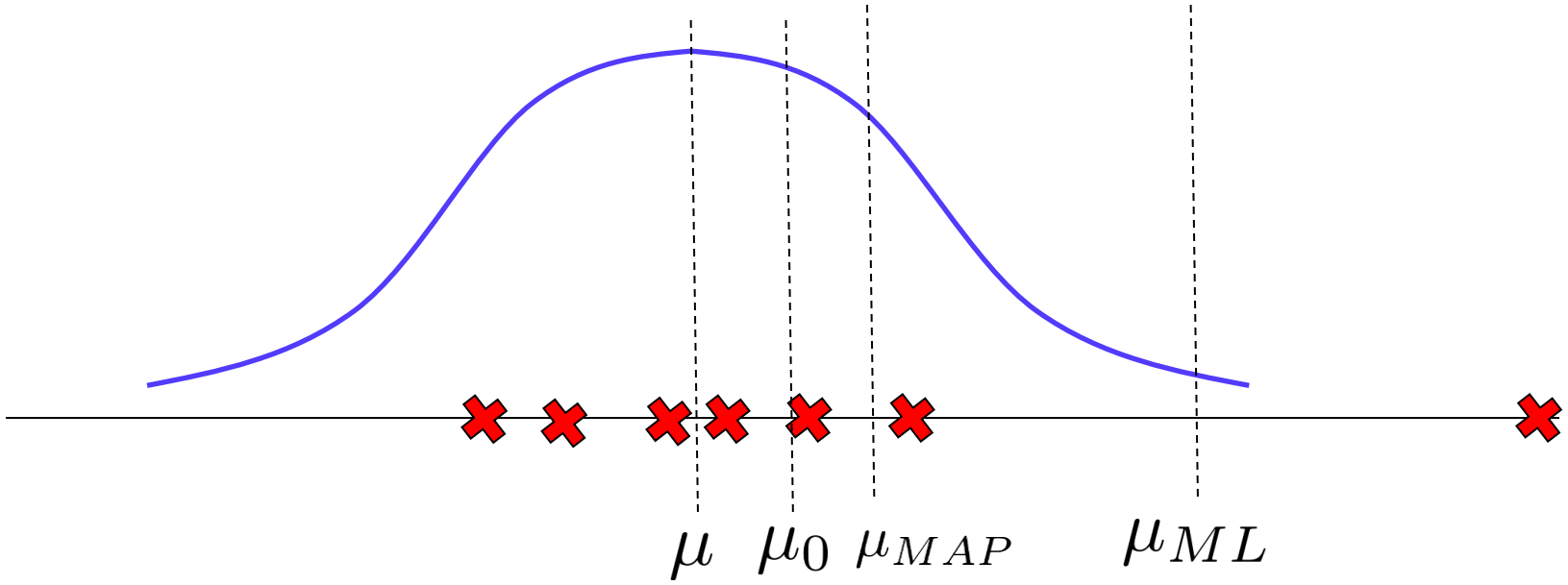
- For Gaussian
  - When we have a few outliers

$$\sigma_0^2 = 5, \sigma^2 = 100$$

$$\mu_n = \underbrace{\frac{25}{5 \cdot 5 + 100}}_{\text{Dominant (learn from } \mu_0\text{)}} \widehat{\mu}_{ML} + \frac{100}{5 \cdot 5 + 100} \mu_0$$

$$\sigma_n = \frac{5 \cdot 100}{25 + 100} \fallingdotseq 4$$

MLE vs. MAP



Bayesian Inference (by Integration)

- The final standard method of prediction is to use Bayesian inference instead of estimating the parameter point.
  - Do not insert a point $\widehat{\mu}_{MAP}$ directly but marginalize.

$$p(x|X) = \int p(x|\mu)p(\mu|X)d\mu$$

$$= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{1}{2\sigma_n}(\mu-\mu_n)^2\right) d\mu$$

$$= \frac{1}{\sqrt{2\pi(\sigma^2+\sigma_n^2)}} \exp\left(-\frac{1}{2(\sigma^2+\sigma_n^2)}(x-\mu)^2\right)$$

$$= \mathcal{N}(\mu_n, \sigma^2 + \sigma_n^2)$$

Uncertainty for prediction

Inference using Gaussians

Decomposition for Inference

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \begin{matrix} \mathbf{x}_a \in \mathbb{R}^{D_a} \\ \mathbf{x}_b \in \mathbb{R}^{D_b} \end{matrix} \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix}$$

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$\mu_{a|b} = \Sigma_{ab}\Sigma_b^{-1}(\mathbf{x}_b - \mu_b)$$

$$= C \exp\left(-\frac{1}{2}\left(\mathbf{x}_a - \Sigma_{ab}\Sigma_b^{-1}(\mathbf{x}_b-\mu_b)\right)^\top \left(\Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ba}\right)^{-1} \left(\mathbf{x}_a - \Sigma_{ab}\Sigma_b^{-1}(\mathbf{x}_b-\mu_b)\right)\right.$$

$$\left. -\frac{1}{2}(\mathbf{x}_b-\mu_b)^\top \Sigma_b^{-1}(\mathbf{x}_b-\mu_b)\right)$$

$$\Sigma_{a|b} = \Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ba}$$

$$= C \exp\left(-\frac{1}{2}(\mathbf{x}_a - \mu_{a|b})^\top \Sigma_{a|b}^{-1}(\mathbf{x}_a - \mu_{a|b})\right.$$

$$\left. -\frac{1}{2}(\mathbf{x}_b-\mu_b)^\top \Sigma_b^{-1}(\mathbf{x}_b-\mu_b)\right)$$

Decomposition for Inference

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \begin{matrix} \mathbf{x}_a \in \mathbb{R}^{D_a} \\ \mathbf{x}_b \in \mathbb{R}^{D_b} \end{matrix} \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix}$$

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$= C_1 \exp\left(-\frac{1}{2}(\mathbf{x}_a - \mu_{a|b}(\mathbf{x}_b))^\top \Sigma_{a|b}^{-1}(\mathbf{x}_a - \mu_{a|b}(\mathbf{x}_b))\right) \cdot$$

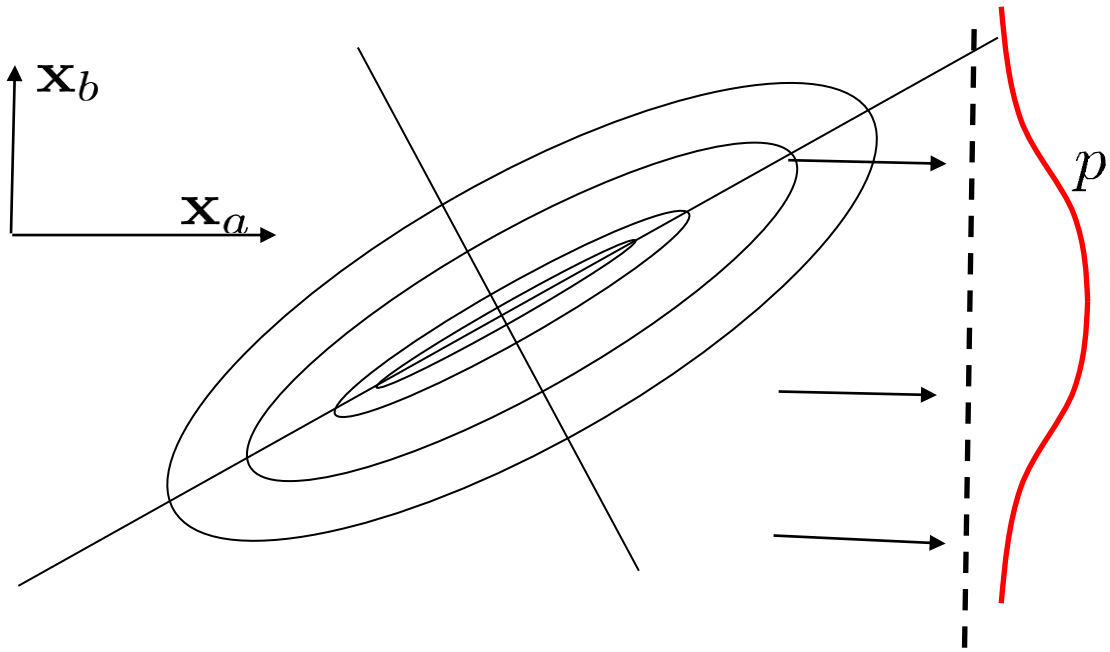
$$C_2 \exp\left(-\frac{1}{2}(\mathbf{x}_b - \mu_b)^\top \Sigma_b^{-1}(\mathbf{x}_b - \mu_b)\right)$$

$$p(\mathbf{x}) = p(\mathbf{x}_a, \mathbf{x}_b) = p(\mathbf{x}_a|\mathbf{x}_b)p(\mathbf{x}_b)$$

Gaussian Random Variable – Marginal

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \begin{matrix} \mathbf{x}_a \in \mathbb{R}^{D_a} \\ \mathbf{x}_b \in \mathbb{R}^{D_b} \end{matrix} \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix}$$

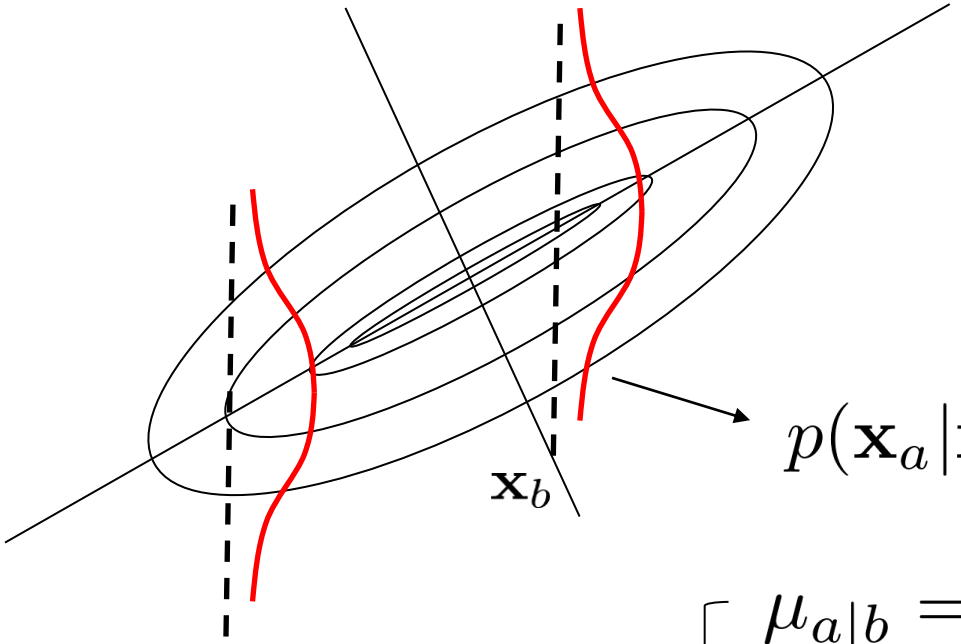
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^{D} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^{\top} \Sigma^{-1} (\mathbf{x} - \mu) \right)$$



$$\begin{aligned} p(\mathbf{x}_b) &= \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_a \\ &= \int p(\mathbf{x}_a | \mathbf{x}_b) p(\mathbf{x}_b) d\mathbf{x}_a \\ &= \mathcal{N}(\mu_b, \Sigma_b) \end{aligned}$$

Gaussian Random Variable – Conditional

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^{D}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)\right)$$



$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \begin{matrix} \mathbf{x}_a \in \mathbb{R}^{D_a} \\ \mathbf{x}_b \in \mathbb{R}^{D_b} \end{matrix}$$

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mu_{a|b}, \Sigma_{a|b})$$

$$\begin{cases} \mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_b^{-1}(\mathbf{x}_b - \mu_b) \\ \Sigma_{a|b} = \Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ba} \end{cases}$$

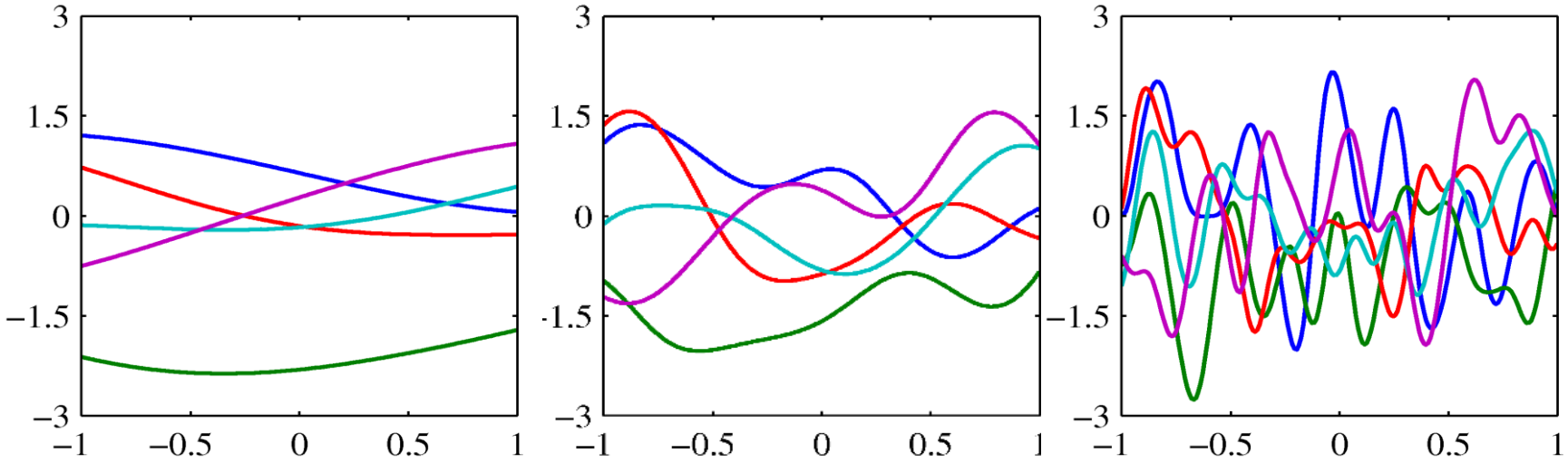
Gaussian Processes – Function Space View

$$y(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$$

$$m(\mathbf{x}) = \mathbb{E}[y(\mathbf{x})] = 0$$

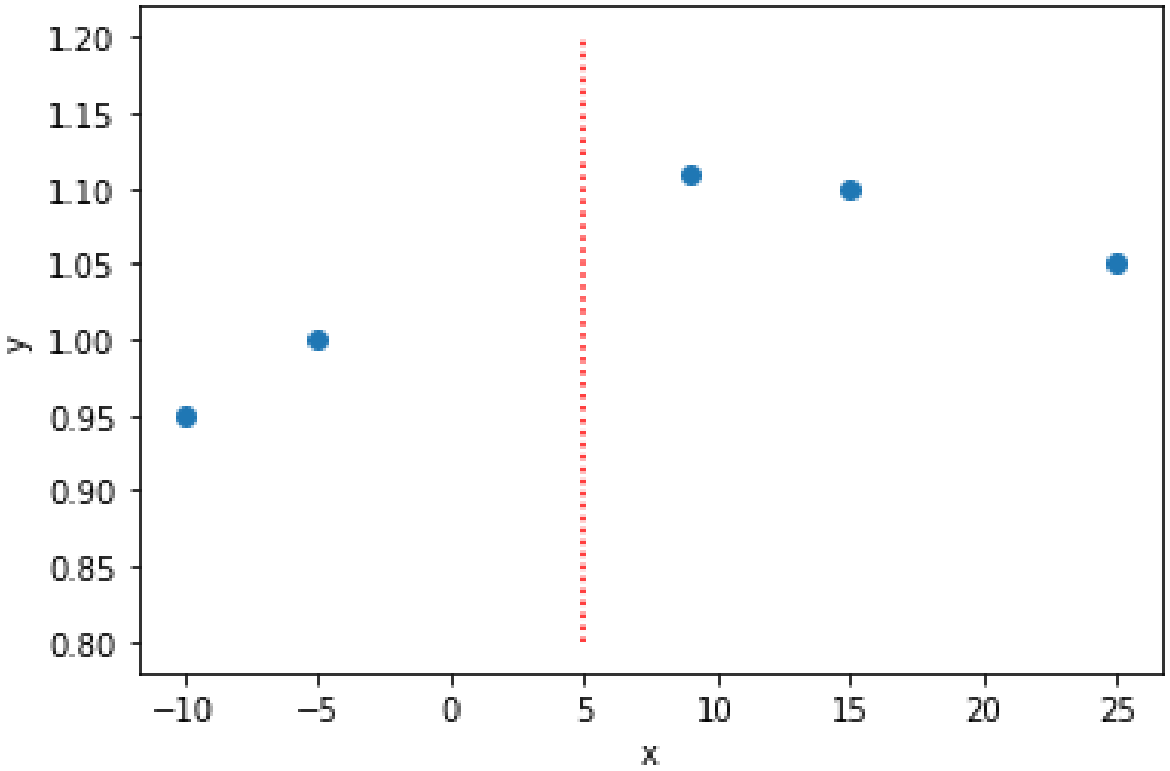
$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(y(\mathbf{x}) - m(\mathbf{x}))(y(\mathbf{x}') - m(\mathbf{x}'))]$$

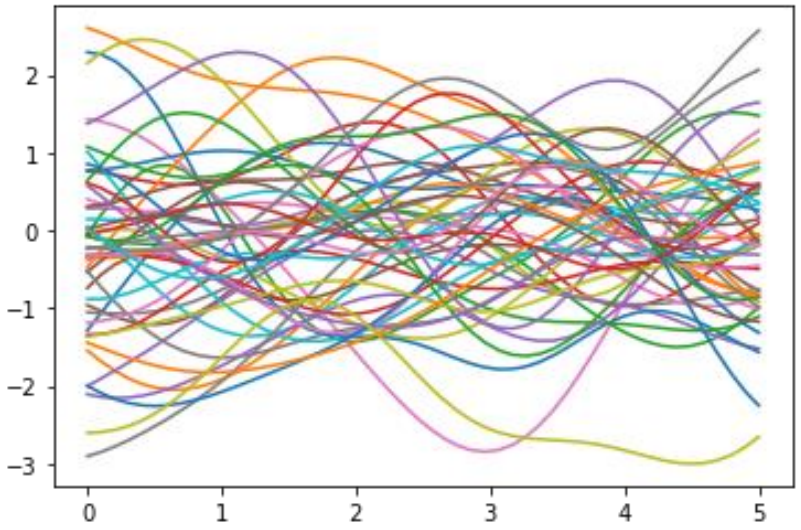
$$= \theta_1 \exp\left\{-\frac{\theta_2}{2}||\mathbf{x} - \mathbf{x}'||^2\right\} + \theta_3 + \theta_4 \mathbf{x}^\top \mathbf{x}'$$

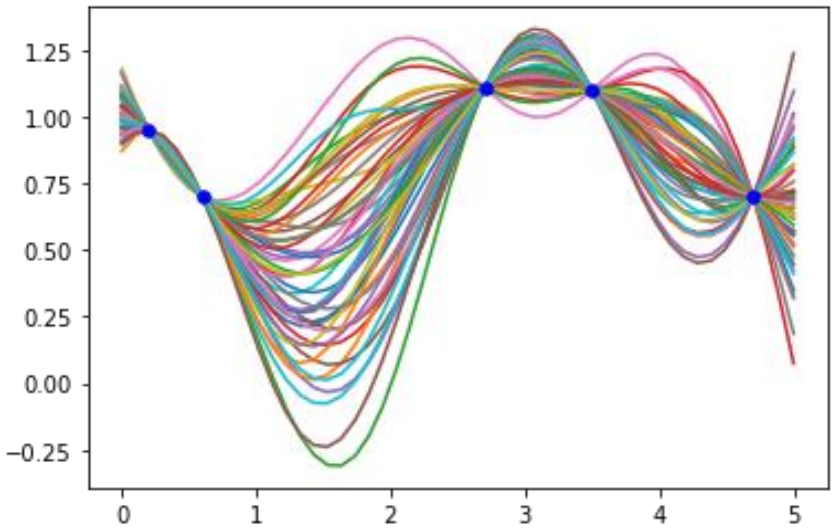


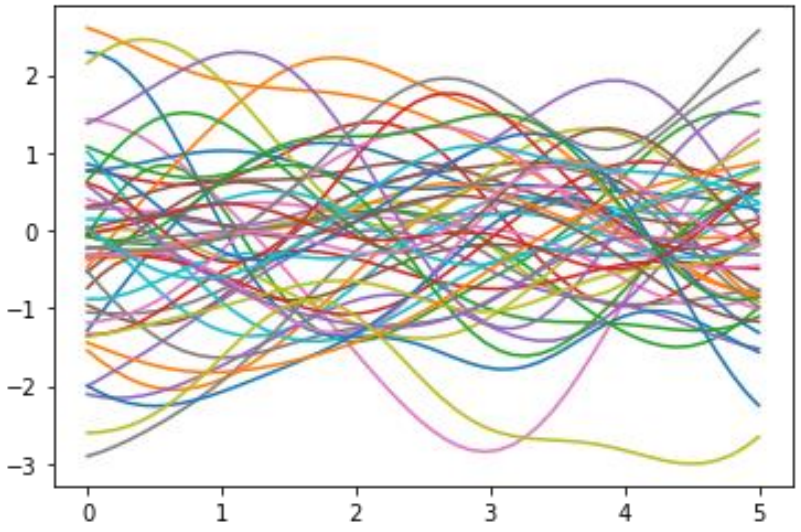
C. Bishop (2007) *Pattern Recognition and Machine Learning*, Springer

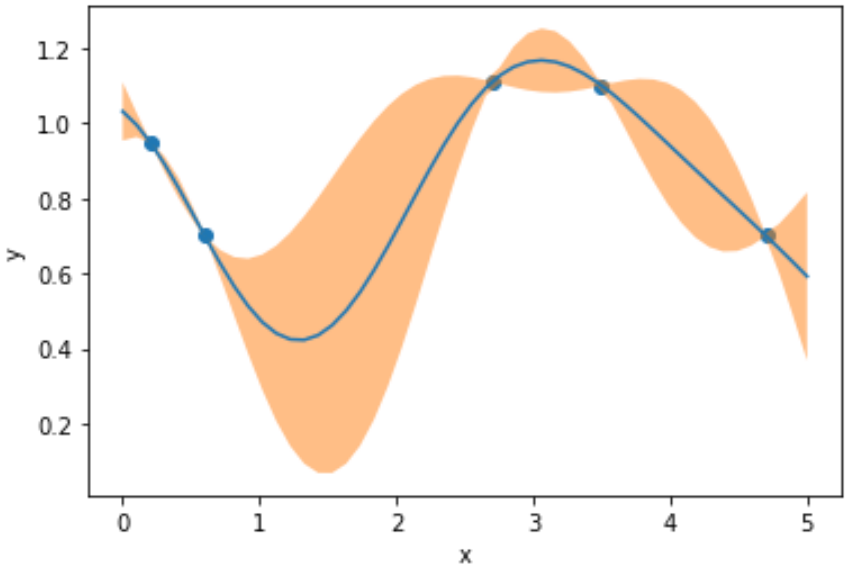
Regression




2
1
0
−1
−2
−3
0
1
2
3
4
5


1.25
1.00
0.75
0.50
0.25
0.00
−0.25
0
1
2
3
4
5

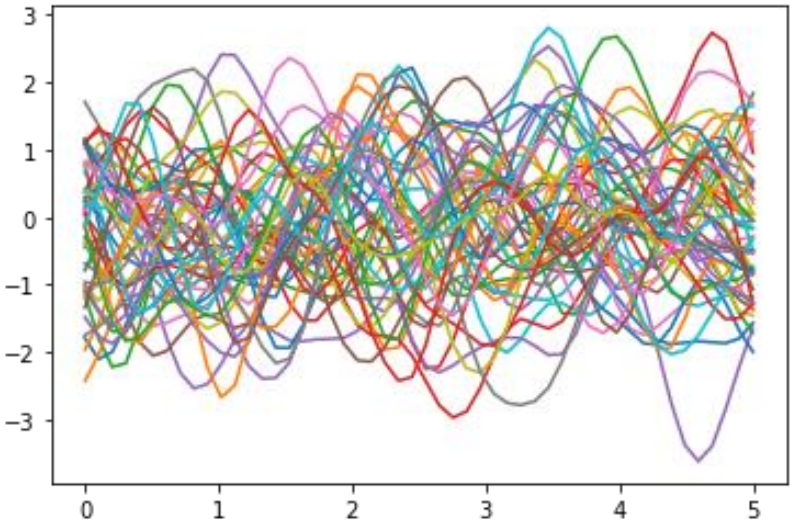


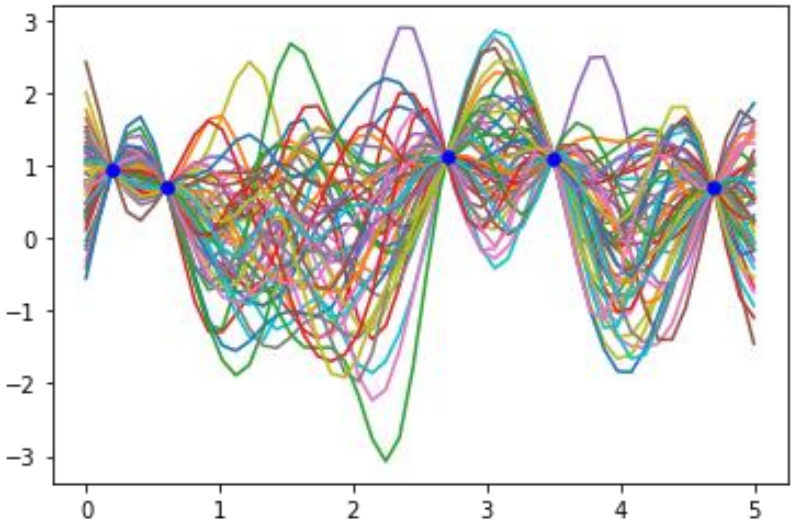


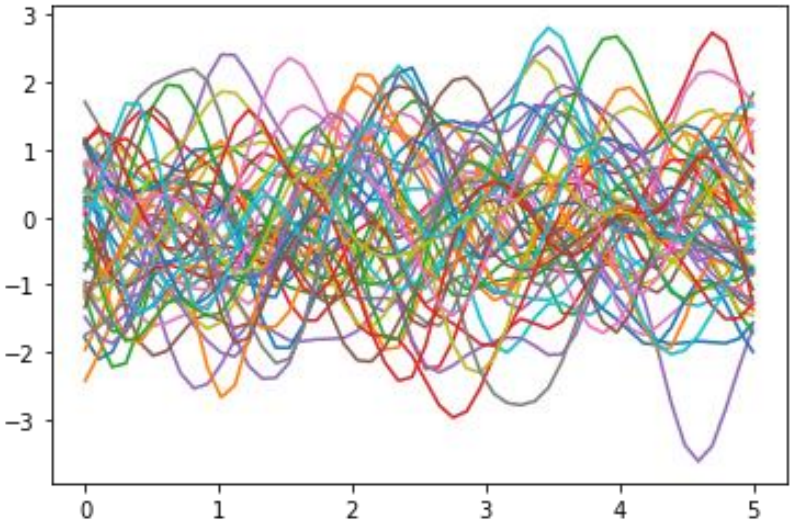
$$m(x) = \mathbf{k}^\top K^{-1}\mathbf{y}$$

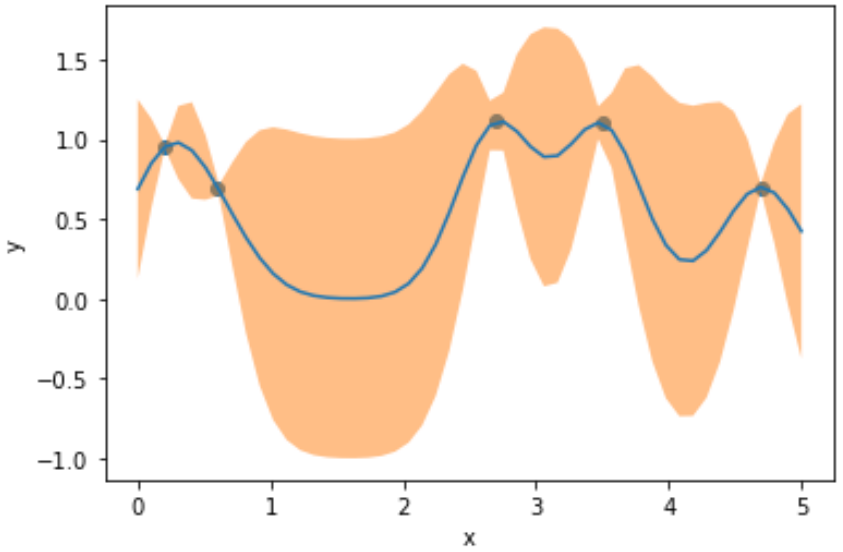
$$\sigma^2(\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - \mathbf{k}^\top K^{-1}\mathbf{k}$$

$$[\mathbf{k}]_i = k(\mathbf{x},\mathbf{x}_i)$$
$$[\mathbf{K}]_{i,j} = k(\mathbf{x}_i,\mathbf{x}_j)$$
$$[\mathbf{y}]_i = y(\mathbf{x}_i)$$


3
2
1
0
−1
−2
−3
0
1
2
3
4
5


3
2
1
0
−1
−2
−3
0
1
2
3
4
5





$$m(x) = \mathbf{k}^\top K^{-1}\mathbf{y}$$

$$\sigma^2(\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - \mathbf{k}^\top K^{-1}\mathbf{k}$$

$$[\mathbf{k}]_i = k(\mathbf{x},\mathbf{x}_i)$$
$$[\mathbf{K}]_{i,j} = k(\mathbf{x}_i,\mathbf{x}_j)$$
$$[\mathbf{y}]_i = y(\mathbf{x}_i)$$

Bayesian linear regression

Review – Two Paradigms

Data:

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N \qquad \mathbf{x} \in \mathbb{R}^D, y_i \in \{1, \ldots, C\} \text{ or } y_i \in \mathbb{R}$$

Model:

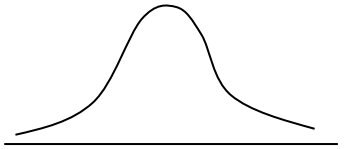
$$f(\mathbf{x}; \theta) \in \mathcal{H} \quad \text{or} \quad p(\mathbf{x}|\theta) \in \mathcal{H}$$

- 1) Choose the best fit to the data in terms of $L(y, f)$

$$\theta^* = \arg\min_\theta \sum_{i=1}^N L(y_i, f(\mathbf{x}_i))$$

Prediction: $y = f(\mathbf{x}; \theta^*)$

- 2) Choose the best guess with likelihood



Likelihood: $p(\mathcal{D}|\theta)$ Prior: $p(\theta)$

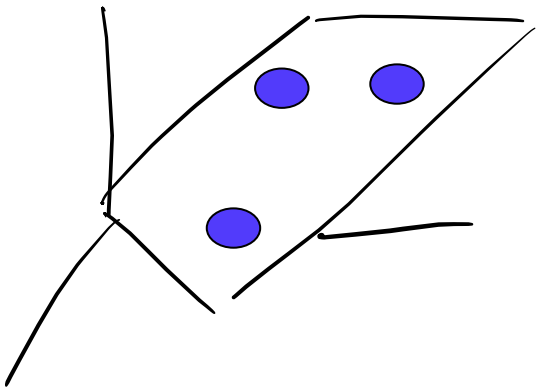
Posterior: $p(\theta|\mathcal{D}) = \dfrac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$

Prediction: $p(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x}; \theta)p(\theta|\mathcal{D})d\theta$

Linear Regression

$$\widehat{L}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N} ||f(\mathbf{x}_i;\mathbf{w}) - y_i||^2$$

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^\top \mathbf{x}$$



$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \widehat{L}(\mathbf{w})$$

$$\frac{\partial \widehat{L}}{\partial \mathbf{w}} = \frac{1}{2}\frac{\partial}{\partial \mathbf{w}}\sum_{i=1}^{N} ||f(\mathbf{x}_i;\mathbf{w}) - y_i||^2 = 0$$

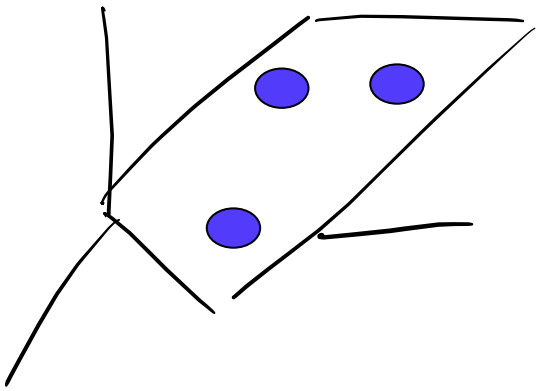
Linear Regression

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^\top \mathbf{x}$$

$$\frac{\partial \widehat{L}}{\partial \mathbf{w}} = \frac{1}{2}\frac{\partial}{\partial \mathbf{w}} \sum_{i=1}^{N} ||f(\mathbf{x}_i;\mathbf{w}) - y_i||^2$$

$$= \sum_{i=1}^{N} (f(\mathbf{x}_i;\mathbf{w}) - y_i) \frac{\partial f(\mathbf{x}_i;\mathbf{w})}{\partial \mathbf{w}}$$

$$= \sum_{i=1}^{N} (\mathbf{w}^\top \mathbf{x}_i - y_i)\mathbf{x}_i = 0$$



$$X = \begin{pmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_N \\ | & & | \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

$$XX^\top \mathbf{w} - X\mathbf{y} = 0 \quad \longrightarrow \quad \mathbf{w}^* = (XX^\top)^{-1} Xy$$

Bayesian Linear Regression

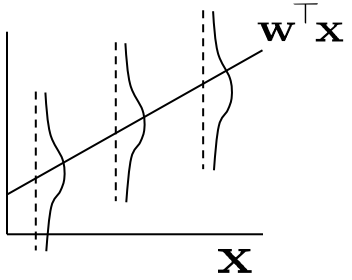
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; 0, \sigma_0^2 I) = \frac{1}{\sqrt{2\pi\sigma_0^2}^D} \exp\left(-\frac{||\mathbf{w}||^2}{2\sigma_0^2}\right)$$

Model:

$$y = \mathbf{w}^\top \mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\Rightarrow \quad p(y|\mathbf{x}, \mathbf{w}) = \mathcal{N}(y;\ \mathbf{w}^\top \mathbf{x},\ \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \mathbf{w}^\top \mathbf{x})^2}{2\sigma^2}\right)$$



Posterior:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{Y}|\mathcal{X}, \mathbf{w})p(\mathbf{w})}{p(\mathcal{Y})}$$

$$\mathcal{X} = \{\mathbf{x}_i, \ldots, \mathbf{x}_N\}$$

$$\mathcal{Y} = \{y_i, \ldots, y_N\}$$

$$\propto \left(\prod_{i=1}^{N} p(y_i|\mathbf{x}_i, \mathbf{w})\right) \cdot p(\mathbf{w})$$

Posterior:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{Y}|\mathcal{X}, \mathbf{w})p(\mathbf{w})}{p(\mathcal{Y})}$$

$$\propto \frac{1}{\sqrt{2\pi\sigma^2}^N} \prod_{i=1}^{N} \exp\left(-\frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2}\right) \cdot$$

$$\frac{1}{\sqrt{2\pi\sigma_0^2}^D} \exp\left(-\frac{||\mathbf{w}||^2}{2\sigma_0^2}\right)$$

Quadratic w.r.t. $\mathbf{w}$

The posterior should be in a form of

$$p(\mathbf{w}|\mathcal{D}) = C \cdot \exp\left(-\frac{1}{2}(\mathbf{w} - \mu_\mathbf{w})^\top \Sigma_\mathbf{w}^{-1}(\mathbf{w} - \mu_\mathbf{w})\right)$$

$$p(\mathbf{w}|\mathcal{D}) = C \cdot \exp\left(-\frac{1}{2}(\mathbf{w}-\mu_{\mathbf{w}})^\top \Sigma_{\mathbf{w}}^{-1}(\mathbf{w}-\mu_{\mathbf{w}})\right)$$

$$\begin{cases} \mu_{\mathbf{w}} = \left(XX^\top + \dfrac{\sigma^2}{\sigma_0^2} I\right)^{-1} X\mathbf{y} \\ \Sigma_{\mathbf{w}} = \dfrac{\sigma^2}{N}\left(\dfrac{1}{N}XX^\top + \dfrac{\sigma^2}{N\sigma_0^2} I\right)^{-1} \end{cases}$$

$$X = \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \dots & \mathbf{x}_N \\ | & & | \end{bmatrix} \qquad \mathbf{y} = [y_i, \dots, y_N]^\top$$

Target noise (Aleatoric uncertainty)

$$\frac{\sigma^2}{N}\left(\frac{1}{N}XX^\top + \frac{\sigma^2}{N\sigma_0^2} I\right)^{-1}$$

Number of data

Covariance of x from *0*

Regularization becomes small with large *N* or small target noise.

Predictive Density

$$p(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathcal{D}) d\mathbf{w}$$

Likelihood:

$$p(y|\mathbf{x}, \mathbf{w}) = \mathcal{N}(y;\ \mathbf{w}^\top \mathbf{x},\ \sigma^2)$$

Posterior:

$$p(\mathbf{w}|\mathcal{D}) = \mathcal{N}(\mathbf{w}; \mu_\mathbf{w}, \Sigma_\mathbf{w})$$

$$\mu_\mathbf{w} = \left(XX^\top + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} X\mathbf{y}, \quad \Sigma_\mathbf{w} = \frac{\sigma^2}{N} \left(\frac{1}{N} XX^\top + \frac{\sigma^2}{N\sigma_0^2} I \right)^{-1}$$

Predictive Density

$$p(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathcal{D}) d\mathbf{w}$$

Predictive density has quadratic polynomial w.r.t. y inside exponential
→ Gaussian predictive density

$$\mathbb{E}[y|\mathbf{x}, \mathcal{D}] = \mathbf{x}^\top X \left(X^\top X + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} \mathbf{y}$$

$$Var[y|\mathbf{x}, \mathcal{D}] = \sigma^2 + \sigma_0^2 \mathbf{x}^\top \mathbf{x} - \sigma_0^2 \, \mathbf{x}^\top X \left(X^\top X + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} X^\top \mathbf{x}$$

Kernelization

Make a matrix $K \in \mathbb{R}^{N \times N}$ s. t. $[K]_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$,

& a vector $\mathbf{k} \in \mathbb{R}^N$ s. t. $[\mathbf{k}]_i = \mathbf{x}_i^\top \mathbf{x}$

$$i, j \in \{1, \ldots, N\}$$

$$\mathbb{E}[y|\mathbf{x}, \mathcal{D}] = \mathbf{x}^\top X \left(X^\top X + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} \mathbf{y}$$

$$= \mathbf{k}^\top \left(K + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} \mathbf{y}$$

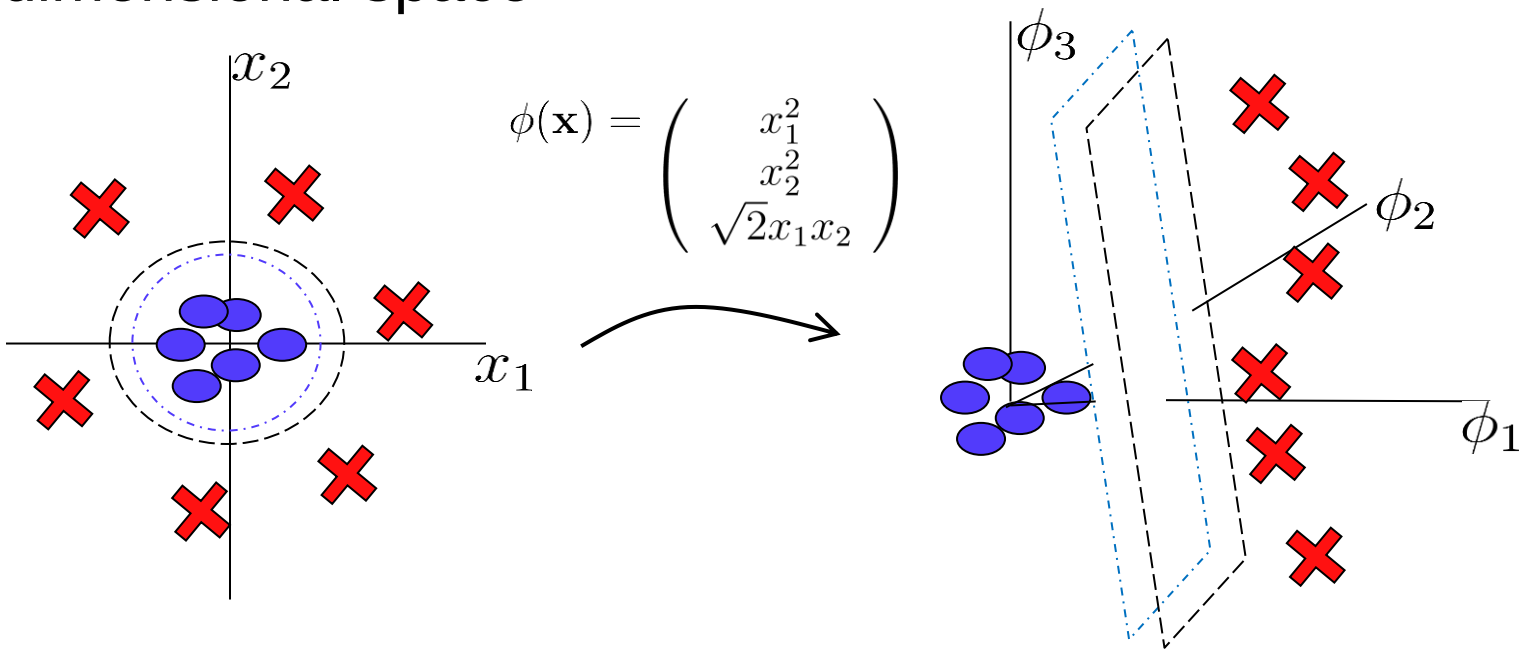
$$Var[y|\mathbf{x}, \mathcal{D}] = \sigma^2 + \sigma_0^2 \mathbf{x}^\top \mathbf{x} - \sigma_0^2 \, \mathbf{x}^\top X \left(X^\top X + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} X^\top \mathbf{x}$$

$$= \sigma^2 + \sigma_0^2 k - \sigma_0^2 \mathbf{k}^\top \left(K + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} \mathbf{k}$$

$k = \mathbf{x}^\top \mathbf{x}$

Mapping Data to a High Dimensional Space

- Consider a nonlinear mapping to a higher-dimensional space



- And consider a function

$$k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z})^2 = (x_1z_1 + x_2z_2)^2 = x_1^2z_1^2 + 2x_1z_1x_2z_2 + x_2^2z_2^2$$

$$= \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix}^\top \begin{pmatrix} z_1^2 \\ z_2^2 \\ \sqrt{2}z_1z_2 \end{pmatrix} = \phi(\mathbf{x})^\top \phi(\mathbf{z})$$

Mercer's Theorem

A symmetric function $K(x, y)$ can be expressed as an inner product

$$K(x, y) = \langle \phi(x), \phi(y) \rangle$$

for some ϕ if and only if $K(x, y)$ is positive semidefinite, i.e.

$$\int K(x, y)g(x)g(y)dxdy \geq 0 \qquad \forall g$$

or, equivalently:

$$\begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots \\ K(x_2, x_1) & \ddots & \\ \vdots & & \end{bmatrix}$$ is psd for any collection $\{x_1 \ldots x_n\}$

Therefore you can either explicitly map the data with a ϕ and take the dot product, or you can take any kernel and use it right away, without knowing nor caring what ϕ looks like. For example:

- Gaussian Kernel: $K(x, y) = e^{\frac{1}{2}||x-y||^2}$
- Spectrum Kernel: count the number of substrings in common. It is a kernel since it is a dot product between vectors of indicators of all the substrings.

https://people.eecs.berkeley.edu/~jordan/courses/281B-spring04/lectures/lec3.pdf

Mapping Data to High Dimensional Space

- Consider functions *k(.,.)* to which the corresponding mapping *Φ(.)* exists:
- Condition: if a function *k(.,.)* is **positive definite (P.D.)**, there exists a mapping function *Φ(.)* satisfying

$$k(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^\top \phi(\mathbf{z}) \qquad \text{(Mercer theorem)}$$

- Such functions include

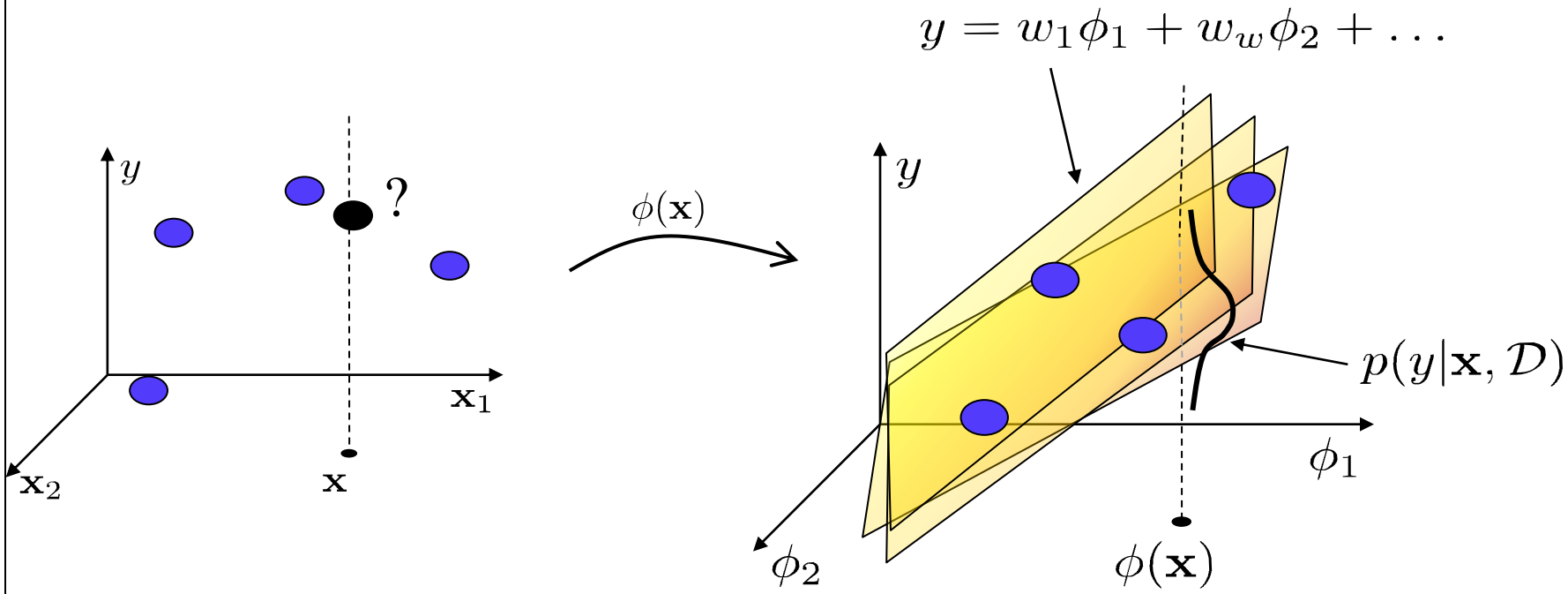
$$k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z} + 1)^d \qquad \text{(Polynomial kernel)}$$

$$k(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{(\mathbf{x} - \mathbf{z})^2}{\gamma}\right) \qquad \text{(Gaussian kernel)}$$

$$\vdots$$

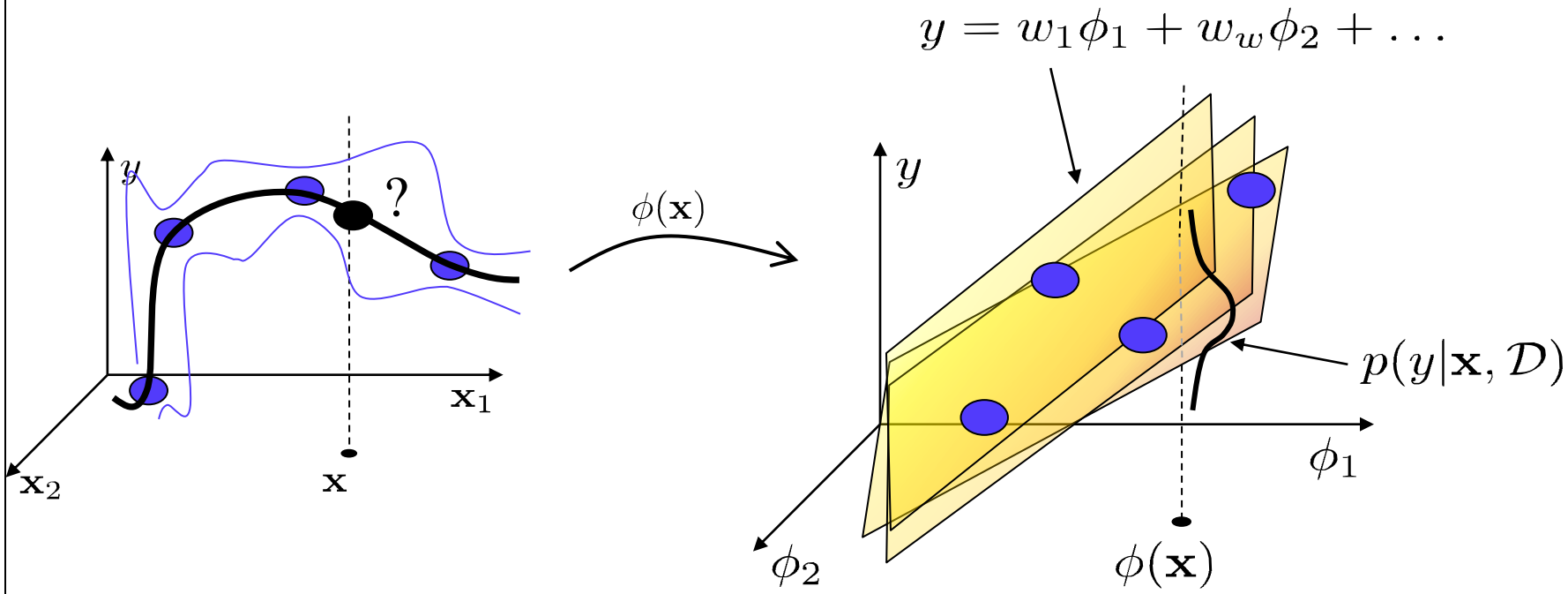
Parameter Space View

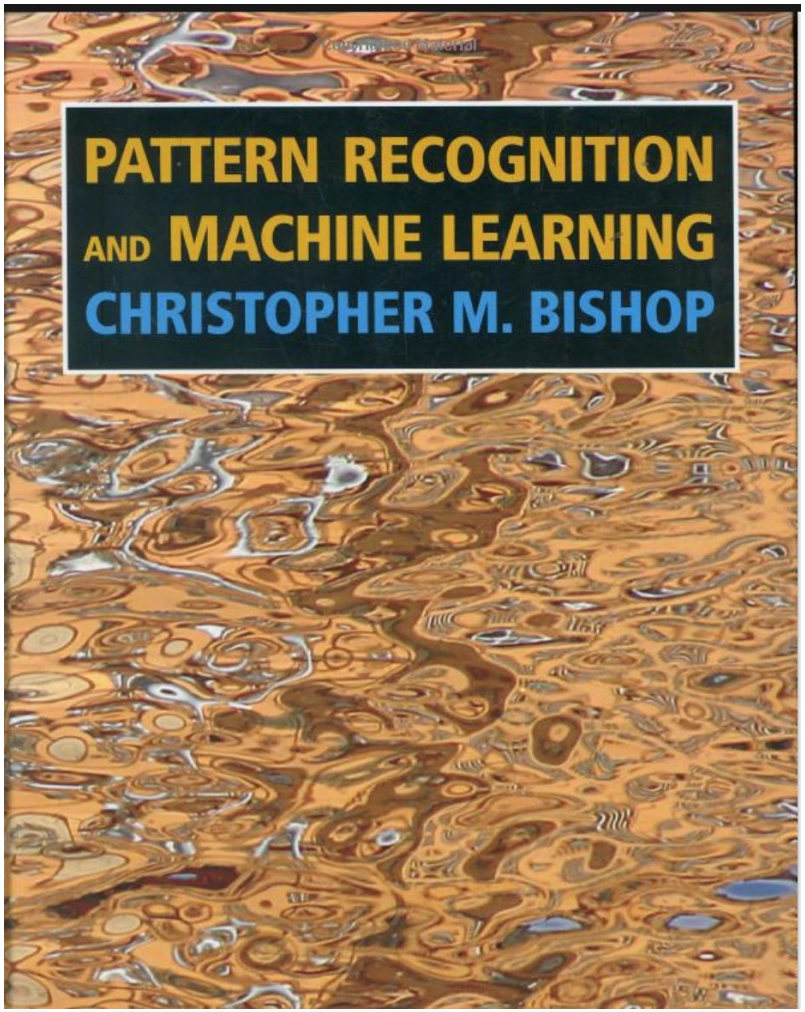
$$k_{GP}(\mathbf{x}, \mathbf{z}) = \sigma_0^2 k_{BL}(\mathbf{x}, \mathbf{z}) + \sigma^2 \delta_{\mathbf{x},\mathbf{z}}$$
$$k_{BL}(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^\top \phi(\mathbf{z})$$

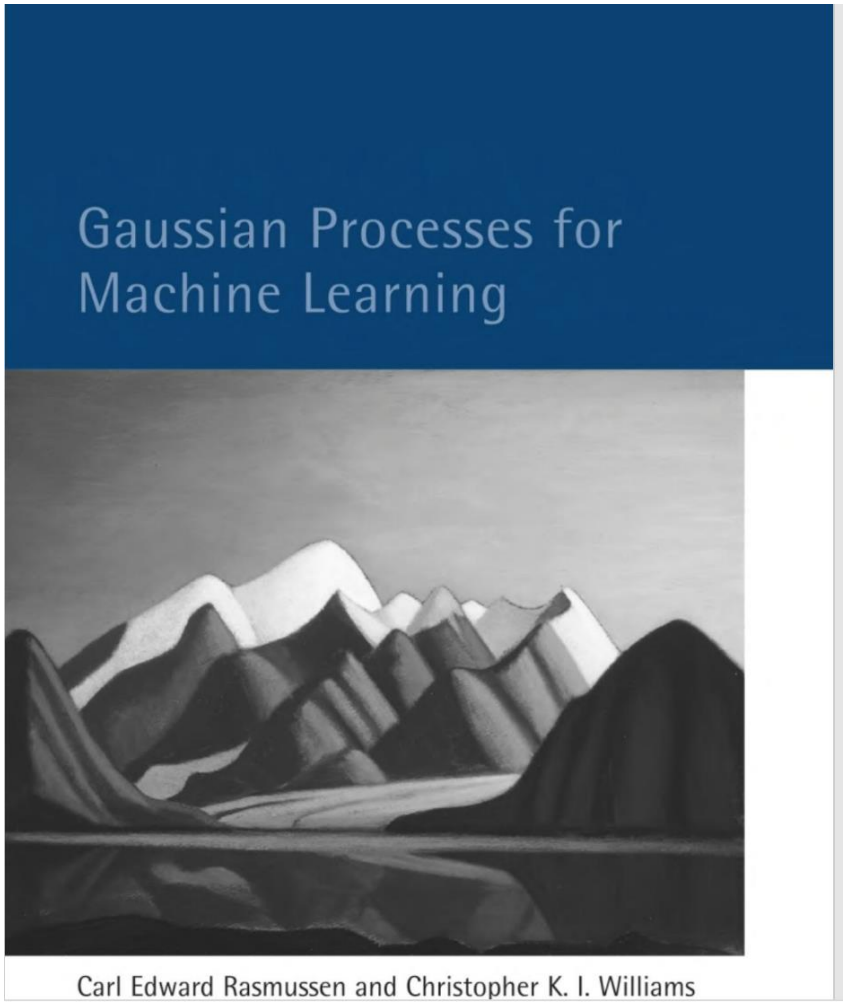


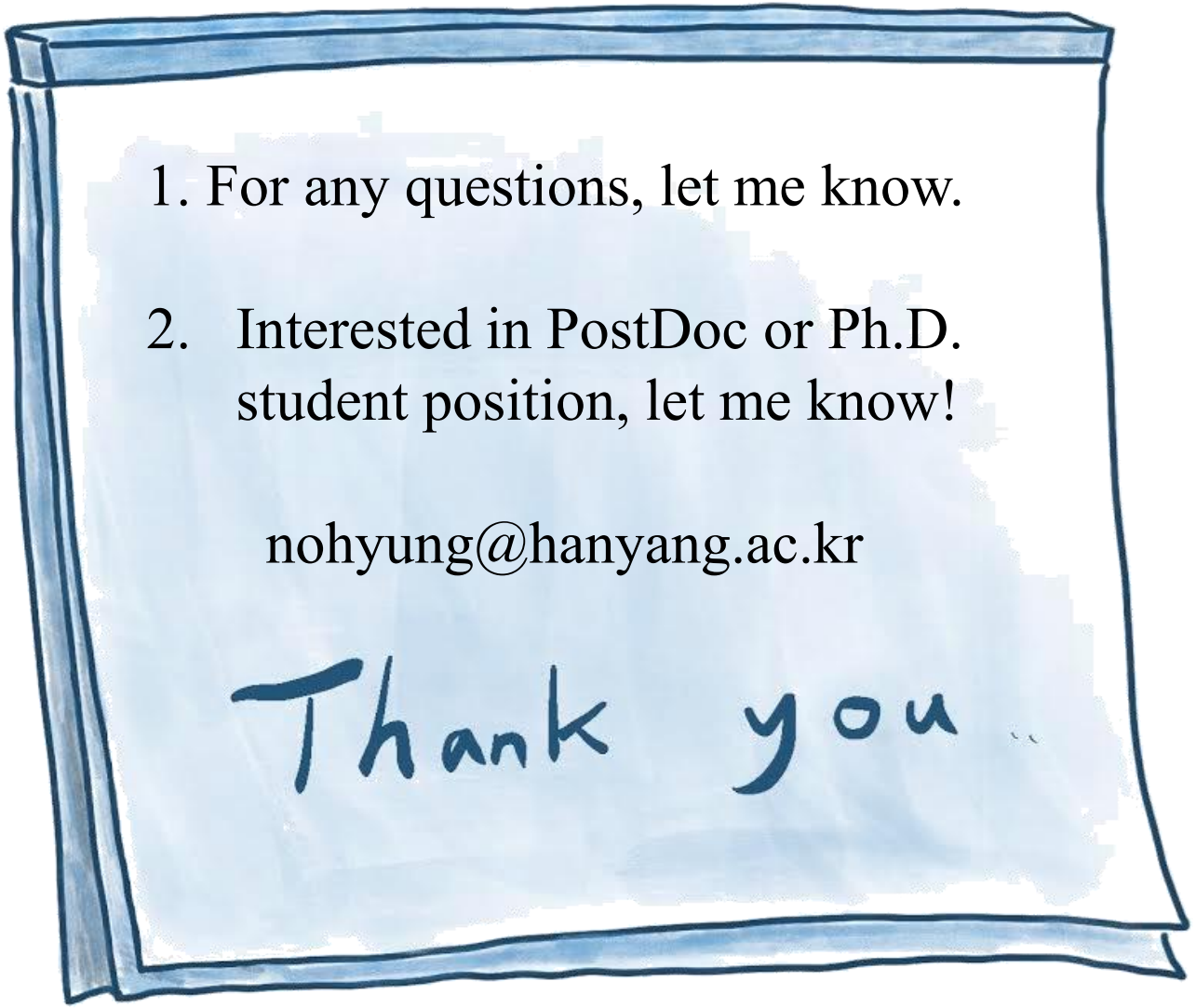
Parameter Space View

$$k_{GP}(\mathbf{x}, \mathbf{z}) = \sigma_0^2 k_{BL}(\mathbf{x}, \mathbf{z}) + \sigma^2 \delta_{\mathbf{x},\mathbf{z}}$$
$$k_{BL}(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^\top \phi(\mathbf{z})$$




PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP


Gaussian Processes for
Machine Learning
Carl Edward Rasmussen and Christopher K. I. Williams

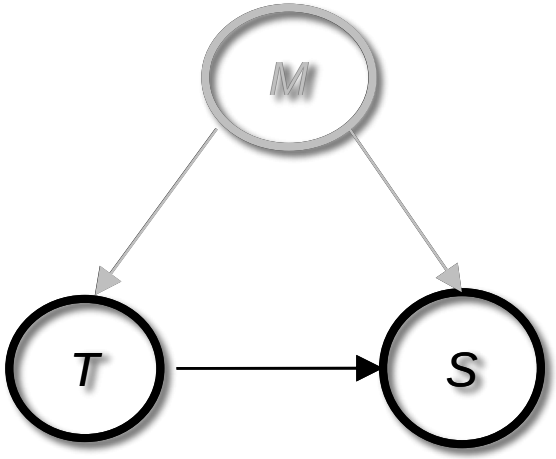

1. For any questions, let me know.
2. Interested in PostDoc or Ph.D. student position, let me know!
nohyung@hanyang.ac.kr
Thank you

Our World is Not Easy.

- Confounding Effect and Data

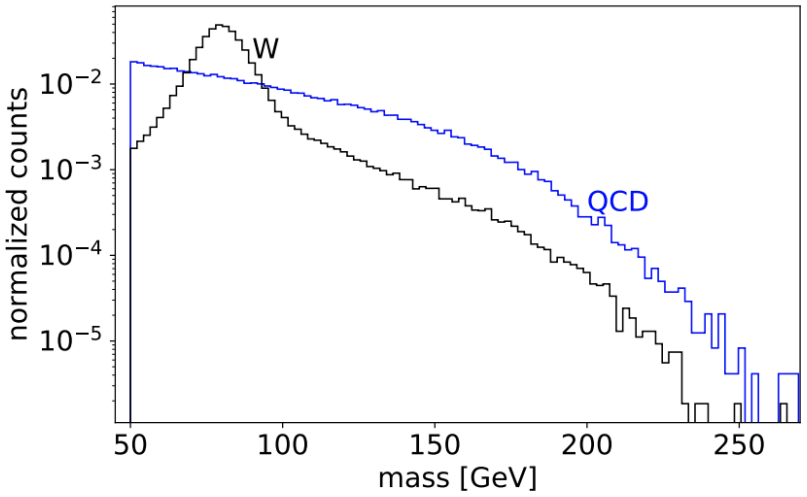
	Full Population, N = 52			Men (M), N = 20			Women (¬M), N = 32		
	Success (S)	Failure (¬S)	Success Rate	Success	Failure	Success Rate	Success	Failure	Success Rate
Treatment (T)	20	20	50%	8	5	≈ 61%	12	15	≈ 44%
Control (¬T)	6	6	50%	4	3	≈ 57%	2	3	≈ 40%

TABLE 1: Simpson's Paradox: the type of association at the population level (positive, negative, independent) changes at the level of subpopulations. Numbers taken from Simpson's original example (1951).



(Simpson's Paradox)

https://plato.stanford.edu/entries/paradox-simpson/



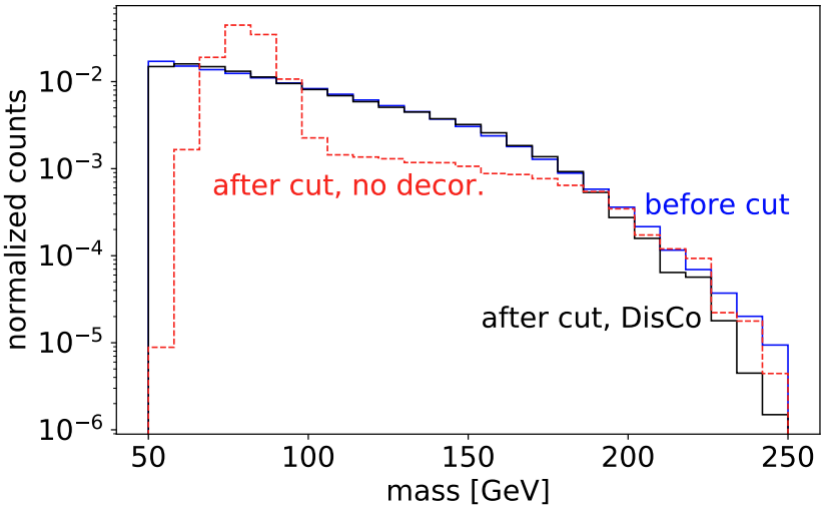
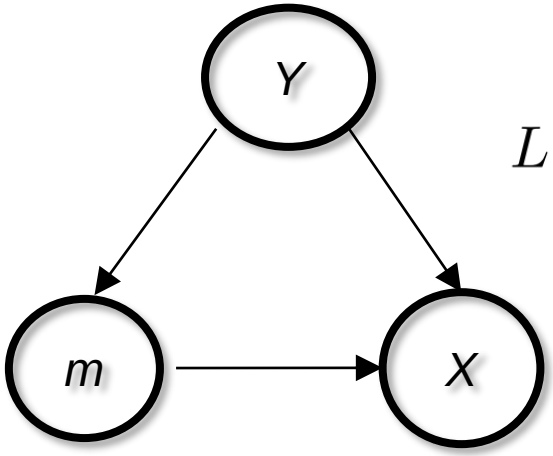


FIG. 1: Invariant mass distribution for the inclusive W and QCD samples.



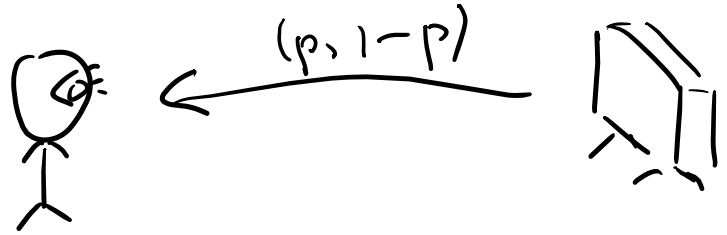
$$L = L_{classifier}(\vec{y}, \vec{y}_{true}) + \lambda \, \mathrm{dCorr}^2_{y_{true}=0}(\vec{m}, \vec{y})$$

DisCo Fever: Robust Networks Through Distance Correlation **arXiv:2001.05310** [hep-ph]

Information Theory for Machine Learning

- Information-theoretic approach (method, …)
  - A machine learning algorithm that uses the formulation from information theory
- Information theory
  - A theory about the amount of information accumulated when we observe a sample
  - In machine learning, we are interested in constructing a "system" that an observation of a test sample has the information about the prediction target as much as possible.
- Applications
  - Loss functions (cross entropy loss)
  - Criteria for finding representations (dimensionality reduction, information bottleneck, decorrelation…)

Probability/Observation/Information



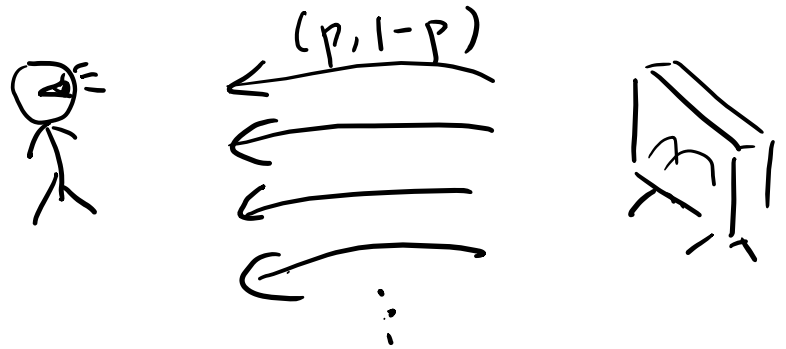
뉴스
입시 발표
성적
Love letter에 대한 답
지이잇는 소리
⋮

Information at every observation

$$I(X) = -\log P(X)$$

Question: Why $\log \frac{1}{p}$?

Entropy and Observation

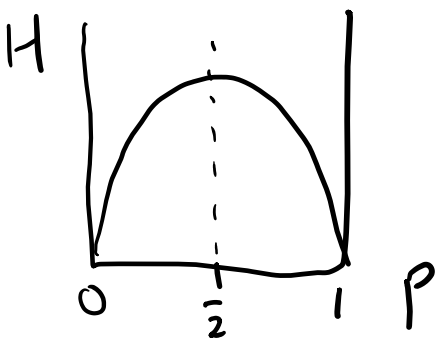


있서 발견
성적
Love letter의 대한 답

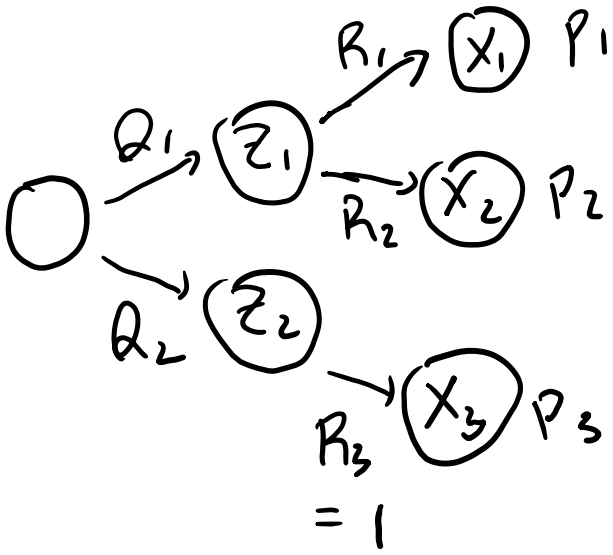
Entropy: Expectation of the information
(The information we get on average.)

$$H(X) = E_{p(x)}[-\log p(x)]$$

$$= -\sum_{x} p(x) \log p(x)$$



Partial Observation



$P_1 = Q_1 R_1$

$P_2 = Q_1 R_2$

$P_3 = Q_2 R_3$

When we observe Z, we partially observe X.

Def: $H(X|Z) = -\sum_x P(x|z) \log P(x|z)$

← Caution: Different from conventional def. in information theory community

$H(X|Z) \equiv \sum_{x,z} -P(x,z) \log P(x|z)$

$$H(X) = H(P_1, P_2, P_3)$$
$$= -P_1 \log P_1 - P_2 \log P_2 - P_3 \log P_3$$
$$= -Q_1 R_1 \log Q_1 R_1 - Q_1 R_2 \log Q_1 R_2 - Q_2 R_3 \log Q_2 R_3$$
$$= H(Z) + Q_1 H(X|Z=z_1) + Q_2 H(X|Z=z_2)$$

$$H(X) = H(P_1, P_2, P_3)$$

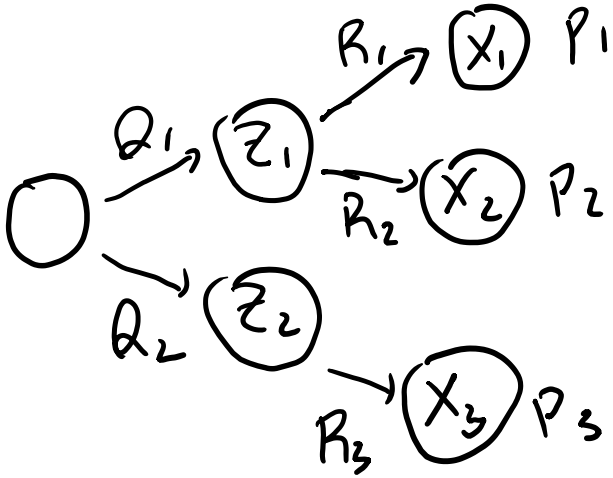
$$= -P_1 \log P_1 - P_2 \log P_2 - P_3 \log P_3$$

$$= -Q_1 R_1 \log Q_1 R_1 - Q_1 R_2 \log Q_1 R_2 - Q_2 R_3 \log Q_2 R_3$$

$$= -Q_1 (R_1 + R_2) \log Q_1 + Q_1 (-R_1 \log R_1 - R_2 \log R_2) - Q_2 \log Q_2$$

$$= -Q_1 \log Q_1 - Q_2 \log Q_2 + Q_1 (-R_1 \log R_1 - R_2 \log R_2)$$

$$= H(Z) + Q_1 H(X|Z=z_1) + Q_2 H(X|Z=z_2)$$



$$H(X) = H(P_1, P_2, P_3)$$

$$= -P_1 \log P_1 - P_2 \log P_2 - P_3 \log P_3$$

$$= -Q_1 R_1 \log Q_1 R_1 - Q_1 R_2 \log Q_1 R_2 - Q_2 R_3 \log Q_2 R_3$$

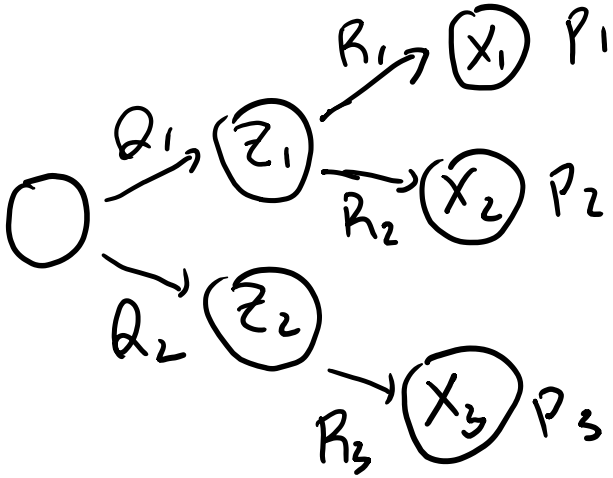
$$= -Q_1 (R_1 + R_2) \log Q_1 + Q_1 (-R_1 \log R_1 - R_2 \log R_2) - Q_2 \log Q_2$$

$$= -Q_1 \log Q_1 - Q_2 \log Q_2 + Q_1 (-R_1 \log R_1 - R_2 \log R_2)$$

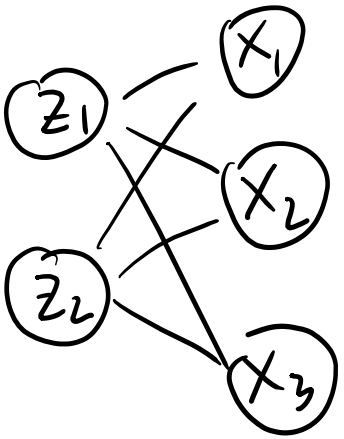
$$= H(Z) + Q_1 H(X|Z = z_1)$$

(Nested) Partial Observation

$$H(X) = H(Z) + E_Z[H(X|Z)]$$



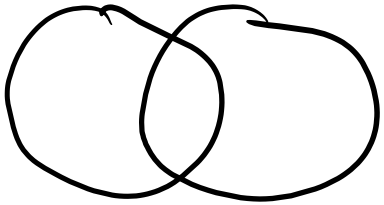
Mutual Information

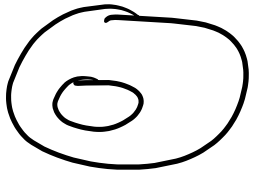


Not a nested partial observation

Observation of z changes distr. of X,

$p(x) \rightarrow p(x|z)$



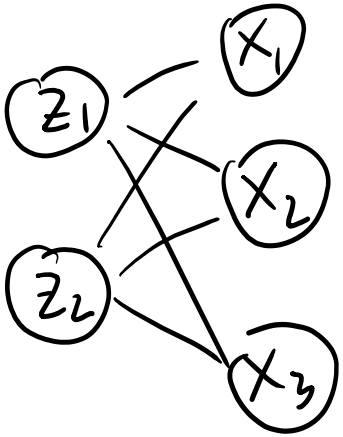


$$H(X) = \underbrace{I(X;Z)}_{\text{Reduced entropy by the observaton of } z} + \underbrace{E_z[H(X|z)]}_{\text{Expectatim of the entropy after observation}}$$

$$H(x) = -\sum_{i=1}^{3} p(x_i) \log p(x_i)$$

$$= -\sum_{j=1}^{2} \sum_{i=1}^{3} p(x_i, z_j) \log p(x_i)$$

$$= I(x; z) + \sum_{i,j} p(x_i, z_j) \log p(x_i | z_j)$$

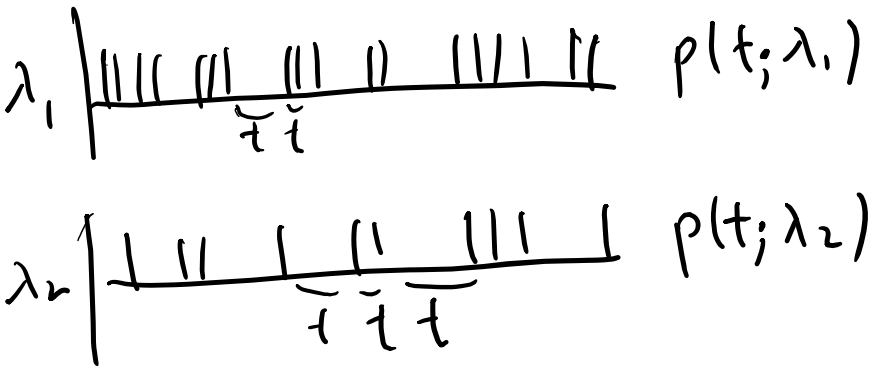


$$I(x; z) = -\sum_{i,j} p(x_i, z_j) \left\{ \log p(x_i) - \log p(x_i | z_j) \right\}$$

$$= -\sum_{i,j} p(x_i, z_j) \log \frac{p(x_i) p(z_j)}{p(x_i, z_j)}$$

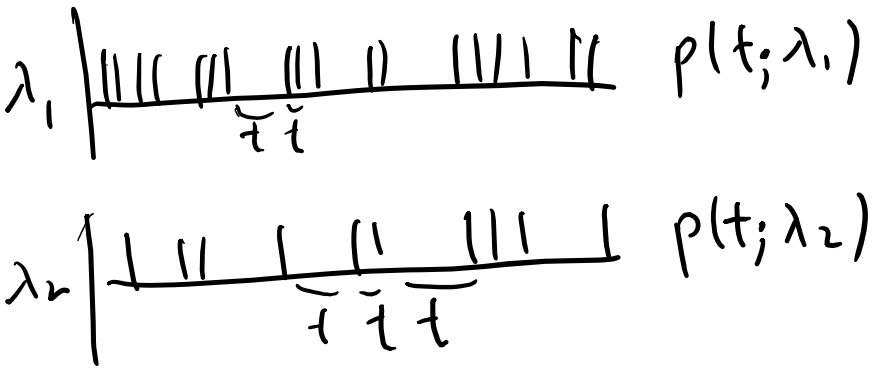
Kullback-Leibler Divergence

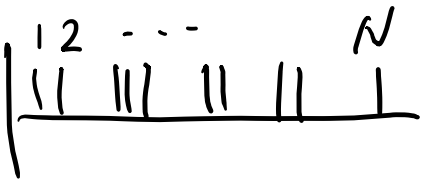
Expectation of log-likelihood ratio. (= Relative Entropy in I.T. community.)



Kullback-Leibler Divergence

Expectation of log-likelihood ratio. (= Relative Entropy in I.T. community.)





Accumulation of likelihood ratio information $\sum_{i=1}^{N} \log \frac{p(t_i;\lambda_1)}{p(t_i;\lambda_2)}$

Accumulation speed of the information $E\left[\log \frac{p(t;\lambda_1)}{p(t;\lambda_2)}\right]$

Kullback-Leibler Divergence

$$E\left[\log \frac{p(t;\lambda_1)}{p(t;\lambda_2)}\right] = -\int p(t|\lambda_1) \log \frac{p(t|\lambda_2)}{p(t|\lambda_1)}\, dt$$

KL-divergence

$$D_{KL}(p_1 \| p_2) = -\int p_1(x) \log \frac{p_2(x)}{p_1(x)}\, dx$$

⇐ How much $p_2(x)$ is different from $p_1(x)$

Note: $$I(x;z) = -\int p(x,z) \log \frac{p(x)p(z)}{p(x,z)}\, dx\, dz$$

$$= D_{KL}(p(x,z) \| p(x)p(z))$$

f-Divergences

$$D_f(p_1, p_2) = E_{p_1}\left[f\left(\frac{p_2}{p_1}\right)\right] = \int p_1(x) f\left(\frac{p_2(x)}{p_1(x)}\right) dx$$

↑ $f(t)$: convex.

① If $f(1) = 0$,

$$D_f(p_1, p_2) = \int p_1(x) f\left(\frac{p_2(x)}{p_1(x)}\right) dx$$

$$\geq f\left(\int p_1(x) \frac{p_2(x)}{p_1(x)} dx\right)$$

$$= f\left(\int p_2(x)\, dx\right) = f(1) = 0.$$

② $D_f(p_1, p_2)$ is minimized when $p_1(x) = p_2(x)$ for all x.

f-Divergences

$$f(t) = -\log t \Rightarrow D_f(p_1, p_2) = -\int p_1(x) \log \frac{p_2(x)}{p_1(x)} dx$$

KL-divergence

$$f(t) = 1 - \sqrt{t} \Rightarrow D_f(p_1, p_2) = \int p_1(x) \left(1 - \sqrt{\frac{p_2(x)}{p_1(x)}}\right) dx$$

$$= 1 - \int \sqrt{p_1(x) p_2(x)}\, dx$$

Hellinger distance

Bhattacharyya coefficient

$$f(t) = \frac{1}{1+t} \Rightarrow D_f(p_1, p_2) = \int p_1(x) \left(\frac{1}{1 + p_2(x)/p_1(x)}\right) dx$$

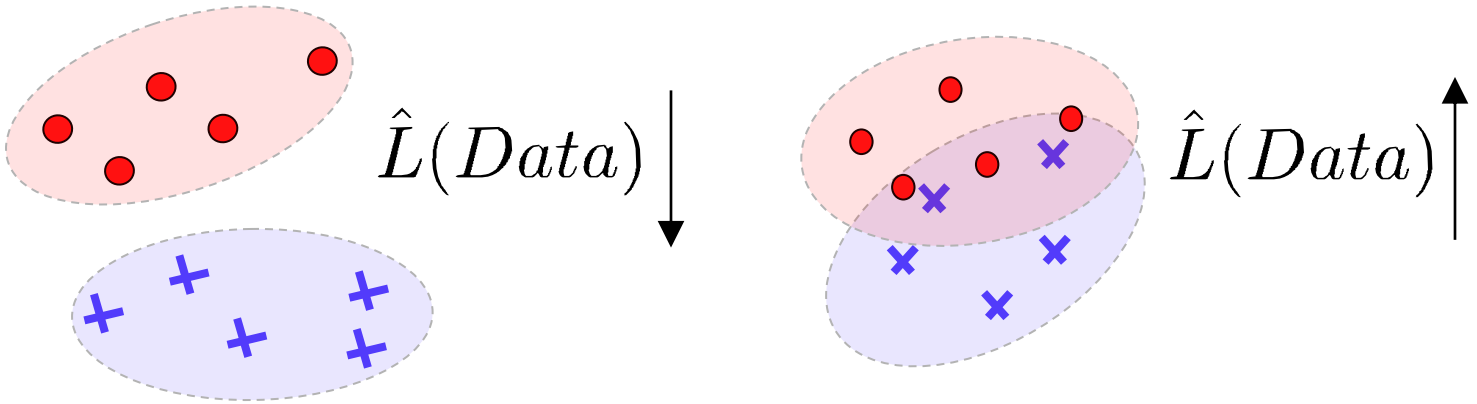
$$= 1 - \int \frac{p_1(x) p_2(x)}{p_1(x) + p_2(x)}\, dx$$

Lecam distance

Triangular discrimination distance

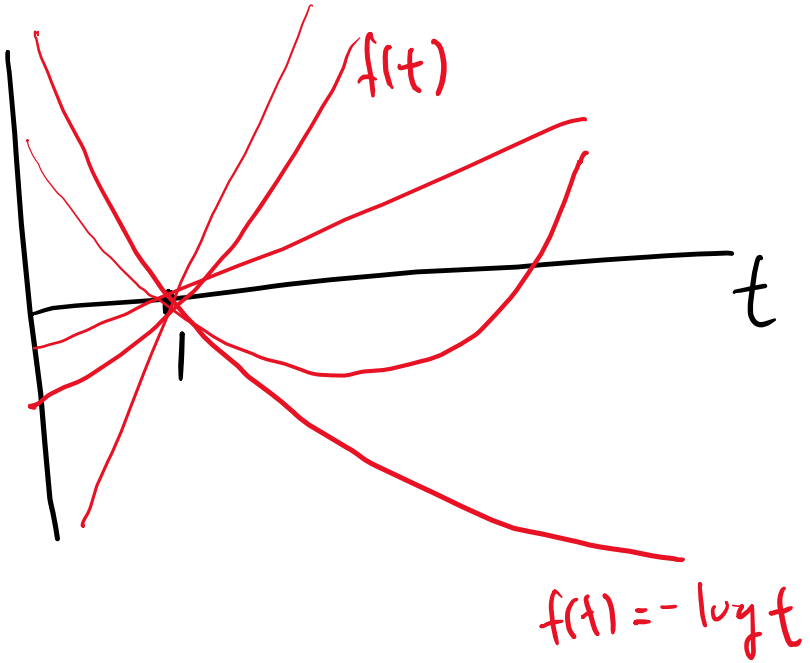
$$D_f(p_1(\mathbf{x}), p_2(\mathbf{x})) = \int p_1(\mathbf{x}) f\left(\frac{p_2(\mathbf{x})}{p_1(\mathbf{x})}\right) d\mathbf{x}$$

f-divergences



f-Divergences

$$D_f(p_1, p_2) = \int p_1(x) f\left(\frac{p_2(x)}{p_1(x)}\right) dx$$



Reparametrization-free Measure

$z = L^T x$ Volume $dz = |L| dx$

$p(x)dx = p(z)dz$

$p(x) = p(z)\frac{dz}{dx}$

$= |L| p(z)$

$$D_f(p_1(x) \| p_2(x))$$

$$= \int p_1(x) f\left(\frac{p_2(x)}{p_1(x)}\right) dx$$

$$= \int |L| p_1(z) f\left(\frac{|L| p_2(z)}{|L| p_1(z)}\right) \frac{dz}{|L|}$$

$$= \int p_1(z) f\left(\frac{p_2(z)}{p_1(z)}\right) dz$$

$$= D_f(p_1(z) \| p_2(z))$$

Role of Loss

- $\hat{L}(Data)$: Empirical loss
  - Consider two separate roles
    - (1) Find the true posterior for given features
    - (2) Improve the features (or representations)

- Optimizing *f*-divergence already encompasses the result of the first role of optimizing $\hat{L}(Data)$

- Use *f*-divergence in case we are interested in (2) without performing (1).

Loss function

$$\widehat{L}(g) = \frac{1}{N}\sum_{i=1}^{N}\phi(y_i, g(\mathbf{x}_i))$$

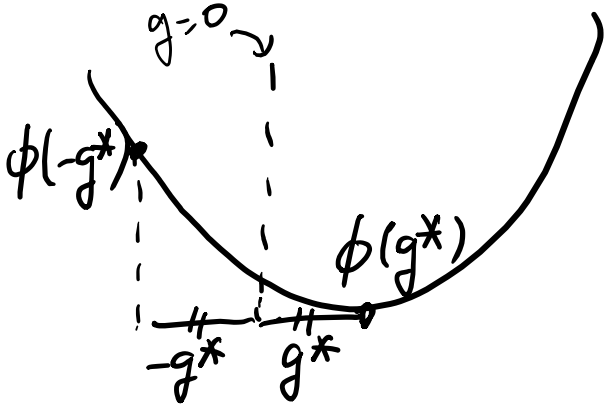
Margin-based

$\phi(y, g(x)) = \phi(yg(x))$

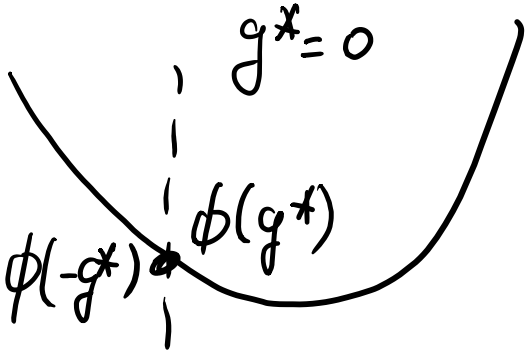
$y \in \{-1, 1\}$

- Optimal g with convex ϕ

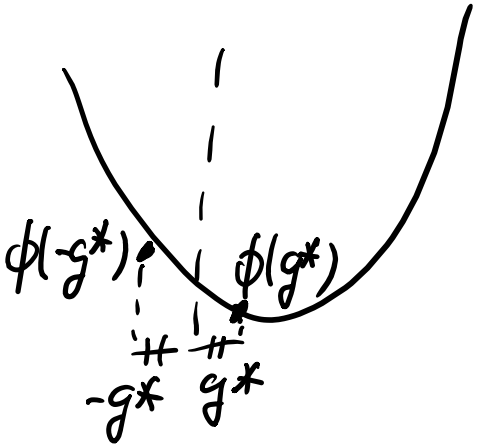
$P(y=1|\mathbf{x}) = 1$
$P(y=-1|\mathbf{x}) = 0$



$P(y=1|\mathbf{x}) = 0$
$P(y=-1|\mathbf{x}) = 1$



$P(y=1|\mathbf{x}) = 0.3$
$P(y=-1|\mathbf{x}) = 0.7$



Ex) Logistic Loss (Cross Entropy)

$$\phi(\alpha) = \log(1+\exp(-\alpha)) \qquad \alpha = y\,g(x)$$

Expectation at x:

$$E[L(g(x))] = P(y=1|x)\,\phi(g(x)) + P(y=-1|x)\,\phi(-g(x))$$

$$= \frac{p_1}{p_1+p_{-1}}\log(1+\exp(-g)) + \frac{p_{-1}}{p_1+p_{-1}}\log(1+\exp(g))$$

$$\frac{d}{dg}E[L] = 0 \quad \Rightarrow \quad g^* = \log\frac{p_{-1}}{p_1}$$

Plug-in

$$E[L] = E_{y_i}\left[\frac{1}{N}\sum_{i=1}^{N} \log\left(1+\exp\left(-y_i g^*(x_i)\right)\right)\right]$$

$$= E_{y_i}\left[\frac{1}{N}\sum_{i=1}^{N} \log\left(1+\exp\left(-y_i \log\frac{p_1(x_i)}{p_{-1}(x_i)}\right)\right)\right]$$

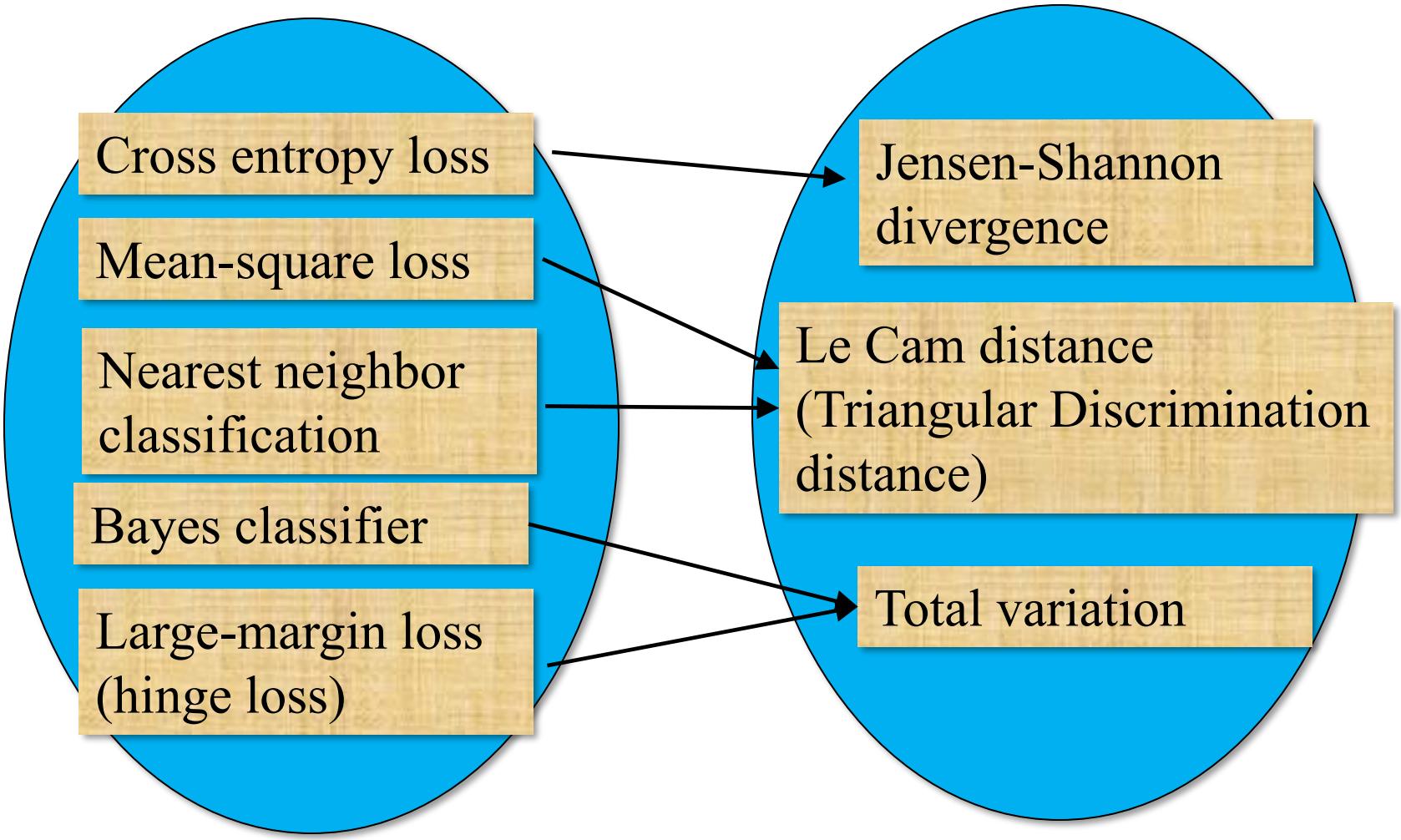
$$\approx \frac{1}{N}\sum_{i=1}^{N} \frac{p_1}{p_1+p_{-1}} \log\left(1+\frac{p_{-1}}{p_1}\right) + \frac{p_{-1}}{p_1+p_{-1}} \log\left(1+\frac{p_1}{p_{-1}}\right)$$

($\frac{p_{-1}}{p_1}$: $y_i = 1$; $\frac{p_1}{p_{-1}}$: $y_i = -1$)

$$\approx -\frac{1}{2} KL\left(p_1 \,\|\, \frac{p_1+p_{-1}}{2}\right) - \frac{1}{2} KL\left(p_{-1} \,\|\, \frac{p_1+p_{-1}}{2}\right) + \log 2$$

- We can generalize that a loss function with an optimal prediction function is related to its corresponding *f*-divergence. (Many to one relationship)

Loss functions – *f*-divergences



Nearest Neighbor Density Functional Estimation From Inverse Laplace Transform

J. Jon Ryu, *Student Member, IEEE*, Shouvik Ganguly, *Member, IEEE*, Young-Han Kim, *Fellow, IEEE*, Yung-Kyun Noh, *Member, IEEE*, and Daniel D. Lee, *Fellow, IEEE*

***Abstract*— A new approach to L_2-consistent estimation of a general density functional using k-nearest neighbor distances is proposed, where the functional under consideration is in the form of the expectation of some function f of the densities at each point. The estimator is designed to be asymptotically unbiased, using the convergence of the normalized volume of a k-nearest neighbor ball to a Gamma distribution in the large-sample limit, and naturally involves the inverse Laplace transform of a scaled version of the function f. Some instantiations of the proposed estimator recover existing k-nearest neighbor based estimators of Shannon and Rényi entropies and Kullback–Leibler and Rényi divergences, and discover new consistent estimators for many other functionals such as logarithmic entropies and divergences. The L_2-consistency of the proposed estimator is established for a broad class of densities for general functionals, and the convergence rate in mean squared error is established as a function of the sample size for smooth, bounded densities.**

***Index Terms*— Density functional estimation, information measure, nearest neighbor, inverse Laplace transform.**

I. Introduction

THIS paper studies the problem of estimating an entropy functional of the form

where $f: \mathbb{R}_+ \to \mathbb{R}$ is a given function and p is a probability density over $\mathbb{R}^d$. Table I lists examples of f and the corresponding functional T_f. The goal is to estimate $T_f(p)$ based on independent and identically distributed (i.i.d.) samples $\mathbf{X}_{1:m} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$ from p by forming an estimator $\hat{T}_f^m(\mathbf{X}_{1:m})$ that converges to $T_f(p)$ in L_2 as the sample size m grows to infinity, that is,

$$\lim_{m\to\infty} \mathbb{E}\left[\left(\hat{T}_f^m(\mathbf{X}_{1:m}) - T_f(p)\right)^2\right] = 0.$$

More generally, let $f: \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ and consider a divergence functional

$$T_f(p,q) := \mathbb{E}_{\mathbf{X}\sim p}[f(p(\mathbf{X}), q(\mathbf{X}))] = \int f(p(\mathbf{x}), q(\mathbf{x}))p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

of a pair of probability densities p and q over $\mathbb{R}^d$. Table II lists examples of f and the corresponding T_f. In this case, the main problem is to construct an estimator $\hat{T}_f^{m,n}(\mathbf{X}_{1:m}, \mathbf{Y}_{1:n})$ based on i.i.d. samples $\mathbf{X}_{1:m}$ from p and $\mathbf{Y}_{1:n}$ from q, independent of each other, such that

Typical Set

- Asymptotic Equipartition Property
  - The probability of generated data

$$\mathbf{x}_1, \ldots, \mathbf{x}_N \sim p(\mathbf{x})$$

$$-\frac{1}{N} \log p(\mathbf{x}_1, \ldots, \mathbf{x}_N) \rightarrow H(\mathbf{x}) \quad \text{in probability}$$

(weak law of large numbers)

Typical set $A_\epsilon^{(N)}$: Set of sequences $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{X}$ with condition

$$e^{-N(H(\mathbf{x})+\epsilon)} \leq p(\mathbf{x}_1, \ldots, \mathbf{x}_N) \leq e^{-N(H(\mathbf{x})-\epsilon)}$$

For any δ, there exists N s.t. $\Pr(A_\epsilon^{(N)}) > 1 - \delta$.

Typical Set

- $(\mathbf{x}_1, \ldots, \mathbf{x}_N) \in A_\epsilon^{(N)}$

$$\rightarrow \quad H(\mathbf{x}) - \epsilon \leq -\frac{1}{N} \log p(\mathbf{x}_1, \ldots, \mathbf{x}_N) \leq H(\mathbf{x}) + \epsilon$$

- $\Pr(A_\epsilon^{(N)}) > 1 - \epsilon$ for N sufficiently large

Because δ in previous page is any positive value

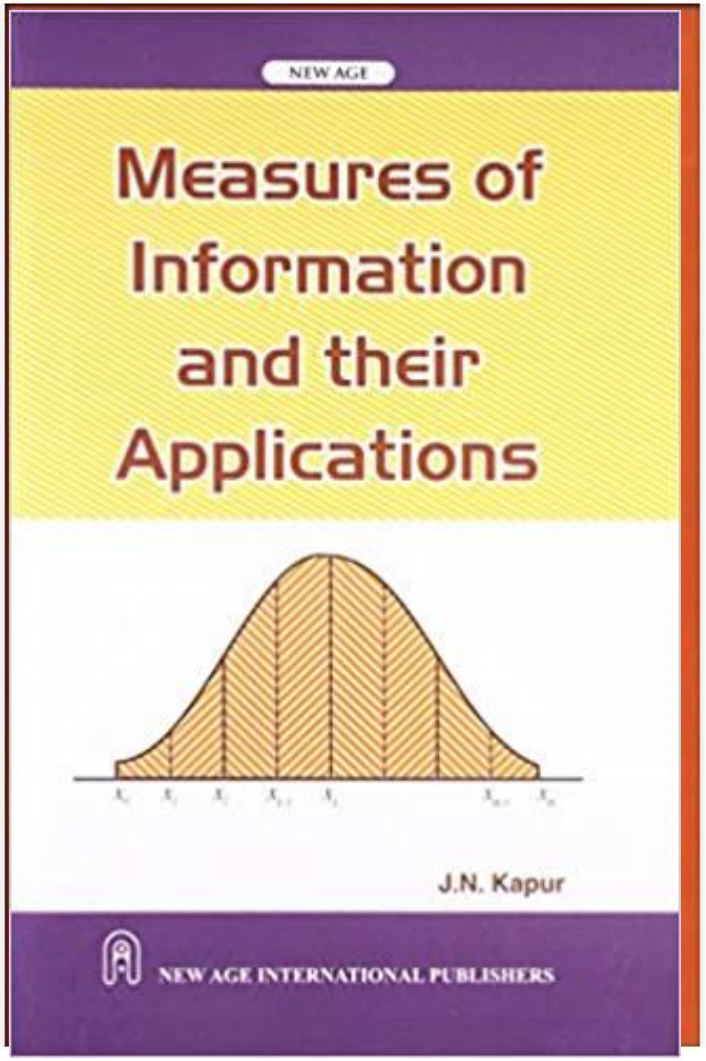
- $|A_\epsilon^{(N)}| \leq e^{N(H(\mathbf{x})+\epsilon)}$ for any N

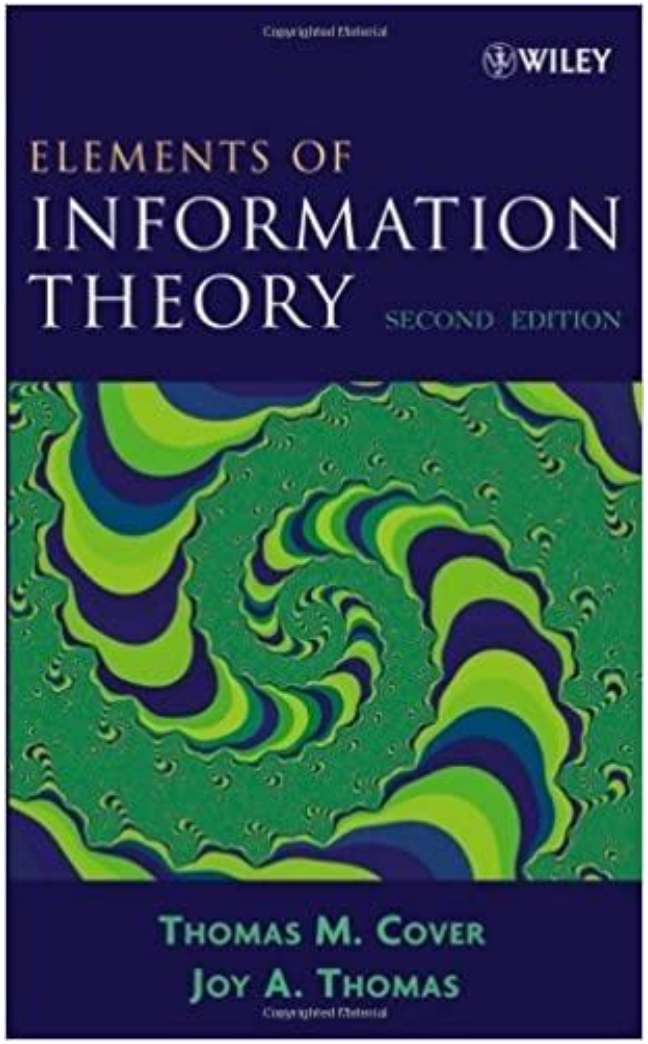
- $|A_\epsilon^{(N)}| \geq (1-\epsilon) e^{N(H(\mathbf{x})-\epsilon)}$ for N sufficiently large

$$\Pr\left((\mathbf{x}_1,\ldots,\mathbf{x}_N) : p(\mathbf{x}_1,\ldots,\mathbf{x}_N) = e^{-N(H(\mathbf{x})\pm\epsilon)}\right)$$

About 2^{NH} number of elements have almost equal probability

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_N) \approx 2^{-NH}$$


NEW AGE
Measures of Information and their Applications
J.N. Kapur
NEW AGE INTERNATIONAL PUBLISHERS


WILEY
ELEMENTS OF INFORMATION THEORY
SECOND EDITION
THOMAS M. COVER
JOY A. THOMAS

Nearest Neighbor Algorithms in High Dimensions

Theory and practice

Cambridge University Press

Yung-Kyun Noh & Masashi Sugiyama


Thank you