Lecture 5: Sparse Penalties Beyond Lasso

Statistical Learning and Data Mining

Xingye Qiao

Department of Mathematical Sciences
Binghamton University

E-mail: qiao@math.binghamton.edu

Read: ESLII Chs. 3.6 & 3.8 and SLS Ch. 6

Outline

The next section would be . . . . . .

Replace the ℓ_1 penalty

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\operatorname{argmin}} \; RSS + \lambda P(\boldsymbol{\beta})$$

- For Lasso, $P(\boldsymbol{\beta}) = \sum_j |\beta_j|$.
- For Ridge regression, $P(\boldsymbol{\beta}) = \sum_j \beta_j^2$.
- In general, assume the penalty takes a simple form

$$P(\boldsymbol{\beta}) = \sum_j P(\beta_j)$$

Variable selection consistency

Suppose $\hat{\boldsymbol{\beta}}_n$ is the coefficient estimate from a certain procedure. Let $\mathcal{A}_n = \{j : \hat{\beta}_j \neq 0\}$ be the indices for non-zero estimates, and $\mathcal{A} = \{j : \beta_j \neq 0\}$ be the true non-zero variables.

To say that this procedure has **variable selection consistency** means that

$$\lim_{n \to \infty} P(\mathcal{A}_n = \mathcal{A}) = 1$$

Lasso can be improved.

- There are scenarios in which a necessary condition for the variable selection consistency for Lasso cannot be satisfied.
- We introduce **adaptive Lasso**:

$$P(\beta_j) = w_j|\beta_j|$$

where the weight $w_j = 1/|\tilde{\beta}_j|^\nu$, $\nu > 0$ and $\tilde{\beta}_j$ is a $\sqrt{n}$-consistent estimator of the true β_j.

  - If $p < n$, then can use OLS for the pilot estimate $\tilde{\beta}_j$.
  - If $p > n$, can use univariate regression coefficients (i.e., regress Y on X_j individually.)

Why Adaptive Lasso?

- Adaptive Lasso is variable selection consistent under more general conditions than Lasso.
- Key idea: penalize more those estimates which are more likely to be zero.
- An approximation to the ℓ_q penalties with $q = 1 - \nu$, which has better variable selection performance than ℓ_1. Recall the star-shaped ℓ_q ball.
- On the other hand, at the second step (after the pilot estimate), we still have an convex optimization (hence easy to compute.)

Combining ℓ_1 and ℓ_2 penalties

- ℓ_1 (lasso): leads to sparse solution. But may arbitrarily select one variable among a group of highly correlated ones.
- ℓ_2 (ridge): shrink correlated variables simultaneously (hence avoid shrinking only one of them). But does not conduct variable selection.
- Elastic net makes a compromise between the ridge and the lasso penalties.
- Introduce the elastic net penalty:

$$P(\beta_j) = \frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|$$

where an additional tuning parameter $\alpha \in [0, 1]$ is a slider between both penalties.

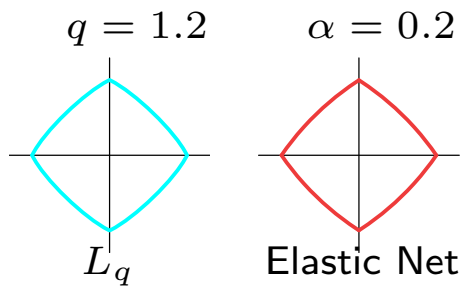


FIGURE 3.13. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for* $q = 1.2$ *(left plot), and the elastic-net penalty* $\sum_j (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$ *(right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the* $q = 1.2$ *penalty does not.*

Advantage of Elastic net

- Selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge.
- Considerable computational advantages over the ℓ_q penalties.
- Can be used with any linear model, in particular for regression or classification.

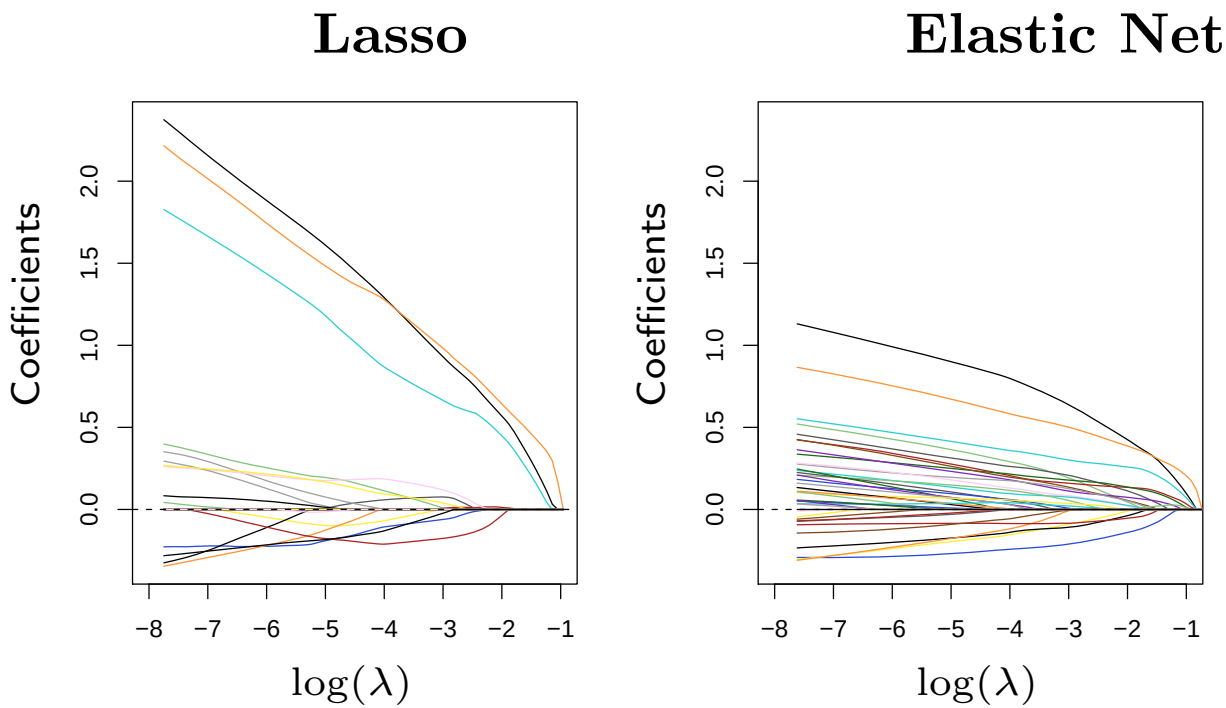


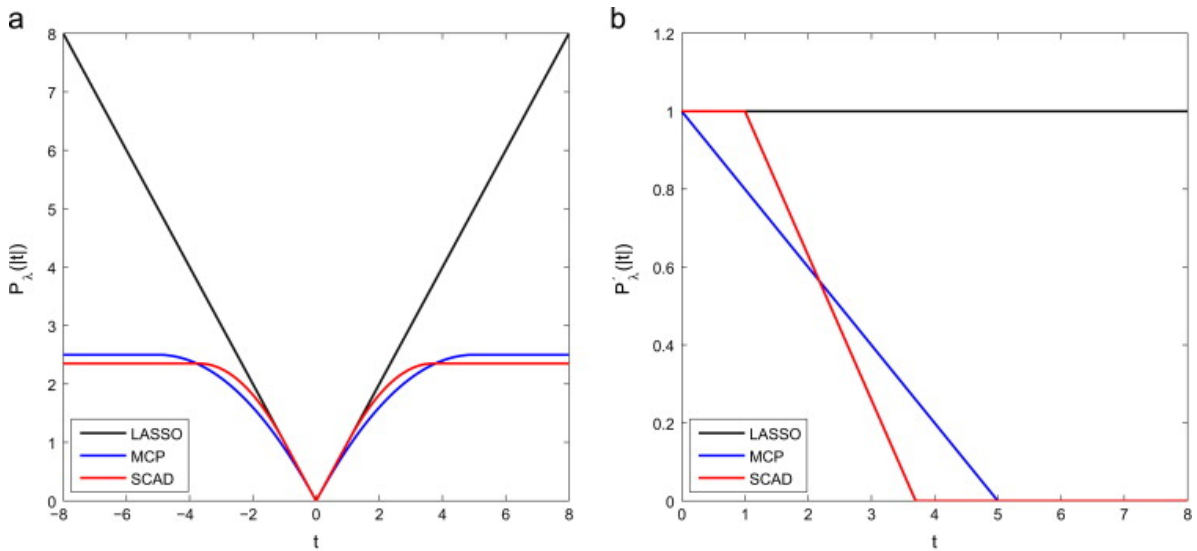
FIGURE 18.5. *Regularized logistic regression paths for the leukemia data. The left panel is the lasso path, the right panel the elastic-net path with* $\alpha = 0.8$. *At the ends of the path (extreme left), there are* 19 *nonzero coefficients for the lasso, and* 39 *for the elastic net. The averaging effect of the elastic net results in more non-zero coefficients than the lasso, but with smaller magnitudes.*

Nonconvex Penalties

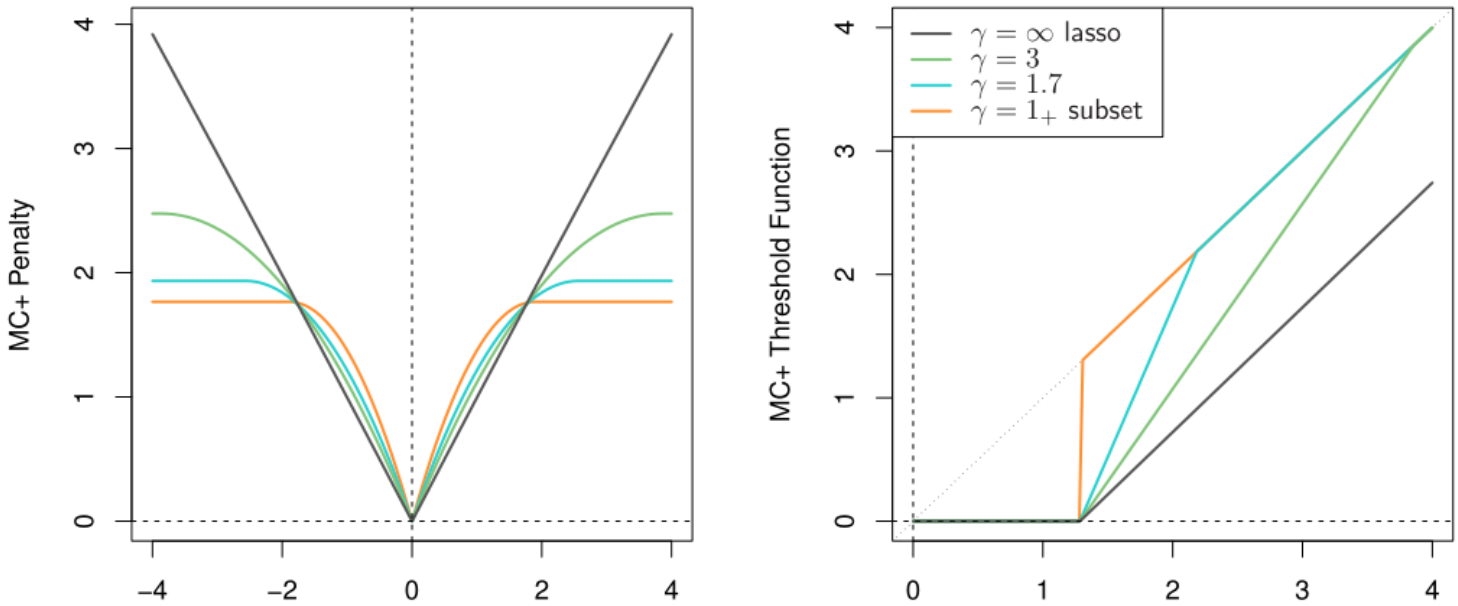
- Lasso, ridge and elastic net penalties are all convex penalties.
- Adaptive lasso tries to approximate ℓ_q penalty $(0 < q < 1)$; But it is a two-step procedure.
- When p is very large, there is an interest in nonconvex penalties
  - Natural choice is ℓ_q penalty $(0 < q < 1)$: spiky nonconvex nature of the ℓ_q ball favors variable selection.
  - Huge computational complexity.

Alternative nonconvex penalties

- SCAD (*Smoothly Clipped Absolute Deviation*) penalty
- MC+ (minimax concave and PLUS: penalized linear unbiased selection algorithm) penalty
- Both are concave (on the positive/negative half of the domain), and flat for large $|\beta_j|$.



More on MC+



As γ grows from 1 to ∞, the penalty function turns from subset selection (hard thresholding) to Lasso (soft-thresholding.)

- Small values of $\tilde{\beta}_j$ are set to 0.
- Large values are left alone.
- Intermediate zone: shrunk by an affine function.
- As γ gets smaller, the intermediate zone gets narrower.

The next section would be . . . . . .

Why Group Lasso?

- Features may be structurally grouped.
- E.g. dummy variables that are used to code a multilevel categorical predictor, or parallel sets of coefficients in a multiple regression problem.
- Natural to select or omit all the coefficients within a group together.
- Group lasso and overlap group lasso achieve these effects by using sums of (un-squared) ℓ_2 penalties.

Group Lasso

- Linear regression setting.
- Suppose that among p variables, there are L groups. Use $\mathbf{X}_\ell$ to represent the design matrix corresponding to the ℓth group, $\boldsymbol{\beta}_\ell$ the corresponding coefficient vector, and p_ℓ its size.

$$P(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L) = \sum_{\ell=1}^{L} \sqrt{p_\ell} \|\boldsymbol{\beta}_\ell\|_2$$

- $\sqrt{p_\ell}$ counts for different group size effect. (subjective choice)
- Note that $\|\boldsymbol{\beta}_\ell\|_2 = 0$ if and only if all $\beta_{\ell,j} = 0$.
- For some λ value, $\|\boldsymbol{\beta}_\ell\|_2$ is shrunk to 0, leading to a whole group of zero coefficients.
- Generalizations: more general norms such as $(\boldsymbol{\beta}^T \mathbf{K} \boldsymbol{\beta})^{1/2}$; allow overlapping groups of predictors.

Example Applications of Group Lasso

- Dummy variable due to multilevel factors
- Highly correlated genes from the same biological pathways
- Multivariate regression: $\mathbf{Y} \in \mathbb{R}^{n \times K}$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{B} \in \mathbb{R}^{p \times K}$.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

it is natural to solve the K regression problems jointly, imposing some type of group structure on the coefficients. For example, there may be unknown subsets of the covariates that are relevant for prediction, preserved across all K components of the response, as well as subsets of variables that are irrelevant for all K components.

Total variation regularization

Consider a 1-D signal $y_1, \ldots, y_n$.

- Total variation is defined as $\sum_{i=1}^{n} |y_i - y_{i+1}|$
- Consider an input signal $x_1, \ldots, x_n$ with large total variation.
- Goal: an approximated signal (close to the input) that has small total variation

$$\min_{y_1,\ldots,y_n} \sum_i (x_i - y_i)^2 + \lambda \sum_{i=1}^{n} |y_i - y_{i+1}|$$

Fused lasso motivation

- Comparative genomic hybridization (CGH) experiment

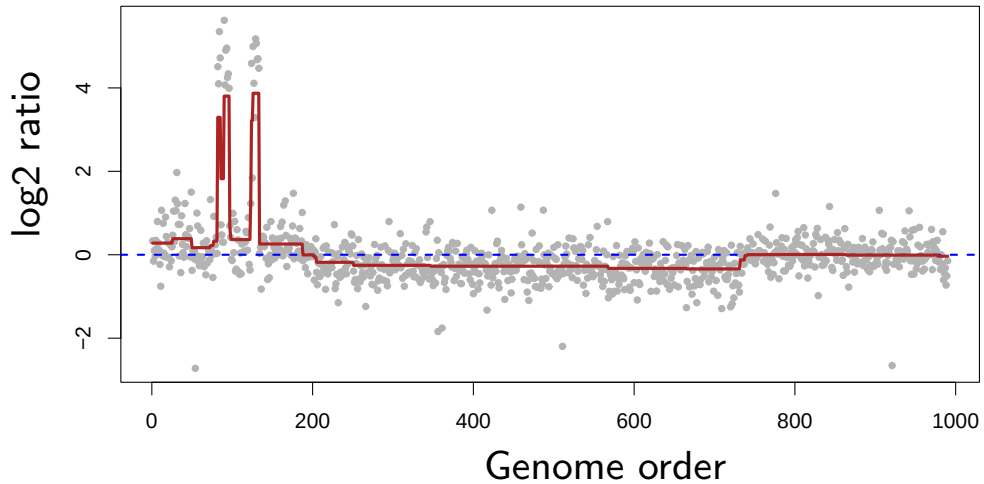


FIGURE 18.8. *Fused lasso applied to CGH data. Each point represents the copy-number of a gene in a tumor sample, relative to that of a control (on the log base-2 scale).*

- Very noisy, so that some kind of smoothing is essential.
- Biology: Typically segments of a chromosome - rather than individual genes - are replicated
- Might expect true copy numbers to be piecewise-constant.

Fused lasso signal approximator

$$\min_{\theta_1,\ldots,\theta_n} \sum_{i=1}^{n}(y_i-\theta_i)^2 + \lambda_1 \sum_{i=1}^{n}|\theta_i| + \lambda_2 \sum_{i=2}^{N}|\theta_i-\theta_{i-1}|$$

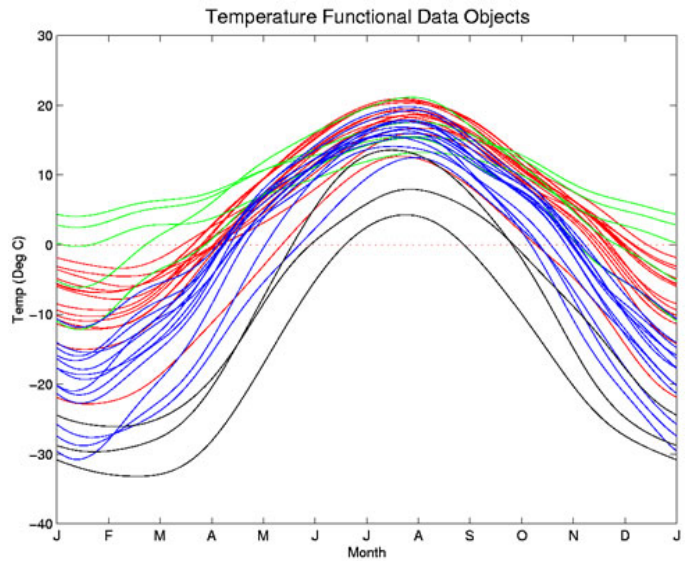
- First penalty is the familiar ℓ_1 norm
- Second penalty encourages neighboring coefficients θ_i to be similar, and hence will force some to be identical.

Generalization of fused lasso

- From linear ordering to more general neighborhoods.
  - For example, adjacent pixel in an image, or
  - nodes in a graph which are connected via an edge.
  - The former is the total variation regularization discussed earlier.
- Fused lasso for functional data. See next slide.

Functional data.

- Sometimes, the input variable is a function, $x_i(t)$.
- For example, the energy consumption or volume of incoming calls to a call center over a 24-hour period.



- Since one may take infinitely many sampling time points, each function may be viewed as an infinite-dimensional vector.
- In reality, $0 \leq t_1 \leq \cdots \leq t_p$. Define $x_{ij} = x_i(t_j)$

Fused lasso for functional data.

- We may want to predict customer satisfactory score y_i based on $x_{i1}, \ldots, x_{ip}$, using a linear regression model.
- **May expect that the coefficient function is smooth**, in the sense that $\beta(t_j)$ and $\beta(t_{j+1})$ are close to each other. For example, there is no reason to believe that a customer would care more about the waiting time at 2:01 PM than at 2:00 PM.

$$\min_{\boldsymbol{\beta}} \sum_i (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j) + \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|$$

The next section would be . . . . . .

Types of settings

- Classical/traditional: fixed p, large n.
- $p \gg n$: there is no hope of any guarantee without imposing a structure to the model.
  - hard sparsity: the real β has at most k nonzeros.
  - weak sparsity: for example,

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

Type of results

- Bound on the ℓ_2 estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$
- Bound on the ℓ_2 (in-sample) prediction error $\frac{1}{n}\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\|_2^2$
- Support Recovery, or Variable-Selection Consistency

Condition on estimation error results

- Prediction loss function

$$f_n(\hat{\beta}) := \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\hat{\beta}\|_2^2$$

- In classical setting, usually we need f_n to be **strongly convex**:

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') \geq \nabla f(\boldsymbol{\theta})^T(\boldsymbol{\theta} - \boldsymbol{\theta}') + \frac{\gamma}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2$$

- If $f(\boldsymbol{\theta})$ is twice continuously differentiable, then equivalent to: minimum eigenvalue of the Hessian matrix $\geq \gamma$ everywhere.
- Hessian matrix in this case is $\mathbf{X}^T\mathbf{X}/n$.
- However, for large p, this condition is usually impossible.

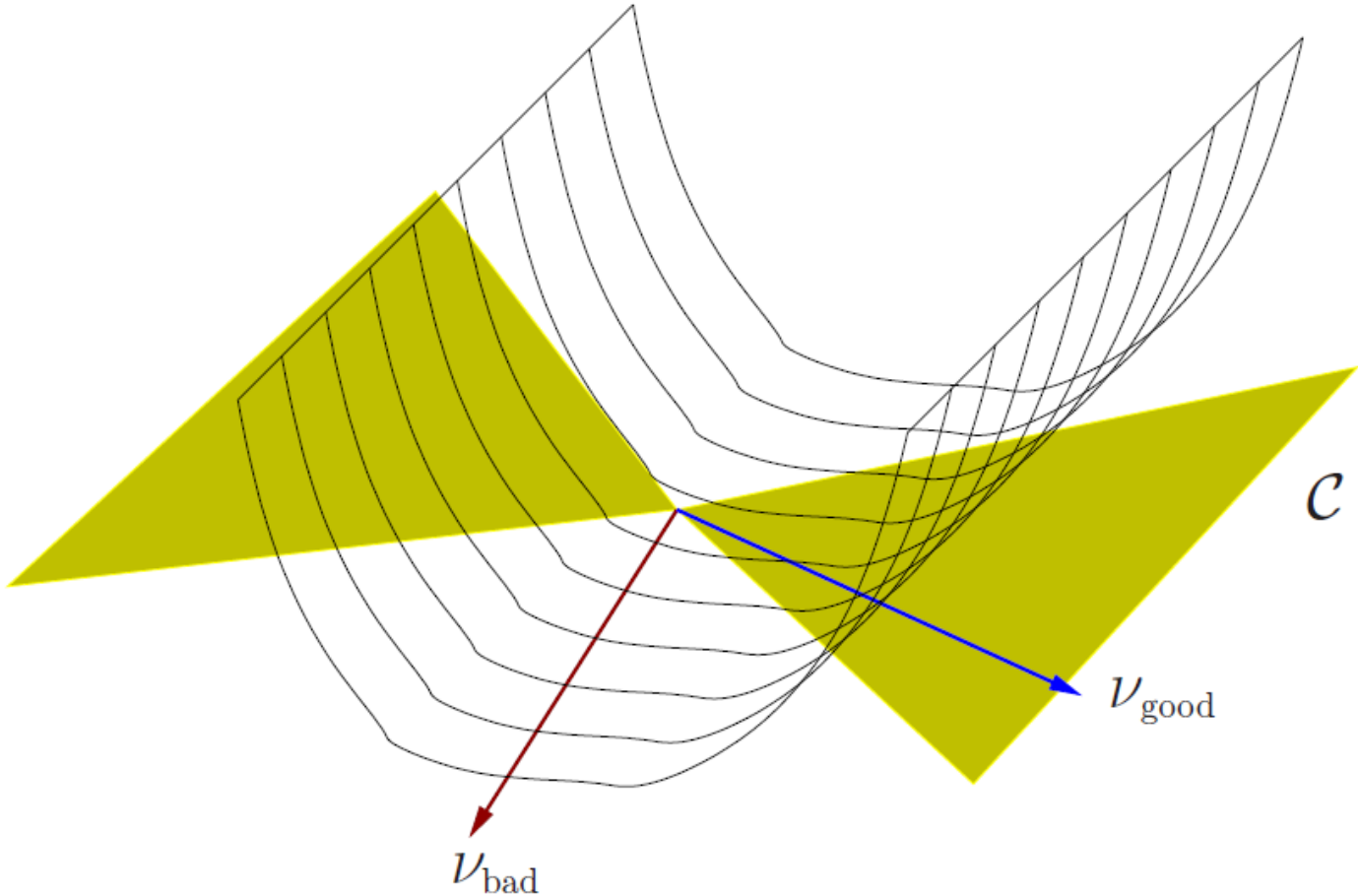


Figure 11.3 *A convex loss function in high-dimensional settings (with $p \gg N$) cannot be strongly convex; rather, it will be curved in some directions but flat in others. As shown in Lemma 11.1, the lasso error $\widehat{\nu} = \widehat{\beta} - \beta^*$ must lie in a restricted subset $\mathcal{C}$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

Restricted Eigenvalues (RE) condition

- We actually only need f_n to be strictly convex for a region of possible $\hat{\beta}$ that is worth looking at.
- **Lower bounding the restricted eigenvalues**

$$\frac{\frac{1}{n}\nu^T \mathbf{X}^T \mathbf{X}\nu}{\|\nu\|_2^2} \geq \gamma \text{ for all nonzero } \nu \in \mathcal{C}$$

where

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

$\mathcal{A}$ is the true support of β and
ν_A is the subvector of ν indexed by set A.

Bounds on Lasso ℓ_2 error

- True support has size k
- Suppose $\mathbf{X}$ satisfies the γ-RE condition over $\mathcal{C}(\mathcal{A}; 3)$
- **constrained lasso**: if constraint $t = \|\beta\|_1$, then we have

$$\|\hat{\beta} - \beta\|_2 \leq \frac{4}{\gamma}\sqrt{\frac{k}{n}}\left\|\frac{\mathbf{X}^T\varepsilon}{\sqrt{n}}\right\|_\infty$$

- **regularized lasso**: if regularization parameter $\lambda \geq 2\frac{\|\mathbf{X}^T\varepsilon\|_\infty}{n} > 0$, then

$$\|\hat{\beta} - \beta\|_2 \leq \frac{3}{\gamma}\sqrt{\frac{k}{n}}\sqrt{n}\lambda$$

Bounds on Prediction Error

Consider regularized lasso only, with parameter $\lambda = c\sigma\sqrt{\frac{\log p}{n}}$.

- $\|\beta\|_1 \leq R$, then we have

$$\|\mathbf{X}(\hat{\beta} - \beta)\|_2/n \leq c_1 \sigma R \sqrt{\frac{\log p}{n}}$$

- if β is supported on a set $\mathcal{A}$, and $\mathbf{X}$ satisfies the γ-RE condition over $\mathcal{C}(\mathcal{A}; 3)$, then

$$\|\mathbf{X}(\hat{\beta} - \beta)\|_2/n \leq c_2 \frac{\sigma^2}{\gamma} \frac{|\mathcal{A}| \log p}{n}$$

- The second rate is faster, but it requires much stronger assumptions.

When can Lasso exactly recover the support

- Small $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$ does not mean that the supports of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ are close.
- Hence, need new assumption.
- Key assumption: mutual incoherence or irrepresentability: $\exists \gamma > 0$ such that

$$\max_{j \in \mathcal{A}^c} \|(\mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}})^{-1} \mathbf{X}_{\mathcal{A}}^T \boldsymbol{x}_j\|_1 \leq 1 - \gamma$$

- Essentially: important variables should not explain the trash variables too well.
- A relaxation of orthogonality in high dimensional space.

Variable selection consistency of Lasso

- Additional assumptions:
  - $\max_j \|\boldsymbol{x}_j\|_2/\sqrt{n} \leq K$
  - $\lambda_{min}(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}/n) > C$

Theorem

Under the above three assumptions, consider regularized lasso with $\lambda \geq \frac{8K\sigma}{\gamma}\sqrt{\frac{\log p}{n}}$. *Then with probability* $1 - c_1 e^{-c_2 n\lambda^2}$,

1. *The lasso solution* $\hat{\boldsymbol{\beta}}$ *is unique.*
2. *The unique solution does not include false variable.*
3. $\|\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}\|_\infty \leq \lambda[\frac{4\sigma}{\sqrt{C}} + \|(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}/n)^{-1}\|_\infty]$
4. *The lasso solution includes all true variables with* $|\beta_j| > \lambda[\frac{4\sigma}{\sqrt{C}} + \|(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}/n)^{-1}\|_\infty]$

Comparison of Penalized Regression Methods

Strength and Weakness?

- Lasso
- Ridge regression
- ℓ_q $(0 < q < 1)$
- Elastic net
- Adaptive lasso
- SCAD or MC+
- Fused lasso
- Group lasso