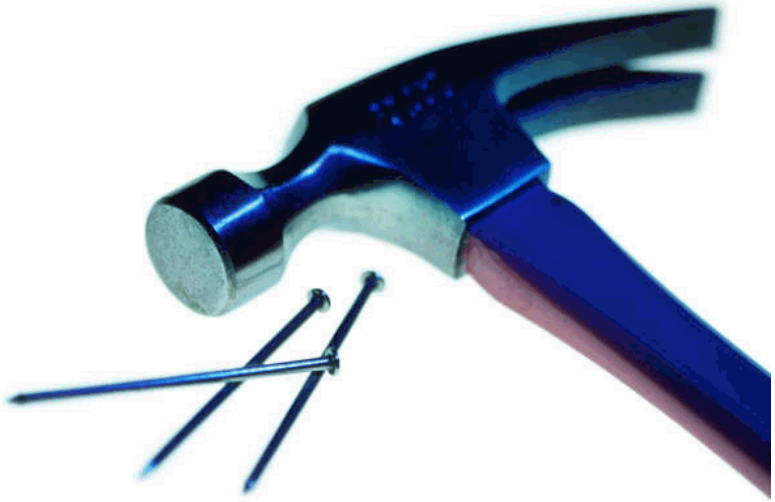


Machine Learning: The Optimization Perspective

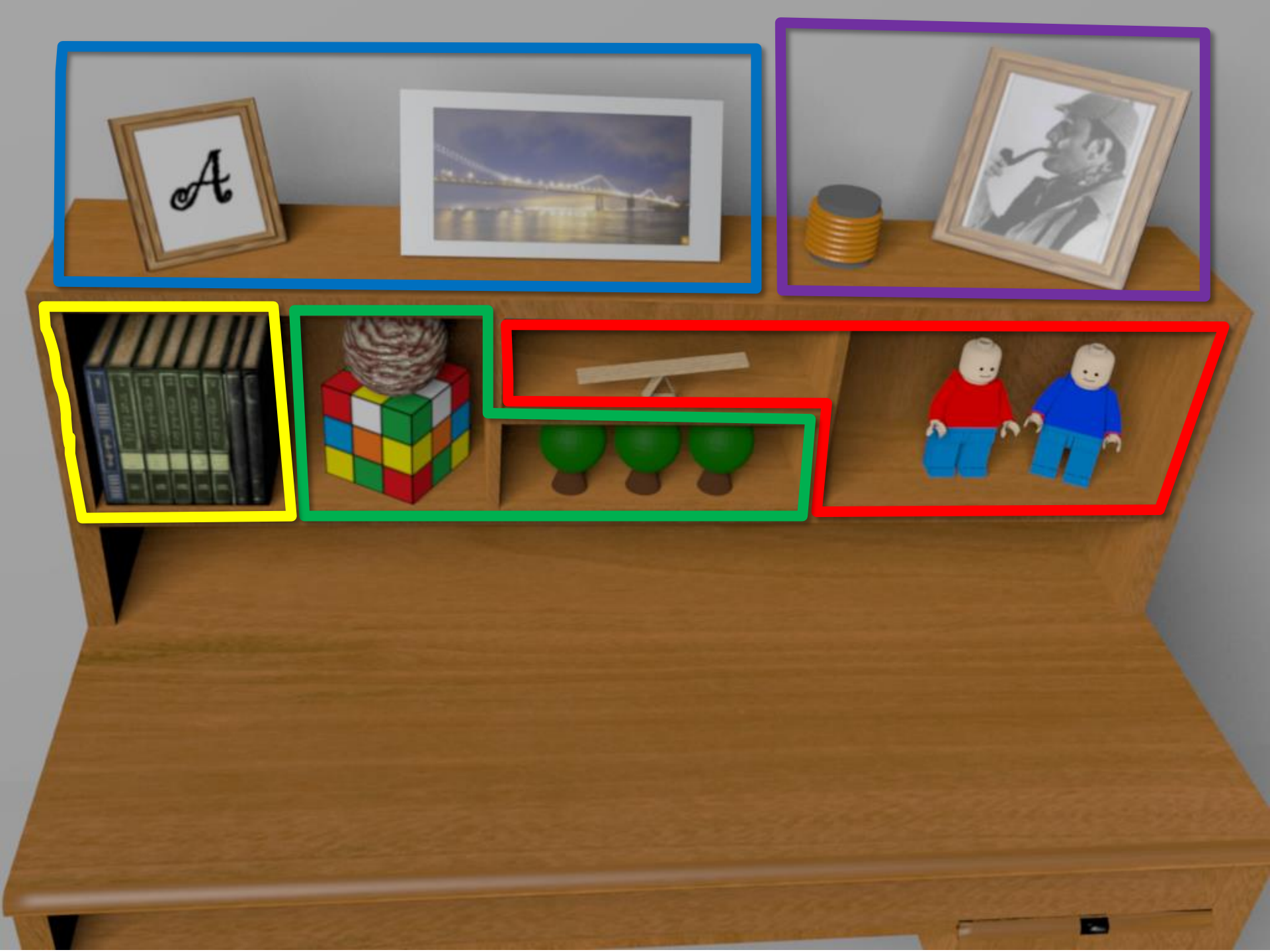
Konstantin Tretyakov
http://kt.era.ee

IFI Summer School 2014

STACC Software Technology and Applications Competence Center

BIT

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632



So far…

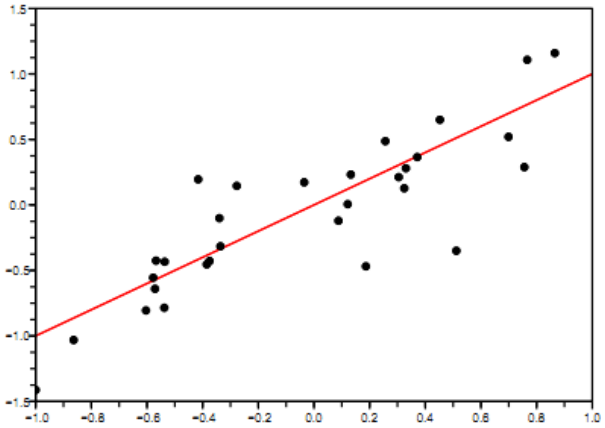
- Machine learning is important and interesting
- The general concept:

Fitting models to data

So far…

- Machine learning is important and interesting
- The general concept:

Searching for the best model fitting to data

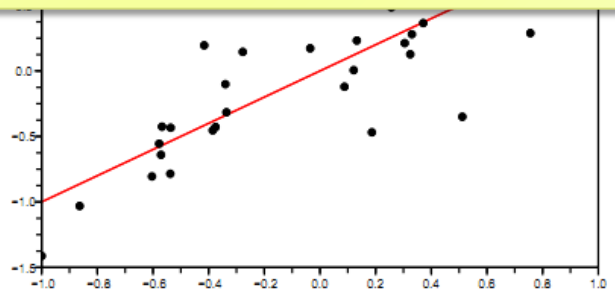


So far…

- Machine learning is important and interesting
- The g

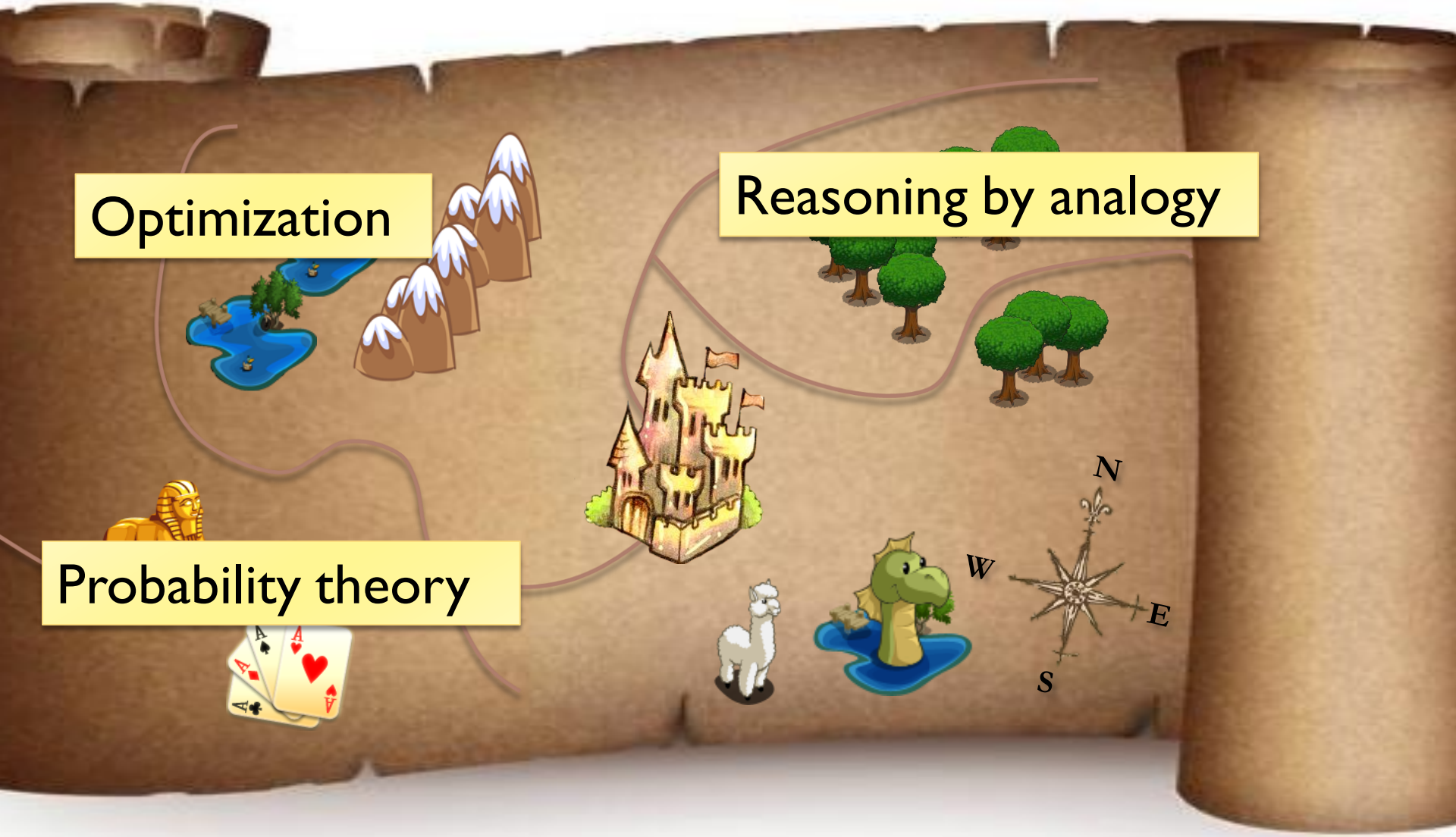
Optimization

Probability Theory

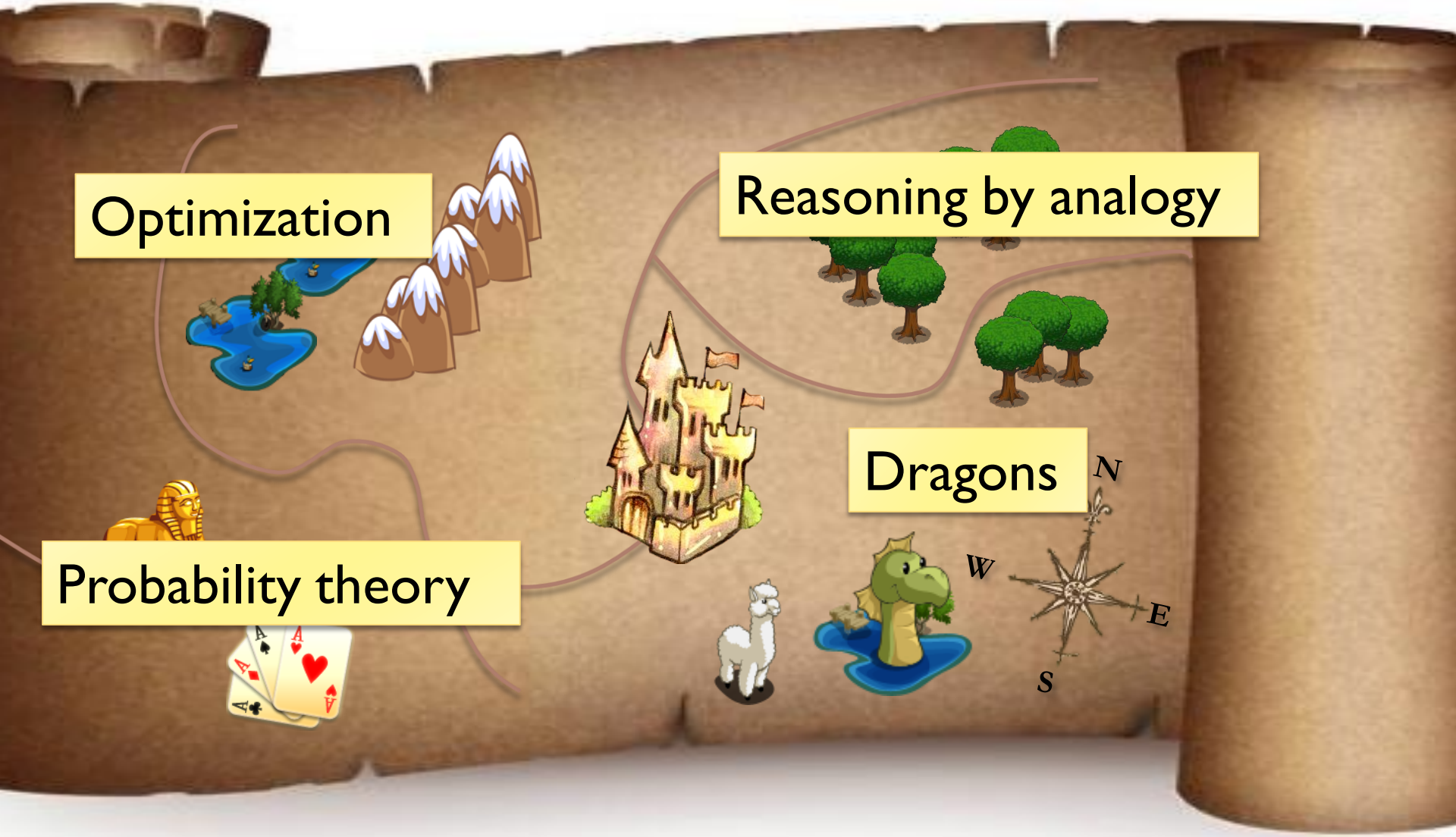


The Land of Machine Learning

The Land of Machine Learning



The Land of Machine Learning



What do you need to know about optimization?

What do you need to know about optimization?

1. Optimization is **important**
2. Optimization is **possible**

What do you need to know about optimization?

1. Optimization is **important**
2. Optimization is **possible***

* Basic **techniques**

- Constrained / Unconstrained
- Analytic / Iterative
- Continuous / Discrete

Special cases of optimization

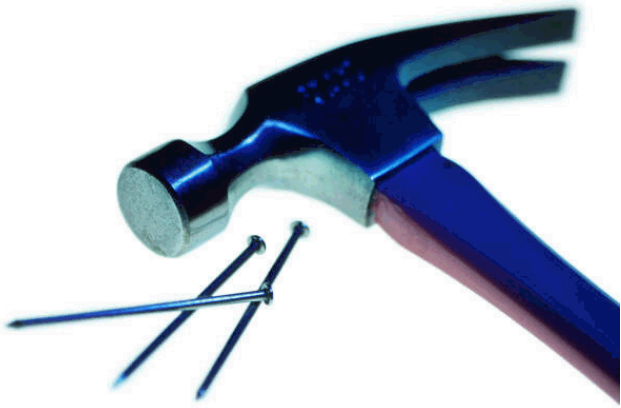
- Machine learning
- …

Special cases of optimization

- Machine learning
- Algorithms and data structures
- General problem-solving
- Management and decision-making

Special cases of optimization

- Machine learning
- Algorithms and data structures
- General problem-solving
- Management and decision-making
- Evolution
- *The Meaning of Life?*



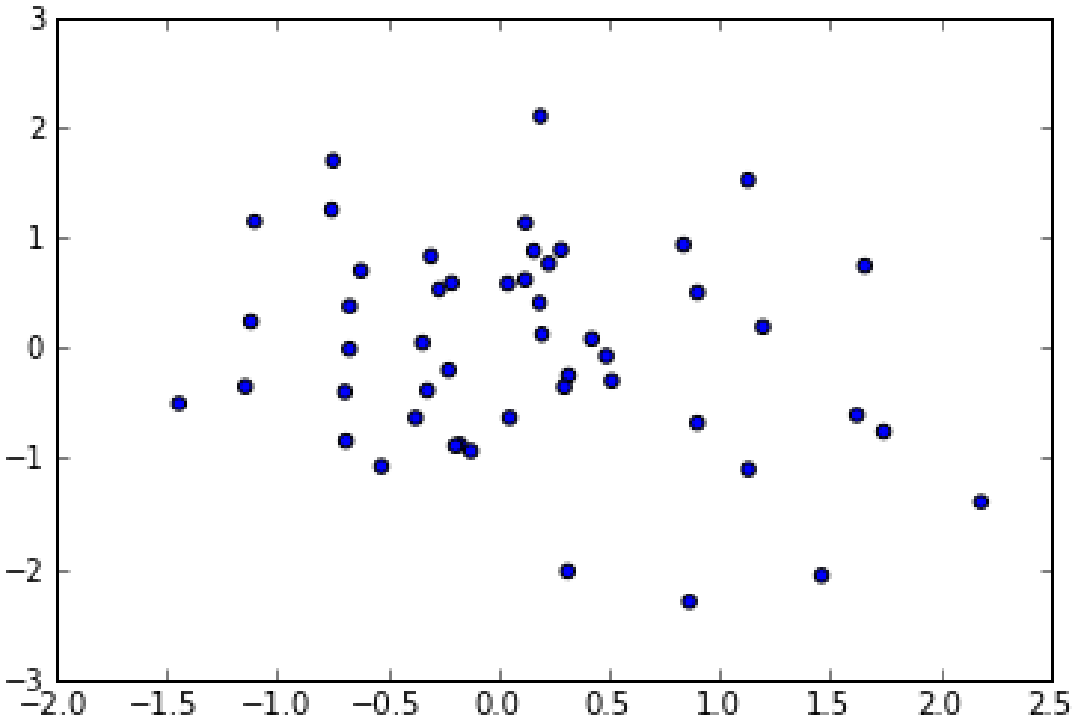
Example

Problem. Given a dataset $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbf{R}^m$, find $\mathbf{w}$, that minimizes

$$f(\mathbf{w}) = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{w}\|^2$$

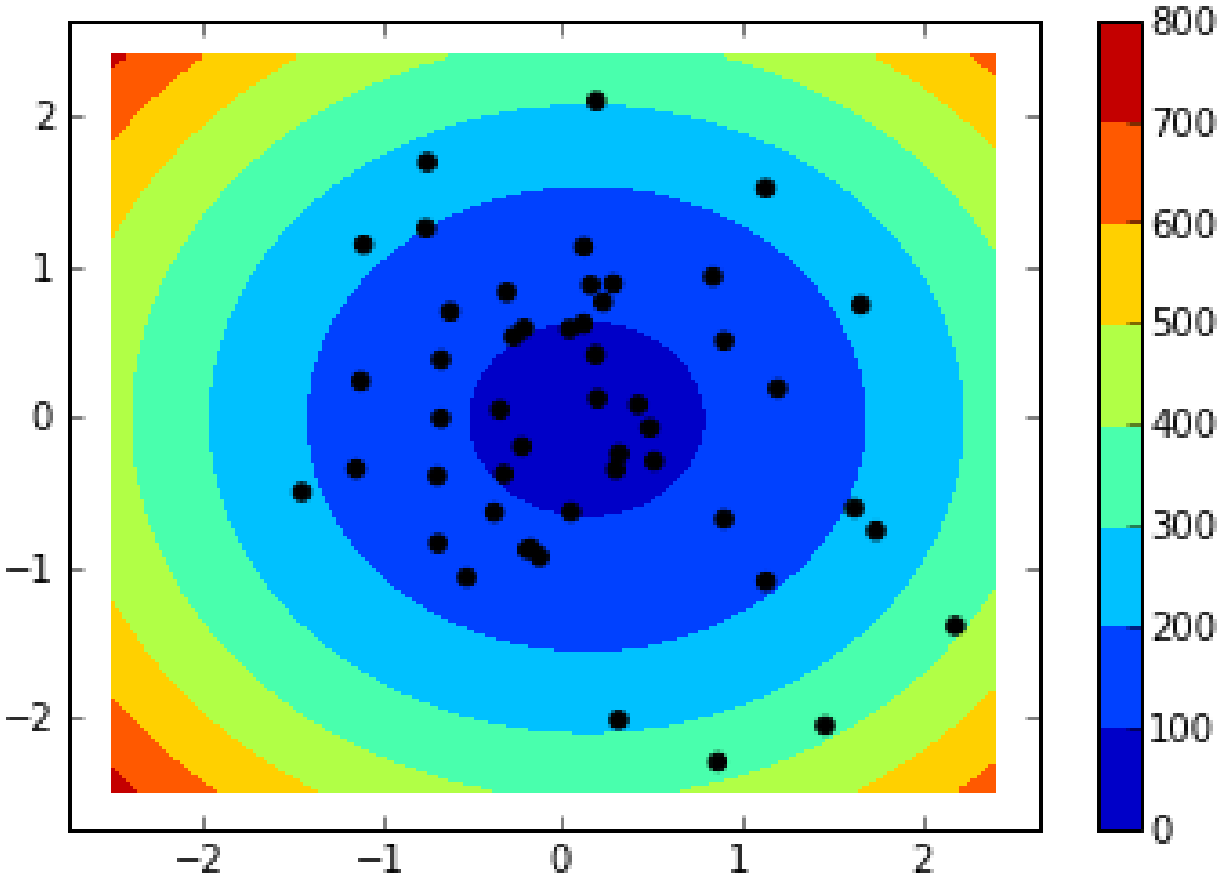
Propose an *analytical* as well as an *iterative* solution.

Example



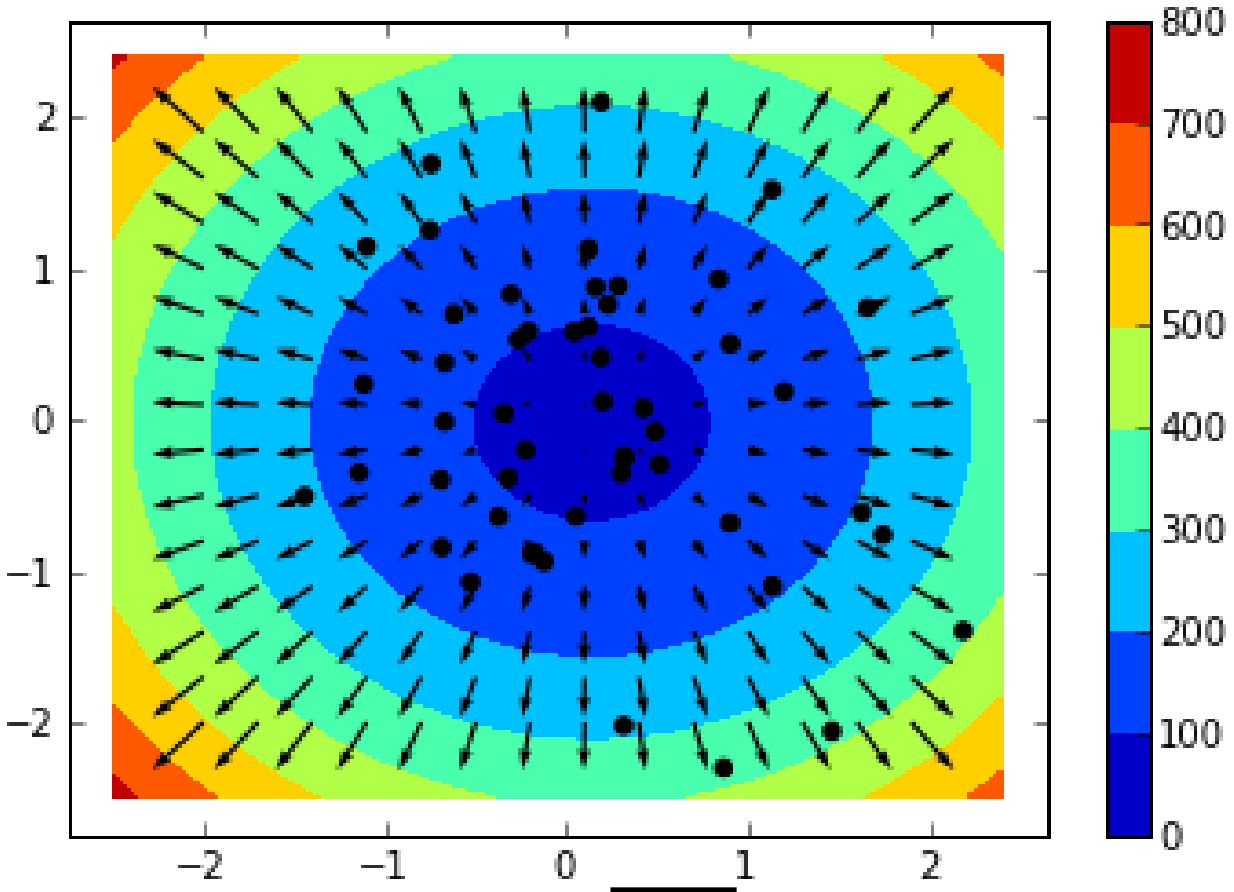
$$x_1, x_2, \ldots, x_{50}$$

Example



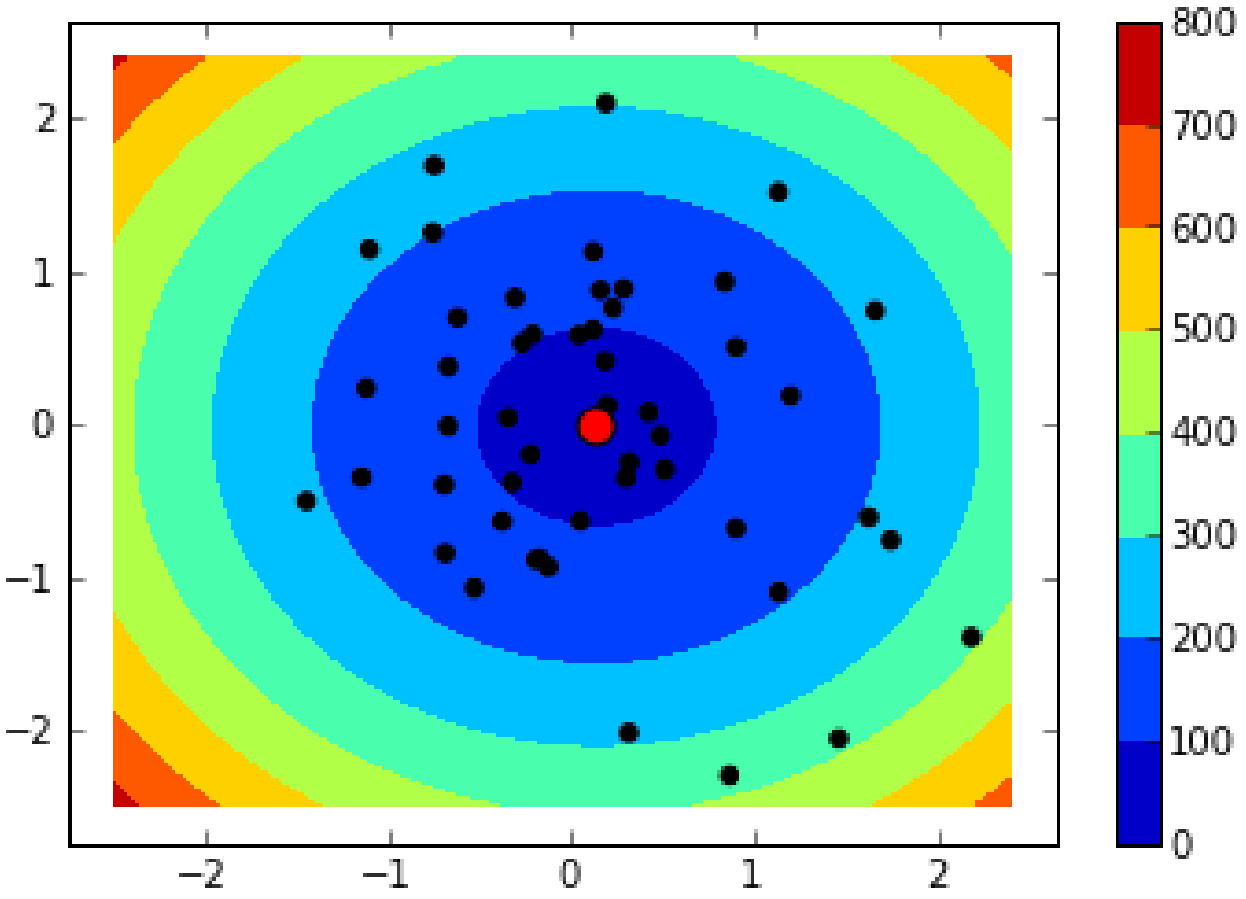
$$f(\boldsymbol{w}) = \sum_{\boldsymbol{i}} \|\boldsymbol{x_i} - \boldsymbol{w}\|^2$$

Example



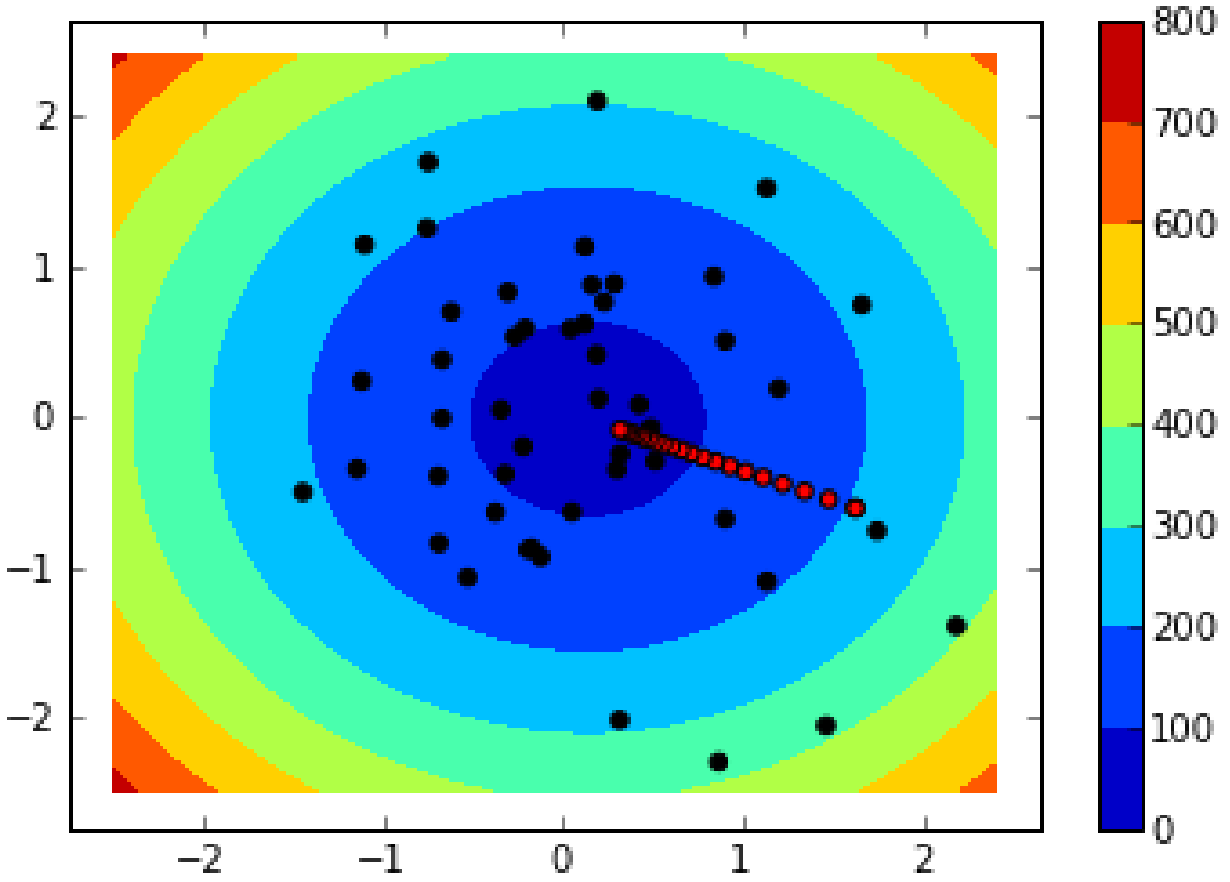
$$\nabla f(\boldsymbol{w}) = -\sum_{\boldsymbol{i}} 2(\boldsymbol{x_i} - \boldsymbol{w})$$

Example



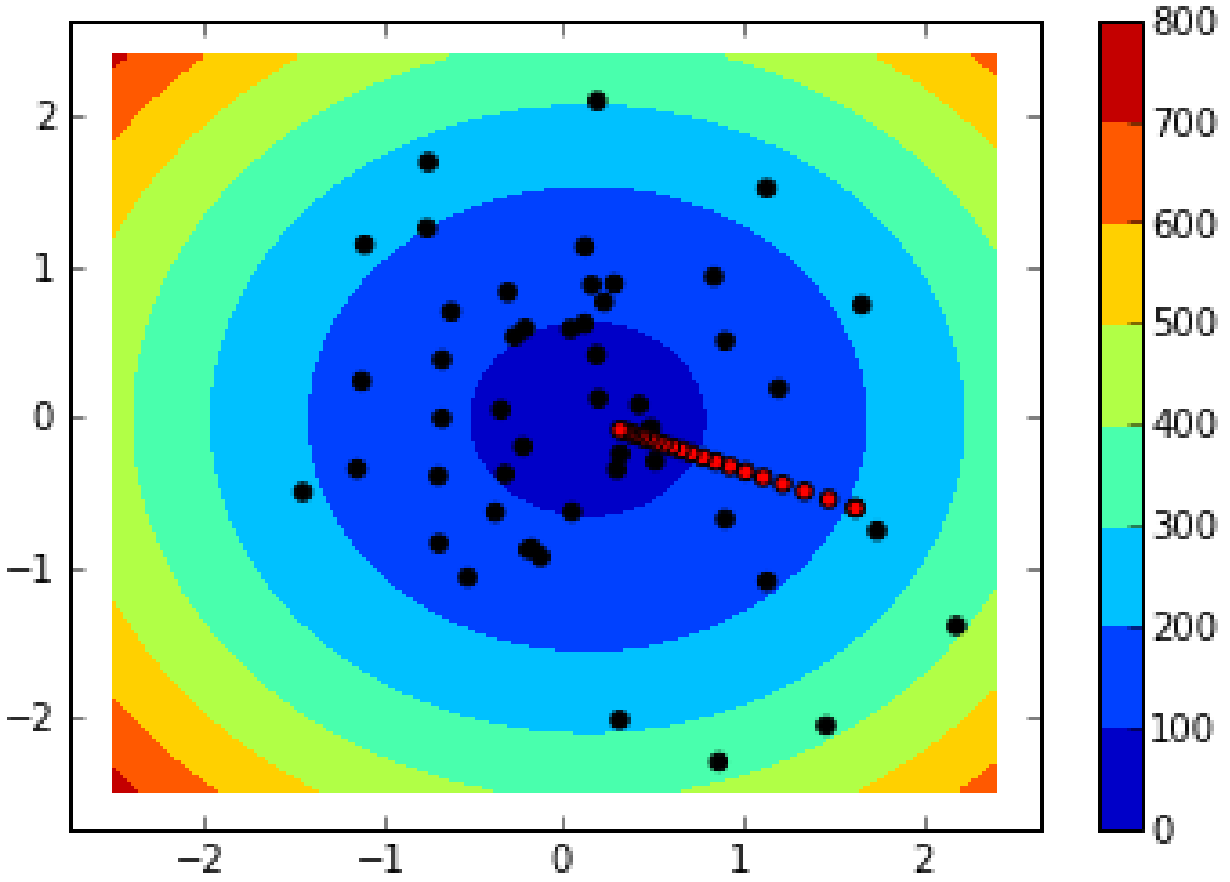
$$\nabla f(\boldsymbol{w}) = \mathbf{0}$$

Example



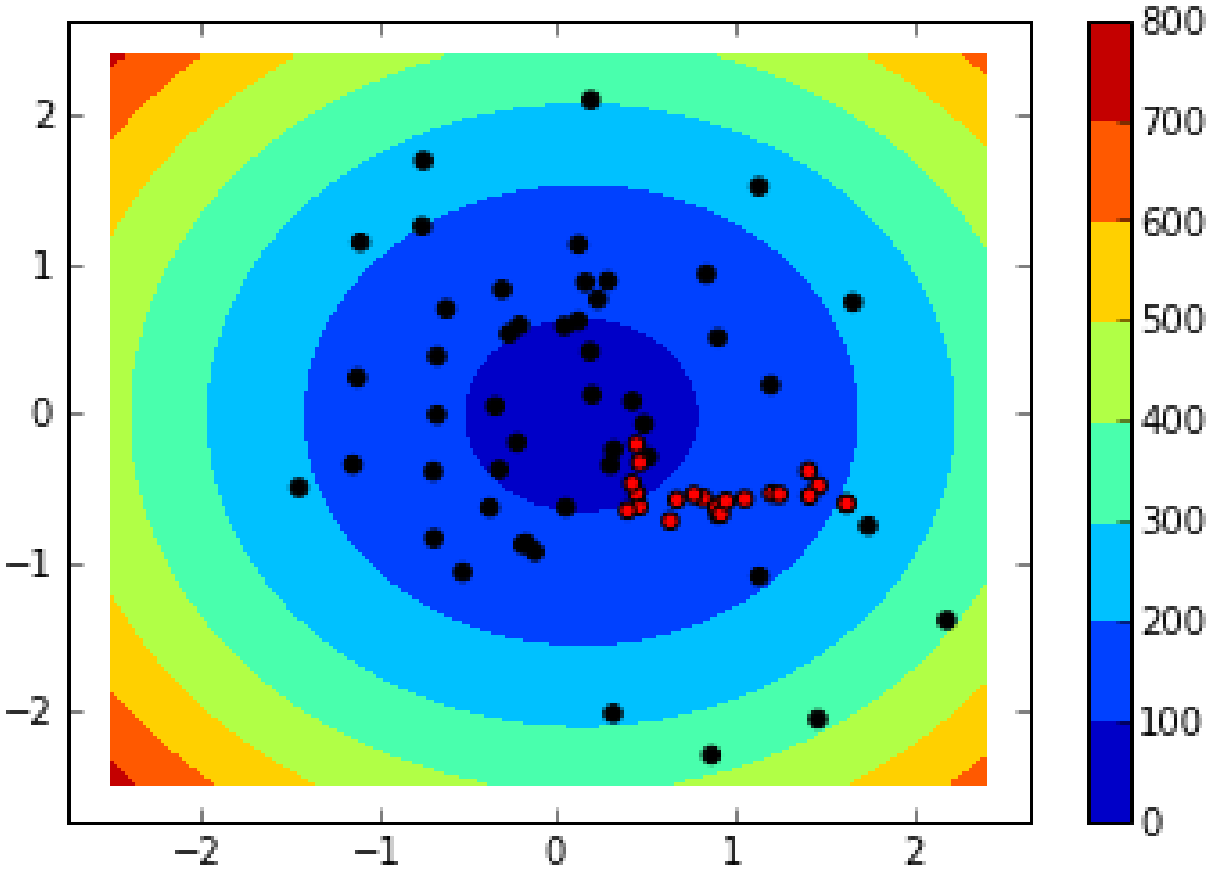
$$\Delta \mathbf{w} = -\mu \nabla f(\boldsymbol{w})$$

Example

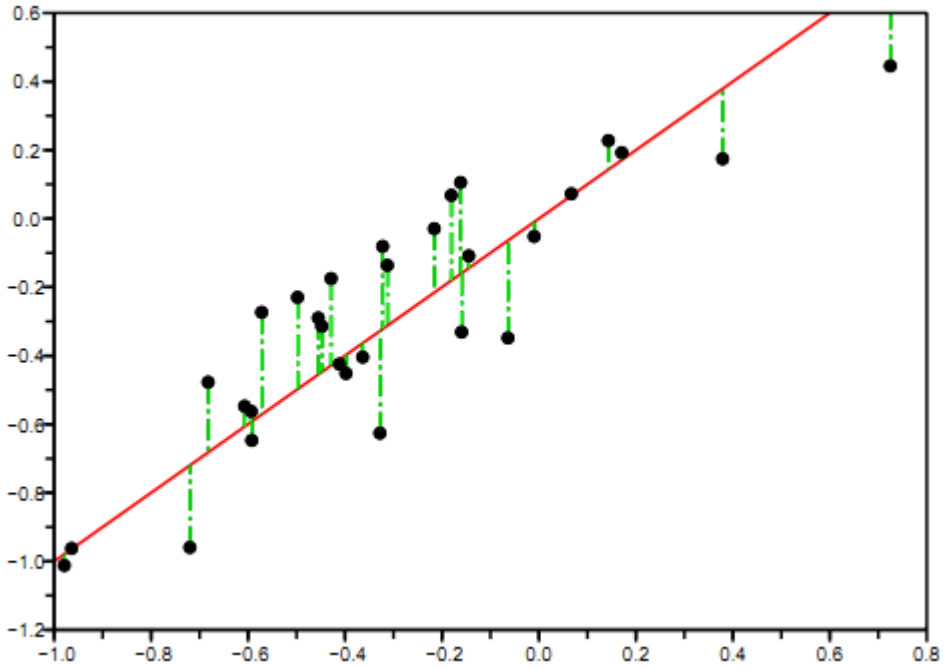


$$\Delta \mathbf{w} = \mu \sum_{i} 2(\boldsymbol{x_i} - \boldsymbol{w})$$

Example



$$\Delta \mathbf{w} = \mu \cdot 2(\boldsymbol{x_i} - \boldsymbol{w})$$



Linear Regression

Konstantin Tretyakov
http://kt.era.ee

IFISS Summmer School 2014

STACC Software Technology and Applications Competence Center

BIT

TARTU ÜLIKOOL 1632 UNIVERSITAS TARTUENSIS

Supervised Learning

- Let X and Y be some sets.
- Let there be a training dataset:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$
$$x_i \in X, y_i \in Y$$

Supervised Learning

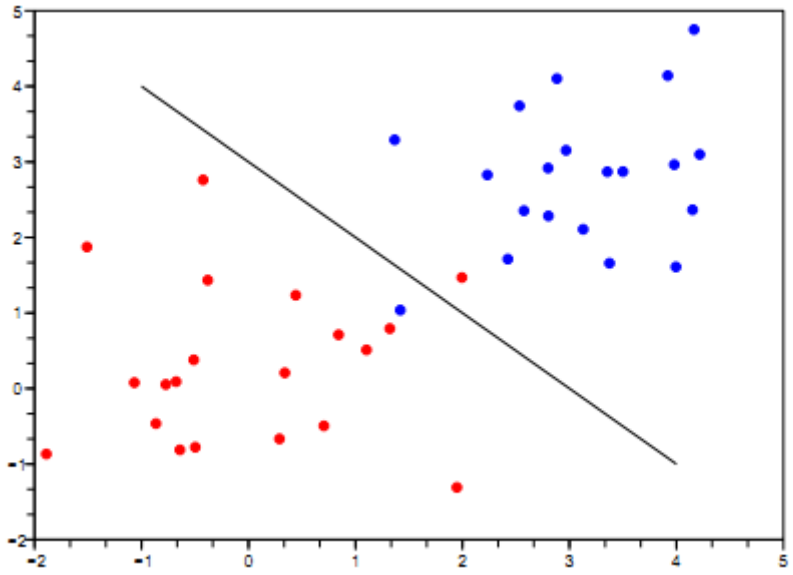
- Let X and Y be some sets.
- Let there be a training dataset:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$
$$x_i \in X, y_i \in Y$$

- **Supervised learning:**

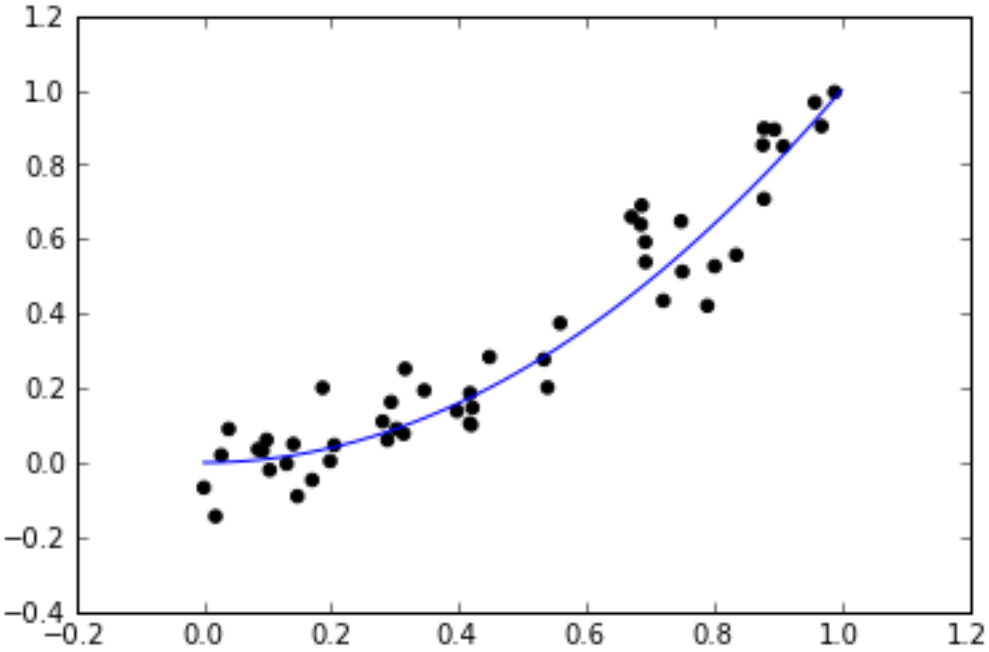
Find **a function** $f: X \to Y$, **generalizing the dependency** present in the data.

Classification



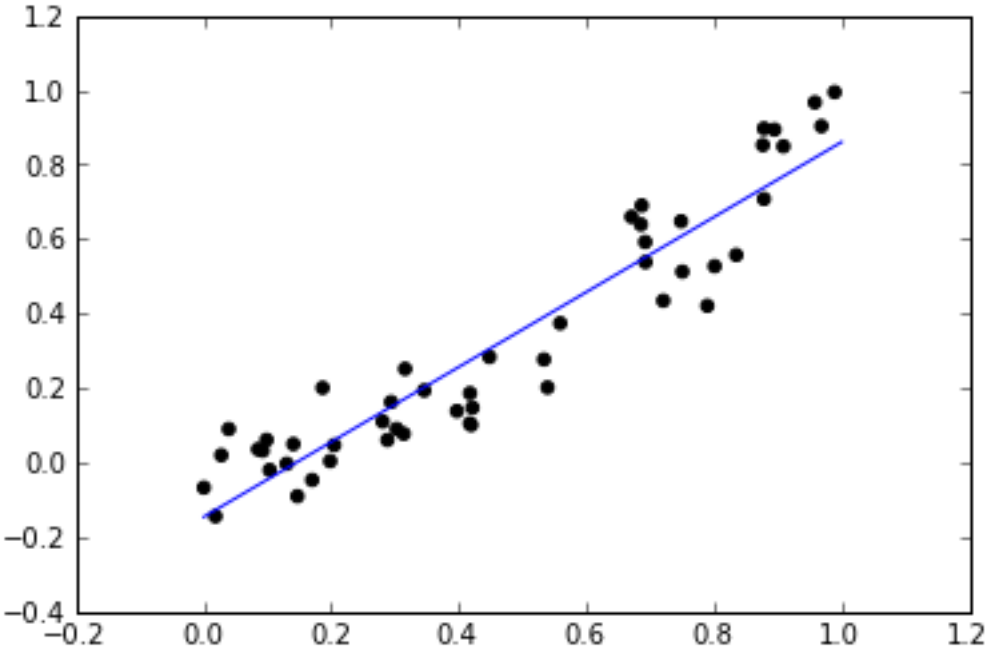
- $X = \mathbb{R}^2, \quad Y = \{\text{blue}, \text{red}\}$
- $D = \{\big((1.3, 0.8), \text{red}\big), \big((2.5,2.3), \text{blue}\big), \dots\}$
- $f(x_1, x_2) = \text{if } (x_1 + x_2) > 3 \text{ then blue else red}$

Regression



- $X = \mathbb{R}, \quad Y = \mathbb{R}$
- $D = \{(0.5, 0.26), (0.43, 0.08), \dots\}$
- $f(x) = x^2$

Linear Regression



- $X = \mathbb{R}, \quad Y = \mathbb{R}$
- $D = \{(0.5, 0.26), (0.43, 0.08), \dots\}$
- $f(x) = -0.14 + 1.01\,x$

Linear Regression

$$X = \mathbb{R}^m, \qquad Y = \mathbb{R}$$

$$f(x_1, \ldots, x_m) = w_0 + w_1 x_1 + \cdots + w_m x_m$$

Linear Regression

$$X = \mathbb{R}^m, \qquad Y = \mathbb{R}$$

$$f(x_1, \ldots, x_m) = w_0 + w_1 x_1 + \cdots + w_m x_m$$

$$f(x_1, \ldots, x_m) = w_0 + \langle \boldsymbol{w}, \boldsymbol{x} \rangle$$

Inner product

Linear Regression

$$X = \mathbb{R}^m, \qquad Y = \mathbb{R}$$

$$f(x_1, \ldots, x_m) = w_0 + w_1 x_1 + \cdots + w_m x_m$$

$$f(x_1, \ldots, x_m) = w_0 + \langle \boldsymbol{w}, \boldsymbol{x} \rangle$$

$$f(x_1, \ldots, x_m) = w_0 + (w_1, \ldots, w_m) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

Linear Regression

$$X = \mathbb{R}^m, \qquad Y = \mathbb{R}$$

$$f(x_1, \dots, x_m) = w_0 + w_1 x_1 + \cdots + w_m x_m$$

$$f(x_1, \dots, x_m) = w_0 + \langle \boldsymbol{w}, \boldsymbol{x} \rangle$$

$$f(x_1, \dots, x_m) = w_0 + (w_1, \dots, w_m) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

$$f(x_1, \dots, x_m) = w_0 + \boldsymbol{w}^T \boldsymbol{x}$$

Linear Regression

$$f(\boldsymbol{x}) = w_0 + \boldsymbol{w}^T \boldsymbol{x}$$

Linear Regression

$$f(\boldsymbol{x}) = w_0 + \boldsymbol{w}^T \boldsymbol{x}$$

Bias term

Linear Regression

$$f(\boldsymbol{x}) = w_0 + \boldsymbol{w}^T \boldsymbol{x}$$

$$f(x_1, \ldots, x_m) = (w_0, w_1, \ldots, w_m) \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_m \end{pmatrix}$$

Linear Regression

$$f(\boldsymbol{x}) = w_0 + \boldsymbol{w}^T \boldsymbol{x}$$

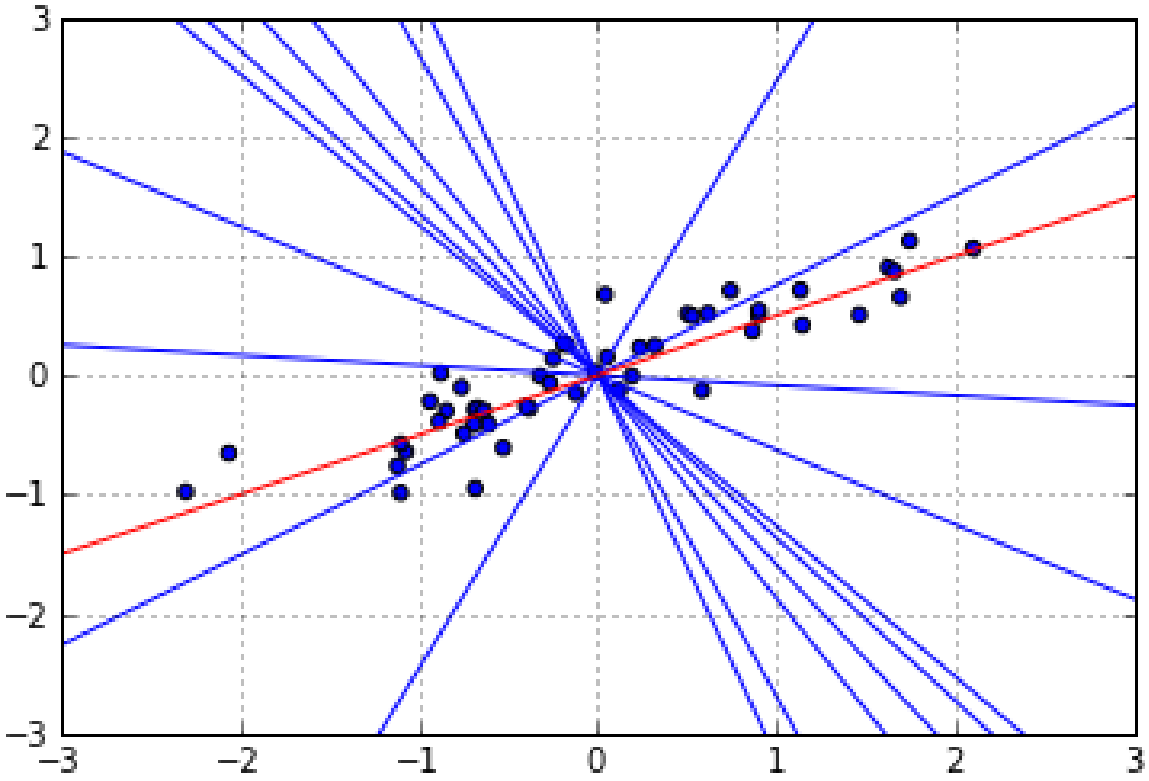
$$f(x_1, \ldots, x_m) = (w_0, w_1, \ldots, w_m) \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_m \end{pmatrix}$$

$$f(\boldsymbol{x}) = \tilde{\boldsymbol{w}}^T \tilde{\boldsymbol{x}}$$

Linear Regression

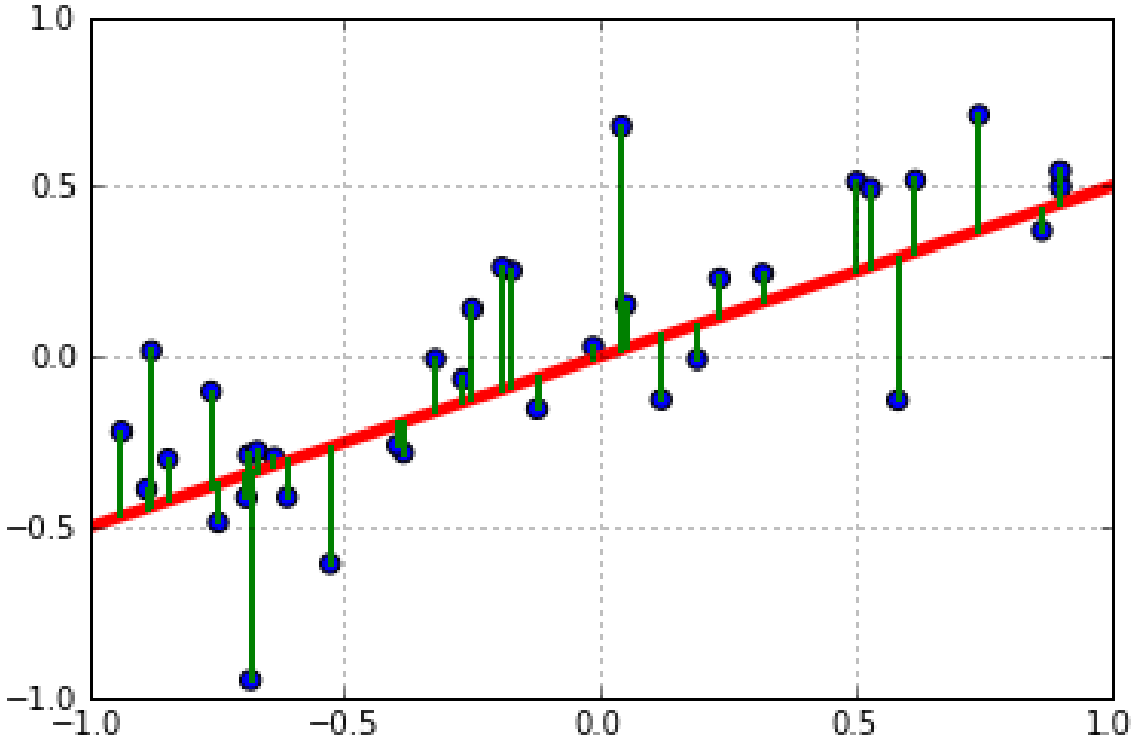
$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

Linear Regression



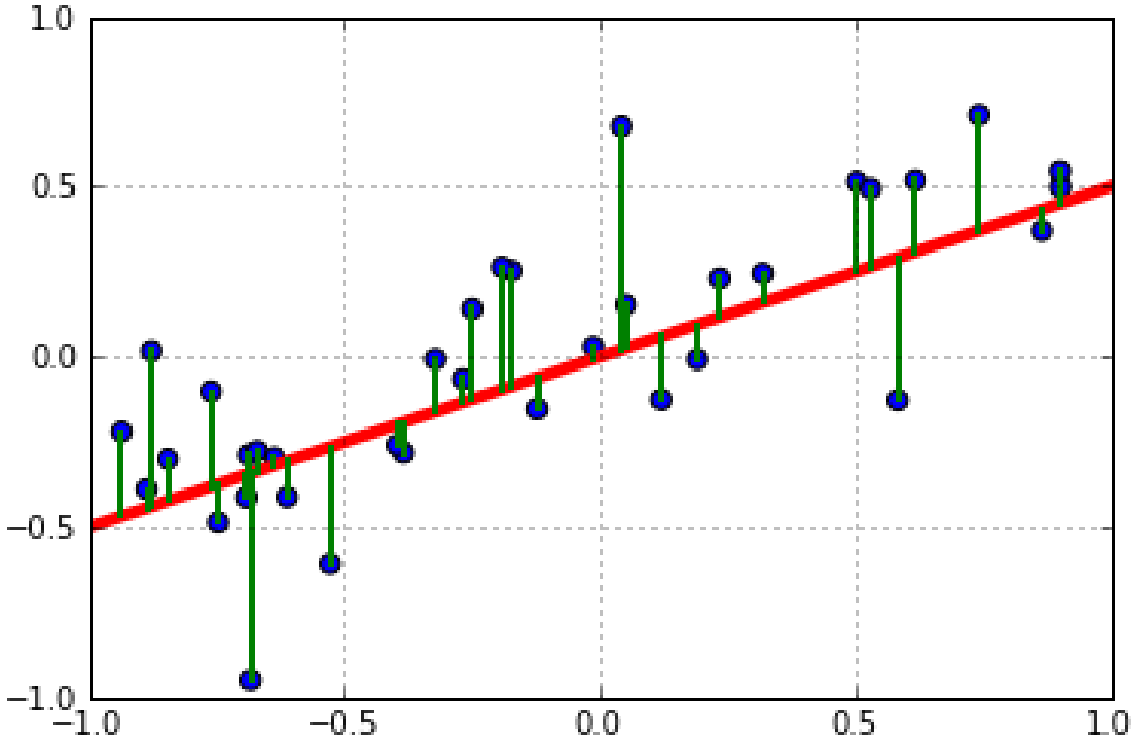
$$f(x) = wx$$

Objective Function



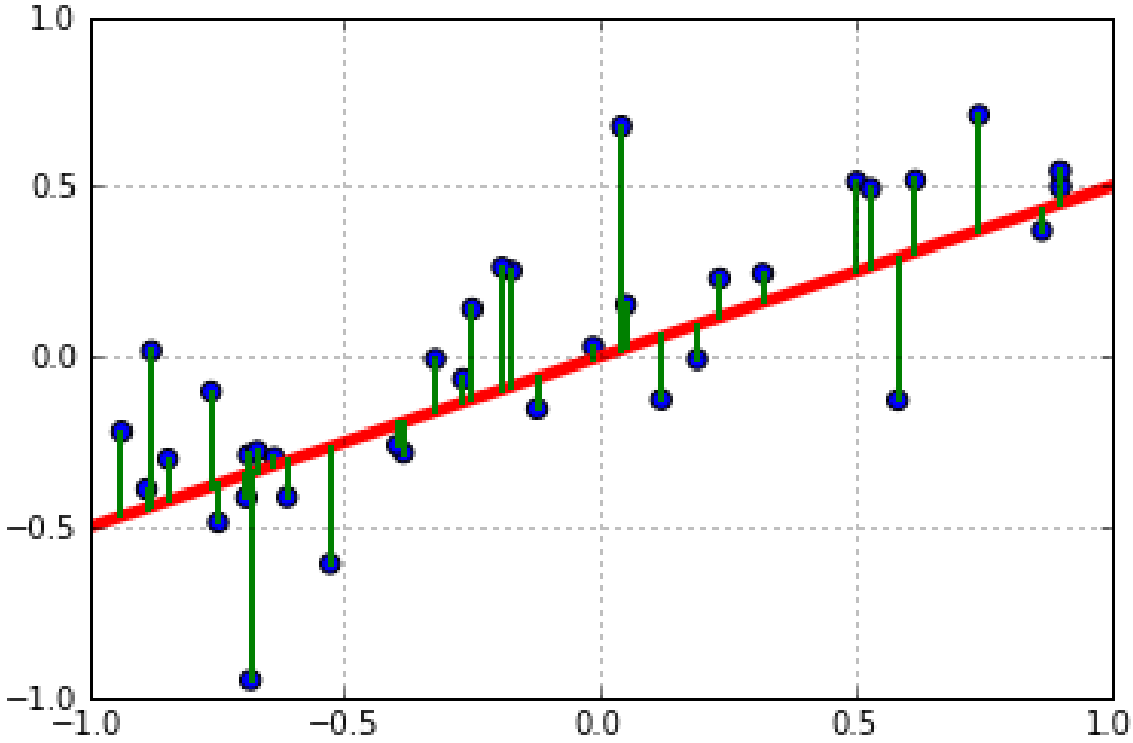
$$e_i = (f(x_i) - y_i)^2$$

Objective Function



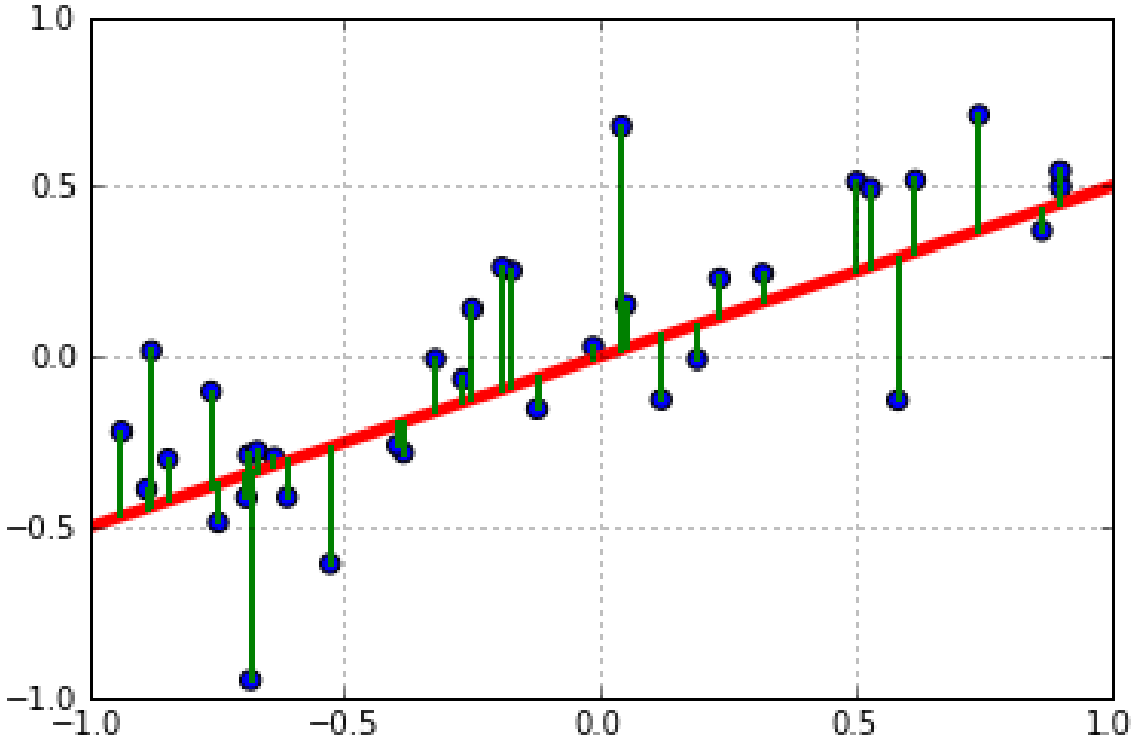
$$e_i = (\hat{y}_i - y_i)^2$$

Objective Function



$$e_i = (wx_i - y_i)^2$$

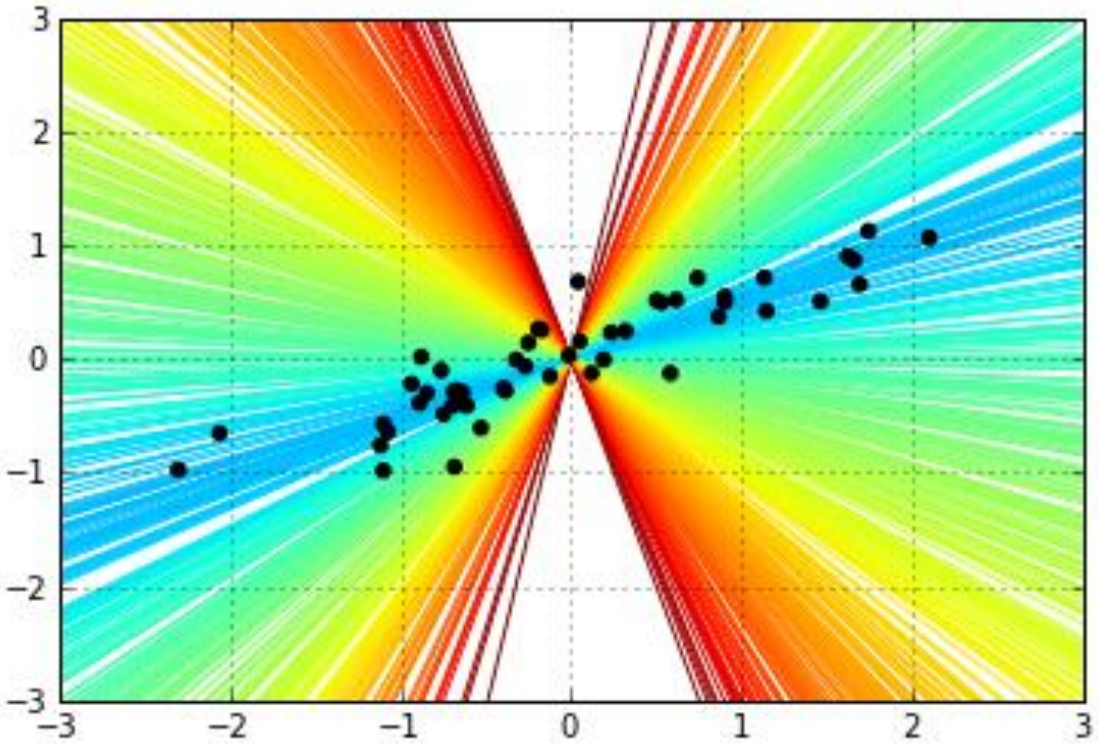
Objective Function



"OLS"

$$E_D(w) = \sum_i (wx_i - y_i)^2$$

Objective Function



$$E_D(w) = \sum_i (wx_i - y_i)^2$$

Objective Function

$$E_D(\boldsymbol{w}) = \sum_i \left(\boldsymbol{w}^T \boldsymbol{x_i} - y_i\right)^2$$

Objective Function

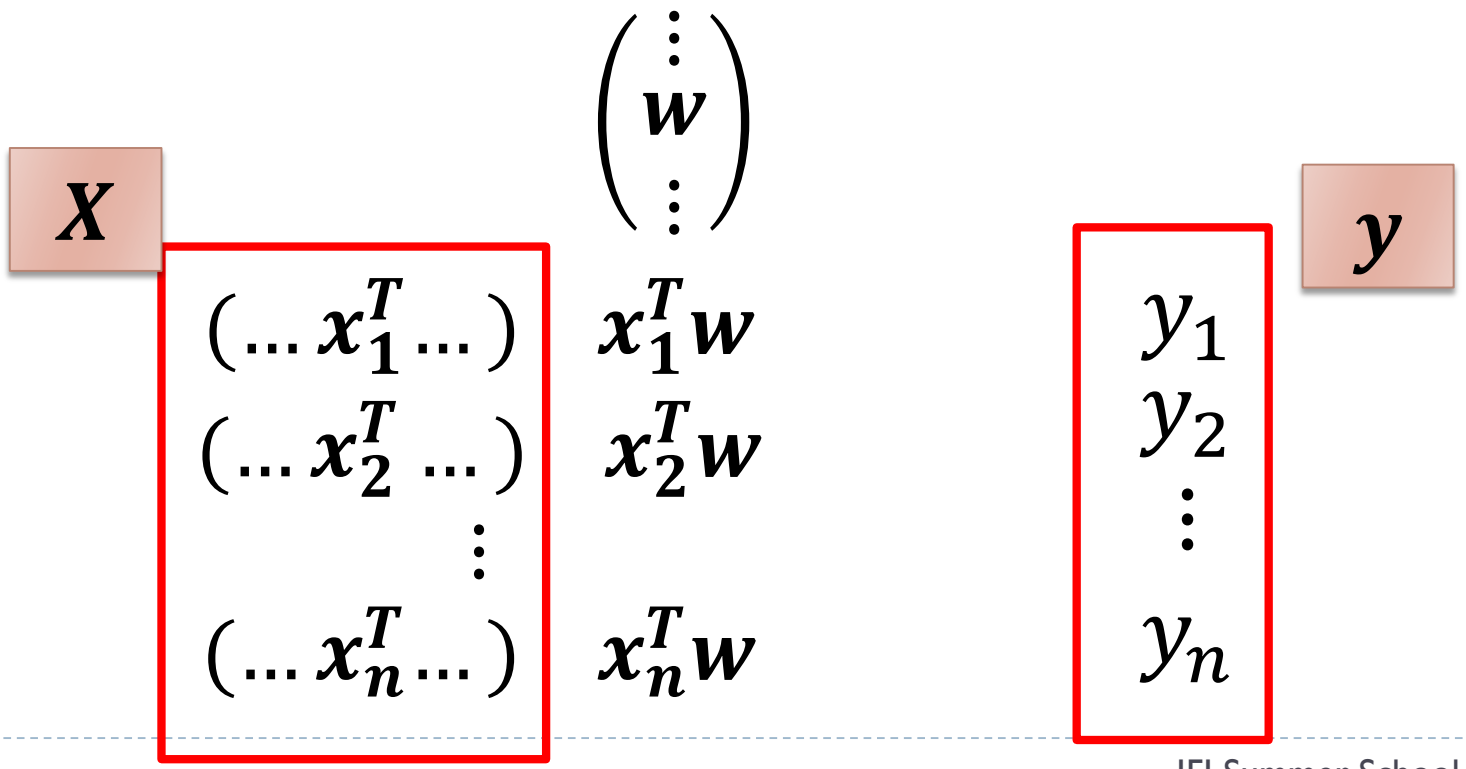
$$E_D(\boldsymbol{w}) = \sum_i \left(\boldsymbol{x_i^T w} - y_i\right)^2$$

$$\begin{pmatrix} \vdots \\ \boldsymbol{w} \\ \vdots \end{pmatrix}$$

$$\begin{matrix} (\ldots \boldsymbol{x_1^T} \ldots) & \boldsymbol{x_1^T w} & y_1 \\ (\ldots \boldsymbol{x_2^T} \ldots) & \boldsymbol{x_2^T w} & y_2 \\ \vdots & & \vdots \\ (\ldots \boldsymbol{x_n^T} \ldots) & \boldsymbol{x_n^T w} & y_n \end{matrix}$$

Objective Function

$$E_D(\boldsymbol{w}) = \sum_i \left(\boldsymbol{x}_i^T \boldsymbol{w} - y_i\right)^2$$



Objective Function

$$E_D(\boldsymbol{w}) = \sum_i \left(\boldsymbol{x_i^T w} - y_i\right)^2$$

$$\boldsymbol{Xw - y}$$

Objective Function

$$E_D(\boldsymbol{w}) = \sum_i \left(\boldsymbol{x_i^T w} - y_i\right)^2$$

$$\|\boldsymbol{Xw} - \boldsymbol{y}\|^2$$

Objective Function

$$E_D(\boldsymbol{w}) = \sum_i \left(\boldsymbol{x_i^T w} - y_i\right)^2$$

$$\|\boldsymbol{Xw} - \boldsymbol{y}\|^2$$

$$\boldsymbol{Xw} \approx \boldsymbol{y}$$

Objective Function

$$E_D(\boldsymbol{w}) = \sum_i \left(\boldsymbol{x}_i^T \boldsymbol{w} - y_i\right)^2$$

$$\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$$

$$\boldsymbol{X}\boldsymbol{w} \approx \boldsymbol{y}$$

$\boldsymbol{w} \approx \boldsymbol{X}^{-1}\boldsymbol{y}$?

Objective Function

$$E_D(\boldsymbol{w}) = \sum_i \left(\boldsymbol{x}_i^T \boldsymbol{w} - y_i\right)^2$$

$$\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$$

$$\boldsymbol{X}\boldsymbol{w} \approx \boldsymbol{y}$$

$\boldsymbol{w} = \boldsymbol{X}^+\boldsymbol{y}$!

Optimization

$$\text{argmin}_{\mathbf{w}} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$$

Optimization

$$\text{argmin}_{\mathbf{w}} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$$

Optimization

$$\text{argmin}_{\mathbf{w}} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$$

$$\|\boldsymbol{a}\|^2 = \boldsymbol{a}^T \boldsymbol{a}$$

Optimization

$$\text{argmin}_{\mathbf{w}} \frac{1}{2} \|\boldsymbol{Xw} - \boldsymbol{y}\|^2$$

$$\frac{1}{2}(\boldsymbol{Xw} - \boldsymbol{y})^T(\boldsymbol{Xw} - \boldsymbol{y})$$

$$(\boldsymbol{a} + \boldsymbol{b})^T = (\boldsymbol{a}^T + \boldsymbol{b}^T)$$

Optimization

$$\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{X}^T - \boldsymbol{y}^T)(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$$

$$\boldsymbol{a}(\boldsymbol{b}+\boldsymbol{c}) = \boldsymbol{a}\boldsymbol{b} + \boldsymbol{a}\boldsymbol{c}$$

$$\frac{1}{2}\left(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y}\right)$$

Optimization

$$\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{X}^T - \boldsymbol{y}^T)(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$$

$$\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y})$$

$\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} =$ scalar

$$\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^T\boldsymbol{y})$$

Optimization

$$E(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^T \boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^T \boldsymbol{y})$$

Optimization

$$E(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^T \boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^T \boldsymbol{y})$$

$$\nabla(f + g) = \nabla \mathbf{f} + \nabla \mathbf{g}$$
$$\nabla(\boldsymbol{w}^T \boldsymbol{A}\boldsymbol{w}) = 2\boldsymbol{A}\boldsymbol{w}$$
$$\nabla(\boldsymbol{a}^T \boldsymbol{w}) = \boldsymbol{a}$$

$$\nabla E(\boldsymbol{w}) = \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{y}$$

Optimization

$$E(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^T \boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^T \boldsymbol{y})$$

$$\nabla E(\boldsymbol{w}) = \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{y}$$
$$\boldsymbol{0} = \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{y}$$
$$\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T \boldsymbol{y}$$

Optimization

$$E(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^T\boldsymbol{y})$$

$$\nabla E(\boldsymbol{w}) = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y}$$
$$\boldsymbol{0} = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y}$$
$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$$

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Linear Regression Solution

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Linear Regression Solution

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

```
X = matrix(X)
y = matrix(y)
w =(X.T * X).I * X.T * y
```

Linear Regression Solution

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Moore-Penrose pseudoinverse

```
X = matrix(X)
y = matrix(y)
w =(X.T * X).I * X.T * y
```

Linear Regression Solution

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$\boldsymbol{w} = \boldsymbol{X}^{+}\boldsymbol{y}$$

Moore-Penrose pseudoinverse

```
X = matrix(X)
y = matrix(y)
w =(X.T * X).I * X.T * y
w = pinv(X) * y
```

Linear Regression & SKLearn

```
from sklearn.linear_model import
                        LinearRegression


model = LinearRegression()
model.fit(X, y)

w=(model.intercept_,model.coef_)
model.predict(X_new)
```

Stochastic Gradient Regression

$$\boldsymbol{\Delta w} = -\mu(\boldsymbol{w}^T \boldsymbol{x}_i - y_i)\boldsymbol{x}_i$$

Stochastic Gradient Regression

$$\boldsymbol{\Delta w} = -\mu e_i \boldsymbol{x_i}$$

Stochastic Gradient Regression

$$\Delta \boldsymbol{w} = -\mu e_i \boldsymbol{x_i}$$

```
from sklearn.linear_model import SGDRegressor

model = SGDRegressor(alpha=0, n_iter=30)
model.fit(X, y)
```

Polynomial Regression

Say we'd like to fit a model:

$$f(x_1, x_2) = w_0 + w_1 x_1 + w_2 x_2^2 + w_3 x_1 x_2$$

Polynomial Regression

Say we'd like to fit a model:

$$f(x_1, x_2) = w_0 + w_1 x_1 + w_2 x_2^2 + w_3 x_1 x_2$$

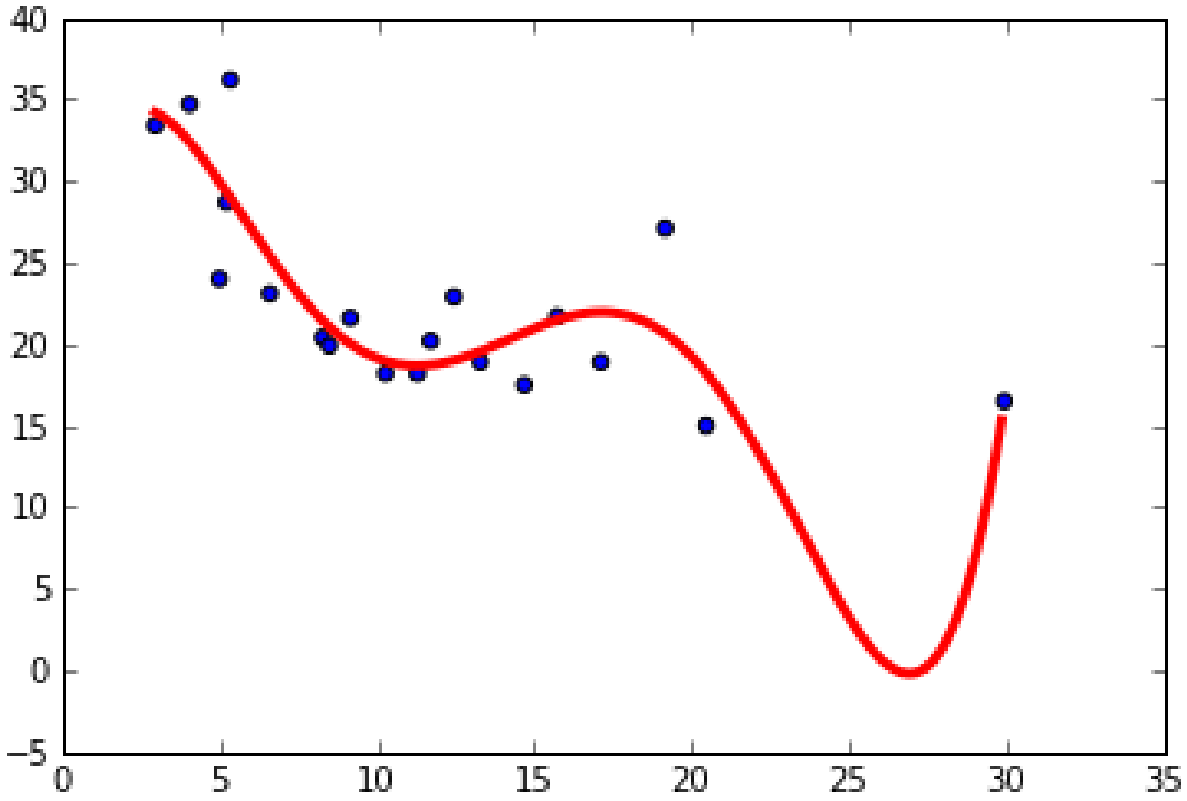
Simply transform the features and proceed as normal:

$$(x_1, x_2) \rightarrow (x_1, x_2^2, x_1 x_2)$$

Single-variable Polynomial OLS

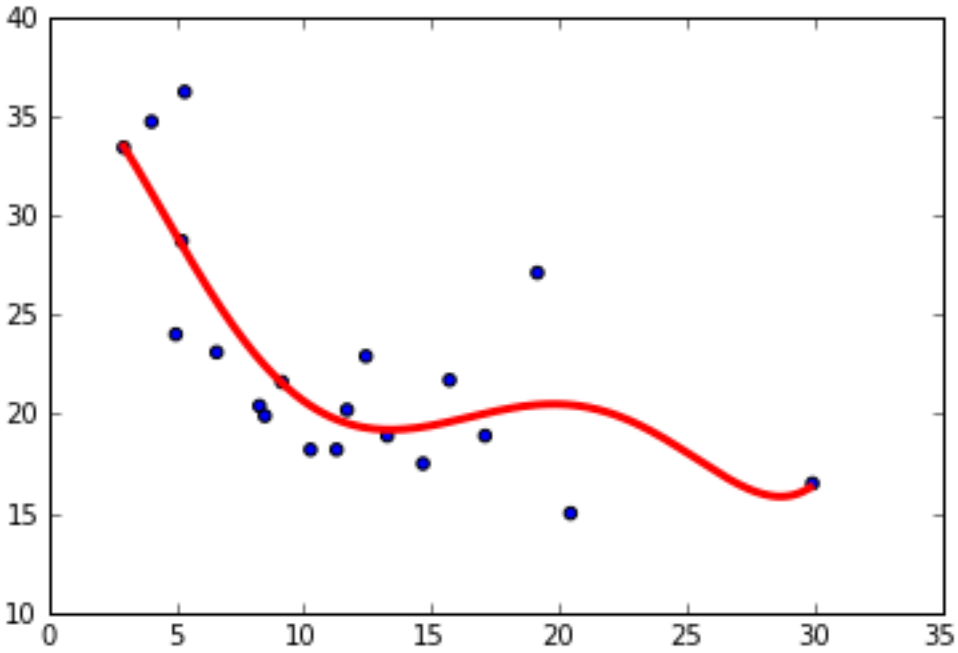
```
# n x 1 matrix
x = matrix(…)

# Add bias & square features
X = hstack([x**0, x**1, x**2])

# Solve for w
w = pinv(X) * y
```

Overfitting



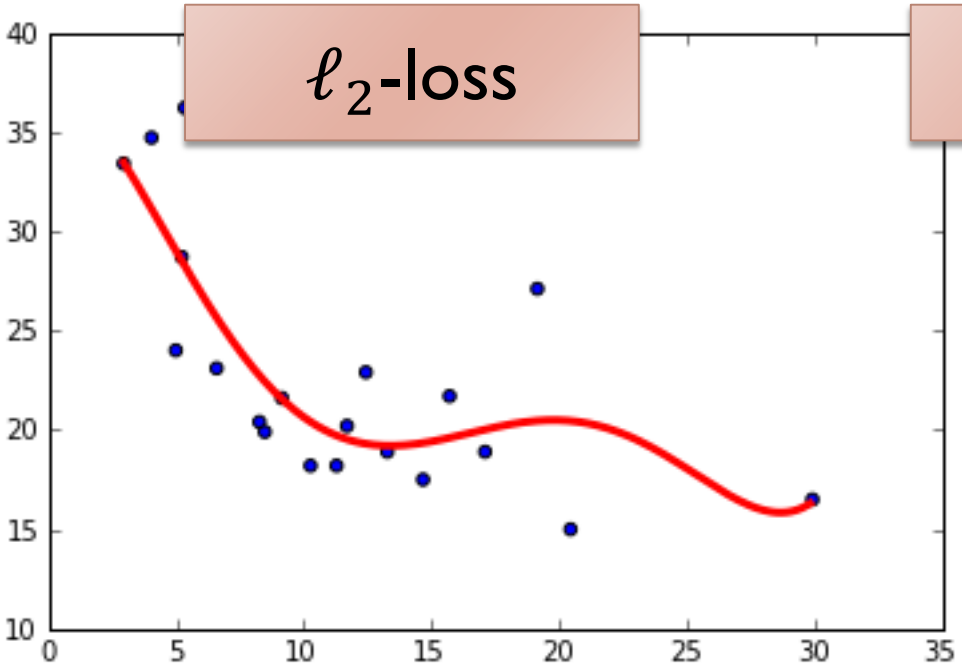
Regularization

$$E(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{w}\|^2$$



Regularization

$$E(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{w}\|^2$$



Regularization

$$E(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda|w|_1$$

ℓ_2-loss

ℓ_1-penalty

Regularization

$$E(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda|w|_0$$

ℓ_2-loss

ℓ_0-penalty

Regularization

$$E(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda|w|_0$$

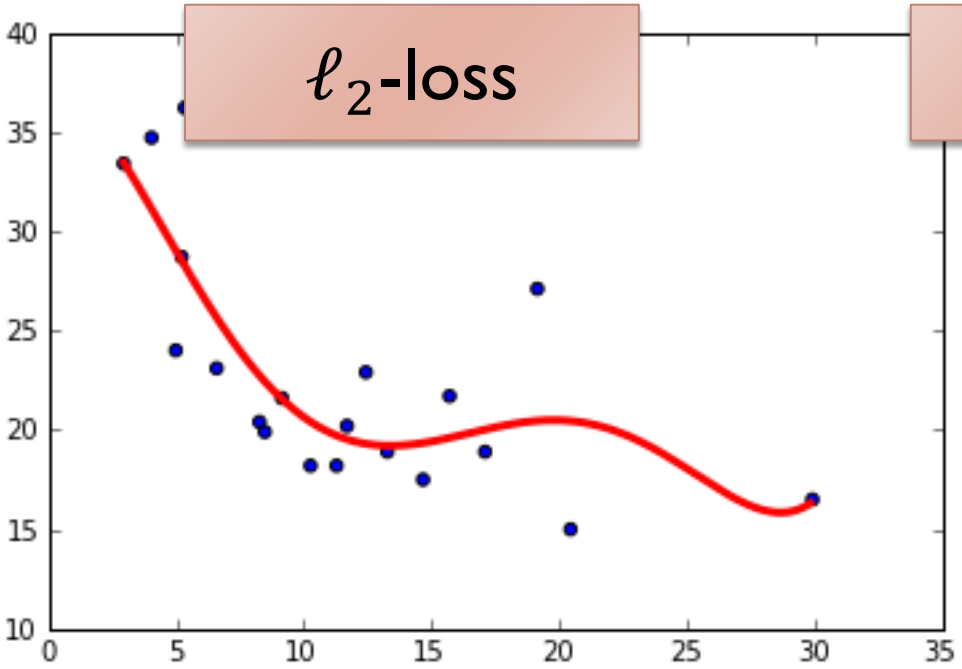
ℓ_2-loss

ℓ_0-penalty

```
>>> SGDRegressor?
Parameters
----------
loss : str, 'squared_loss' or 'huber' ...
...
penalty : str, 'l2' or 'l1' or 'elasticnet'
...
```

Regularization

$$E(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{w}\|^2$$



Ridge regression

$$\text{argmin}_{\boldsymbol{w}} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{w}\|^2$$

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\, \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Ridge regression

$$\text{argmin}_{\boldsymbol{w}} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{w}_*\|^2$$

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\, \boldsymbol{I}_*)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

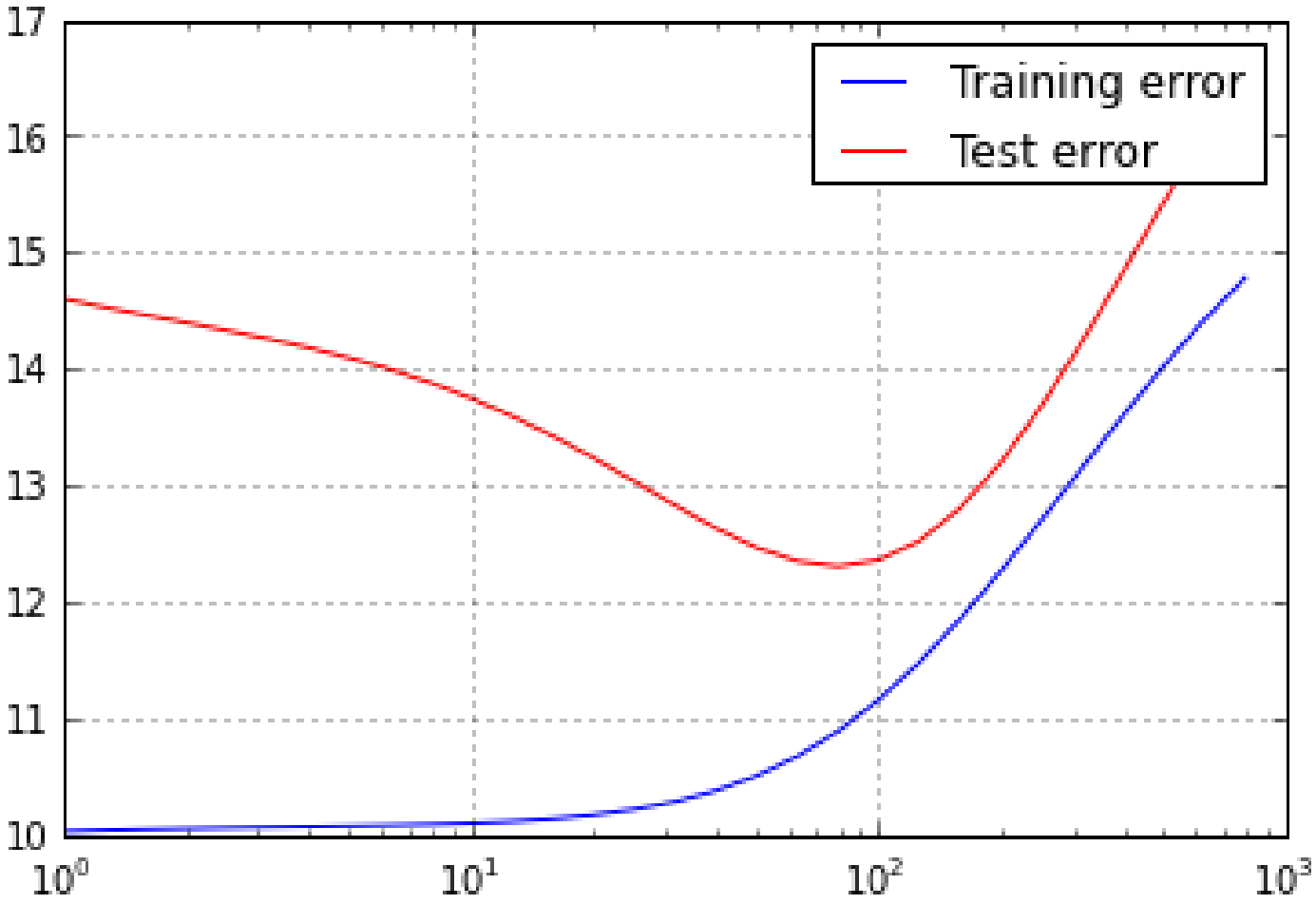
The bias term w_0 is usually **not penalized.**

$$\boldsymbol{I}_* = \begin{pmatrix} \mathbf{0} & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Exercise

Derive an SGD algorithm for Ridge Regression.

Effects of Regularization



Quiz

- OLS linear regression searches for a _____ model that has the best _____.
- Analytic solution for OLS regression:

$$\boldsymbol{w} = \_\_\_\_\_$$

- Stochastic gradient solution for OLS regression:

$$\boldsymbol{\Delta w} = \_\_\_\_\_$$

Quiz

- Large number of model parameters and/or small data may lead to _____.
- We address overfitting by _____.
- "Ridge regression" means __-loss and ___-penalty.
- Analytic solution for Ridge regression:

$$\boldsymbol{w} = \_\_\_\_\_$$

Quiz

- As we increase regularization strength (i.e. increase λ), the training error _____.

- … and the test error _____.

Questions?