Lecture 5

Dr. Sebastian Vollmer

- Recap MLJ
- Regularisation
- Gradient Descent
- Autodiff and more optimisation

Motivation

In the old days

Typically $n > p$ (much more data than predictors)

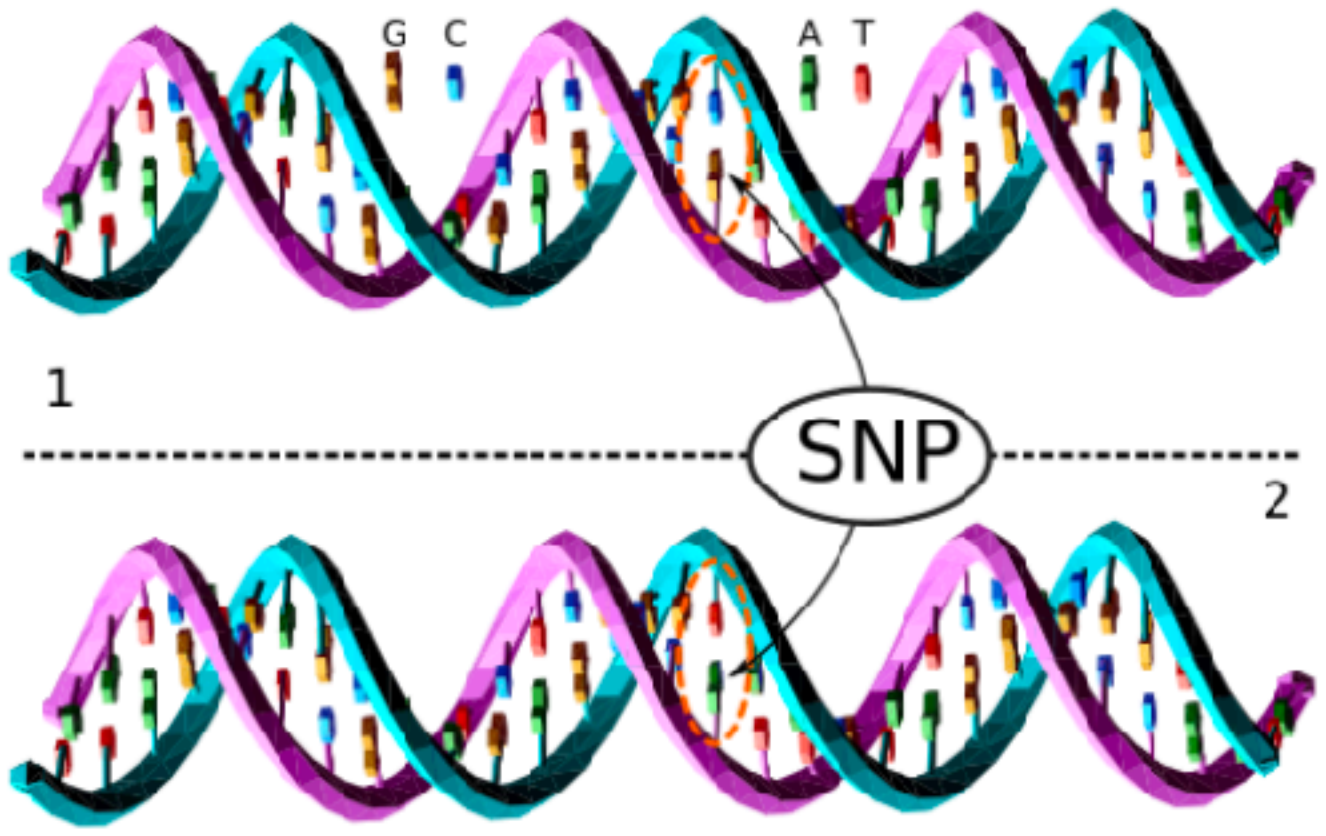
For example: predict blood pressure based on age, gender and body mass index (BMI) (e.g. $n = 200$ patients, $p = 3$).

Nowadays: Big Data

Often $n \approx p$ or $n < p$

For example: predict blood pressure based on 500 000 single nucleotide polymorphisms (SNP) ($n = 200$, $p = 500\,000$).

$\Rightarrow$ **Linear Model perfectly fits the training data.**



Beyond Least Squares Regression

Sebastian Vollmer

Recall: A linear regression model is given by

$$Y = X\beta + \epsilon$$

where

$X \in \mathbb{R}^{n \times p}$ is a matrix of covariates,

$\beta \in \mathbb{R}^p$ is the vector of regression coefficients,

$\epsilon \sim \mathcal{N}(0,\sigma^2 I)$.

Ordinary least squares regression:

$$\hat{\beta}_{OLS} = \underset{\beta}{\textbf{arg min}}\, S(\beta)$$

$$S(\beta) = \frac{1}{2} \parallel Y - X\beta \parallel_2^2 = \sum_{i=1}^{n} (Y_i - \sum_{k=1}^{p} X_{ik}\beta_k)^2$$

Motivation

1. predictive accuracy - trade bias against variance
2. In case $n \ll p$, $\hat{\beta}$ is not uniquely identified

   Genetics: n patients participate and p genes observed
3. interpretation - a small number of predictors captures the main effect
4. robustness
5. non-linear

Hence, even if a model is correctly specified, **we should consider alternative approaches!**

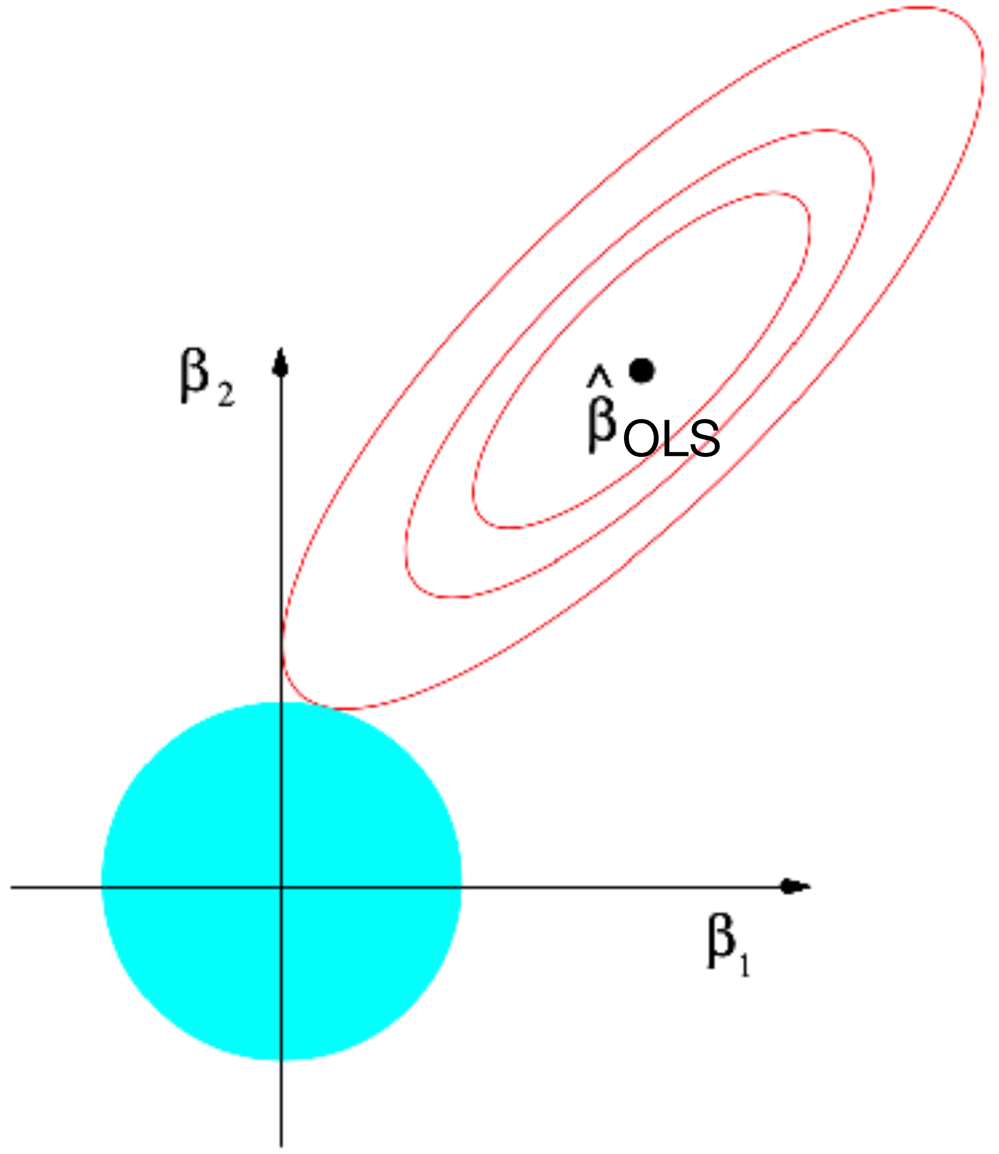
Question: How can we adapt the approach to reduce variance?

Let's add a constraint

Ridge Regression

$$\min_{\beta} \frac{1}{2} \parallel Y - X\beta \parallel_2^2$$

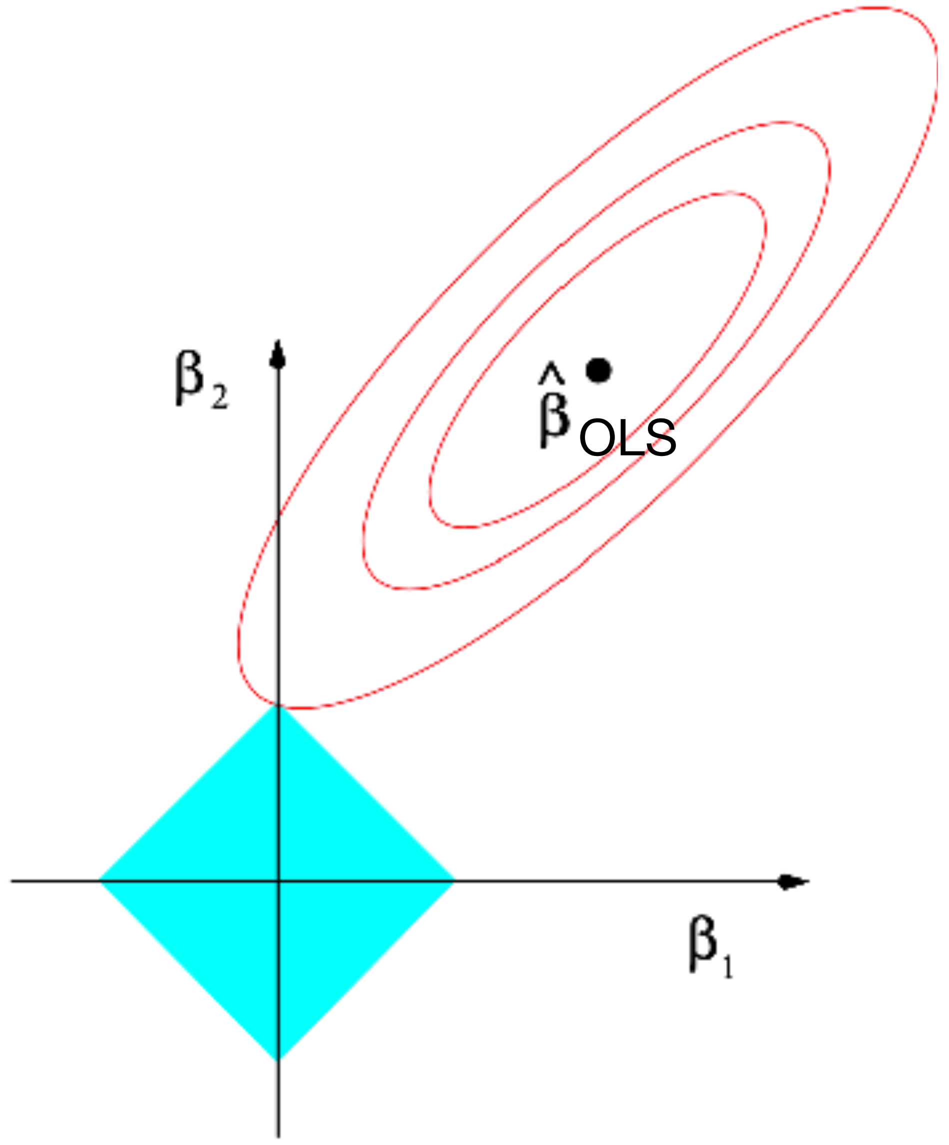
subject to $\sum_{i=1}^{p} |\beta_i|^2 \leq c$



LASSO: least absolute shrinkage and selection operator;

$$\min_{\beta} \frac{1}{2} \parallel Y - X\beta \parallel_2^2$$

subject to $\sum_{i=1}^{p} |\beta_i| \leq c$



Source: Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity: the lasso and generalizations. CRC press.

Example

Sebastian Vollmer

Example: Analysing the crime-rate in US states with respect to education and deprivation

Covariates:

funding: annual police funding in dollars per resident

hs: percent of people 25 years and older with four years of high school

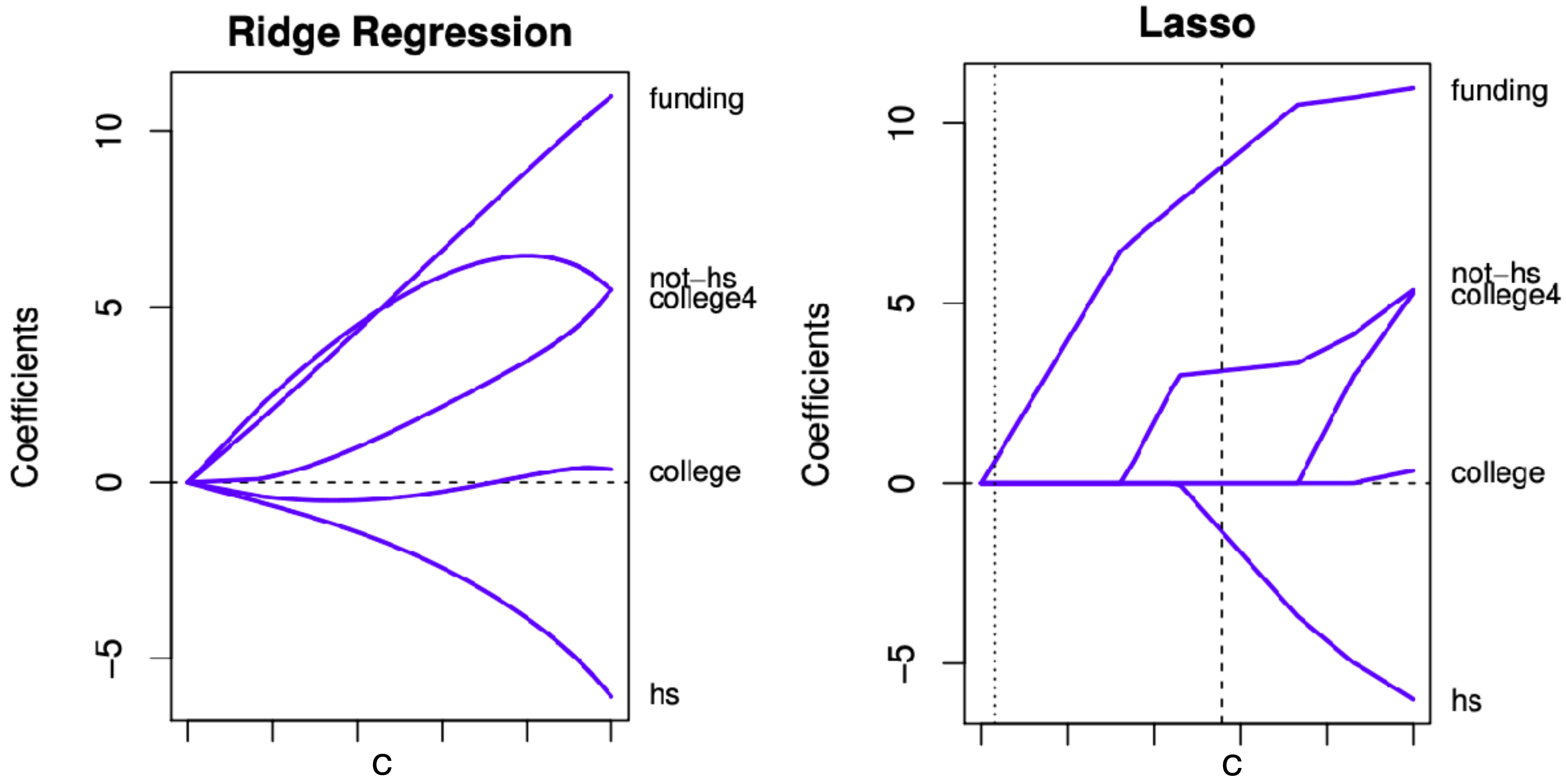
not-hs: percent of 16- to 19-year olds not in high school and not high school graduates

college: percent of 18- to 24-year olds in college

college4 and percent of people 25 years and older with at least four years of college

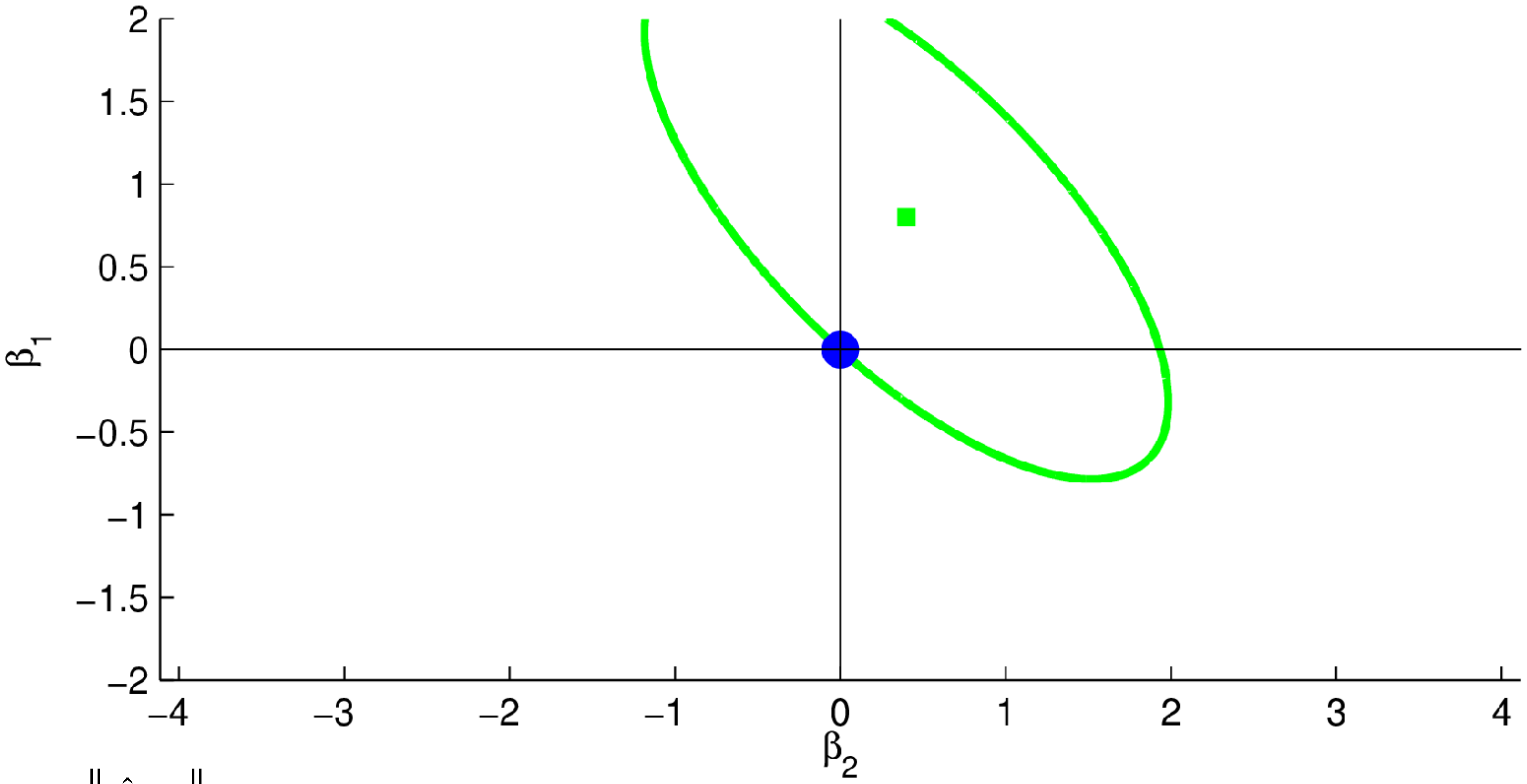
$$\min_{\beta} \frac{1}{2} \| Y - X\beta \|_2^2$$

subject to $\sum_{i=1}^{p} |\beta_i|^2 \leq c$ **subject to** $\sum_{i=1}^{p} |\beta_i| \leq c$



(Hastie et al., “Statistical learning with sparsity: the lasso and generalizations.”)

Varying the constraint $\|\beta\|_1 \leq c$



Notice if $c \geq \left\| \hat{\beta}_{OLS} \right\|_1$,
then the blue area contains $\hat{\beta}_{OLS}$

Level sets of the OLS problem

Border of the L^1 constraint

Solution of the constrained problem

$\hat{\beta}_{OLS}$ minimiser of OLS

Penalised regression

Generalisation: Let Ω be a constraint on β such that

$$\Omega(\beta) \leq c.$$

Examples:

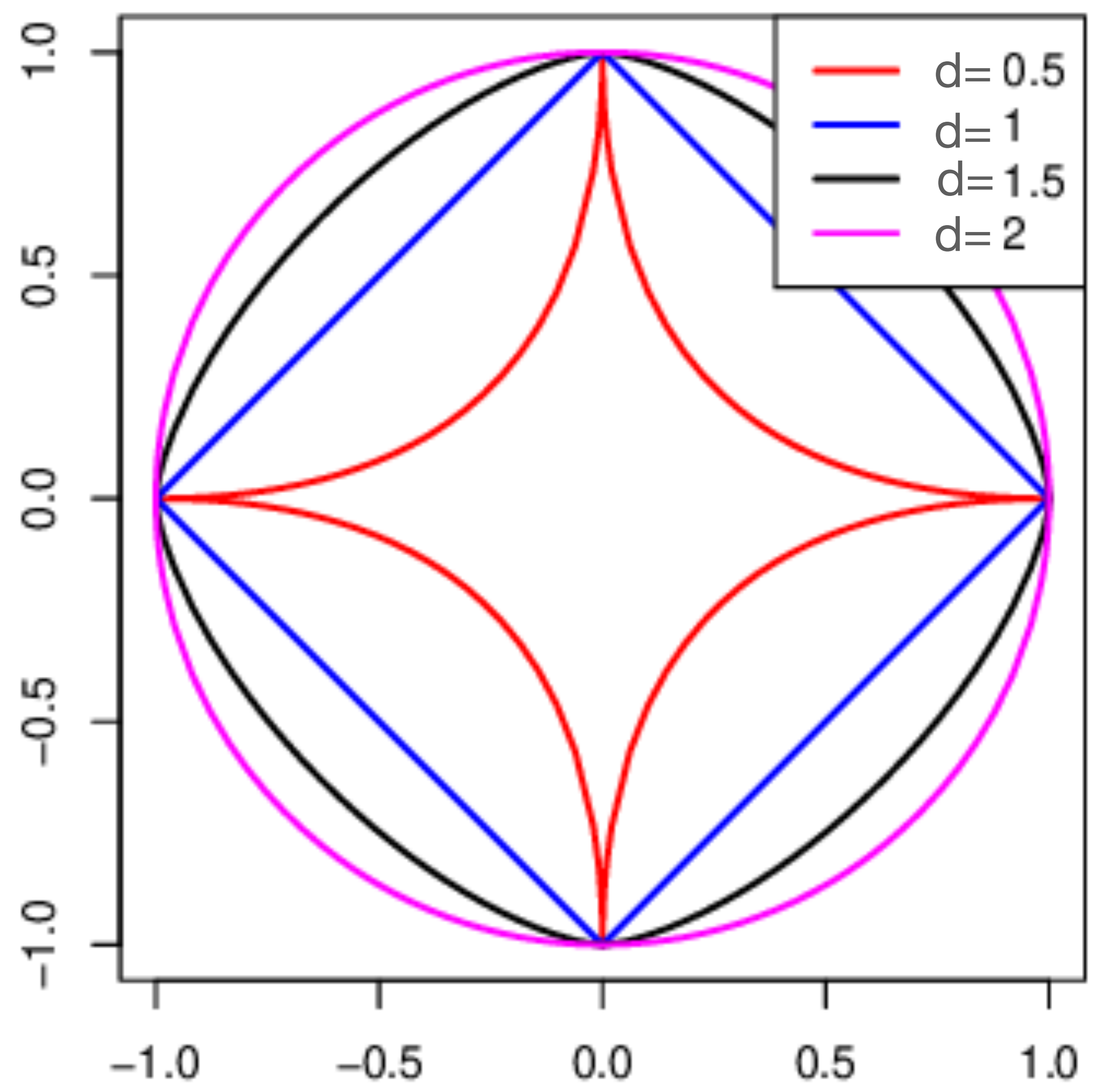
- $\Omega(\beta) = \|\beta\|_0 = \#\{i \mid \beta_i \neq 0\}$

Sparsity: means $\|\beta\|_0 \ll p$
e.g. genes effecting an illness.

- bridge regression

$$\sum_{i=1}^{p} |\beta_p|^d$$

where $d = 1$ **Lasso**
and $d = 2$ **Ridge**



Definition:
A function $f : X \to \mathbb{R}$ is convex if $\forall x_1, x_2 \in X, \forall t \in [0,1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

Remarks:

- $f \in C^2(X)$ convex iff $\nabla^2 f$ is positive-semi-definite
- every local minima is a global minima (proof by contradiction)

Let Ω be convex. Then

$$\min_{\beta} S(\beta)$$
subject to $\Omega(\beta) \leq c$

is equivalent to

$$\min_{\beta} S(\beta) + \lambda\Omega(\beta)$$
for convex Ω

Aim: Explore the computational complexity of solving LASSO

$$\min_{\beta} S(\beta) + \lambda\|\beta\|_1$$

Coordinate descent

AIM: To minimize $f(\beta) = S(\beta) + |\beta|_1 = \frac{1}{2} \parallel Y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_1$

Initialise $\beta_0 \in \mathbb{R}^p$

Repeat

$$\beta_1^{(k)} = \textbf{arg min}_{\beta_1} f(\beta_1, \beta_2^{(k-1)}, \beta_3^{(k-1)}, \ldots, \beta_p^{(k-1)})$$

$$\beta_2^{(k)} = \textbf{arg min}_{\beta_2} f(\beta_1^{(k)}, \beta_2, \beta_3^{(k-1)}, \ldots, \beta_p^{(k-1)})$$

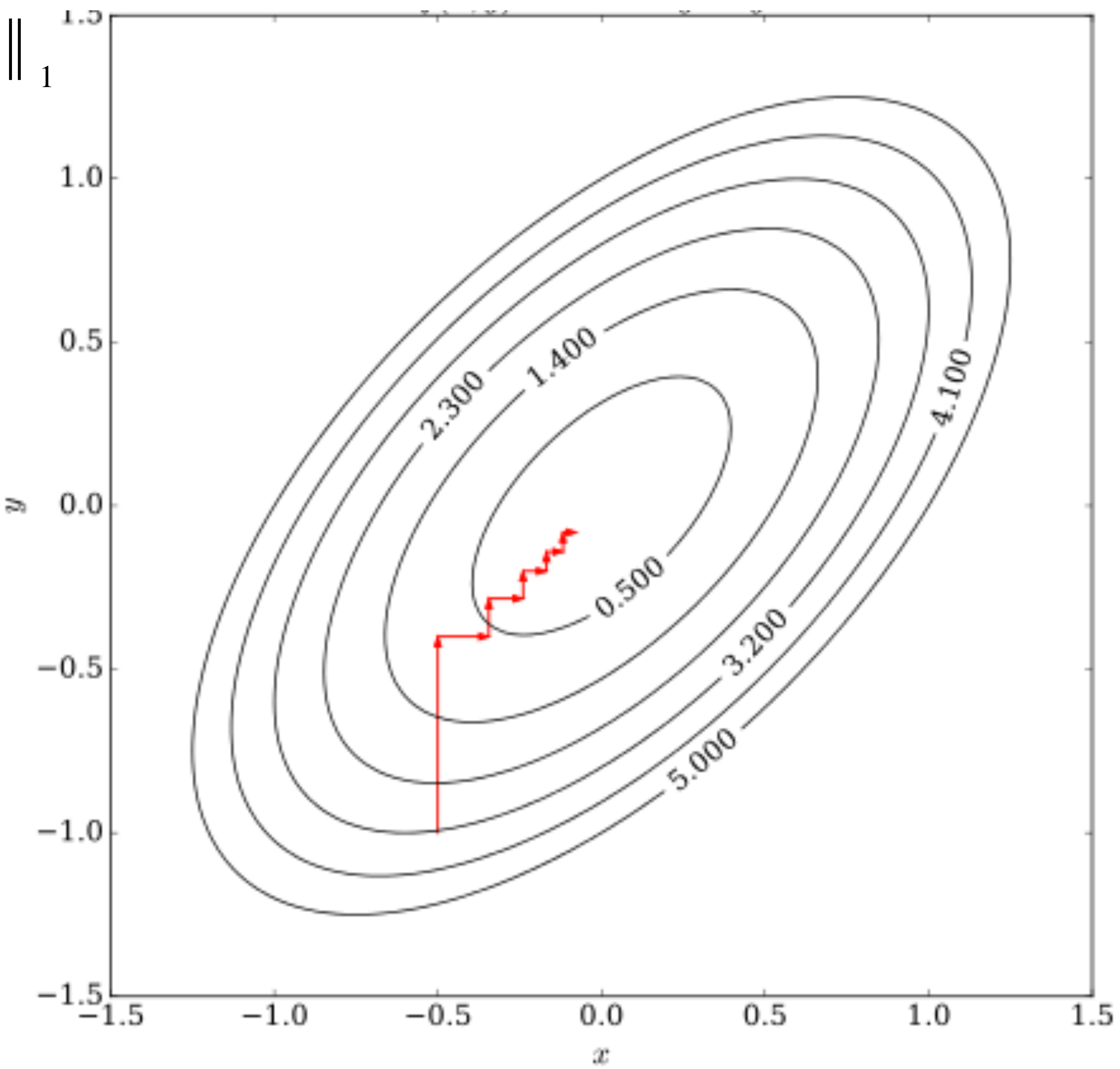
$$\beta_3^{(k)} = \textbf{arg min}_{\beta_3} f(\beta_1^{(k)}, \beta_2^{(k)}, \beta_3, \beta_4^{(k-1)} \ldots, \beta_p^{(k-1)})$$

$$\vdots \quad \vdots$$

$$\beta_p^{(k)} = \textbf{arg min}_{\beta_p} f(\beta_1^{(k)}, \beta_2^{(k)}, \beta_3^{(k)}, \ldots, , \beta_p^{(k)} \beta_p^{(k-1)})$$

until $\|\beta^k - \beta^{k-1}\| \leq \epsilon$

Note: Order can be randomised



Source: WikiCommons

> **Exercise 1:** Given a **convex** differentiable function f, and a point x such that $f(x)$ is minimised along each coordinate axis. Have we found a local minimiser?

Coordinate descent for LASSO

Lemma

The update for the j-th coordinate is given in closed form. For $r_i^{(j)} = y_i - \sum_{k \neq j} x_{ik}\beta_k^t$

$$\beta_j^{t+1} = \frac{1}{\sum_{i=1}^n x_{ij}^2} \begin{cases} 0 & \left| \sum_{i=1}^n r_i^{(j)} x_{ij} \right| < \lambda \\ \sum_{i=1}^n r_i^{(j)} x_{ij} - \lambda & \sum_{i=1}^n r_i^{(j)} x_{ij} \geq \lambda \\ \sum_{i=1}^n r_i^{(j)} x_{ij} + \lambda & \sum_{i=1}^n r_i^{(j)} x_{ij} \leq -\lambda \end{cases}$$

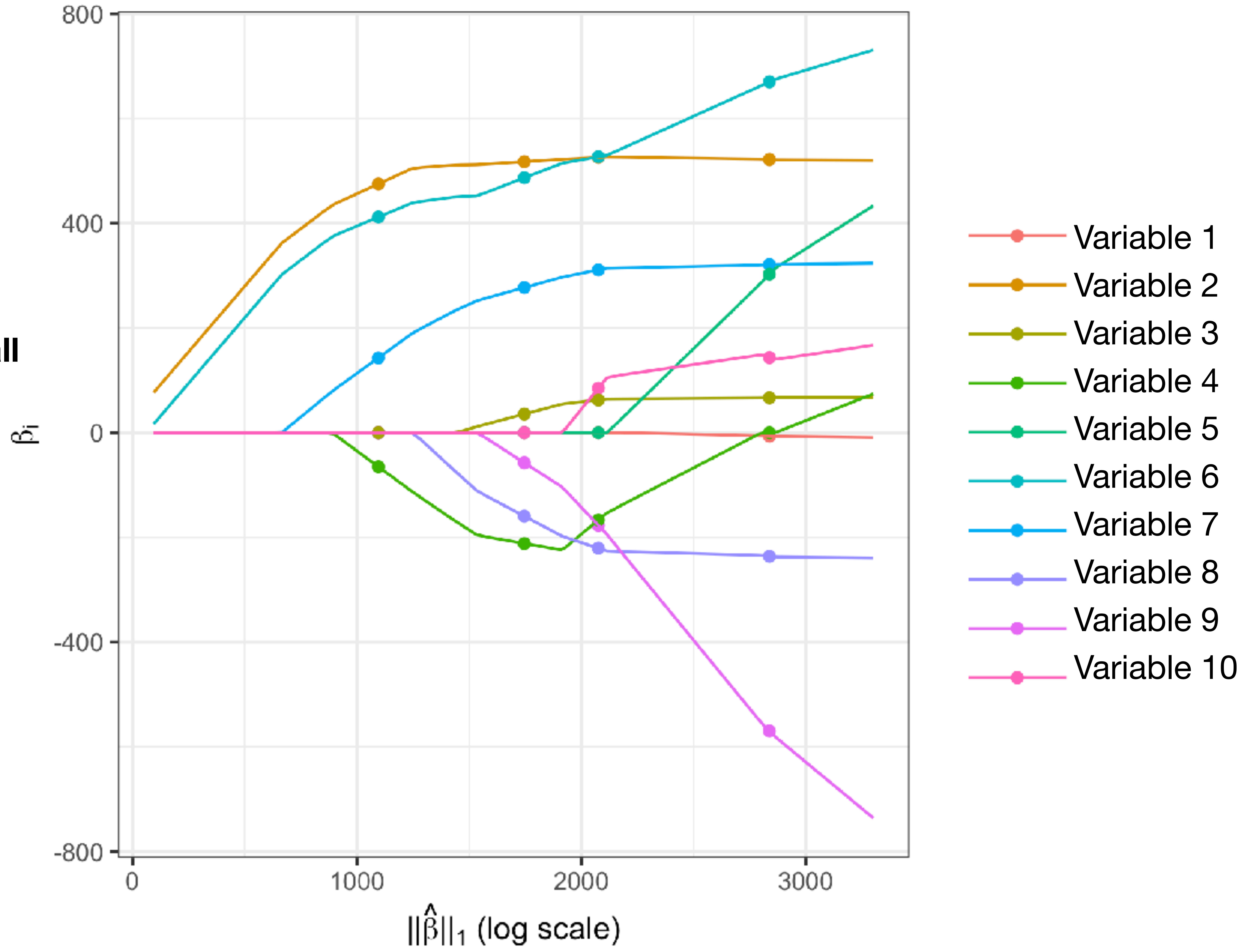
Exercise 2: Finish the proof

LASSO path

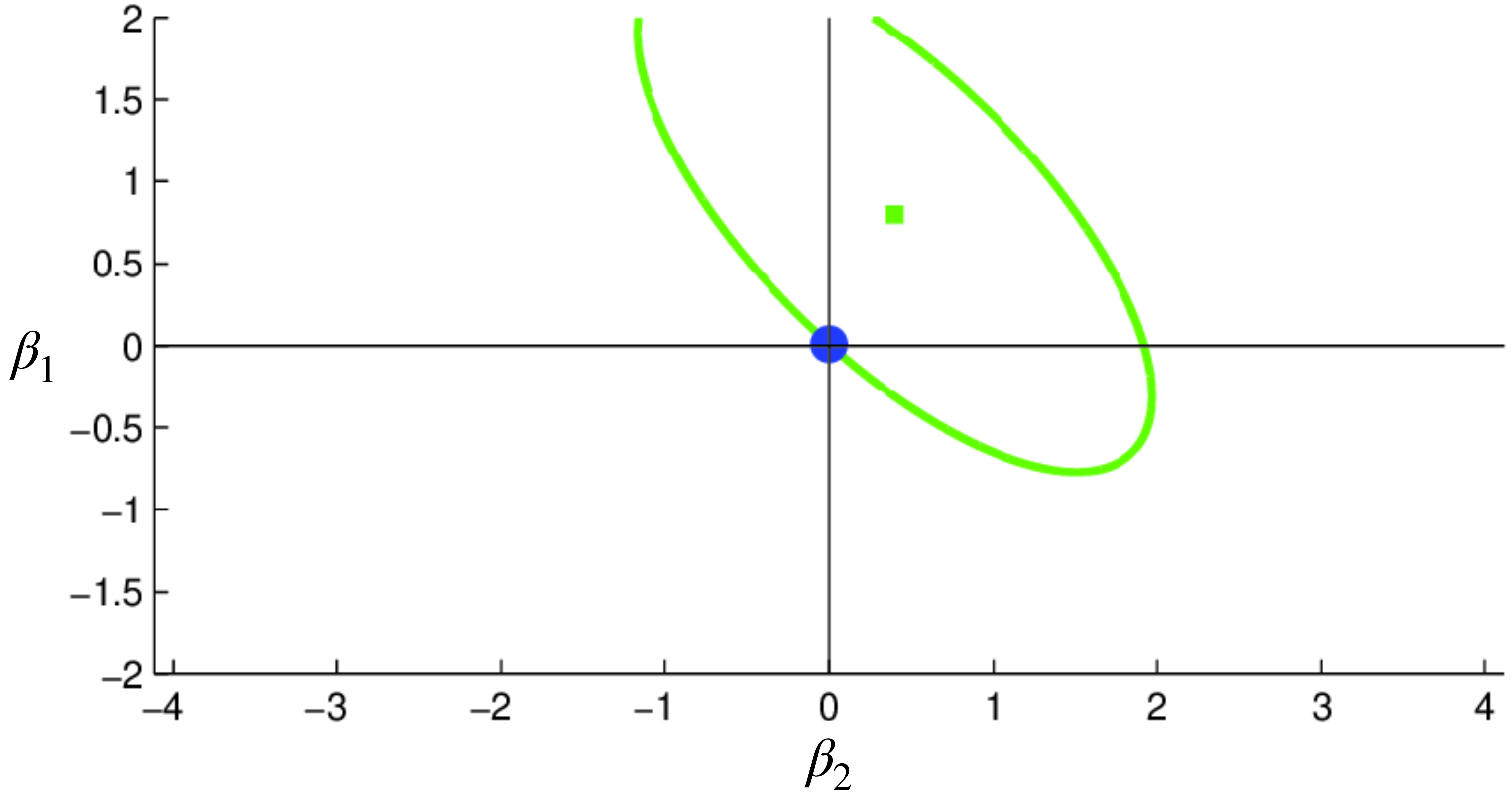
Aim: Find the whole path for all values of λ?

But what about the cost?

After running coordinate descent for multiple λ



Least Angle Regression (Efron et al., 2004)



Let $x_1, \ldots, x_p$ be the columns of X. After a change of coordinates we may assume that $\sum_{i=1}^{n} X_{ij} = 0$ and $\sum_{i=1}^{n} X_{ij}^2 = 1$

$$\min_{\beta} S(\beta) + \lambda \|\beta\|_1$$

- **Let A be the set of active covariates (i.e. those coordinates of β that are currently changing).**
- **Initially, let $A = \{x_{j_1}\}$ with the smallest angle with Y**
- **Step in the direction of x_{j_1} until another predictor enters A (equal same angle).**
- **Continue in the direction such that the angle from x_{j_1} to the residual and x_{j_2} to the residual are equal. Add new predictor x_{j_3} if it has the same same angle and add it to A.**
- **NOTE: For the LASSO direction can drop out of active set.**

Computing the equiangular direction

Recall computational complexity in the big O notation: $f(x)$ **is** $O(g(x))$ **as** $x \to \infty$
$\exists M \exists C > 0$ **such that for all**
x **with** $x_i \geq M$ **for some** $i, |f(x)| \leq C|g(x)|$

Equiangular direction

$$X_A(X_A^t X_A)^{-1} X_A^T (Y - \beta_{current} X)$$

The active set grows $A = \{x_{j_1}\}, A = \{x_{j_1}, x_{j_2}\}, \ldots, A = \{x_{j_1}, x_{j_2}, \ldots, x_{j_k}\}$ **for** $k = 1, \ldots, p$

Before: Computation of $(X_A^T X_A)^{-1}$ $O(k^3 + \ldots)$
Cholesky decomposition
$X^T X = LL^T$ with L being lower triangular $Lx = \begin{pmatrix} l_{11} & & \\ l_{21} & l_{22} & \\ l_{i1} & \cdots & l_{n} \end{pmatrix} x = Y$

Thus solving linear equation $O(k^2)$

We need to compute L at cost $O(k^3)$.

But we can use Matrix algebra to update L

Empirical complexity for n=p

Estimation of polynomial complexity

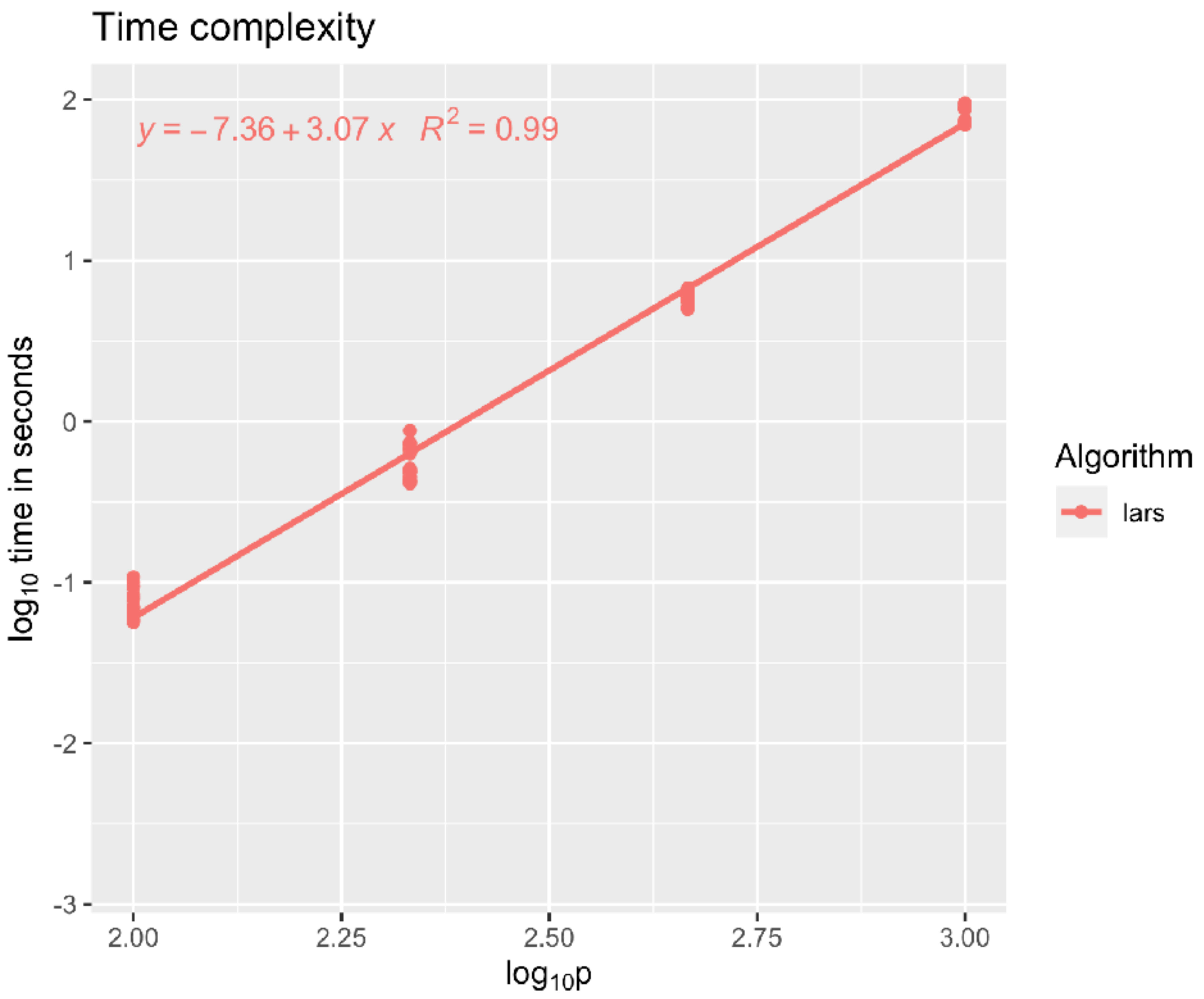
$$\log y = a \log x + b \Leftrightarrow y = C \cdot x^a$$

```r
library(tidyverse)
library(lars)
library(microbenchmark)
library(matlab)

eps=0.2;p=10000;n=20
ps=round(logspace(2,4,n=4))
df=data.frame()
for(p in ps){
  n=p

truth=matrix(1.0*rbernoulli(p,p=0.4),ncol=1)*rnorm(n=
p)
  X=matrix(rnorm(n=p*n),ncol=p)
  y=as.numeric(X %*% truth[1:p]+eps*rnorm(n=n))

df=bind_rows(df,data.frame(n=n,p=p,t=microbenchmark(
   lars(X,y,type="lasso",max.steps = 100*p),times =
   R),
   alg="lars"))
}
```



Suggests $\mathcal{O}(p^3)$ instead of $\mathcal{O}(p^4)$

Computing the equiangular direction

Dr. Sebastian Vollmer

Recall: the computational complexity in the "big O notation"

$f(x)$ is $O(g(x))$ as $x \to \infty$

if a fixed multiple of $g(x)$ is an upper bound for large values of x

The active set grows $A = \{x_{j_1}\}, A = \{x_{j_1}, x_{j_2}\}, \ldots, A = \{x_{j_1}, x_{j_2}, \ldots, x_{j_k}\}$ **for** $k = 1,\ldots,p$

Equiangular direction (check)

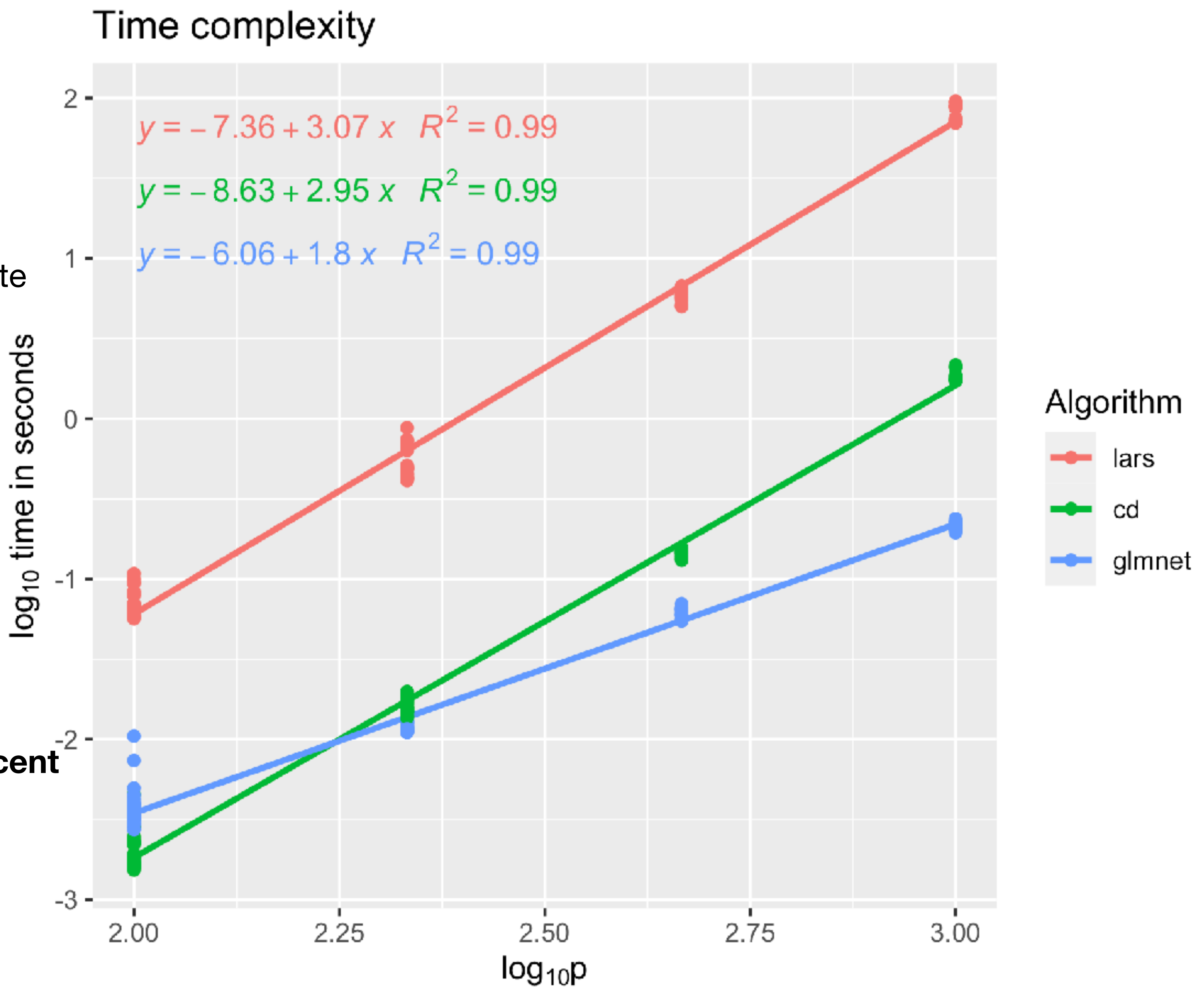
$$X_A(X_A^t X_A)^{-1} X_A^T (Y - \beta_{current} X)$$

More details in Tibshirani, R., Johnstone, I., Hastie, T. & Efron, B. Least angle regression. The Annals of Statistics vol. 32 407–499 (2004)

Empirical comparison LARS vs Coordinate Descent

- Comparison
  - Lars to compute entire path
  - Coordinate descent for a single $\lambda = 1$ -
    - stopping criteria parameter changes less than 1e-

- R-package glmnet (Friedman et. all) is based on coordinate descent to compute the entire LASSO path. Speedup is gained through
  - warm starts
  - “strong rules” - temporarily leave out portion of variables



GLMNET is much faster and based on coordinate descent

Practical considerations

Problem

Assume we find in multiple linear regression on the weather data the following parameters

X_1	LUZ_pressure	[hPa]	$\theta_1 = -1$	[km/h/hPa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

We could have measured the pressure in Pa and get the equivalent result

X_1	LUZ_pressure	[Pa]	$\theta_1 = -1/100$	[km/h/Pa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

With regularization $\lambda(\theta_1^2 + \theta_2^2)$ we would get different results for measurements in hPa and in Pa, because θ_1 contributes less to the penalty in the latter case.

Solution

Standardize all predictors, such that they have mean 0 and variance 1:

$$\tilde{X}_i = (X_i - \bar{X}_i)/\sqrt{\mathrm{Var}(X_i)}$$

Practical considerations

Problem

Assume we find in multiple linear regression on the weather data the following parameters

X_1	LUZ_pressure	[hPa]	$\theta_1 = -1$	[km/h/hPa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

We could have measured the pressure in Pa and get the equivalent result

X_1	LUZ_pressure	[Pa]	$\theta_1 = -1/100$	[km/h/Pa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

With regularization $\lambda(\theta_1^2 + \theta_2^2)$ we would get different results for measurements in hPa and in Pa, because θ_1 contributes less to the penalty in the latter case.

Solution

Standardize all predictors, such that they have mean 0 and variance 1:

$$\tilde{X}_i = (X_i - \bar{X}_i)/\sqrt{\mathrm{Var}(X_i)}$$

Practical considerations

With loss $L(\theta) = \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda\|\theta\|_2^2$
the effective regularization depends on the size of the data set.

One can use instead an average loss $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda\|\theta\|_2^2$ or
(equivalently) scale the regularization term $L(\theta) = \sum_{i=1}^{n} \ell(y_i, f(x_i)) + n \cdot \lambda\|\theta\|_2^2$

Summary

Today's lecture

1. Penalised regression and its induced sparsity
2. Coordinate descent
3. LARS and the importance of getting linear algebra right

Further topics:

Non-convex penalises

Thank you for your attention

coordinate descent being stuck

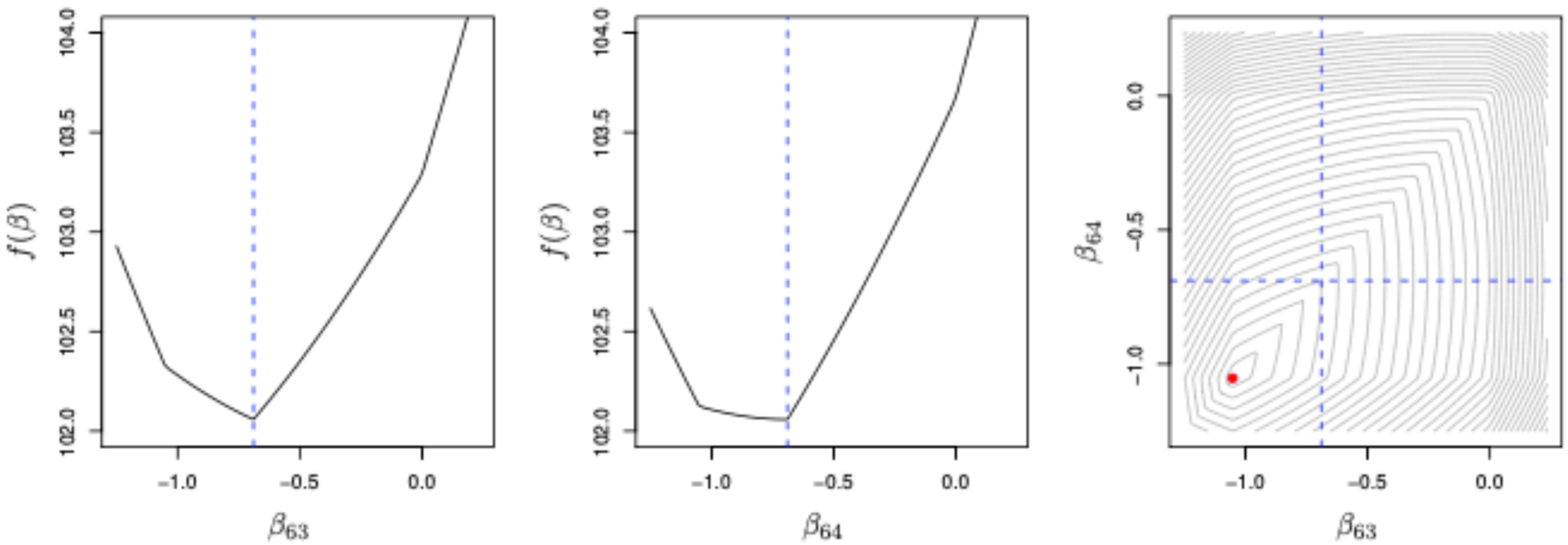
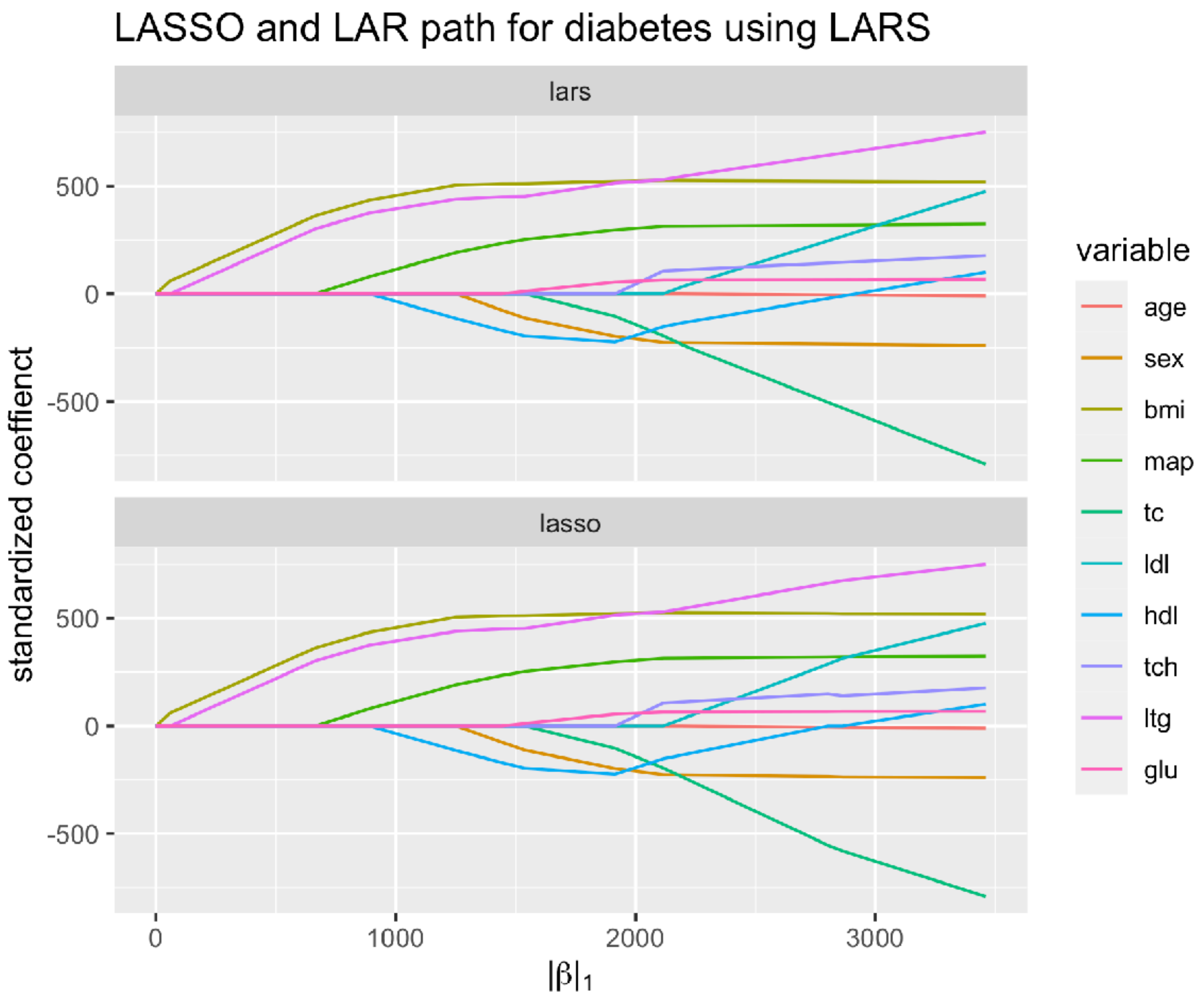


Figure 5.8 *Failure of coordinate-wise descent in a fused lasso problem with 100 parameters. The optimal values for two of the parameters, β_{63} and β_{64}, are both -1.05, as shown by the dot in the right panel. The left and middle panels show slices of the objective function f as a function of β_{63} and β_{64}, with the other parameters set to the global minimizers. The coordinate-wise minimizer over both β_{63} and β_{64} (separately) is -0.69, rather than -1.05. The right panel shows contours of the two-dimensional surface. The coordinate-descent algorithm is stuck at the point $(-0.69, -0.69)$. Despite being strictly convex, the surface has corners, in which the*

LARS vs LASSO



Cholesky updating

$$\begin{pmatrix} X^t \\ v^t \end{pmatrix} (X \; v) = \begin{pmatrix} X^tX & X^tv \\ v^tX & v^tv \end{pmatrix}$$

X^tX contains

$= (\langle x_i, x_j \rangle)$

$O(k \cdot n + k^2)$

1) For solving linear equation

2) rank-one update $O(k^2)$

Given $\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{13} \\ \mathbf{A}_{13}^{\mathrm{T}} & \mathbf{A}_{33} \end{pmatrix}$ $\quad \mathbf{L} = \begin{pmatrix} \mathbf{L}_{11} & \mathbf{L}_{13} \\ 0 & \mathbf{L}_{33} \end{pmatrix}$,

Update $\tilde{\mathbf{A}} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \mathbf{A}_{12}^{\mathrm{T}} & \mathbf{A}_{22} & \mathbf{A}_{23} \\ \mathbf{A}_{13}^{\mathrm{T}} & \mathbf{A}_{23}^{\mathrm{T}} & \mathbf{A}_{33} \end{pmatrix}$ $\quad \tilde{\mathbf{S}} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \mathbf{S}_{13} \\ 0 & \mathbf{S}_{22} & \mathbf{S}_{23} \\ 0 & 0 & \mathbf{S}_{33} \end{pmatrix}$.

$$\begin{aligned}
\mathbf{S}_{11} &= \mathbf{L}_{11}, \\
\mathbf{S}_{12} &= \mathbf{L}_{11}^{\mathrm{T}} \backslash \mathbf{A}_{12}, \\
\mathbf{S}_{13} &= \mathbf{L}_{13}, \\
\mathbf{S}_{22} &= \textbf{chol}(\mathbf{A}_{22} - \mathbf{S}_{12}^{\mathrm{T}} \mathbf{S}_{12}), \\
\mathbf{S}_{23} &= \mathbf{S}_{22}^{\mathrm{T}} \backslash (\mathbf{A}_{23} - \mathbf{S}_{12}^{\mathrm{T}} \mathbf{S}_{13}), \\
\mathbf{S}_{33} &= \textbf{chol}(\mathbf{L}_{33}^{\mathrm{T}} \mathbf{L}_{33} - \mathbf{S}_{23}^{\mathrm{T}} \mathbf{S}_{23}).
\end{aligned}$$

Title Text

Sketch of Proof: Let us consider (1) first. Let $\hat{\boldsymbol{\theta}}_{Ridge}$ be the minimum of $g_\lambda(\boldsymbol{\theta})$. Necessarily, the gradient of g_λ at $\hat{\boldsymbol{\theta}}_{Ridge}$ is 0:

$$\nabla g_\lambda(\hat{\boldsymbol{\theta}}_{Ridge}) = -2\boldsymbol{y}^T\boldsymbol{X} + 2(\hat{\boldsymbol{\theta}}_{Ridge})^\top\boldsymbol{X}^T\boldsymbol{X} + 2\lambda(\hat{\boldsymbol{\theta}}_{Ridge})^\top = 0.$$

We show that we can find a value t such that $\hat{\boldsymbol{\theta}}_{Ridge}$ is also the optimal solution to problem (2).

Problem

Assume we find in multiple linear regression on the weather data the following parameters

X_1	LUZ_pressure	[hPa]	$\theta_1 = -1$	[km/h/hPa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

We could have measured the pressure in Pa and get the equivalent result

X_1	LUZ_pressure	[Pa]	$\theta_1 = -1/100$	[km/h/Pa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

With regularization $\lambda(\theta_1^2 + \theta_2^2)$ we would get different results for measurements in hPa and in Pa, because θ_1 contributes less to the penalty in the latter case.

Solution

Standardize all predictors, such that they have mean 0 and variance 1:

$$\tilde{X}_i = (X_i - \bar{X}_i)/\sqrt{Var(X_i)}$$

Convergence

$$\frac{m\,d\,\beta\,\|\mathbf{w}^{\star}\|_2^2}{\varepsilon},$$

$O\left(m\,d\,\beta\,\|\mathbf{w}^{\star}\|_1^2/\varepsilon\right).$

-
-

Title Text

Dr. Sebastian Vollmer

We calculate the Lagrangian of (2)

$$L(\boldsymbol{\theta}, \alpha) \quad = \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 + \alpha(\|\boldsymbol{\theta}\|_2^2 - t).$$

The first KKT condition says:

$$\nabla_\theta L(\boldsymbol{\theta}, \alpha) = -2\boldsymbol{y}^T\boldsymbol{X} + 2\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X} + 2\alpha\boldsymbol{\theta}^T = 0.$$

Since $\nabla g_\lambda(\hat{\boldsymbol{\theta}}_{Ridge}) = 0$, this condition is satisfied if we set $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{Ridge}$ and $\alpha = \lambda$.

The KKT-conditions also require that complementarity is fulfilled:

$$\alpha(\|\boldsymbol{\theta}\|_2^2 - t) = 0.$$

This is satisfied if we set $t = \|\hat{\boldsymbol{\theta}}_{Ridge}\|^2$.

The converse is also true: The optimal solution to problem (2) is also a solution to problem (1) if we set $\lambda = \alpha$.