

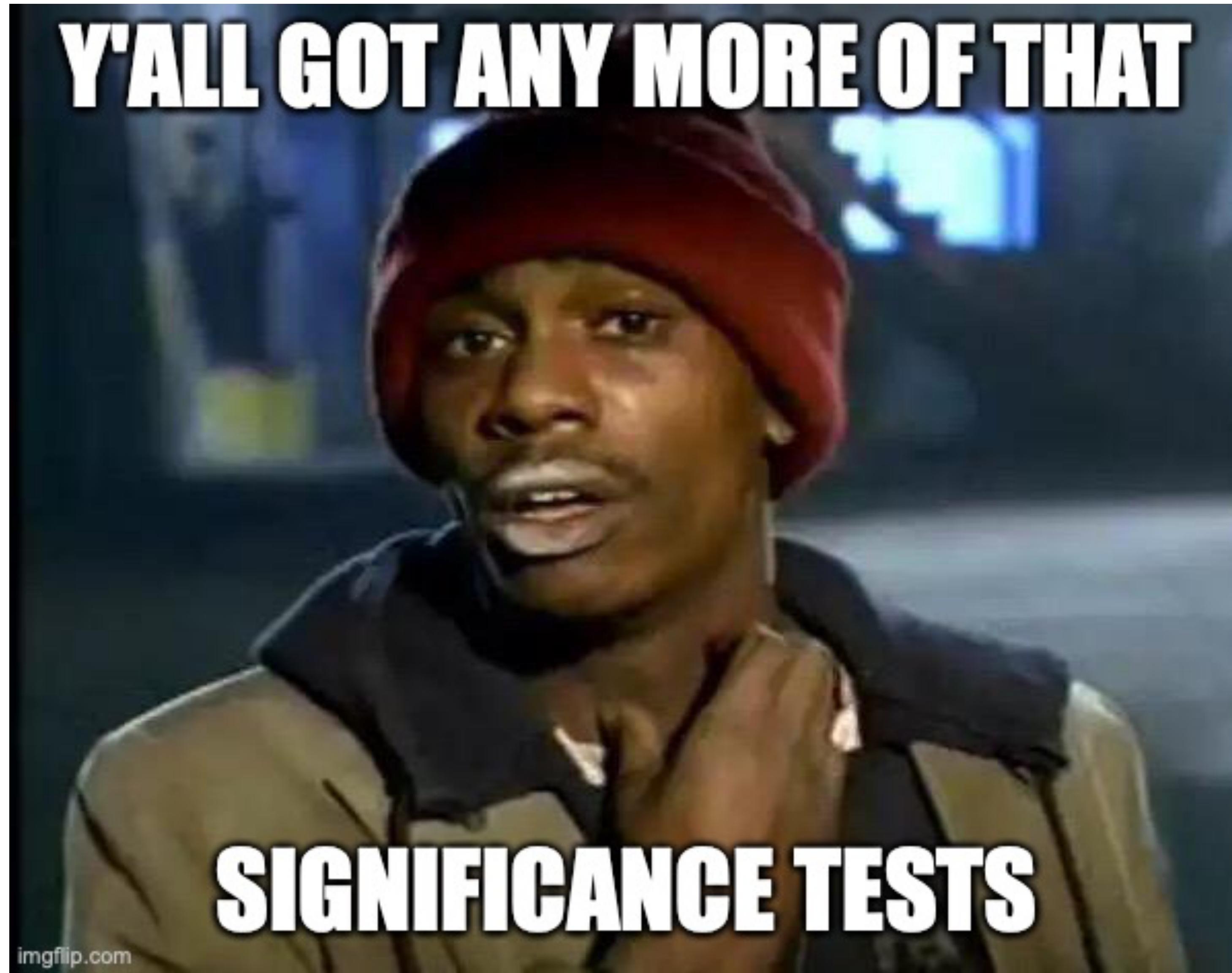
Statistical Techniques for Data Science & Robotics

Week 4

Objectives (for today)

- Recap of the last week
- Decision Theory / Classification
- Regression Models:
 - Loss function for regression
 - Linear Regression
 - The Bias-Variance Decomposition

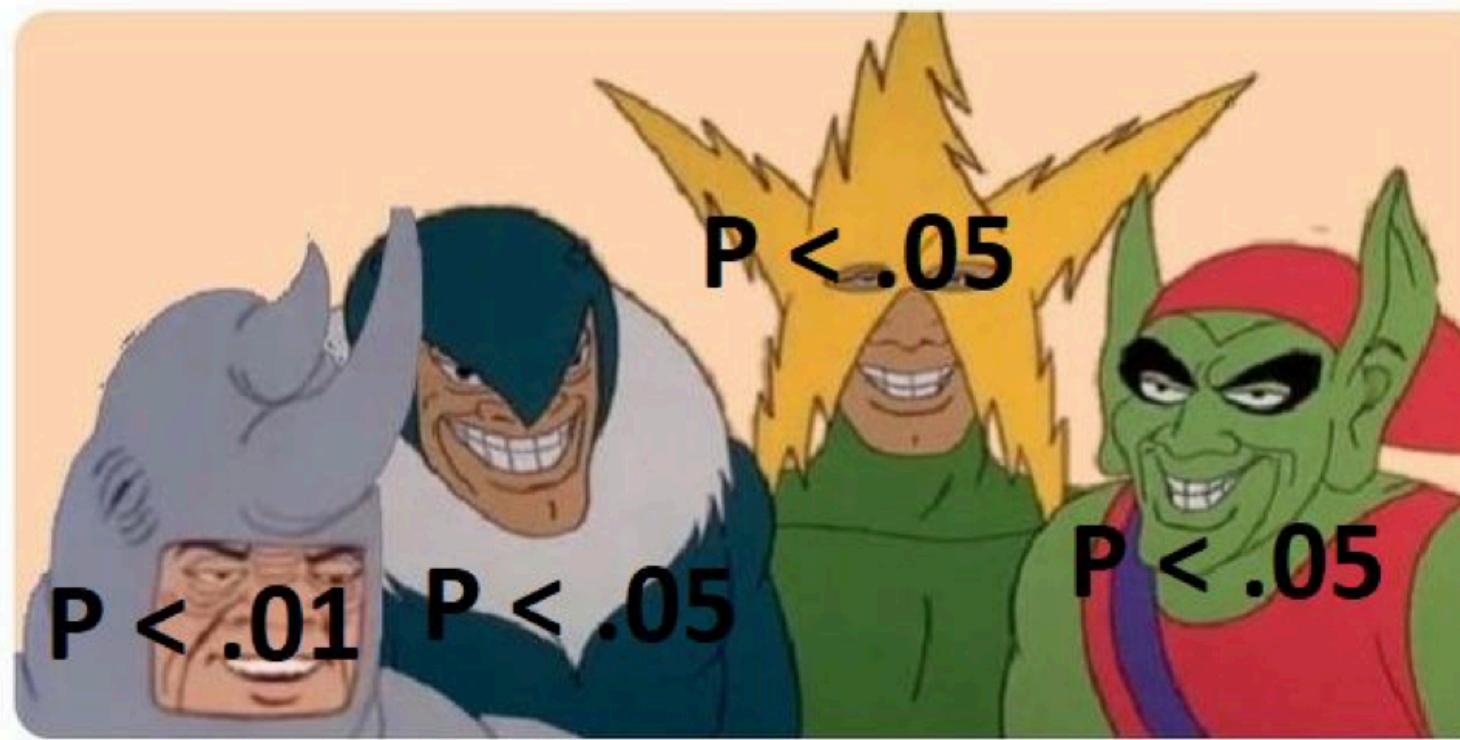
Last week debt / recap



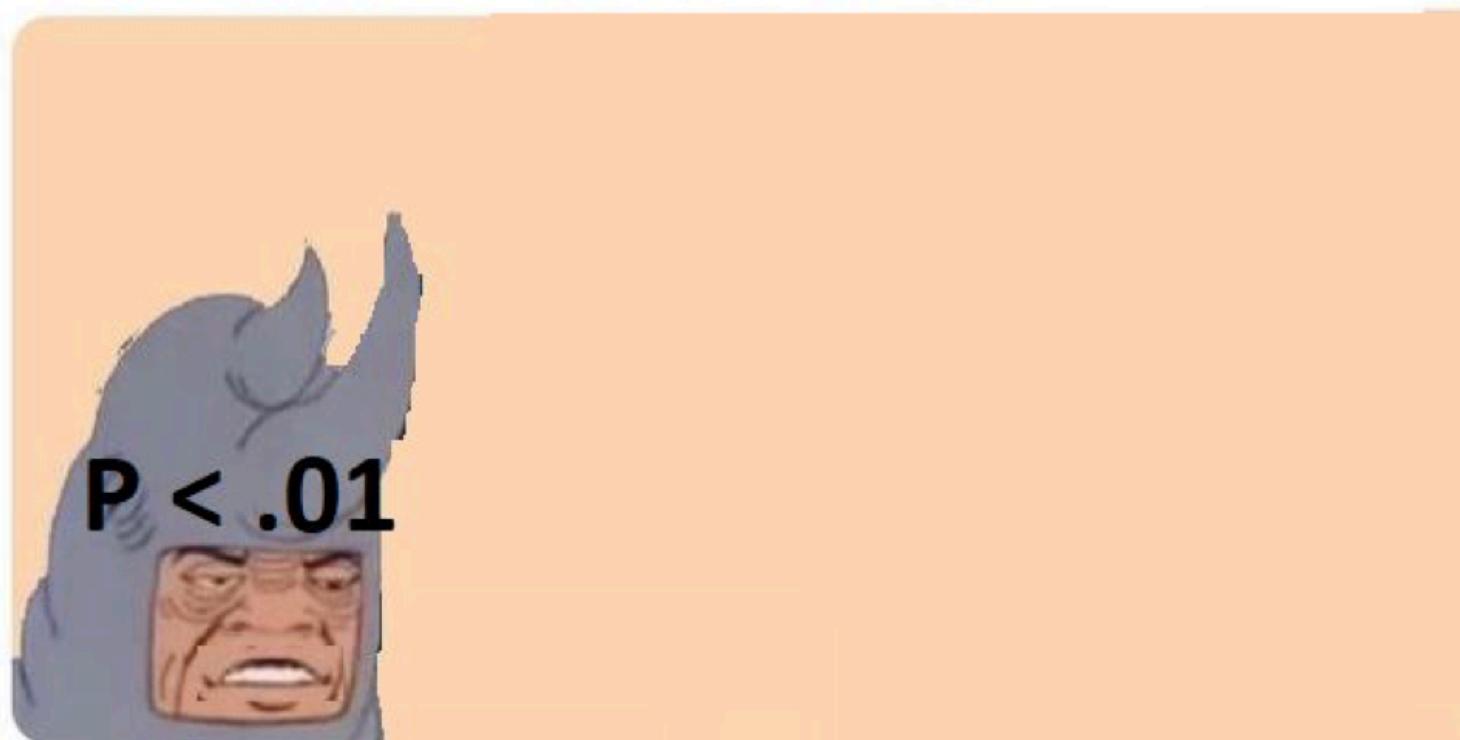
Multiple testing and Correction

- Significance tests:
 - Z-test,
 - t-test,
 - Chi-square GoF
 - ANOVA

Me and the significant boys



Me and the significant boys after Bonferroni correction



- Bonferroni Correction

when you do multiple testing, adjust the significance level, α

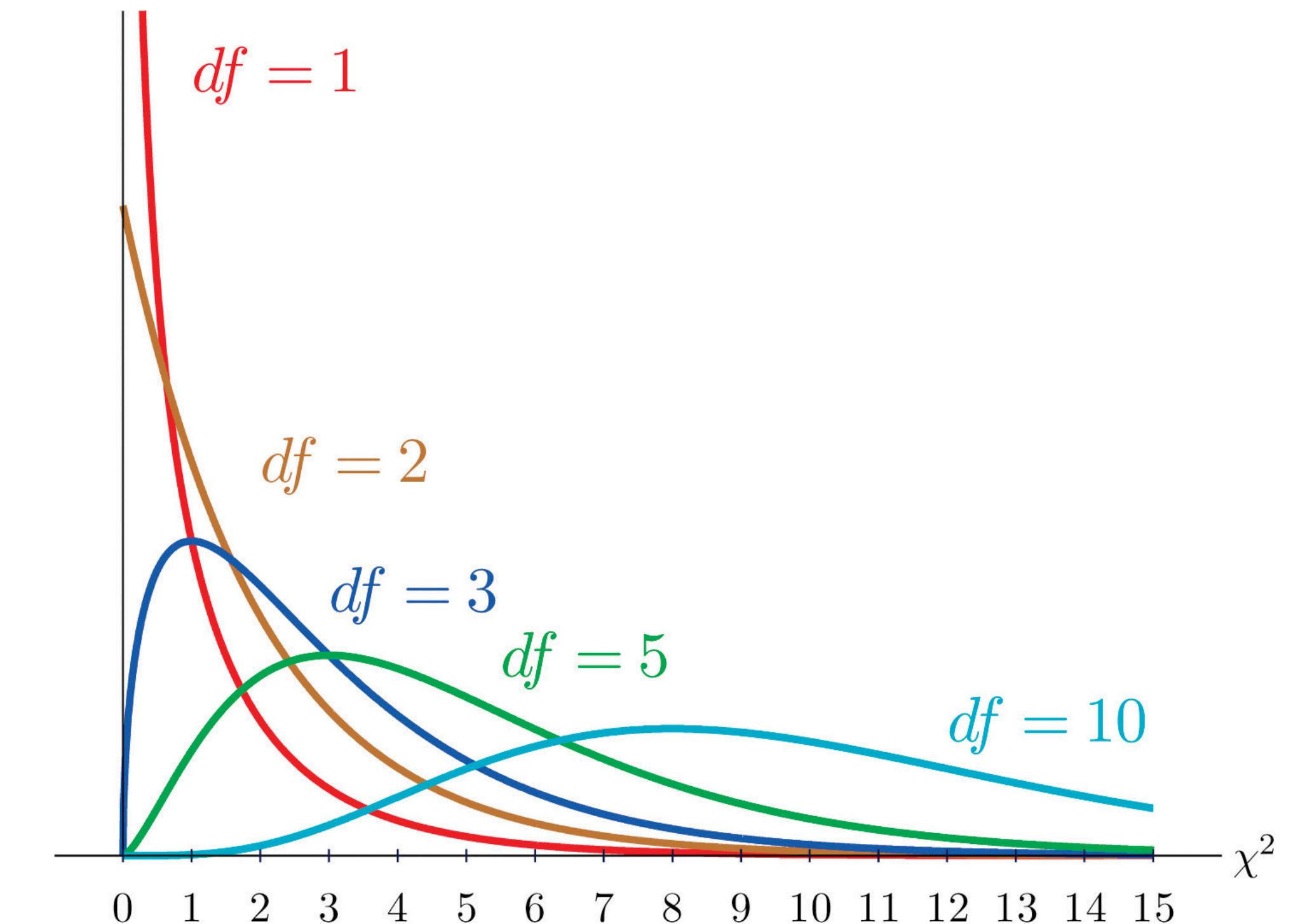
$$\alpha_{corr} = \frac{\alpha}{\# \text{ of tests}}$$

Chi-square distribution, χ^2

X_i ($i = 1, \dots, n$) is an IID sample

$$\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2}$$

this random variable has the chi-square distribution with **n degrees of freedom**



Degrees of freedom, χ^2

X_i ($i = 1, \dots, n$) is an IID sample

- Question: what is the distribution of $\frac{(n - 1)S^2}{\sigma^2}$?

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n [(X_i - \bar{X}) + (\bar{X} - \mu)]^2$$

$$= \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (\bar{X} - \mu)^2 + 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \bar{X})$$

$$= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$

$$\chi_n^2$$

$$\chi_{n-1}^2$$

$$\chi_1^2$$

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = \frac{(n - 1)S^2}{\sigma^2} + \frac{(\bar{X} - \mu)^2}{\sigma^2/n}.$$

Theorem: If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2 , then the statistic

$$\chi^2 = \frac{(n - 1)S^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$$

has a chi-squared distribution with $v = n - 1$ degrees of freedom.

Decision Theory

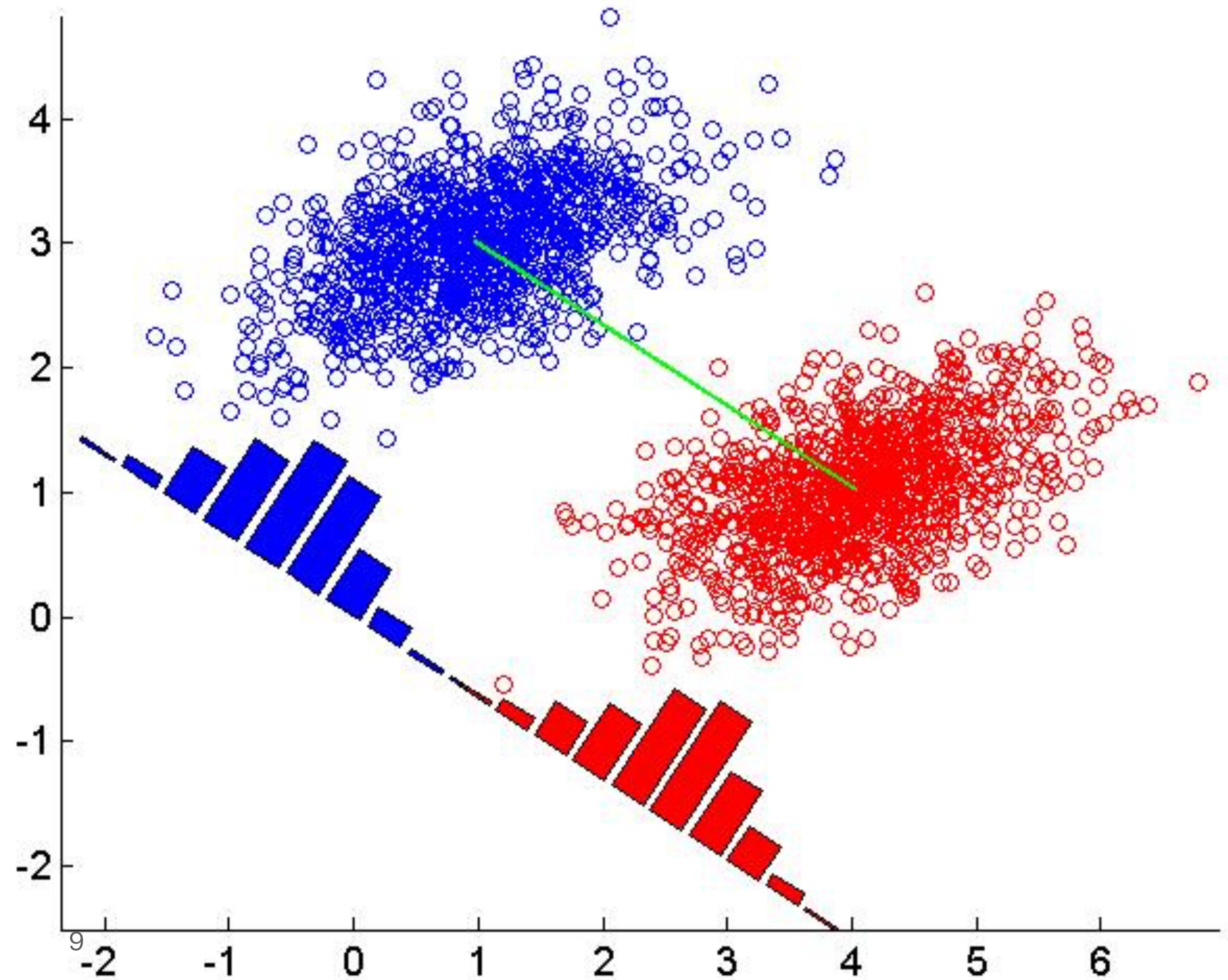
Introduction to Decision Theory

- Decision Theory
 - helps us making optimal decisions
 - in case of uncertainty (that is modeled with Prob.Theory)
- **First step:** to infer (or learn) the $p(x,t)$ - a joint distribution of inputs and targets
- **Second step:** use probabilities to make optimal decisions

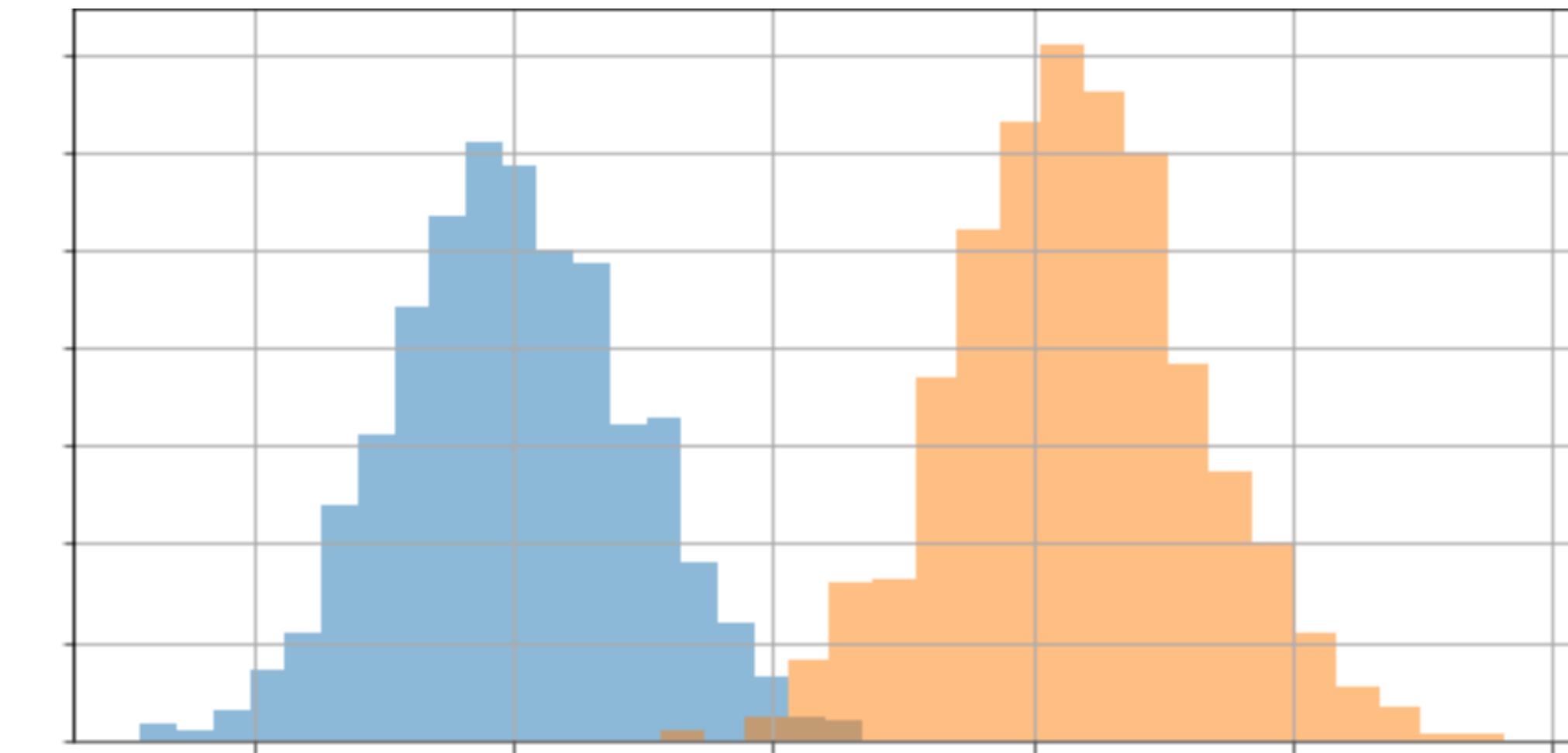
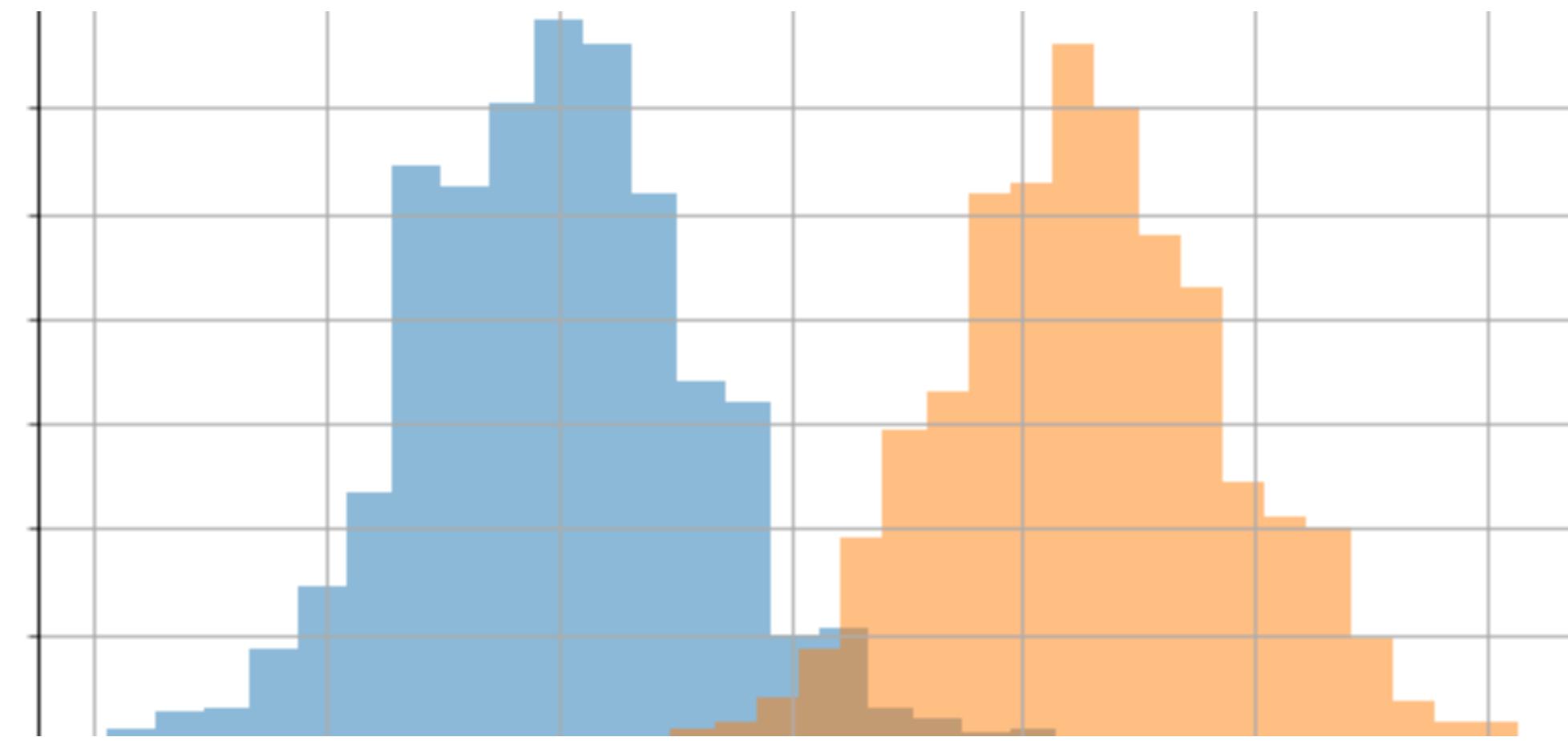
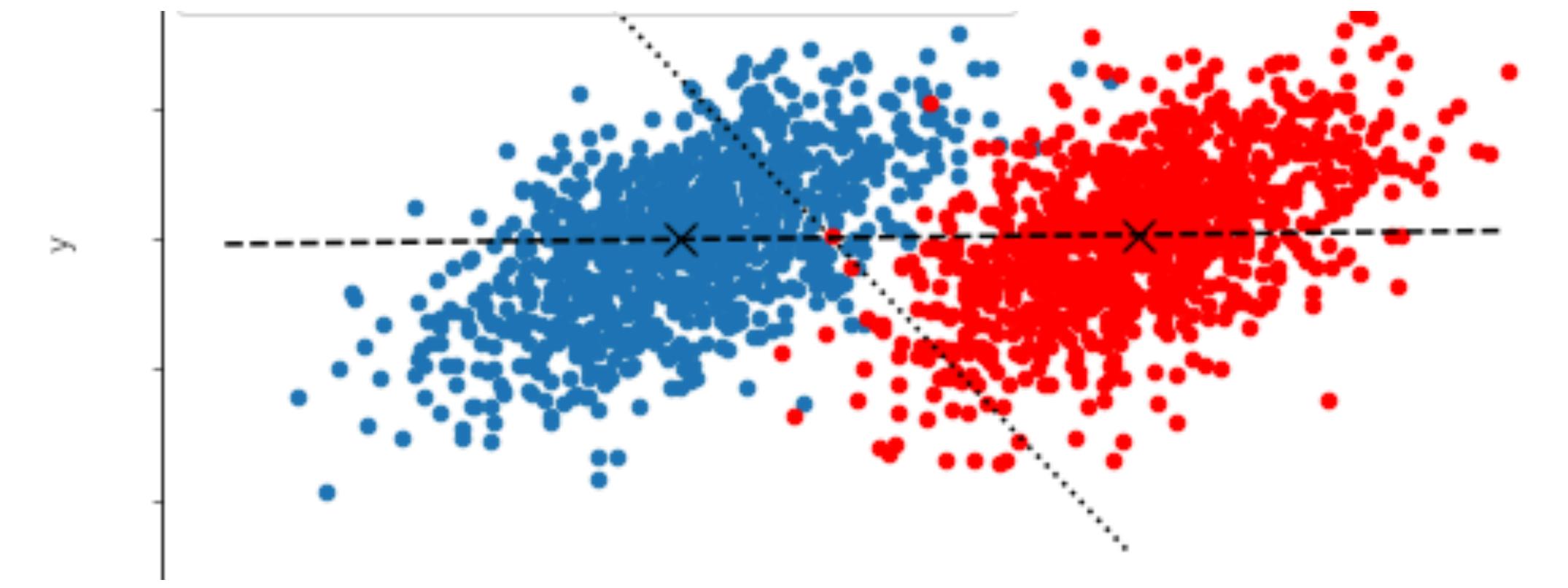
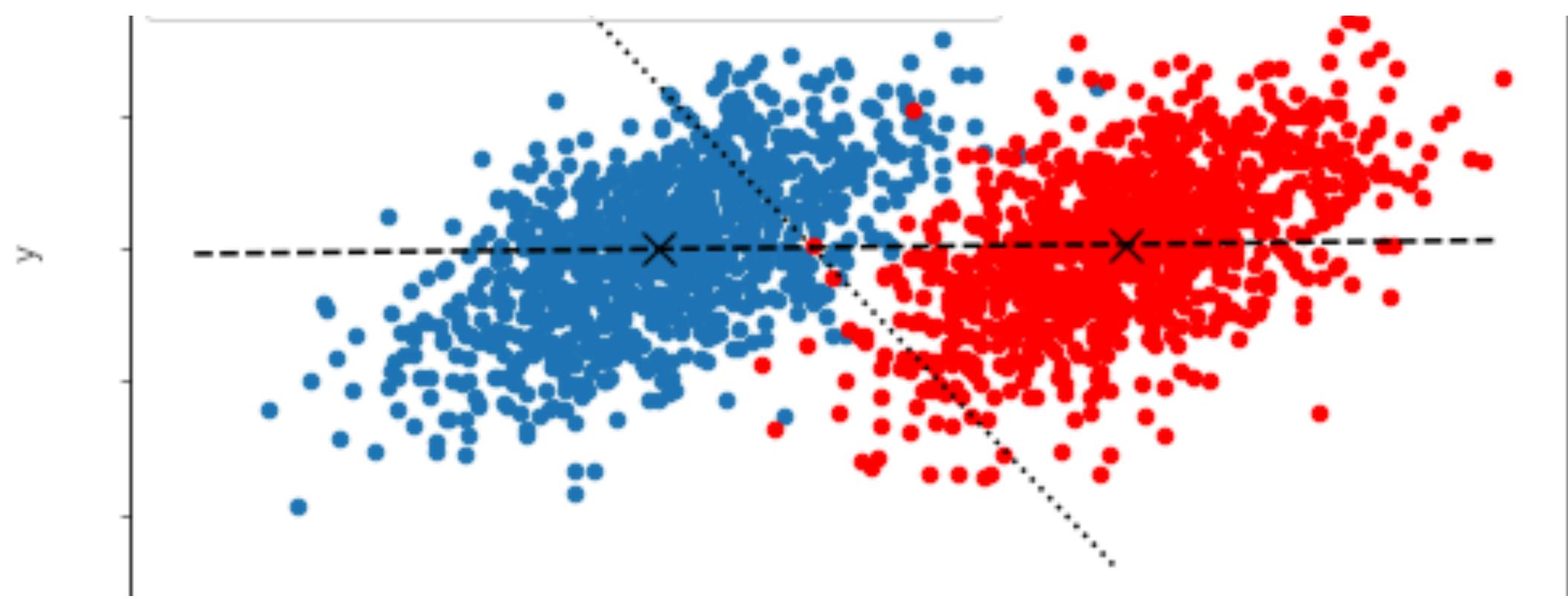
Setup: classification with 2 classes

a medical diagnosis problem

- We need a decision rule:
- \mathbf{x} is an input (a D-dim vector);
- $t \in \{0,1\}$;
- $t = 0 \implies C_1$;
- $t = 1 \implies C_2$



View of data affects the decision



Decision is based on the distribution

a medical diagnosis problem

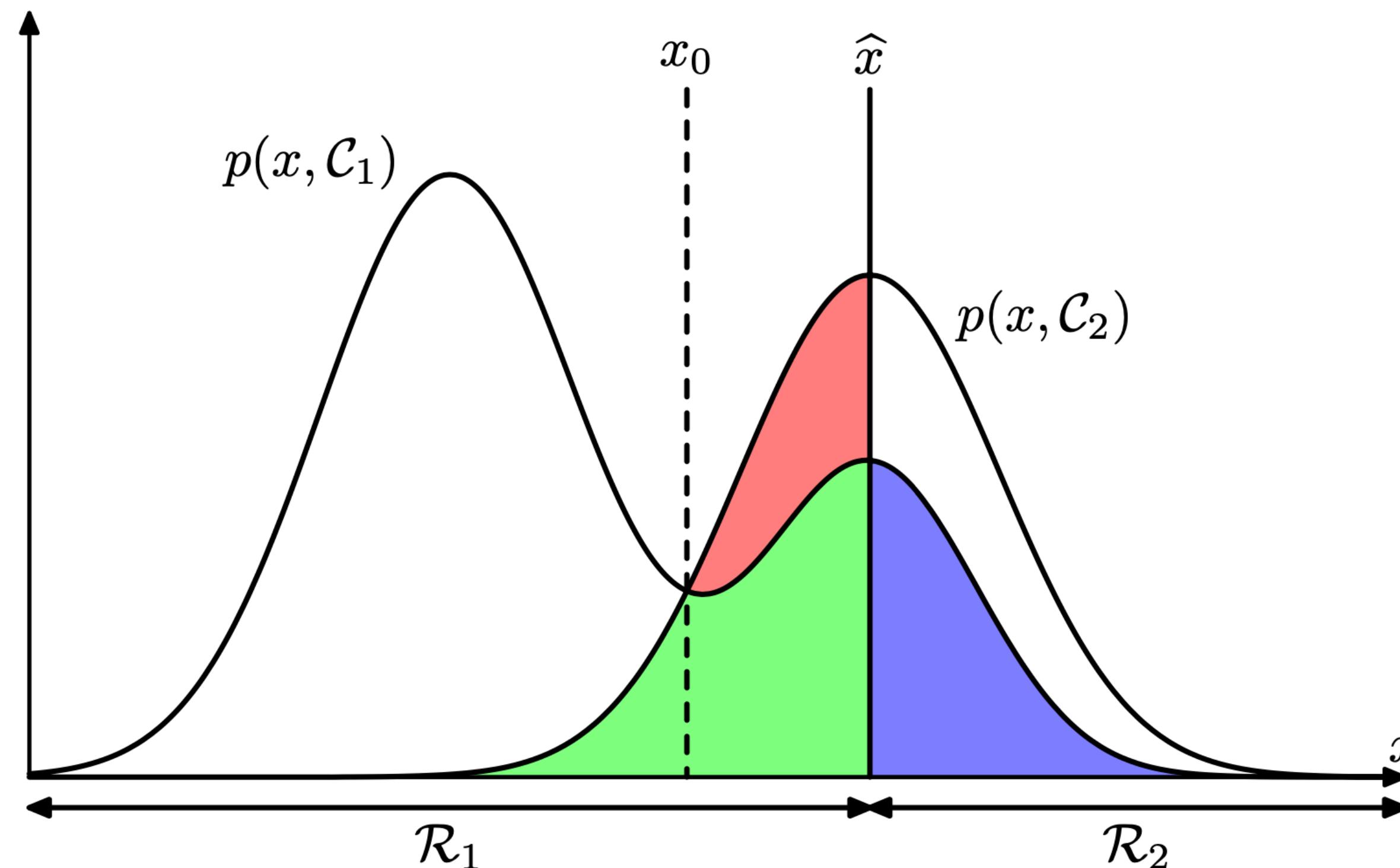
- => we need a decision rule.
- Setup:
 - \mathbf{x} is an input (a vector);
 - $t \in \{0,1\}$;
 - $t = 0 \implies C_1$;
 - $t = 1 \implies C_2$

$$p(C_k | \mathbf{x}) = \frac{p(\mathbf{x} | C_k) p(C_k)}{\sum_{k=1,2} p(\mathbf{x})}$$

Why is the $p(x,t)$ not present here?

Decision regions and boundaries

- R_1, R_2 are decision regions with assignments of \mathbf{x} to C_1, C_2



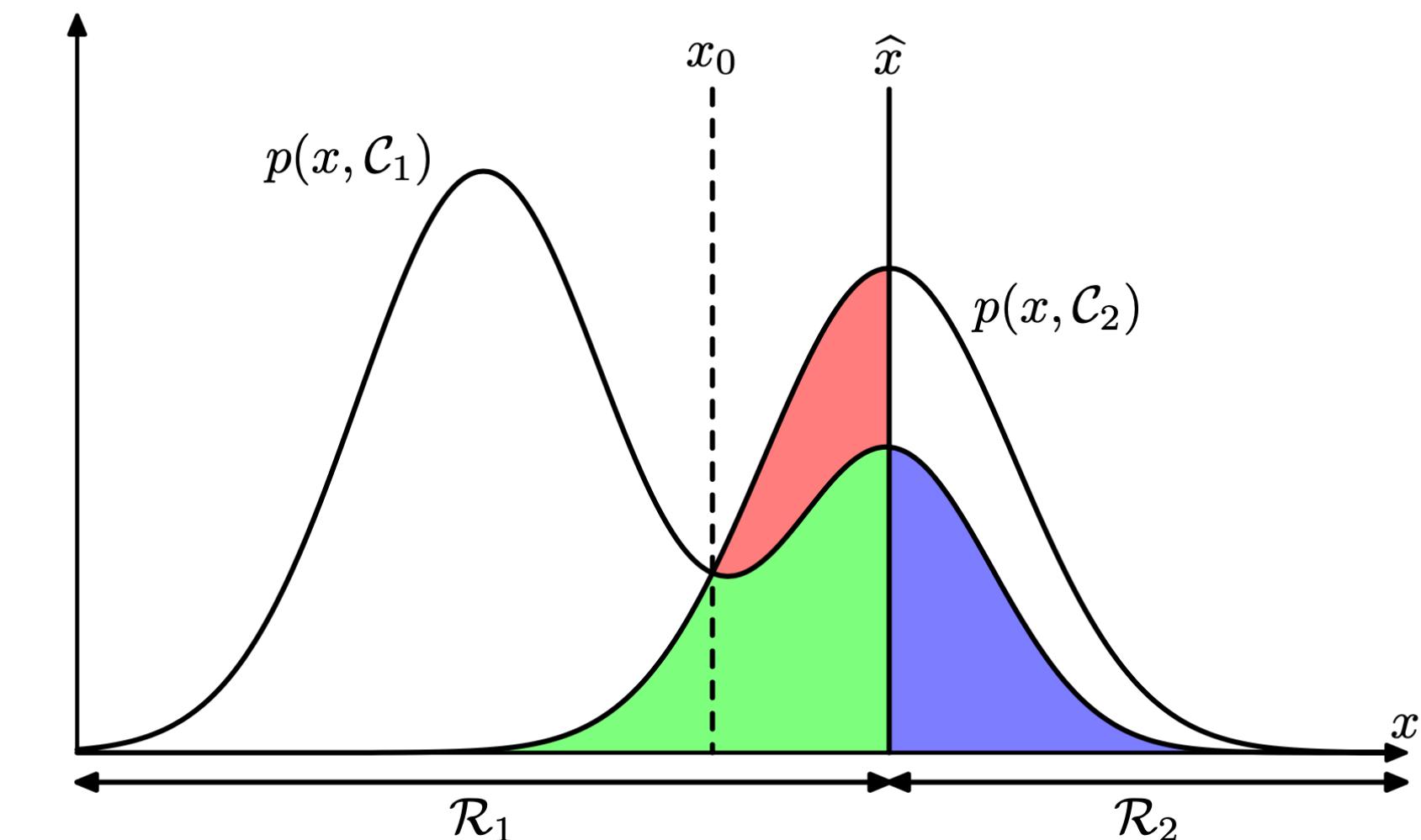
- Is \hat{x} the optimal decision boundary?

Minimizing errors of classification

- our rule should minimize by picking a decision boundary (x_0):

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x}. \end{aligned}$$

- $\mathcal{R}_1, \mathcal{R}_2$ are regions with **wrong** assignments of \mathbf{x} to C_2, C_1



Inference and decision

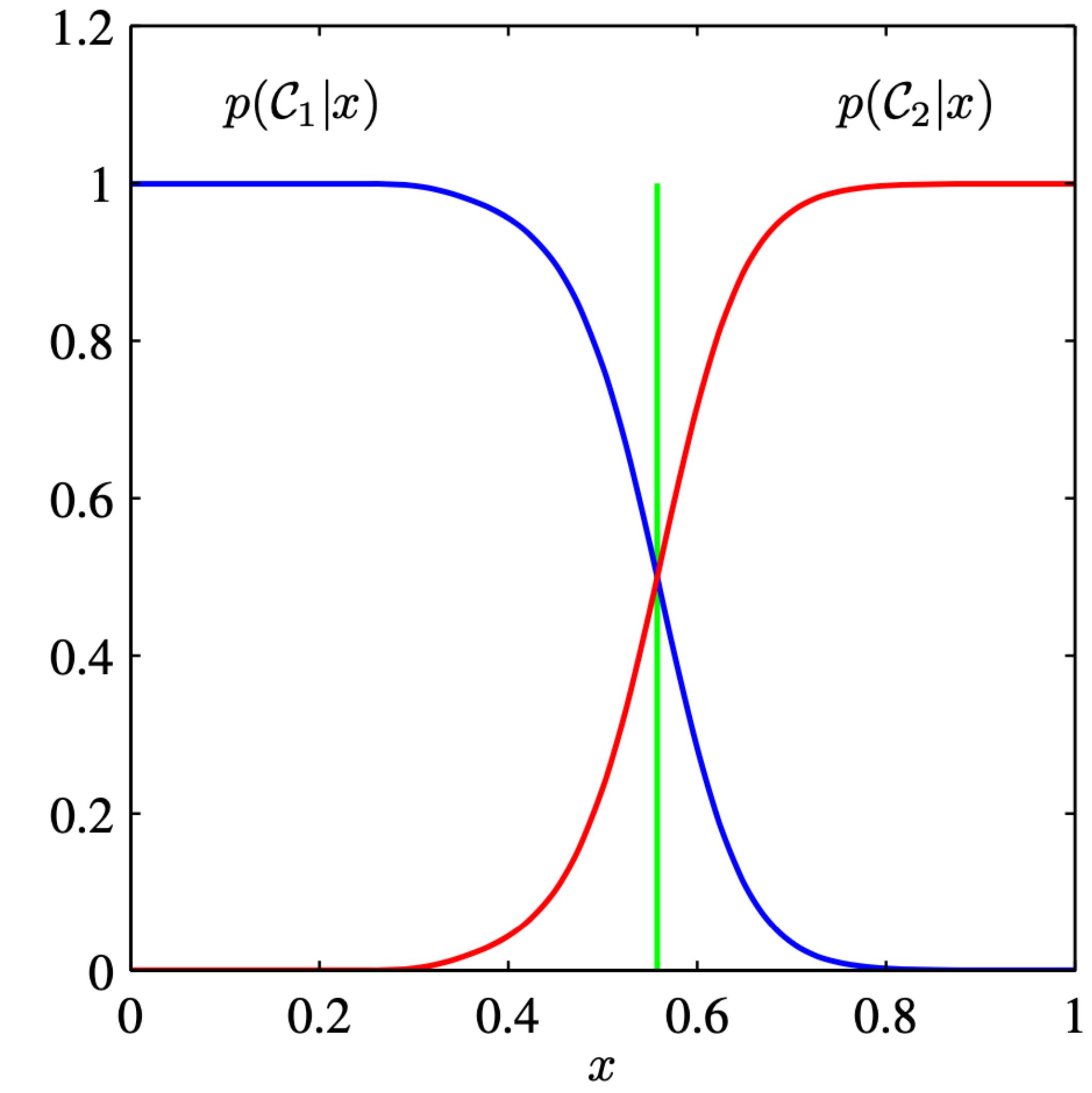
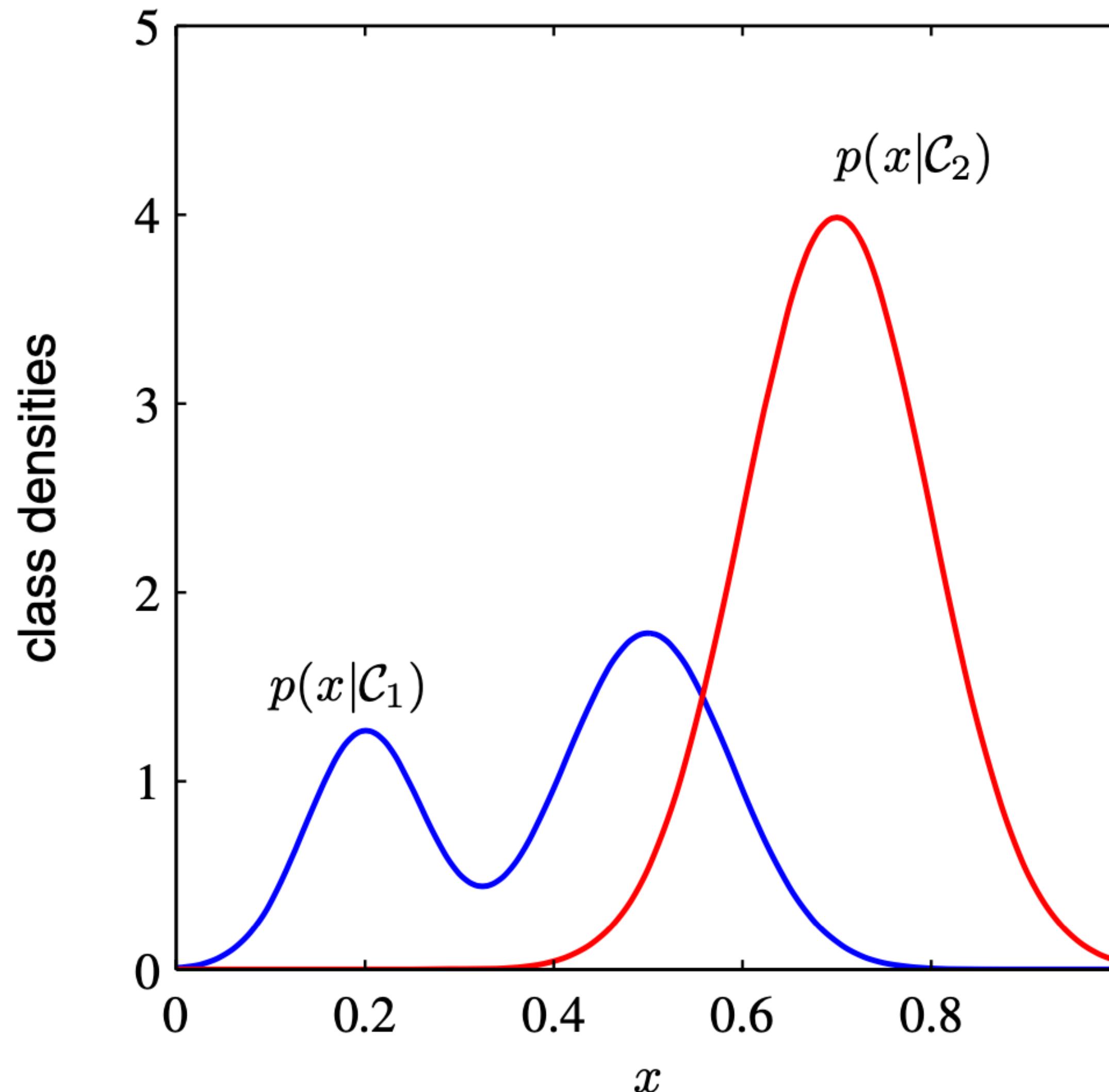
3 distinct approaches to solving decision problems

- Determine **class-conditional densities** $p(x | C_k)$ for each class C_k , then make a decision for the given x using the formula

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

- We get a **generative model**
- Determine **posterior class probabilities** $p(C_k | x)$ and predict class for an x
 - This is a **discriminative model**
- Find a **discriminant function**, $f(x)$, that maps x to class labels (e.g. a linear classifier)

Example



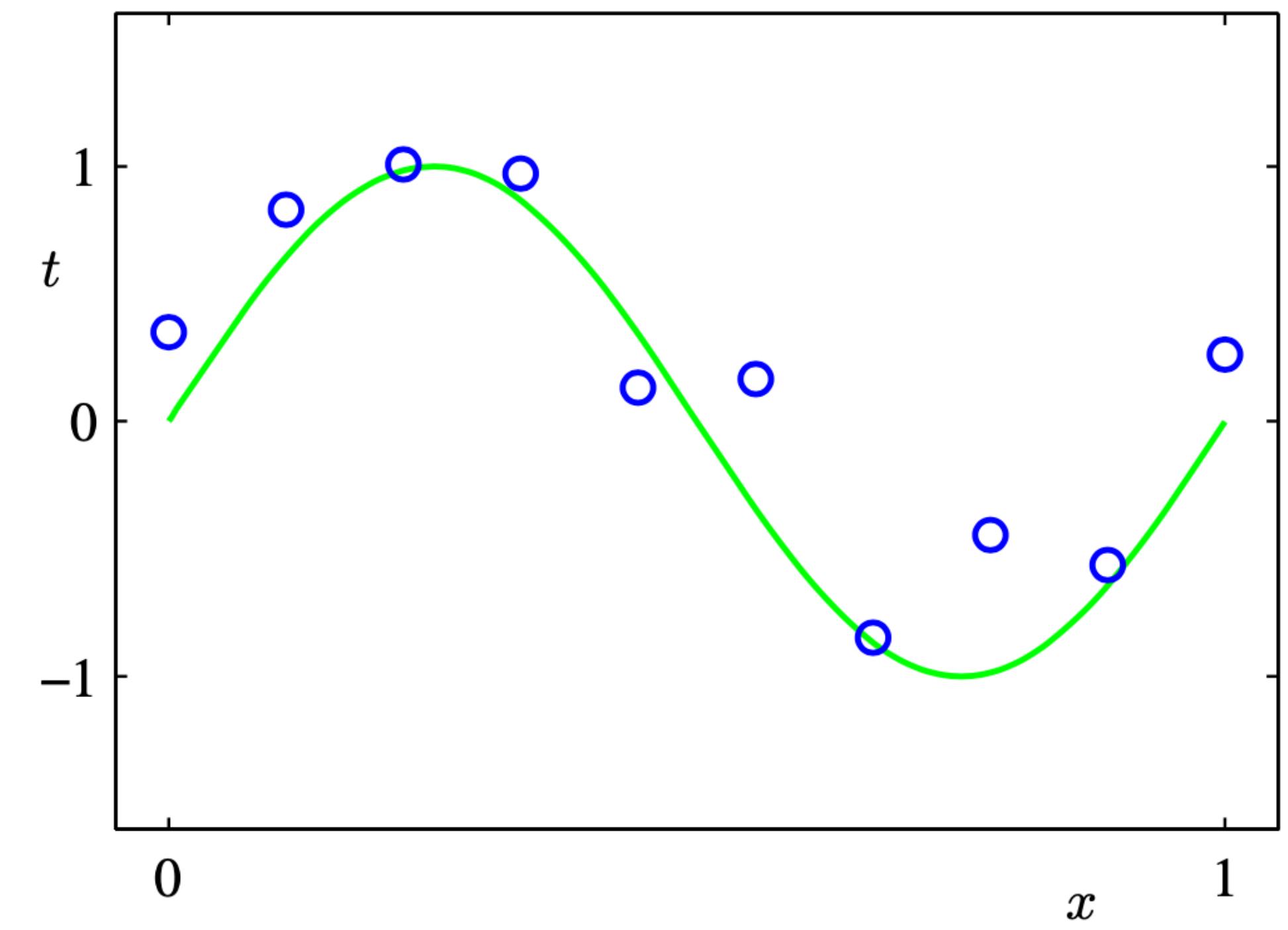
Break, 5 min.

Regression models

Loss functions for regression

Setup

- t is target
- x is input (D-dimensional vector)
- $y(x)$ estimate of t for each x
- $L(t, y(x))$ is a Loss of the estimate
 - common loss: $L(t, y(x)) = (y(x) - t)^2$
- Goal is to Minimize the expected loss

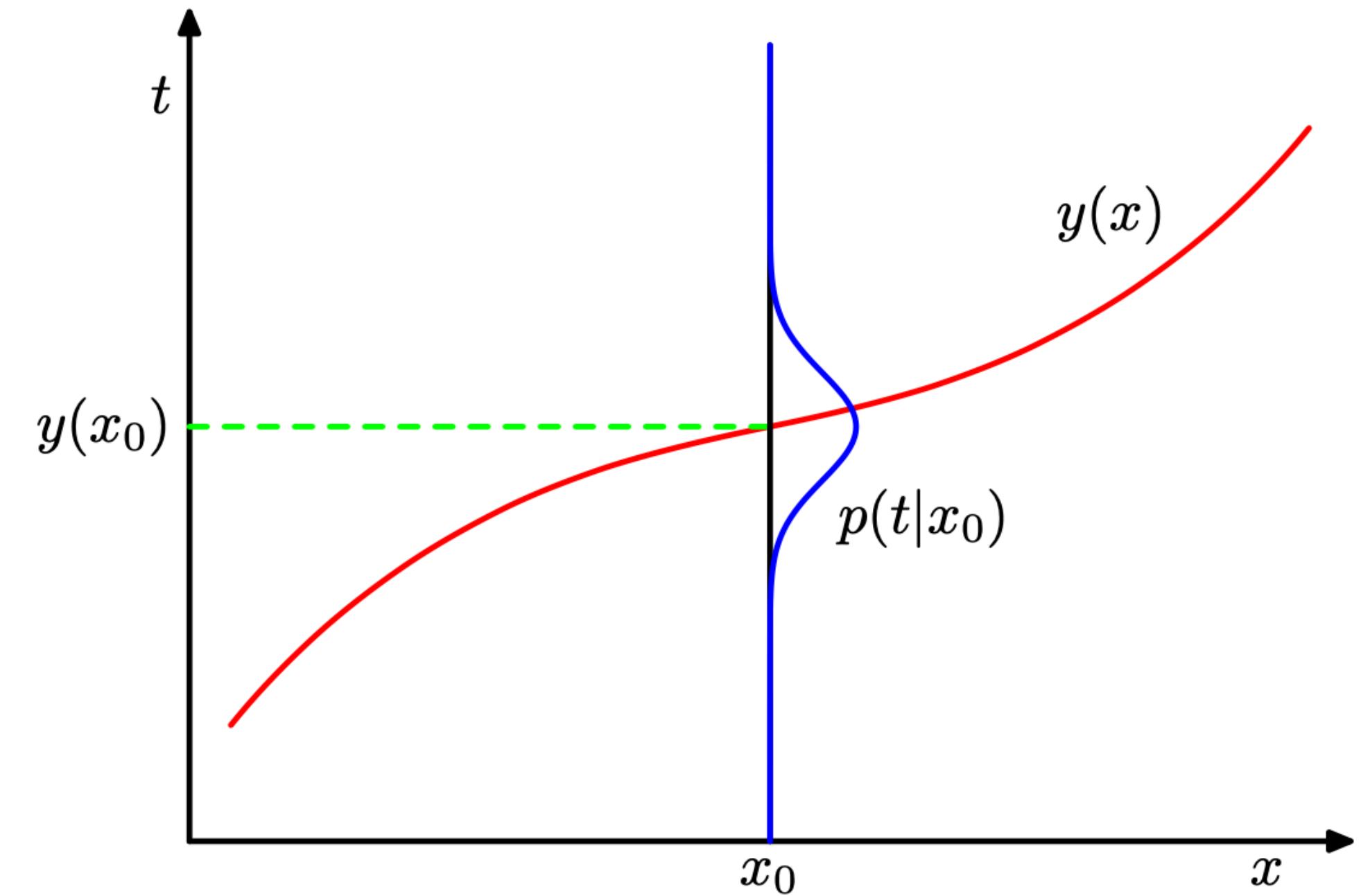


Minimizing Expected Loss

- $\mathbb{E}[L] = \int \int (y(x) - t)^2 p(x, t) dx dt$
- this is a functional of the loss function L: mean squared error for all possible x, t
- To Solve $\min_{y(x)} \mathbb{E}[L]$ $\Rightarrow \frac{\delta \mathbb{E}[L]}{\delta y(x)} = 2 \int (y(x) - t)p(x, t) dt = 0$ (from variational calculus)
- The Solution is: $y(x) = \frac{\int tp(x, t) dt}{p(x)} = \int tp(t | x) dt = \mathbb{E}_t[t | x]$
- So, the optimal regression function is the **conditional expectation** $\mathbb{E}_t[t | x]$

Regression function

- $y(x) = \mathbb{E}_t[t | x]$ – regression function
 - conditional average of t conditioned on x
- Recap:
 - Conditional expectation $\mathbb{E}_t[t | x]$
 - $\mathbb{E}_t[t | x] = \int t \frac{p(x, t)}{p(x)} dt = \int t \frac{p(t | x)p(x)}{p(x)} dt = \int tp(t | x)dt$
 - **it does not depend on t !**



But why?

Why the optimal regression function is the conditional expectation, $\mathbb{E}_t[t|x]$?

- OK, let us derive the expected loss in slightly different way
- for fixed value of x :

$$\begin{aligned}(y(x) - t)^2 &= (y(x) - \mathbb{E}[t|x] + \mathbb{E}[t|x] - t)^2 = \\ &= \{y(x) - \mathbb{E}[t|x]\}^2 + 2\{y(x) - \mathbb{E}[t|x]\}\{\mathbb{E}[t|x] - t\} + \{\mathbb{E}[t|x] - t\}^2\end{aligned}$$

- We put this into the formula below and integrate over t

$$\mathbb{E}[L] = \iint (y(x) - t)^2 p(x, t) dx dt$$

- Hint: the cross-term should disappear

A FEW
MOMENTS
LATER

Great!

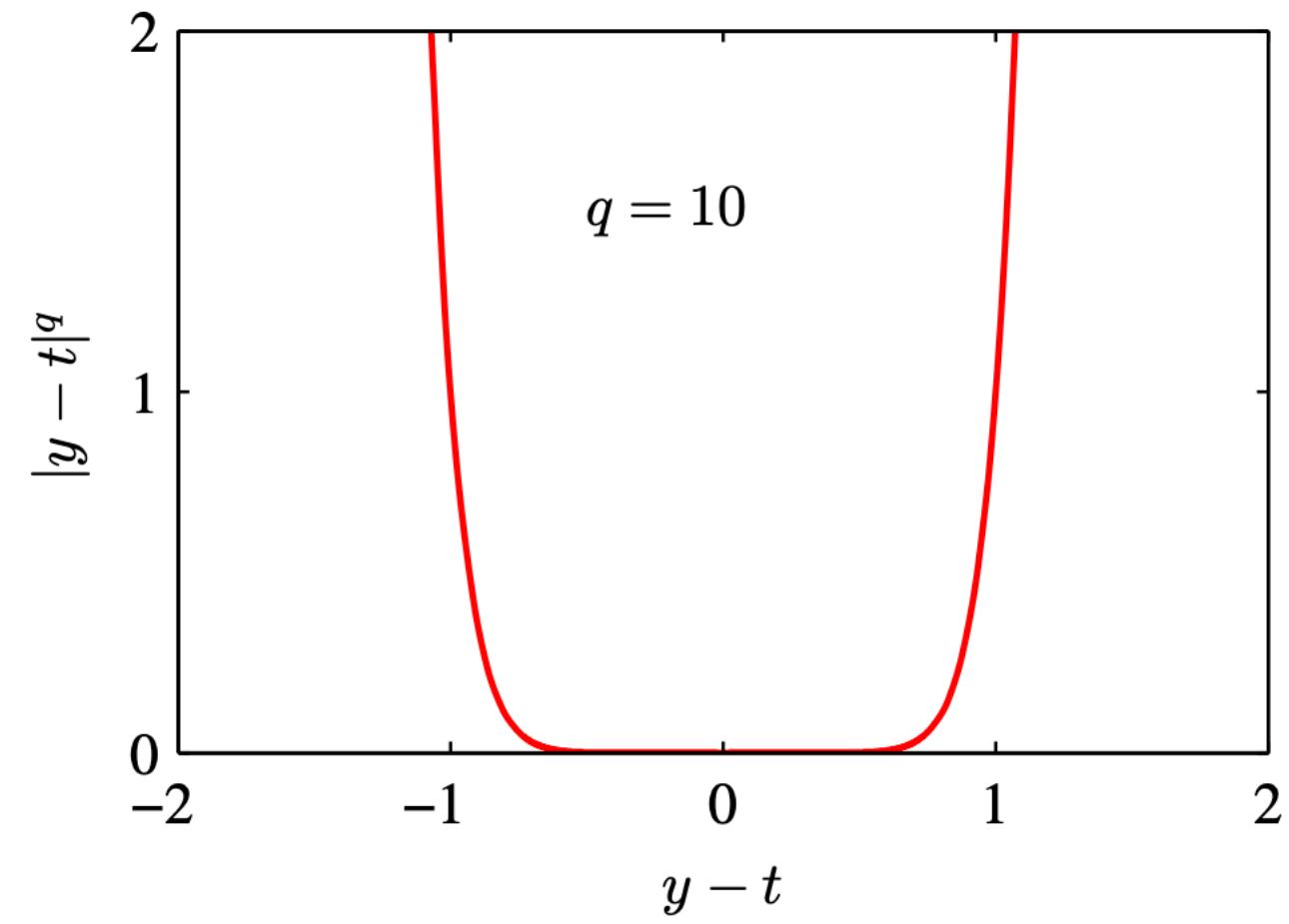
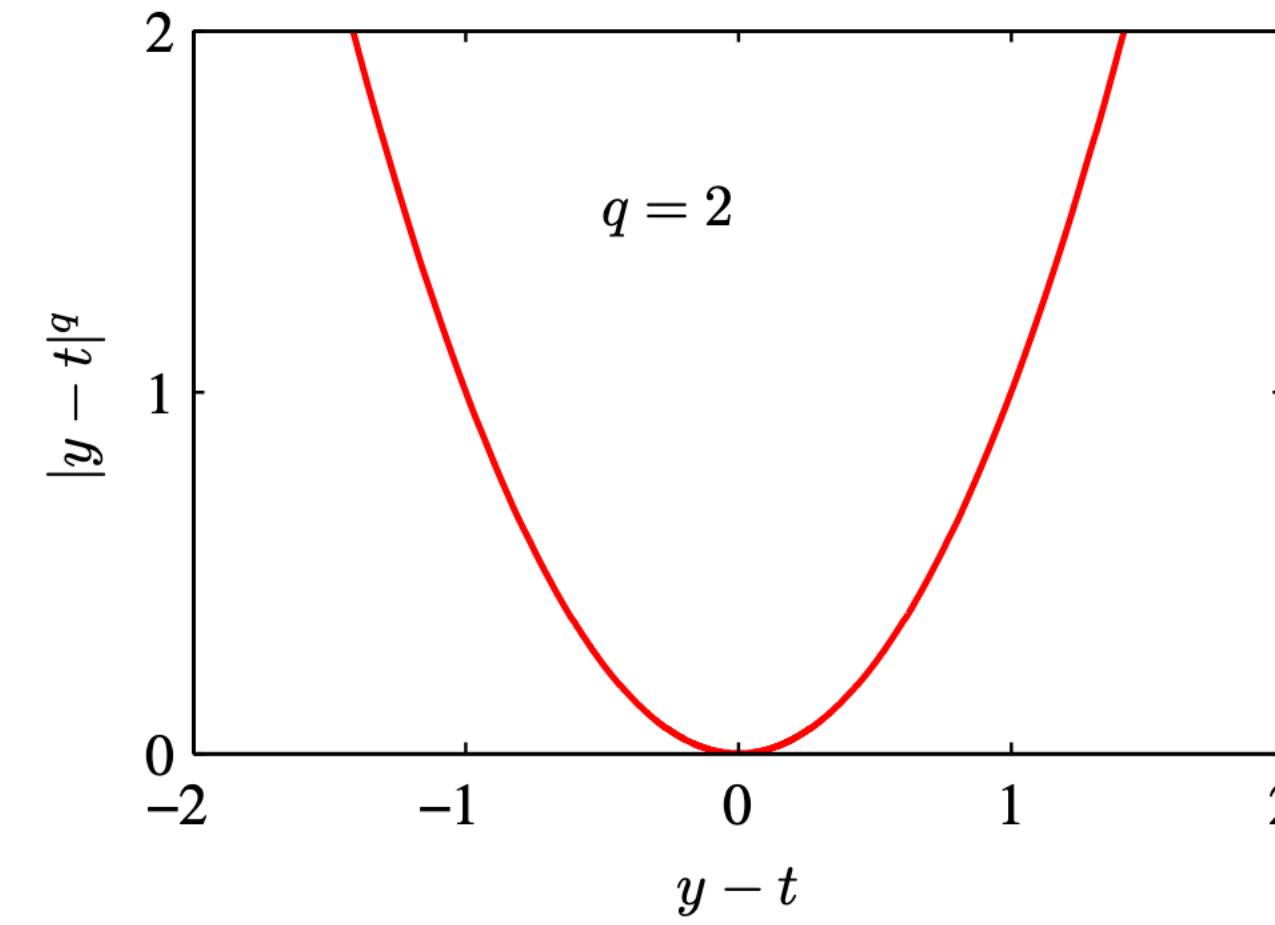
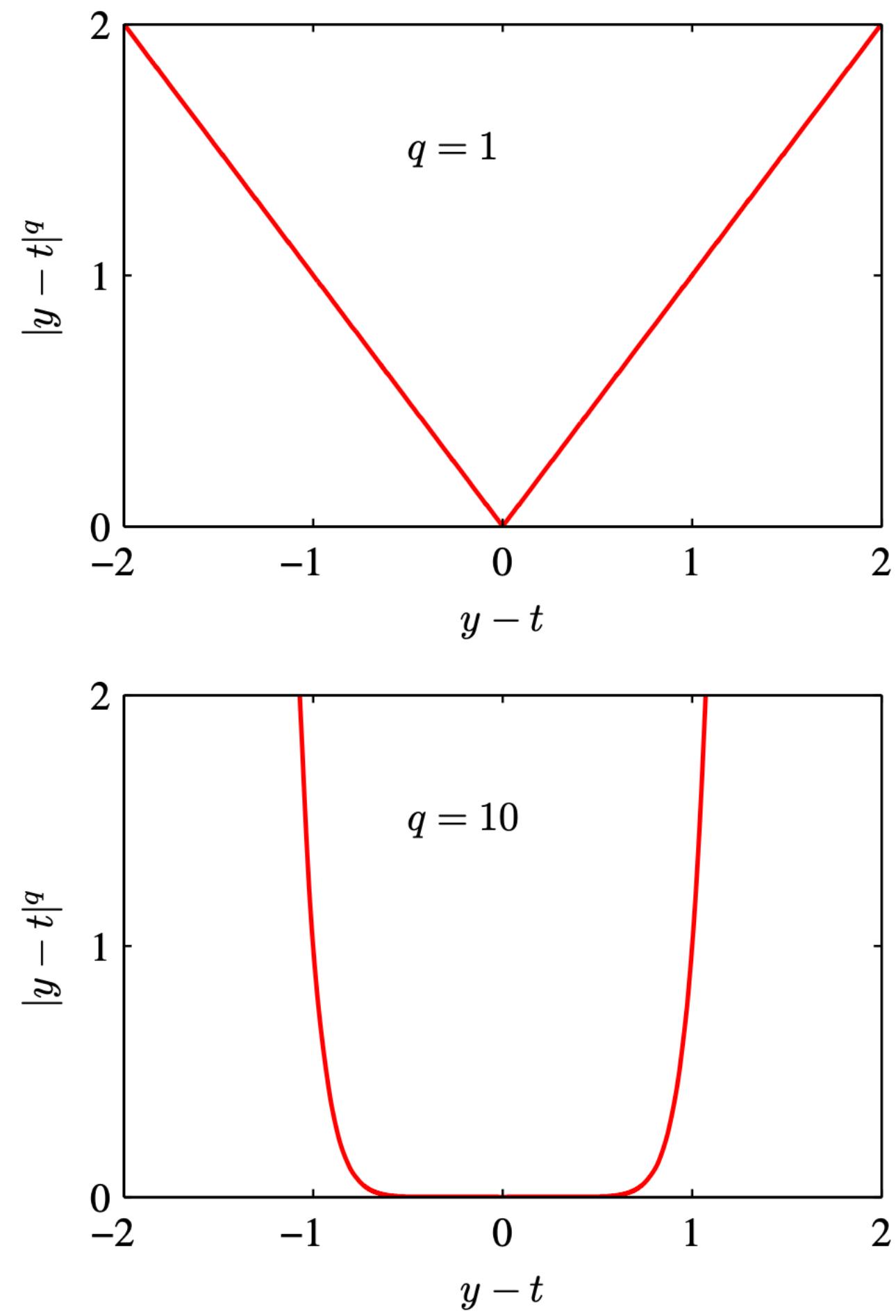
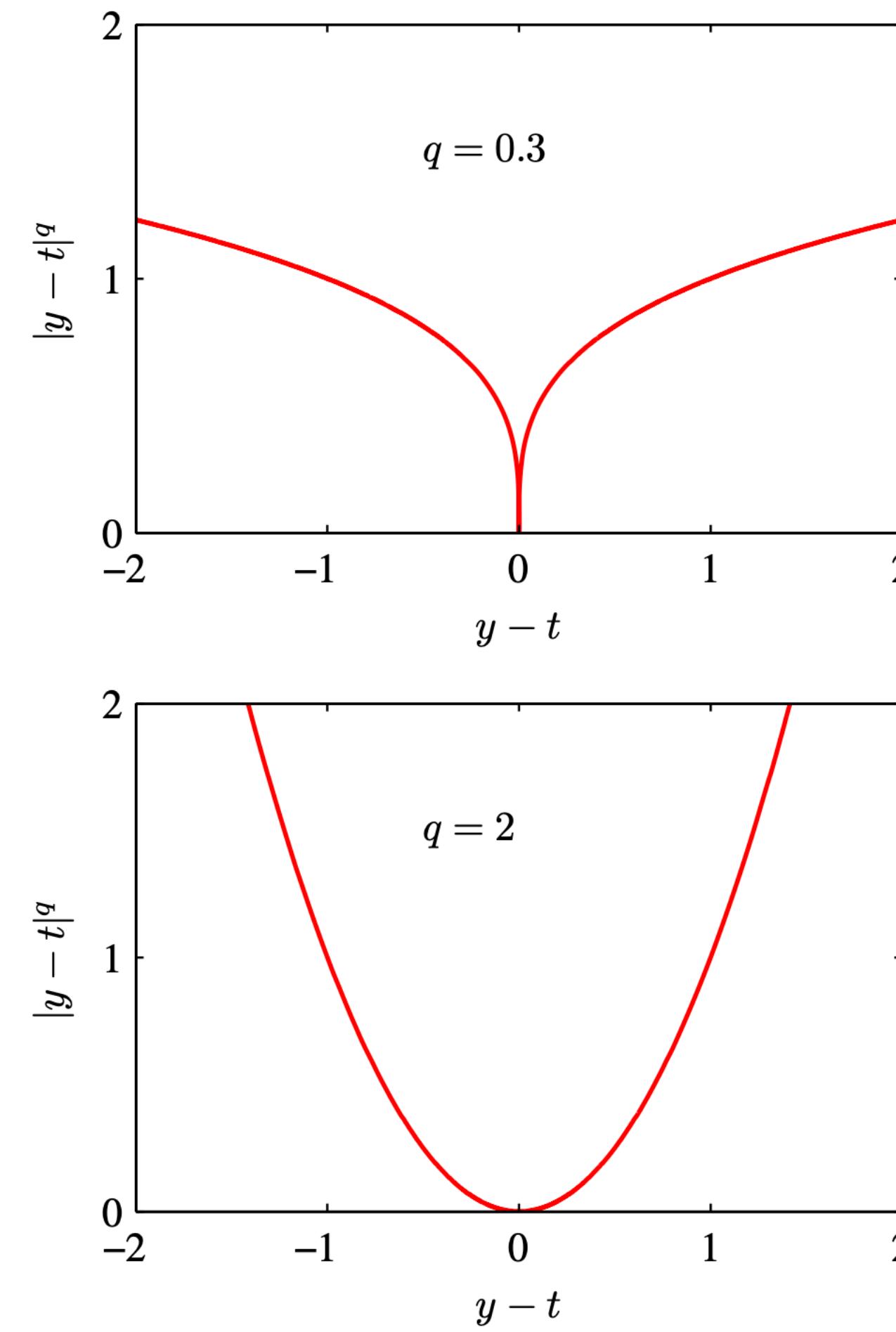
$$\begin{aligned}\bullet \quad \mathbb{E}[L] &= \iint (y(x) - t)^2 p(x, t) dx dt = \\ &= \int \{y(x) - \mathbb{E}[t|x]\}^2 p(x) dx + \int \{\mathbb{E}[t|x] - t\}^2 p(x, t) dx dt\end{aligned}$$

Therefore, the minimum of $\mathbb{E}[L]$ corresponds to $y(x) = \mathbb{E}_t[t|x]$ when the first term vanishes

The second term denotes an irreducible noise, or the variance of the distribution of target t , averaged over (all possible values of) x

Minkowski loss

- $\mathbb{E}[L_q] = \iint |y(x) - t|^q p(x, t) dx dt$
- $q=2$: Conditional **Mean**
- $q=1$: Conditional **Median**
- $q \rightarrow \infty$: Conditional **Mode**



3 Approaches to Regression

1. First solve the inference problem of determining the joint density $p(x, t)$
2. First solve the inference problem of determining the conditional density $p(t|x)$
3. Find a regression function $y(x)$ directly from the training data

Linear regression

Setup

linear regression

- t is target
- \mathbf{x} is input (D-dimensional vector)
- $y(\mathbf{x})$ estimate of t for each \mathbf{x}
- loss: $L(t, y(\mathbf{x})) = (y(\mathbf{x}) - t)^2$

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \dots + w_D x_D$$

In general (linear in w):

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

Basis functions

a.k.a. features

- Polynomial:

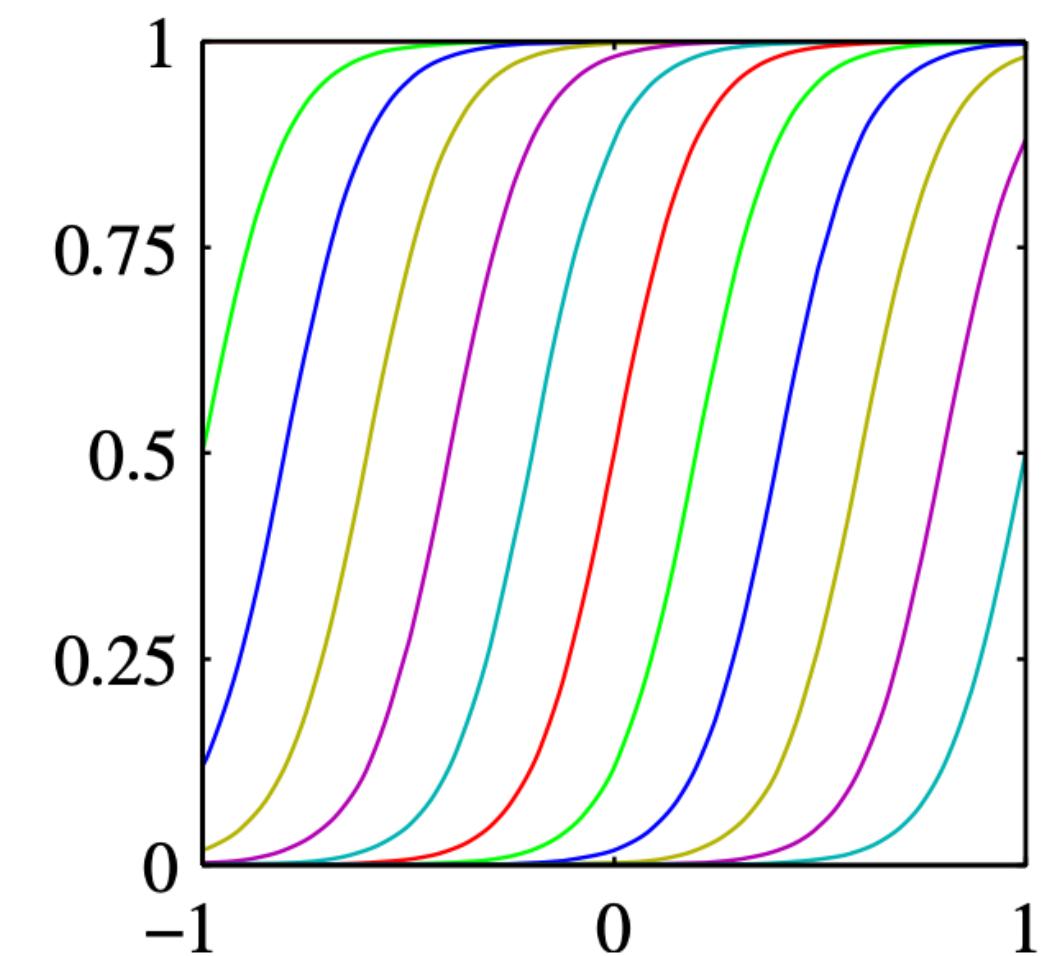
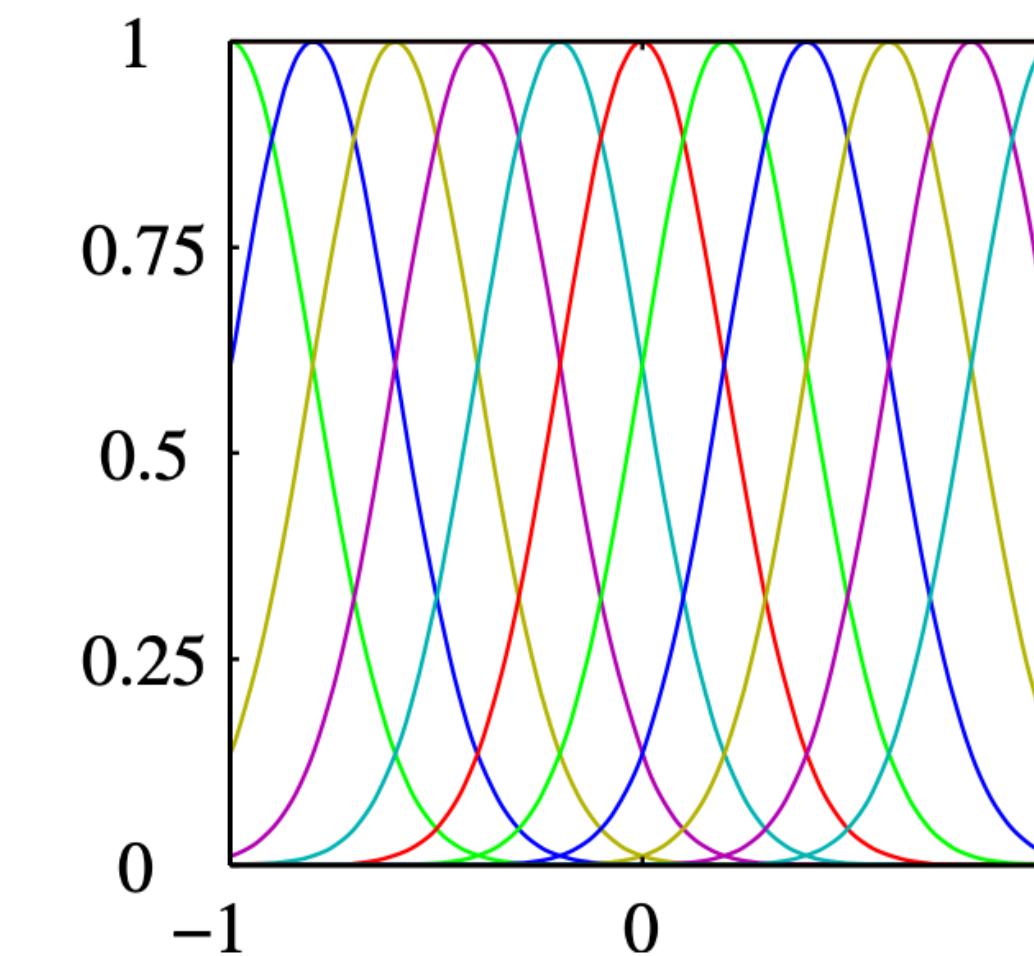
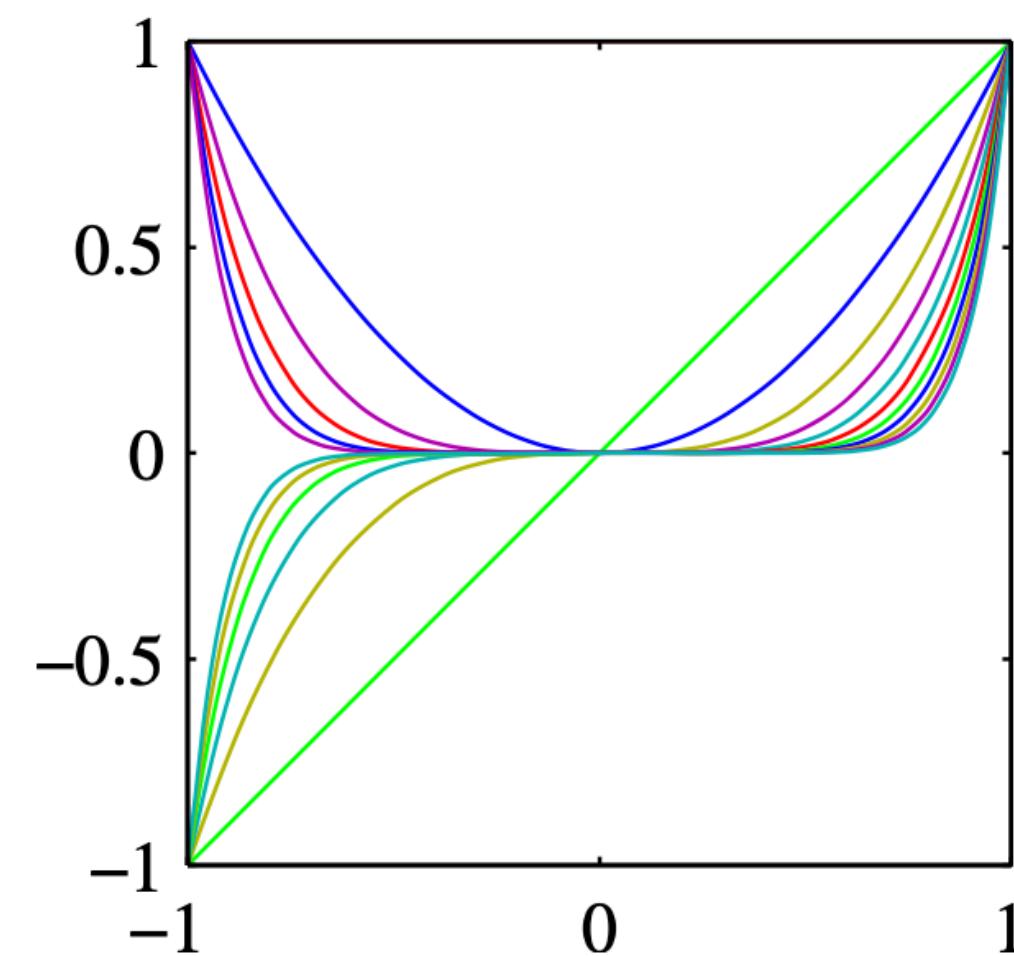
$$\phi_j(x) = x^j$$

- Gaussian:

$$\phi_j(x) = \exp\left\{-\frac{(x - \mu_j)^2}{2s}\right\}$$

- Sigmoidal:

$$\phi_j(x) = \sigma\left\{\frac{x - \mu_j}{s}\right\}$$



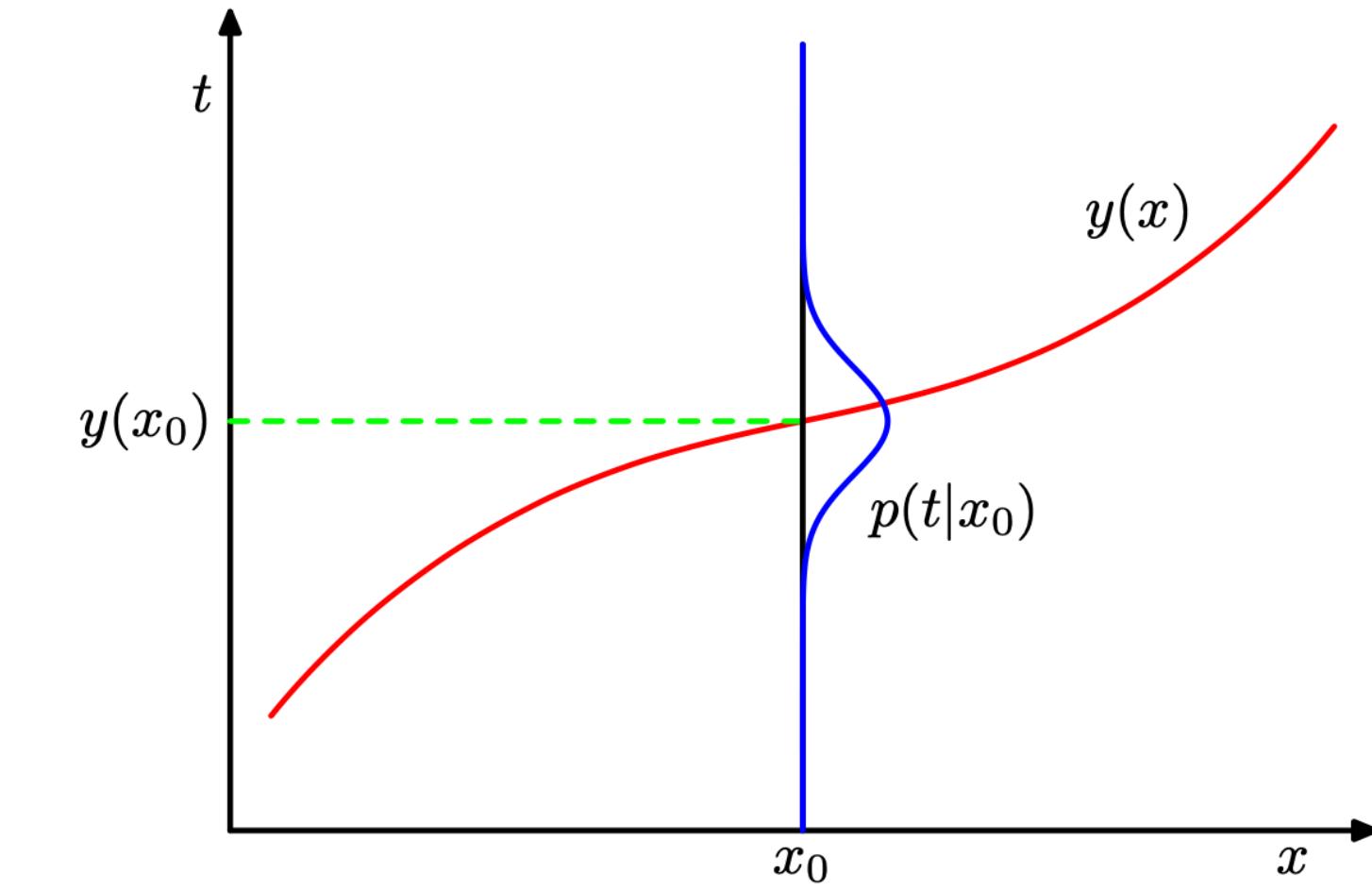
$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

Maximum likelihood and least squares

Build

- $t = y(x, w) + \text{gaussian noise; (standard normal distr. with inverse variance } \beta)$
- $p(t|x, w, \beta) = \mathcal{N}(t|y(x, w), \beta^{-1})$
- likelihood function $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n|\mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1})$
- log-likelihood: $\ln p(\mathbf{t}|\mathbf{w}, \beta) = \sum_{n=1}^N \ln \mathcal{N}(t_n|\mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1})$
 $= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w})$
- where

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2$$



A LITTLE LONGER

THAN A FEW

MINUTES LATER

Solution

- normal equations for the least squares problem

$$\mathbf{w}_{\text{ML}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$$

- Φ is a design matrix ($N \times M$)

$$\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

Assumptions of linear regression

- Here we consider the simplest form:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1x_1 + \dots + w_Dx_D$$

- **Linearity of the data.** The relationship between the predictor (x) and the outcome (y) is assumed to be linear.
- **Normality of residuals.** The residual errors are assumed to be normally distributed.
- **Homogeneity of residuals variance.** The residuals are assumed to have a constant variance (homoscedasticity)
- **Independence** of residuals error terms.

Potential problems

- **Non-linearity** of the outcome - predictor relationships
- **Heteroscedasticity:** Non-constant variance of error terms.
- **Presence of influential values** in the data that can be:
 - **Outliers:** extreme values in the outcome (y) variable
 - **High-leverage points:** extreme values in the predictors (x) variable

Break, 5 min.

The Bias-Variance Decomposition and Trade-off

Recall the Bias and Variance

For some estimator $\hat{\theta}$,

- $\hat{\theta}$ is an estimator of the parameter θ
- Bias:
 - $Bias(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$
- Variance:
 - $V[\hat{\theta}] = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$

Setup

- t is target
 - \mathbf{x} is input (D-dim. vector)
 - $y(\mathbf{x})$ is estimate of t for each \mathbf{x}
 - Squared loss: $L(t, y(\mathbf{x})) = (y(\mathbf{x}) - t)^2$
 - For the squared loss function, the optimal prediction is given by the conditional expectation
- $$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int tp(t|\mathbf{x}) dt.$$

We are minimizing $E[L]$:

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

Goal

approximate the $h(\mathbf{x})$ using the finite dataset,

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

- To minimize the first term we
 - need to find $y(\mathbf{x})$
 - be able to assess its quality,
 - e.g. how good is it on average over possible ensambles of datasets

Build

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

- Given a dataset, \mathcal{D}
- Our interest is this quantity, $\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2$.
- which can be expanded:
$$\begin{aligned} & \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\ &= \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\ &\quad + 2\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}. \end{aligned}$$
- Expectation $\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]$ is calculated via independent sampling of N datasets

Punchline

$$\begin{aligned}\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2 &= \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\ &\quad + 2\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}.\end{aligned}$$

- take the expectation of this expression with respect to \mathcal{D} and note that the final term will vanish

$$\begin{aligned}\mathbb{E}_{\mathcal{D}} [\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2] &= \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2}_{\text{(bias)}^2} + \underbrace{\mathbb{E}_{\mathcal{D}} [\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2]}_{\text{variance}}\end{aligned}$$

Expected loss decomposition

expected loss = (bias)² + variance + noise

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x}) d\mathbf{x}$$

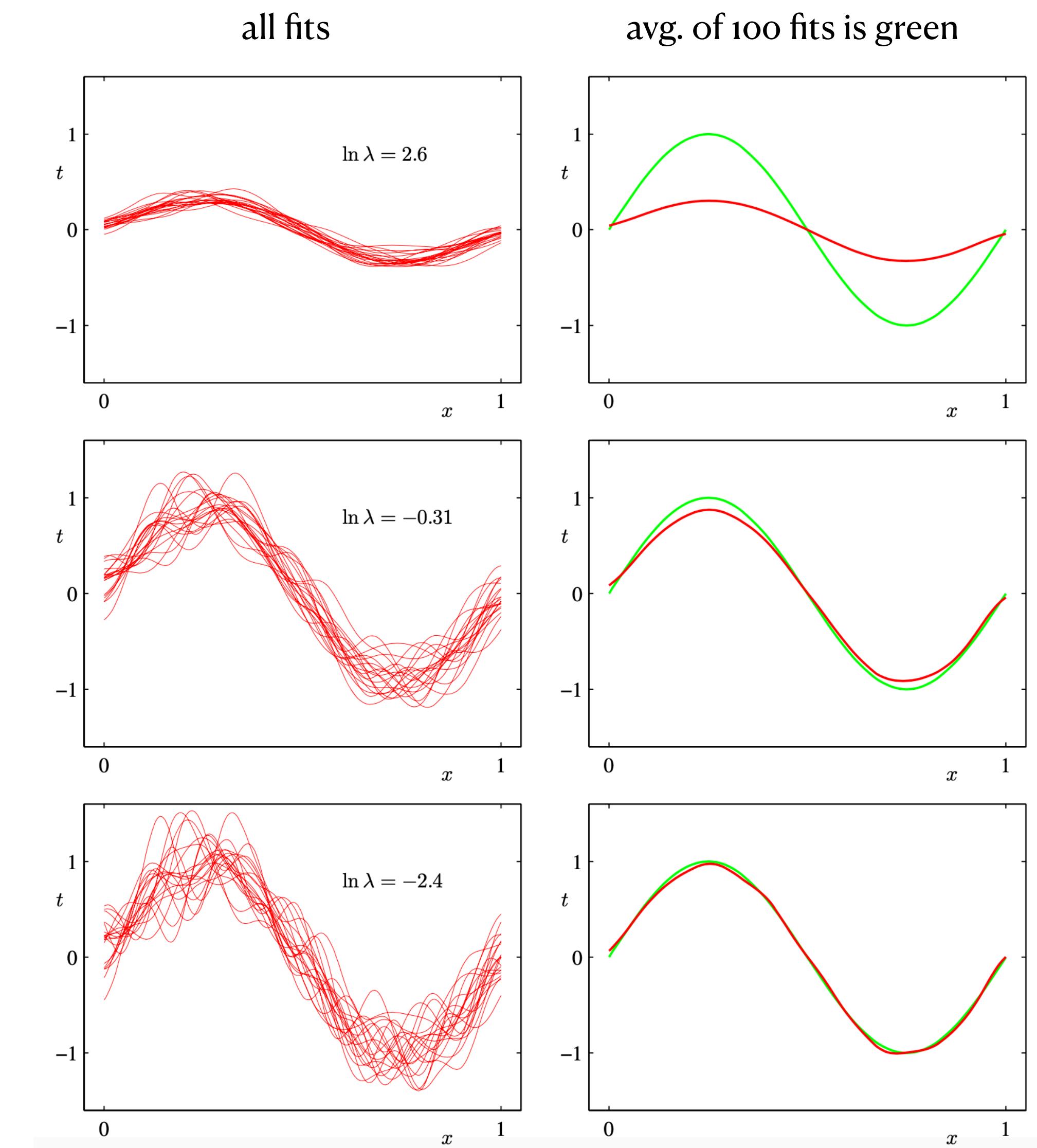
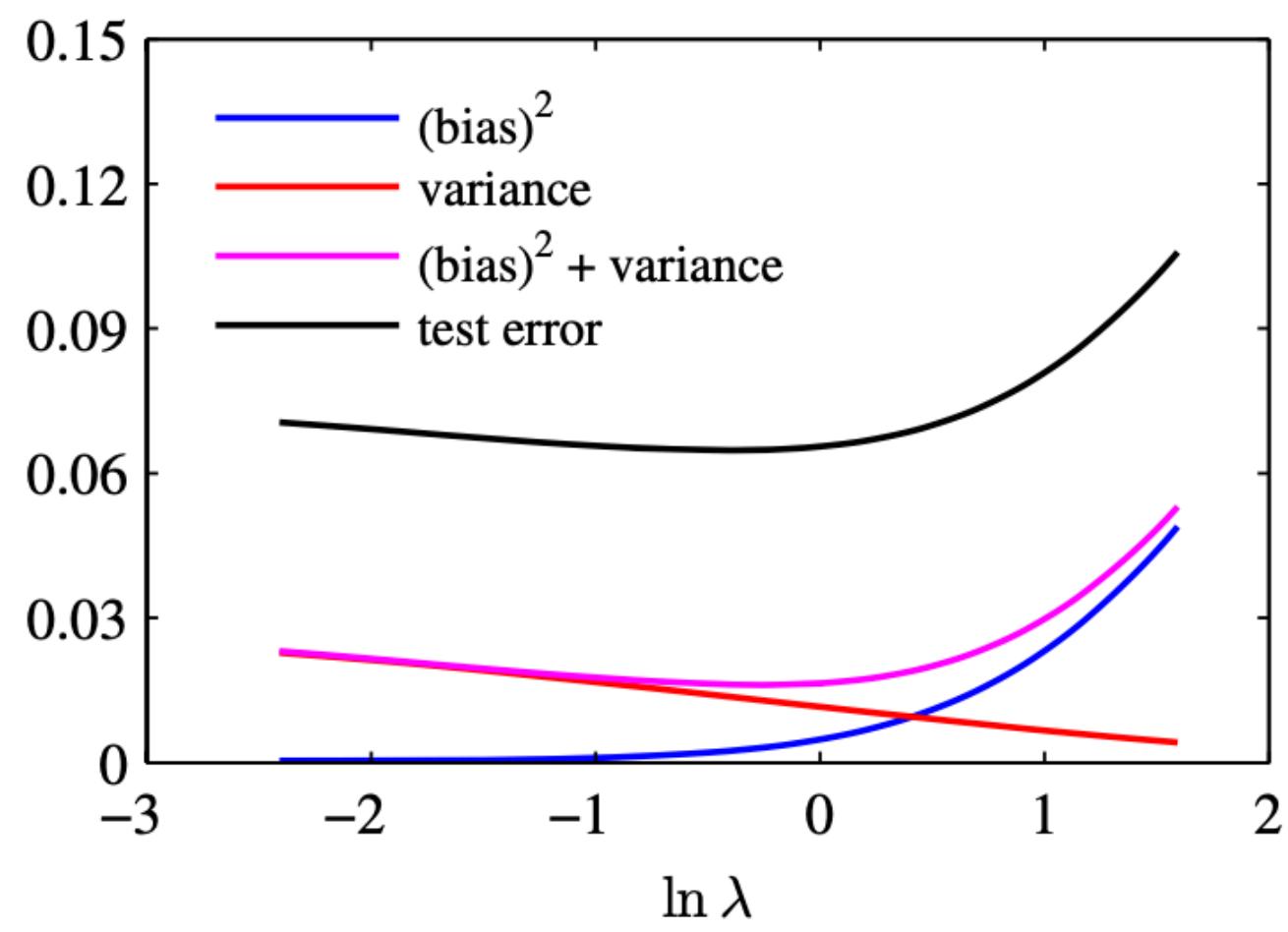
$$\text{variance} = \int \mathbb{E}_{\mathcal{D}} [\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2] p(\mathbf{x}) d\mathbf{x}$$

$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt$$

Example: Regularized least squares

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

- $L = 100$ datasets
- with $N = 25$ training examples each
- 24 Gaussian basis functions
- λ is a regularizer coefficient
- Analysis:



Summary

- Regression Models:
 - Loss function for regression
 - Linear Regression
 - The Bias-Variance Decomposition
- Fill the feedback form with your questions, please