Lasso Regression

Nipun Batra

February 5, 2020

IIT Gandhinagar

Lasso Regression

- LASSO $\longrightarrow$ Least absolute shrinkage and selection operator

Lasso Regression

- LASSO $\longrightarrow$ Least absolute shrinkage and selection operator
- Popular as it leads to a sparse solution.

Constructing the Objective Function

- Find a θ_{opt} such that

$$\theta_{opt} = \arg\min_{\theta} (Y - X\theta)^T (Y - X\theta) : \ \|\theta\|_1 < s \tag{1}$$

Constructing the Objective Function

- Find a θ_{opt} such that

$$\theta_{opt} = \arg\min_{\theta} (Y - X\theta)^T (Y - X\theta) : \; \|\theta\|_1 < s \tag{1}$$

- Using KKT conditions

$$\theta_{opt} = \underbrace{\arg\min_{\theta} (Y - X\theta)^T (Y - X\theta) + \delta^2 \|\theta\|_1}_{\text{convex function}} \tag{2}$$

Solving the Objective

- Since $|\theta|$ is not differentiable, we cannot solve,

$$\frac{\partial (Y - X\theta)^T (Y - X\theta) + \delta^2 \|\theta\|_1}{\partial \theta} = 0 \tag{3}$$

Solving the Objective

- Since $|\theta|$ is not differentiable, we cannot solve,

$$\frac{\partial (Y - X\theta)^T (Y - X\theta) + \delta^2 \|\theta\|_1}{\partial \theta} = 0 \tag{3}$$

- How to Solve? Use Coordinate descent!

Sample Dataset

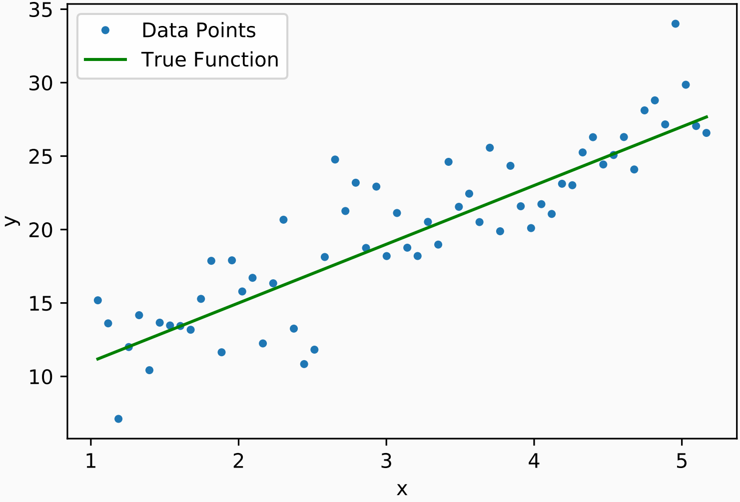


Figure 1: $y = 4x + 7$

Geometric Interpretation

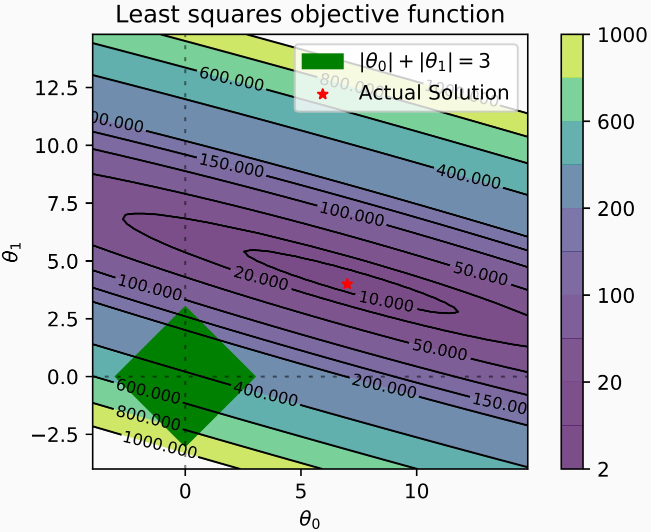
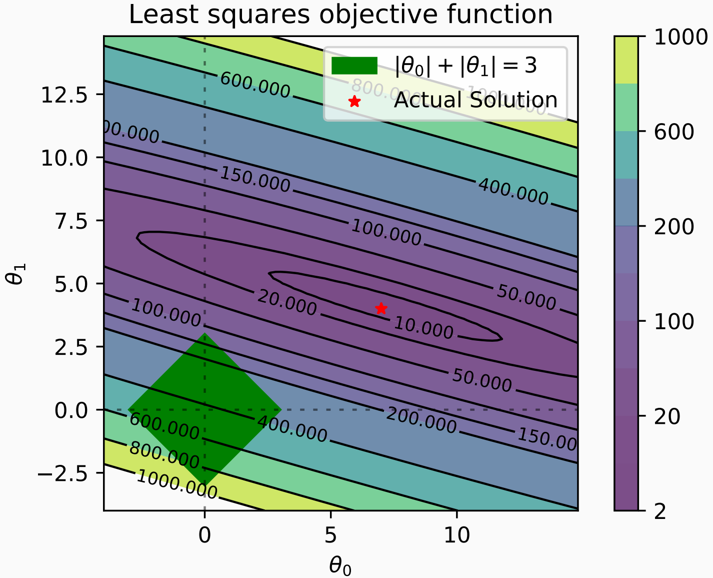


Figure 2: Lasso regression

Effect of μ - Regularization of Parameters



Effect of μ - Regularization of Parameters

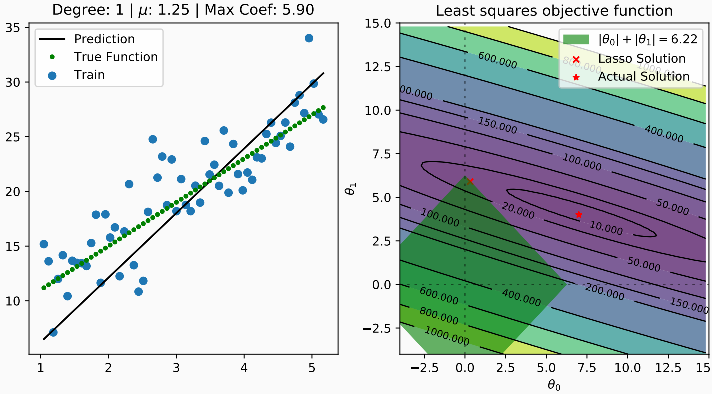


Figure 4: $\mu = 1.25$
(on the *Sample Dataset*)

Effect of μ - Regularization of Parameters

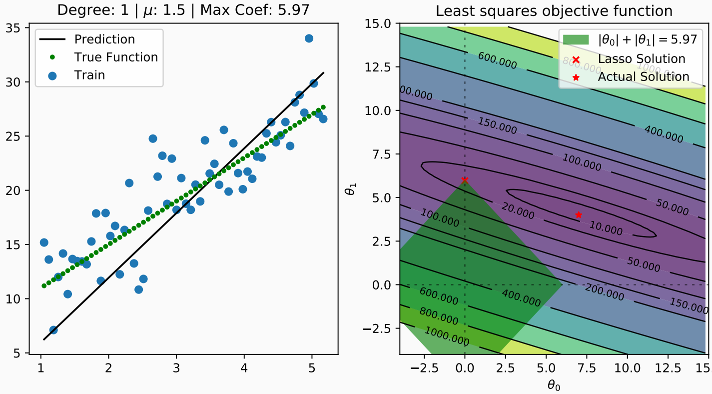


Figure 5: $\mu = 1.5$
(on the *Sample Dataset*)

Effect of μ - Regularization of Parameters

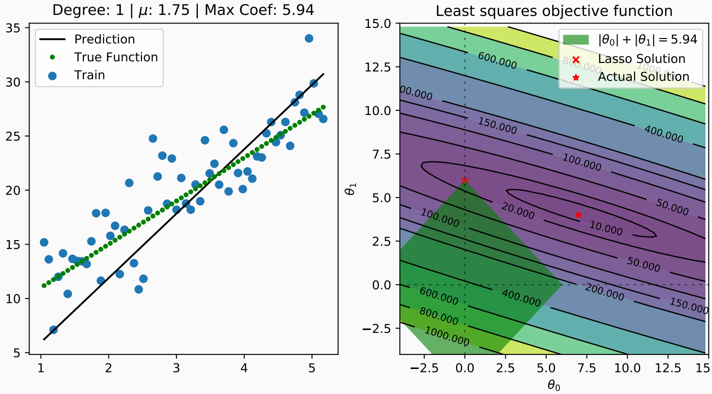


Figure 6: $\mu = 1.75$
(on the *Sample Dataset*)

Effect of μ - Regularization of Parameters

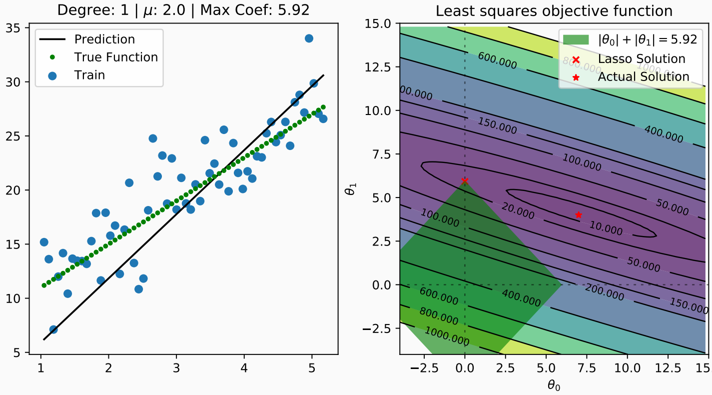
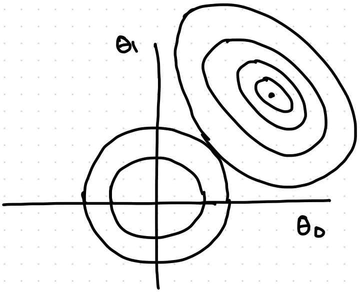


Figure 7: $\mu = 2.0$
(on the *Sample Dataset*)

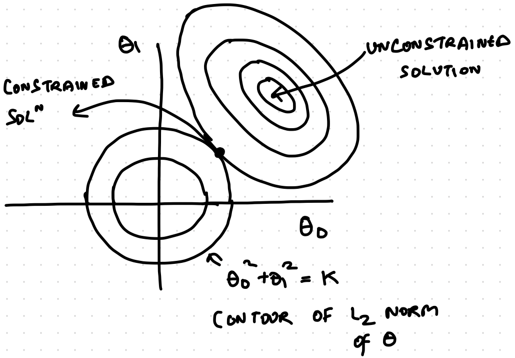
WHY LASSO GIVES SPARSITY

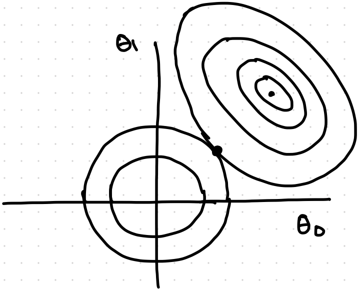
1. GEOMETRIC INTERPRETATION

2. G.D. BASED INTERPRETATION

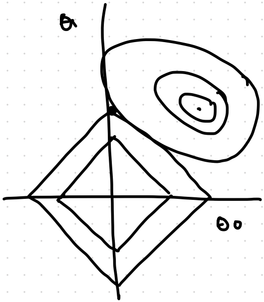


θ1
θ0

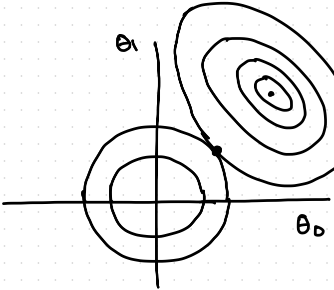

θ1
UNCONSTRAINED
SOLUTION
CONSTRAINED
SOLn
θ0
$\theta_0^2 + \theta_1^2 = K$
CONTOUR OF L2 NORM
OF θ



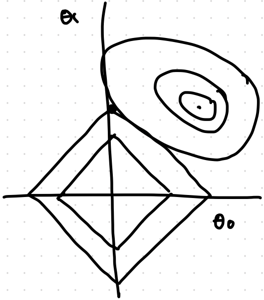
L2 NORM



L1 NORM

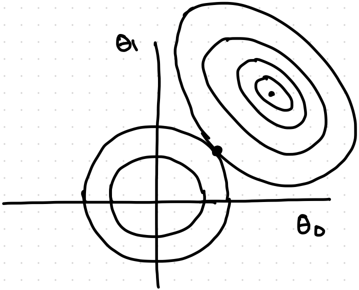


L_2 NORM

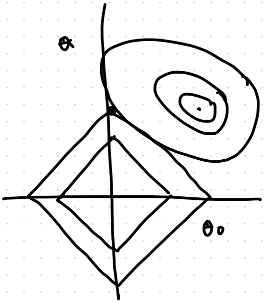


L_1 NORM

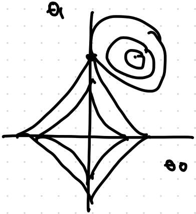
L_p NORM
$(0 < p < 1)$



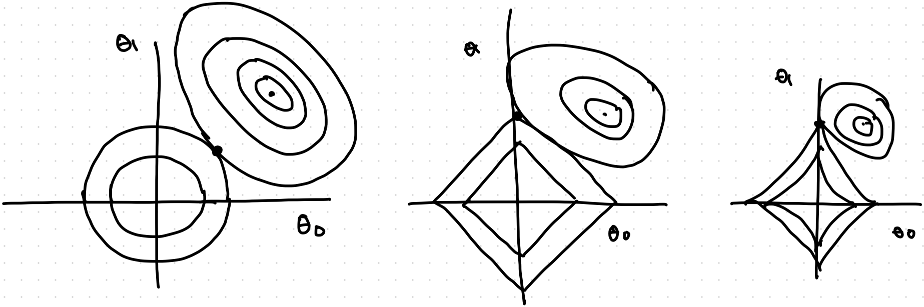
L_2 NORM



L_1 NORM



L_p NORM
$(0 < p < 1)$



L_2 NORM — L_1 NORM — L_p NORM $(0 < p < 1)$

SPARSITY →

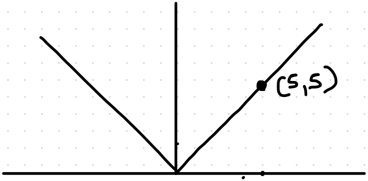
PROB. OF INTERSECTING AXIS →

DIFFICULTY OF SOLVING →

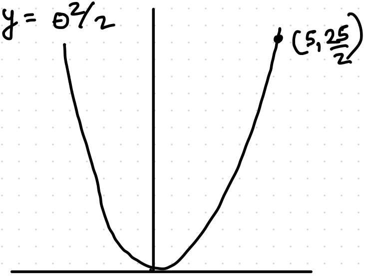
$y = |\theta|$ (FOR NOW ASSUME $\theta > 0$)

$y = \theta^2/2$

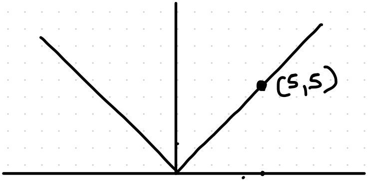
$y = |\theta|$ (FOR NOW ASSUME $\theta > 0$)



$y = \theta^2/2$

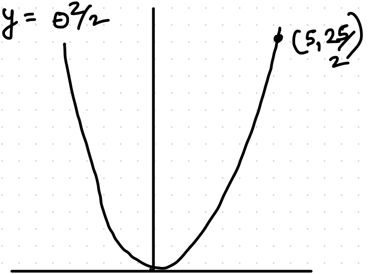


$y = |\theta|$ (FOR NOW ASSUME $\theta > 0$)



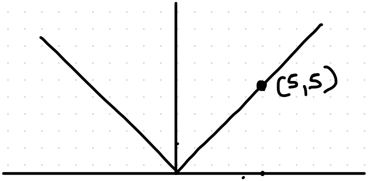
$\frac{\partial y}{\partial \theta} = 1$ (Assume $\theta > 0$)

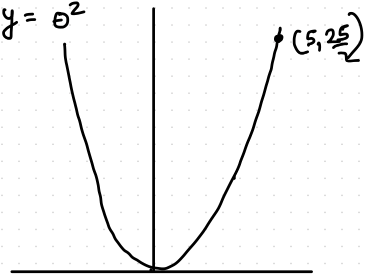
$y = \theta^2/2$



$\frac{\partial y}{\partial \theta} = \frac{2\theta}{2} = \theta$

$y = |\theta|$ (FOR NOW ASSUME $\theta > 0$) $y = \theta^2$





$\frac{\partial y}{\partial \theta} = 1$ (Assume $\theta > 0$)

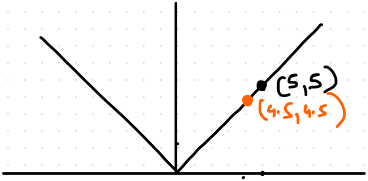
$\frac{\partial y}{\partial \theta} = \frac{2\theta}{2} = \theta$

LET $\alpha = 0.5$

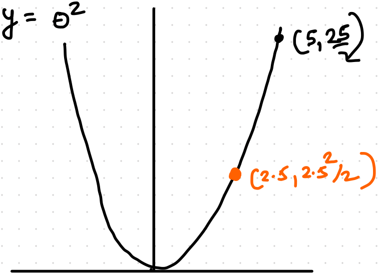
$\theta_0^1 = \theta_0^0 - 0.5 * 1 = 4.5$

$\theta_0^1 = \theta_0^0 - 0.5 * 5 = 2.5$

$y = |\theta|$ (FOR NOW ASSUME $\theta > 0$)



$y = \theta^2$



$\frac{\partial y}{\partial \theta} = 1$ (Assume $\theta > 0$)

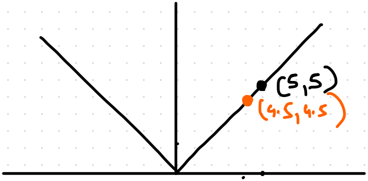
$\frac{\partial y}{\partial \theta} = \frac{2\theta}{2} = \theta$

LET $\alpha = 0.5$

$\theta_0^1 = \theta_0^0 - 0.5 * 1 = 4.5$

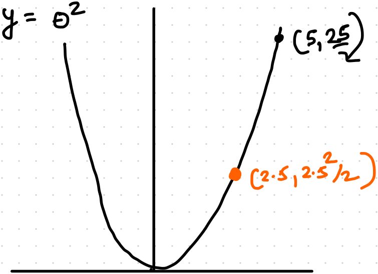
$\theta_0^1 = \theta_0^0 - 0.5 * 5 = 2.5$

$y = |\theta|$ (FOR NOW ASSUME $\theta > 0$)



$\frac{\partial y}{\partial \theta} = 1$ (Assume $\theta > 0$)

$y = \theta^2$



$\frac{\partial y}{\partial \theta} = \frac{2\theta}{2} = \theta$

LET $\alpha = 0.5$

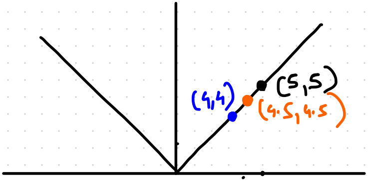
$\theta_0^1 = \theta_0^0 - 0.5 * 1 = 4.5$

$\theta_0^2 = \theta_0^1 - 0.5 \times 1 = 4.0$

$\theta_0^1 = \theta_0^0 - 0.5 * 5 = 2.5$

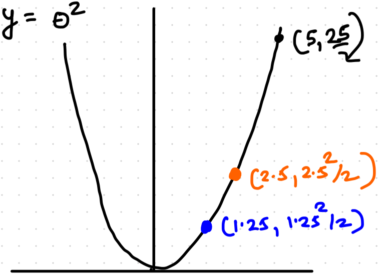
$\theta_0^2 = \theta_0^1 - 0.5 \times 2.5 = 1.25$

$y = |\theta|$ (FOR NOW ASSUME $\theta > 0$)



$\frac{\partial y}{\partial \theta} = 1$ (Assume $\theta > 0$)

$y = \theta^2$



$\frac{\partial y}{\partial \theta} = \frac{2\theta}{2} = \theta$

LET $\alpha = 0.5$

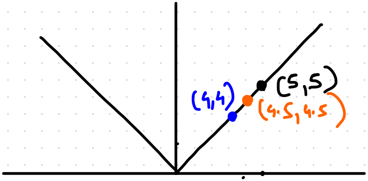
$\theta_0^1 = \theta_0^0 - 0.5 * 1 = 4.5$

$\theta_0^2 = \theta_0^1 - 0.5 \times 1 = 4.0$

$\theta_0^1 = \theta_0^0 - 0.5 * 5 = 2.5$

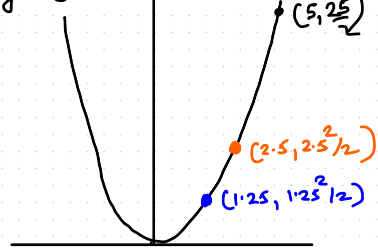
$\theta_0^2 = \theta_0^1 - 0.5 \times 2.5 = 1.25$

$y = |\theta|$ (FOR NOW ASSUME $\theta > 0$)



$\frac{\partial y}{\partial \theta} = 1$ (Assume $\theta > 0$)

$y = \theta^2$



$\frac{\partial y}{\partial \theta} = \frac{2\theta}{2} = \theta$

LET $\alpha = 0.5$

$\theta_0^1 = \theta_0^0 - 0.5 \times 1 = 4.5$

$\theta_0^2 = \theta_0^1 - 0.5 \times 1 = 4.0$

$\vdots$

$\theta_0^t = \theta_0^{t-1} - 0.5$

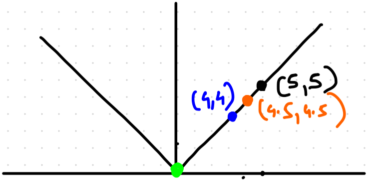
$\theta_0^1 = \theta_0^0 - 0.5 * 5 = 2.5$

$\theta_0^2 = \theta_0^1 - 0.5 \times 2.5 = 1.25$

$\vdots$

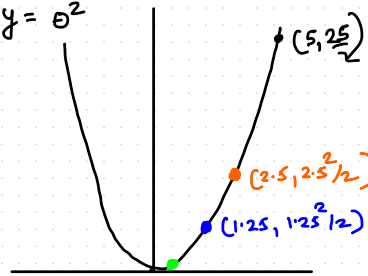
$\theta_0^t = \theta_0^{t-1} - 0.5\,\theta_0^{t-1} = 0.5\,\theta_0^{t-1}$

$y = |\theta|$ (FOR NOW ASSUME $\theta > 0$)



$\frac{\partial y}{\partial \theta} = 1$ (Assume $\theta > 0$)

$y = \theta^2$



$\frac{\partial y}{\partial \theta} = \frac{2\theta}{2} = \theta$

LET $\alpha = 0.5$

$\theta_0^{10} = 0$

$\theta_0^{10} = 5 * (0.5)^{10}$

$= 0.0048$

(Approaching 0 but not exactly zero)

Regularization path of lasso regression

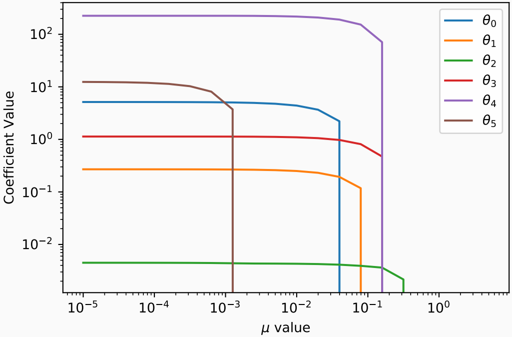


Figure 8: Regularization path of θ_i

LASSO and feature selection

- LASSO inherently does feature selection!

LASSO and feature selection

- LASSO inherently does feature selection!
- Sets coefficients of “less important” features to zero.

LASSO and feature selection

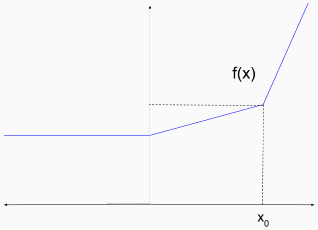
- LASSO inherently does feature selection!
- Sets coefficients of “less important” features to zero.
- Sparse and memory efficient and often more interpretable models.

Subgradient

- Generalizes gradient to convex but non-differentiable problems
- Examples:
  - $f(x) = |x|$

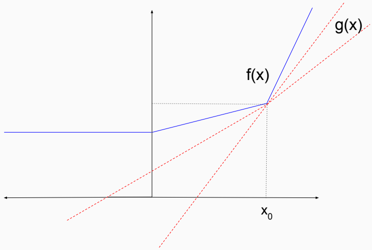
Task at hand

- TASK: find derivative of $f(x)$ at x = x^0



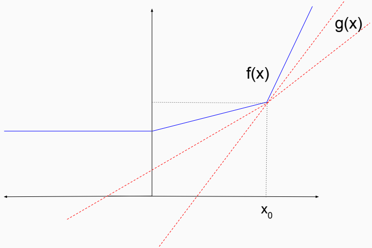
Solution

- Construct a differentiable g(x)
  - Intersecting $f(x)$ at $x = x_0$
  - Below or on $f(x)$ for all x



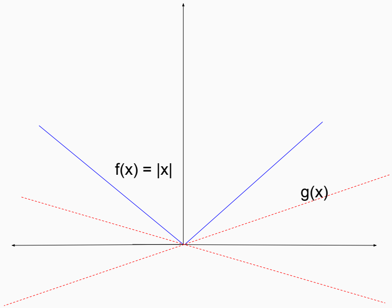
Solution

- Compute slope of g(x) at $x = x_0$



Another Example: f(x) = | x |

- Subgradient of f(x) belongs to [-1, 1]



Coordinate Descent

- Another optimisation method (akin to gradient descent)

Coordinate Descent

- Another optimisation method (akin to gradient descent)
- Objective: $\text{Min}_{\theta} f(\theta)$

Coordinate Descent

- Another optimisation method (akin to gradient descent)
- Objective: $\text{Min}_\theta f(\theta)$
- Key idea: Sometimes difficult to find minimum for all coordinates

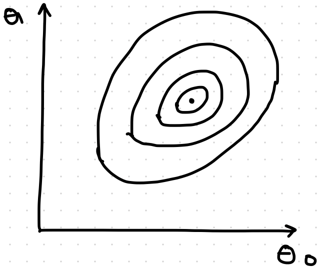
Coordinate Descent

- Another optimisation method (akin to gradient descent)
- Objective: $\min_{\theta} f(\theta)$
- Key idea: Sometimes difficult to find minimum for all coordinates
- ..., but, easy for each coordinate

Coordinate Descent

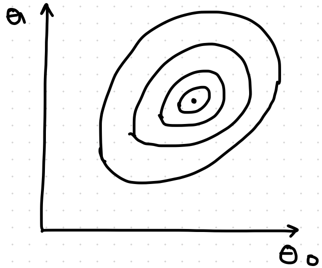
- Another optimisation method (akin to gradient descent)
- Objective: $\text{Min}_\theta f(\theta)$
- Key idea: Sometimes difficult to find minimum for all coordinates
- ..., but, easy for each coordinate
- turns into a 1D optimisation problem

COORDINATE DESCENT ALGORITHM



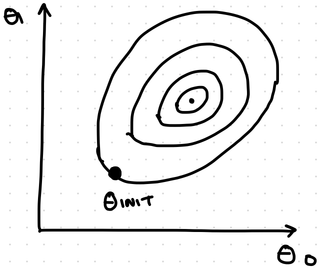
COORDINATE DESCENT ALGORITHM

GOAL: $\min_{\theta} f(\theta)$



COORDINATE DESCENT ALGORITHM

1) INIT θ

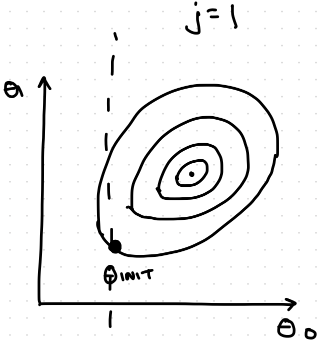


COORDINATE DESCENT ALGORITHM

1) INIT θ

2) WHILE NOT CONVERGED

   2.1) PICK COORDINATE 'j'



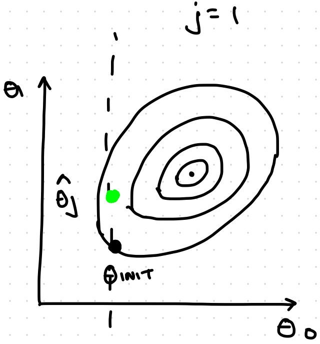
COORDINATE DESCENT ALGORITHM

1) INIT θ

2) WHILE NOT CONVERGED

   2.1) PICK COORDINATE 'j'

   2.2) $\hat{\theta}_j = \min_{\phi} f(\theta_0, \phi)$



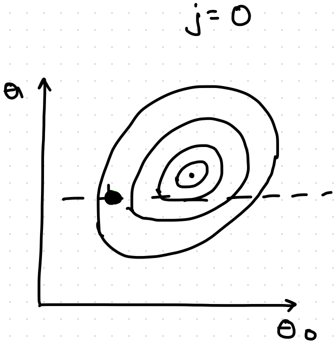
COORDINATE DESCENT ALGORITHM

1) INIT θ

2) WHILE NOT CONVERGED

✓ 2.1) PICK COORDINATE 'j'

2.2) $\hat{\theta}_j = \min_{\phi} f(\phi, \theta_1)$



COORDINATE DESCENT ALGORITHM

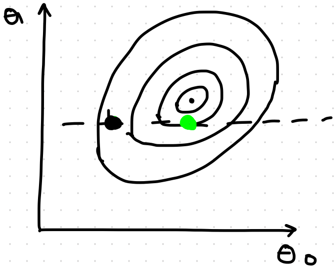
$j = 0$

1) INIT θ

2) WHILE NOT CONVERGED

    2.1) PICK COORDINATE 'j'

    ✓ 2.2) $\hat{\theta}_j = \min_{\phi} f(\theta_0, \phi)$



COORDINATE DESCENT ALGORITHM

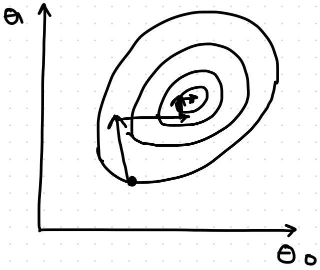
1) INIT θ

2) WHILE NOT CONVERGED

   2.1) PICK COORDINATE 'j'

   2.2) $\hat{\theta}_j = \min_{\phi} f(\theta_0, \phi)$

j = 0



Coordinate Descent

- Picking next coordinate:

Coordinate Descent

- Picking next coordinate:

Coordinate Descent

- Picking next coordinate: random, round-robin
- No step-size to choose!

Coordinate Descent

- Picking next coordinate: random, round-robin
- No step-size to choose!
- Converges for Lasso objective

Coordinate Descent : Example

Learn $y = \theta_0 + \theta_1 x$ on following dataset, using coordinate descent where initially $(\theta_0, \theta_1) = (2, 3)$ for 2 iterations.

x	y
1	1
2	2
3	3

Coordinate Descent : Example

Our predictor, $\hat{y} = \theta_0 + \theta_1 x$

Error for i^{th} datapoint, $\epsilon_i = y_i - \hat{y}_i$

$\epsilon_1 = 1 - \theta_0 - \theta_1$

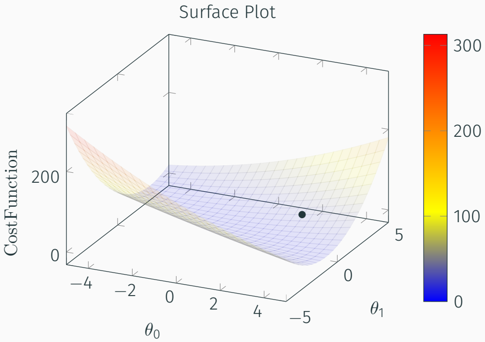
$\epsilon_2 = 2 - \theta_0 - 2\theta_1$

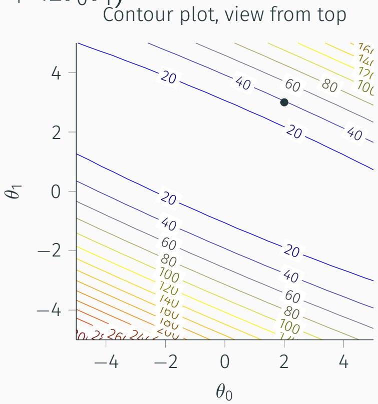
$\epsilon_3 = 3 - \theta_0 - 3\theta_1$

MSE = $\dfrac{\epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2}{3}$ = $\dfrac{14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1}{3}$

Iteration 0

$$\text{MSE} = \frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$





Iteration 1

INIT: $\theta_0 = 2$ and $\theta_1 = 3$

$\theta_1 = 3$ optimize for θ_0

Iteration 1

INIT: $\theta_0 = 2$ and $\theta_1 = 3$

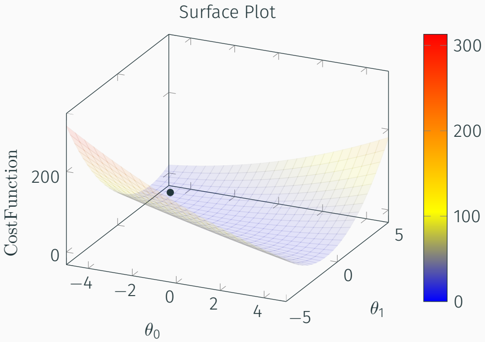
$\theta_1 = 3$ optimize for θ_0

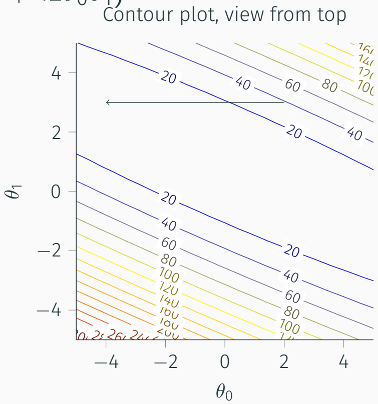
$\frac{\partial MSE}{\partial \theta_0} = 6\theta_0 + 24 = 0$

$\theta_0 = -4$

Iteration 1

$$\text{MSE} = \frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$





Coordinate Descent : Example

Iteration 2

INIT: $\theta_0 = -4$ and $\theta_1 = 3$

$\theta_0 = -4$ optimize for θ_1

Iteration 2

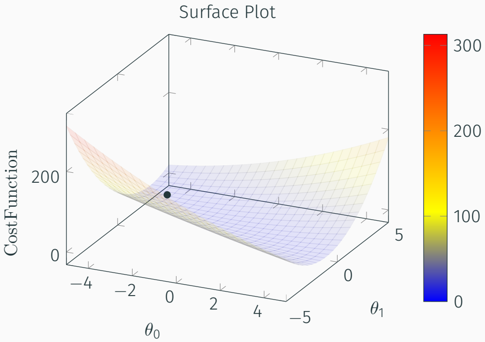
INIT: $\theta_0 = -4$ and $\theta_1 = 3$

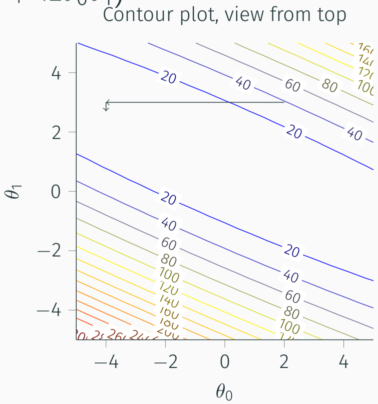
$\theta_0 = -4$ optimize for θ_1

$\theta_1 = 2.7$

Iteration 2

$$\text{MSE} = \frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$





Coordinate Descent : Example

Iteration 3

INIT: $\theta_0 = -4$ and $\theta_1 = 2.7$

$\theta_1 = 2.7$ optimize for θ_0

Iteration 3

INIT: $\theta_0 = -4$ and $\theta_1 = 2.7$

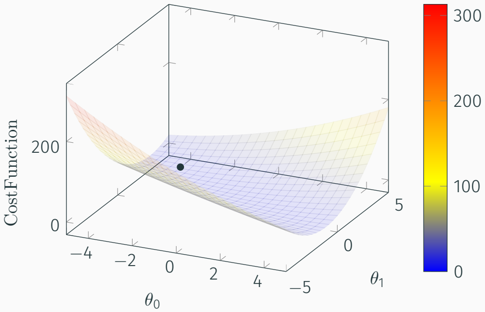
$\theta_1 = 2.7$ optimize for θ_0

$\theta_0 = -3.4$

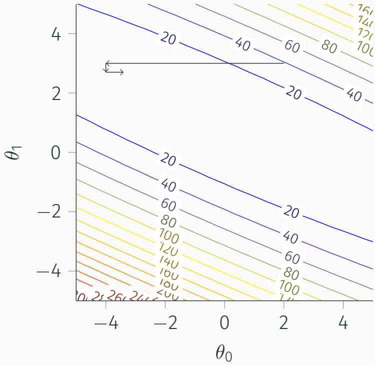
Iteration 3

$$\text{MSE} = \frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$

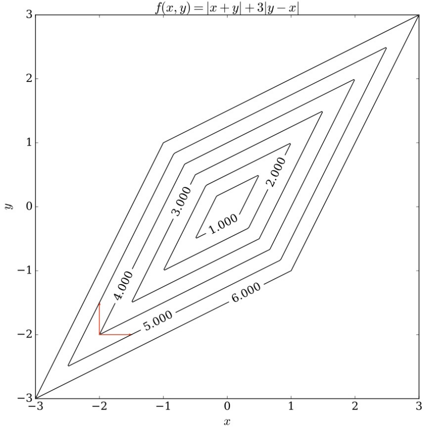
Surface Plot

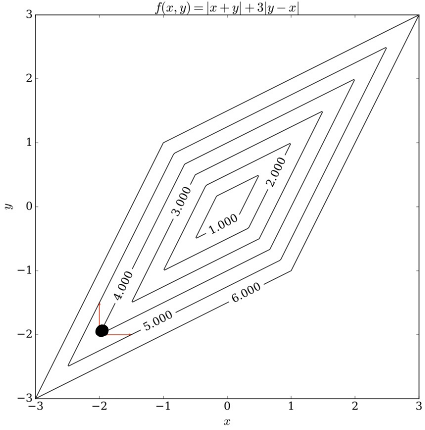


Contour plot, view from top



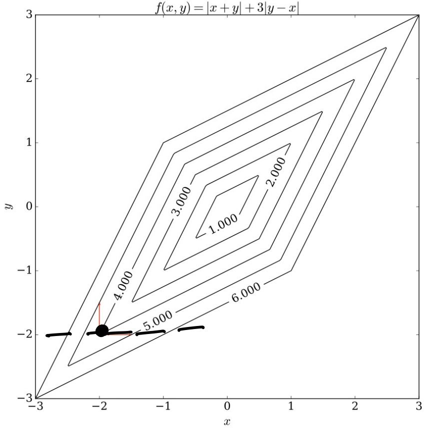
FAILURE OF COORDINATE DESCENT





FAILURE OF COORDINATE DESCENT

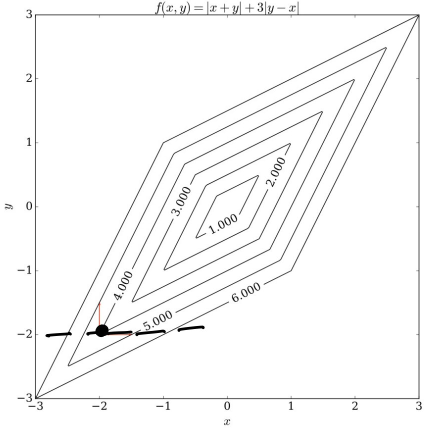
START WITH $(x,y) = (-2,-2)$



FAILURE OF COORDINATE DESCENT

START WITH $(x, y) = (-2, -2)$

FIX $y = -2$, OPTIMIZE ABOUT x.

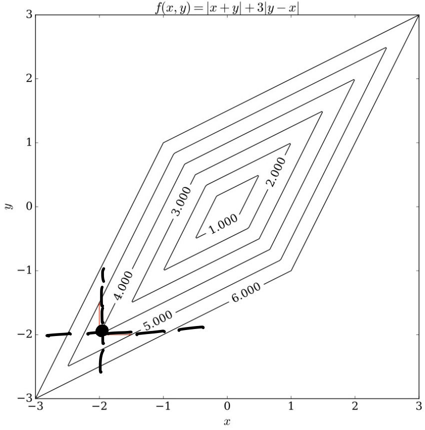


FAILURE OF COORDINATE DESCENT

START WITH $(x, y) = (-2, -2)$

FIX $y = -2$, OPTIMIZE ABOUT x.

OBJECTIVE INCREASES IN BOTH DIRECTIONS



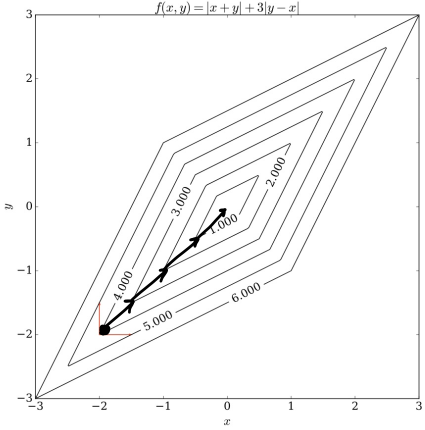
FAILURE OF COORDINATE DESCENT

START WITH $(x, y) = (-2, -2)$

FIX $y = -2$, OPTIMIZE ABOUT x.

OBJECTIVE INCREASES IN BOTH DIRECTIONS

SIMILAR IF WE FIX x and OPTIMIZE ABOUT y.



GRADIENT DESCENT WILL WORK!

- NEED SIMULTANEOUS UPDATE IN BOTH COORDINATES

Coordinate Descent for Unregularized Regression

- Express error as a difference of y_i and $\hat{y}_i$

$$\hat{y}_i = \sum_{j=0}^{d} \theta_j x_i^j = \theta_0 x_i^0 + \theta_1 x_i^1 + \theta_2 x_i^2 ...... + \theta_d x_i^d \tag{4}$$

$$\epsilon_i = y_i - \hat{y}_i \tag{5}$$

$$= y_i - \theta_0 x_i^0 + \theta_1 x_i^1 + ... + \theta_d x_i^d \tag{6}$$

$$= y_i - \sum_{j=0}^{d} \theta_j x_i^j \tag{7}$$

Coordinate Descent for Unregularized regression

$$\sum_{i=1}^{N} \epsilon^2 = RSS = \sum_{i=1}^{N} \left(y_i - \left(\theta_0 x_i^0 + \ldots \quad \theta_j x_i^j + \theta_d x_i^d \right) \right)^2$$

Coordinate Descent for Unregularized regression

$$\sum_{i=1}^{N} \epsilon^2 = RSS = \sum_{i=1}^{N} \left(y_i - \left(\theta_0 x_i^0 + \dots \quad \theta_j x_i^j + \theta_d x_i^d \right) \right)^2$$

$$\frac{\partial \, \mathrm{RSS}\left(\theta_j\right)}{\partial \theta_j} = 2 \sum_{i=1}^{N} \left(y_i - \left(\theta_0 x_i^0 + \dots \quad \theta_j x_i^j + \dots \right) \right) \left(-x_i^j \right)$$

Coordinate Descent for Unregularized regression

$$\sum_{i=1}^{N} \epsilon^2 = RSS = \sum_{i=1}^{N} \left(y_i - \left(\theta_0 x_i^0 + \dots \quad \theta_j x_i^j + \theta_d x_i^d\right)\right)^2$$

$$\frac{\partial \, \mathrm{RSS}\left(\theta_j\right)}{\partial \theta_j} = 2 \sum_{i=1}^{N} \left(y_i - \left(\theta_0 x_i^0 + \dots \quad \theta_j x_i^j + \dots\right)\right)\left(-x_i^j\right)$$

$$= 2 \sum_{i=1}^{N} \left(y_i - \left(\theta_0 x_i^0 + \dots + \theta_d x_i^d\right)\right)\left(-x_i^j\right) + 2 \sum_{i=1}^{N} \theta_j (x_i^j)^2$$

Coordinate Descent for Unregularized regression

$$\sum_{i=1}^{N} \epsilon^2 = RSS = \sum_{i=1}^{N} \left(y_i - \left(\theta_0 x_i^0 + \dots \quad \theta_j x_i^j + \theta_d x_i^d\right)\right)^2$$

$$\frac{\partial \, \mathrm{RSS}\,(\theta_j)}{\partial \theta_j} = 2 \sum_{i=1}^{N} \left(y_i - \left(\theta_0 x_i^0 + \dots \quad \theta_j x_i^j + \dots\right)\right)\left(-x_i^j\right)$$

$$= 2 \sum_{i=1}^{N} \left(y_i - \left(\theta_0 x_i^0 + \dots + \theta_d x_i^d\right)\right)\left(-x_i^j\right) + 2 \sum_{i=1}^{N} \theta_j (x_i^j)^2$$

where:

$$\hat{y}_i^{(-j)} = \theta_0 x_i^0 + \dots + \theta_d x_i^d$$

is $\hat{y}_i$ without θ_j

Coordinate Descent for Unregularized regression

$$Set \frac{\partial \operatorname{RSS}\left(\theta_j\right)}{\partial \theta_j} = 0$$

$$\theta_j = \sum_{i=1}^{N} \frac{\left(y_i - \left(\theta_0 x_i^0 + \ldots + \ldots + \theta_d x_i^d\right)\right)\left(x_i^j\right)}{\left(x_i^j\right)^2} = \frac{\rho_j}{z_j}$$

$$\rho_j = \sum_{i=1}^{N} x_i^j \left(y_i - \hat{y}_i^{(-j)})\right)$$

$$z_j = \sum_{i=1}^{N} \left(x_i^j\right)^2$$

z_j is the squared of ℓ_2 norm of the j^{th} feature

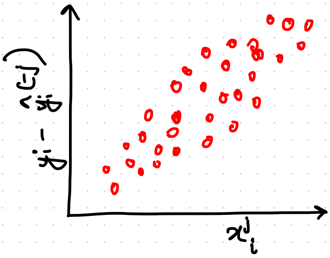
UNDERSTANDING ρ_j IN COORDINATE DESCENT

$$\rho_j = \sum_{i=1}^{N} x_i^j \left(y_i - \hat{y}_i^{(-j)}\right)$$

UNDERSTANDING ρ_j IN COORDINATE DESCENT

$$\rho_j = \sum_{i=1}^{N} x_i^j \left(y_i - \hat{y}_i^{(-j)} \right)$$

CASE 1

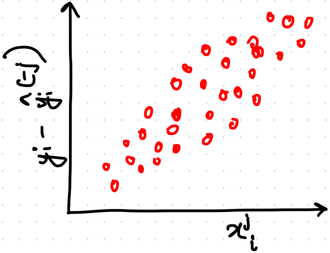


x_i^j STRONG +VE CORR. WITH $y_i - \hat{y}_i^{(-j)}$

UNDERSTANDING ρ_j IN COORDINATE DESCENT

$$\rho_j = \sum_{i=1}^{N} x_i^j \left(y_i - \hat{y}_i^{(-j)} \right)$$

CASE 1



x_i^j STRONG +VE CORR. WITH $y_i - \hat{y}_i^{(-j)}$

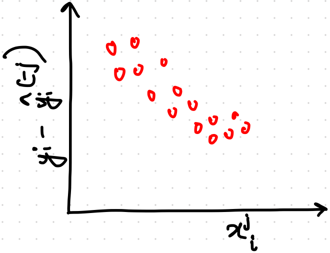
$\Downarrow$

j^{th} FEATURE IS IMPT. AND ITS COEFFICIENT +VE

UNDERSTANDING ρ_j IN COORDINATE DESCENT

$$\rho_j = \sum_{i=1}^{N} x_i^j \left(y_i - \hat{y}_i^{(-j)}\right)$$

CASE II



x_i^j STRONG -VE CORR. WITH $y_i - \hat{y}_i^{(-j)}$

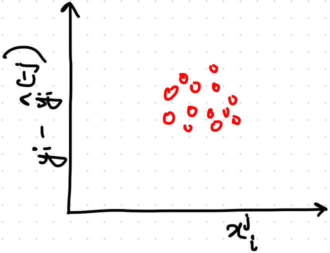
⇓

j^{th} FEATURE IS IMPT. AND ITS COEF. -VE

UNDERSTANDING ρ_j IN COORDINATE DESCENT

$$\rho_j = \sum_{i=1}^{N} x_i^j \left(y_i - \hat{y}_i^{(-j)}\right)$$

CASE II



x_i^j WEAK CORR. WITH $y_i - \hat{y}_i^{(-j)}$

$\Downarrow$

j^{th} FEATURE IS NOT IMPT.

AND ITS COEF. $\rightarrow 0$

Coordinate Descent for Lasso Regression

$$\text{Minimize} \underbrace{\sum_{i=1}^{N} \epsilon^2 + \delta^2 \left\{ |\theta_0| + |\theta_1| + \ldots \left|\theta_j\right| + \ldots |\theta_d| \right\}}_{\textit{LASSO OBJECTIVE}}$$

$$\frac{\partial}{\partial \theta_j}(\text{LASSO OBJECTIVE}) = -2\rho_j + 2\theta_j z_j + \delta^2 \frac{\partial}{\partial \theta_j}\left|\theta_j\right|$$

$$\frac{\partial}{\partial \theta_j}\left|\theta_j\right| = \begin{cases} 1 & \theta_j > 0 \\ [-1, 1] & \theta_j = 0 \\ -1 & \theta_j < 0 \end{cases}$$

Coordinate Descent for Lasso Regression

- Case 1: $\theta_j > 0$

$$2\rho_j + 2\theta_j z_j + \delta^2 = 0$$

$$\theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

$$\rho_j > \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

Coordinate Descent for Lasso Regression

- **Case 1:** $\theta_j > 0$

$$2\rho_j + 2\theta_j z_j + \delta^2 = 0$$
$$\theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$
$$\rho_j > \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

- **Case 2:** $\theta_j < 0$

$$\rho_j < \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j + \delta^2/2}{z_j} \tag{8}$$

Coordinate Descent for Lasso Regression

- Case 3: $\theta_j = 0$

$$\frac{\partial}{\partial \theta_j}(\text{LASSO OBJECTIVE}) = -2\rho_j + 2\theta_j z_j + \delta^2 \underbrace{\frac{\partial}{\partial \theta_j} \left|\theta_j\right|}_{\text{[-1,1]}}$$

$$\epsilon \underbrace{\left[-2\rho_j - \delta^2, -2\rho_j + \delta^2\right]}_{\{0\}\text{ lies in this range}}$$

$$-2\rho_j - \delta^2 \leq 0 \text{ and } -2\rho_j - \delta^2 \leq 0$$

$$-\frac{\delta^2}{2} \leq \rho_j \leq \frac{\delta^2}{2} \Rightarrow \theta_j = 0$$

Summary of Lasso Regression

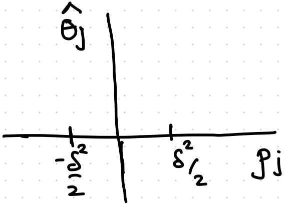
$$\theta_j = \begin{bmatrix} \frac{\rho_j + \frac{\delta^2}{2}}{z_j} & if & \rho_j < -\frac{\delta^2}{2} \\ 0 & if & -\frac{\delta^2}{2} \leq \rho_j \leq \frac{\delta^2}{2} \\ \frac{\rho_j - \frac{\delta^2}{2}}{z_j} & if & \rho_j > \frac{\delta^2}{2} \end{bmatrix} \tag{9}$$

LASSO (SOFT) THRESHOLDING

$$\theta_j = \begin{cases} \frac{\rho_j + \delta^2/2}{z_j} & \text{if} \quad \rho_j < -\delta^2/2 \\ 0 & \text{if} \quad -\frac{\delta^2}{2} \le \rho_j \le \frac{\delta^2}{2} \\ \frac{\rho_j - \delta^2/2}{z_j} & \text{if} \quad \rho_j > \delta^2/2 \end{cases}$$

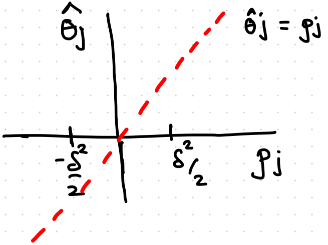
LASSO (SOFT) THRESHOLDING

$$\hat{\theta}_j = \begin{cases} \dfrac{\rho_j + \delta^2/2}{z_j} & \text{if} \quad \rho_j < -\delta^2/2 \\ 0 & \text{if} \quad -\dfrac{\delta^2}{2} \le \rho_j \le \dfrac{\delta^2}{2} \\ \dfrac{\rho_j - \delta^2/2}{z_j} & \text{if} \quad \rho_j > \delta^2/2 \end{cases}$$



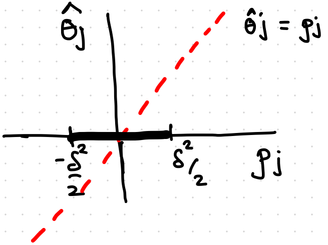
LASSO (SOFT) THRESHOLDING

$$\hat{\theta}_j = \begin{cases} \dfrac{\rho_j + \delta^2/2}{z_j} & \text{if} \quad \rho_j < -\delta^2/2 \\ 0 & \text{if} \quad -\dfrac{\delta^2}{2} \le \rho_j \le \dfrac{\delta^2}{2} \\ \dfrac{\rho_j - \delta^2/2}{z_j} & \text{if} \quad \rho_j > \delta^2/2 \end{cases}$$



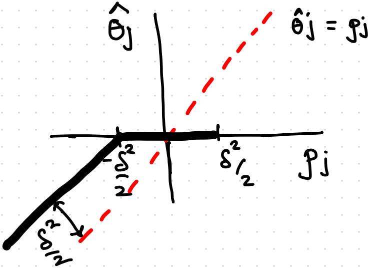
LASSO (SOFT) THRESHOLDING

$$\hat{\theta}_j = \begin{cases} \frac{\rho_j + \delta^2/2}{z_j} & \text{if} \quad \rho_j < -\delta^2/2 \\ 0 & \text{if} \quad -\frac{\delta^2}{2} \le \rho_j \le \frac{\delta^2}{2} \\ \frac{\rho_j - \delta^2/2}{z_j} & \text{if} \quad \rho_j > \delta^2/2 \end{cases}$$



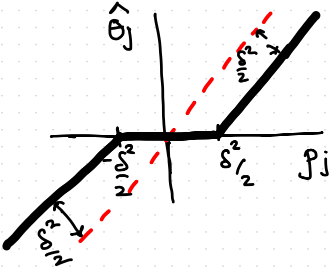
LASSO (SOFT) THRESHOLDING

$$\hat{\theta}_j = \begin{cases} \dfrac{\rho_j + \delta^2/2}{z_j} & \text{if} \quad \rho_j < -\delta^2/2 \\ 0 & \text{if} \quad -\dfrac{\delta^2}{2} \leq \rho_j \leq \dfrac{\delta^2}{2} \\ \dfrac{\rho_j - \delta^2/2}{z_j} & \text{if} \quad \rho_j > \delta^2/2 \end{cases}$$



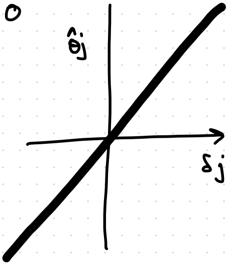
LASSO (SOFT) THRESHOLDING

$$\hat{\theta}_j = \begin{cases} \dfrac{\rho_j + \delta^2/2}{z_j} & \text{if} \quad \rho_j < -\delta^2/2 \\ 0 & \text{if} \quad -\dfrac{\delta^2}{2} \le \rho_j \le \dfrac{\delta^2}{2} \\ \dfrac{\rho_j - \delta^2/2}{z_j} & \text{if} \quad \rho_j > \delta^2/2 \end{cases}$$



LASSO (SOFT) THRESHOLDING

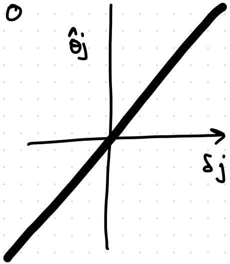
$\delta^2 = 0$



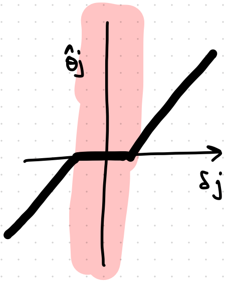
NO REGULARISATION

LASSO (SOFT) THRESHOLDING

$\delta^2 = 0$



NO REGULARISATION



REGULARISATION
⇓
SPARSITY