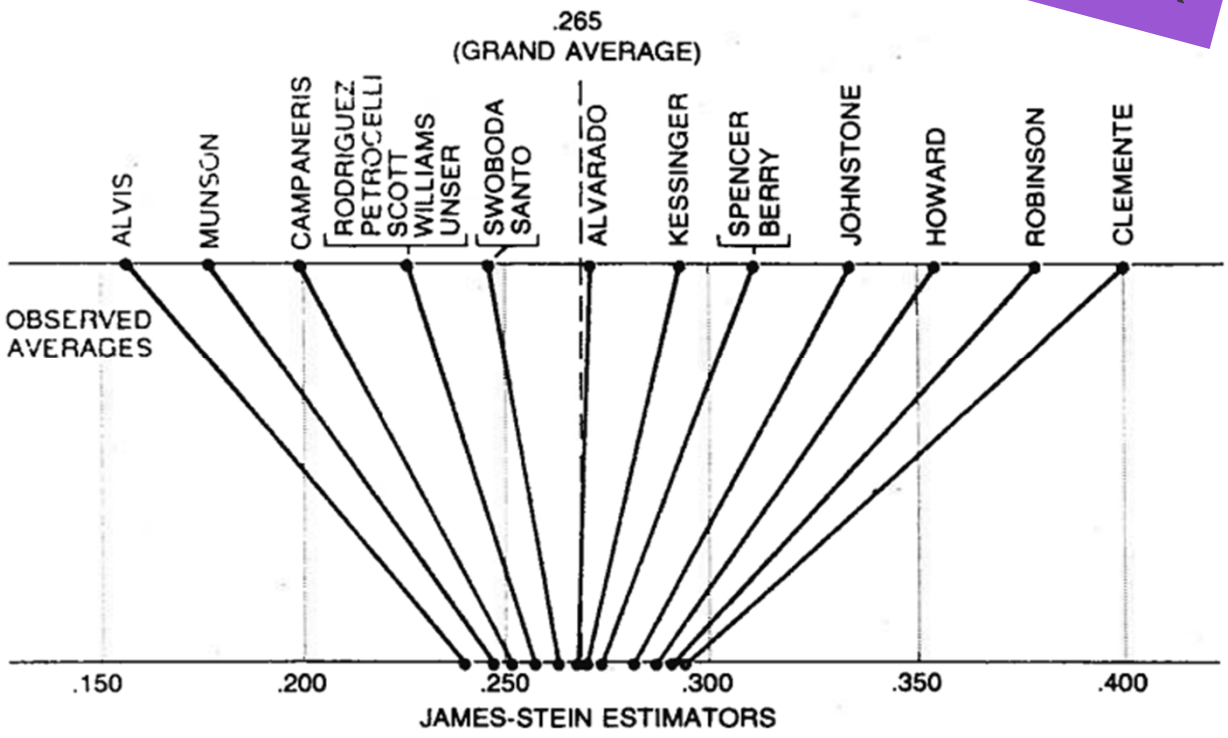
Questions

Stein Paradox

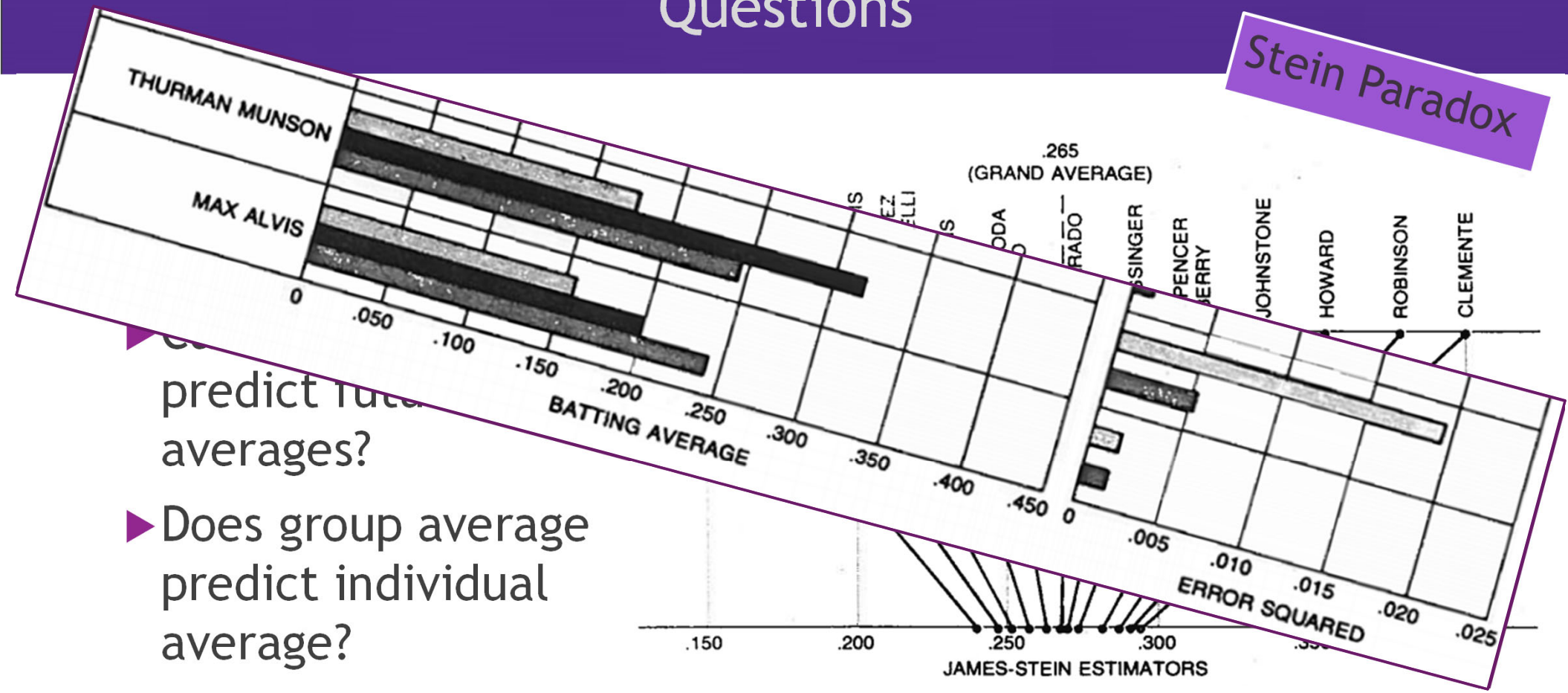
- Questions on Piazza?
- Question for You!
  - Can past averages predict future averages?
  - Does group average predict individual average?



Questions

Stein Paradox

- predict future averages?
- Does group average predict individual average?



NYU Center for Data Science

DS-GA 1003
Machine Learning

Week 4: Lecture 4
Fitting Models – Lasso Regression and Feature Selection

How can sparsity of the weights be helpful? Can regularization be used to determine features?

DS-GA 1003 Machine Learning

Week 4: Lecture 4

Fitting Models – Lasso Regression and Feature Selection

Adapted from Rosenberg, Boyd, Rangan, Singh

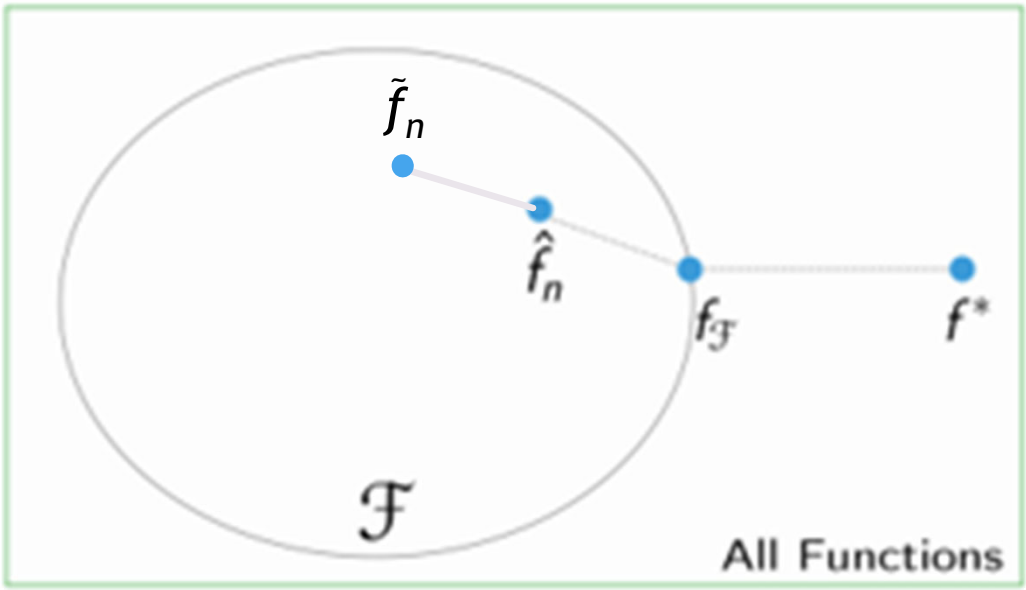
Announcements

- Please check Week 4 agenda on NYU Classes
  - Homework
  - Recordings
  - Tutoring Session
  - Office Hours
- Remember to post to Piazza

Check Calendar linked to NYU Classes for important dates

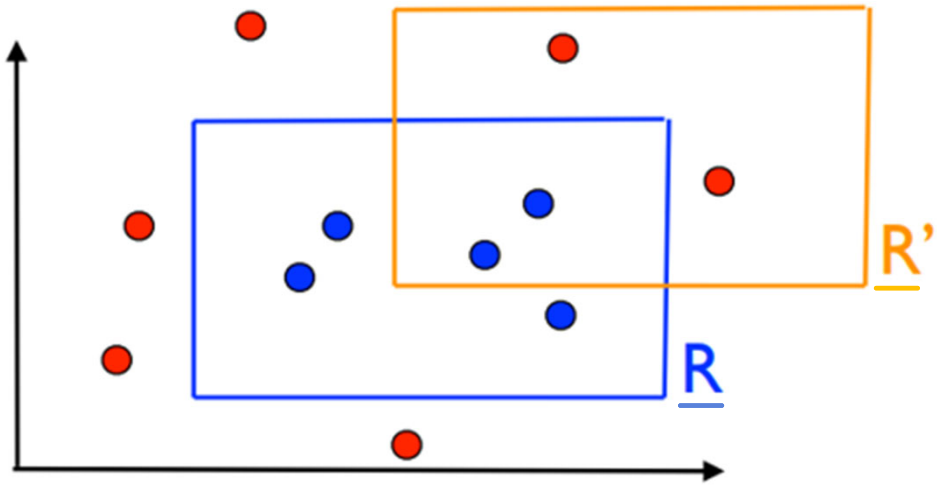
Review

- Consider
  - Approximation Error
  - Estimation Error
  - Optimization Error
- For each try to determine whether it's
  - Random or Not Random
  - Positive or Negative
  - Increases or Decreases with more data
  - Increases or Decreases with more Parameters
  - Can we ever compute it?



Review

- Problem: Suppose you want to predict ripeness of fruit. Does color and firmness impact ripeness?
- Input Space:
  - Fruit
- Features:
  - Encoding of color
  - Encoding of firmness
- Labels:
  - +1 for ripe
  - -1 for not ripe/spoiled



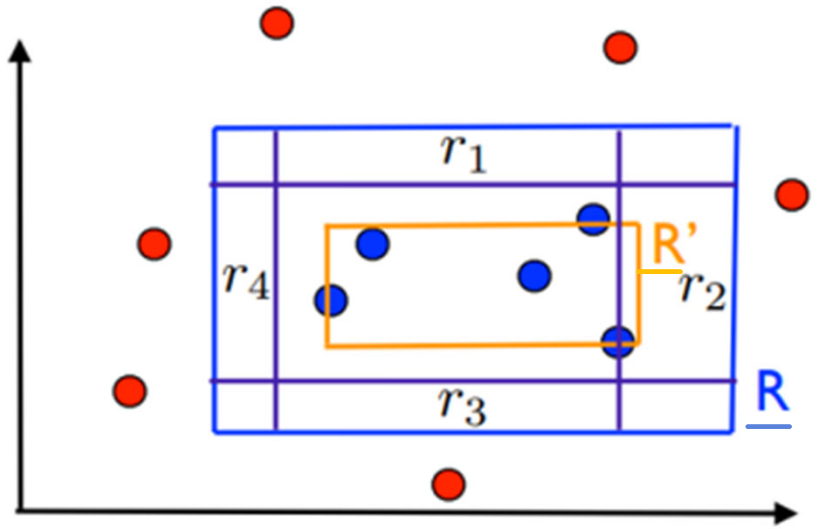
- Action Space:
  - Eat
  - Wait/Compost
- Loss Function:
  - +1 incorrect
  - 0 correct

Review

What is the probability that $R(R') > \epsilon$?

- Fix $0 < \delta < 1$
- For $m \geq \frac{4}{\epsilon} \log \frac{4}{\delta}$

we probably (with respect to δ) have approximately (with respect to ε) no statistical risk



$$\Pr[R(R') > \epsilon] \leq \Pr[\cup_{i=1}^{4} \{R' \text{ misses } r_i\}]$$
$$\leq \sum_{i=1}^{4} \Pr[\{R' \text{ misses } r_i\}]$$
$$\leq 4(1 - \tfrac{\epsilon}{4})^m \leq 4e^{-\frac{m\epsilon}{4}}$$

Agenda

- Ridge Regression
  - Solving for minimizer of square loss
- Lasso Regression
  - Connection to feature selection
  - How to fit parameters to data?
- Elastic Net
  - Behavior of Ridge and Lasso with related features?

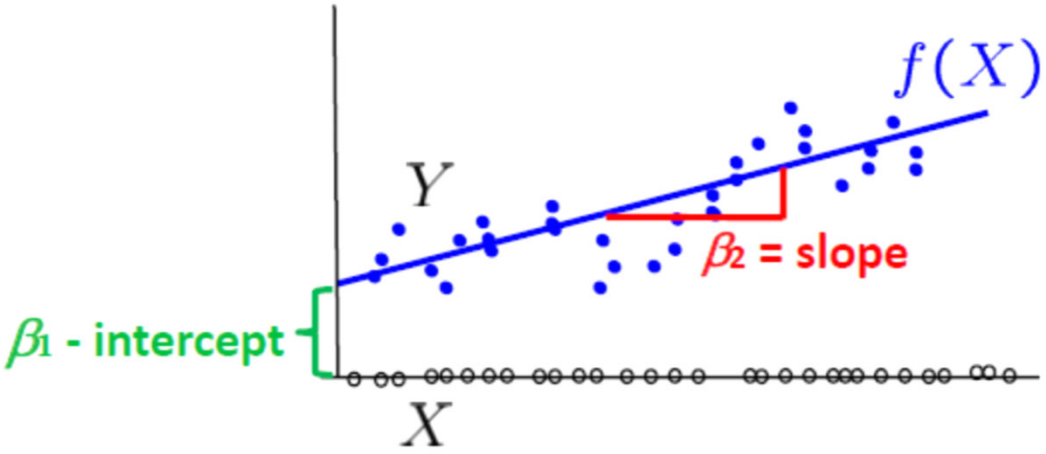
- References
  - T. Hastie, *Elements of Statistical Learning*
    - Chapter 3.2, 3.3, 3.4.2
  - L. Miolane, *DS-GA 1014 Lecture Notes*
    - Lecture 9

Square Loss

- Recall that the objective function for linear regression involves the square loss

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

- Here the collection of hypotheses is *F*



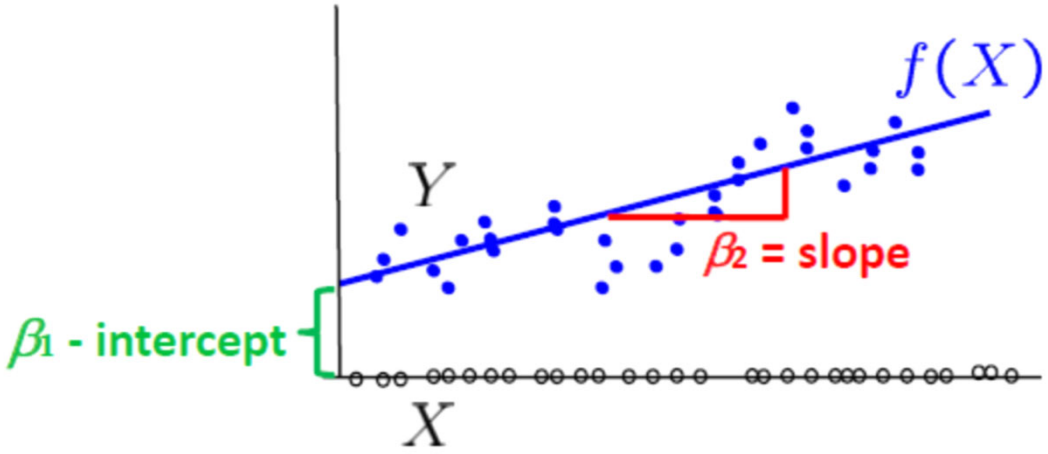
$$f(X) = \beta_1 + \beta_2 X$$

Square Loss

- Recall that the objective function for linear regression involves the square loss

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

- Here the collection of hypotheses is *F*



$$f(X) = X\beta$$

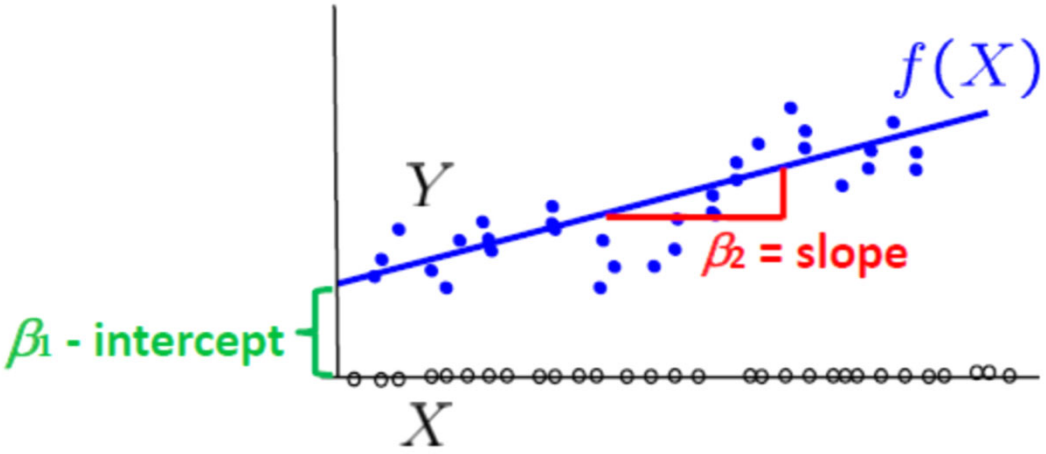
where $X = [X^{(1)} \ldots X^{(p)}]$, $\beta = [\beta_1 \ldots \beta_p]^T$

Square Loss

- Recall that the objective function for linear regression involves the square loss

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

- We must find weights that satisfy…

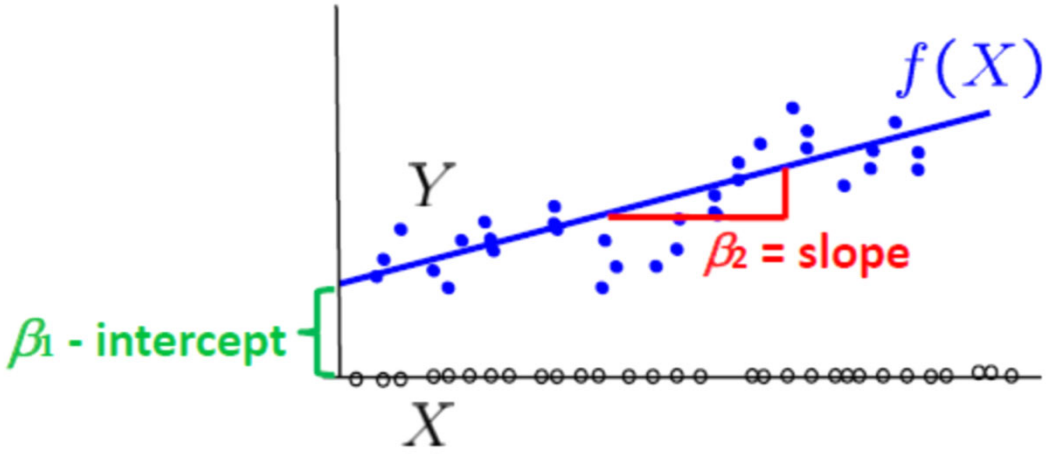


$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (X_i\beta - Y_i)^2$$

Square Loss

- Recall that the objective function for linear regression involves the square loss

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$



- We must find weights that satisfy…or equivalently

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} (\mathbf{A}\beta - \mathbf{Y})^T (\mathbf{A}\beta - \mathbf{Y}) = \arg\min_{\beta} J(\beta)$$

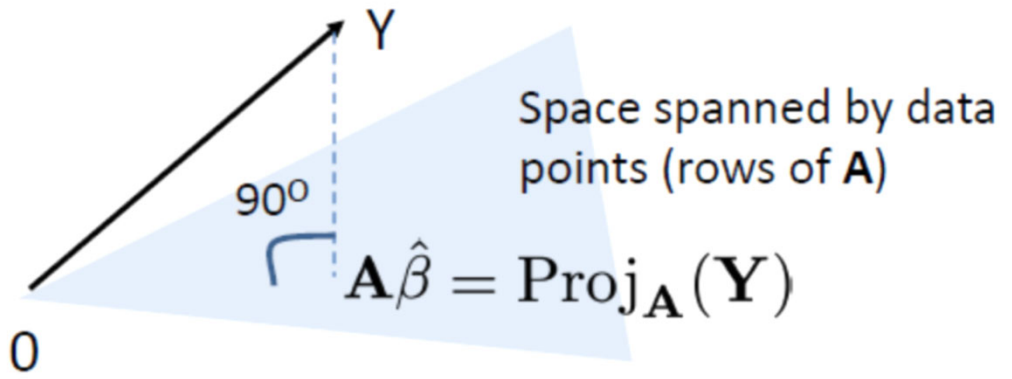
Square Loss

- If A^TA is invertible then we can solve normal equations. We obtain

$$\widehat{\beta} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}$$

- If p is large, then inversion takes O(p) operations. Instead we can use gradient descent

$$\begin{aligned}\beta^{t+1} &= \beta^t - \frac{\alpha}{2}\frac{\partial J(\beta)}{\partial \beta}\Big|_t \\ &= \beta^t - \alpha\ \mathbf{A}^T(\mathbf{A}\beta^t - Y)\end{aligned}$$



$$\frac{\partial J(\beta)}{\partial \beta}\Big|_{\widehat{\beta}} = 0 \qquad \text{gives} \qquad (\mathbf{A}^T\mathbf{A})\widehat{\beta} = \mathbf{A}^T\mathbf{Y}$$

p x p p x1 p x1

Square Loss

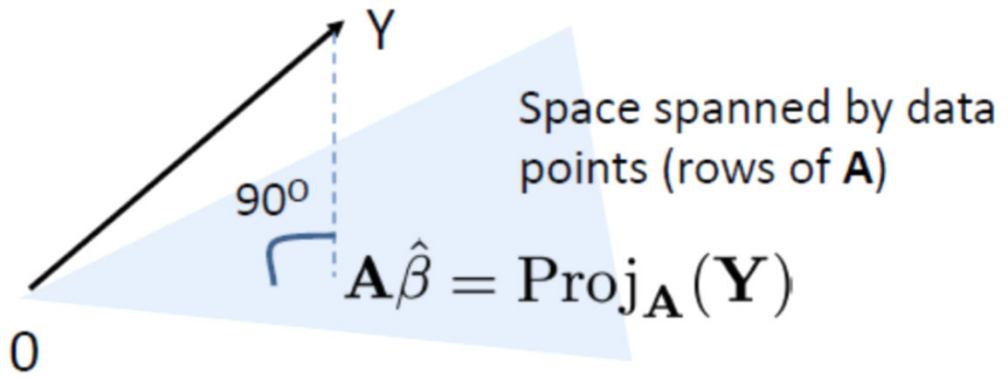
- If A^TA is not invertible then we need another approach.
- We can decompose

$$\underset{\text{S - r x r}}{\mathbf{A}} = \mathbf{U}\mathbf{S}\mathbf{V}^\top,$$

where U, V are orthogonal and S is diagonal. Here r is the rank

- The equations become

$$\underset{\text{r x p}}{(\mathbf{S}\mathbf{V}^\top)}\underset{\text{p x 1}}{\hat{\beta}} = \underset{\text{r x 1}}{(\mathbf{U}^\top\mathbf{Y})}$$



For p > r we have more variables than equations meaning the solutions are not unique

Ridge Regression

- If A^TA is not invertible then we need another approach.
- We can decompose

$$\mathbf{A} = \underset{\text{S - r x r}}{\mathbf{U}\mathbf{S}\mathbf{V}^\top},$$

where U, V are orthogonal and S is diagonal. Here r is the rank

- The equations become

$$\underset{\text{r x p}}{(\mathbf{S}\mathbf{V}^\top)}\underset{\text{p x 1}}{\hat{\beta}} = \underset{\text{r x 1}}{(\mathbf{U}^\top\mathbf{Y})}$$

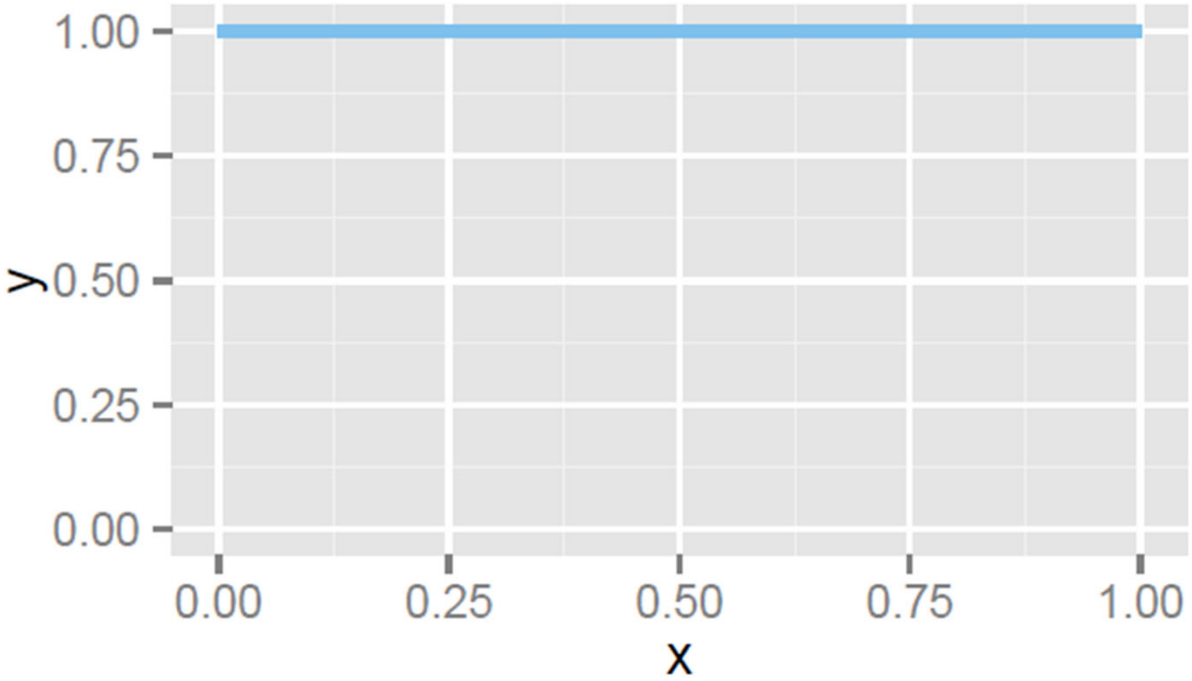
$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_2^2$$

$$= \arg\min_{\beta} \; (\mathbf{A}\beta - \mathbf{Y})^T(\mathbf{A}\beta - \mathbf{Y}) + \lambda\|\beta\|_2^2$$

Here λ is a positive number

Regularization

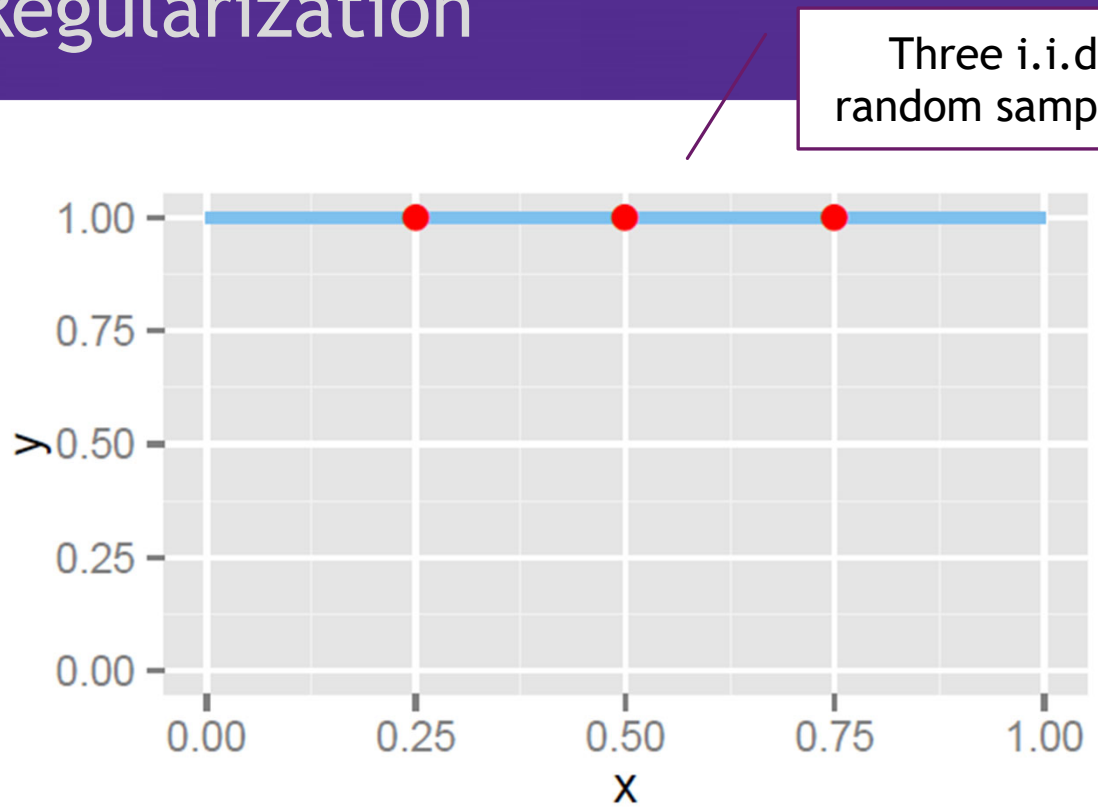
- Problem: How can we prevent against unstable hypotheses?
- Input Space:
  - [0,1] with uniform probability
- Features:
  - Identity transformation
- Output Space:
  - [0,1] with probability concentrated at value 1



- Action Space:
  - [0,1]
- Loss Function:
  - +1 incorrect
  - 0 correct

Regularization

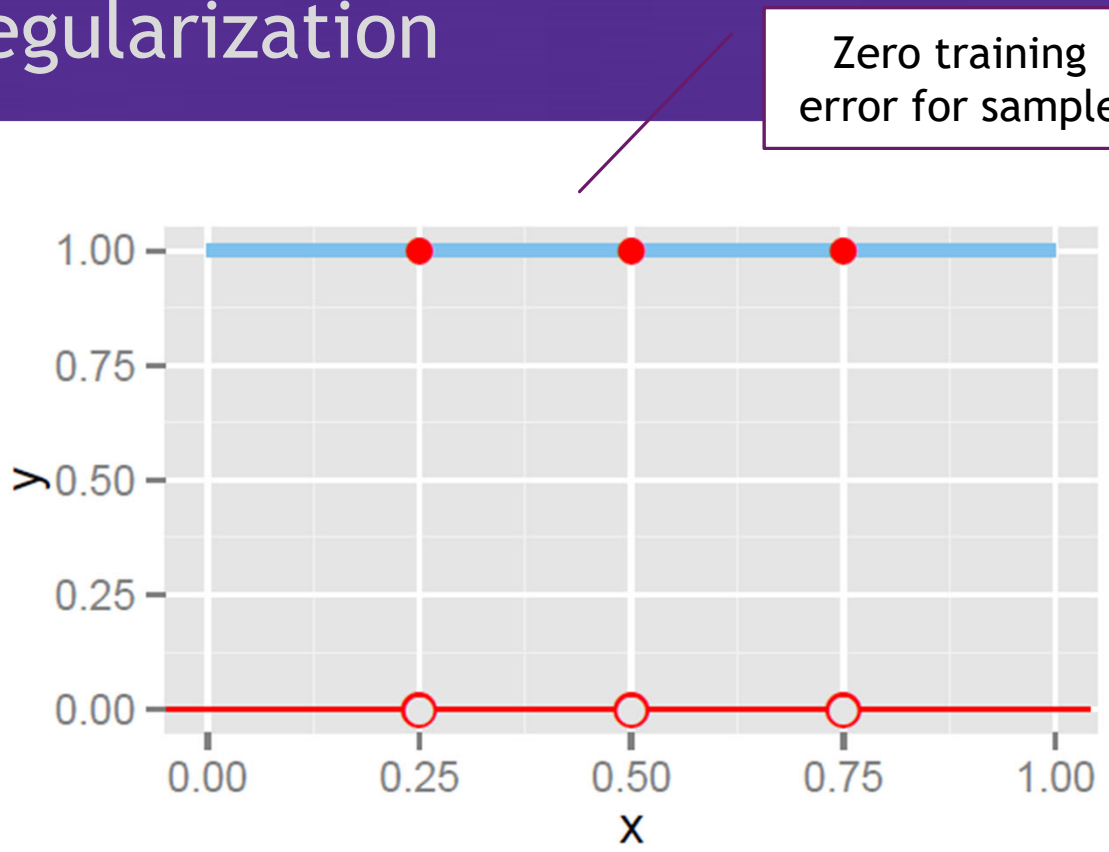
- Problem: How can we prevent against unstable hypotheses?
- Input Space:
  - [0,1] with uniform probability
- Features:
  - Identity transformation
- Output Space:
  - [0,1] with probability concentrated at value 1



- Action Space:
  - [0,1]
- Loss Function:
  - +1 incorrect
  - 0 correct

Regularization

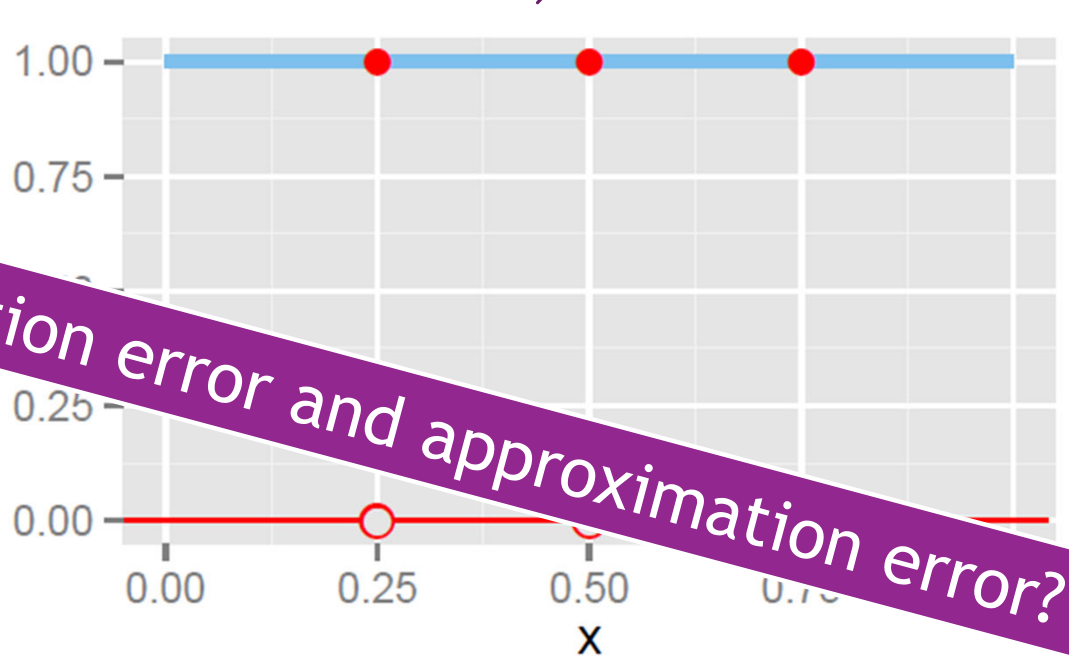
- Problem: How can we prevent against unstable hypotheses?
- Input Space:
  - [0,1] with uniform probability
- Features:
  - Identity transformation
- Output Space:
  - [0,1] with probability concentrated at value 1



- Action Space:
  - [0,1]
- Loss Function:
  - +1 incorrect
  - 0 correct

Regularization

- Problem: How can we [illegible] against unstable [illegible]
- Input Space:
  - [0,1] with uniform probability
- Features:
  - Identity transformation
- Output Space:
  - [0,1] with probability concentrated at value 1



How to balance estimation error and approximation error?

- Action Space:
  - [0,1]
- Loss Function:
  - +1 incorrect
  - 0 correct

Constraint

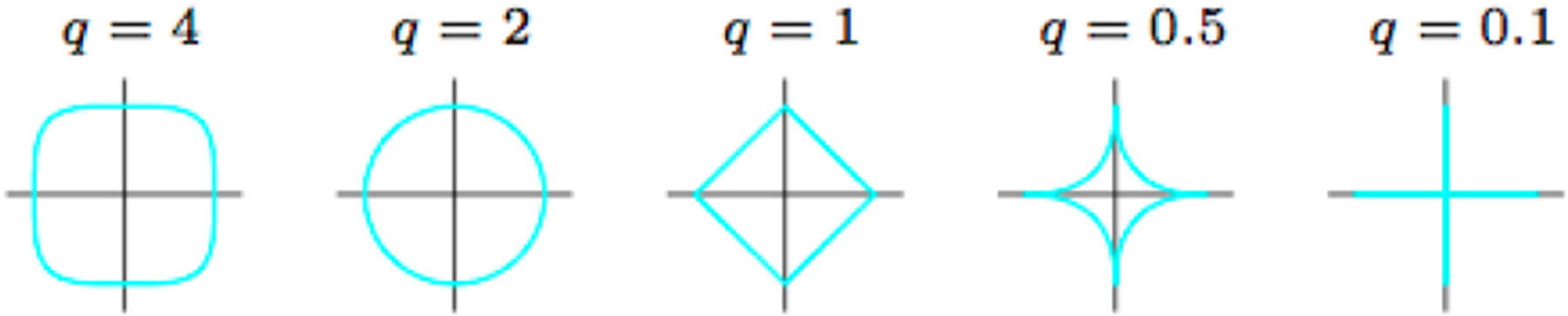
- Define the length function

$$\ell_q = \left(\sum_{i=1}^{n} |w_i|^q \right)^{\frac{1}{q}}$$

- For q = 2 we have the usual length denoted as absolute value

- For q ≥ 1 we have the properties of a norm

$$||w||_q \geq 0 \text{ with equality only for } w = 0$$
$$||c \cdot w||_q = c||w||_q \text{ for any number } c$$
$$||w + v||_q \leq ||w||_q + ||v||_q$$



Constraint

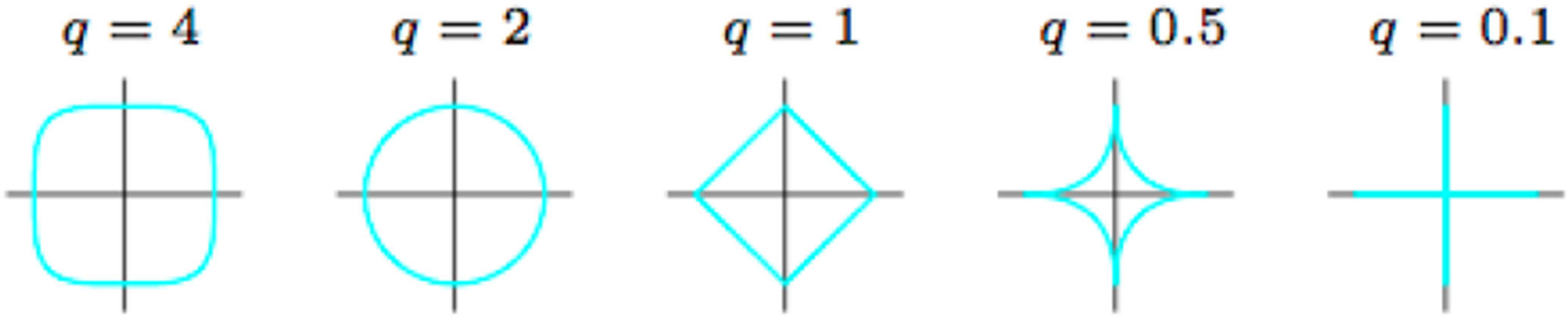
- Note that

$$||w||_2 \leq ||w||_1$$

for all vectors w. Why can we compare the two length functions?

- For q ≥ 1 we have the properties of a norm

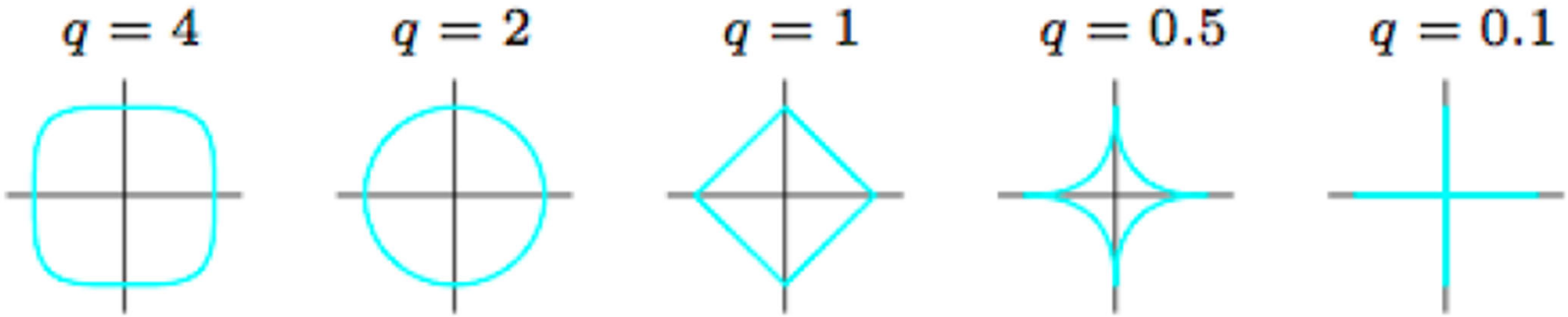
$$||w||_q \geq 0 \text{ with equality only for } w = 0$$
$$||c \cdot w||_q = c||w||_q \text{ for any number } c$$
$$||w + v||_q \leq ||w||_q + ||v||_q$$



Constraint

- For $h(x) = w \cdot x$ we can calculate the change in output caused by change in input

$$|h(x+z) - h(x)| = |w \cdot (x+z) - w \cdot x| = |w \cdot z|$$
$$\leq ||w||_2 \cdot ||z||_2 \leq ||w||_1 \cdot ||z||_2$$

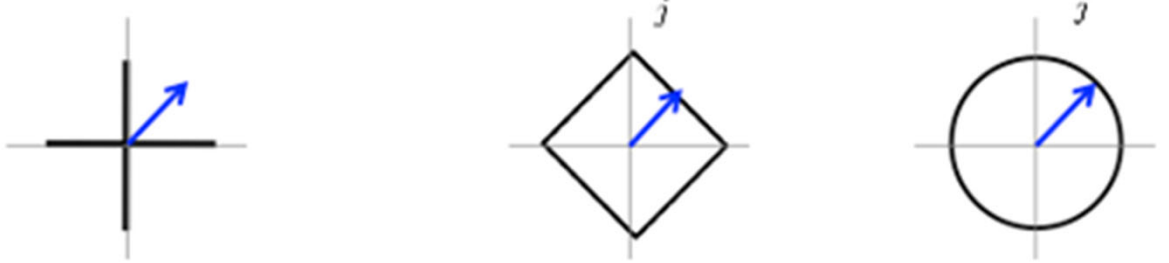


Constraint

- For $h(x) = w \cdot x$ we can calculate the change in output caused by change in input

$$|h(x+z) - h(x)| = |w \cdot (x+z) - w \cdot x| = |w \cdot z|$$
$$\leq ||w||_2 \cdot ||z||_2 \leq ||w||_1 \cdot ||z||_2$$

$\|W\|_0 = \#\{W_j > 0\}$ $\|W\|_1 = \sum_j |W_j|$ $\|W\|_2 = \sum_j W_j^2$



Ridge Regression

Ridge Regression (Tikhonov Form)

The ridge regression solution for regularization parameter $\lambda \geqslant 0$ is

$$\hat{w} = \underset{w \in \mathbf{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda \|w\|_2^2,$$

Penalization Form of Regularization

where $\|w\|_2^2 = w_1^2 + \cdots + w_d^2$ is the square of the ℓ_2-norm.

Constraint Form of Regularization

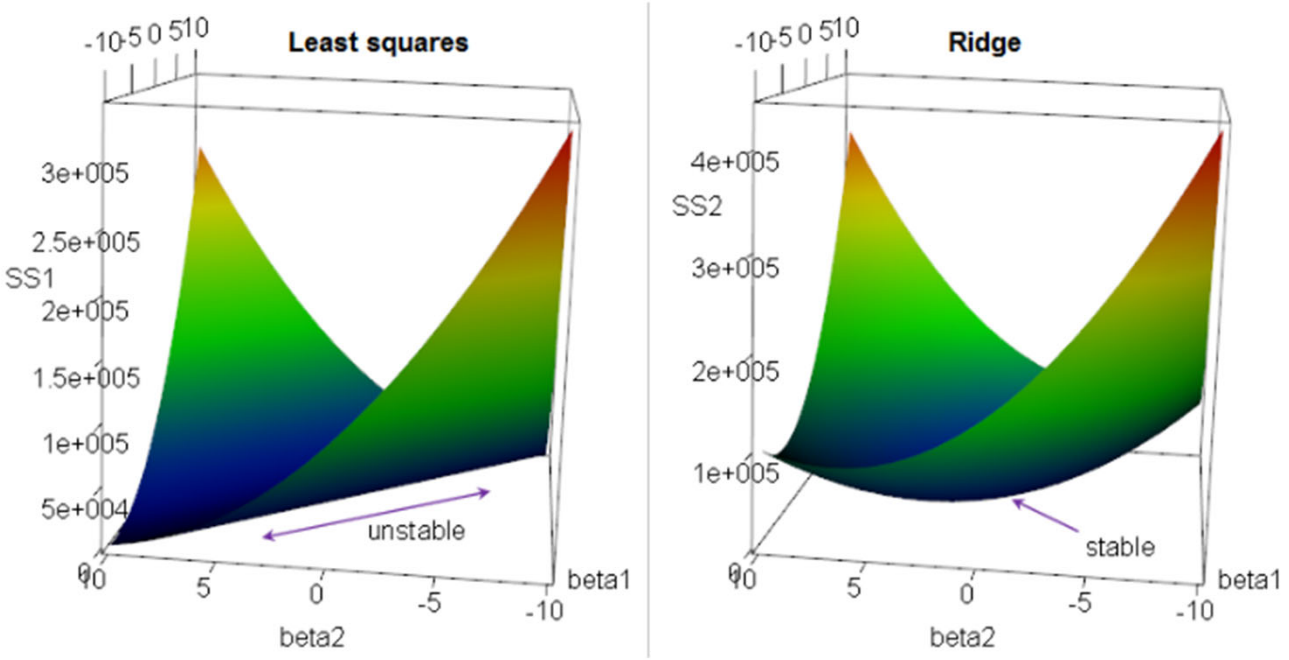
Ridge Regression (Ivanov Form)

The ridge regression solution for complexity parameter $r \geqslant 0$ is

$$\hat{w} = \underset{\|w\|_2^2 \leqslant r^2}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2.$$

Ridge Regression

- Correlated features will yield flat regions in the objective function nicknamed ridges
- Regularization causes the ridge to bend upward away from a global minimum



Penalization

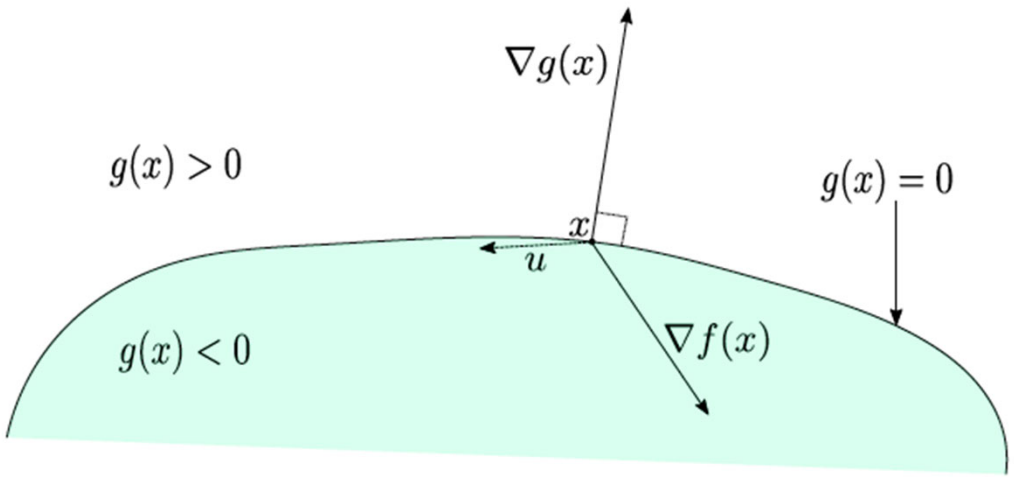
- General minimization problems with constraints take the form

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && g_i(x) \leq 0, \quad i = 1, \ldots, m \end{aligned}$$

where $x \in \mathbb{R}^n$ and

$$g_1(x) \leq 0, \ldots, g_m(x) \leq 0$$

- Suppose that the minimize x occurs at the boundary of the constraint set where g(x) = 0



Penalization

- General minimization problems with constraints take the form

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && g_i(x) \leq 0, \quad i = 1, \ldots, m \end{aligned}$$

where $x \in \mathbb{R}^n$ and

$$g_1(x) \leq 0, \ldots, g_m(x) \leq 0$$

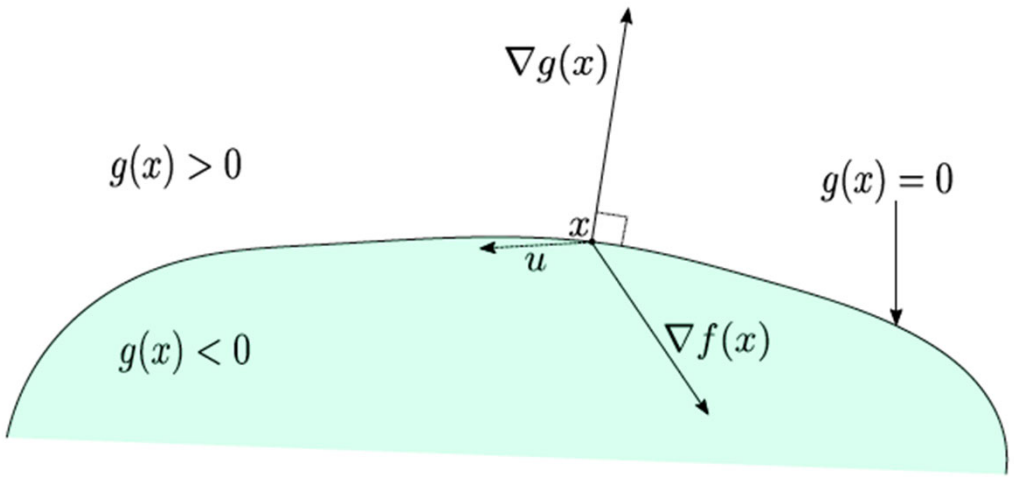
- Suppose that the minimize x occurs at the boundary of the constraint set where g(x) = 0

- If we can find a vector *u* such that

$$\langle u, \nabla g(x) \rangle < 0 \quad \text{and} \quad \langle u, \nabla f(x) \rangle < 0$$

then we can decrease the value of both *g* and *f* for some small number δ > 0

$$g(x + \delta u) \simeq g(x) + \delta \langle u, \nabla g(x) \rangle \leq 0$$
$$f(x + \delta u) \simeq f(x) + \delta \langle u, \nabla f(x) \rangle < f(x)$$



Penalization

- General minimization problems with constraints take the form

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && g_i(x) \le 0, \quad i = 1, \ldots, m \end{aligned}$$

where $x \in \mathbb{R}^n$ and

$$g_1(x) \le 0, \ldots, g_m(x) \le 0$$

- Suppose that the minimize x occurs at the boundary of the constraint set where g(x) = 0

- If we can find a vector *u* such that

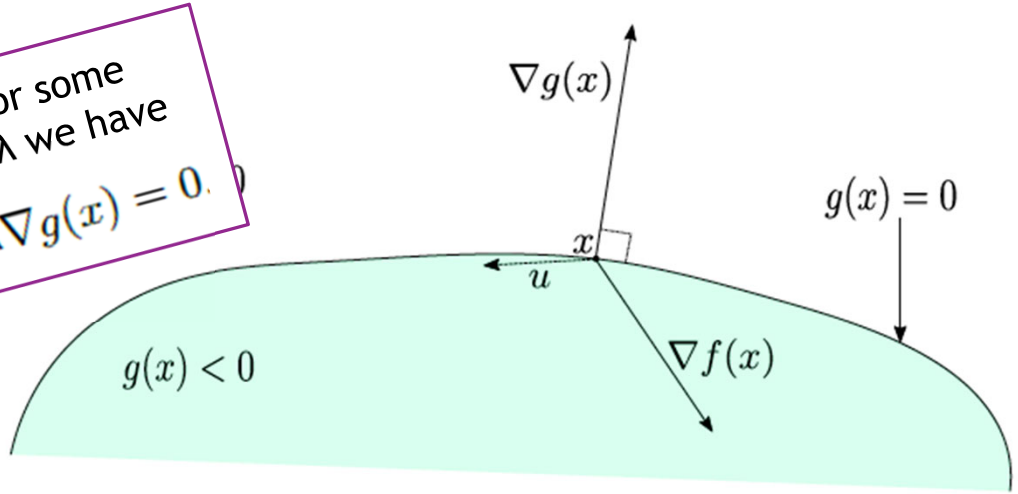
$$\langle u, \nabla g(x) \rangle < 0 \quad \text{and} \quad \langle u, \nabla f(x) \rangle < 0$$

then we can decrease the value of both *g* and *f* for some small number δ > 0

$$g(x + \delta u) \simeq g(x) + \delta \langle u, \nabla g(x) \rangle \le 0$$
$$f(x + \delta u) \simeq f(x) + \delta \langle u, \nabla f(x) \rangle < f(x)$$

Therefore for some nonnegative λ we have

$$\nabla f(x) + \lambda \nabla g(x) = 0$$



Lasso Regression

Lasso Regression (Tikhonov Form)

The lasso regression solution for regularization parameter $\lambda \geqslant 0$ is

$$\hat{w} = \underset{w \in \mathbf{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda \|w\|_1,$$

where $\|w\|_1 = |w_1| + \cdots + |w_d|$ is the ℓ_1-norm.

Penalization Form of Regularization

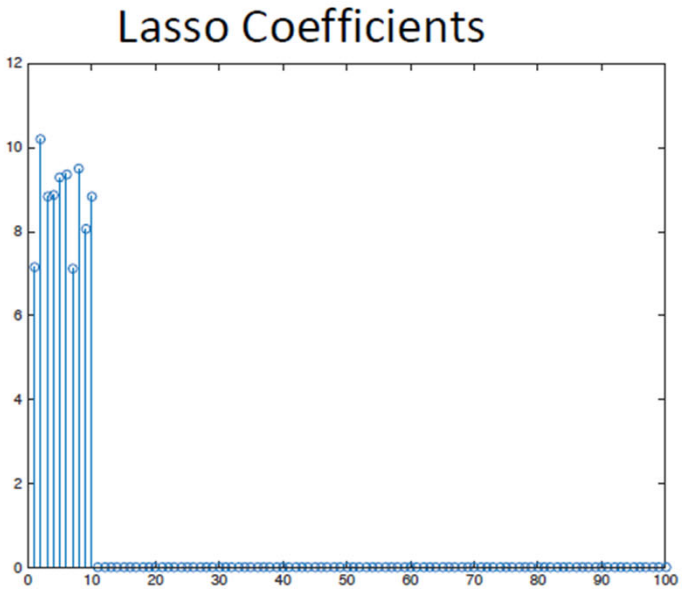
Constraint Form of Regularization

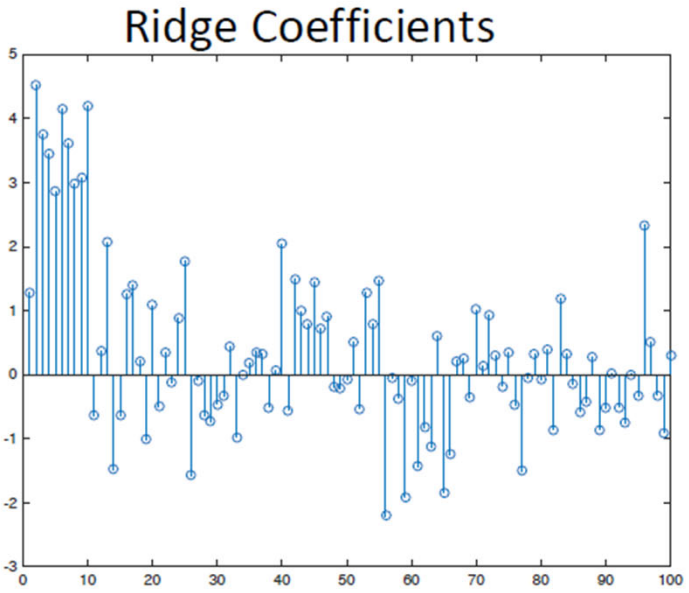
Lasso Regression (Ivanov Form)

The lasso regression solution for complexity parameter $r \geqslant 0$ is

$$\hat{w} = \underset{\|w\|_1 \leqslant r}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2.$$

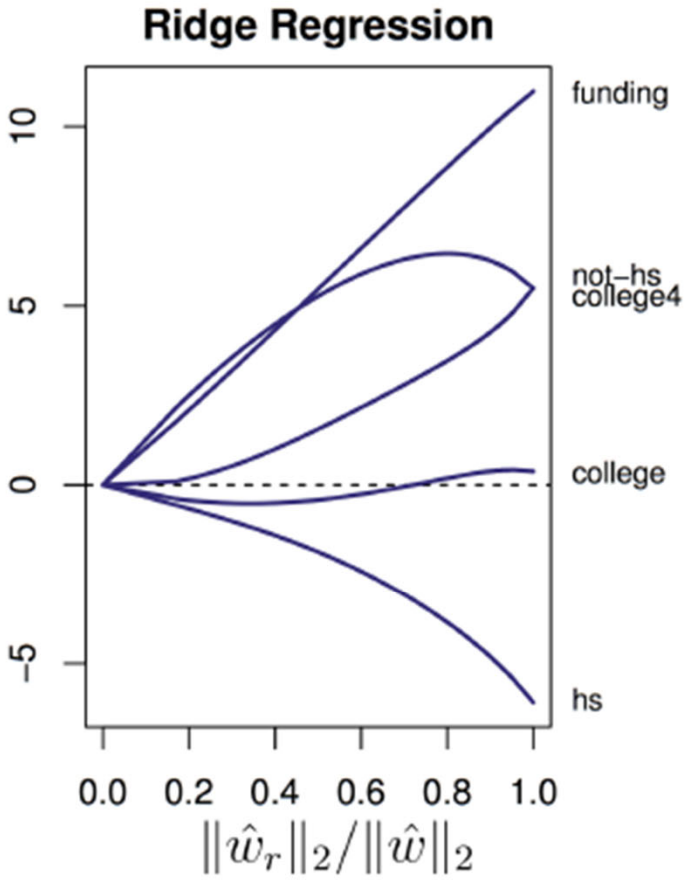
Lasso Regression





- Ridge Regression tends to shrink features
- Lasso Regression tends to reduce features to zero
- Having many features with value zero is called sparsity or parsimony

Regularization Path

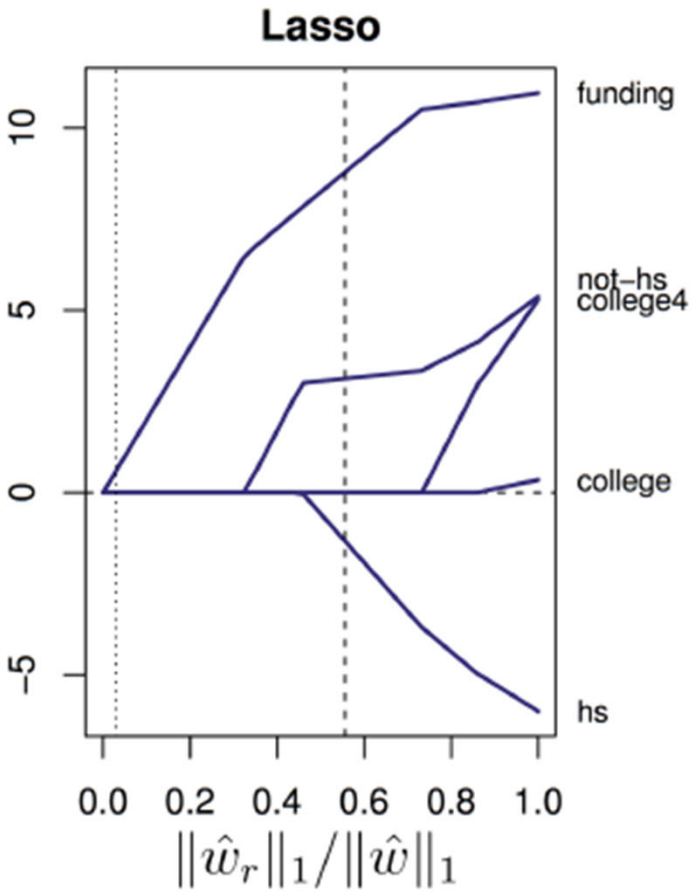


$$\hat{w}_r = \arg\min_{\|w\|_2^2 \le r^2} \frac{1}{n} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2$$

$$\hat{w} = \hat{w}_\infty = \text{Unconstrained ERM}$$

- For $r = 0$, $\|\hat{w}_r\|_2/\|\hat{w}\|_2 = 0$.
- For $r = \infty$, $\|\hat{w}_r\|_2/\|\hat{w}\|_2 = 1$

Regularization Path



$$\hat{w}_r = \arg\min_{\|w\|_1 \le r} \frac{1}{n} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2$$

$$\hat{w} = \hat{w}_\infty = \text{Unconstrained ERM}$$

- For $r = 0$, $\|\hat{w}_r\|_1/\|\hat{w}\|_1 = 0$.
- For $r = \infty$, $\|\hat{w}_r\|_1/\|\hat{w}\|_1 = 1$

Exercise

- Question: Suppose we have the objective from ridge regression

$$\operatorname*{argmin}_{w\in\mathbb{R}^m}\left(\frac{1}{2m}\sum_{i=1}^{m}(x_i w-y_i)^2+\lambda w^2\right).$$

Here we take dimension equal to one.

- Show that the value of w is

$$w=\frac{\langle \mathbf{x},\mathbf{y}\rangle/m}{\|\mathbf{x}\|^2/m+2\lambda}.$$

- Question: Suppose we have the objective function

$$\min_{w\in\mathbb{R}}\left(\tfrac{1}{2}w^2-xw+\lambda|w|\right)$$

- Show that the value of w is

$$w=\operatorname{sign}(x)\,[|x|-\lambda]_+$$

where

$$[a]_+ \stackrel{\text{def}}{=} \max\{a,0\}$$

Exercise

Threshold function

- Suppose we have the objective from lasso regression

$$\underset{w\in\mathbb{R}^m}{\operatorname{argmin}} \left(\frac{1}{2m} \sum_{i=1}^{m} (x_i w - y_i)^2 + \lambda |w| \right)$$

We can rewrite as

$$\underset{w\in\mathbb{R}^m}{\operatorname{argmin}} \left(\frac{1}{2} \left(\frac{1}{m} \sum_i x_i^2 \right) w^2 - \left(\frac{1}{m} \sum_{i=1}^{m} x_i y_i \right) w + \lambda |w| \right)$$

- Taking we $\frac{1}{m}\sum_i x_i^2 = 1$ can apply the Question to obtain

$$w = \operatorname{sign}(\langle \mathbf{x}, \mathbf{y} \rangle) [|\langle \mathbf{x}, \mathbf{y} \rangle|/m - \lambda]_+$$

- Question: Suppose we have the objective function

$$\min_{w\in\mathbb{R}} \left(\tfrac{1}{2} w^2 - x w + \lambda |w| \right)$$

- Show that the value of w is

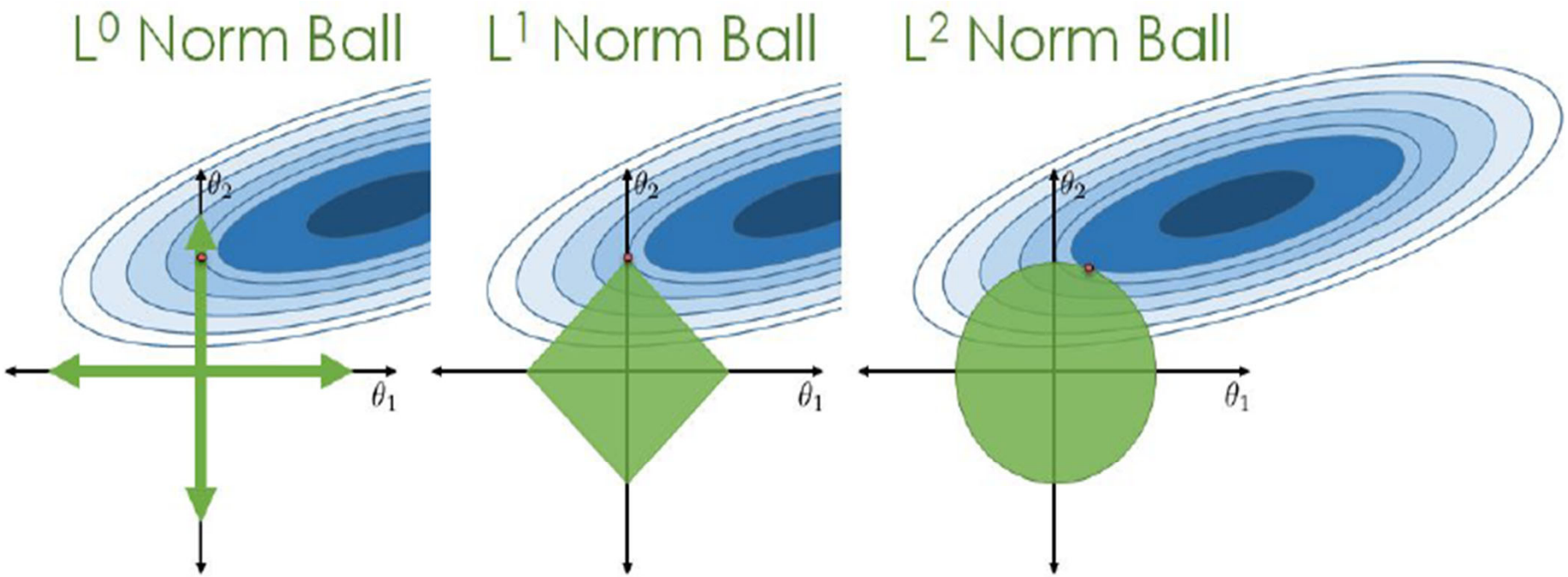
$$w = \operatorname{sign}(x) [|x| - \lambda]_+$$

where

$$[a]_+ \stackrel{\text{def}}{=} \max\{a, 0\}$$

Regularization

Notes



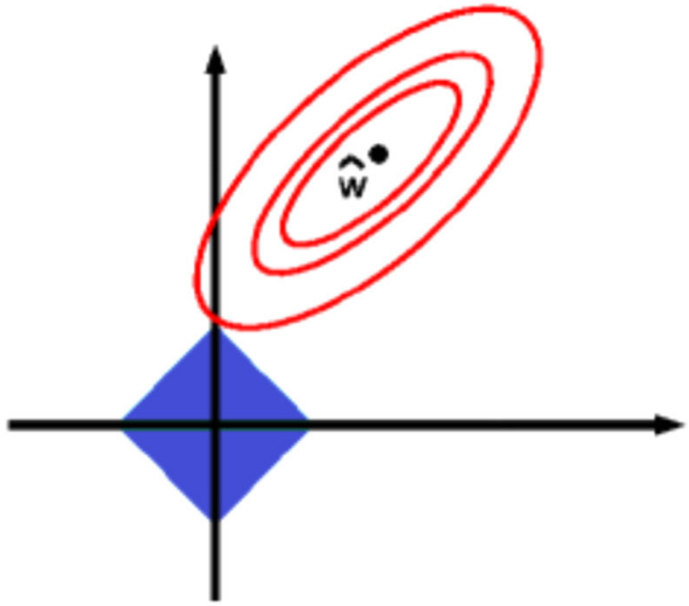
Level Sets

- Note that the level sets of the empirical risk relative to empirical risk of the weights of the target functions are

$$\left\{ w \mid (w - \hat{w})^T X^T X (w - \hat{w}) = nc \right\}$$

For constant c.

- If the matrix X^TX has rank lower than the number of features, then the level sets are lines instead of ellipses



Repeated Features

- Suppose we have dimension two. How do Ridge Regression and Lasso Regression behave with repeated features?
- Take

$$\hat{f}(x_1, x_2) = w_1 x_1 + w_2 x_2$$

where $x_1 = x_2$.

- Suppose the ERM is $4x_1$. So

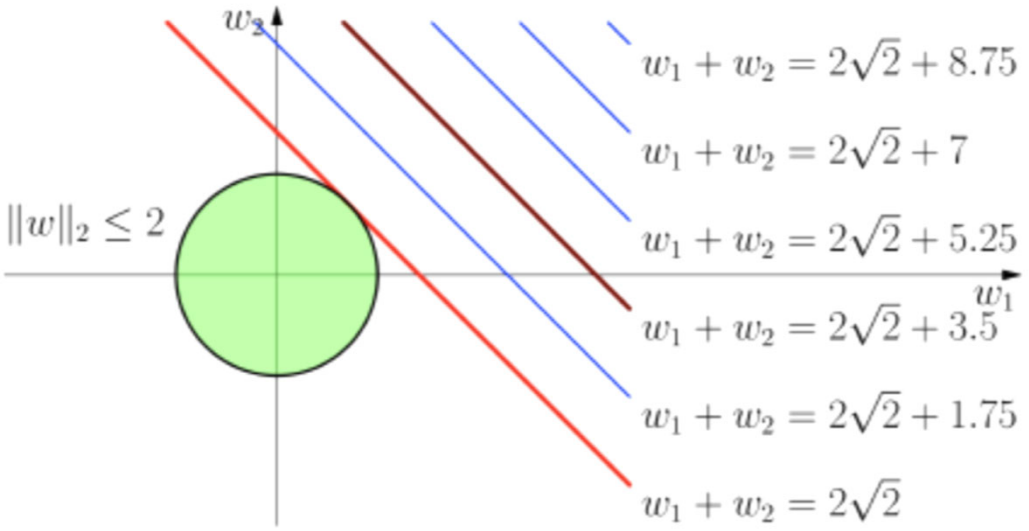
$$w_1 + w_2 = 4$$

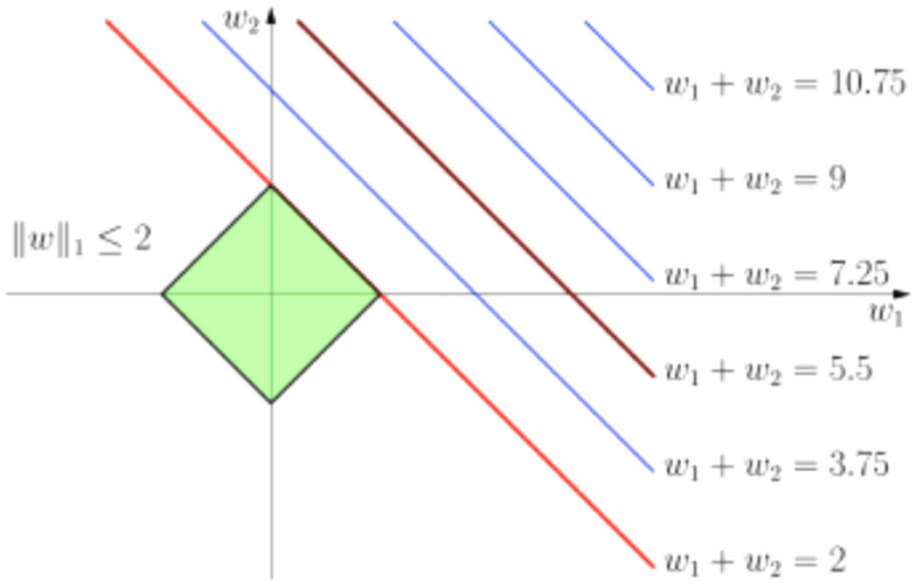
give ERM.

w_1	w_2	$\|w\|_1$	$\|w\|_2^2$
4	0	4	16
2	2	4	8
1	3	4	10
-1	5	6	26

- Lasso Regression does not distinguish granted the weights have the same sign
- Ridge Regression spreads weight evenly

Repeated Features



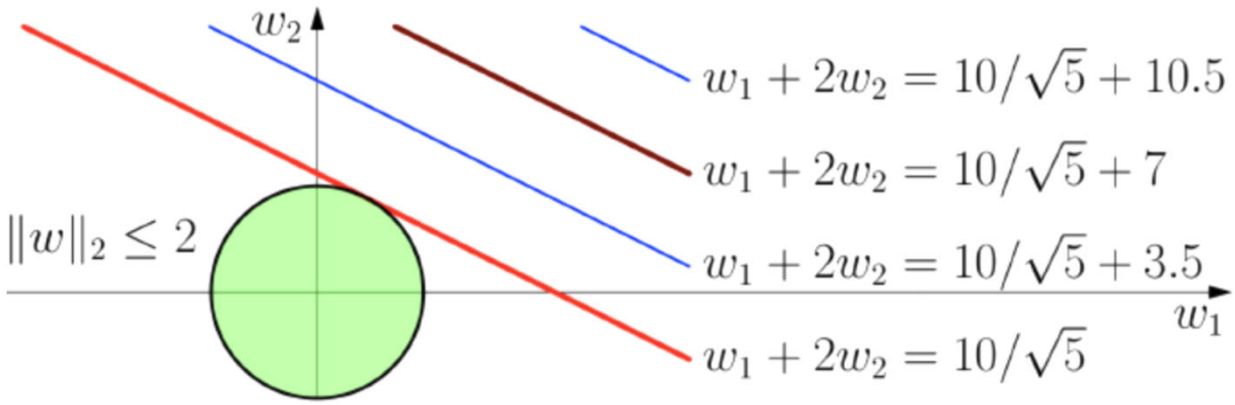


Repeated Features

- Suppose we have dimension two. How do Ridge Regression and Lasso Regression behave with repeated features of different scales?
- Take

$$\hat{f}(x_1, x_2) = w_1 x_1 + w_2 x_2$$

where $2x_1 = x_2$.

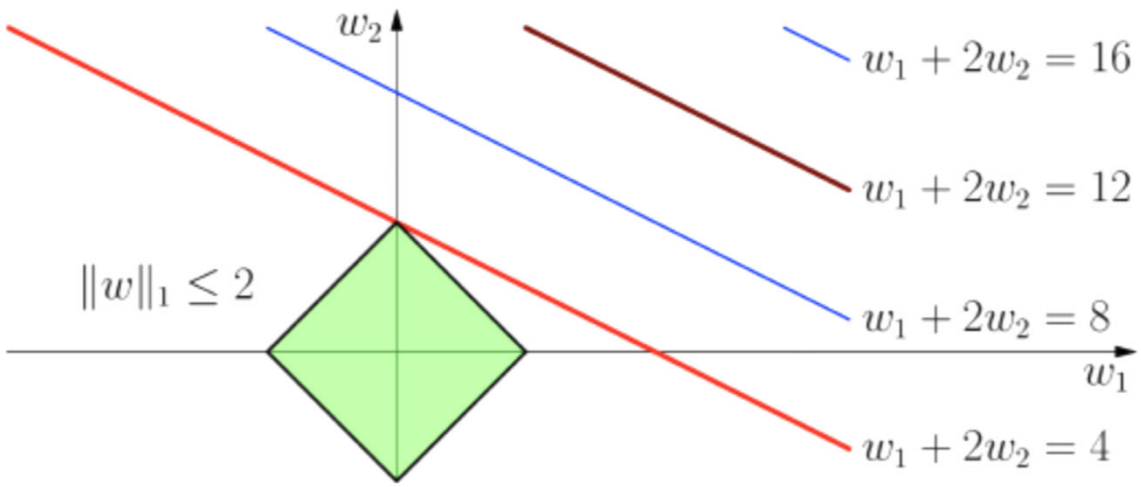


Repeated Features

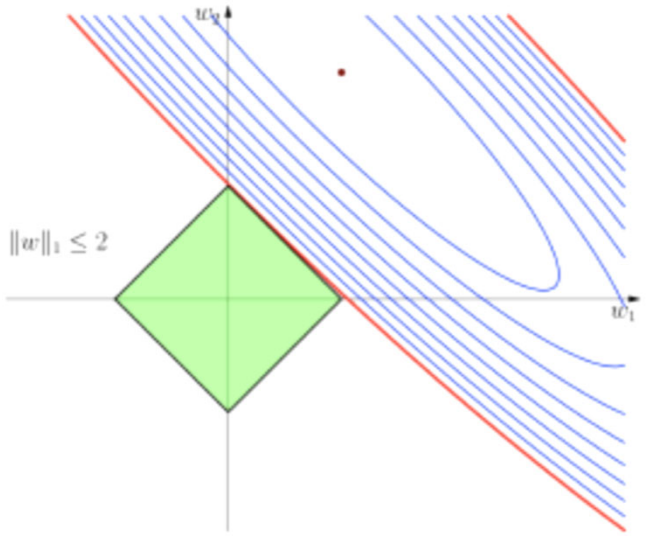
- Suppose we have dimension two. How do Ridge Regression and Lasso Regression behave with repeated features of different scales?
- Take

$$\hat{f}(x_1, x_2) = w_1 x_1 + w_2 x_2$$

where $2x_1 = x_2$.

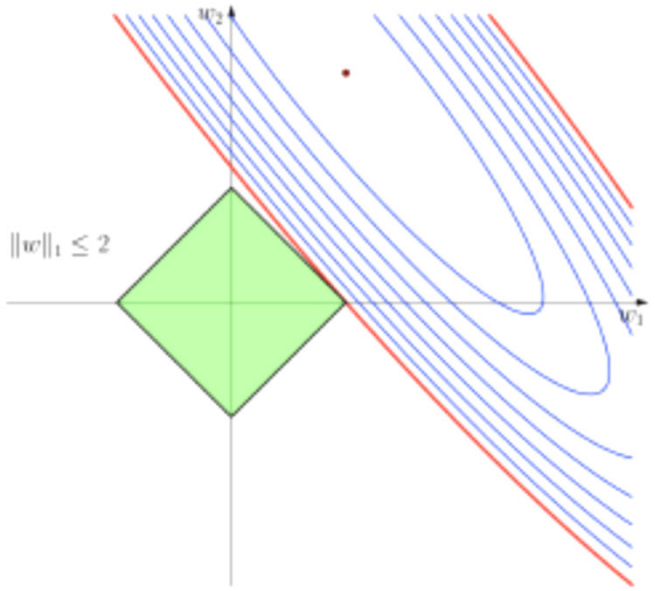


Correlated Features



- With correlated features the level sets are eccentric ellipses. If the features have the same scale then division of weight seemingly arbitrary

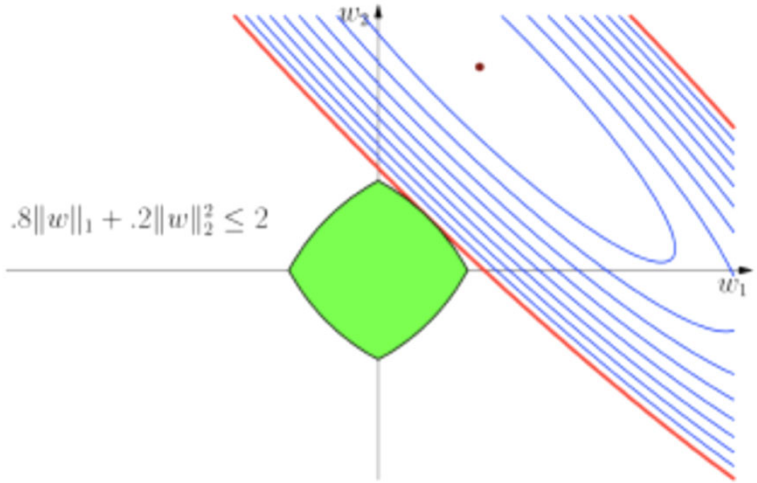
- If the features have different scale, then ellipse will hit a corner unpredictably leading to unstable weights.

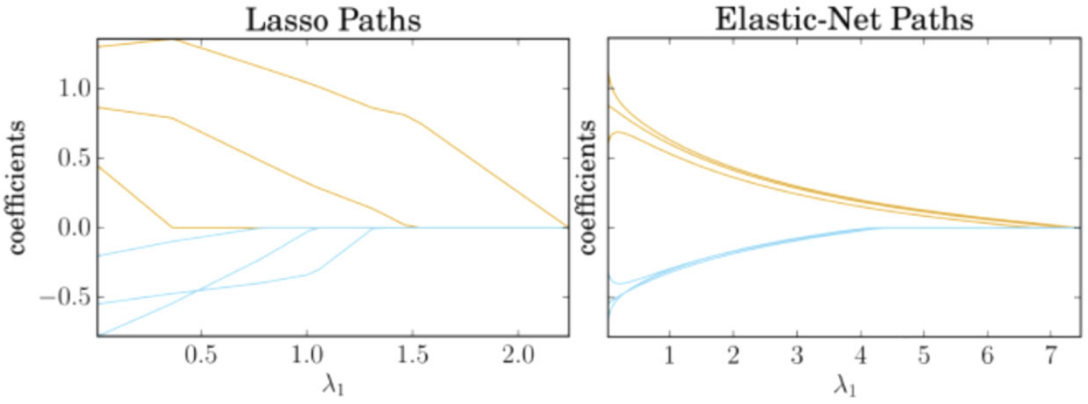


Elastic Net

- We can combine both forms of penalization into the objective function

$$\hat{w} = \underset{w \in \mathbf{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \{w^T x_i - y_i\}^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$$





- With correlated features, we obtain some averaging to prevent against instability

Feature Selection

- What are algorithmic approaches to selecting features besides Lasso Regression
- General approach is
  - Select subset of features for the model
  - Score each set of features
  - Select set of features with high scores

Forward feature selection

- F_0 is empty.
- For F_t choose hypothesis function h_t.
- Select next best feature X_i.
  - Here compared to h_t
- Set $F_{t+1} = F_t \cup X_i$
- Repeat

Feature Selection

- What are algorithmic approaches to selecting features besides Lasso Regression
- General approach is
  - Select subset of features for the model
  - Score each set of features
  - Select set of features with high scores

Backward feature selection

- F_0 contains all features.
- For F_t choose hypothesis function h_t.
- Select next worst feature X_i.
  - Here compared to h_t
- Set $F_{t+1} = F_t - X_i$
- Repeat

Projected Gradient Descent

- How can we fit the parameters in Lasso Regression to the data in the training set?
- We have to handle the absolute value in the objective function

$$\min_{w \in \mathbf{R}^d} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 + \lambda \|w\|_1$$

- Suppose we have a number a. We can split it into two numbers that help us compute the absolute value.

$$a^+ = \begin{cases} a & \text{if } a > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$a^- = \begin{cases} -a & \text{if } a < 0 \\ 0 & \text{otherwise} \end{cases}$$

$$a = a^+ - a^-$$

$$|a| = a^+ + a^-$$

Projected Gradient Descent

- How can we fit the parameters in Lasso Regression to the data in the training set?
- We have to handle the absolute value in the objective function

$$\min_{w \in \mathbf{R}^d} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 + \lambda \|w\|_1$$

- Split the entries of the weight w into positive parts and negative parts. Denote the vectors as a and b.

$$\min_{w} \min_{a,b} \sum_{i=1}^{n} \left((a-b)^T x_i - y_i\right)^2 + \lambda 1^T (a+b)$$

$$\text{subject to} \quad a_i \geqslant 0 \text{ for all } i \qquad b_i \geqslant 0 \text{ for all } i,$$

$$a - b = w$$

$$a + b = |w|$$

Equivalent minimization problem for the lasso objective function. Note **1** means a vector of all one's

Projected Gradient Descent

- Note that we get smaller value through the replacements

$$a' \leftarrow a - \min(a, b)$$

$$b' \leftarrow b - \min(a, b)$$

- So if a and b are minimizing values, then either $a = 0$ or $b = 0$
- Since $a - b = w$, this means that

$$a = w^+$$

$$b = w^-$$

- Therefore we must have $a + b = |w|$ for a and b minimizing values.

$$\min_{w} \min_{a,b} \sum_{i=1}^{n} \left((a-b)^T x_i - y_i \right)^2 + \lambda 1^T (a+b)$$

$$\text{subject to} \quad a_i \geqslant 0 \text{ for all } i \qquad b_i \geqslant 0 \text{ for all } i,$$

$$a - b = w$$

$$a + b = |w|$$

Equivalent minimization problem for the lasso objective function. Note **1** means a vector of all one's

Projected Gradient Descent

- Note that we get smaller value through the replacements

$$a' \leftarrow a - \min(a, b)$$

$$b' \leftarrow b - \min(a, b)$$

- So if a and b are minimizing values, then either $a = 0$ or $b = 0$
- Since $a - b = w$, this means that

$$a = w^+$$

$$b = w^-$$

- Therefore we must have $a + b = |w|$ for a and b minimizing values.

$$\min_w \min_{a,b} \sum_{i=1}^{n} \left((a-b)^T x_i - y_i \right)^2 + \lambda 1^T (a+b)$$

$$\text{subject to} \quad a_i \geqslant 0 \text{ for all } i \qquad b_i \geqslant 0 \text{ for all } i,$$

$$a - b = w$$

Equivalent minimization problem dropping the constraint $a + b = |w|$.

Projected Gradient Descent

- Note that we get smaller value through the replacements

$$a' \leftarrow a - \min(a, b)$$

$$b' \leftarrow b - \min(a, b)$$

- So if *a* and *b* are minimizing values, then either *a* = 0 or *b* = 0
- Since *a* – *b* = w, this means that

$$a = w^+$$

$$b = w^-$$

- Therefore we must have a + b = |w| for *a* and *b* minimizing values.

$$\min_{a,b} \min_{w} \sum_{i=1}^{n} \left((a-b)^T x_i - y_i \right)^2 + \lambda 1^T (a+b)$$

$$\text{subject to} \quad a_i \geqslant 0 \text{ for all } i \qquad b_i \geqslant 0 \text{ for all } i,$$

$$a - b = w$$

Switch the order of minimization to show that w no longer necessary.

Projected Gradient Descent

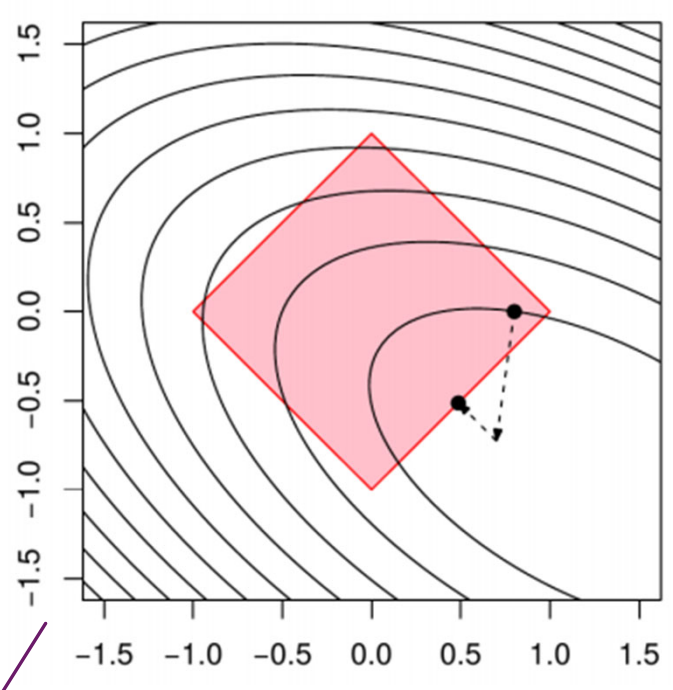
- We have show that the minimization problem for lasso regression

$$\min_{w \in \mathbf{R}^d} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 + \lambda \|w\|_1$$

is equivalent to

$$\min_{a,b} \quad \sum_{i=1}^{n} \left((a-b)^T x_i - y_i\right)^2 + \lambda 1^T (a+b)$$

$$\text{subject to} \quad a_i \geqslant 0 \text{ for all } i \qquad b_i \geqslant 0 \text{ for all } i,$$

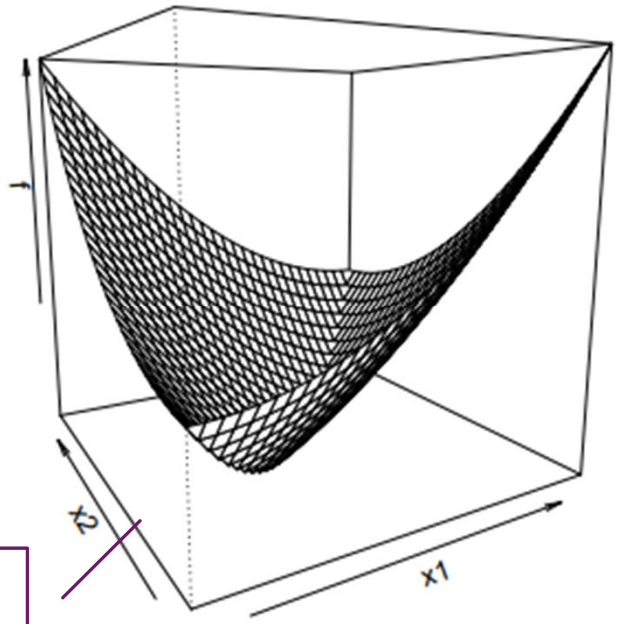


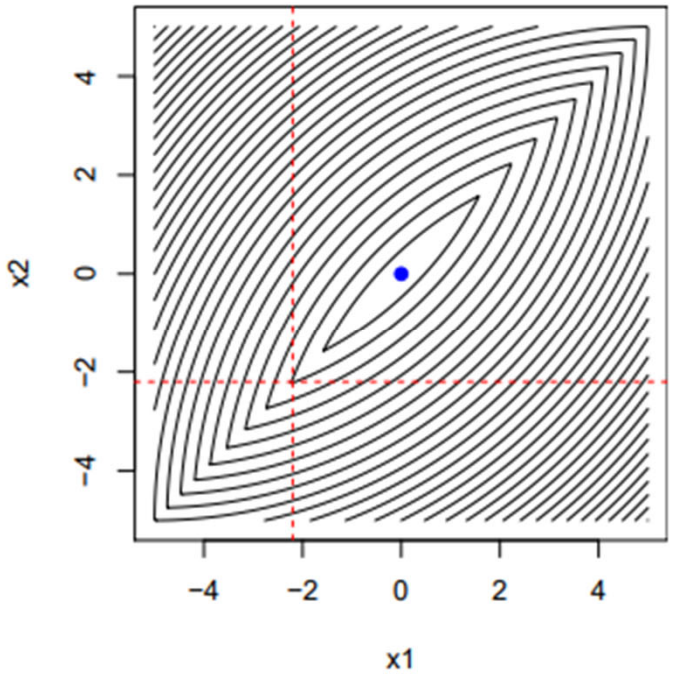
Run gradient descent projecting the weights back to 0 for negative values. See numpy.clip

Coordinate Descent

Notes

Can we apply gradient descent for each coordinate?





Updating coordinate by coordinate might not converge for functions with corners

Summary

- Ridge Regression
  - Solving for minimizer of square loss
- Lasso Regression
  - Connection to feature selection
  - How to fit parameters to data?
- Elastic Net
  - Behavior of Ridge and Lasso with related features?

- Goals
  - What are the differences between ridge and lasso particularly for related features?
  - How does projected gradient descent or coordinate descent help us with lasso regression?
  - What are other methods for feature selection?

Questions

ADALINE Algorithm

- Questions on Piazza?
  - Please provide your feedback
- Question for You!
  - Could square Loss be used for Classification?