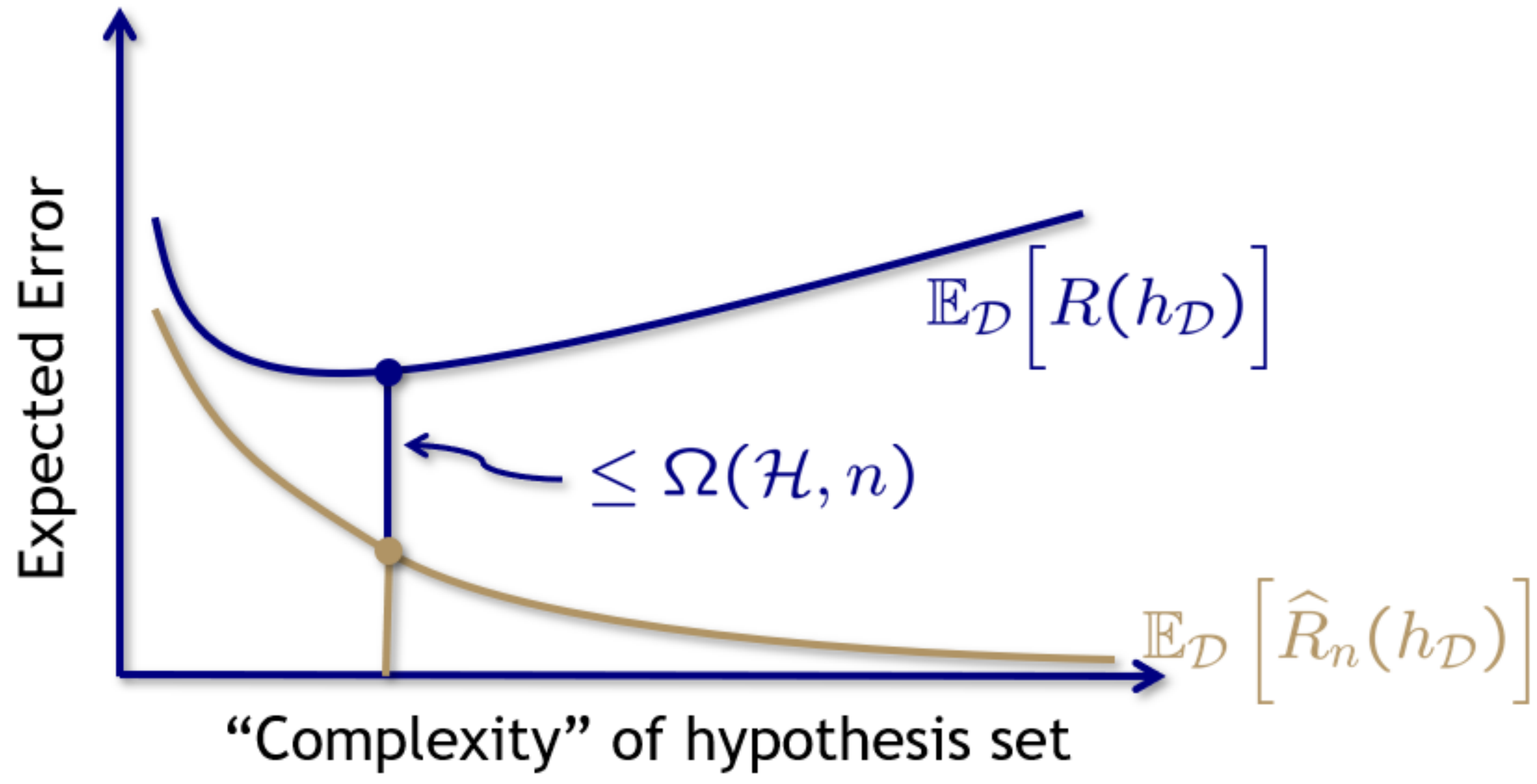
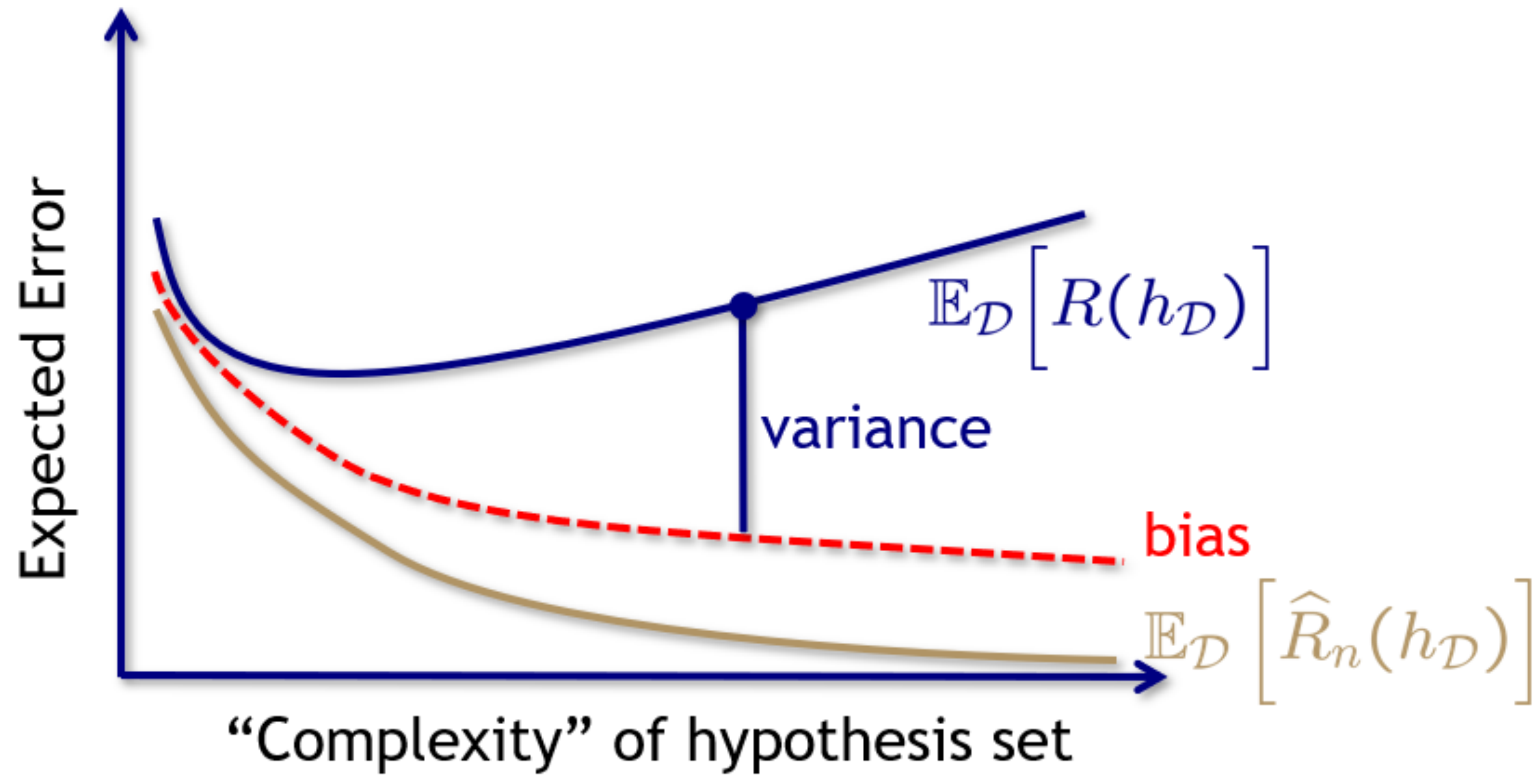
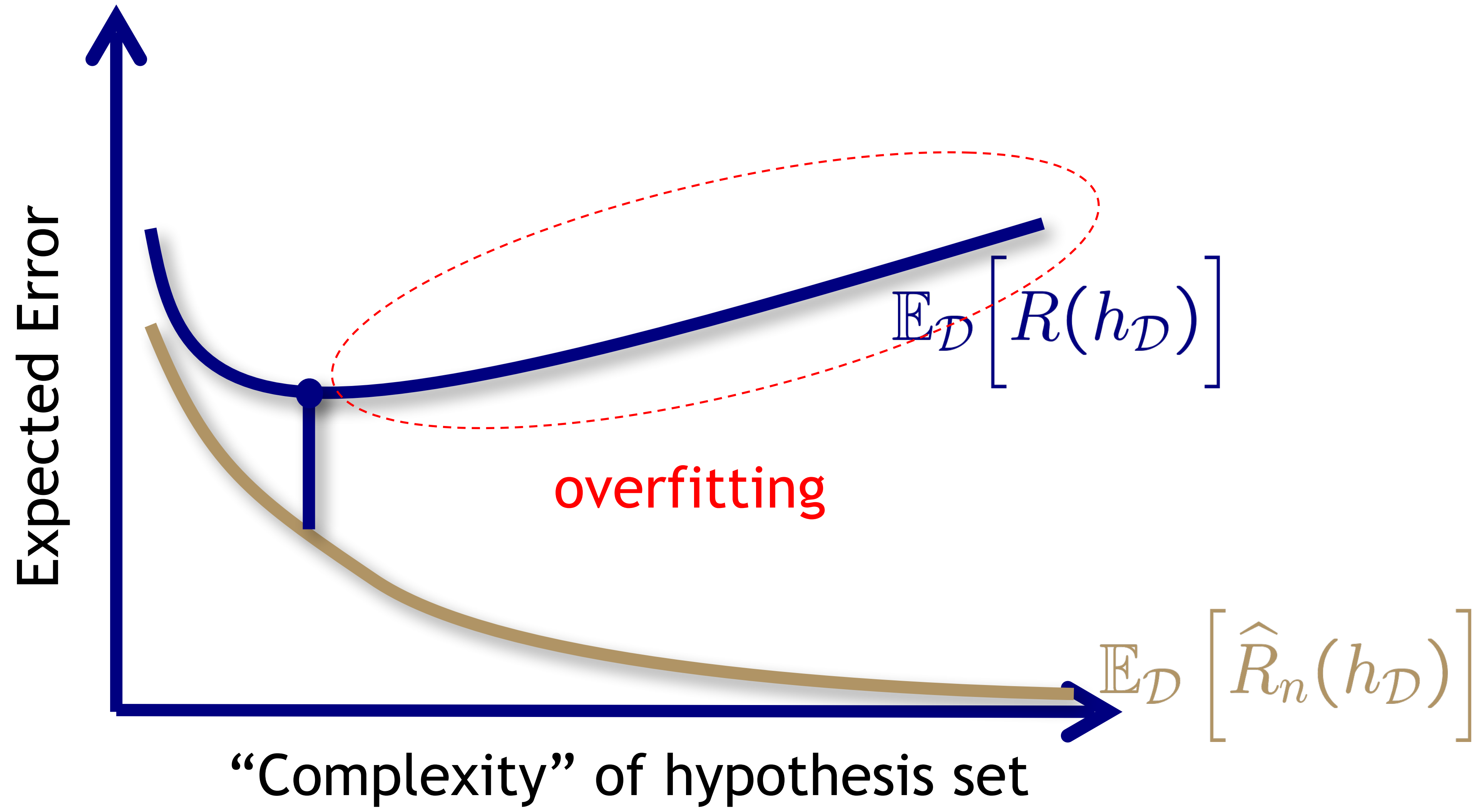
Review of Lecture 8

VC analysis



Bias-variance analysis





Review of Lecture 8

- ***Least squares*** linear regression

Select $\theta(\boldsymbol{\beta}_1, \ldots, \beta_n, \beta_0)$ to minimize

$$\mathrm{SSE}(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^T\mathbf{x}_i - \beta_0)^2 \;\; = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{A} = \begin{bmatrix} 1 & x_1(1) & \cdots & x_1(d) \\ 1 & x_2(1) & \cdots & x_2(d) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n(1) & \cdots & x_n(d) \end{bmatrix} \quad \boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta(1) \\ \vdots \\ \beta(d) \end{bmatrix}$$

- Minimizer given by

$$\widehat{\boldsymbol{\theta}} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

provided that $\boldsymbol{A}^T\boldsymbol{A}$ is ***nonsingular***

Review of Lecture 8

- ***Least squares*** linear regression

Select $\theta(\boldsymbol{\beta}_1, \ldots, \beta_n, \beta_0)$ to minimize

$$\mathsf{SSE}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0)^2 \;\; = \|\boldsymbol{y} - \boldsymbol{A\theta}\|_2^2$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{A} = \begin{bmatrix} 1 & x_1(1) & \cdots & x_1(d) \\ 1 & x_2(1) & \cdots & x_2(d) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n(1) & \cdots & x_n(d) \end{bmatrix} \quad \boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta(1) \\ \vdots \\ \beta(d) \end{bmatrix}$$

- Minimizer given by

$$\widehat{\boldsymbol{\theta}} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

provided that $\boldsymbol{A}^T\boldsymbol{A}$ is ***nonsingular***

- Overfitting $n \approx d$

too many degrees of freedom

- **Idea:** penalize candidate solutions with too many features

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$$

- ***Tikhonov regularization***

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A\theta}\|_2^2 + \|\boldsymbol{\Gamma\theta}\|_2^2$$

Ridge

$$\boldsymbol{\Gamma} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{\lambda} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{\lambda} \end{bmatrix}$$

ECE 6254
Statistical Machine Learning

Professor: Amirali Aghazadeh
Office: Coda S1209
Georgia Institute of Technology

Lecture 9: Linear Models
Regression and Regularization II

Outline

- Alternative regularization for regression (LASSO)
- A general formulation of regularization
- Robust regression

Alternative regularizers

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$$

- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)

$$r(\boldsymbol{\theta}) \approx \|\boldsymbol{\theta}\|_0 := |\mathrm{supp}(\boldsymbol{\theta})|$$

Alternative regularizers

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$$

- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)

$$r(\boldsymbol{\theta}) \approx \|\boldsymbol{\theta}\|_0 := |\mathrm{supp}(\boldsymbol{\theta})|$$

- Least absolute shrinkage and selection operator (LASSO)

$$r(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1 = \sum_j |\theta(j)|$$

Alternative regularizers

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$$

- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)

$$r(\boldsymbol{\theta}) \approx \|\boldsymbol{\theta}\|_0 := |\mathrm{supp}(\boldsymbol{\theta})|$$

- Least absolute shrinkage and selection operator (LASSO)

$$r(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1 = \sum_j |\theta(j)|$$

  - shrinkage of each coordinate (city-block norm)
  - promotes sparsity
  - can think of $\|\boldsymbol{\theta}\|_1$ as a more computationally tractable replacement for $\|\boldsymbol{\theta}\|_0$

The LASSO

LASSO

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$$

Can also be stated in a constrained form

$$\begin{aligned} \widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} & \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 \\ \text{s.t.} \quad & \|\boldsymbol{\theta}\|_1 \leq \tau \end{aligned}$$

$$\begin{aligned} \widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} & \|\boldsymbol{\theta}\|_1 \\ \text{s.t.} \quad & \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 \leq \sigma \end{aligned}$$

Tikhonov has closed form solution, but LASSO requires solving an optimization

Note: Just like in ridge regression, in practice just penalize $\boldsymbol{\beta}$ (not β_0)

LASSO as a quadratic program

Formulate LASSO as a convex quadratic program with linear inequality constraints

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$$

LASSO as a quadratic program

Formulate LASSO as a convex quadratic program with linear inequality constraints

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$$

Becomes

$$\text{minimize}_{\theta,u} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \lambda \sum_{i=1}^{d+1} u(i)$$

$$\text{subject to } -u(i) \leq \theta(i) \leq u(i), \quad i = 1, \ldots, d+1$$

LASSO as a quadratic program

Formulate LASSO as a convex quadratic program with linear inequality constraints

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$$

Becomes

$$\text{minimize}_{\theta,u} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A\theta}\|_2^2 + \lambda \sum_{i=1}^{d+1} u(i)$$

$$\text{subject to } -u(i) \leq \theta(i) \leq u(i), \quad i = 1, \ldots, d+1$$

Equivalently

$$\text{minimize}_{\theta,u} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A\theta}\|_2^2 + \lambda \sum_{i=1}^{d+1} u(i)$$

$$\text{subject to} \quad \begin{array}{r} \theta - u \leq 0 \\ -\theta - u \leq 0 \end{array}$$

LASSO as a quadratic program

Formulate LASSO as a convex quadratic program with linear inequality constraints

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$$

Becomes

$$\text{minimize}_{\theta,u} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \lambda \sum_{i=1}^{d+1} u(i)$$

$$\text{subject to } -u(i) \leq \theta(i) \leq u(i), \quad i = 1, \ldots, d+1$$

Equivalently

$$\text{minimize}_{\theta,u} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \lambda \sum_{i=1}^{d+1} u(i)$$

$$\text{subject to} \quad \begin{array}{r} \theta - u \leq 0 \\ -\theta - u \leq 0 \end{array}$$

Use “slack” terms to change <u>piecewise-linear</u> ℓ_1 to <u>linear</u> with linear constraints

Numerous algs for solving LASSO (also known as BPDN): LARS, IHT, ISTA, FISTA…

Sparsity and the LASSO

Recall: when $n \ll d$: ***very*** susceptible to overfitting

- fewer observations than unknowns
- $\boldsymbol{A}$ has nontrivial null space
- infinitely many different choices of $\boldsymbol{\theta}$ with no obvious best solution

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$$

➔ LASSO: limiting the number of non-zeros addresses this problem

In practice, the number of non-zeros is usually much smaller than n

Sparsity and the LASSO

Recall: when $n \ll d$: ***very*** susceptible to overfitting

- fewer observations than unknowns
- $\boldsymbol{A}$ has nontrivial null space
- infinitely many different choices of $\boldsymbol{\theta}$ with no obvious best solution

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$$

➔ LASSO: limiting the number of non-zeros addresses this problem

In practice, the number of non-zeros is usually much smaller than n

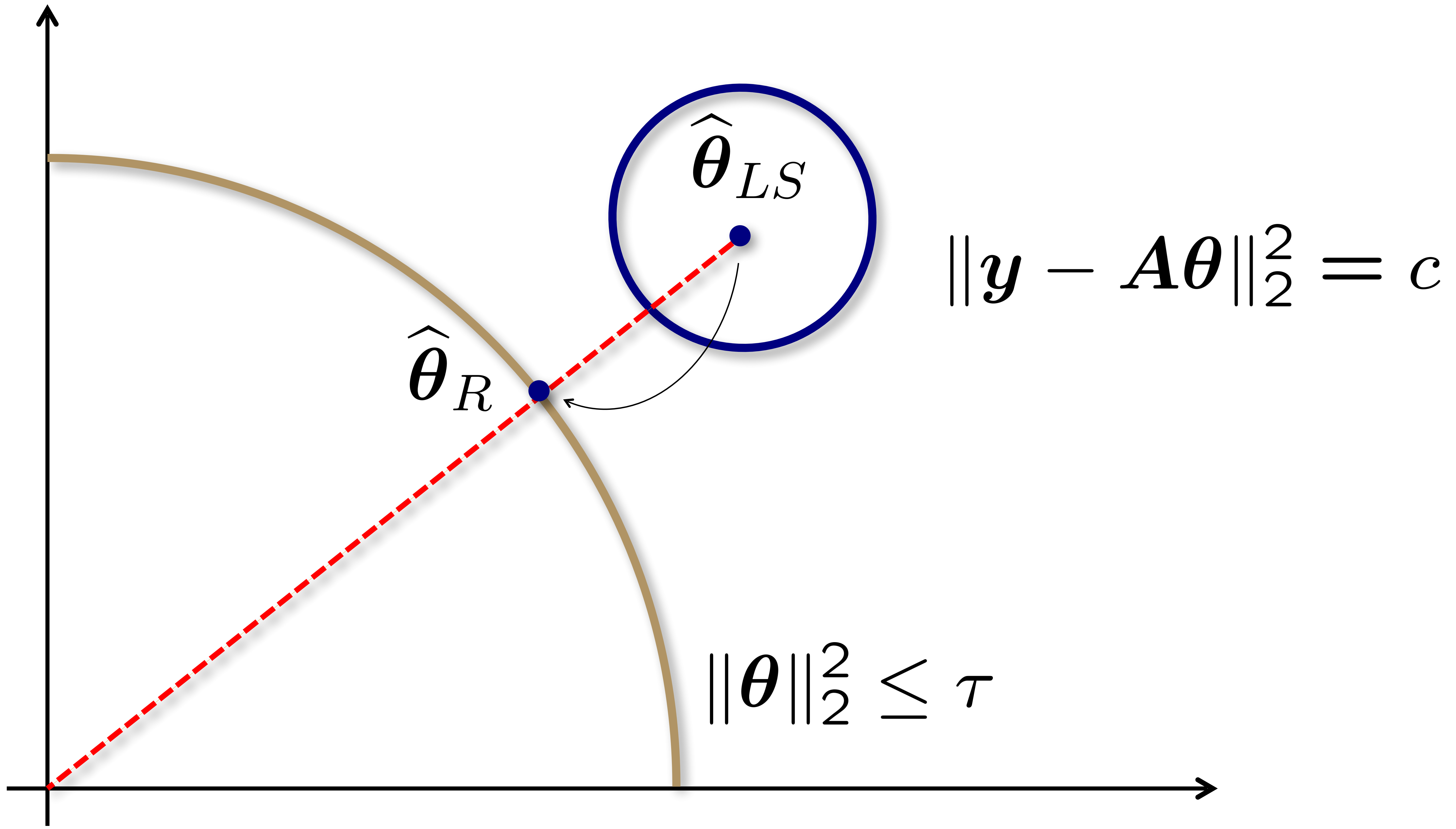
Does LASSO regularization have shrinkage similar to Tikhonov ?

Tikhonov versus least squares (review)

Assume $\Gamma = I$ and that A has orthonormal columns

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2$$

$$\text{subject to } \|\boldsymbol{\Gamma}\boldsymbol{\theta}\|_2^2 \leq \tau$$

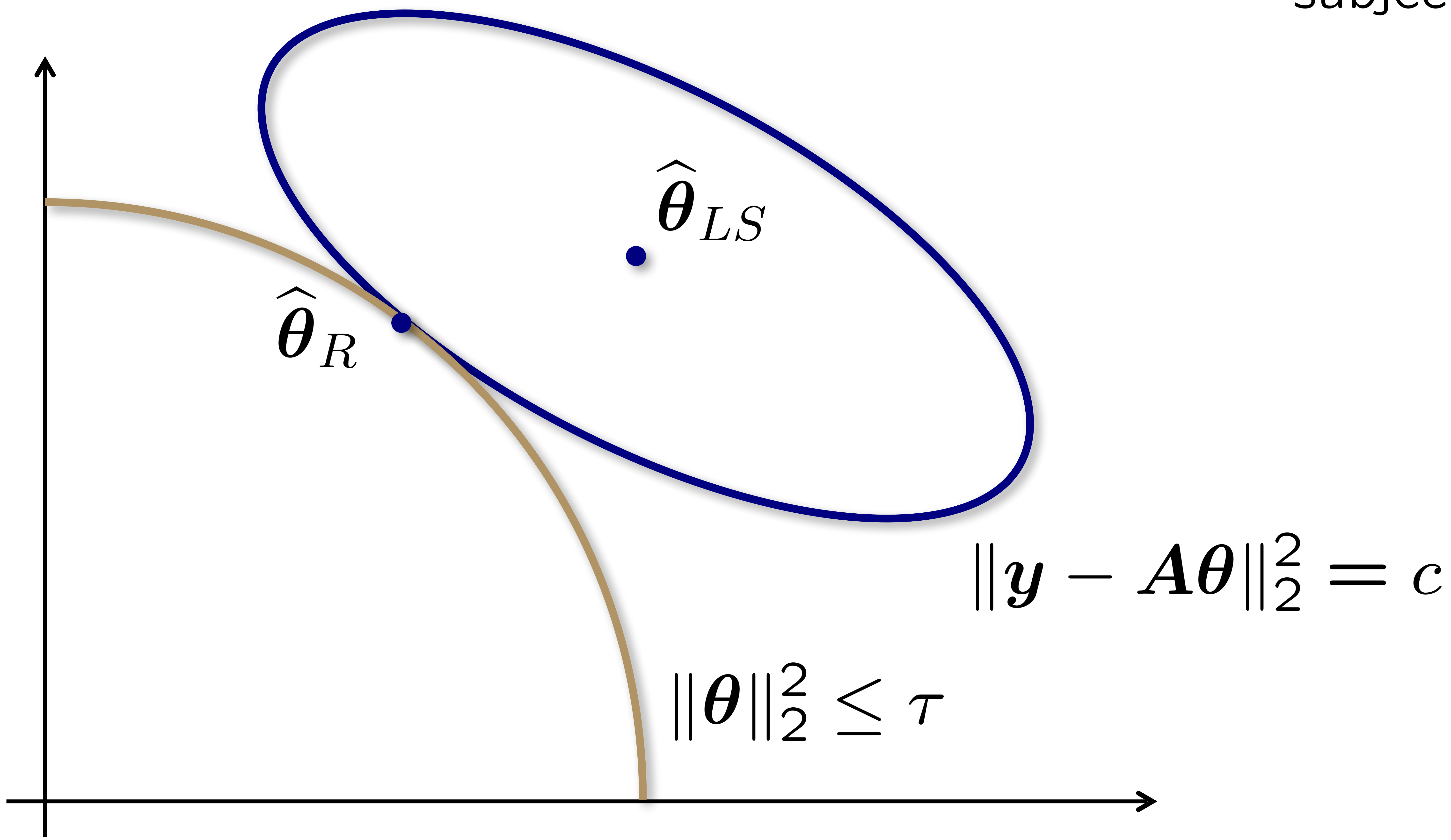


Tikhonov regularization is equivalent to shrinking LS solution towards origin

Tikhonov versus least squares (review)

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2$$
$$\text{subject to } \|\boldsymbol{\Gamma}\boldsymbol{\theta}\|_2^2 \le \tau$$

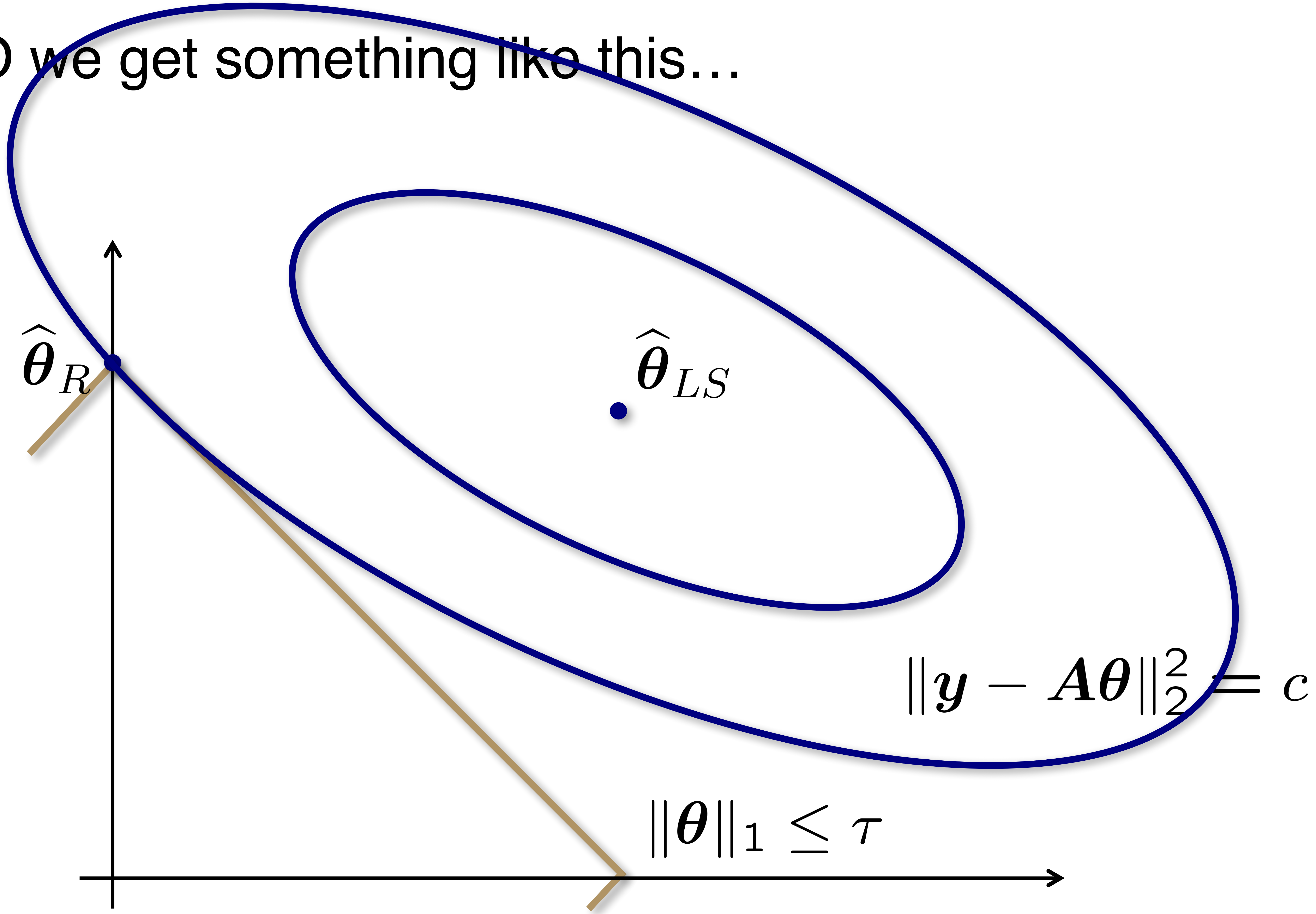
In general, we have this picture



Tikhonov regularization still shrinking the LS solution towards the origin

Lasso versus least squares

For the LASSO we get something like this…



LASSO still shrinks LS solution towards origin, but in a way that promotes sparsity

Outline

- Alternative regularization for regression (LASSO)
- A general formulation of regularization
- Robust regression

A general approach to regression

Least squares, ridge regression, and the LASSO can all be viewed as particular instances of the following general approach to regression

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \; L(\boldsymbol{\theta}, X, y) + \lambda r(\boldsymbol{\theta})$$

- $L(\boldsymbol{\theta}, X, y)$ is a ***loss function*** and enforces data fidelity $h_{\boldsymbol{\theta}}(\mathbf{x}_i) \approx y_i$
- $r(\boldsymbol{\theta})$ is a ***regularizer*** which serves to quantify the “complexity” of $\boldsymbol{\theta}$

We have seen some examples of regularizers, what about other loss functions?

Outliers in regression

The squared error loss function is sensitive to ***outliers***

If $h_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i$ is small, then $(h_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i)^2$ is not too large

But if $h_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i$ is big, then $(h_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i)^2$ is ***really*** big

Normally good – we penalize big errors – but solution is sensitive to large outliers

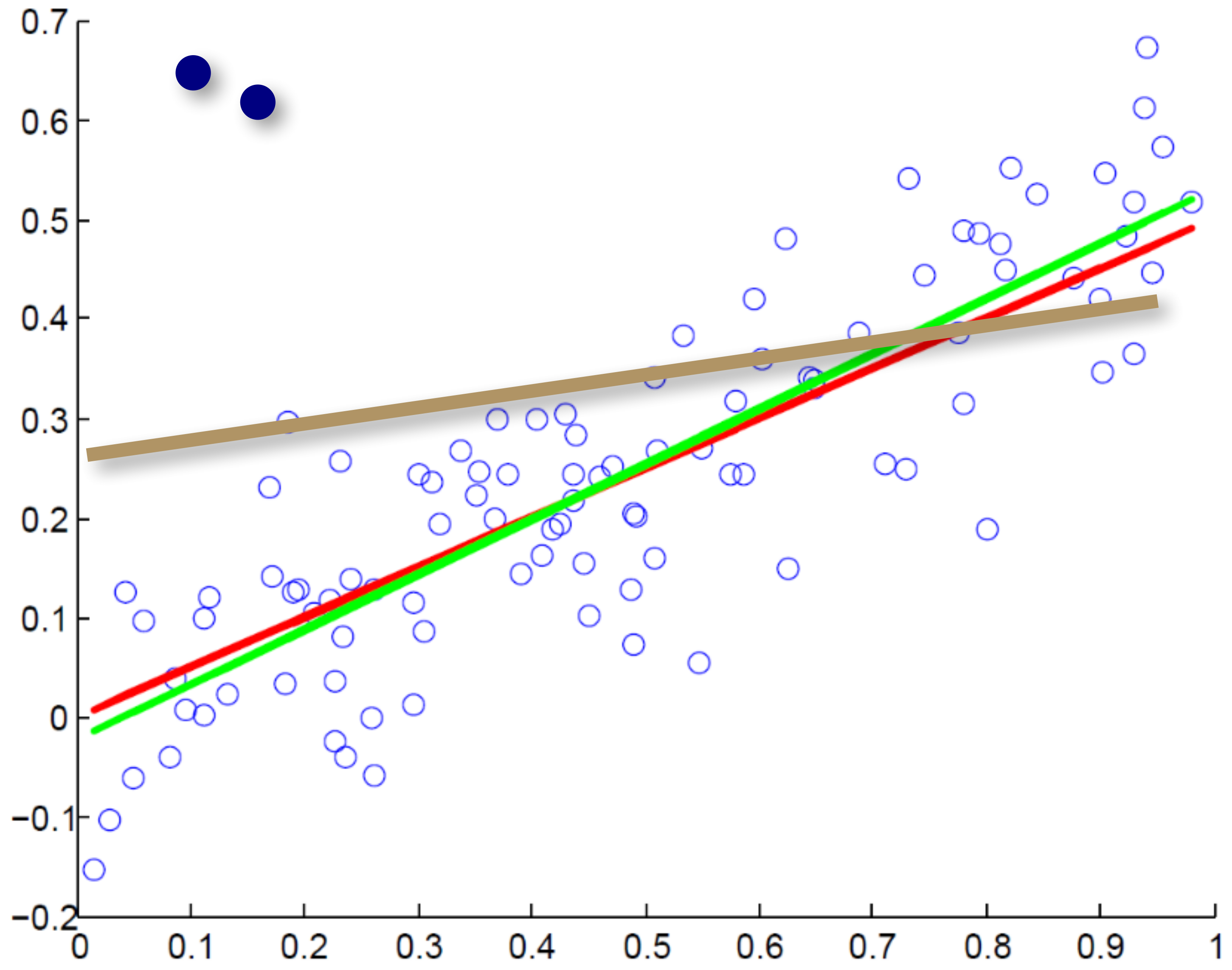
Outliers in regression

The squared error loss function is sensitive to ***outliers***

If $h_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i$ is small, then $(h_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i)^2$ is not too large

But if $h_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i$ is big, then $(h_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i)^2$ is ***really*** big

Normally good – we penalize big errors – but solution is sensitive to large outliers



Outline

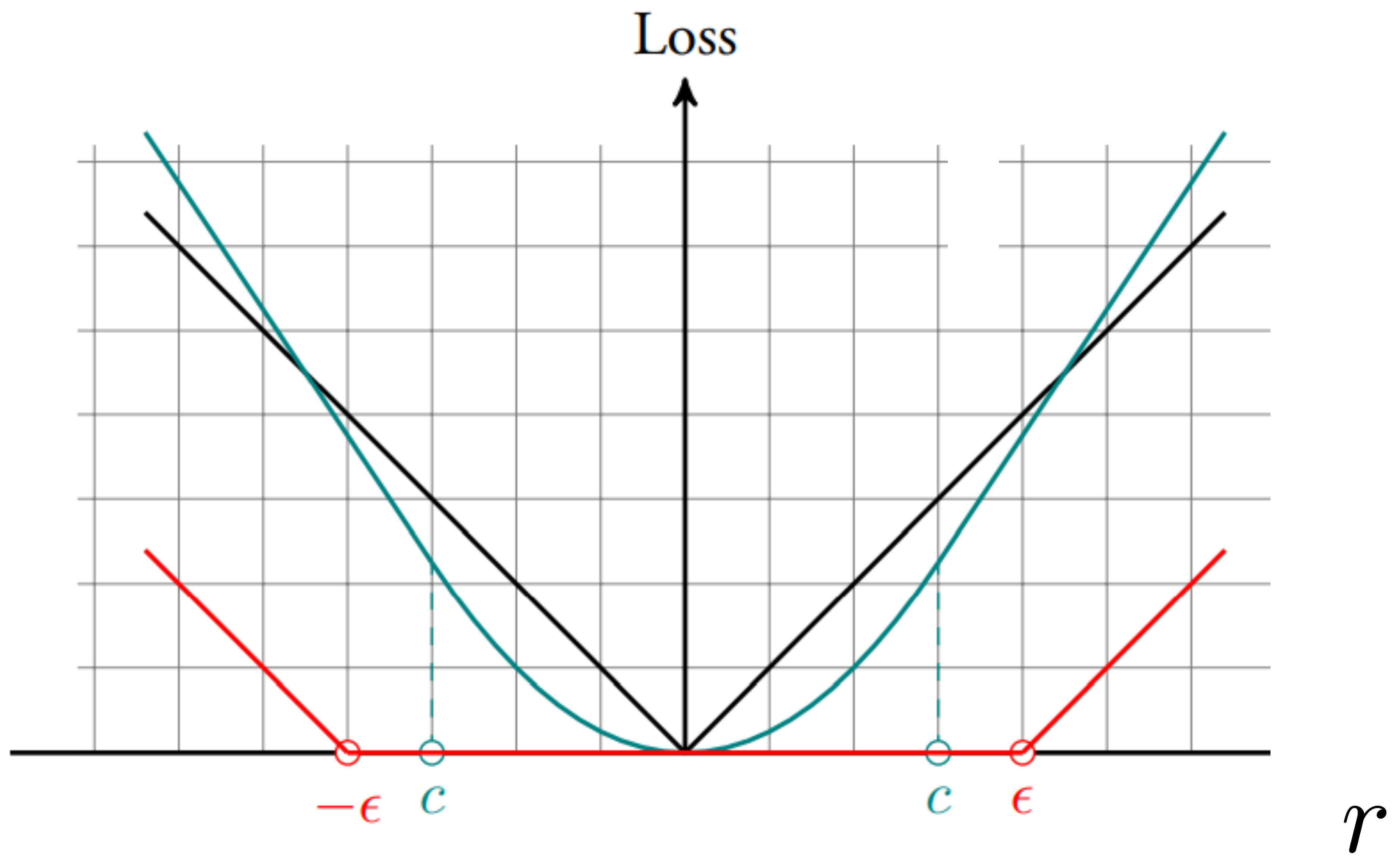
- Alternative regularization for regression (LASSO)
- A general formulation of regularization
- Robust regression

Robust regression

$$r = h_\theta(x_i) - y_i$$

Least squares

$$L_{LS}(r) = r^2$$



Robust regression

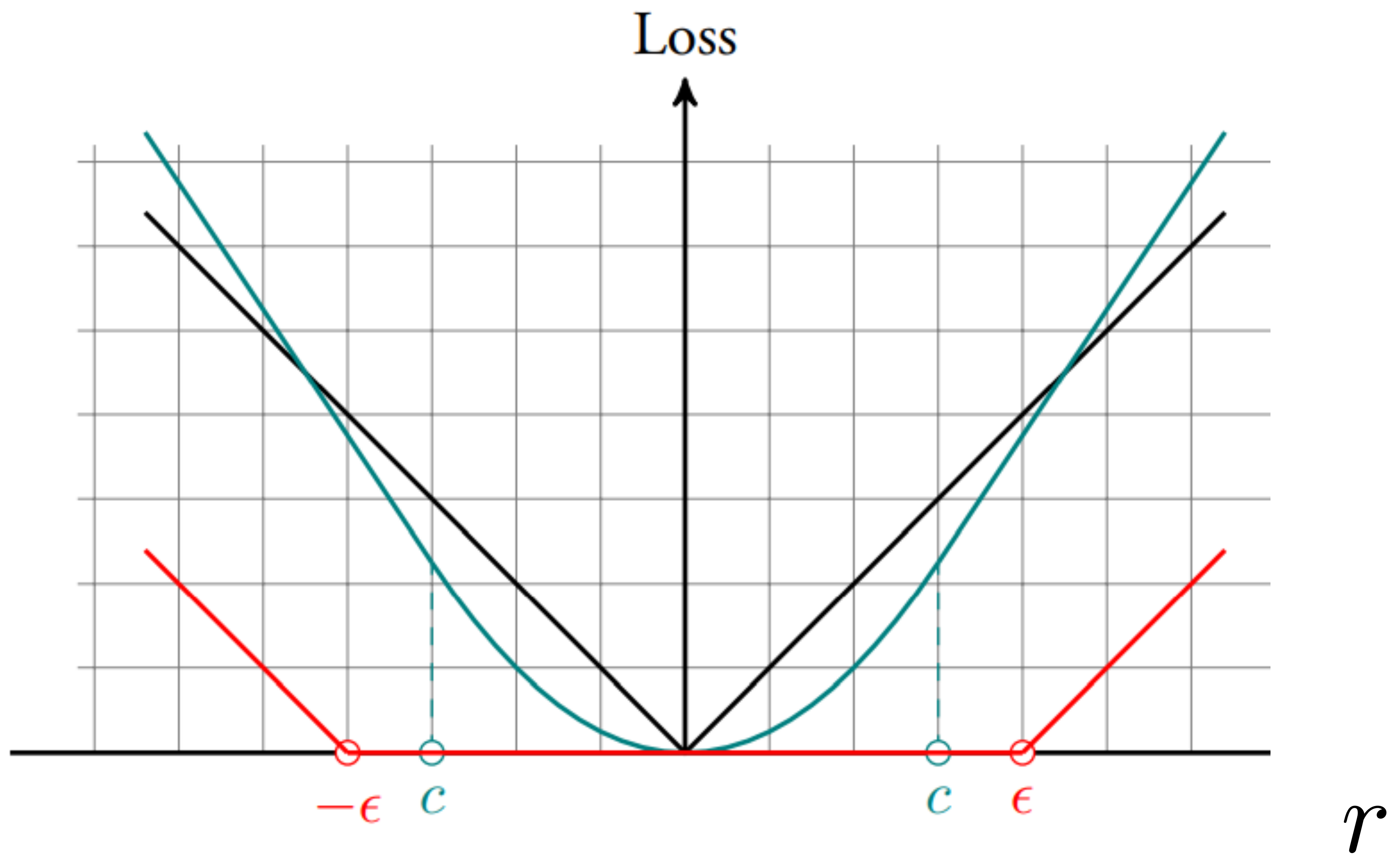
$$r = h_\theta(x_i) - y_i$$

Least squares

$$L_{LS}(r) = r^2$$

Mean absolute error

$$L_{AE}(r) = |r|$$



Robust regression

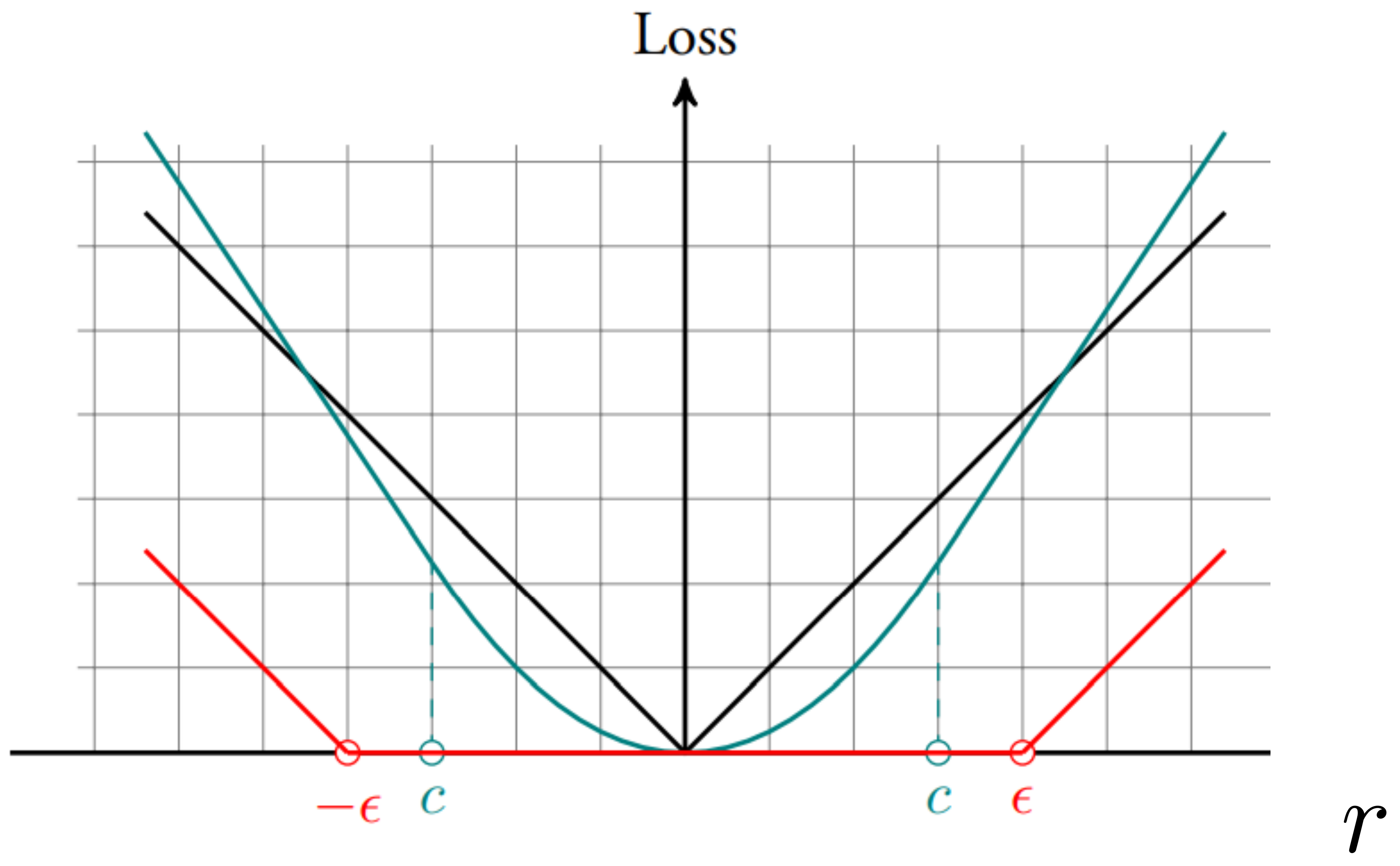
$$r = h_\theta(x_i) - y_i$$

Least squares

$$L_{LS}(r) = r^2$$

Mean absolute error

$$L_{AE}(r) = |r|$$



Huber loss $L_H(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \leq c \\ c|r| - \frac{c^2}{2} & \text{if } |r| > c \end{cases}$

Robust regression

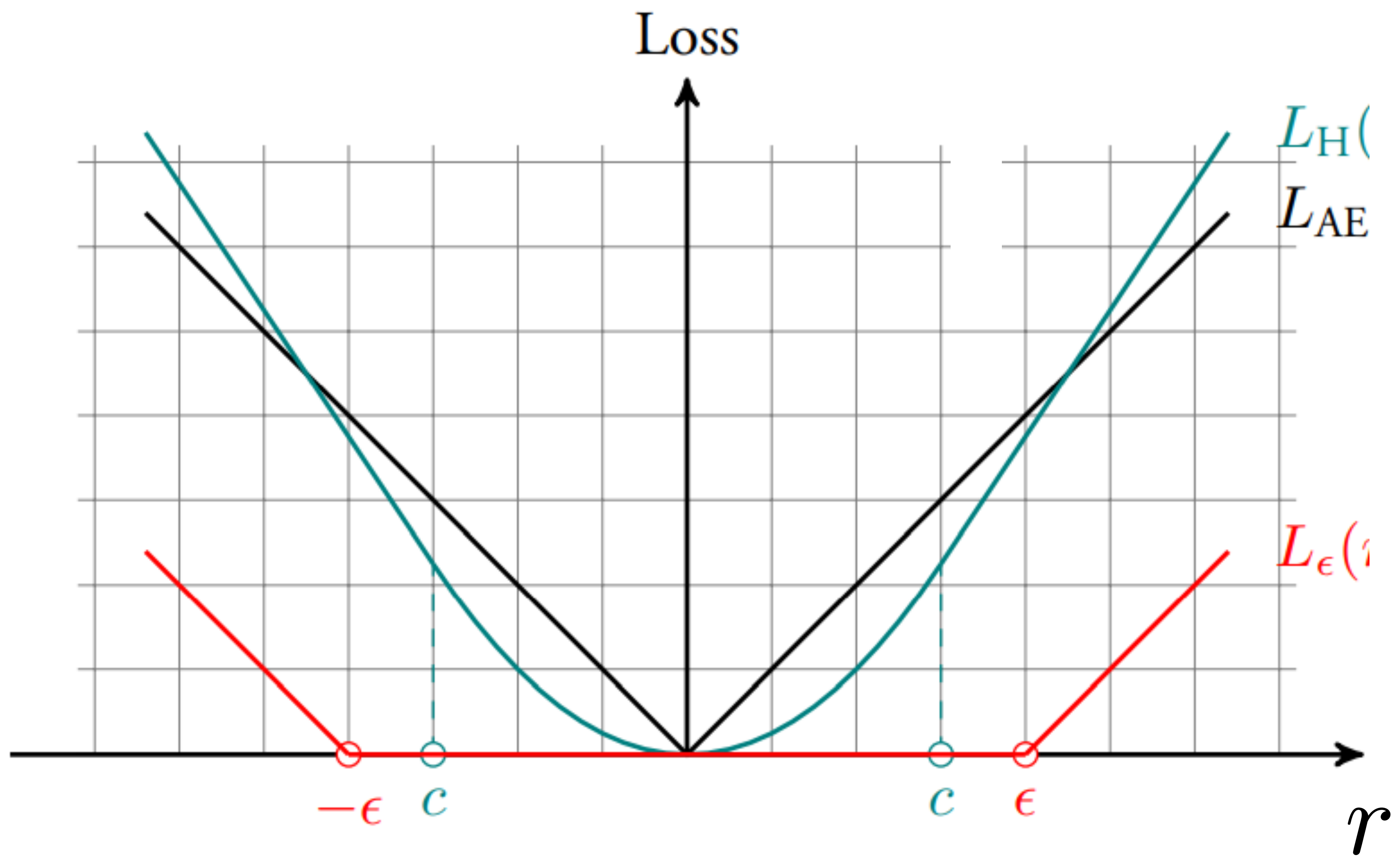
$$r = h_\theta(x_i) - y_i$$

Least squares

$$L_{LS}(r) = r^2$$

Mean absolute error

$$L_{AE}(r) = |r|$$



Huber loss $$L_H(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \leq c \\ c|r| - \frac{c^2}{2} & \text{if } |r| > c \end{cases}$$

ϵ -insensitive loss $$L_\epsilon(r) = \begin{cases} 0 & \text{if } |r| \leq \epsilon \\ |r| - \epsilon & \text{if } |r| > \epsilon \end{cases}$$

Regularized robust regression

Combine ϵ -insensitive loss with ℓ_2 regularizer

$$\widehat{\boldsymbol{\beta}}, \beta_0 = \arg\min_{(\boldsymbol{\beta}, \beta_0)} \sum_{i=1}^{n} L_\epsilon(y_i - (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0)) + \frac{\lambda}{2} ||\boldsymbol{\beta}||_2^2$$

ϵ -insensitive loss has no penalty as long as prediction is within “margin” of ϵ

This looks like a Support Vector Machine (SVM)

Exercise 9.1

Given the general formulation of the regression problem:

$$\hat{\theta} = \arg\min_{\theta} L(h_\theta(x_i) - y_i) + \lambda r(\theta)$$

where $L(h_\theta(x_i) - y_i)$ is a loss function, $h_\theta(x_i)$ is the regression model, y_i are the true values, and $r(\theta)$ is the regularizer. Given the ϵ-sensitive robust regularizer, what is the cumulative loss with $\epsilon = 0.25$ for the following values of the regression model $h_\theta(x_i)$ and the true values y_i:

$h_\theta(x_i)$	y_i
0	0.1
0.25	0.2
0.5	0.3
0.75	0.4
1.0	0.5

Logistics

- Quiz #2: HW3 and HW4
- Midterm: next week: covers everything: ALL HWs and Assignments

    (There will be questions from decision trees, etc.)

- No cheat sheets, etc.