PRML을 위한 기초 확률 이론
(Basic Probability Theory for PRML)

패턴인식 및 기계학습 겨울학교
(Pattern Recognition and Machine Learning Winter School)
2016. 1. 20 (Wed.)

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

Contents

- Probability / Probability density
- Conditional probability (density)

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

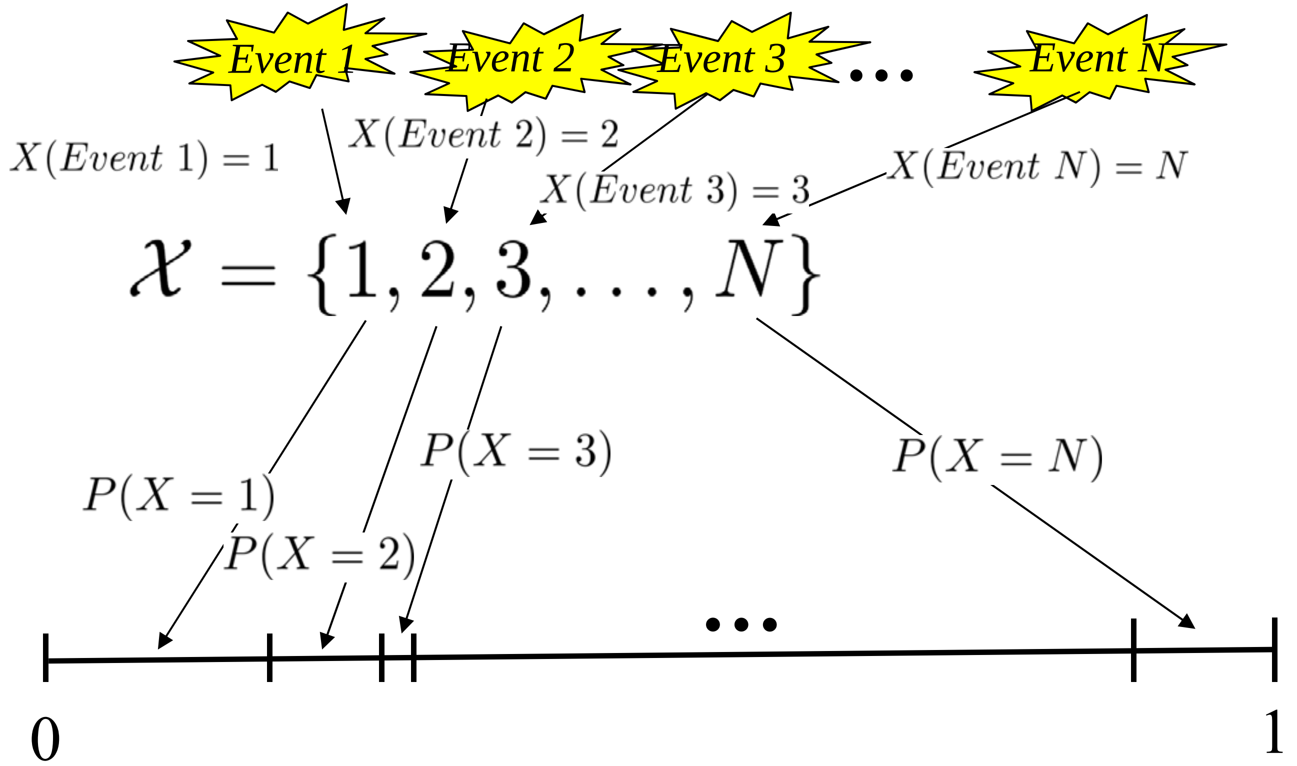
$$\mathbf{x}_1 \in \mathbb{R}^{D_1}, \mathbf{x}_2 \in \mathbb{R}^{D_2}, \mathbf{x} \in \mathbb{R}^{D}, y \in \{1, 2\}$$

- Marginal probability (density)
- Joint probability (density)
- Inference and classification
- Gaussian Processes

Probability

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

- Mapping from a random variable to a number



Probability

X: random variable $\quad X_1$: set of outputs of random variables

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

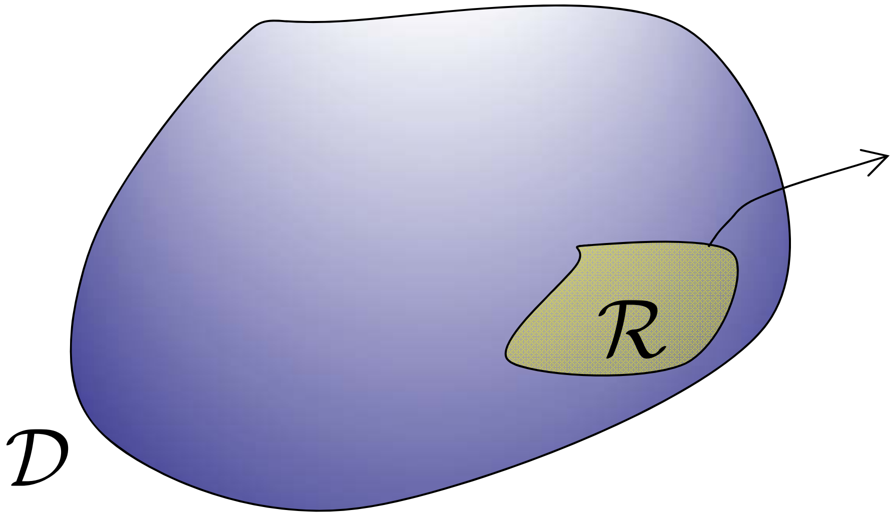
$$P(X_1 \cup X_2) = P(X_1) + P(X_2) - P(X_1 \cap X_2)$$

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

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

Probability and Probability Density

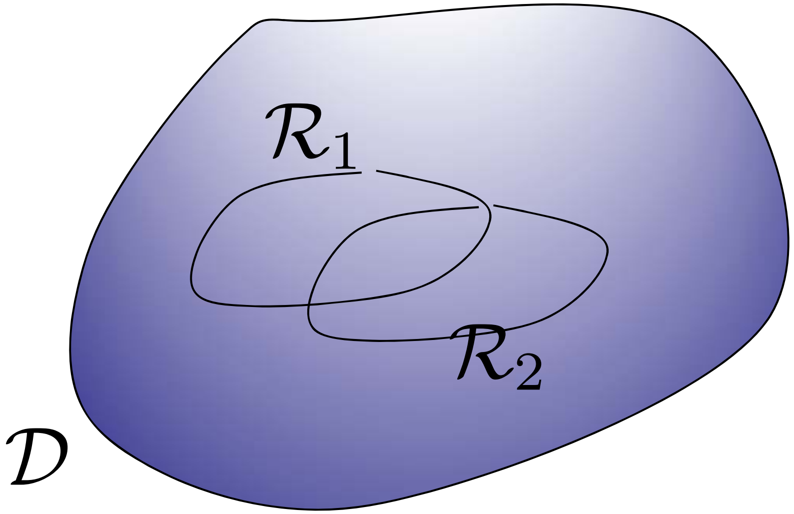
$$p(\mathbf{x}) \in \mathbb{P} \qquad \int_{\mathcal{D}} p(\mathbf{x})d\mathbf{x} = 1 \qquad p(\mathbf{x}) \geq 0$$



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

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

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?

$$P(X = 1)$$

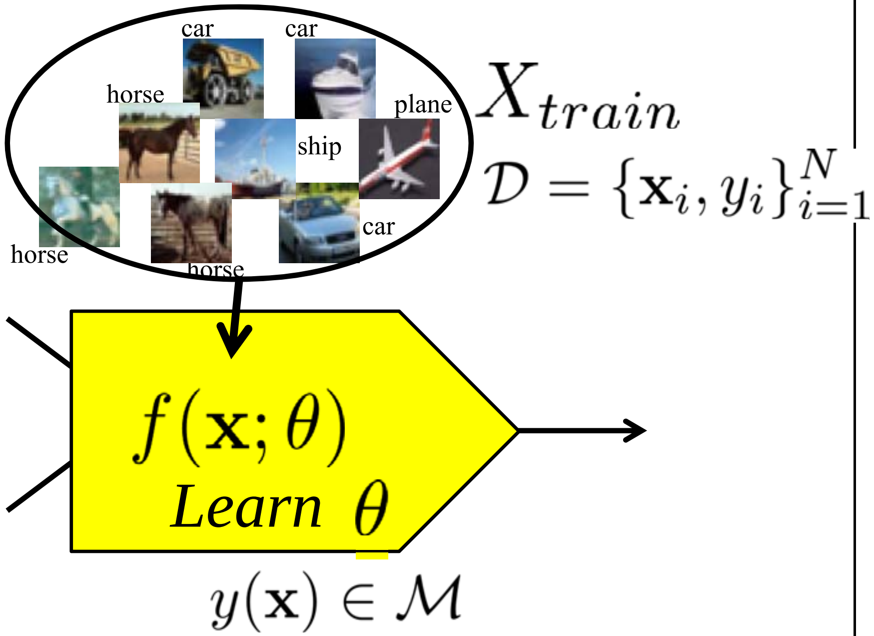
$$P(X = 1 | Y = 2)$$

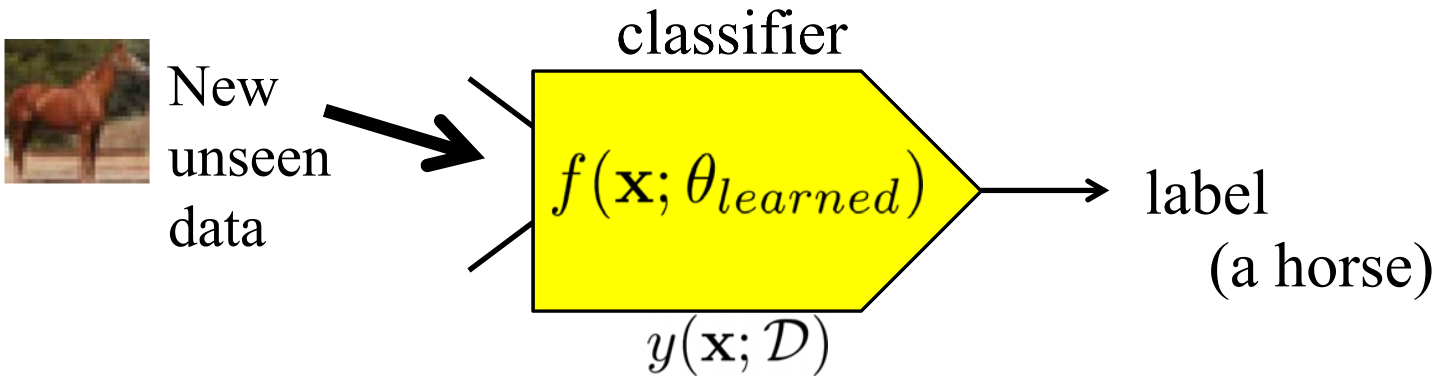
$$p(x = 1)$$ Compare with $P(x = 1)$?

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

Supervised Learning (Prediction)

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

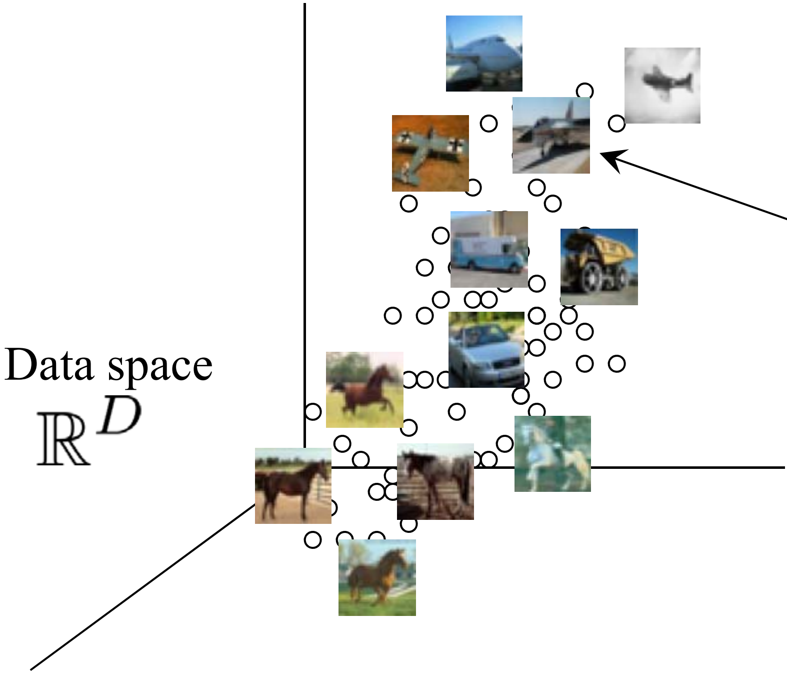


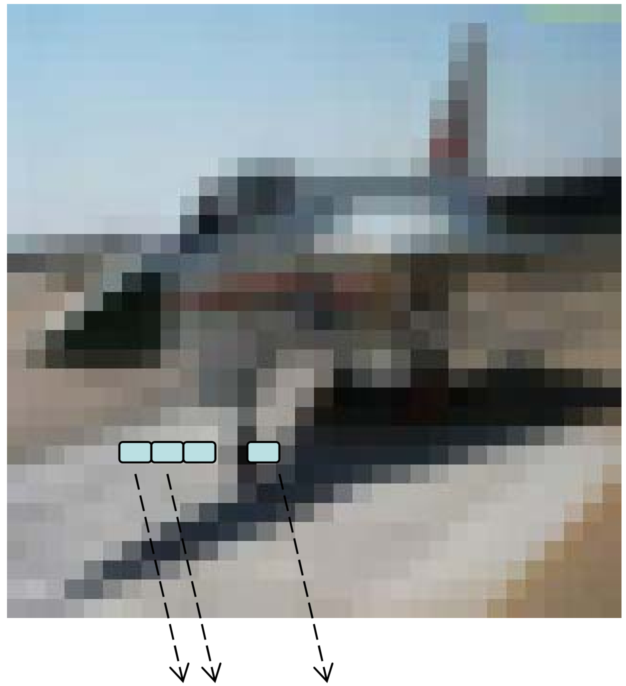


[Based on the assumption of regularity]

Representation of Data

Data space $\mathbb{R}^D$





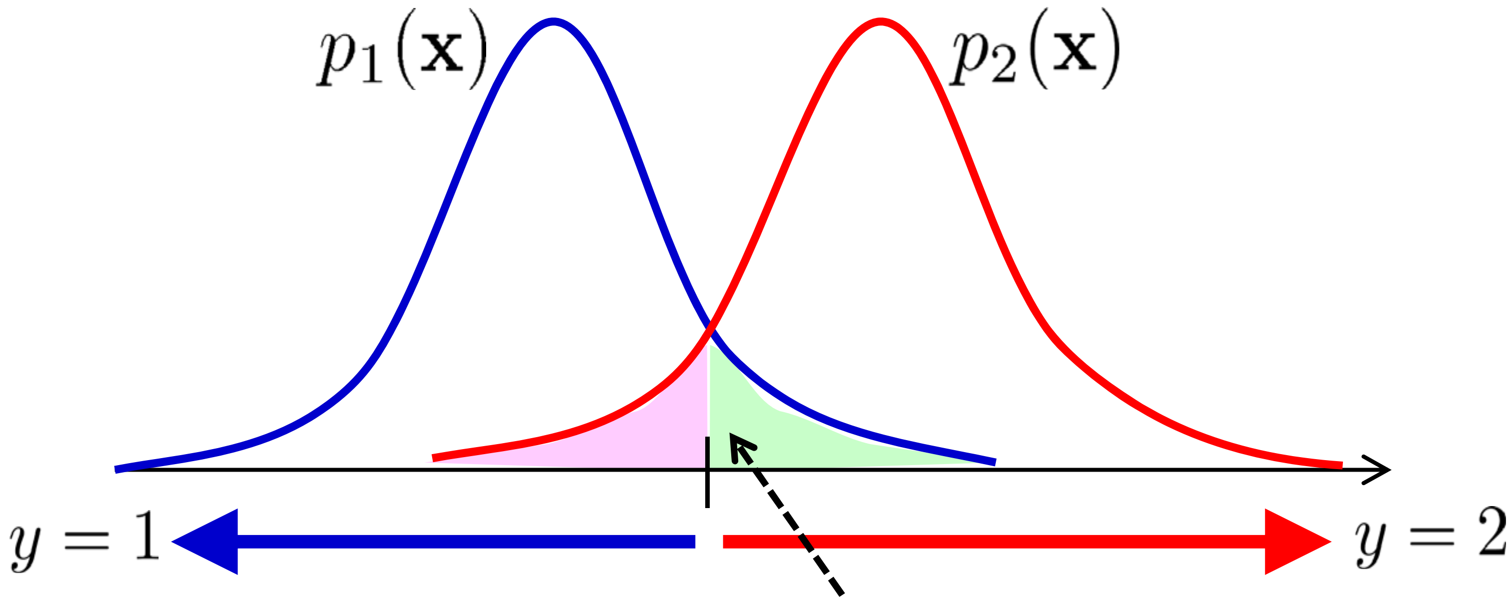
$$=\underbrace{[1, 2, 5, 10, \ldots]^T}_{D \text{ elements}}$$

- Each datum is one point in a data space

Classification

Bayes Optimal Classifier

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

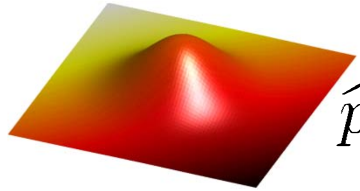


(Optimal) Bayes error

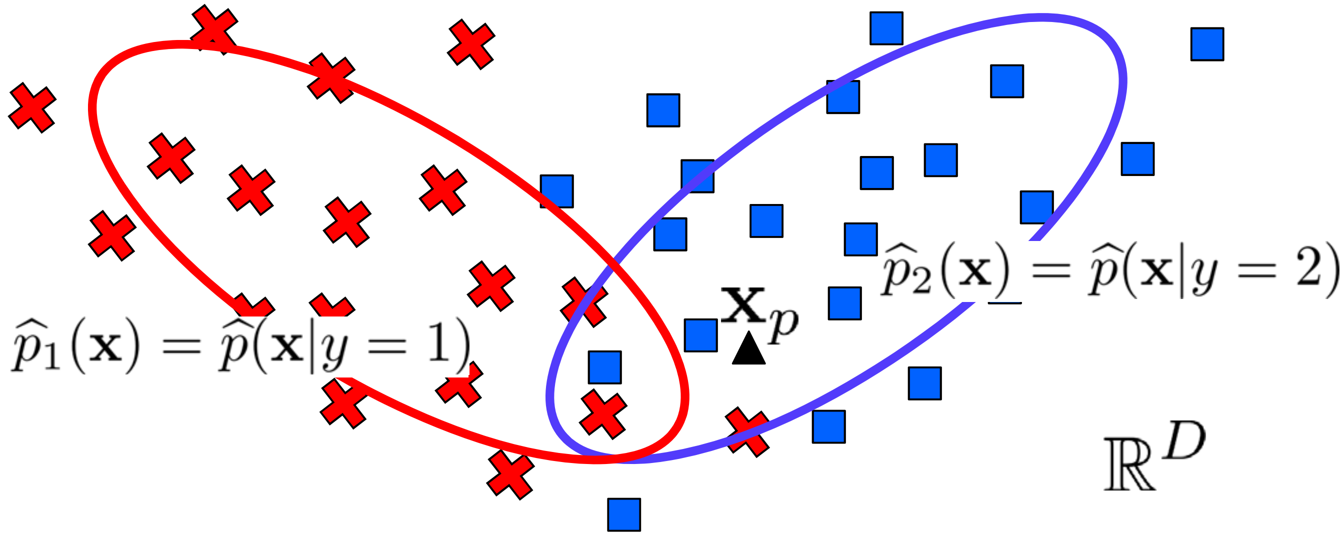
$$E_{Bayes} = \frac{1}{2} \int \min[p_1(\mathbf{x}), p_2(\mathbf{x})] d\mathbf{x}$$

Figure credit: Masashi Sugiyama

Model on Each Class



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



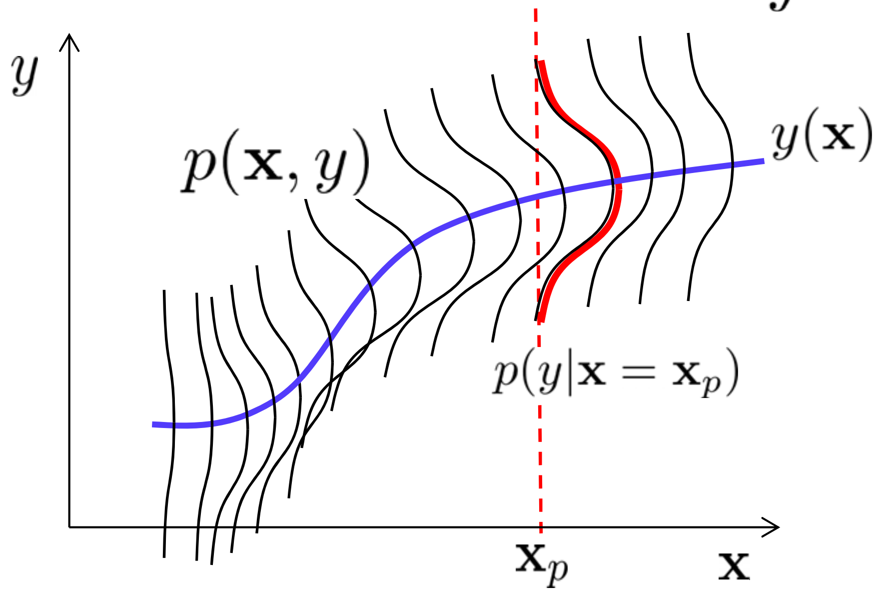
$$\widehat{p}_1(\mathbf{x}_p) \geq \widehat{p}_2(\mathbf{x}_p) \rightarrow y_p = 1$$
$$\widehat{p}_1(\mathbf{x}_p) < \widehat{p}_2(\mathbf{x}_p) \rightarrow y_p = 2$$

- Model: Class-conditional density as a Gaussian

Optimal Regression

- Minimizing mean square error

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



Minimize

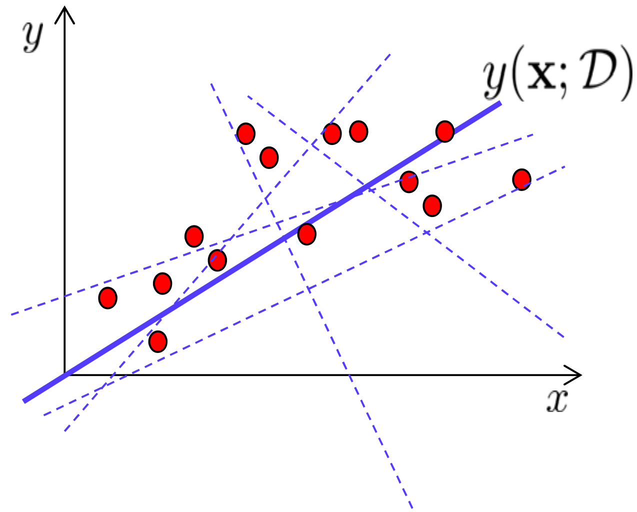
$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - y\}^2\, p(\mathbf{x}, y)d\mathbf{x}dy$$

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

→ 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 $y(\mathbf{x};\mathcal{D}) = \mathbb{E}\left[y|\mathbf{x}\right]$

- Bias-Variance tradeoff

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

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

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:

$Y \backslash X$	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

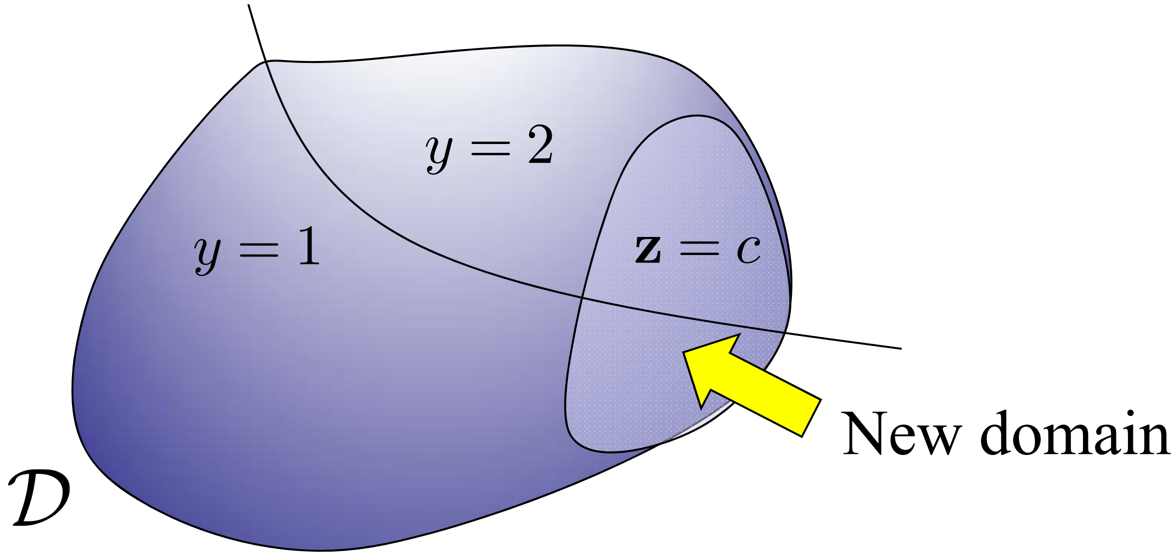
$Y \backslash X$	X_1	X_2	X_3	
Y_1	$P(X_1, Y_1)$	$P(X_2, Y_1)$	$P(X_3, Y_1)$	$P(Y_1)$
Y_2	$P(X_1, Y_2)$	$P(X_2, Y_2)$	$P(X_3, Y_2)$	$P(Y_2)$
Y_3	$P(X_1, Y_3)$	$P(X_2, Y_3)$	$P(X_3, Y_3)$	$P(Y_3)$
	$P(X_1)$	$P(X_2)$	$P(X_3)$	1

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

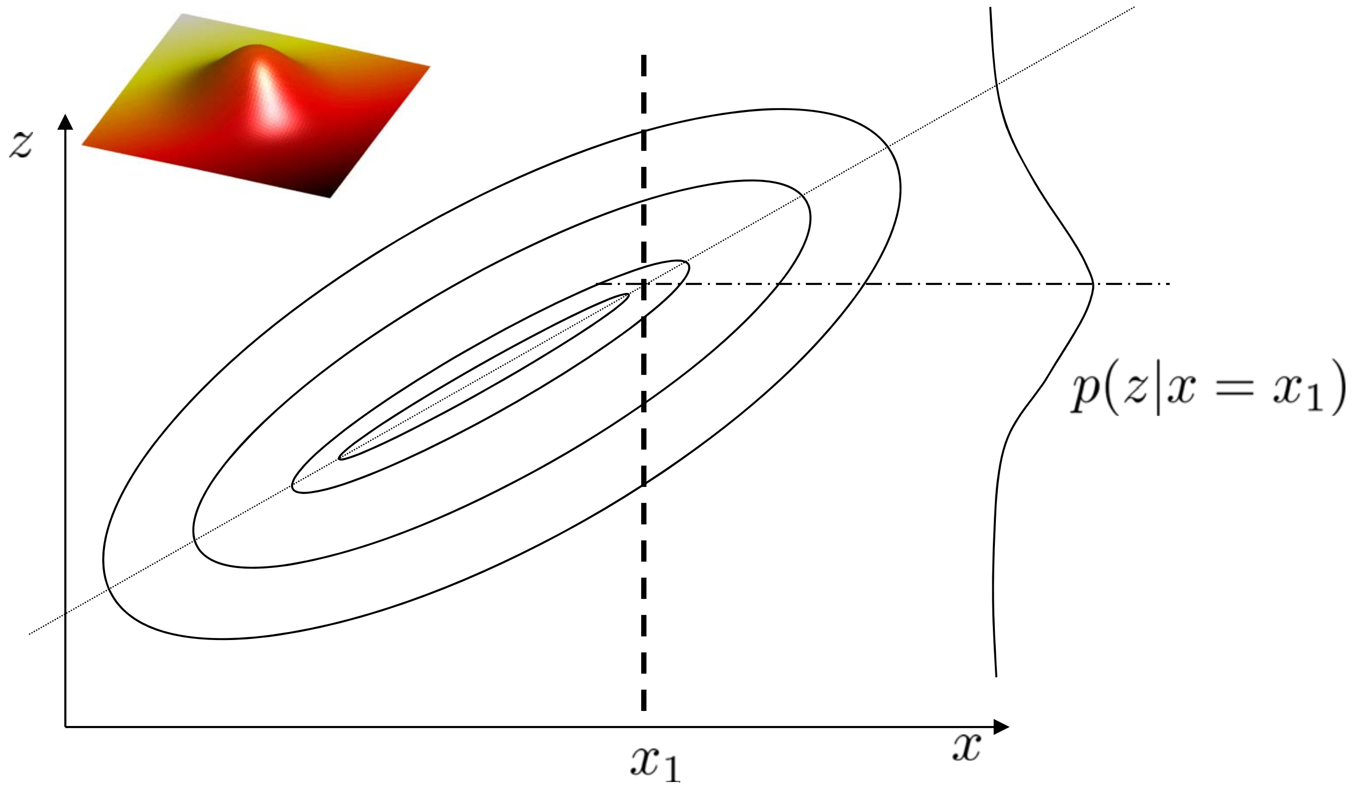
Conditional Probability Density

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

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

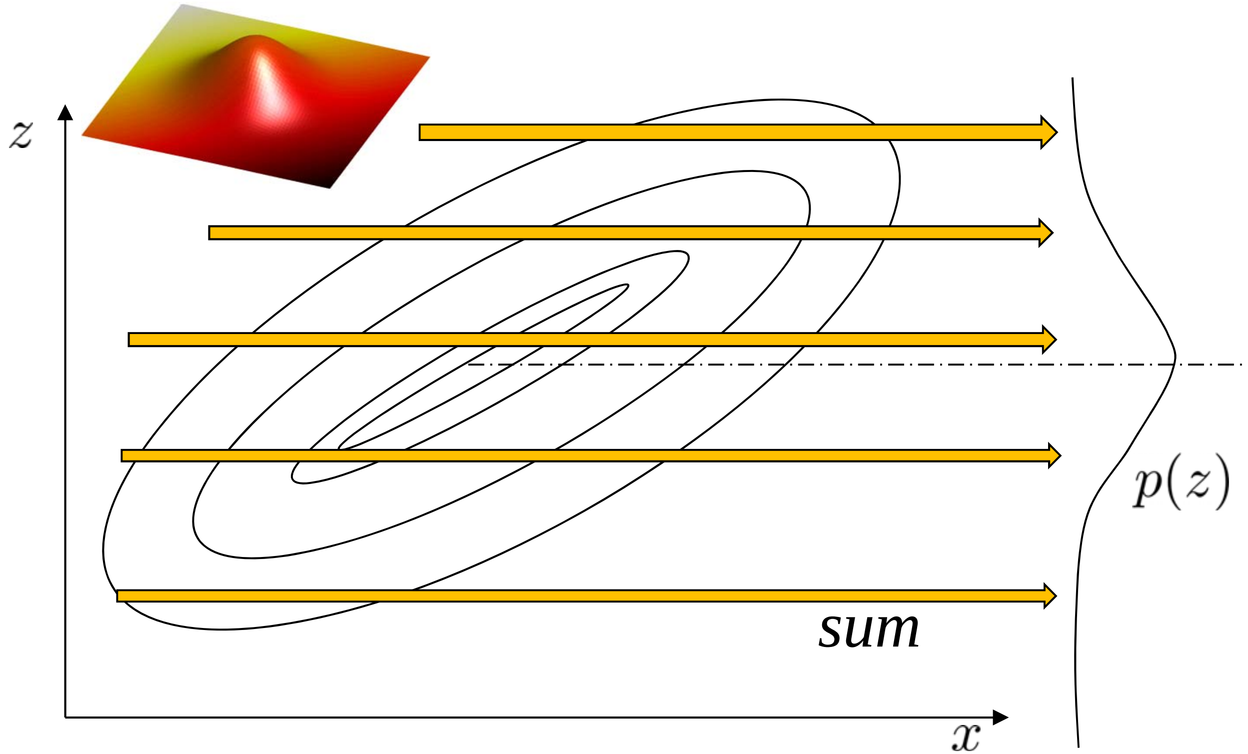


Conditional Probability Density



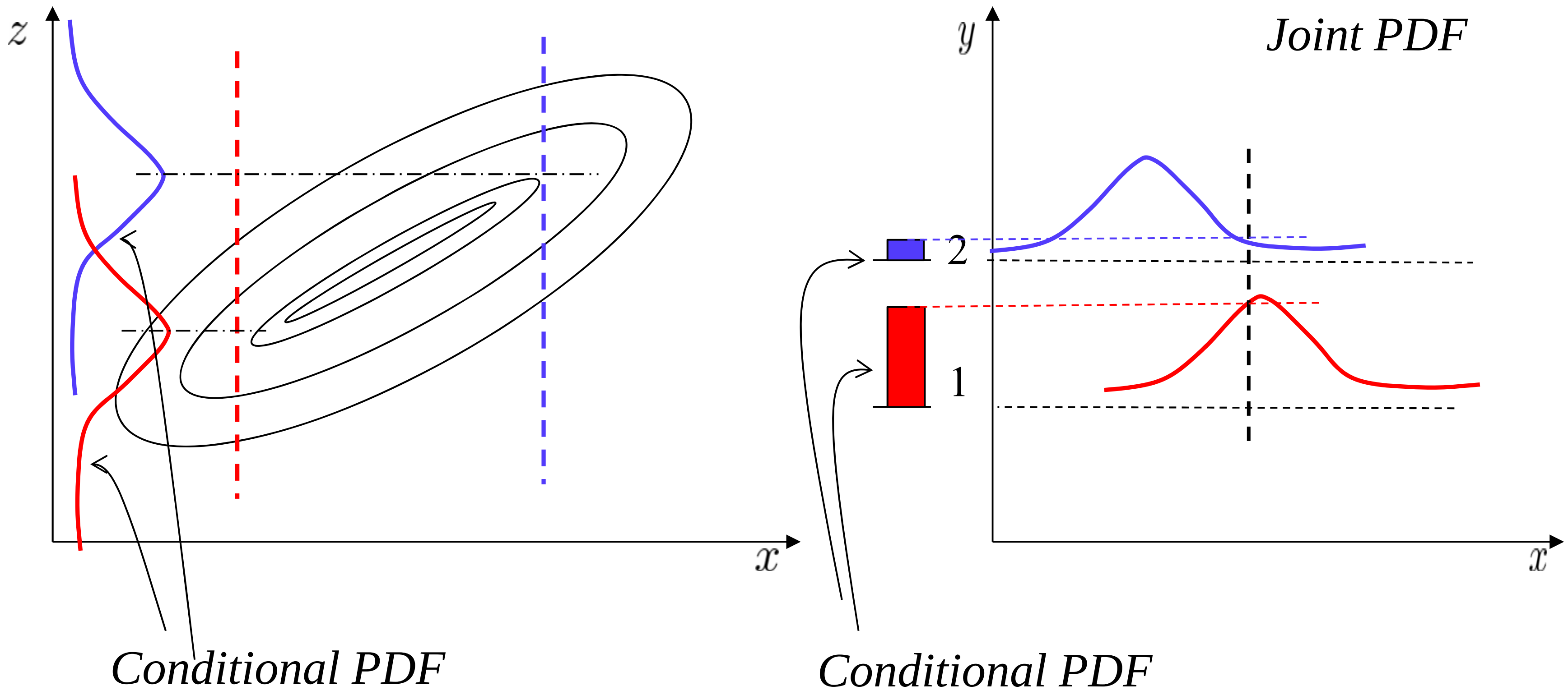
$$p(z|x = x_1) = \frac{p(z, x = x_1)}{\int_{x=x_1} p(z, x) dz} = \frac{p(z, x)}{p(x)}$$

Marginal Probability Density

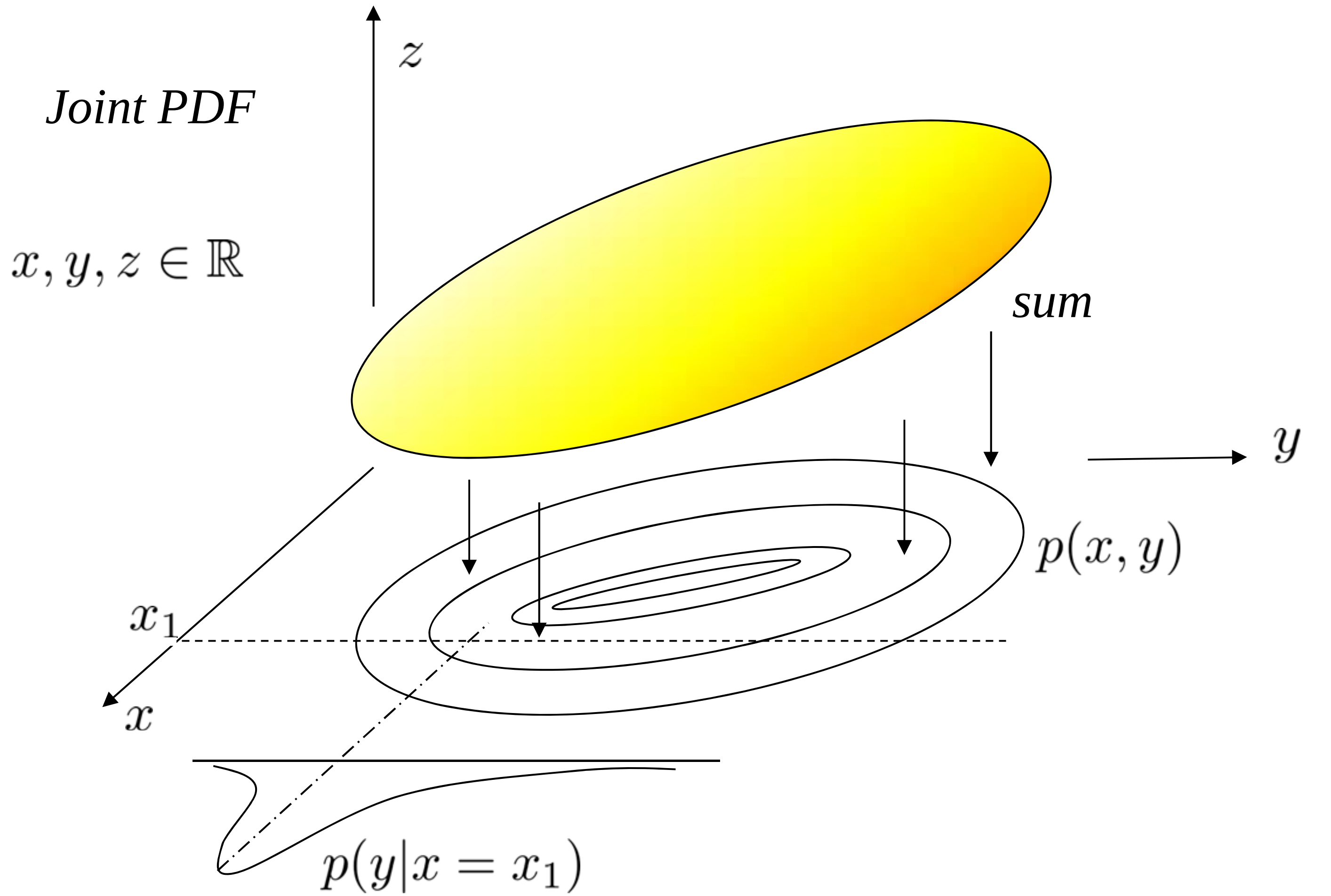


$$p(z) = \int p(z, x) dx$$

Marginal Probability Density and Conditional Probability Density in Machine Learning

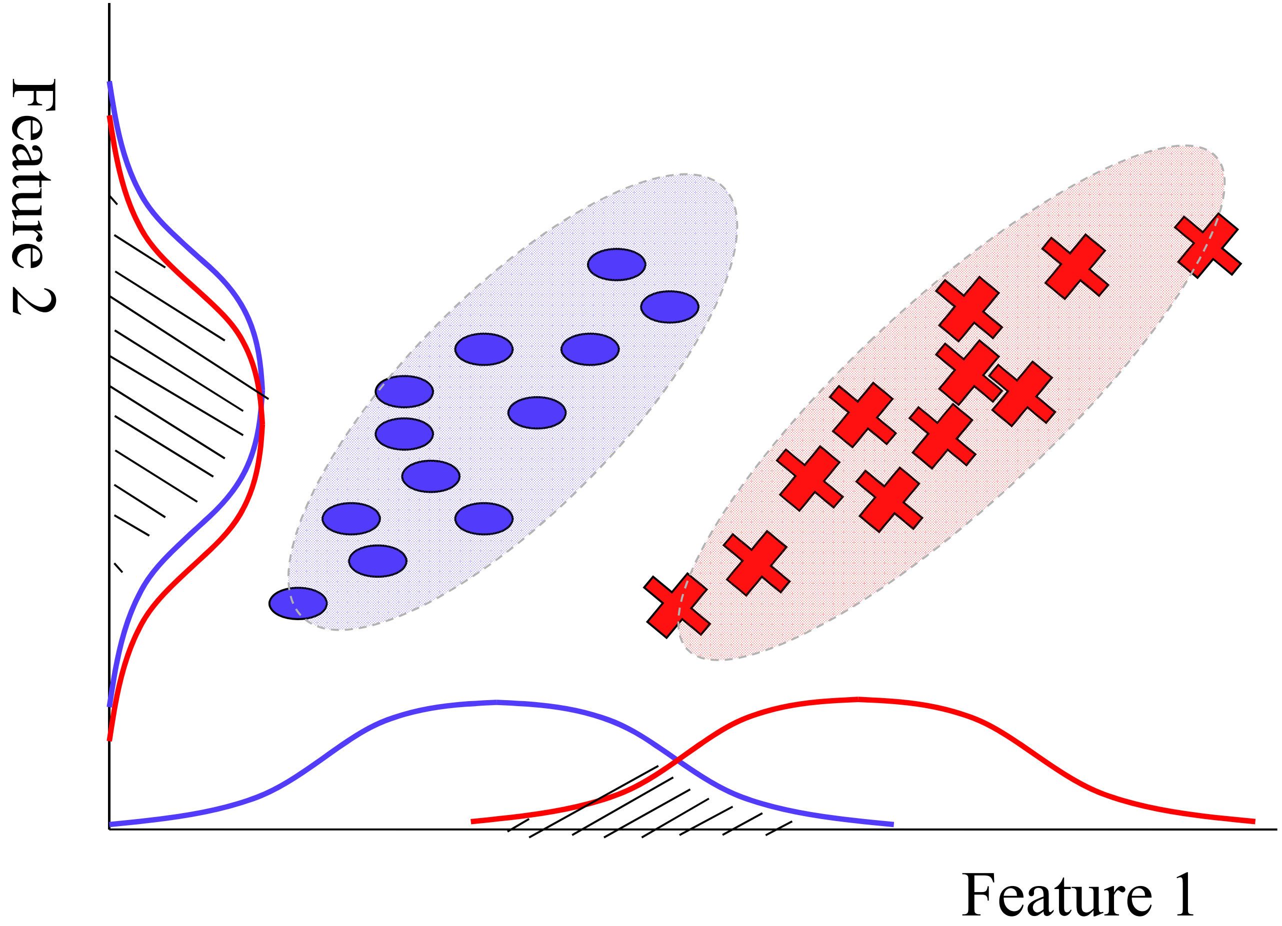


Marginal Probability Density and Conditional Probability Density in Machine Learning



Benefits of Using High Dimensionalities

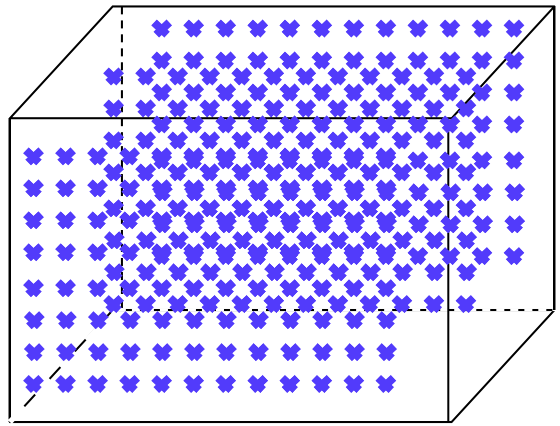
- Feature 1 and Feature 2 have correlation



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.

GAUSSIAN DENSITY FUNCTION

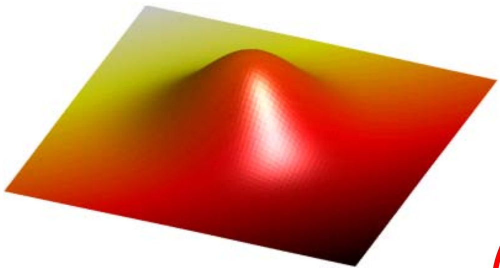
Gaussian Random Variable

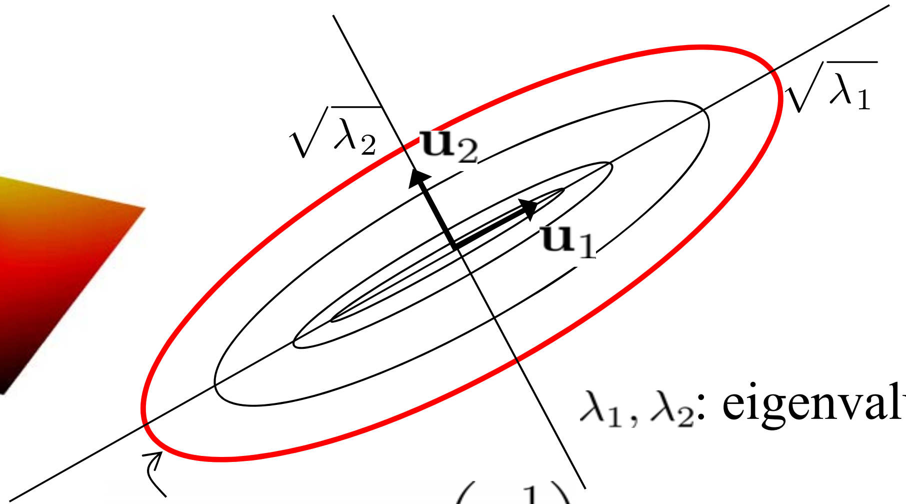
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^{D} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1} (\mathbf{x}-\mu)\right)$$

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

Principal axes are the eigenvector directions of Σ

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





λ_1, λ_2: eigenvalues

$$p(\mathbf{x}) = p(\mu)\exp\left(-\frac{1}{2}\right)$$

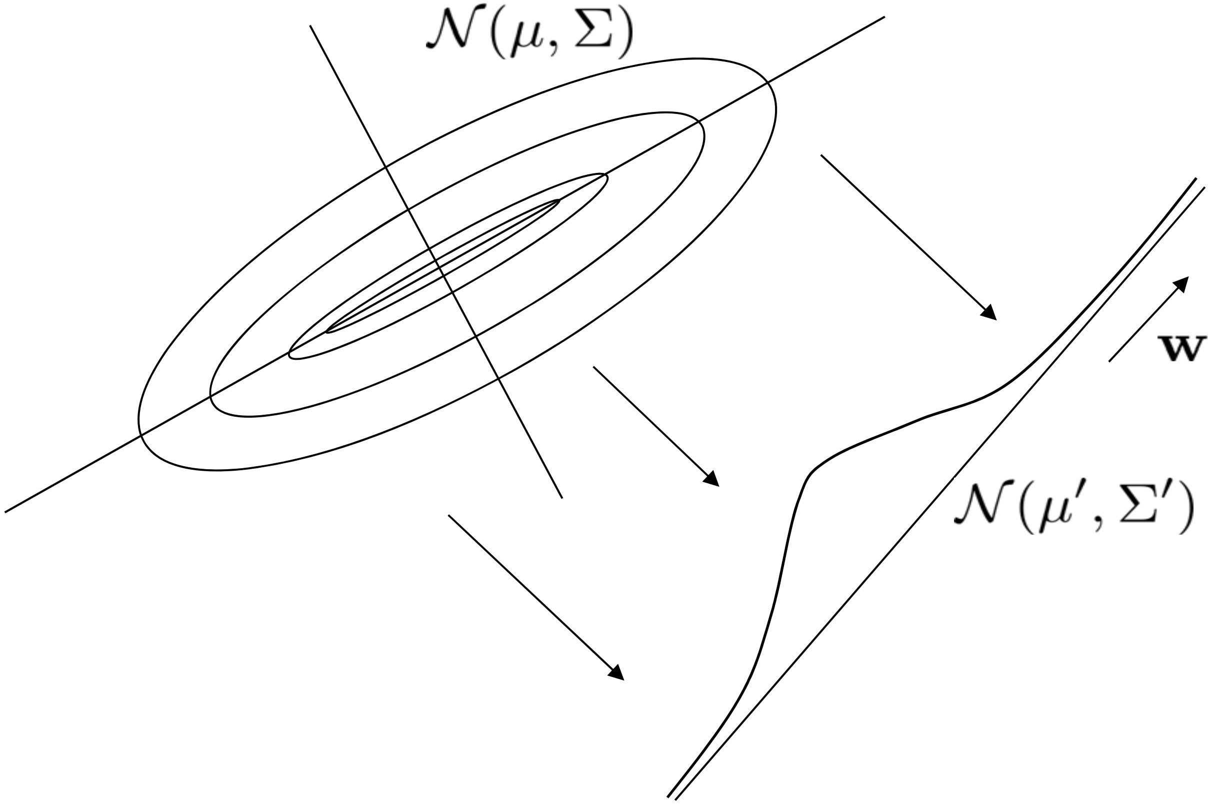
Gaussian Random Variable - Projection

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

Projection to any direction is Gaussian.

$$\mu' = \mathbf{w}^\top \mu$$

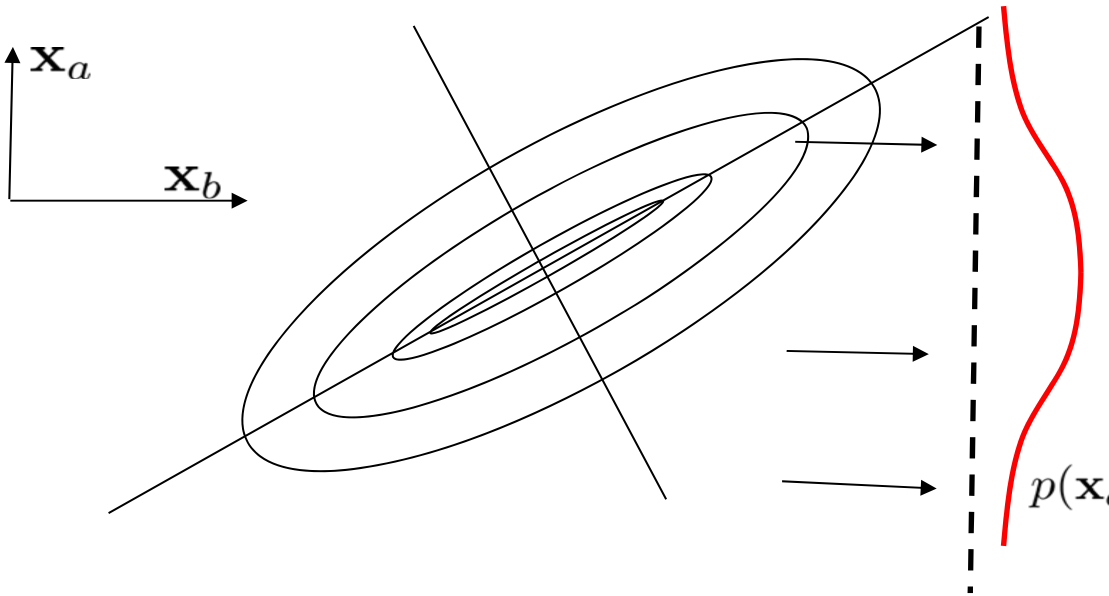
$$\Sigma' = \mathbf{w}^\top \Sigma \mathbf{w}$$



Gaussian Random Variable – Marginal

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

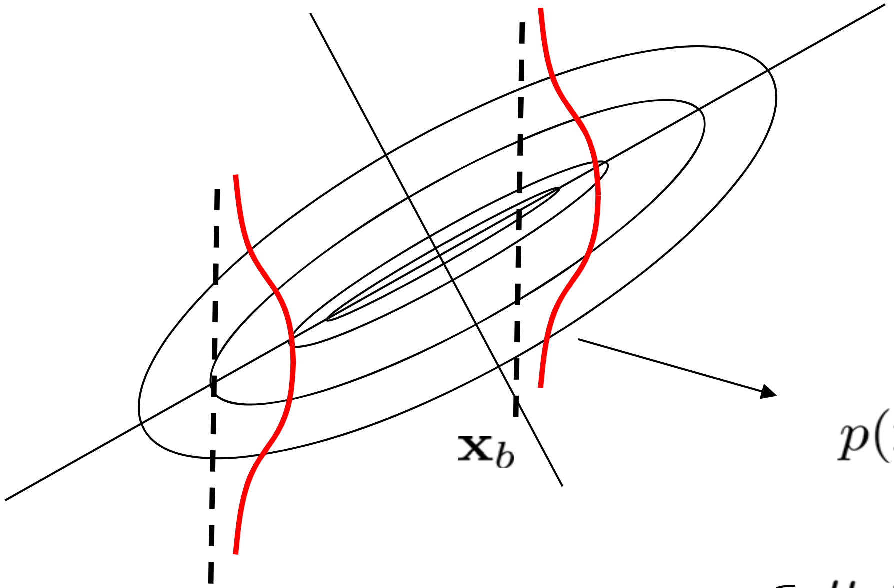
$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \begin{matrix} \mathbf{x}_a \in \mathbb{R}^{D_a} \\ \mathbf{x}_b \in \mathbb{R}^{D_b} \end{matrix} \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba}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



$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b)\,d\mathbf{x}_b$$
$$= \mathcal{N}(\mu_a, \Sigma_a)$$

Gaussian Random Variable – Conditional

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



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

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

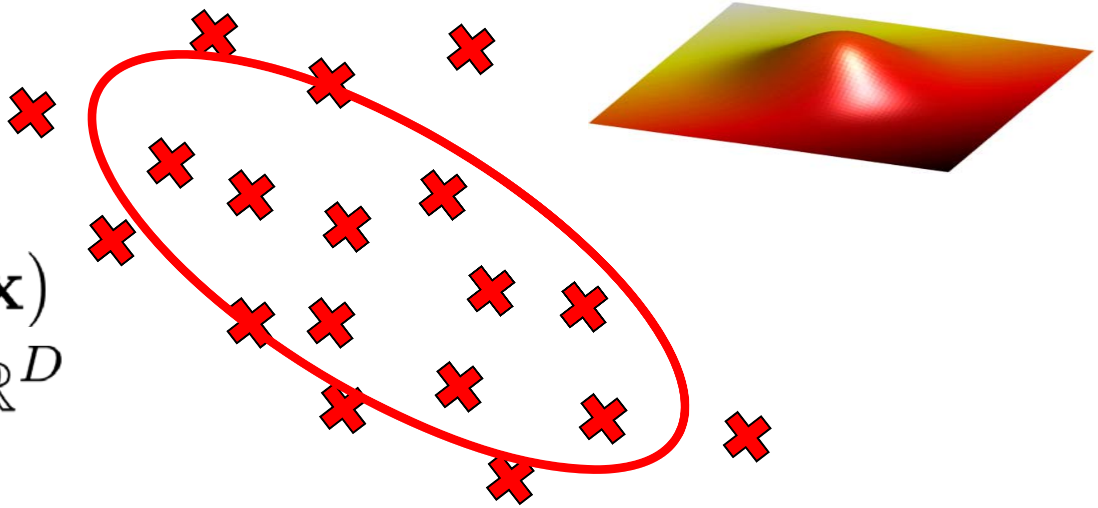
$$\begin{cases} \mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_b^{-1}(\mathbf{x}_b - \mu_b) \\ \Sigma_{a|b} = \Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ba} \end{cases}$$

PARAMETER ESTIMATION

Motivation – Parameter Estimation

- Parameter estimation is an optimization problem

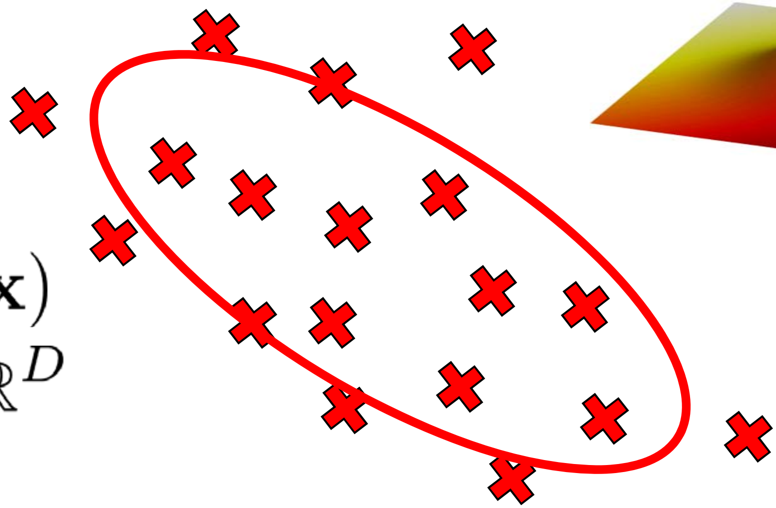
$$\mathbf{x} \sim p(\mathbf{x})$$
$$\mathbf{x} \in \mathbb{R}^D$$



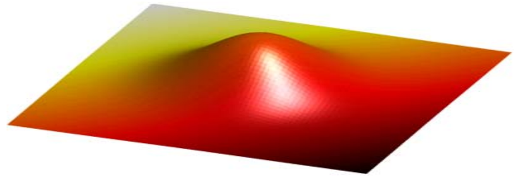
$\widehat{p}(\mathbf{x})$: estimated probability density function,
in other words, density function that fits data the most

Maximum Likelihood Estimation

- Parameter estimation is an optimization problem



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



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

Maximum Likelihood for Gaussian

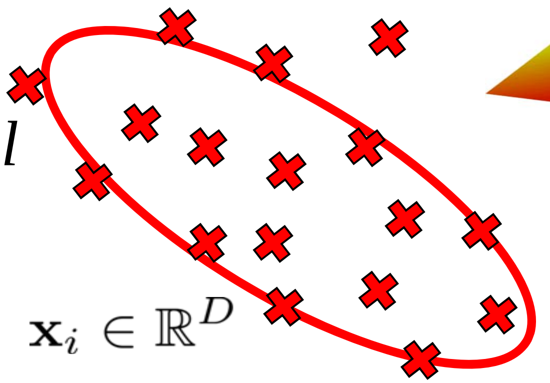
$$p(\mathbf{x}|\mu,\Sigma) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

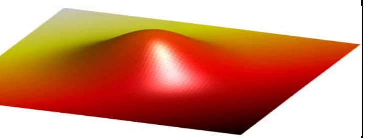
- With optimal parameters satisfying

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

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

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

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

  - 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),

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

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

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

MLE vs. MAP

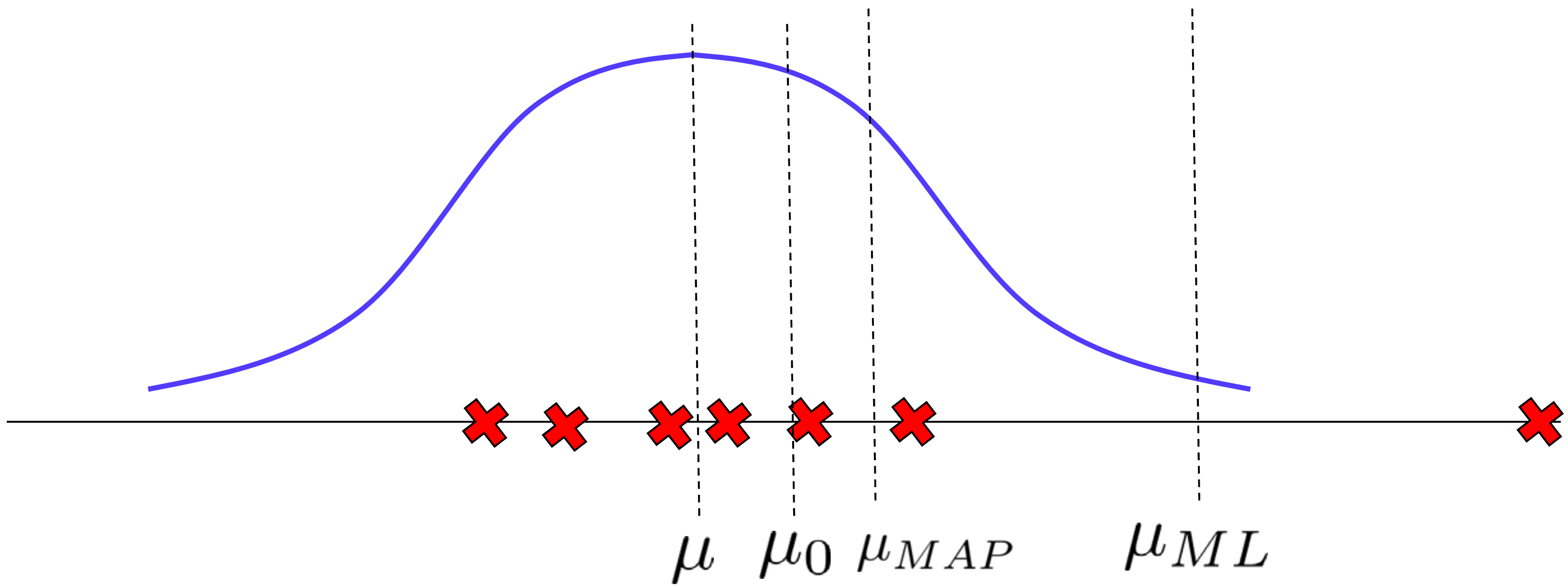
- For Gaussian
  - When we have a few outliers

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

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

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

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.

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

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

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

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

Uncertainty of μ

Conjugate Priors

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

$$\text{Prior } p(\theta) \xrightarrow{\text{Likelihood } p(x|\theta)} \text{Posterior } p(\theta|x)$$

Two distributions
Have the same form

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

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

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

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

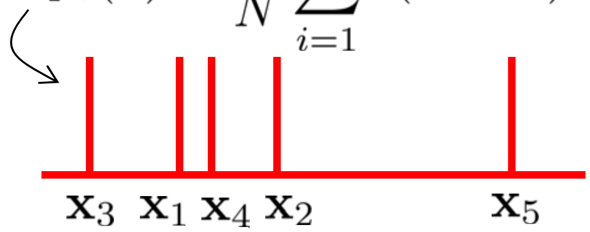
$$= \arg\max_{p_\theta} \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(\mathbf{x}_i)$$

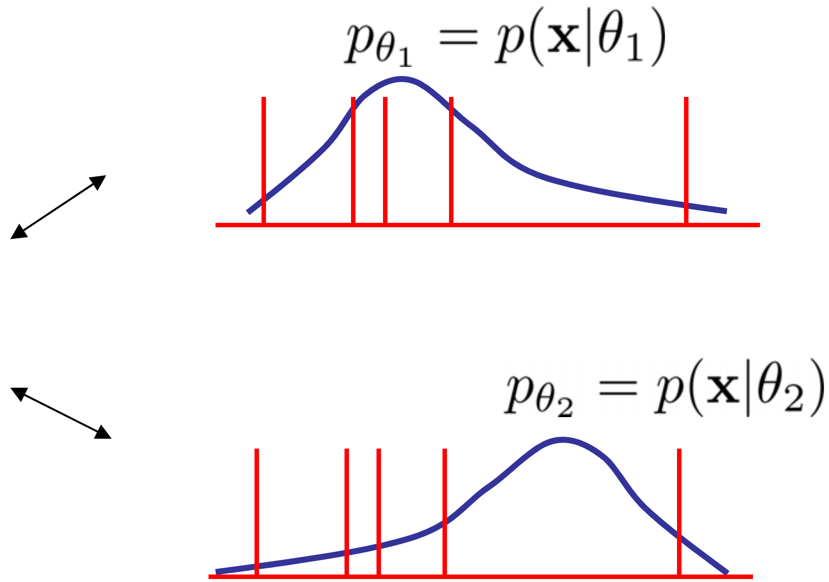
$$= \arg\max_{p_\theta} \log \prod_{i=1}^{N} p_\theta(\mathbf{x}_i) = \arg\max_{p_\theta} p(\mathcal{D}|\theta)$$

Kullback-Leibler Divergence

$$N = 5$$

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

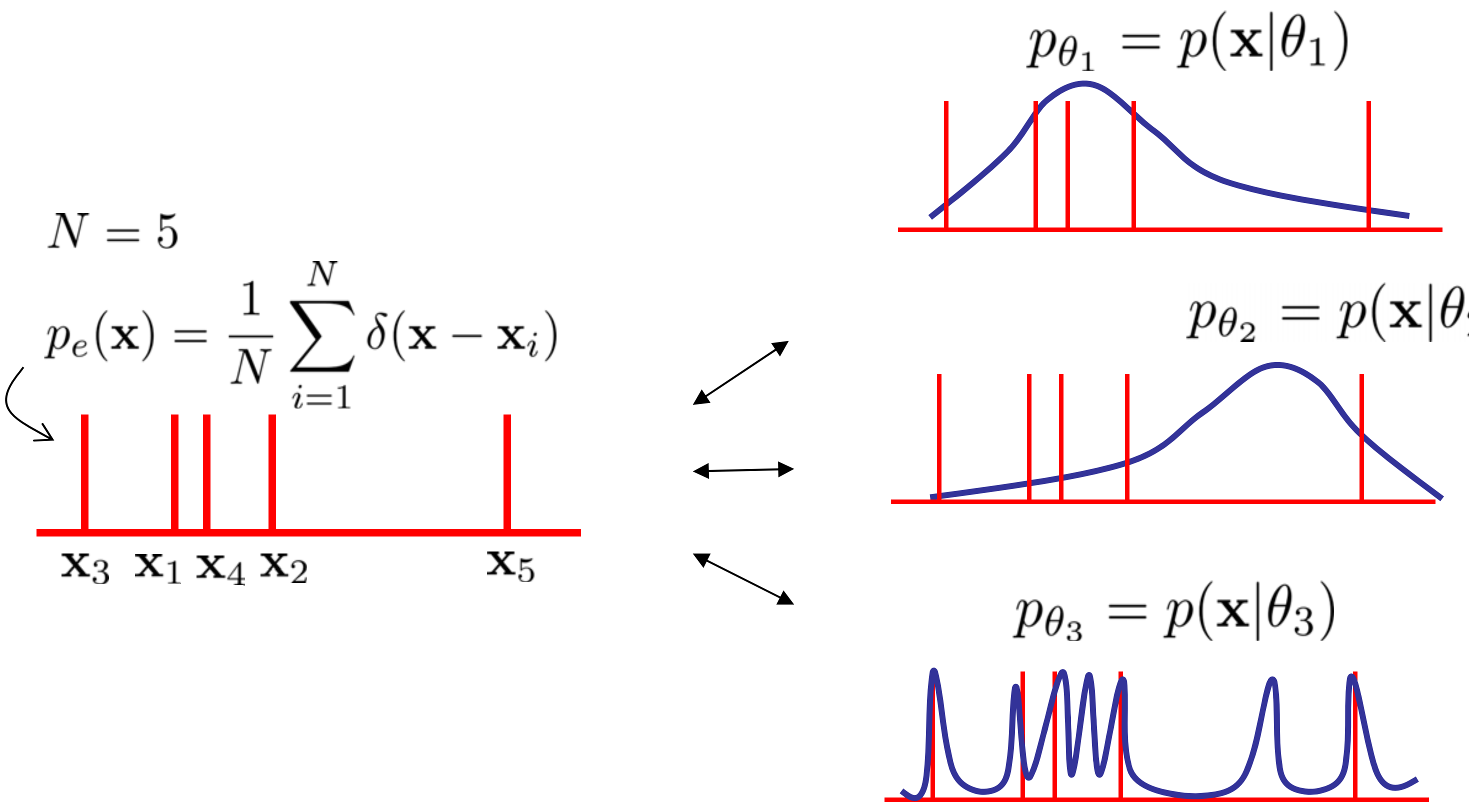




KL Divergence: $KL(p_e||p_{\theta_1}) < KL(p_e||p_{\theta_2})$

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

Kullback-Leibler Divergence

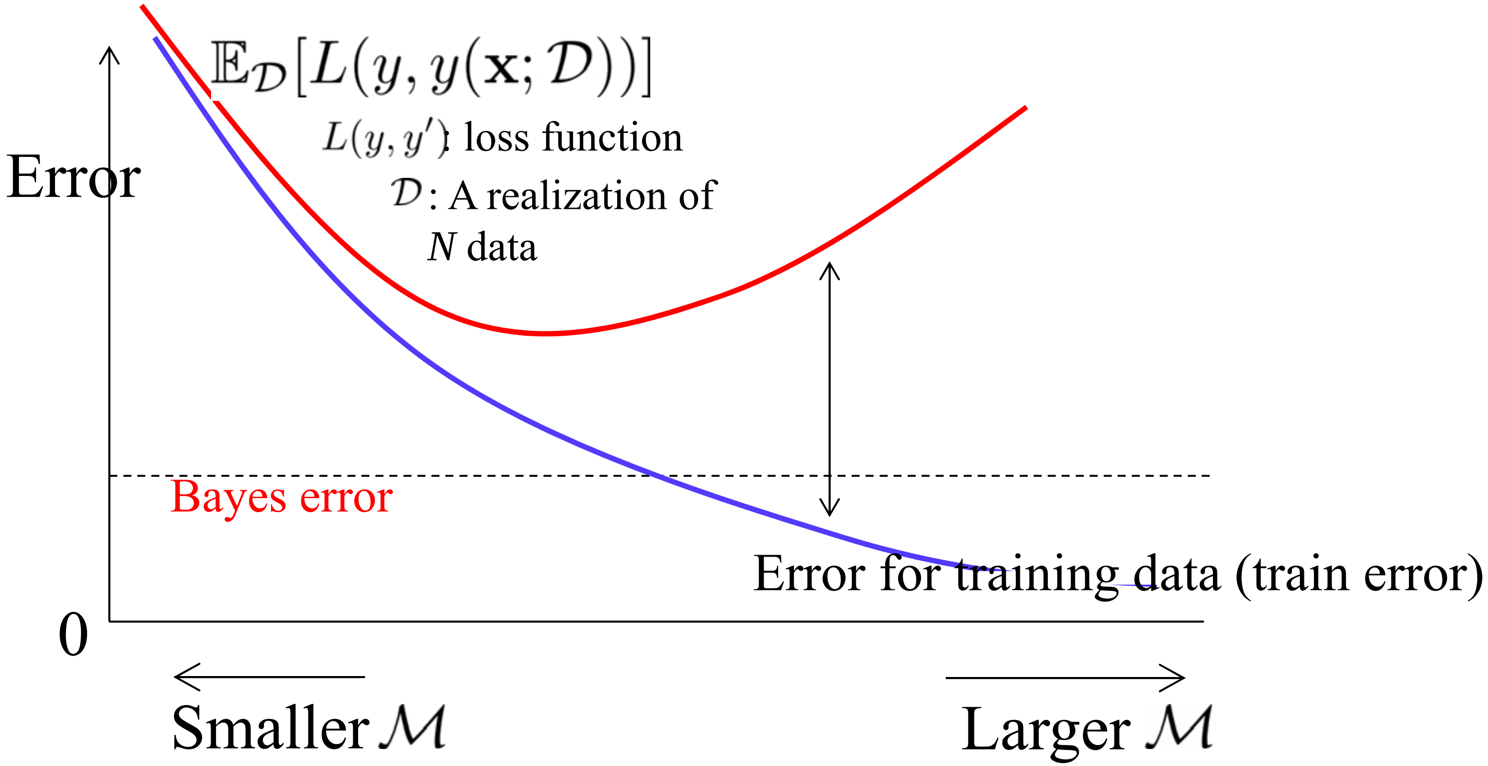


$$\theta_3 = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

Model with complex function will capture the noise.

Consistency and Bayes Error

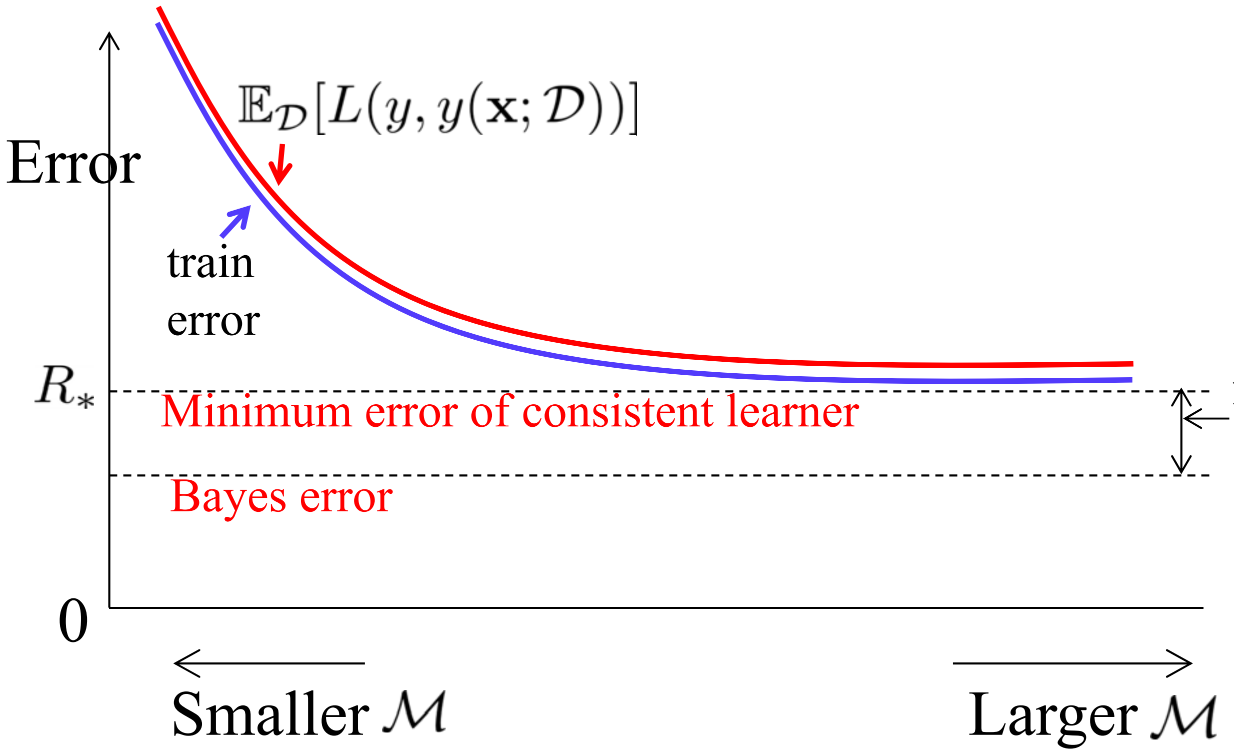
- Objective: minimizing expected error



(For example, a linear classifier with regularization)

Consistency and Bayes Error

- Consistent learner with many data

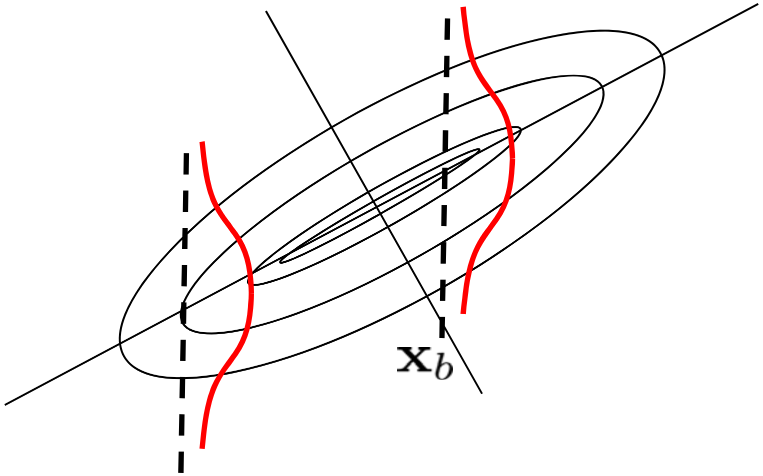


Consistency does not necessarily imply achieving the Bayes error!!!

(For example, a linear classifier with regularization)

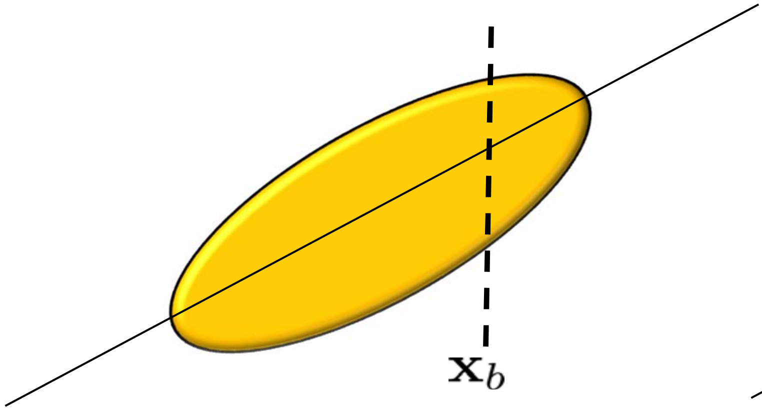
GAUSSIAN PROCESS REGRESSION

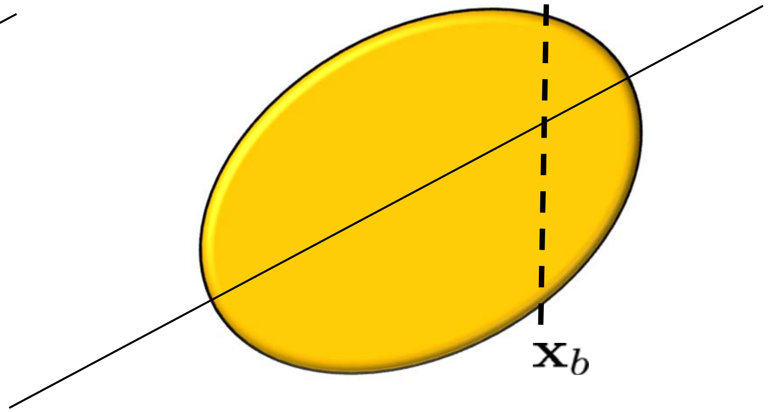
Prediction Using Correlation



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

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





Gaussian Process

$$y(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$$
$$m(\mathbf{x}) = \mathbb{E}[y(\mathbf{x})] = 0$$
$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(y(\mathbf{x}) - m(\mathbf{x}))(y(\mathbf{x}') - m(\mathbf{x}'))]$$
$$= \exp\left\{-\frac{\theta}{2}||\mathbf{x} - \mathbf{x}'||^2\right\}$$

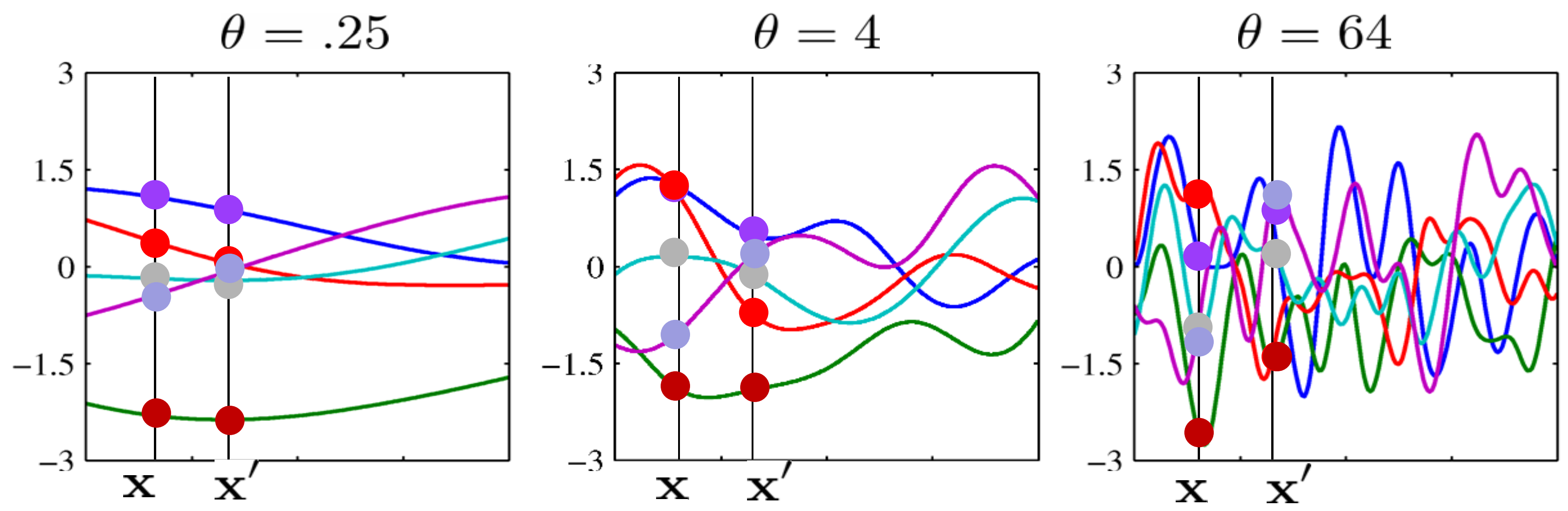
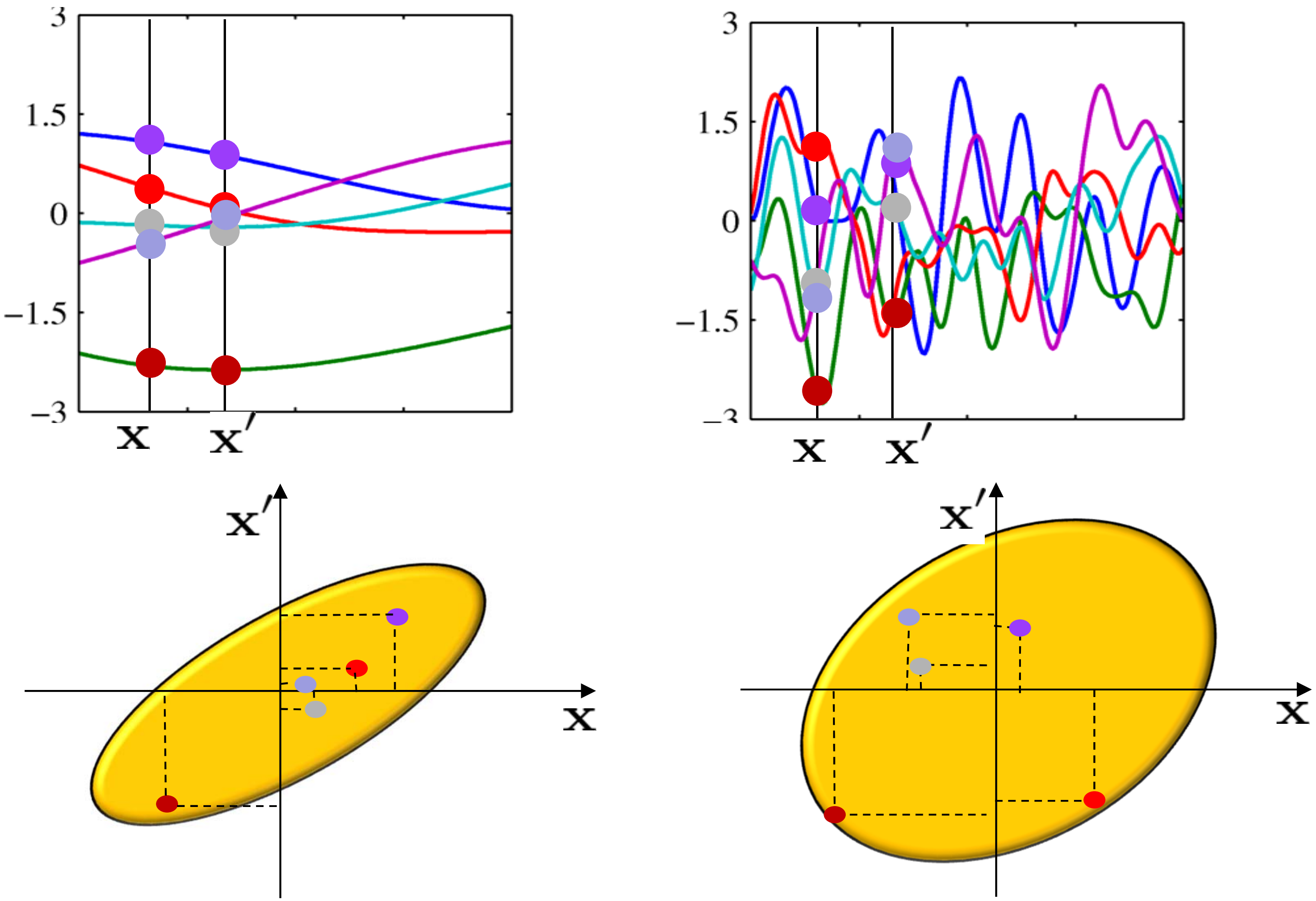
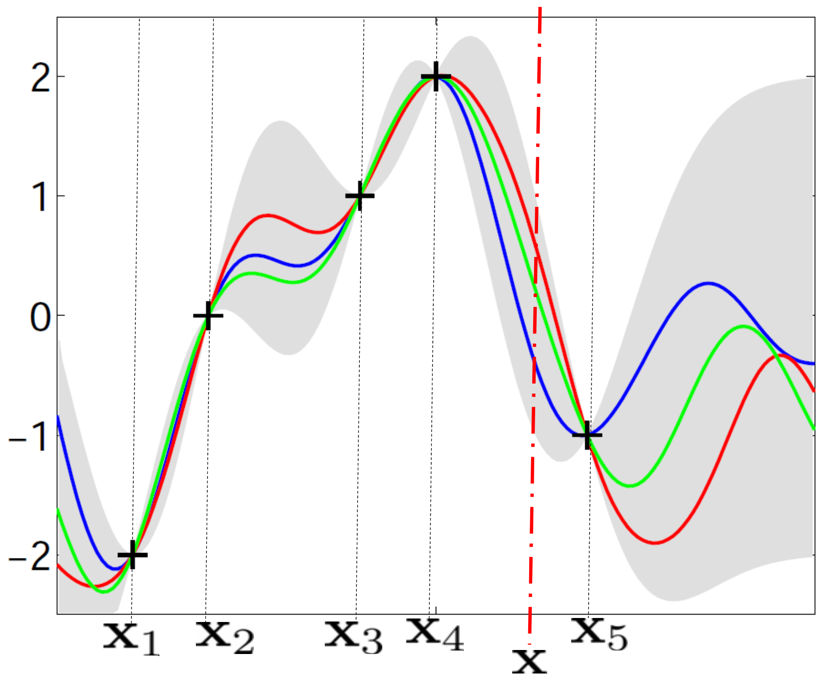


Figure credit: Christopher Bishop

Gaussian Process as an Infinite Dimensional Gaussian



Gaussian Process Regression



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

$$\mathbb{E}[y(\mathbf{x})] = \mathbf{k}^\top K^{-1}\mathbf{y}$$

$$\mathbb{V}[y(\mathbf{x})] = k(\mathbf{x},\mathbf{x}) - \mathbf{k}^\top K^{-1}\mathbf{k}$$

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

$$K = \begin{pmatrix} k(\mathbf{x}_1,\mathbf{x}_1) & \cdots & k(\mathbf{x}_1,\mathbf{x}_N) \\ k(\mathbf{x}_2,\mathbf{x}_1) & \cdots & k(\mathbf{x}_2,\mathbf{x}_N) \\ \cdots & \cdots & \cdots \\ k(\mathbf{x}_N,\mathbf{x}_1) & \cdots & k(\mathbf{x}_N,\mathbf{x}_N) \end{pmatrix}$$

$$\mathbf{k} = \begin{pmatrix} k(\mathbf{x}_1,\mathbf{x}) \\ k(\mathbf{x}_2,\mathbf{x}) \\ \vdots \\ k(\mathbf{x}_N,\mathbf{x}) \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{pmatrix}$$

Figure credit: C. E. Rasmussen & C. K. I. Williams

Summary

- What we did:
  - Probability and probability density
  - Conditional density, marginalized density
  - Model construction
  - Gaussian model
  - Parameter estimation
  - Gaussian process

- What we didn't do:
  - Multinomial distribution and Dirichlet distribution
  - Convergence of estimation
  - Generative model vs. Discriminative model

THANK YOU

Yung-Kyun Noh
nohyung@snu.ac.kr