# The Lasso for Linear Models

Isabel Stransky, Camilla Gerboth

24 September 2018

# Content

# Linear Regression

Given:

- N samples $\{(x_i, y_i)\}_{i=1}^{N}$
- $x_i = (x_{i1}, ..., x_{ip})$ p-dimensional vector of predictors and each $y_i \in \mathbb{R}$ is the associated response variable

**Goal:**

approximate the response variable $y_i$ using a linear combination of the predictors

# Linear Regression

## Definition

Linear Regression Model:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i$$

- $\beta_0$ and $\beta = (\beta_1, ..., \beta_p)$: unknown parameters
- $e_i$ : error term

## Definition:

Method of least-squares:

$$\underset{\beta_0,\beta}{\text{minimize}} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

**Remarks**:

- if the least-squares estimates will be nonzero and statistically significant $\Longrightarrow$ interpretation difficult if p large
- if $p > N \Longrightarrow$ least-squares estimates are not unique (infinite set of solutions)

**Problem:**

"We are drowning in information and starving for knowledge."

$\longrightarrow$ there is a crucial need to sort through this mass of information
$\longrightarrow$ we need to hope that the complex processes can be described using relatively simple models

**Example:**

We hope that not all of the approx. 30'000 genes in the human body are directly involved in the process that leads to the development of cancer.

**Question:**

Why do we might want to consider an alternative to the least-squares estimate?

1. **Prediction accuracy:**
   - the least-squares estimate often has low bias but large variance
   - prediction accuracy can sometimes be improved by shrinking the values of the regression coefficients or setting some coefficients to zero
   - then the bias increases but the variance of the predicted values decreases
2. **Interpretation:** with a large number of predictors, we often would like to identify a smaller subset of these predictors that exhibit the strongest effects

# Lasso estimator

## Definition

given $N$ predictor-response pairs $\{(x_i, y_i)_{i=1}^N$, the lasso finds the solution $(\hat{\beta}_0, \hat{\beta})$ to the optimization problem

$$\underset{\beta_0, \beta}{\text{minimize}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

rewritten with $l_1$-norm

$$||\beta||_1 \leq t$$

# Lasso estimator

## Definition

Lagrangian form:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} ||\boldsymbol{y} - \boldsymbol{X}\beta||_2^2 + \lambda ||\beta||_1$$

for some $\lambda \geq 0$

- $||\cdot||_2$: Euclidean norm
- $\boldsymbol{y} = (y_1, ..., y_N)$ N-vector of responses
- $\boldsymbol{X}$ $N \times p$ matrix with $x_i \in \mathbb{R}^p$ in its $i^{th}$ row

- first term is a measure of the fit of the model to the data
- second term is a penalty term
- Goal: to get the fit as small as possible, and at the same time get the penalty part as small as possible
- in order to get the first part as small as possible, the formula tells us that we should have as many $\beta$'s as possible to get a good approximation $\Longrightarrow$ penalty value in the second part will be large

# Lasso estimator

**Remark:**

- if $\lambda = 0$: penalty term has no effect $\hat{\beta}^L = \hat{\beta}^{LS}$

# Lasso estimator

1)

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}$$

subject to

$$||\beta||_1 \leq t$$

2)

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} ||\boldsymbol{y} - \boldsymbol{X}\beta||_2^2 + \lambda||\beta||_1$$

by Lagrangian duality, there is a one-to-one correspondence between the constrained problem 1) and the Lagrangian form 2):

$\Longrightarrow$ for each value of $t$, there is a corresponding value of $\lambda$ that yields the same solution from the Lagrangian form and vice versa

# Bound

## Definition

bound $t$ is kind of a "budget":
since a shrunken parameter estimate corresponds to a more heavily-constrained model, this budget limits how well we can fit the data

**Remark:**
budget must be specified separately (see later)

# Ridge estimator

Ridge regression uses a similar criterion but with the $l_2$-norm as constraint

## Definition

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}$$

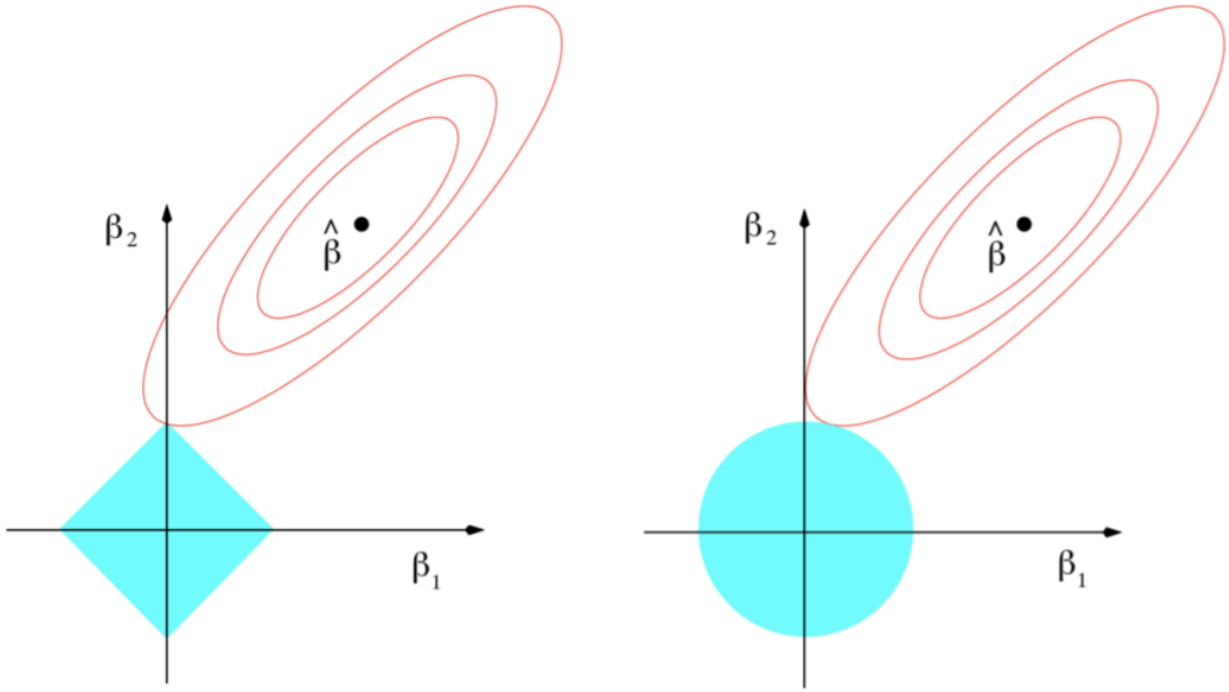
subject to

$$\sum_{j=1}^{p} \beta_j^2 \leq t$$

**Question:**

What is the difference between the lasso and ridge regression?

# Estimation picture for the lasso (left) and ridge regression (right)




Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the Lasso and generalizations*. CRC Press, 2015.

- solid blue areas: constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t$
- red ellipses: contours of the residual sum of squares
- $\hat{\beta}$: usual (unconstrained) least-squares estimate

$\Longrightarrow$ both methods find the first point where the elliptical contours hit the constraint region

**Difference:**
unlike the disk, the diamond has corners
$\Longrightarrow$ if the solution occurs at a corner, then it has one coefficient $\beta_j$ equal to zero

# Sparsity

**Definition**

a sparse statistical model is one with only a few nonzero coefficients

$\Longrightarrow$ **the lasso yields sparse models but ridge regression does not**

# Bound

**Recall:**
bound $t$ in the lasso criterion controls complexity of the model

**Question:**

What happens for larger values of $t$?

more coefficients are free up and allow the model to adapt more closely to the data

**Question:**

What happens for smaller values of $t$?

coefficients are more restricted $\Longrightarrow$ sparser, more interpretable models that fit data less closely

# Bound

looking for the value of $t$ that gives the most accurate model for predicting independent test data from the same population

**Question**

How can we find this best value for $t$?

$\Longrightarrow$ can use Cross-Validation

# Cross-Validation

**Definition**

Cross-Validation is used to estimate the test error associated with a given statistical learning method in order to evaluate its performance or to select the appropriate level of flexibility

# Cross-Validation

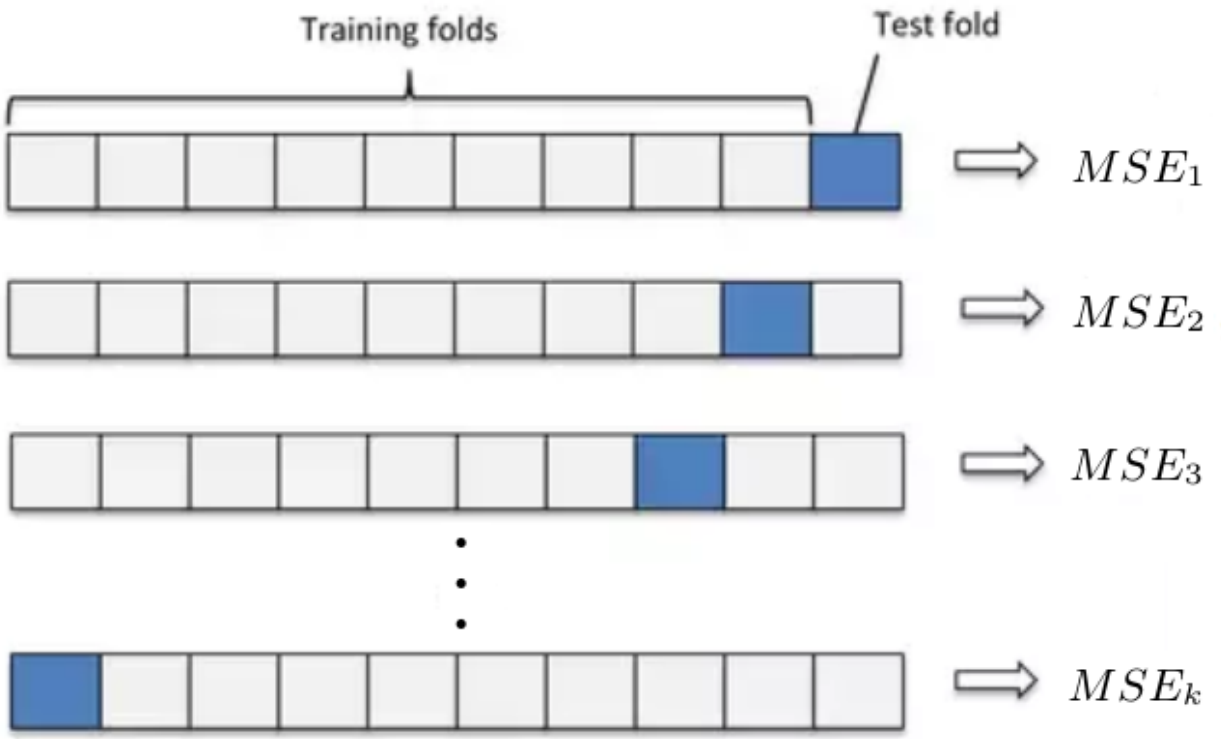
**Procedure:**

- randomly divide the full dataset into $k > 1$ folds
  (typical choices are $k = 5$ or $k = 10$)
- fix one fold as test set and remaining $k - 1$ folds as training sets
- apply the lasso to the training data for a range of different $t$ values
- use each fitted model to predict the response in the test set
- determine the mean-squared prediction errors for each value of $t$
  Mean-squared error for test fold $j$:

$$MSE_j = \frac{1}{|F_j|} \sum_{i \in F_j} (y_i - \hat{f}(x_i))^2$$

- repeat process $k$ times such that each fold is once the test set
- obtain $k$ estimates of the prediction error over a range of values of $t$

# Cross-Validation




Source: https://medium.com/@sebastiannorena/some-model-tuning-methods-bfef3e6544f0

# Cross-Validation

- average $k$ mean-squared errors for each value of $t$

$$CV_{(k)}(t) = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

- obtain cross-validation error curve
- in order to choose which $t$ is best we take the value of $t$ for which the CV-error $CV_{(k)}(t)$ is smallest

**One standard error rule:**
we take the simplest (most regularized) model whose error is within one standard error of the minimal CV-error

# Example

**Setup:**

Look at baseball data from 1986-1987. The Hitters dataset contains information about 322 baseball players and 20 attributes as follows:

1. AtBat: Number of assists in 1986
2. Hits: Number of hits in 1986
3. HmRun: Number of home runs in 1986

⋮

**target variable of interest:** the players' salaries

# Uniqueness of the Lasso Solutions

- if the columns of $\boldsymbol{X}$ are in general position (columns $\{x_j\}_{j=1}^{p}$ are in general position if any affine subspace $\mathbb{L} \subset \mathbb{R}^N$ of dimension $k < N$ contains at most $k + 1$ elements of the set $\{\pm x_1, ..., \pm x_p\}$)
  $\Longrightarrow$ for $\lambda > 0$ the solution to the lasso problem is unique
- also true for $p \geq N$, although then the number of nonzero coefficients in any lasso solution is at most $N$
- if the predictor matrix $\boldsymbol{X}$ is not of full column rank, then the parameter estimates are not unique

**Question**

When can the non-full rank case occur?

- if $p \leq N$, due to collinearity
- it always occurs if $p > N$ (infinite number of solutions)

# Consistency

Assumption: $\boldsymbol{X}$ is fixed

If $\beta^*$ and $\hat{\beta}$ are the true and the lasso-estimated parameters, it can be shown that as $p, N \to \infty$

$$\frac{||\boldsymbol{X}(\hat{\beta} - \beta^*)||_2^2}{N} \leq C \cdot ||\beta^*||_1 \sqrt{\frac{log(p)}{N}}$$

with high probability.

Thus if $||\beta^*||_1 = o(\sqrt{\frac{N}{log(p)}})$ (i.e. the true parameter vector must be sparse relative to the ratio $\frac{N}{log(p)}$), then the lasso is consistent for prediction in terms of the MSE.

# Recap

**Recall:**

## Definition

Standard linear regression model:

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

where $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ (model matrix), $\boldsymbol{w} \in \mathbb{R}^N$ (vector of noise variables), and $\beta^* \in \mathbb{R}^p$ (unknown coefficient vector)

## Definition

Constrained form of the Lasso:

$$\underset{||\beta||_1 \leq R}{\text{minimize}} ||\boldsymbol{y} - \boldsymbol{X}\beta||_2^2$$

## Definition

Lagrangian form:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \{||\boldsymbol{y} - \boldsymbol{X}\beta||_2^2 + \lambda_N ||\beta||_1\}$$

for some $\lambda_N \geq 0$

# Loss Functions

Given lasso estimate $\hat{\beta} \in \mathbb{R}^p$ we want to assess it's quality with two loss functions:

**Definition**

Prediction loss function:

$$\mathcal{L}_{\text{pred}}(\hat{\beta}, \beta^*) = \frac{1}{N}||\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta^*||_2^2$$

- corresponds to mean-squared error of $\hat{\beta}$ over given samples of $\boldsymbol{X}$
- helpful if interested in predictive performance of $\hat{\beta}$

More appropriate if $\beta^*$ is of primary interest:

**Definition**

Parameter estimation loss ($\ell_2$-error):

$$\mathcal{L}_2(\hat{\beta}, \beta^*) = ||\hat{\beta} - \beta^*||_2^2$$

# Sparsity Models

- Classical analysis of a method such as lasso would fix number of covariates $p$, and then take sample size $N$ to infinity.
- Often $p$ of same order or substantially larger than $N$.
- Want to come up with theory for $p >> N$.
- **Note**: If model lacks any additional structure $\Longrightarrow$ not able to recover useful information
- Indeed: $N \leq p \implies$ linear model is unidentifiable (i.e. the solution is not unique).
- Thus: add additional constraints on unknown regression vector $\beta^* \in \mathbb{R}^p$

# Sparsity Models

**Sparsity constraints:**

**Definition**

**Hard sparsity**: Assume $\beta^*$ has at most $k \leq p$ nonzero entries.

Can consider prediction and $\ell_2$-norm losses in this case.

**Definition**

**Weak sparsity**: Assume $\beta^*$ can be closely approximated by vectors with few nonzero entries. Formalization:
For a parameter $q \in [0, 1]$ and radius $R_q > 0$, define the set

$$\mathbb{B}(R_q) = \{\beta \in \mathbb{R}^q | \sum_{j=1}^{p} |\beta_i|^q \leq R_q\}$$

# Bounds on Lasso $\ell_2$-Error

- Now: Some results on the $\ell_2$-norm loss between a lasso solution $\hat{\beta}$ and the true regression vector $\beta^*$.
- Consider $\beta^*$ $k$-sparse, i.e. its entries are nonzero on a subset $S(\beta^*) \subset \{1, 2, ..., p\}$ of cardinality $k$.

# Strong Convexity in the Classical Setting

- Want to establish conditions on the model matrix $\boldsymbol{X}$ that are needed to establish bounds on $\ell_2$-error.
- For intuition: Consider one route for proving $\ell_2$-consistency where $p$ is fixed, $N$ tends to infinity.
- Suppose we estimate some parameter vector $\beta^*$ by minimizing a data-dependent objective function $f_N(\beta)$ over some constraint set.
- Suppose the difference in function values $\Delta f_N = |f_N(\hat{\beta}) - f_N(\beta^*)|$ converges to zero as sample size $N$ increases.

**Question:**

What additional conditions are necessary to ensure the $\ell_2$-norm of $\Delta\beta = ||\hat{\beta} - \beta^*||_2$ also converges to zero?

Answer: **Strong convexity**. Because if the function is strongly convex, then $\Delta\beta = ||\hat{\beta} - \beta^*||_2$ also converges to zero.

# Strong Convexity in Classical Setting

## Definition

**Strong convexity:**

- $f : \mathbb{R}^p \to \mathbb{R}$ differentiable function, is strongly convex with parameter $\gamma > 0$ at $\theta \in \mathbb{R}^p$ if $\forall \theta' \in \mathbb{R}^p : f(\theta') - f(\theta) \geq \nabla f(\theta)^T(\theta' - \theta) + \frac{\gamma}{2}||\theta' - \theta||_2^2$.
- If $f : \mathbb{R}^p \to \mathbb{R}$ is twice continuously differentiable, then:

$$f \text{ strongly convex with parameter } \gamma \text{ around } \beta^* \in \mathbb{R}^p$$

$$\Longleftrightarrow$$

$$\text{minimum eigenvalue of the Hessian matrix } \nabla^2 f(\beta) \text{ is at least } \gamma$$

$$\text{for all vectors } \beta \text{ in a neighbourhood of } \beta^*$$

# Strong Convexity in Classical Setting

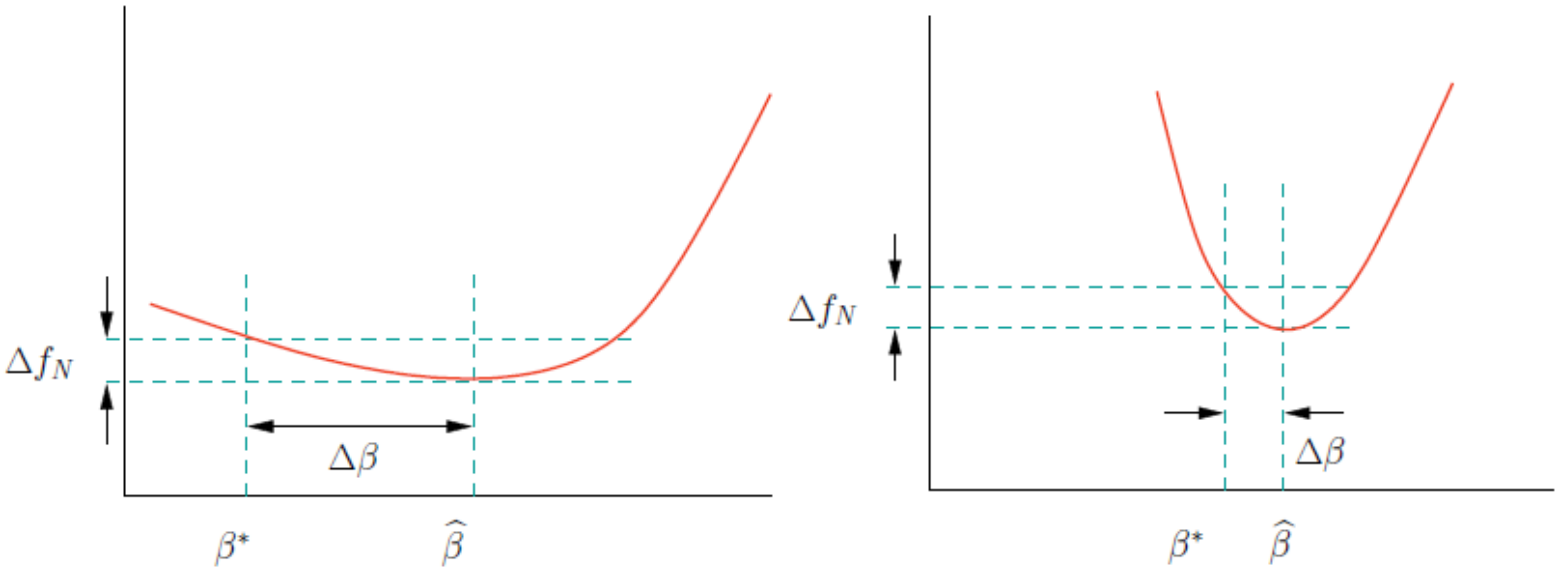



**Figure 11.2** *Relation between differences in objective function values and differences in parameter values. Left: the function $f_N$ is relatively "flat" around its optimum $\widehat{\beta}$, so that a small function difference $\Delta f_N = |f_N(\widehat{\beta}) - f_N(\beta^*)|$ does not imply that $\Delta\beta = \|\widehat{\beta} - \beta^*\|_2$ is small. Right: the function $f_N$ is strongly curved around its optimum, so that a small difference $\Delta f_N$ in function values translates into a small difference in parameter values.*

Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the Lasso and generalizations*. CRC Press, 2015.

# Restricted Eigenvalues for Regression

- Return to high-dimensional setting, i.e. number of parameters $p$ might be larger than sample size $N$.
- In this setting the least-squares objective function $f_N(\beta) = \frac{1}{2N}||\boldsymbol{y} - \boldsymbol{X}\beta||_2^2$ is always convex.

**Question:**

Under what conditions is the objective function $f_N(\beta)$ also strongly convex?

**Answer:** We can observe $\nabla^2 f(\beta) = \boldsymbol{X}^T\boldsymbol{X}/N$ for all $\beta \in \mathbb{R}^p$. Hence, the least-square loss is strongly convex iff eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$ are uniformly bounded away from zero.

# Restricted Eigenvalues for Regression

- **Problem:** Any matrix of the form $\boldsymbol{X}^T\boldsymbol{X}$ has rank $\leq \min\{N, p\}$, so it is always rank-deficient, hence not strongly convex, whenever $N < p$.
- Need to relax our notion of strong convexity.
- Only need to impose a type of strong convexity condition for some subset $C \subset \mathbb{R}^p$ of possible perturbation vectors $\nu \in \mathbb{R}^p$ (as we will see soon).

# Restricted Eigenvalues for Regression

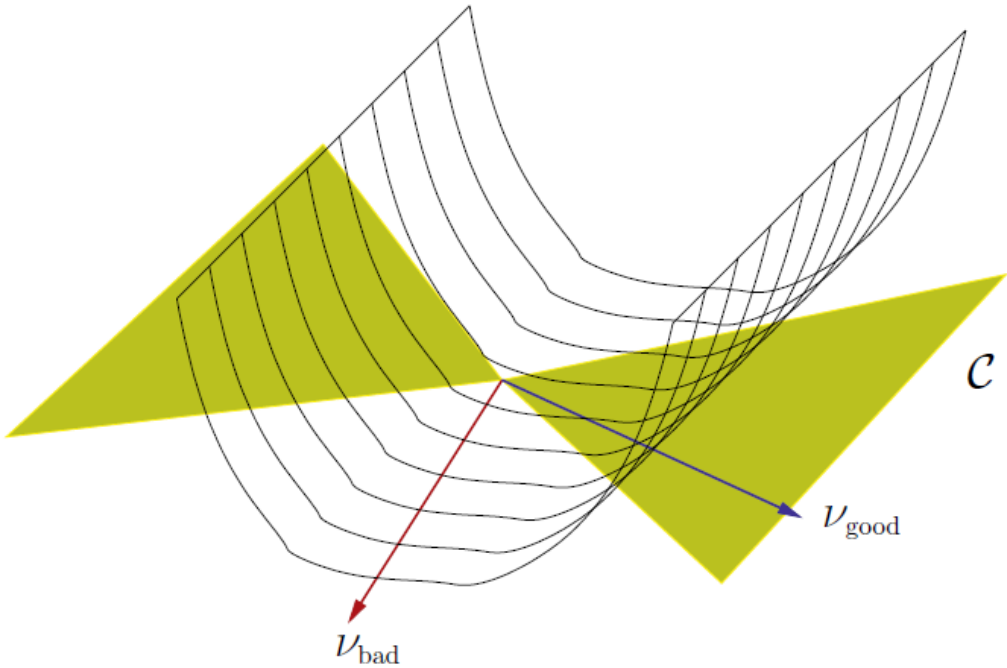



**Figure 11.3** *A convex loss function in high-dimensional settings (with $p \gg N$) cannot be strongly convex; rather, it will be curved in some directions but flat in others. As shown in Lemma 11.1, the lasso error $\widehat{\nu} = \widehat{\beta} - \beta^*$ must lie in a restricted subset $\mathcal{C}$ of $\mathbb{R}^p$. For this reason, it is only necessary that the loss function be curved in certain directions of space.*

Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the Lasso and generalizations*. CRC Press, 2015.

# Restricted Eigenvalues for Regression

**Definition**

A function $f$ satisfies **restricted strong convexity** at $\beta^*$ with respect to $C$ if there is a constant $\gamma > 0$ such that

$$\frac{\nu^T \nabla^2 f(\beta) \nu}{||\nu||_2^2} \geq \gamma \text{ for all } \nu \in C,$$

and for all $\beta \in \mathbb{R}^p$ in a neighbourhood of $\beta^*$.

In the specific case of linear regression, this is equivalent to lower bounding the **restricted eigenvalues** of the model matrix, in particular requiring:

$$\frac{\frac{1}{N}\nu \boldsymbol{X}^T \boldsymbol{X} \nu}{||\nu||_2^2} \geq \gamma \text{ for all nonzero } \nu \in C.$$

# Restricted Eigenvalues for Regression

**Question:**

What constraint sets $C$ are relevant?

**Answer:** $C(S, \alpha) := \{\nu \in \mathbb{R}^p \mid ||\nu_{S^c}||_1 \leq \alpha ||\nu_S||_1\}$

# Restricted Eigenvalues for Regression

**Derivation:**

- Suppose the parameter vector $\beta^*$ is sparse, supported on subset $S = S(\beta^*)$.
- Define the lasso error $\hat{\nu} = \hat{\beta} - \beta^*$.
- Let $\hat{\nu_S} \in \mathbb{R}^{|S|}$ denote the subvector indexed by elements of $S$, with $\hat{\nu}_{S^c}$ defined analogously.

**Lemma**

For appropriate choices of $\ell_1$-ball radius, or equivalently of the regularization parameter $\lambda_N$, the lasso error satisfies a **cone constraint** of the form

$$||\hat{\nu}_{S^c}||_1 \leq \alpha ||\hat{\nu}_S||_1$$

for some constant $\alpha \geq 1$.

# Restricted Eigenvalues for Regression

**Conclusion:**
In its constrained or regularized form, the lasso error is restricted to a set of the form

$$C(S;\alpha) := \{\nu \in \mathbb{R}^p \mid ||\nu_{S^c}||_1 \leq \alpha||\nu_S||_1, \}$$

for some parameter $\alpha \geq 1$.
(Which is what we were looking for when asking which constraint sets $S$ are relevant.)

# Basic Consistency Result

Take a look at a result that provides a bound on the lasso error $||\hat{\beta} - \beta^*||_2$, based on the linear model $\boldsymbol{y} = \boldsymbol{X}\beta^* + \boldsymbol{w}$, where $\beta^*$ is $k$-sparse, supported on the subset $S$.

## Theorem 11.1

Suppose the model matrix $\boldsymbol{X}$ satisfies the restricted eigenvalue bound with parameter $\gamma > 0$ over $C(S; 3)$.Then:

(a) Any estimate $\hat{\beta}$ based on the constrained lasso with $R = ||\beta^*||_1$ satisfies the bound

$$||\hat{\beta} - \beta^*||_2 \leq \frac{4}{\gamma}\sqrt{\frac{k}{N}}\left\|\frac{\boldsymbol{X}^T\boldsymbol{w}}{\sqrt{N}}\right\|_\infty$$

(b) Given a regularization parameter $\lambda_N \geq 2||\boldsymbol{X}^T\boldsymbol{w}||_\infty/N > 0$, any estimate $\hat{\beta}$ from the regularized lasso satisfies the bound

$$||\hat{\beta} - \beta^*||_2 \leq \frac{3}{\gamma}\sqrt{\frac{k}{N}}\sqrt{N}\lambda_N$$

# Basic Consistency Result

- These results are deterministic.
- They apply to any set of linear regression equations with a given observed noise vector $w$.
- Assumptions on the noise vector $w$ and/or the model matrix affect the rate through the restricted eigenvalue constant $\gamma$, and the terms $||\boldsymbol{X}^T \boldsymbol{w}||_\infty$ and $\lambda_N$ in the two bounds.
- The two terms $||\boldsymbol{X}^T \boldsymbol{w}||_\infty$ and $\lambda_N$ reflect the interaction of the observation noise $\boldsymbol{w}$ with the model matrix $\boldsymbol{X}$.

# Example: Classical Linear Gaussian Model

- Let the observation noise $\boldsymbol{w} \in \mathbb{R}^N$ be Gaussian, with i.i.d. $N(0, \sigma^2)$ entries.
- Fix the design matrix $\boldsymbol{X}$, with columns $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_p\}$.
- Then for any given column $j \in \{1, ..., p\}$ the random variable $\boldsymbol{x}_j^T \boldsymbol{w}/N$ is distributed as $N(0, \frac{\sigma^2}{N} \cdot \frac{||\boldsymbol{x}_j||_2^2}{N})$.
- Hence, if the columns of the design matrix $\boldsymbol{X}$ are normalized, (i.e. $||\boldsymbol{x}_j||_2/\sqrt{N} = 1 \ \forall j \in \{1, ..., p\}$), then the variable $\boldsymbol{x}_j^T \boldsymbol{w}/N$ is stochastically dominated by a $N(0, \frac{\sigma^2}{N})$ variable, so that we have the

**Gaussian tail bound:**

$$\mathbb{P}\left[\frac{|\boldsymbol{x}_j^T \boldsymbol{w}|}{N} \geq t\right] \leq 2e^{-\frac{Nt^2}{2\sigma^2}}$$

# Basic Consistency Result

- Since $\frac{||\boldsymbol{X}^T\boldsymbol{w}||_\infty}{N}$ corresponds to the maximum over $p$ such variables, the union bound yields

$$\mathbb{P}\left[\frac{||\boldsymbol{X}^T\boldsymbol{w}||_\infty}{N} \geq t\right] \leq 2e^{-\frac{Nt^2}{2\sigma^2}+\log p} = 2e^{-\frac{1}{2}(\tau-2)\log p},$$

  when we set $t = \sigma\sqrt{\frac{\tau \log p}{N}}$ for some $\tau > 2$.
- Conclusion: The lasso error satisfies the bound

$$||\hat{\beta} - \beta^*||_2 \leq \frac{c\sigma}{\gamma}\sqrt{\frac{\tau k \log p}{N}}$$

  with probability at least $1 - 2e^{-\frac{1}{2}(\tau-2)\log p}$.
- Gives us a valid choice of regularization parameter $\lambda_N$ which is valid for Theorem 11.1(b).

  (Namely $\lambda_N = 2\sigma\sqrt{\tau\frac{\log p}{N}}$ for some $\tau > 2$ is a valid choice with same high probability.)

# Basic Consistency Result

The rate $\frac{c\sigma}{\gamma}\sqrt{\frac{\tau k \log p}{N}}$ is reasonable:

- Suppose $S(\beta^*)$, the support set, would be known.
- Then estimation of $\beta^*$ would require approximating a total of $k$ parameters, namely the elements $\beta_i^*$ for all $i \in S(\beta^*)$.
- But even with knowledge of support set, since model has $k$ free parameters, no method can achieve squared $\ell_2$-error that decays more quickly than $\frac{k}{N}$.

  $\Longrightarrow$ apart from logarithmic factor, lasso rate matches the best possible one could achieve, even if $S(\beta^*)$ were known a priori.
- In fact, the rate $\frac{c\sigma}{\gamma}\sqrt{\frac{\tau k \log p}{N}}$ cannot be substantially improved by any estimator.

# Bounds on Prediction Error

- So far we studied performance of lasso in recovering true regression vector, as assessed by $||\hat{\beta} - \beta^*||_2$.
- Now: theoretical guarentees on relatively low (in-sample) prediction error $\mathcal{L}_{pred}(\hat{\beta}, \beta^*) = \frac{1}{N}||\boldsymbol{X}(\hat{\beta} - \beta^*)||_2^2$
- We consider Lagrangian lasso, but analogous results could be derived for other forms of lasso.

# Bounds on Prediction Error

**Theorem 11.2**

Consider the Lagrangian lasso with regularization parameter $\lambda_N \geq \frac{2}{N}||\boldsymbol{X}^T\boldsymbol{w}||_\infty$.

(a) If $||\beta^*||_1 \leq R_1$, then any optimal solution $\hat{\beta}$ satisfies

$$\frac{||\boldsymbol{X}(\hat{\beta} - \beta^*)||_2^2}{N} \leq 12 R_1 \lambda_N$$

(b) If $\beta^*$ is supported on a subset $S$, and the design matrix $\boldsymbol{X}$ satisfies the restricted eigenvalue bound over $C(S; 3)$, then any optimal solution $\hat{\beta}$ satisfies

$$\frac{||\boldsymbol{X}(\hat{\beta} - \beta^*)||_2^2}{N} \leq \frac{144}{\gamma}|S|\lambda_N^2.$$

# Bounds on Prediction Error

- As before, the choice $\lambda_N = c\sigma\sqrt{\frac{\log p}{N}}$ is valid for Theorem 11.2 with high probability, hence the two bounds take the form

$$\frac{||\boldsymbol{X}(\hat{\beta} - \beta^*)||_2^2}{N} \leq c_1 \sigma R_1 \sqrt{\frac{\log p}{N}}$$

$$\frac{||\boldsymbol{X}(\hat{\beta} - \beta^*)||_2^2}{N} \leq c_2 \frac{\sigma^2}{\gamma} \frac{|S| \log p}{N}.$$

- The former bound is known as the 'slow rate' for the lasso, since the squared prediction error decays as $1/\sqrt{N}$.
- The latter bound is known as the 'fast rate' since it decays as $1/N$.
- The latter is based on much stronger Assumptions:
  - hard sparsity condition: $\beta^*$ is supported on a small subset $S$
  - restricted eigenvalue bound on design matrix $\boldsymbol{X}$

# Summary

- **Interpretation of the final model:**
  the $l_1$-penalty provides a natural way to encourage sparsity and simplicity in the solution
- **Statistical efficiency:**
  if the true underlying model is sparse, then the lasso works well
  if the true underlying model is not sparse, then the lasso will not work well
- **Computational efficiency:**
  resulting optimization problem is convex & can be solved efficiently for large problems

# References


- Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the Lasso and generalizations*. CRC Press, 2015.
- G. James, D. Witten, T. Hasti, R. Tibshirani: *An Introduction to Statistical Learning. With Applications in R*. Springer.

# Questions