Basic Probability Theory for Machine Learning

OSIA 동계 단기강좌
2017. 2. 13 (Mon.)

***Yung-Kyun Noh* (노영균)**
Seoul National University

Contents

- Probability / Probability density
- Conditional probability (density)

$$p(\mathbf{x}_2|\mathbf{x}_1)$$

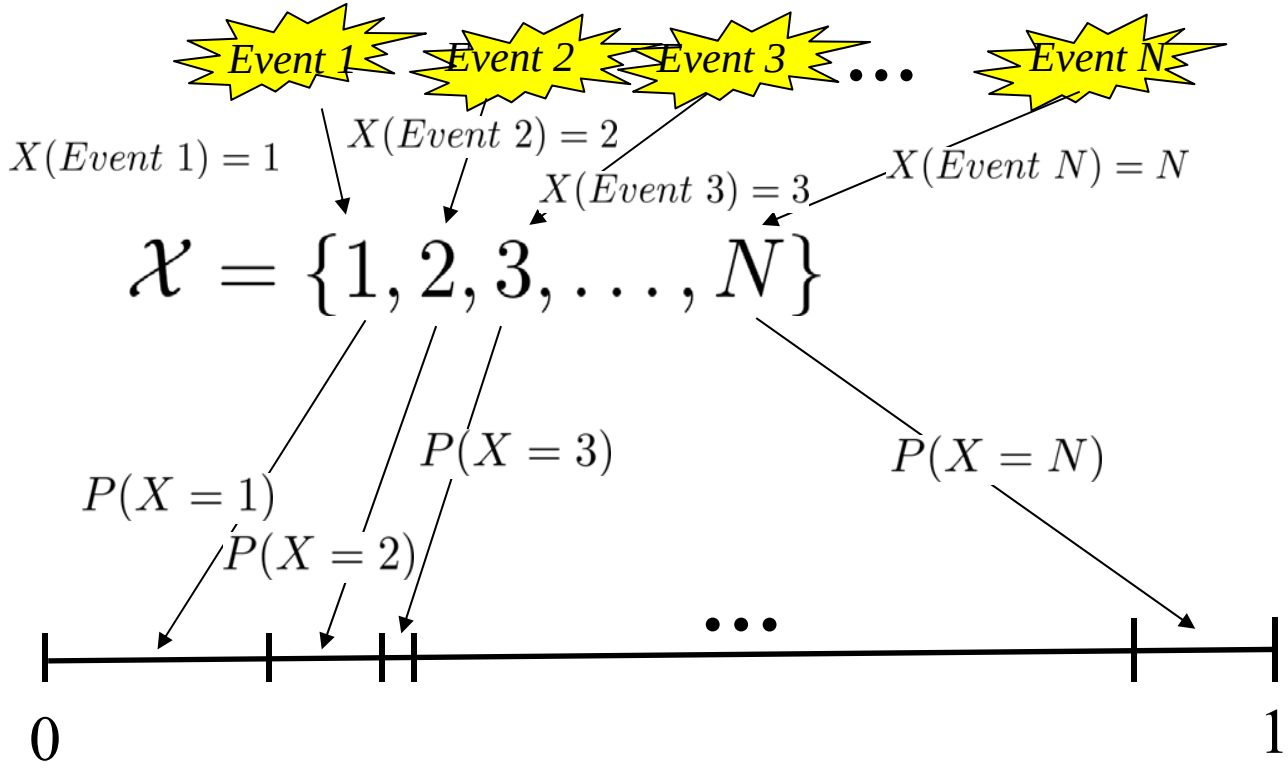
$$P(y|\mathbf{x}) \quad \mathbf{x} \in \mathbb{R}^D, y \in \{1, 2\}$$

- Marginal probability (density)
- Model for consistent learner
- Parameter estimation

Probability

$$P(X) : \mathcal{X} \rightarrow [0, 1]$$

- Mapping from a random variable to a number



Probability

X: random variable X_1, X_2: set of outputs of random variables

$$P(X_1) \equiv P(X \in X_1)$$

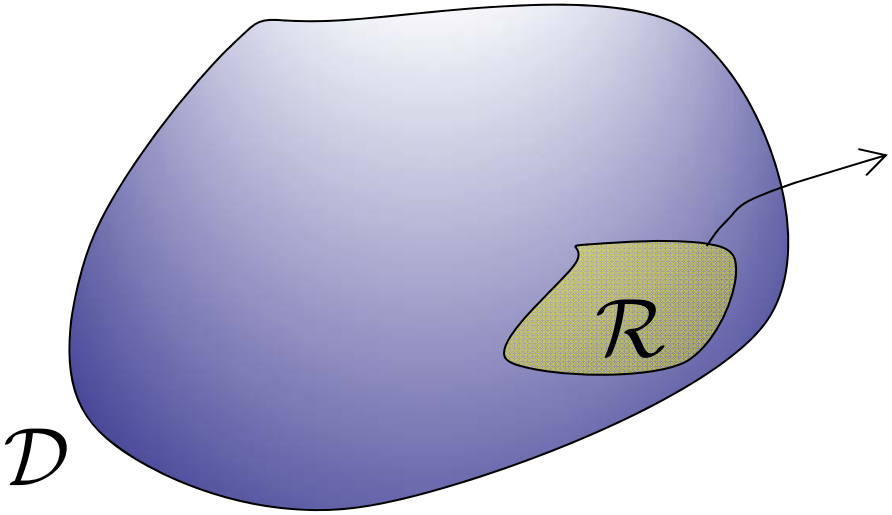
$$P(X_1 \cup X_2) = P(X_1) + P(X_2) - P(X_1 \cap X_2)$$

$$X_1 = \{1, 2, 3, 4\}, \; X_2 = \{3, 4, 5\}$$
$$P(1, 2, 3, 4, 5) = P(1, 2, 3, 4) + P(3, 4, 5) - P(3, 4)$$

$$P(X_1 \cup X_2) = P(X_1) + P(X_2) \text{ if } X_1 \cap X_2 = \phi$$

Probability and Probability Density

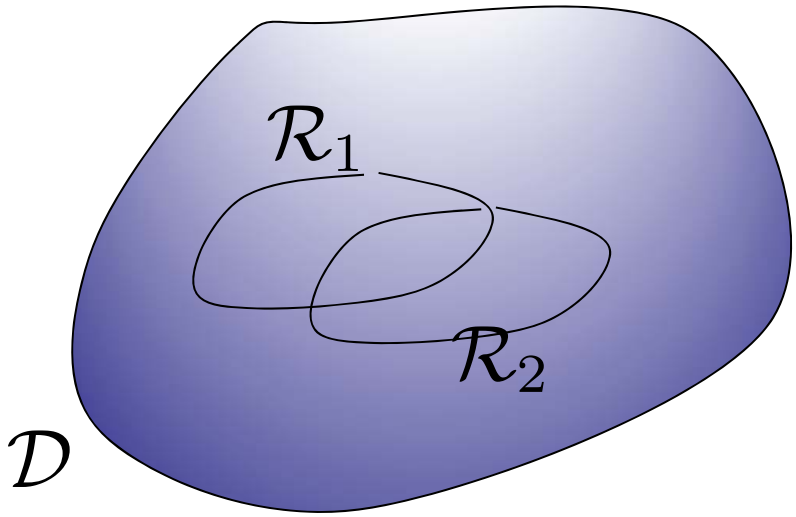
$$p(\mathbf{x}) \in \mathbb{P} \qquad \int_{\mathcal{D}} p(\mathbf{x})d\mathbf{x} = 1 \qquad p(\mathbf{x}) \geq 0$$



$$P(\mathbf{x} \in \mathcal{R}) = \int_{\mathcal{R}} p(\mathbf{x})d\mathbf{x}$$

$$\text{Probability} = \int_{\mathcal{R}} p(\mathbf{x})d\mathbf{x}$$

Probability and Probability Density



$$P(\mathbf{x} \in \mathcal{R}_1 \cup \mathbf{x} \in \mathcal{R}_2) = \int_{\mathcal{R}_1 \cup \mathcal{R}_2} p(\mathbf{x}) d\mathbf{x}$$

$$= P(\mathcal{R}_1) + P(\mathcal{R}_2) - P(\mathcal{R}_1 \cap \mathcal{R}_2)$$

Event is defined infinitesimally:

$\mathcal{R}$: set of infinitesimal events

Can you explain the meaning of these functions?

$P(X = 1)$

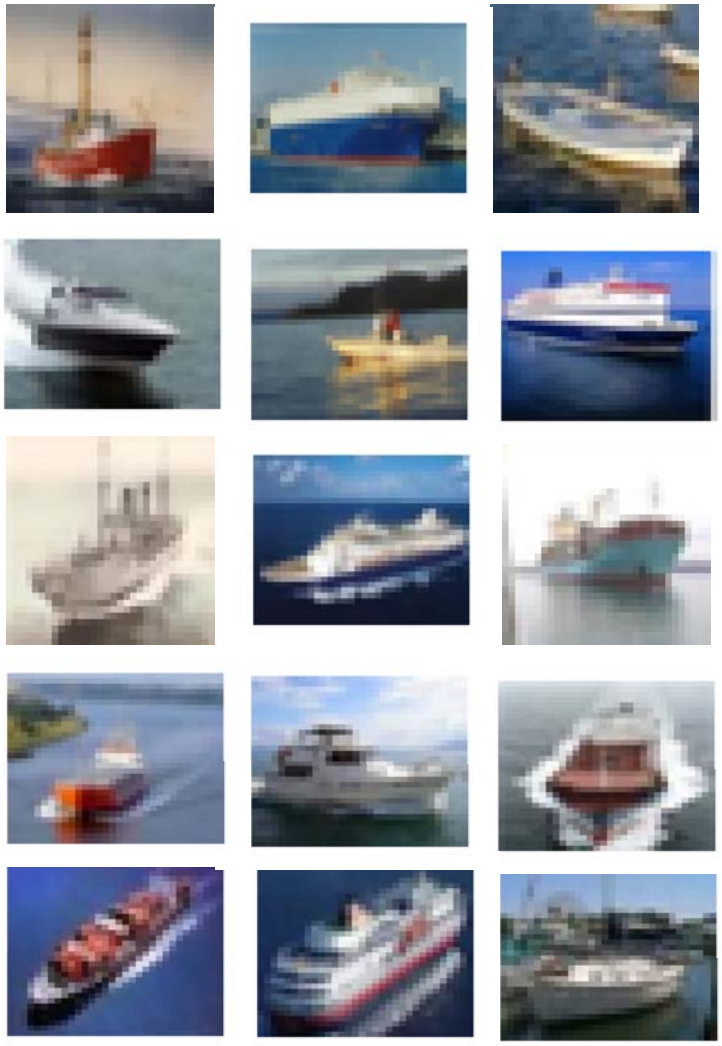
$P(X = 1|Y = 2)$

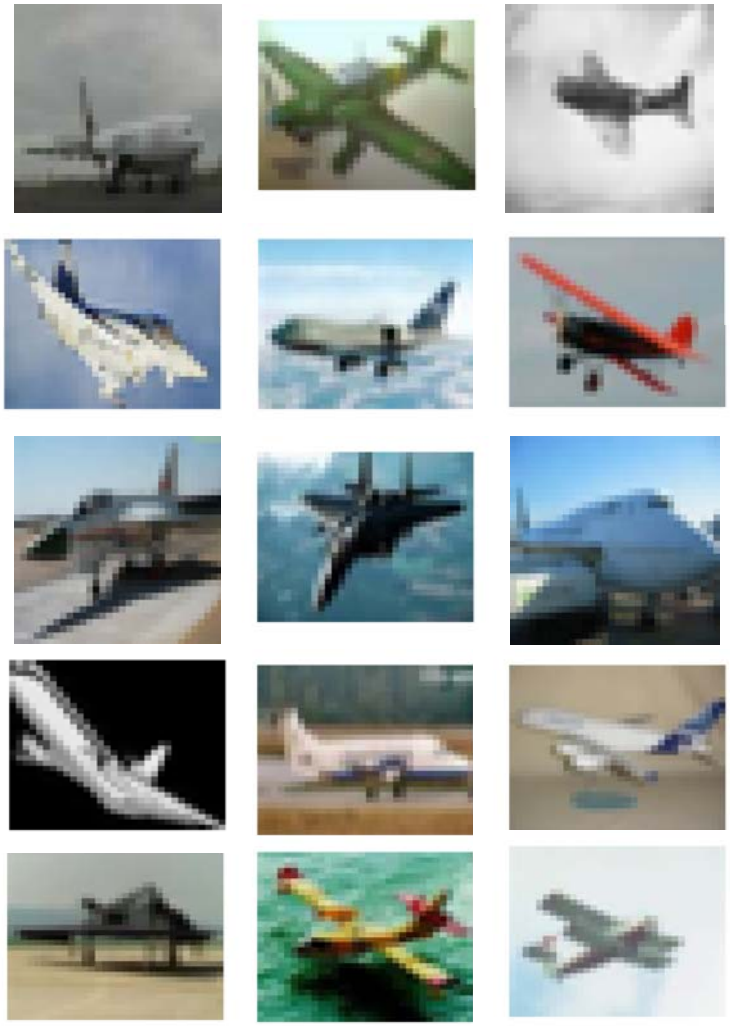
$p(x = 1)$ Compare with $P(x = 1)$?

$p(x = 1|y = 2)$

What is the *rule* of discriminating ship images from airplane images?

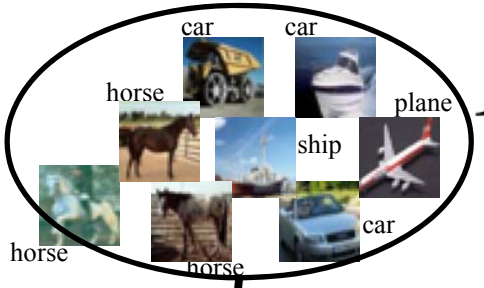
Can you still make rules?





Supervised Learning (Prediction)

- Method:
  - Learning from *examples* and can classify an *unseen data*



X_{train}

$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$

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

Learn θ

$y(\mathbf{x}) \in \mathcal{M}$

New unseen data

classifier

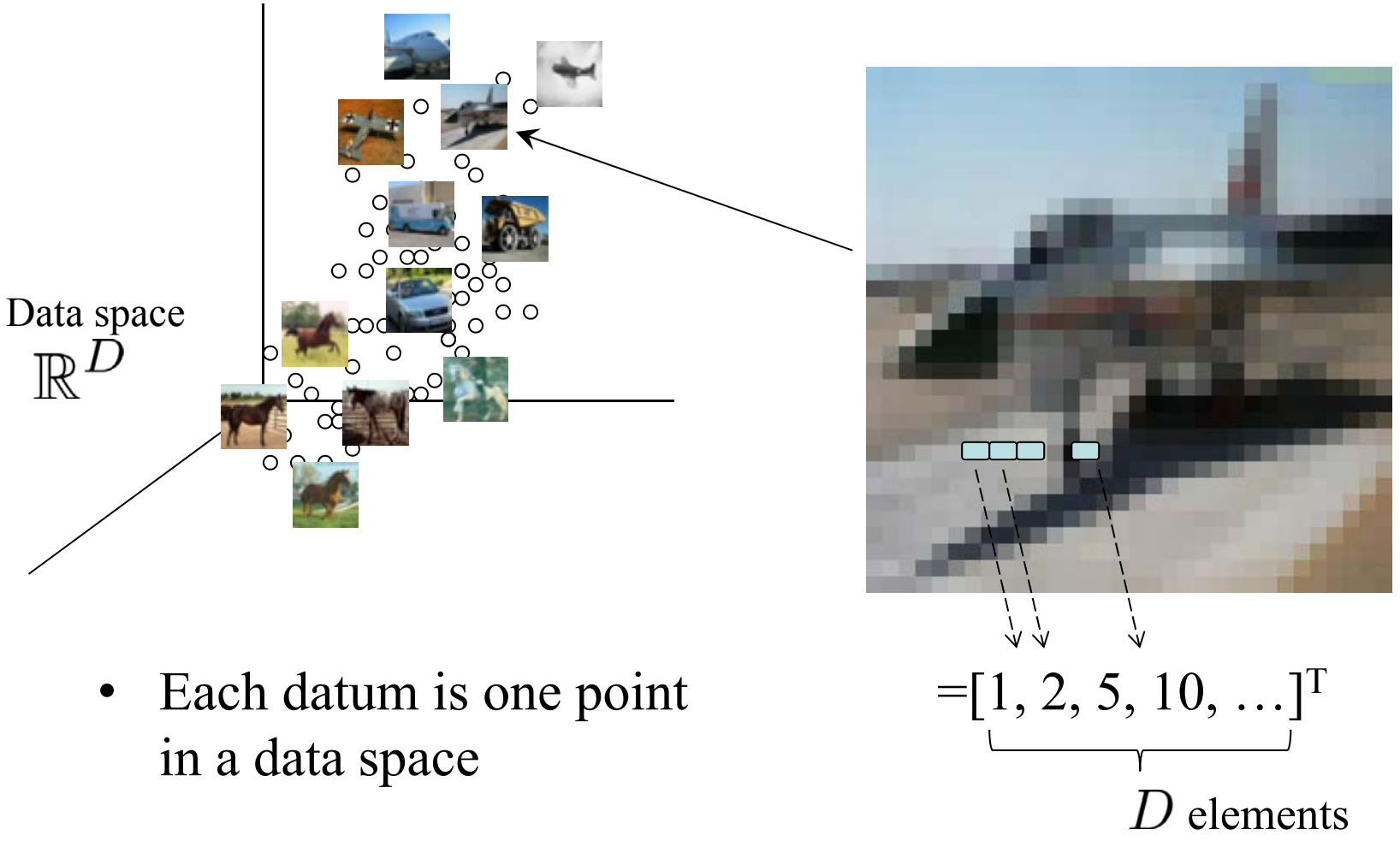
$f(\mathbf{x};\theta_{learned})$

label (a horse)

$y(\mathbf{x};\mathcal{D})$

[Based on the assumption of regularity]

Representation of Data

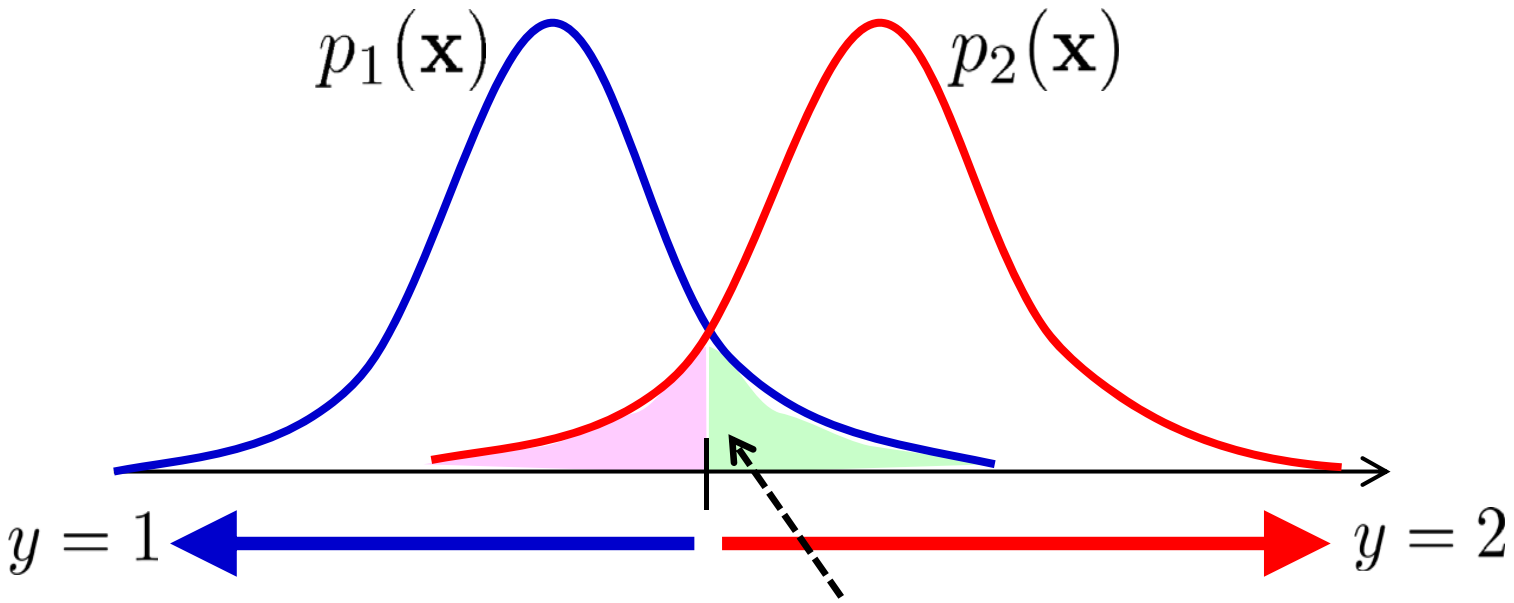


- Each datum is one point in a data space

Classification

Bayes Optimal Classifier

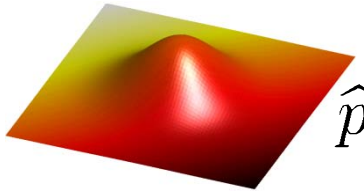
- Our ultimate goal is *not a zero error.*



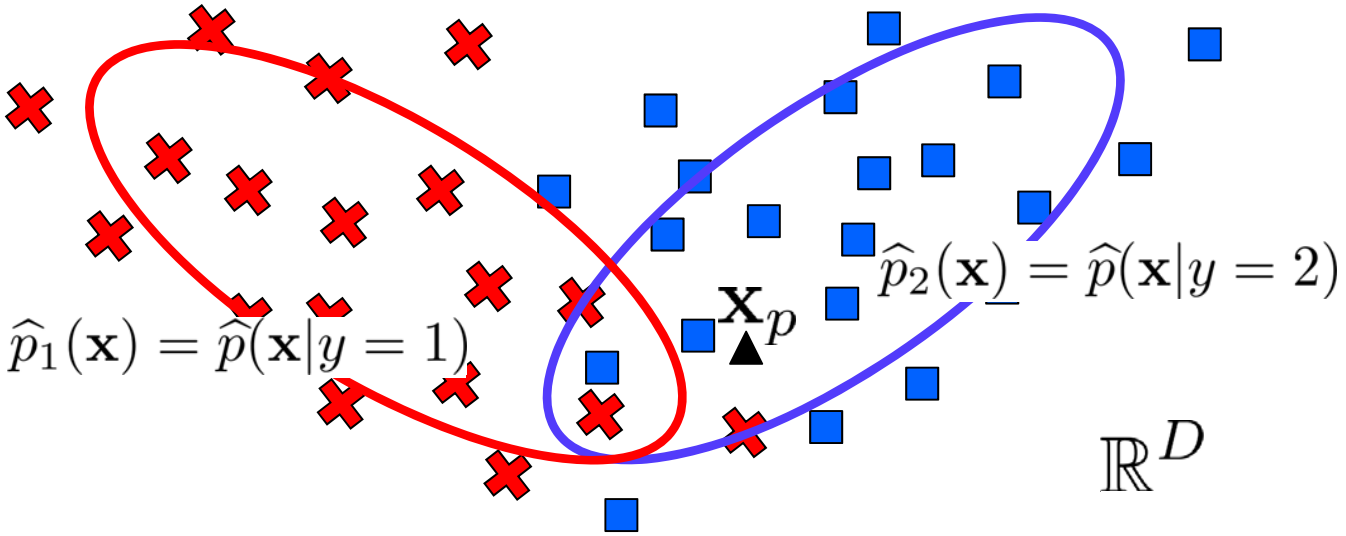
$$E_{Bayes} = \frac{1}{2}\int \min[p_1(\mathbf{x}), p_2(\mathbf{x})]d\mathbf{x}$$

Figure credit: Masashi Sugiyama

Model on Each Class



$$\widehat{p}_c(\mathbf{x}),\ c \in \{1,2\}$$



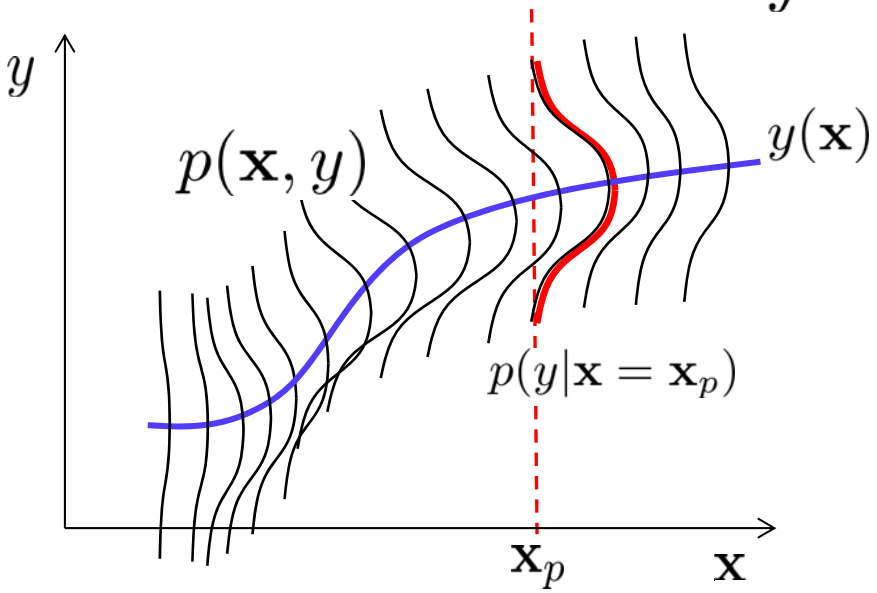
$$\widehat{p}_1(\mathbf{x}_p) \geq \widehat{p}_2(\mathbf{x}_p) \rightarrow y_p = 1$$
$$\widehat{p}_1(\mathbf{x}_p) < \widehat{p}_2(\mathbf{x}_p) \rightarrow y_p = 2$$

- Model: Class-conditional density as a Gaussian

Optimal Regression

- Minimizing mean square error

$$y(\mathbf{x}) = \mathbb{E}\left[y|\mathbf{x}\right] = \int y\, p(y|\mathbf{x})dy$$



Minimize

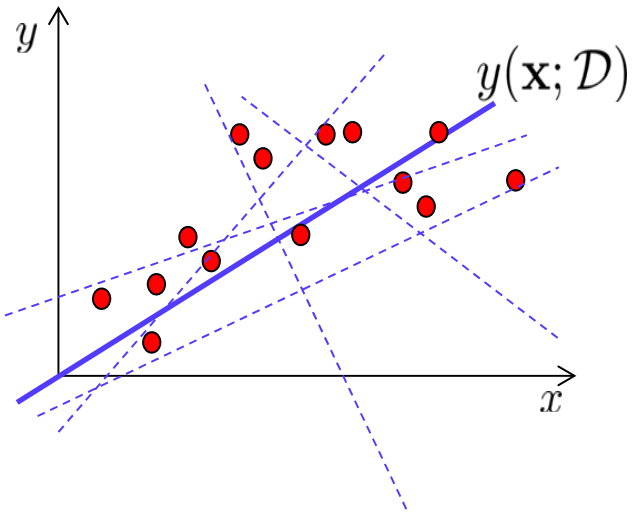
$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - y\}^2\, p(\mathbf{x}, y)d\mathbf{x}dy$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}]\}^2\, p(\mathbf{x})d\mathbf{x} + \int \{\mathbb{E}[y|\mathbf{x}] - y\}^2\, p(\mathbf{x}, y)d\mathbf{x}dy$$

→ Minimized when $y(\mathbf{x}) = \mathbb{E}\left[y|\mathbf{x}\right]$

Model for Regression

- Obtain regression function $y(\mathbf{x};\mathcal{D}) \in \mathcal{M}$ from data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N} \sim p(\mathbf{x}, y)$



- Choose a model $\mathcal{M}$ where the following expectation is minimized:

$$\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}[y|\mathbf{x}]\}^2\right]$$

  - Minimized for $y(\mathbf{x};\mathcal{D}) = \mathbb{E}[y|\mathbf{x}]$

- Bias-Variance tradeoff

$$\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}[y|\mathbf{x}]\}^2 = \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - \mathbb{E}[y|\mathbf{x}]\}^2$$

$$\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}[y|\mathbf{x}]\}^2\right]$$

Variance

Bias^2

$$= \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - \mathbb{E}[y|\mathbf{x}]\}^2$$

Several Rules

$$\sum_{X_i \in all\ \ disjoint\ \ set} P(X = X_i) = 1$$

$$\sum_{X_i \in all\ \ disjoint\ \ set} P(X = X_i | Z = Z_j) = 1$$

$$\sum_{Z_j \in all\ \ disjoint\ \ set} P(X = X_i | Z = Z_j) = ?$$

For More Than Two Random Variables

- For three disjoint sets X_1, X_2, X_3 for a random variable X and another three disjoint sets Y_1, Y_2, Y_3 for a random variable Y:

Y \ X	X_1	X_2	X_3	
Y_1	$P(X_1, Y_1)$	$P(X_2, Y_1)$	$P(X_3, Y_1)$	$P(Y_1)$
Y_2	$P(X_1, Y_2)$	$P(X_2, Y_2)$	$P(X_3, Y_2)$	$P(Y_2)$
Y_3	$P(X_1, Y_3)$	$P(X_2, Y_3)$	$P(X_3, Y_3)$	$P(Y_3)$
	$P(X_1)$	$P(X_2)$	$P(X_3)$	1

$P(X \in \{X_1, X_2\}, Y \in Y_1)$

Conditional Probability

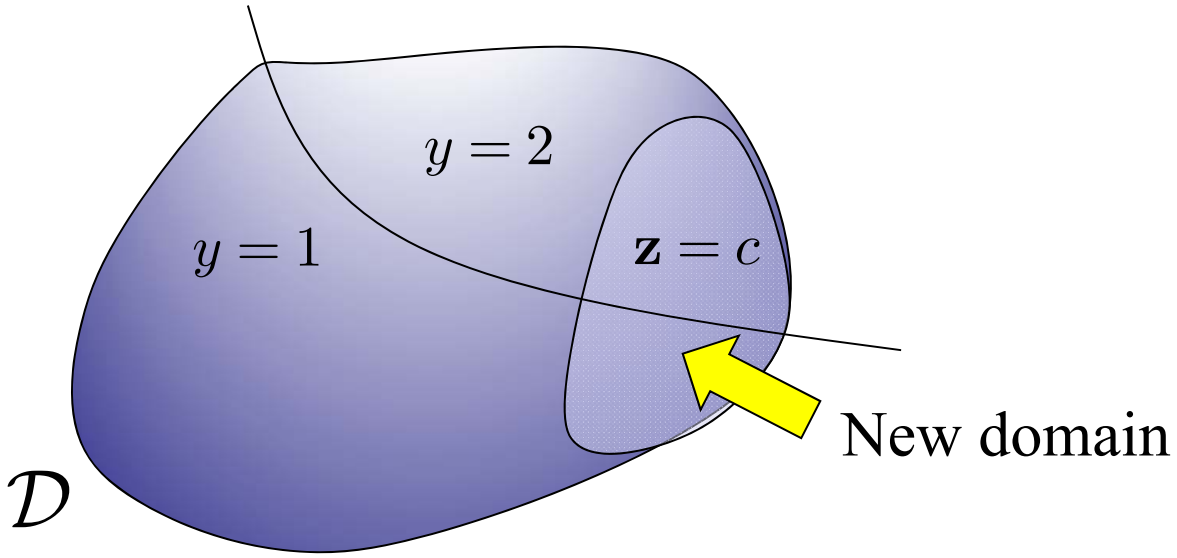
Y \ X	X_1	X_2	X_3	
Y_1	$P(X_1, Y_1)$	$P(X_2, Y_1)$	$P(X_3, Y_1)$	$P(Y_1)$
Y_2	$P(X_1, Y_2)$	$P(X_2, Y_2)$	$P(X_3, Y_2)$	$P(Y_2)$
Y_3	$P(X_1, Y_3)$	$P(X_2, Y_3)$	$P(X_3, Y_3)$	$P(Y_3)$
	$P(X_1)$	$P(X_2)$	$P(X_3)$	1

$$P(X = X_1|Y = Y_1) = \frac{P(X_1, Y_1)}{P(X_1, Y_1) + P(X_2, Y_1) + P(X_3, Y_1)}$$
$$= \frac{P(X_1, Y_1)}{P(Y_1)}$$

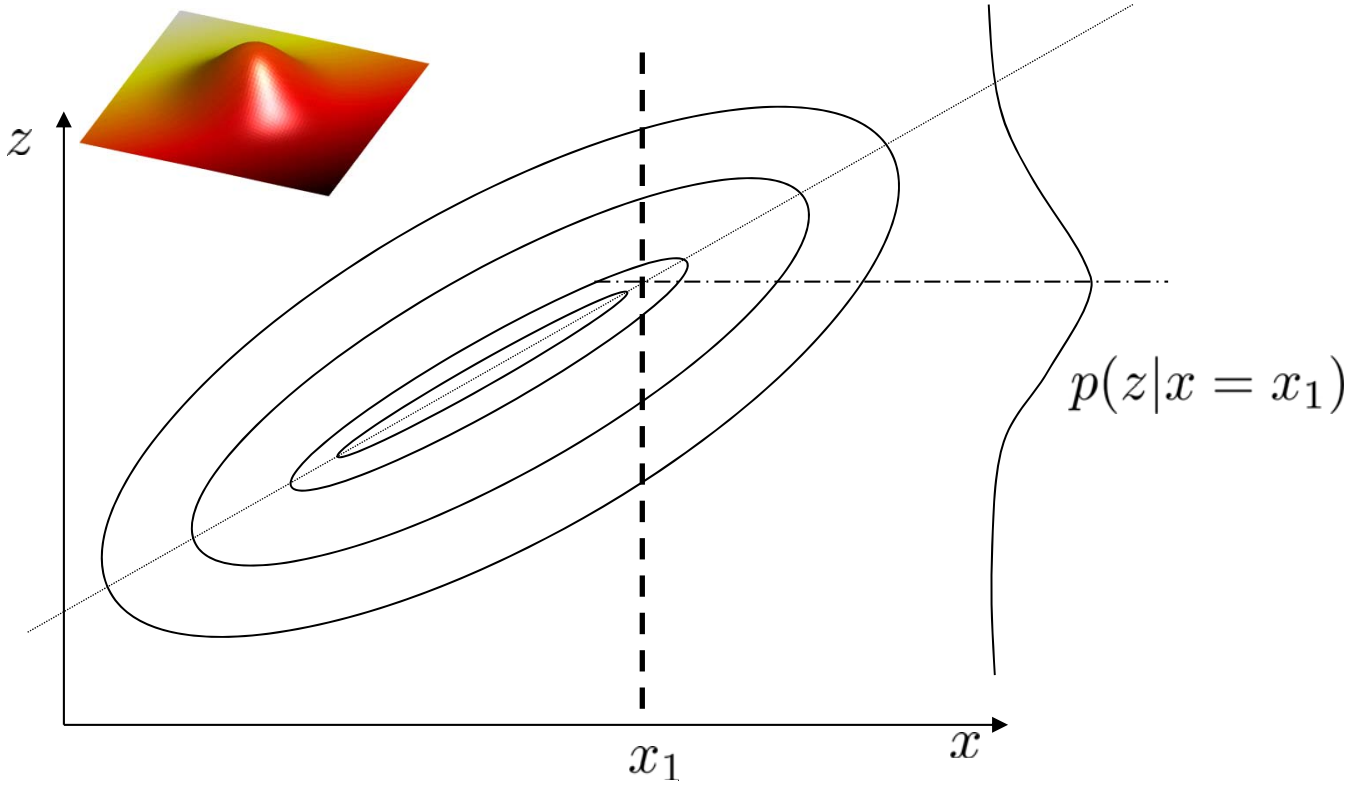
Conditional Probability Density

$$p(\mathbf{x}, \mathbf{z}) \qquad \mathbf{x} \in \mathbb{R}^{D_\mathbf{x}}, \mathbf{z} \in \mathbb{R}^{D_\mathbf{z}}$$

$$\rightarrow \quad p(\mathbf{x}|\mathbf{z} = c) = \frac{p(\mathbf{x}, \mathbf{z} = c)}{\int p(\mathbf{x}, \mathbf{z} = c) d\mathbf{x}}$$

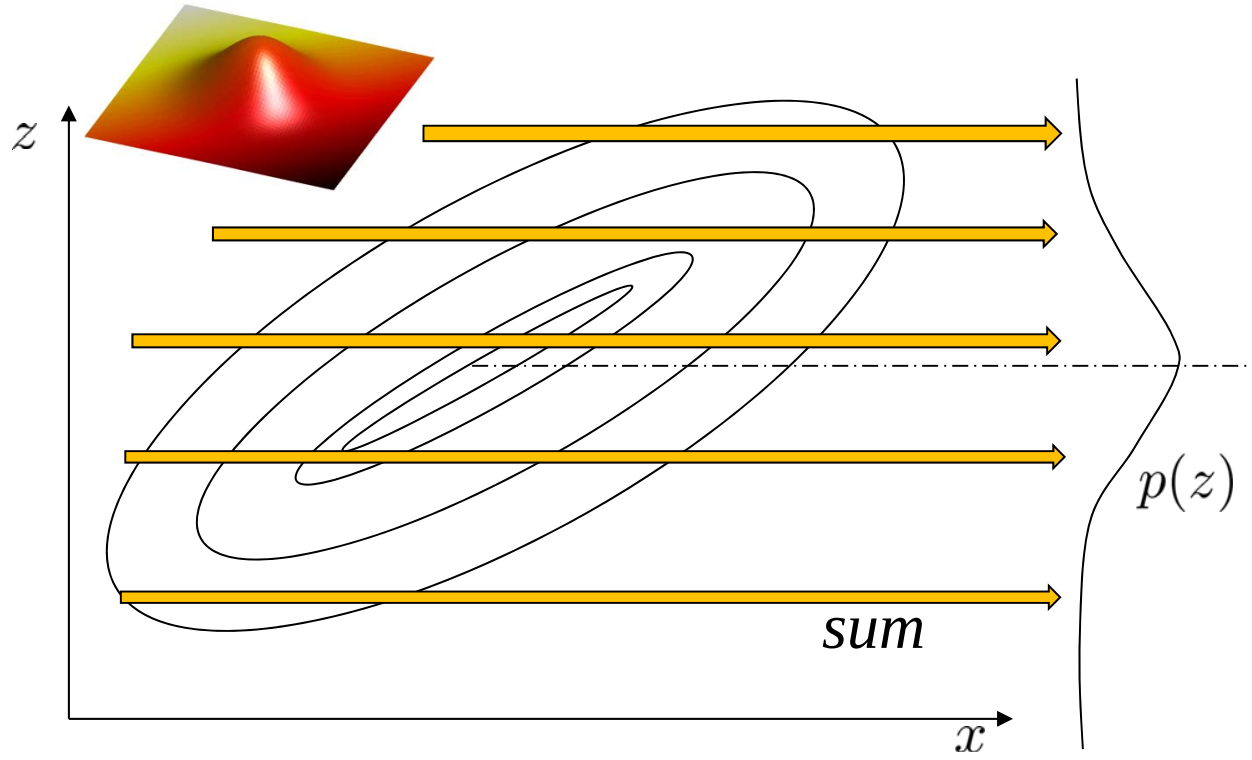


Conditional Probability Density



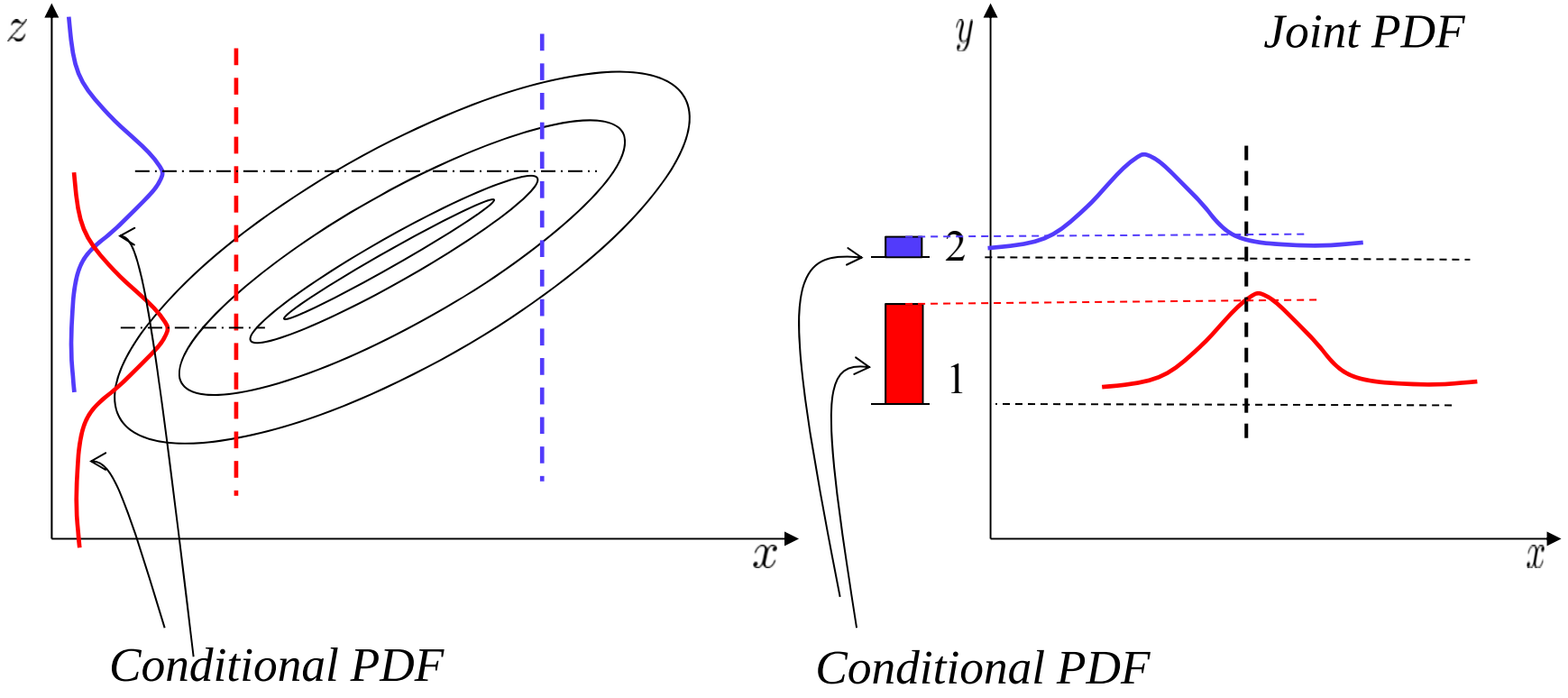
$$p(z|x = x_1) = \frac{p(z, x = x_1)}{\int_{x=x_1} p(z, x)dz} = \frac{p(z, x)}{p(x)}$$

Marginal Probability Density

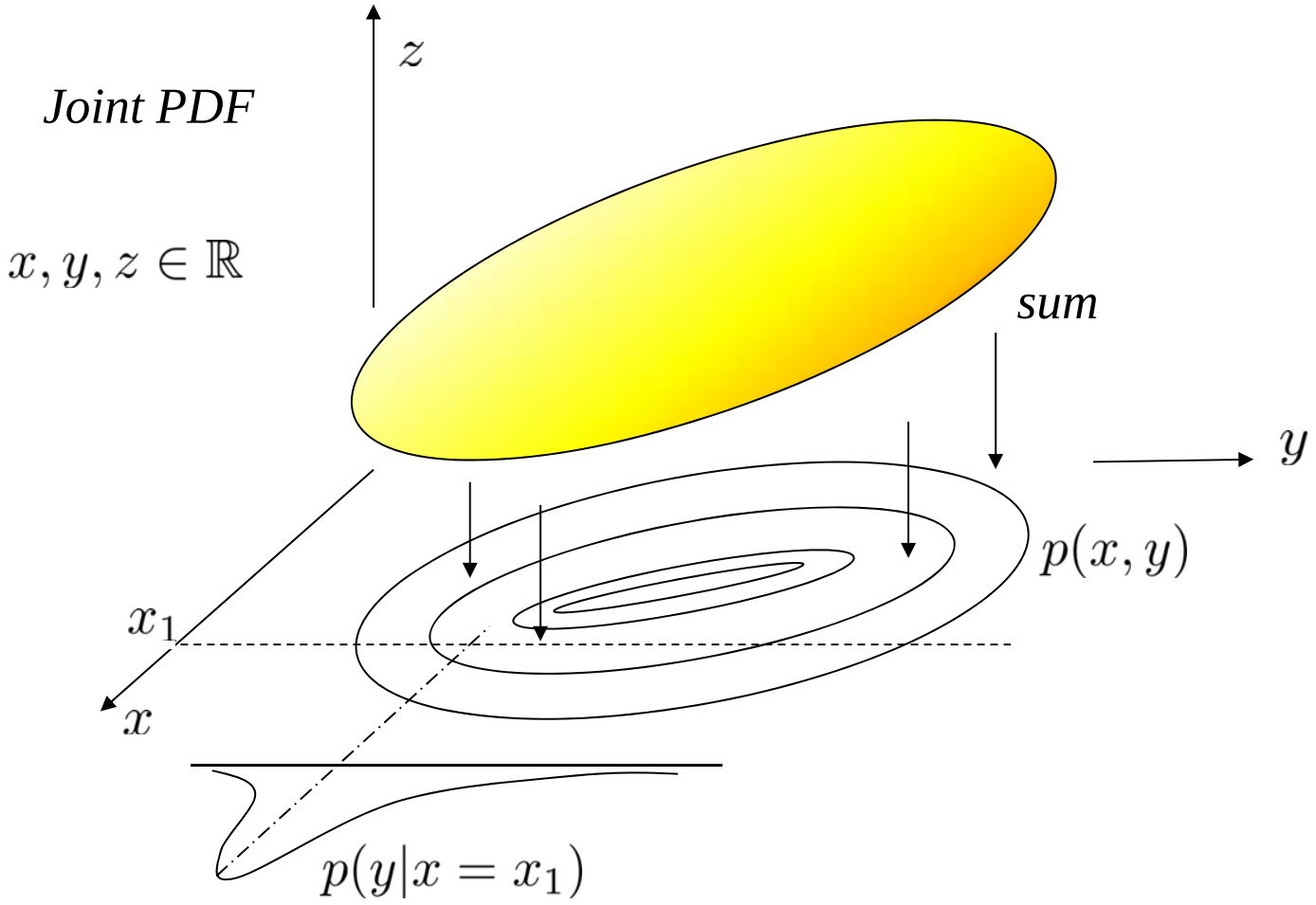


$$p(z) = \int p(z, x) dx$$

Marginal Probability Density and Conditional Probability Density in Machine Learning

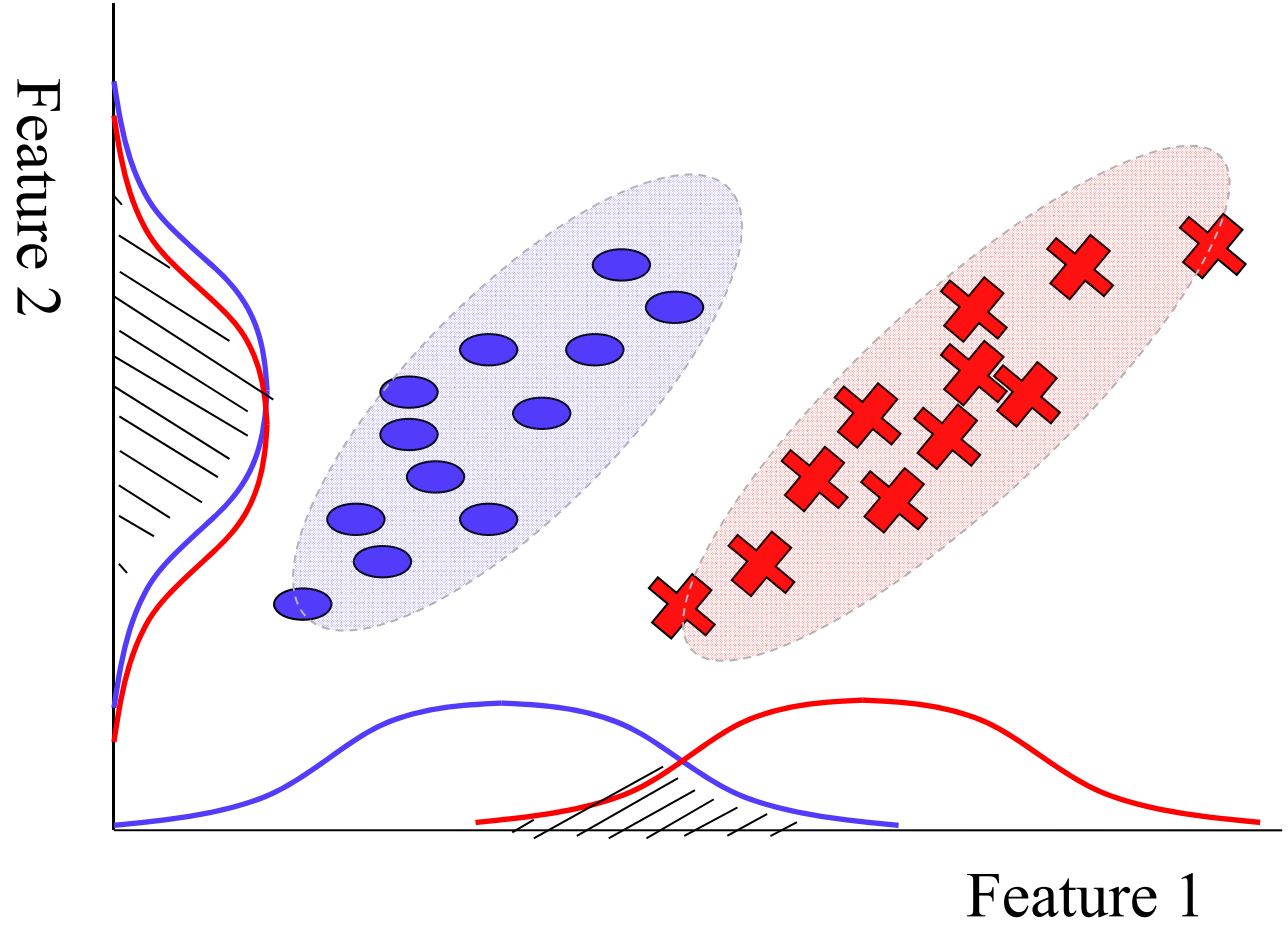


Marginal Probability Density and Conditional Probability Density in Machine Learning



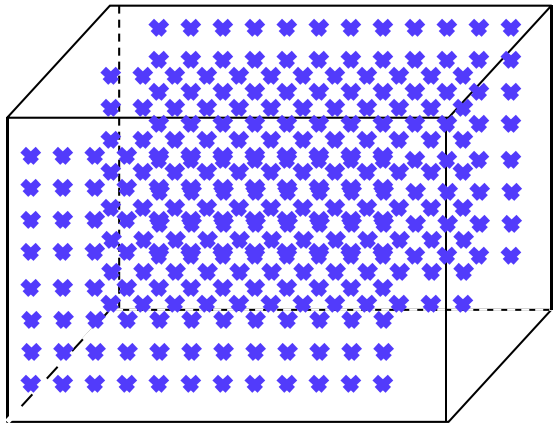
Benefits of Using High Dimensionalities

- Feature 1 and Feature 2 have correlation



Curse of Dimensionality

- To achieve same density as N = *100* for *1*-variable
- We need N = 100^D for *D* variables



- Conversely, when we have *60,000* data for *10*-dimensional space, the density is the same as *3* data in *1*-dimensional space.

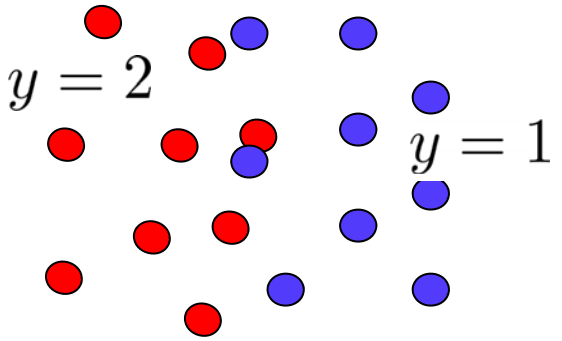
CONSISTENT LEARNER

Learning

- Data

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N} \sim P$$

(Regularity)



- Prediction

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

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

Quantify the Evaluation

- Measure of quality: expected loss

$$L = \mathbb{E}_P[l(y, f(\mathbf{x}))] \qquad l(y, y'): \text{loss function}$$

- Estimated error

$$\hat{L} = \sum_n l(y_n, f(\mathbf{x}_n)), \quad f(\mathbf{x}) \in \mathcal{H}$$

- Examples
  - Classification $\widehat{L} = \sum_n \mathbb{I}(y_n \neq f(\mathbf{x}_n))$
  - Regression $\widehat{L} = \sum_n ||y_n - f(\mathbf{x}_n)||^2$
  - Clustering $\widehat{L} = \sum_n \min_c ||y_n - f(\mathbf{x}_n)||^2$

Consistent Learner

- $\mathcal{H}$ satisfies $\widehat{L} \underset{N \to \infty}{\to} L$

$$P\{\sup_{f \in \mathcal{H}} (L(f) - \widehat{L}(f)) > \epsilon\} \to 0 \quad \text{for} \quad \epsilon > 0$$

<Uniform convergence>

- Caution:
  - The definition of consistency is *not*

$$\widehat{L}(f) \to L(f) \quad \text{for} \quad f \in \mathcal{H}$$

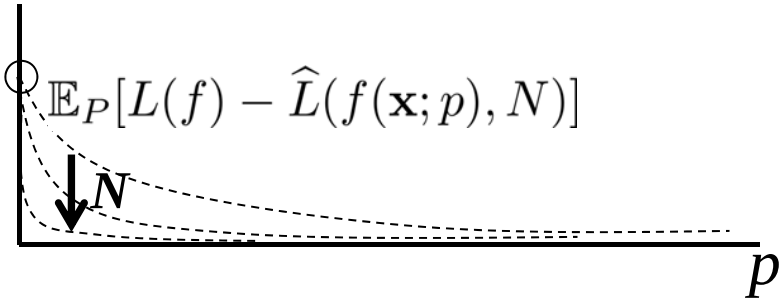
Quiz 1: Counter Example

- Consider a hypothesis set $\mathcal{H}$ which satisfies

$$\mathbb{E}_P[L(f) - \widehat{L}(f(\mathbf{x};p),N)] = \left(\frac{1}{N}\right)^p$$

$$\mathcal{H} = \{f(\mathbf{x};p)|p > 0\}$$

Explain that learning with $\mathcal{H}$ is *not* consistent even when it satisfies $\widehat{L}(f) \to L(f)$.



What is the possible problem in this case?

Solution 1:

- For any fixed large *N*, there are functions where the expected loss and the estimated error still show severe difference.

- The number of data *N* does *not* exist that guarantees the expected error of selected function is greater than the expected error of the optimal function with the amount less than a small number $\epsilon > 0$.

Motivation for More Assumptions

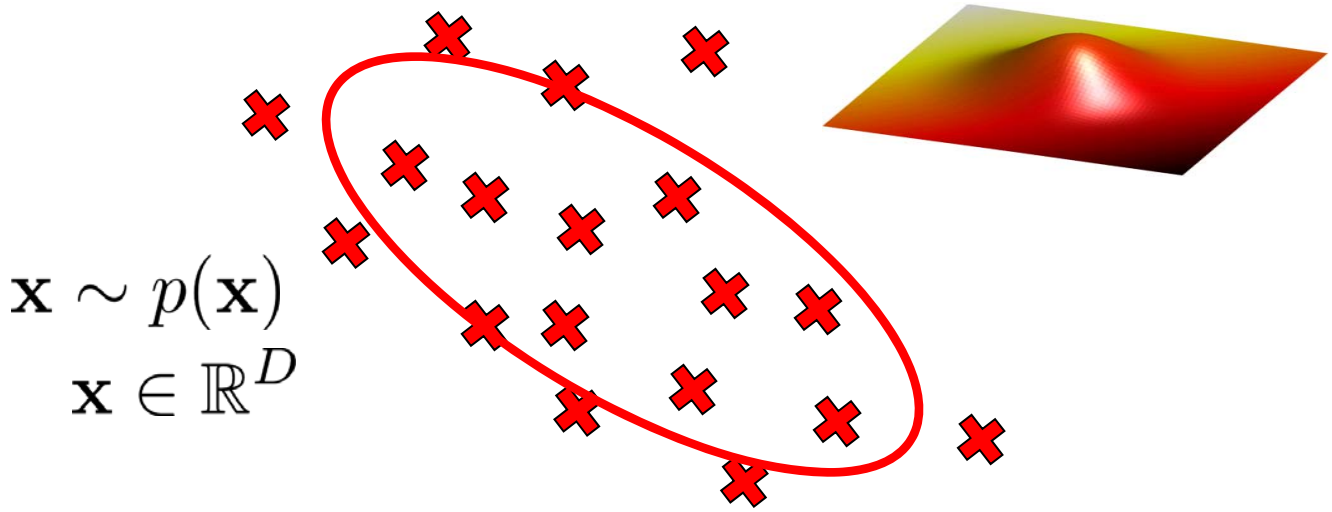
- If we can have sufficient (or infinite) data
  - A consistent learner will produce an optimal classifier with high probability by reducing the empirical error.
  - An arbitrarily complex $f(\mathbf{x})$ will be chosen producing good classification performance for testing data.

  - ***One serious problem***: In real situation, we cannot have enough data. → We need a hypothesis set $\mathcal{H}$ which is more applicable to the *finite sampling situation*.

PARAMETER ESTIMATION

Motivation – Parameter Estimation

- Parameter estimation is an optimization problem

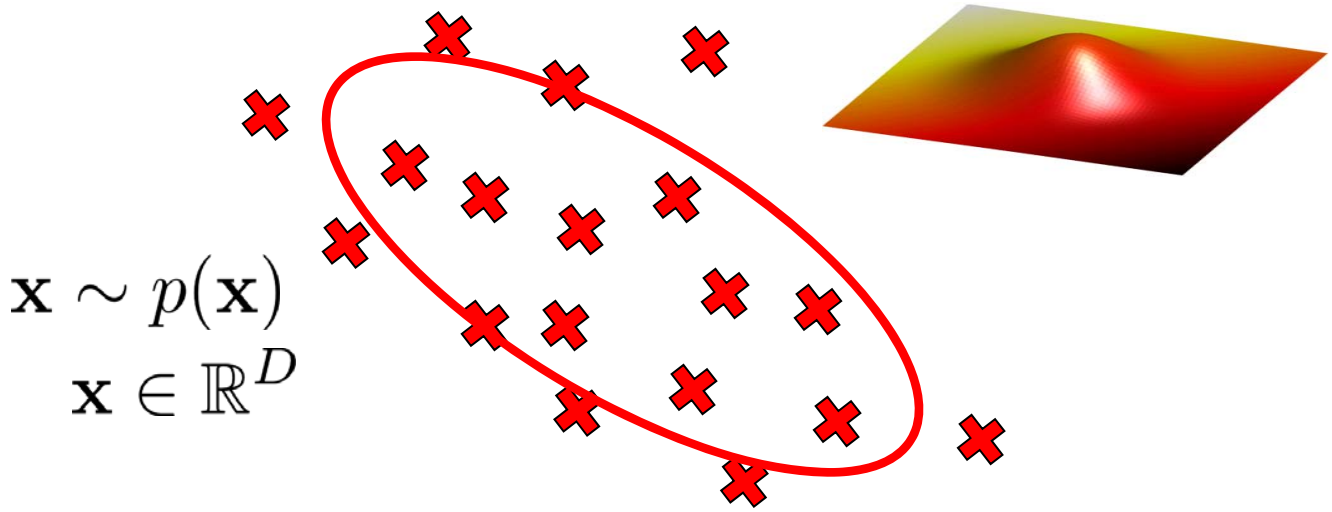


$$\mathbf{x} \sim p(\mathbf{x})$$
$$\mathbf{x} \in \mathbb{R}^D$$

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

Maximum Likelihood Estimation

- Parameter estimation is an optimization problem



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

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

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

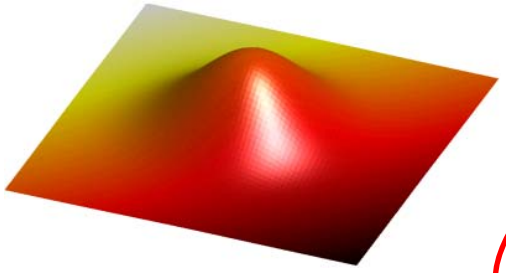
Gaussian Model

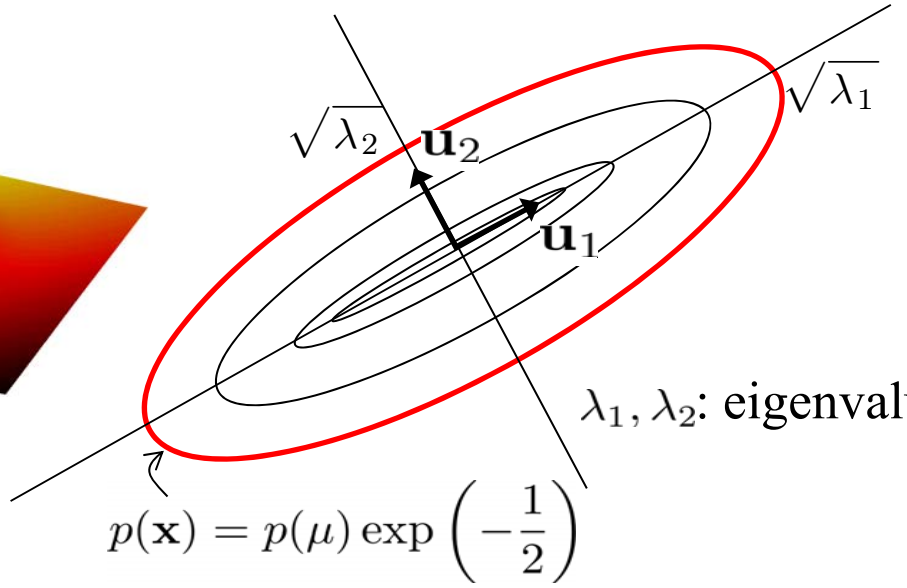
$$p(\mathbf{x}|\mu,\Sigma) = \frac{1}{\sqrt{2\pi}^{D}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \in \mathbb{R}^D$$

Principal axes are the eigenvector directions of Σ

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





Maximum Likelihood for Gaussian

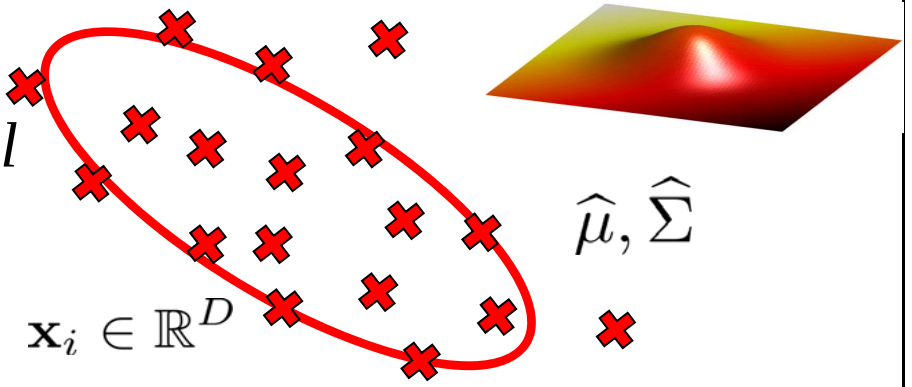
$$p(\mathbf{x}|\mu,\Sigma) = \frac{1}{\sqrt{2\pi}^{D}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

- 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

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

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

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

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

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

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

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

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

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

Maximum A Posteriori (MAP) Estimation for Gaussian

- Posterior density

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

  - 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

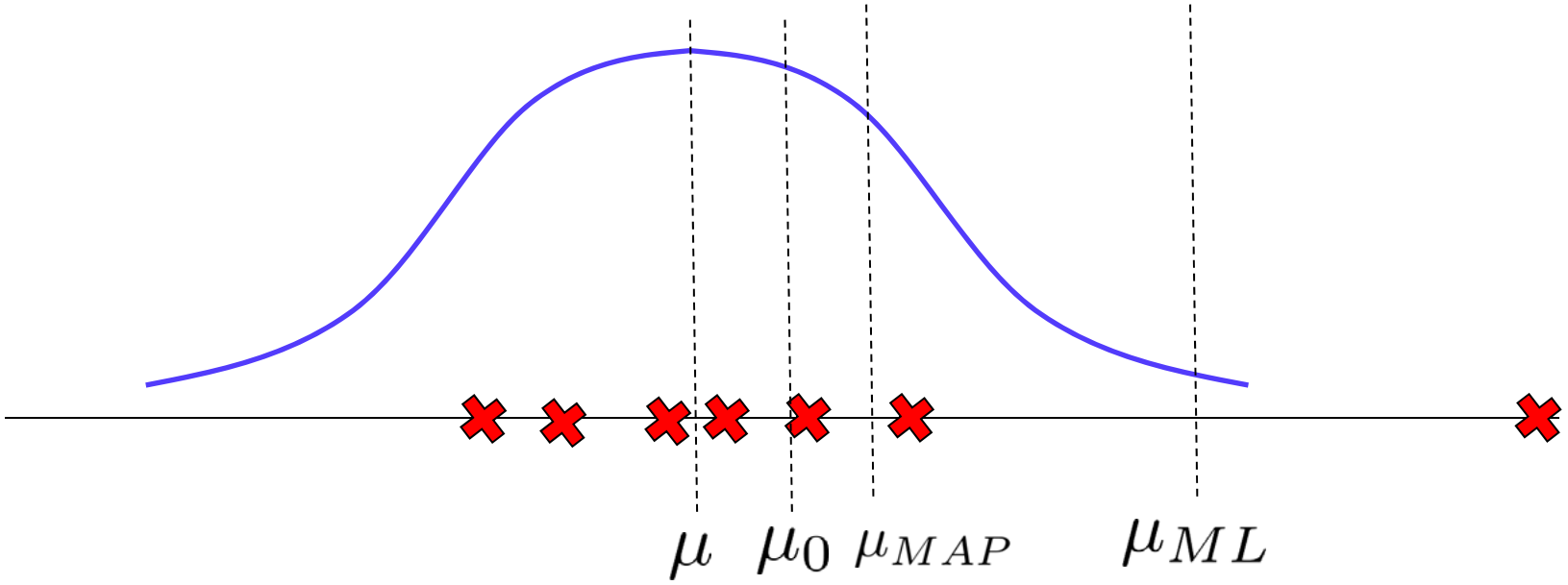
- For Gaussian
  - When we have a few outliers

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

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

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

MLE vs. MAP



Bayesian Integration

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

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

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

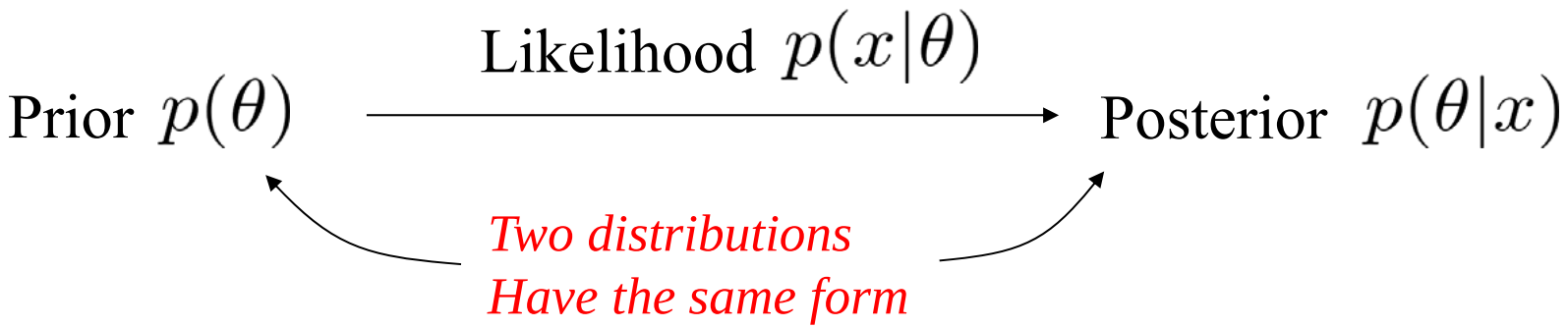
$$= \frac{1}{\sqrt{2\pi(\sigma^2+\sigma_n^2)}} \exp\left(-\frac{1}{2(\sigma^2+\sigma_n^2)}(x-\mu)^2\right)$$

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

Uncertainty of μ

Conjugate Priors

- Given a likelihood pdf, $p(x|\theta)$, posterior $p(\theta|x)$ has the same form as the prior $p(\theta)$.



Conjugate Priors

$$\text{Gaussian} \xrightarrow{\text{Gaussian } (\mu)} \text{Gaussian}(\mu)$$

$$\text{Dirichlet} \xrightarrow{\text{Multinomial}} \text{Dirichlet}$$

$$\text{Beta} \xrightarrow{\text{Binomial}} \text{Beta}$$

$$\text{Gamma} \xrightarrow{\text{Poisson}} \text{Gamma}$$

$$\text{Gamma } (\lambda = \frac{1}{\sigma^2}) \xrightarrow{\text{Gaussian}(\sigma^2)} \text{Gamma } (\lambda = \frac{1}{\sigma^2})$$

Kullback-Leibler Divergence

$$KL(p_e||p_\theta) = -\int p_e \log \frac{p_\theta}{p_e} d\mathbf{x}$$

p_e: Empirical density function

p_θ: Model density function

$$= -\int \left[p_e \log p_\theta - p_e \log p_e\right] d\mathbf{x}$$

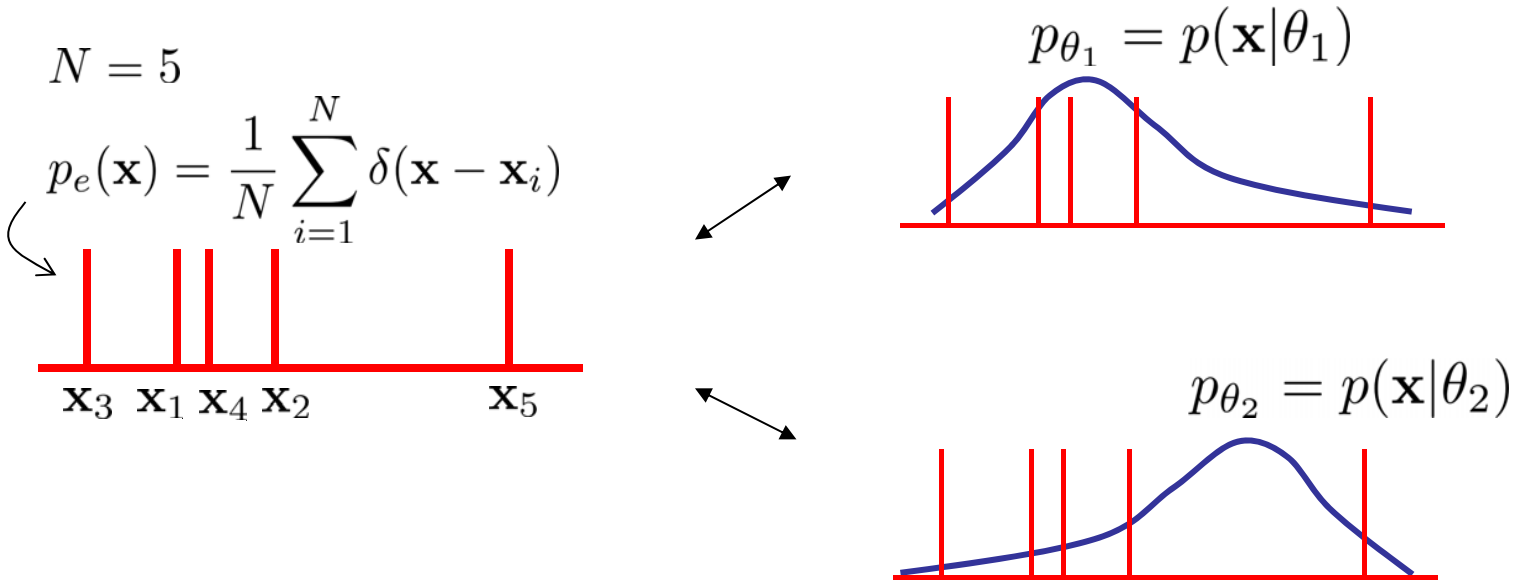
$$p_e = \frac{1}{N}\sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i)$$

$$\arg\min_{p_\theta} KL(p_e||p_\theta) = \arg\min_{p_\theta} -\int \frac{1}{N}\sum_{i=1}^{N} \delta(\mathbf{x}-\mathbf{x}_i) \log p_\theta(\mathbf{x})\, d\mathbf{x}$$

$$= \arg\max_{p_\theta} \frac{1}{N}\sum_{i=1}^{N} \log p_\theta(\mathbf{x}_i)$$

$$= \arg\max_{p_\theta} \log \prod_{i=1}^{N} p_\theta(\mathbf{x}_i) = \arg\max_{p_\theta} p(\mathcal{D}|\theta)$$

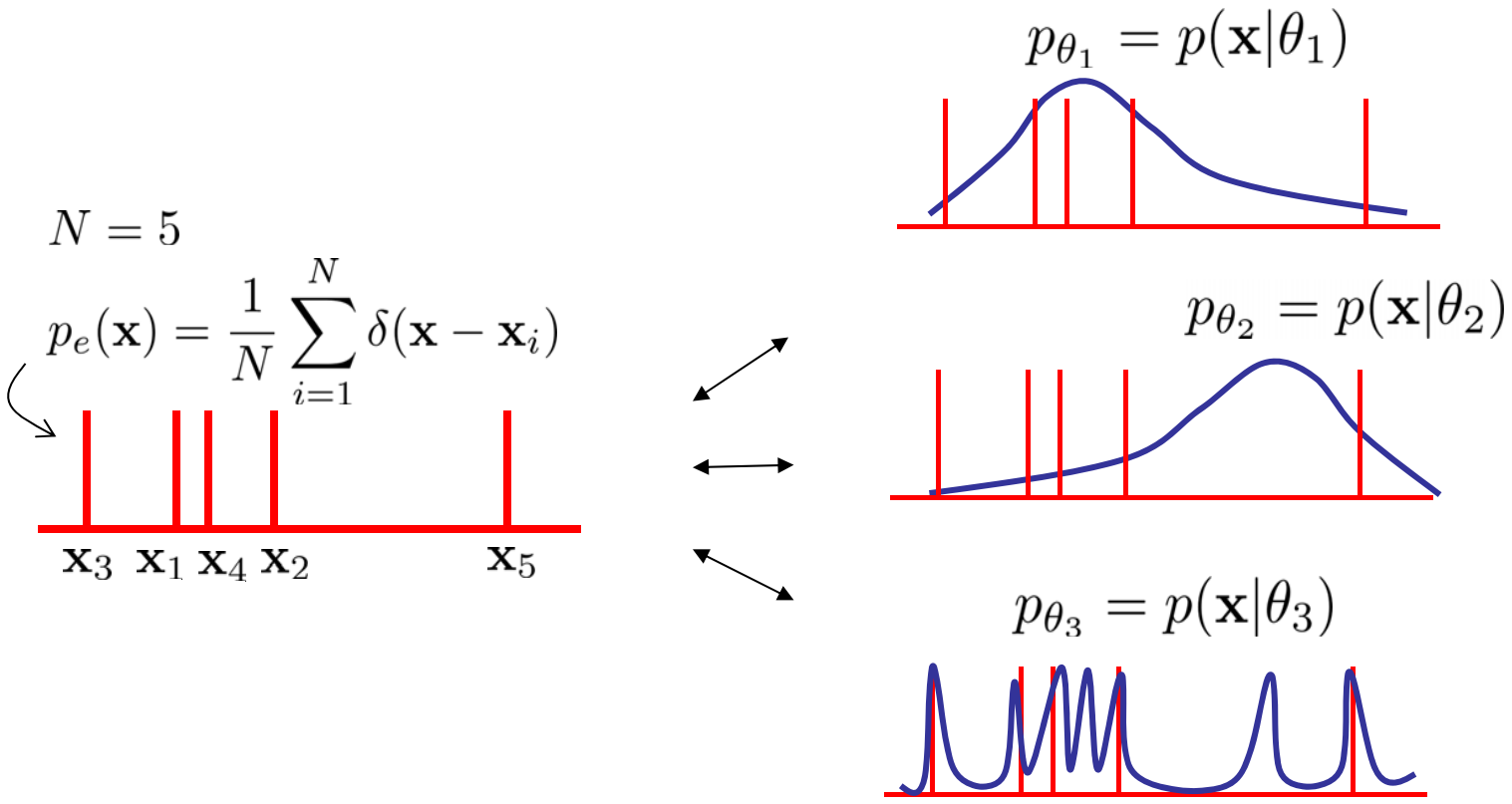
Kullback-Leibler Divergence



KL Divergence: $KL(p_e||p_{\theta_1}) < KL(p_e||p_{\theta_2})$

Likelihood: $p(\mathcal{D}|\theta_1) > p(\mathcal{D}|\theta_2)$

Kullback-Leibler Divergence

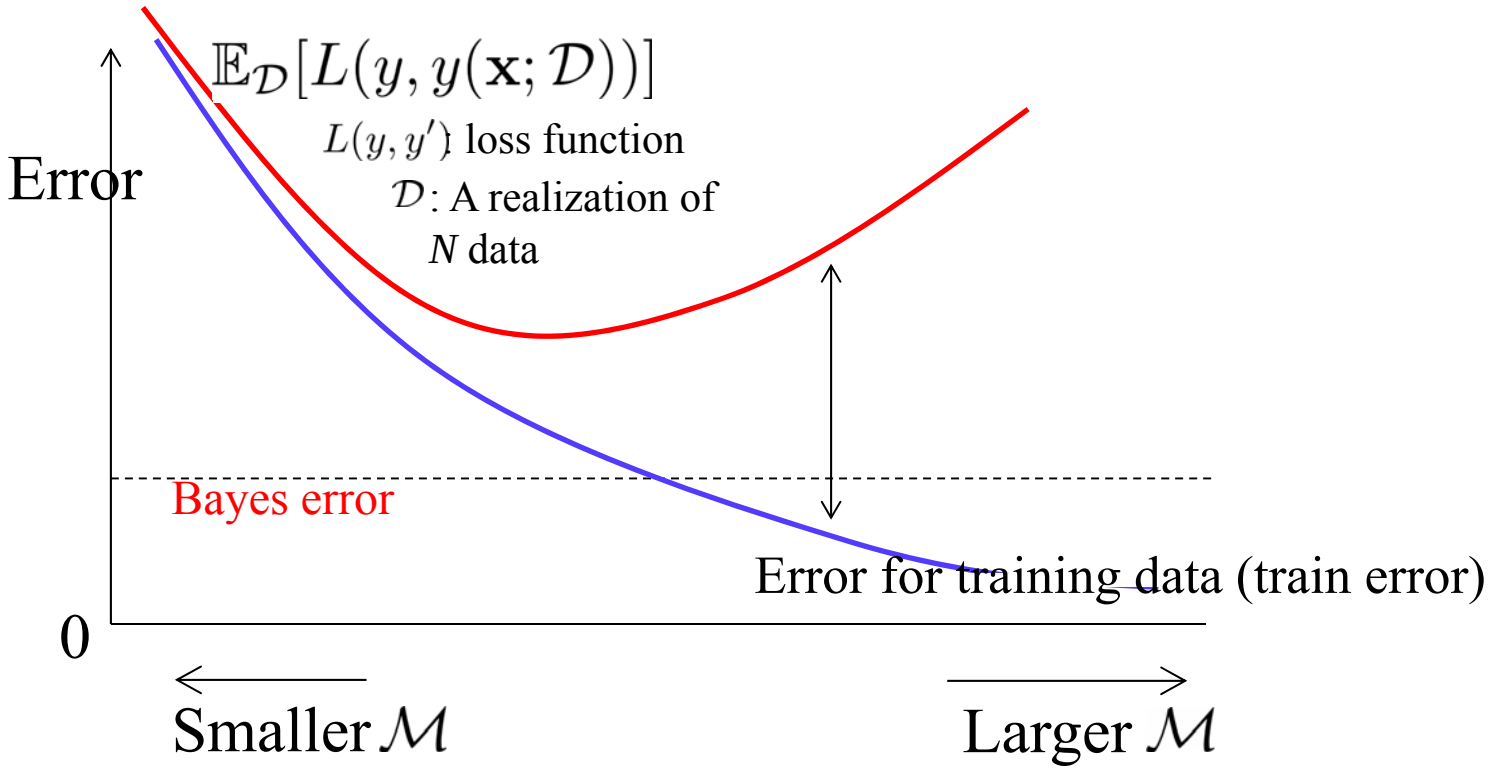


$$\theta_3 = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

Model with complex function will capture the noise.

Consistency and Bayes Error

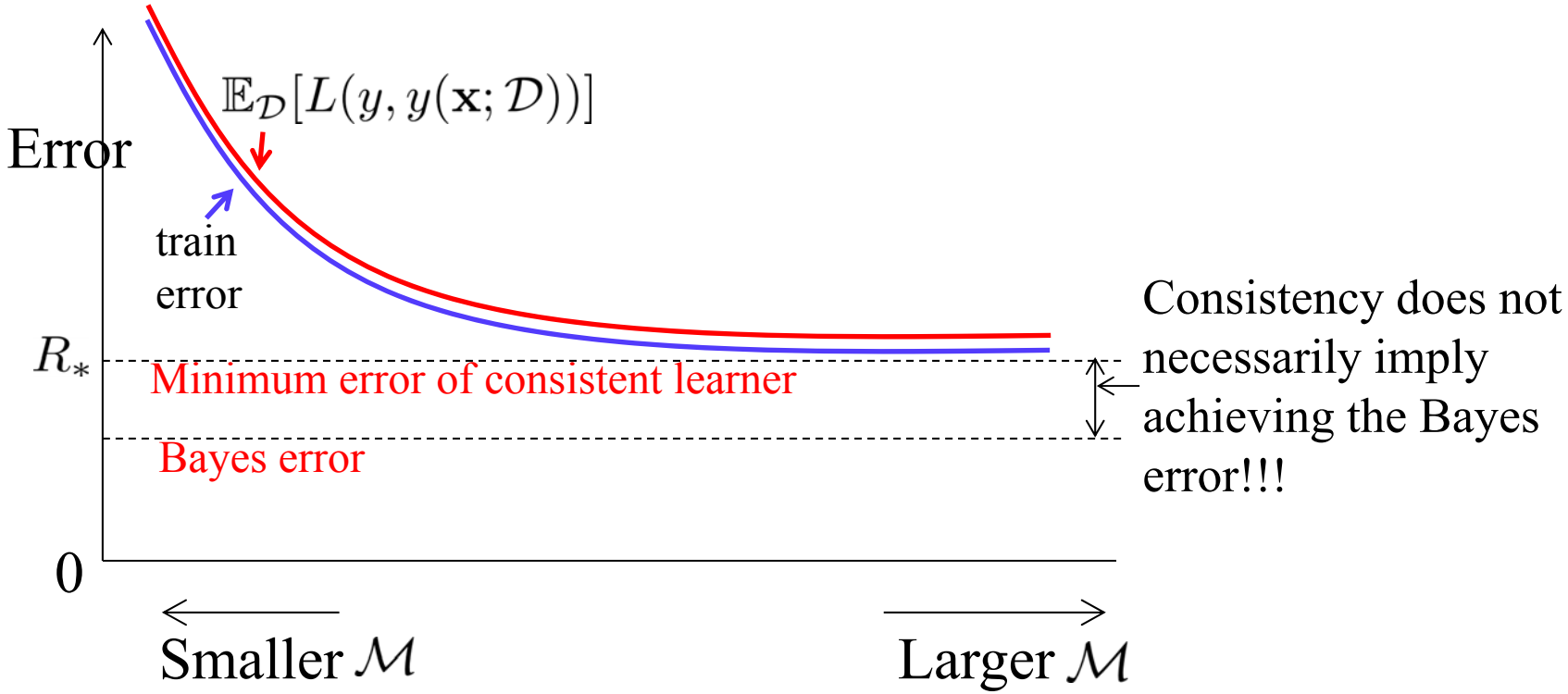
- Objective: minimizing expected error



(For example, a linear classifier with regularization)

Consistency and Bayes Error

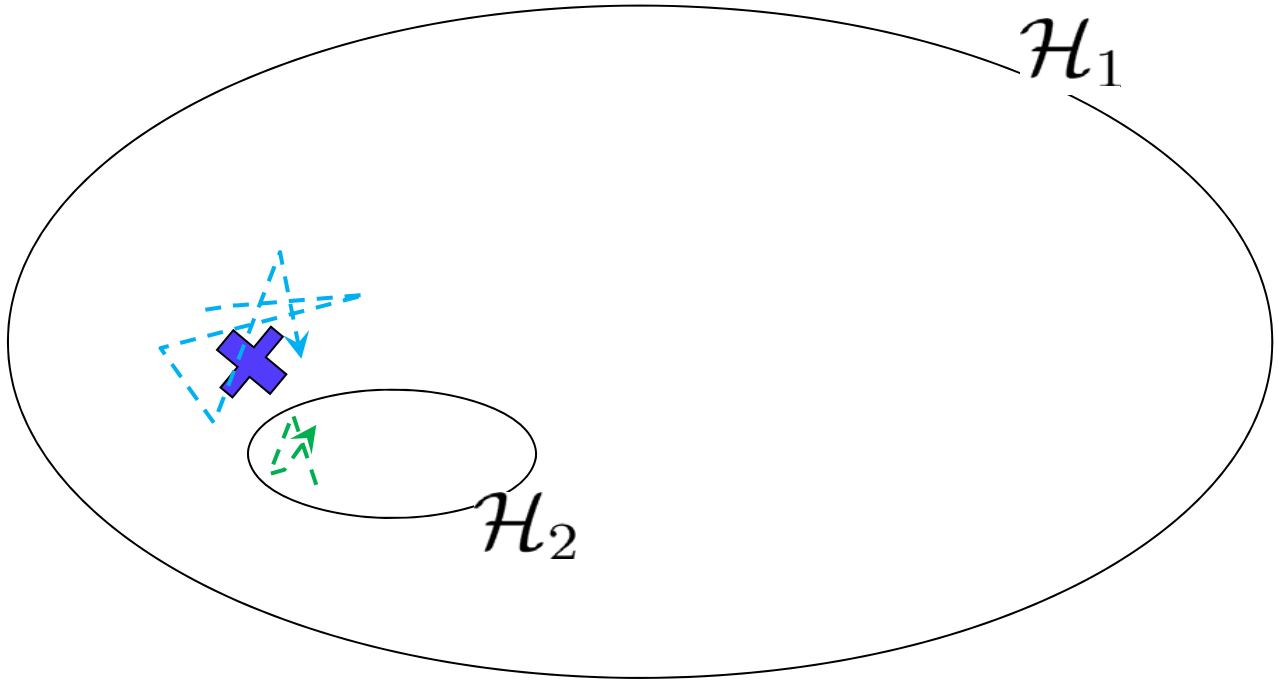
- Consistent learner with many data



(For example, a linear classifier with regularization)

Confining $\mathcal{H}$

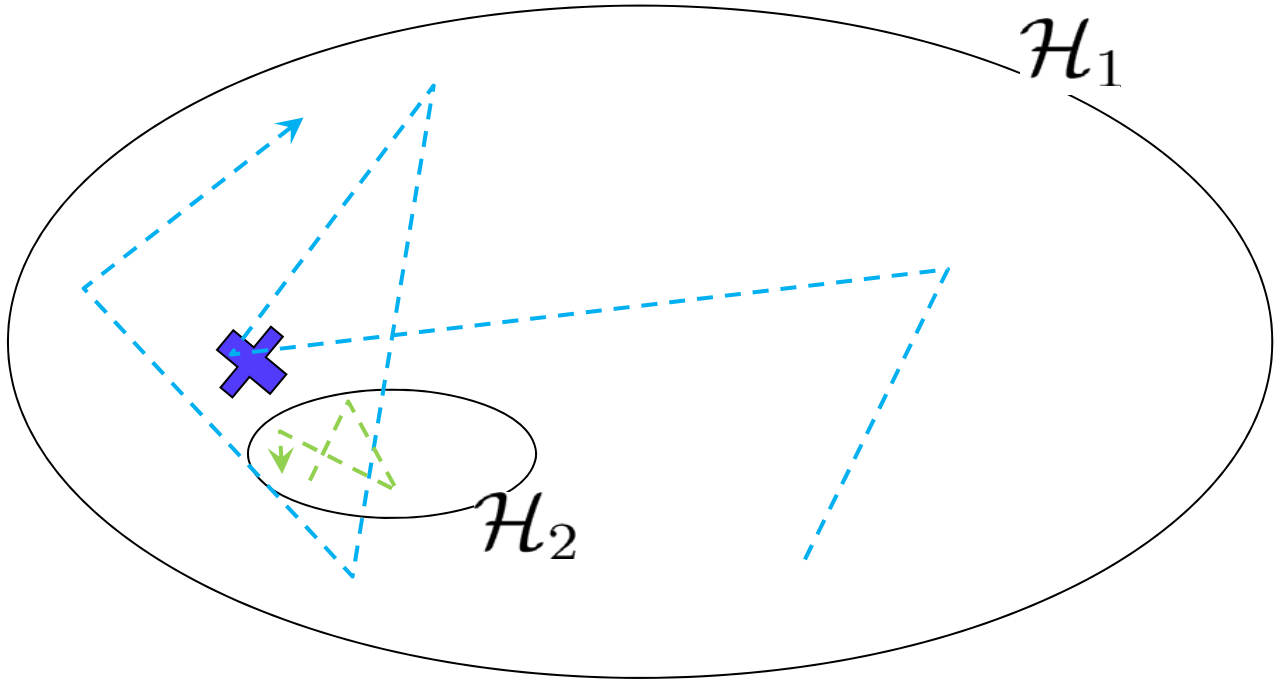
- Estimation with large number of data



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

Confining $\mathcal{H}$

- Estimation with small number of data



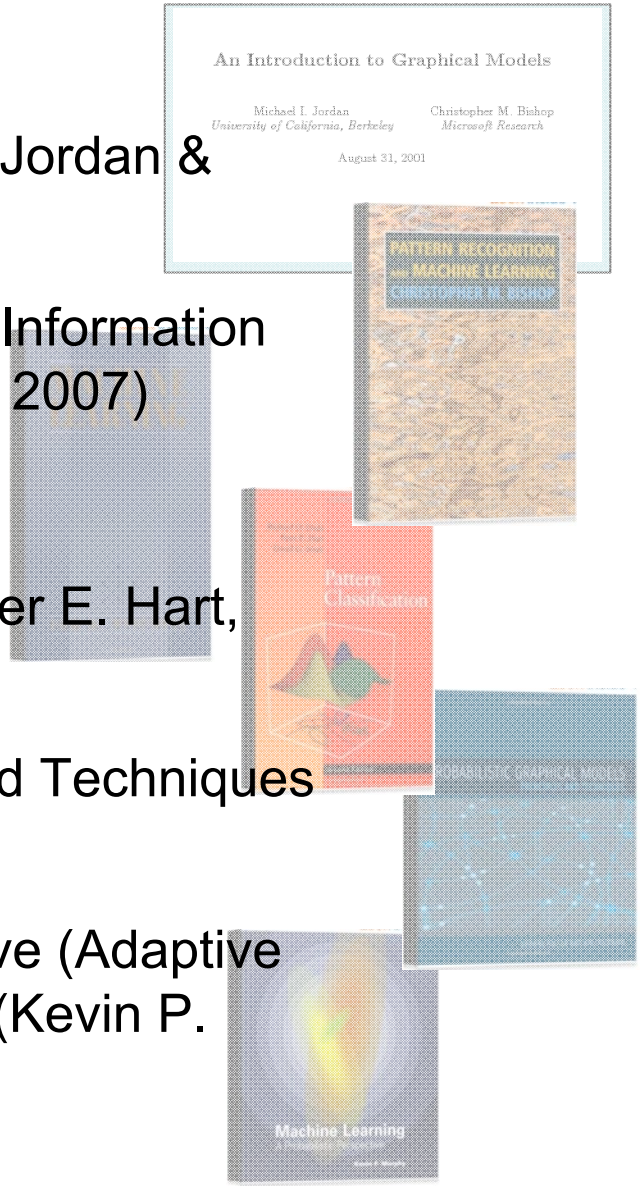
Optimal solution (✖) and the selected solutions of different realizations

Summary

- What we did:
  - Probability and probability density
  - Conditional density, marginal density
  - Model construction
  - Parameter estimation
- What we didn't do:
  - Convergence of estimation
  - Graphical models
  - Inference

Books

- Introduction to Graphical Models (Michael I. Jordan & Christopher Bishop), unpublished
- Pattern Recognition and Machine Learning (Information Science and Statistics) (Christopher Bishop, 2007)
- Machine Learning (Tom M. Mitchell, 1997)
- Pattern Classification (Richard O. Duda, Peter E. Hart, David G. Stork, 2000)
- Probabilistic Graphical Models Principles and Techniques (Daphne Koller, Nir Friedman, 2009)
- Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series) (Kevin P. Murphy, 2012)



THANK YOU

Yung-Kyun Noh
nohyung@snu.ac.kr