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,

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

- How to Solve? Use Coordinate descent!

Sample Dataset

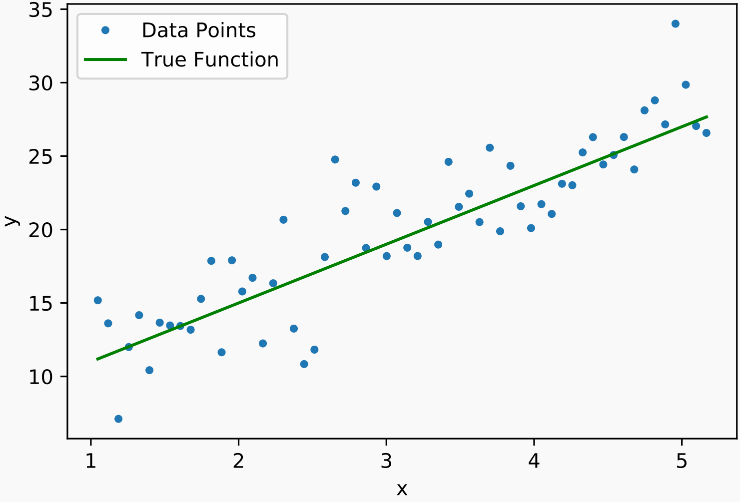


Figure 1: $y = 4x + 7$

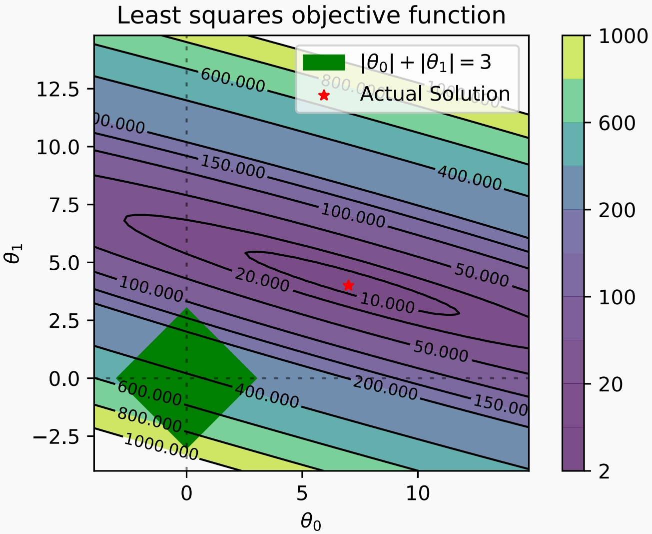
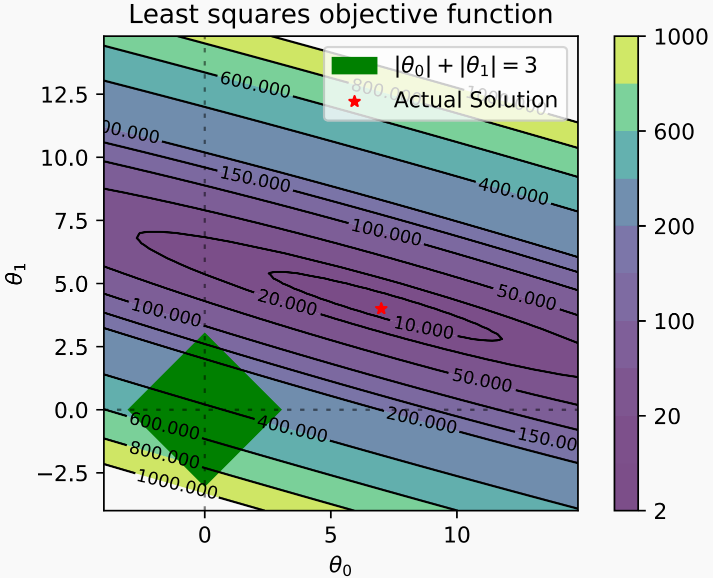


Figure 2: Lasso regression

Effect of μ - Regularization of Parameters



Effect of μ - Regularization of Parameters

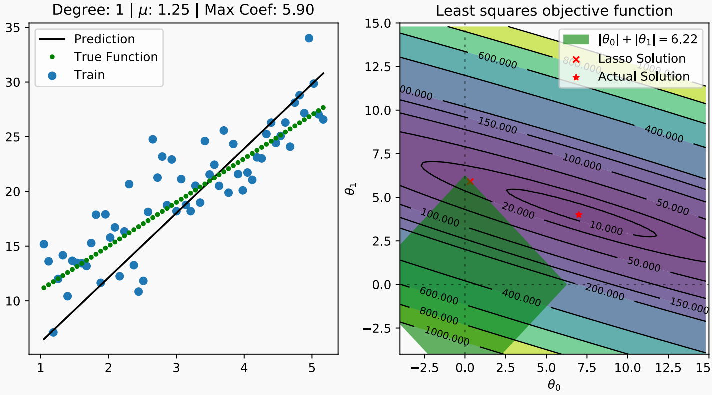


Figure 4: $\mu = 1.25$
(on the *Sample Dataset*)

Effect of μ - Regularization of Parameters

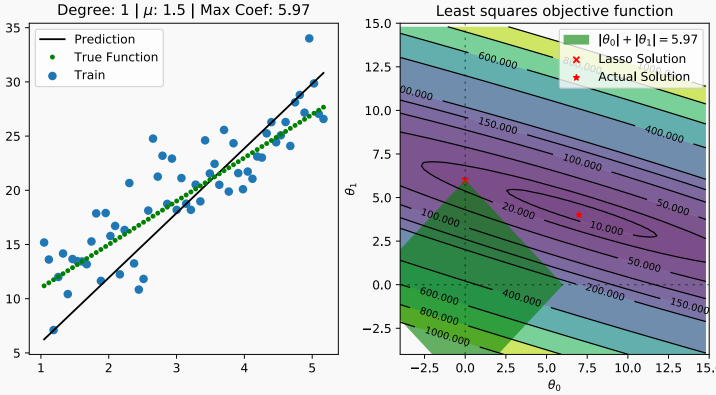


Figure 5: $\mu = 1.5$
(on the *Sample Dataset*)

Effect of μ - Regularization of Parameters

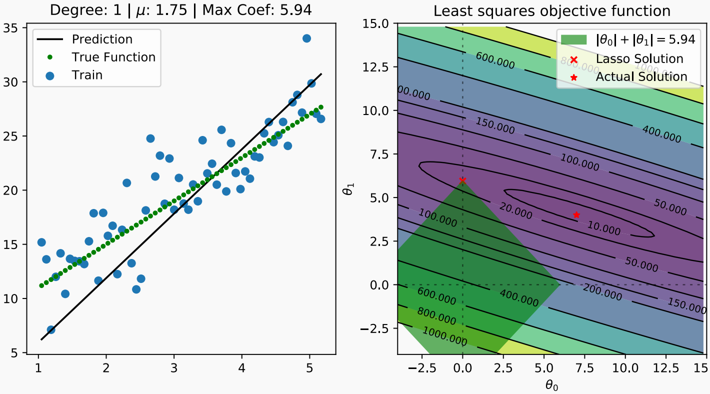


Figure 6: $\mu = 1.75$
(on the *Sample Dataset*)

Effect of μ - Regularization of Parameters

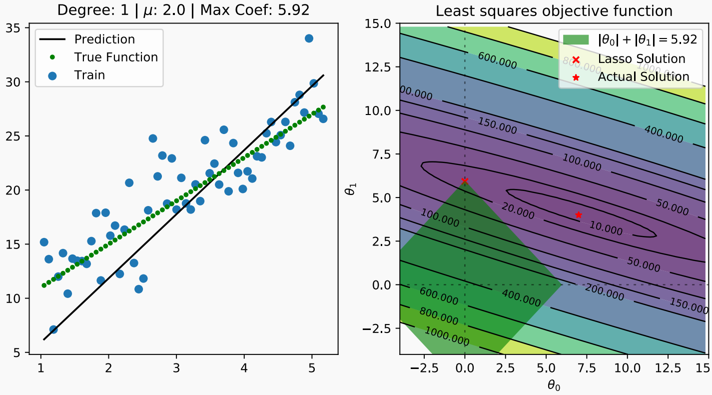
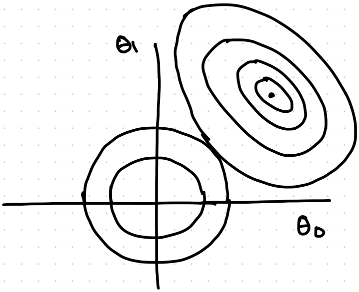


Figure 7: $\mu = 2.0$
(on the *Sample Dataset*)

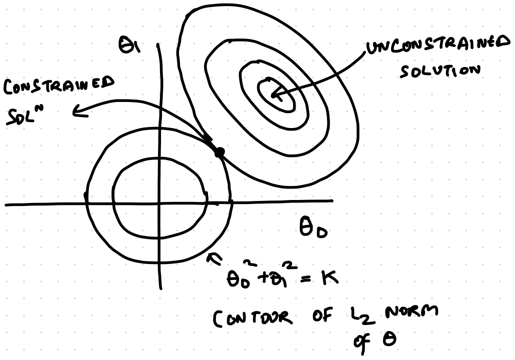
WHY LASSO GIVES SPARSITY

1. GEOMETRIC INTERPRETATION

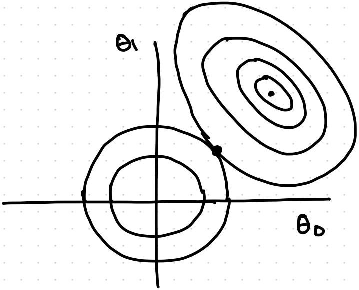
2. G.D. BASED INTERPRETATION



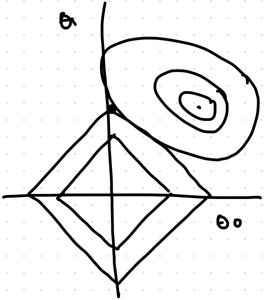
θ1
θ0



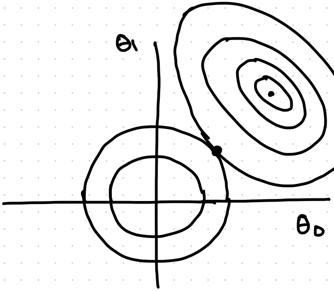
θ1
UNCONSTRAINED
SOLUTION
CONSTRAINED
SOLⁿ
θ0
θ0² + θ1² = K
CONTOUR OF L2 NORM
OF θ



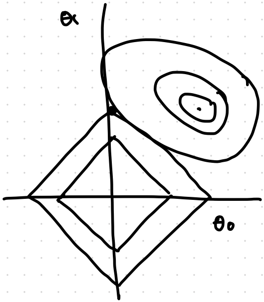
L_2 NORM



L_1 NORM

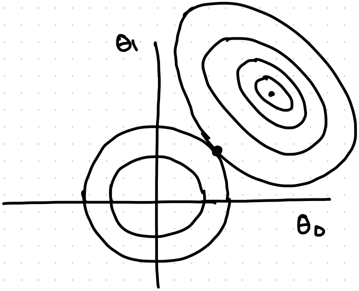


L_2 NORM

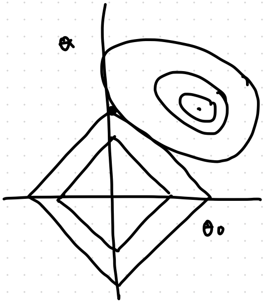


L_1 NORM

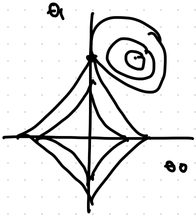
L_p NORM
($0 < p < 1$)



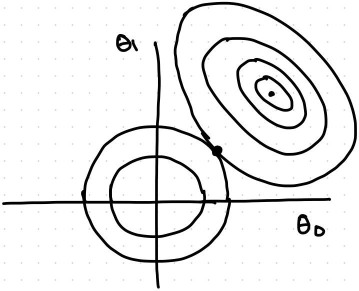
L_2 NORM



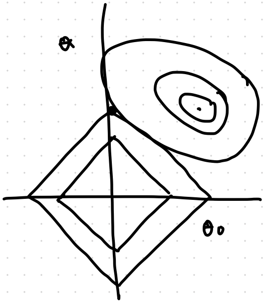
L_1 NORM



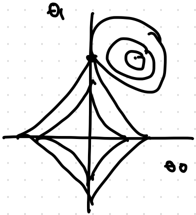
L_p NORM
$(0 < p < 1)$



L_2 NORM



L_1 NORM



L_p NORM
($0 < p < 1$)

SPARSITY →

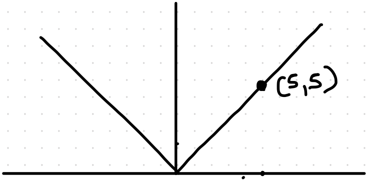
PROB. OF INTERSECTING AXIS →

DIFFICULTY OF SOLVING →

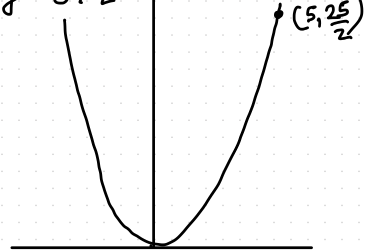
$y = |\theta|$ (FOR NOW ASSUME $\theta > 0$)

$y = \theta^2/2$

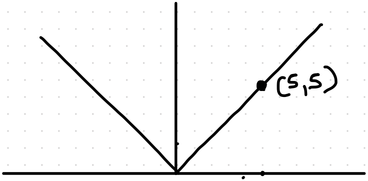
$y = |\theta|$ (FOR NOW ASSUME $\theta > 0$)



$y = \theta^2/2$

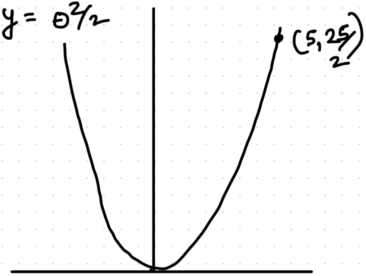


$y = |\theta|$ (FOR NOW ASSUME $\theta > 0$)



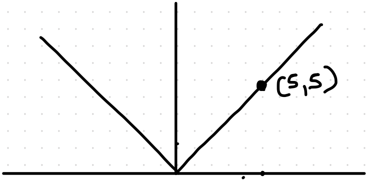
$\frac{\partial y}{\partial \theta} = 1$ (Assume $\theta > 0$)

$y = \theta^2/2$



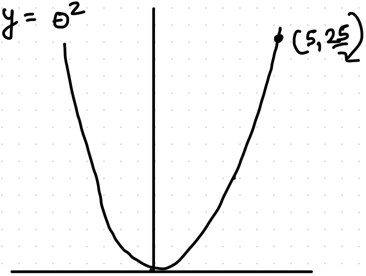
$\frac{\partial y}{\partial \theta} = \frac{2\theta}{2} = \theta$

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



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

$y = \theta^2$



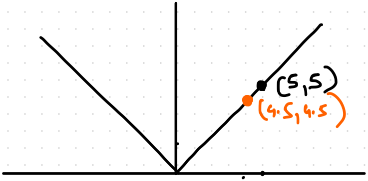
$\frac{\partial y}{\partial \theta} = \frac{2\theta}{2} = \theta$

LET $\alpha = 0.5$

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

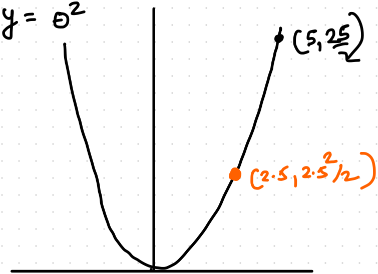
$\theta_0^1 = \theta_0^0 - 0.5 * 5 = 2.5$

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



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

$y = \theta^2$



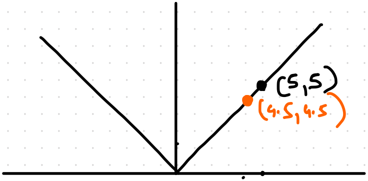
$$\frac{\partial y}{\partial \theta} = \frac{2\theta}{2} = \theta$$

LET $\alpha = 0.5$

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

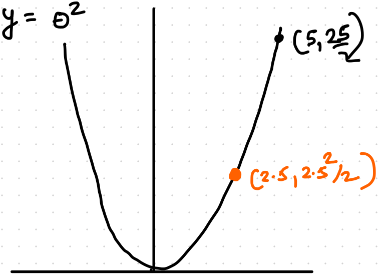
$$\theta_0^1 = \theta_0^0 - 0.5 * 5 = 2.5$$

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



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

$y = \theta^2$



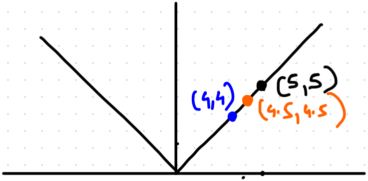
$$\frac{\partial y}{\partial \theta} = \frac{2\theta}{2} = \theta$$

LET $\alpha = 0.5$

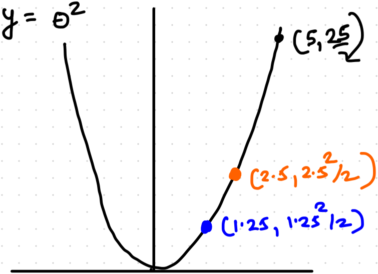
$$\theta_0^1 = \theta_0^0 - 0.5 \times 1 = 4.5$$
$$\theta_0^2 = \theta_0^1 - 0.5 \times 1 = 4.0$$

$$\theta_0^1 = \theta_0^0 - 0.5 * 5 = 2.5$$
$$\theta_0^2 = \theta_0^1 - 0.5 \times 2.5 = 1.25$$

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



$y = \theta^2$



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

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

LET $\alpha = 0.5$

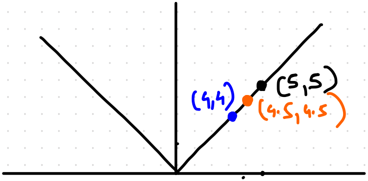
$\theta_0^1 = \theta_0^0 - 0.5 \times 1 = 4.5$

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

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

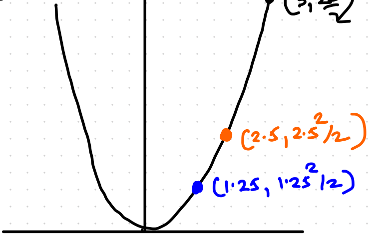
$\theta_0^2 = \theta_0^1 - 0.5 \times 2.5 = 1.25$

$y = |\theta|$ (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 $\alpha = 0.5$

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

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

$\vdots$

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

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

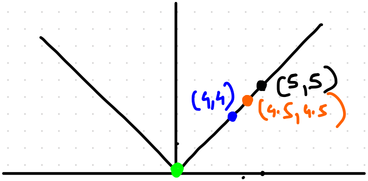
$\theta_0^2 = \theta_0^1 - 0.5 \times 2.5 = 1.25$

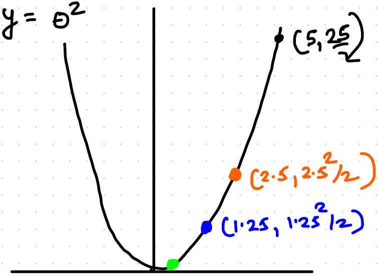
$\vdots$

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

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

$y = \theta^2$





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

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

LET $\alpha = 0.5$

$\theta_0^{10} = 0$

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

$= 0.0048$

(Approaching 0 but not exactly zero)

Regularization path of lasso regression

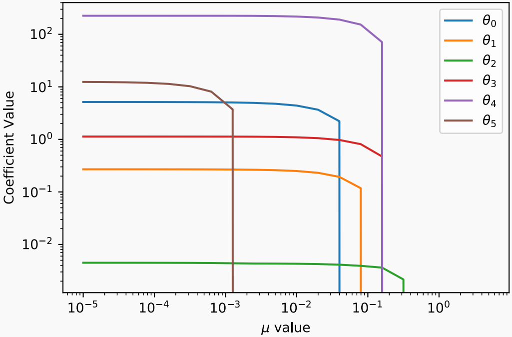


Figure 8: Regularization path of θ_i

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

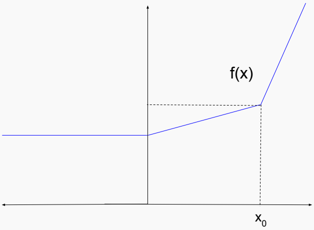
- LASSO inherently does feature selection!
- Sets coefficients of “less important” features to zero.
- Sparse and memory efficient and often more interpretable models.

Subgradient

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

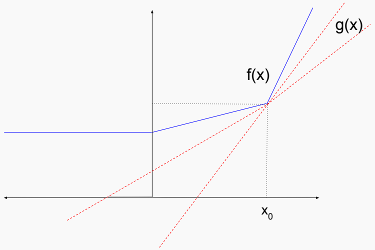
Task at hand

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



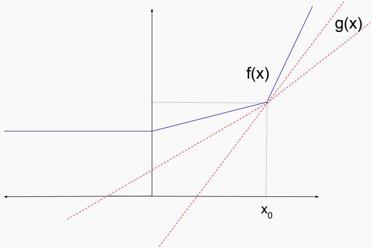
Solution

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



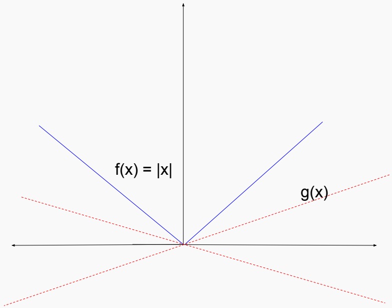
Solution

- 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

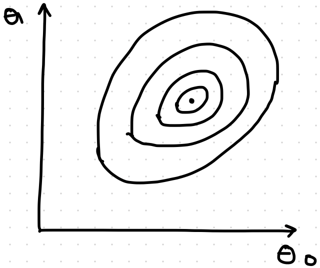
Coordinate Descent

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

Coordinate Descent

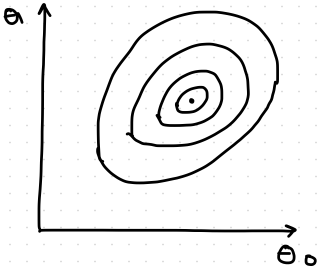
- Another optimisation method (akin to gradient descent)
- Objective: $\text{Min}_{\theta} f(\theta)$
- Key idea: Sometimes difficult to find minimum for all coordinates
- ..., but, easy for each coordinate
- turns into a 1D optimisation problem

COORDINATE DESCENT ALGORITHM



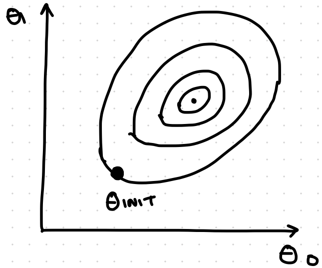
COORDINATE DESCENT ALGORITHM

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



COORDINATE DESCENT ALGORITHM

1) INIT θ

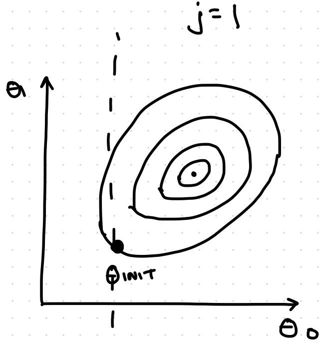


COORDINATE DESCENT ALGORITHM

1) INIT θ

2) WHILE NOT CONVERGED

   2.1) PICK COORDINATE 'j'



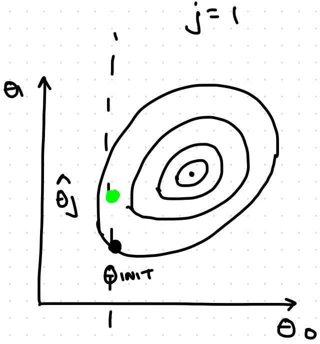
COORDINATE DESCENT ALGORITHM

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

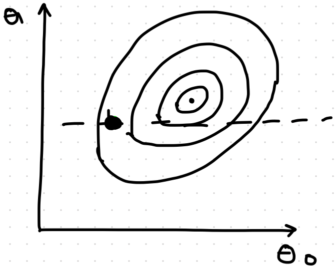
1) INIT θ

2) WHILE NOT CONVERGED

✓ 2.1) PICK COORDINATE 'j'

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

$j = 0$



COORDINATE DESCENT ALGORITHM

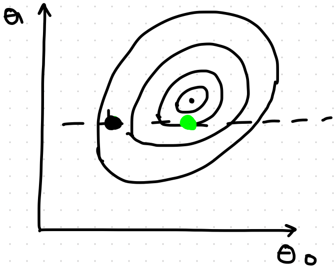
1) INIT θ

2) WHILE NOT CONVERGED

   2.1) PICK COORDINATE 'j'

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

$j = 0$



COORDINATE DESCENT ALGORITHM

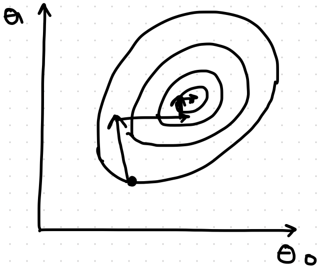
1) INIT θ

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

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

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

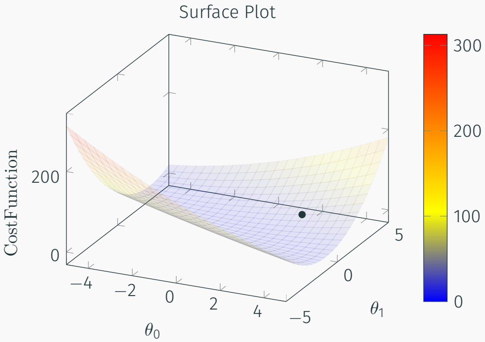
$\epsilon_1 = 1 - \theta_0 - \theta_1$

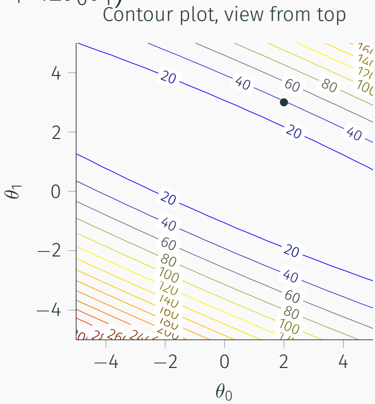
$\epsilon_2 = 2 - \theta_0 - 2\theta_1$

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

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

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





Coordinate Descent : Example

Iteration 1

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

$\theta_1 = 3$ optimize for θ_0

Iteration 1

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

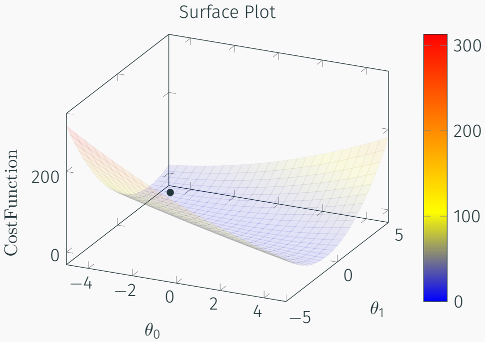
$\theta_1 = 3$ optimize for θ_0

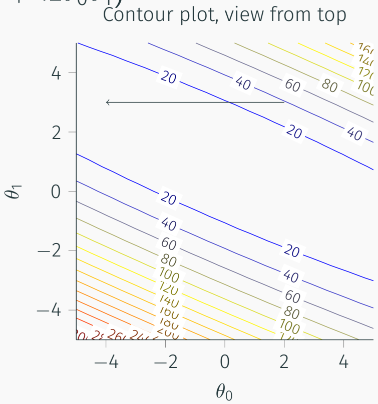
$\frac{\partial MSE}{\partial \theta_0} = 6\theta_0 + 24 = 0$

$\theta_0 = -4$

Iteration 1

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





Iteration 2

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

$\theta_0 = -4$ optimize for θ_1

Iteration 2

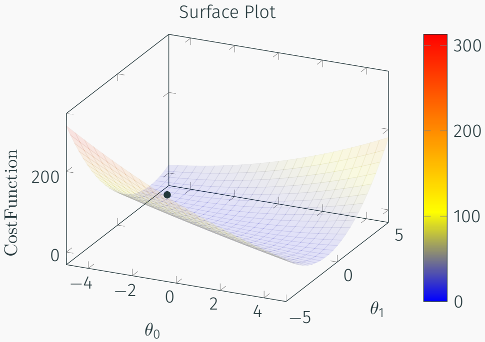
INIT: $\theta_0 = -4$ and $\theta_1 = 3$

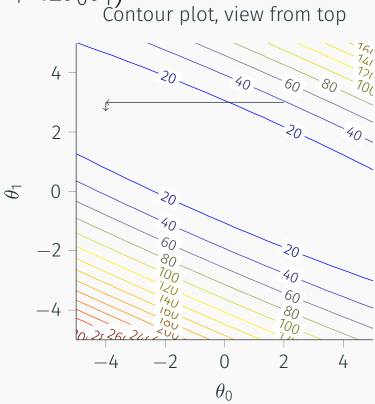
$\theta_0 = -4$ optimize for θ_1

$\theta_1 = 2.7$

Iteration 2

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





Iteration 3

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

$\theta_1 = 2.7$ optimize for θ_0

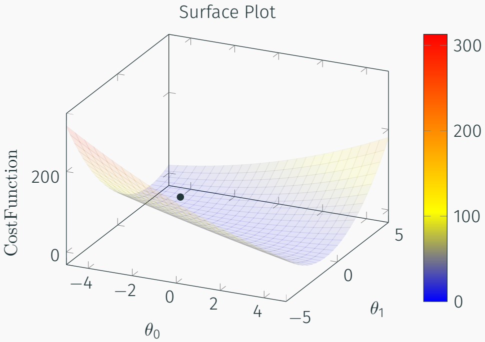
Iteration 3

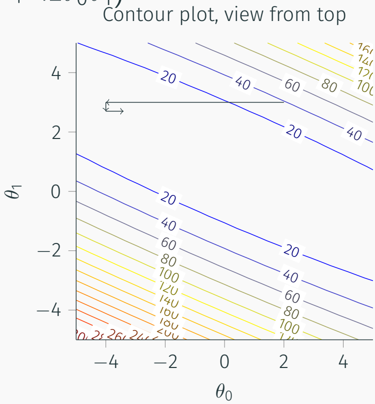
INIT: $\theta_0 = -4$ and $\theta_1 = 2.7$

$\theta_1 = 2.7$ optimize for θ_0

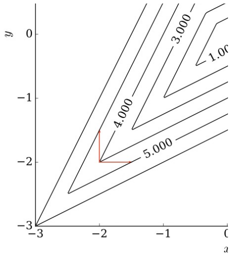
$\theta_0 = -3.4$

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



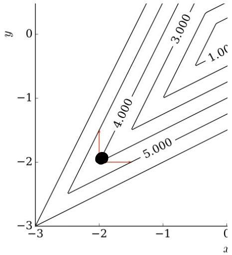


FAILURE OF COORDINATE DESCENT

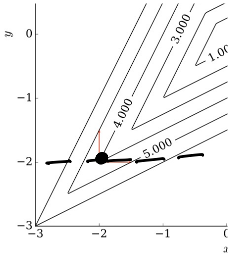


FAILURE OF COORDINATE DESCENT

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



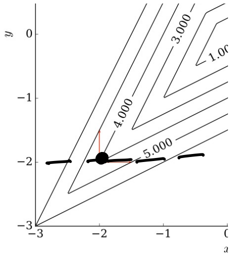
FAILURE OF COORDINATE DESCENT



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

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

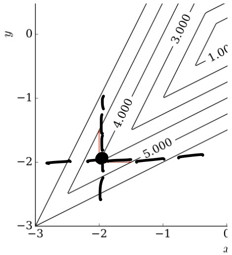
FAILURE OF COORDINATE DESCENT



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

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

OBJECTIVE INCREASES IN BOTH DIRECTIONS



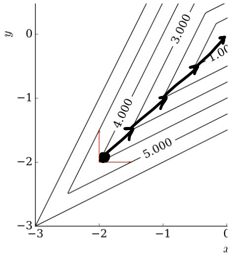
FAILURE OF COORDINATE DESCENT

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