Extra slides regularisation in linear regression

- Regularisation is explained in full on day 2 when we discuss logistic regression.
- These slides explain the principles and apply them to linear regression. They're extra material!

Regularisation

- Change the cost function to apply a cost for complexity

Regularisation

- Change the cost function to apply a cost for complexity

$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x) - y^{(i)}\right)^2$$

Regularisation: ridge regression

- Change the cost function to apply a cost for complexity

$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x) - y^{(i)})^2$$

$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} (\theta_j)^2$$

- Make J a function of both the error of predictions given some parameters *and* the magnitude of those parameters themselves

Regularisation: ridge regression

- Change the cost function to apply a cost for complexity

$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x) - y^{(i)}\right)^2$$

$$\left.J(\theta_0, \ldots, \theta_n) = \underbrace{\frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x) - y^{(i)}\right)^2}_{} + \lambda \underbrace{\sum_{j=1}^{n} (\theta_j)^2}_{}\right\}$$

By convention: don't shrink bias/intercept term

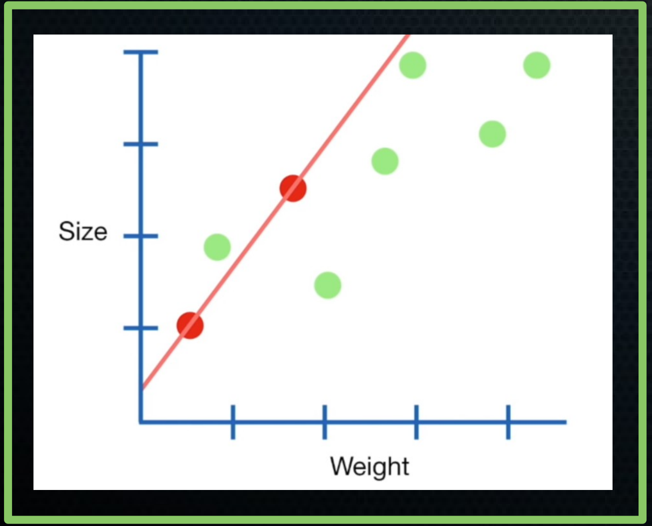
- Make J a function of both the error of predictions given some parameters *and* the magnitude of those parameters themselves

Regularisation: ridge regression

- Change the cost function to apply a cost for complexity

$$J(\theta_0, \ldots, \theta_n) = \boxed{\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x) - y^{(i)}\right)^2} + \lambda\sum_{j=1}^{n}(\theta_j)^2 \qquad h_\theta(x) = \theta_0 + \theta_1 x$$

- Add some **bias** (constrain hypothesis to a set with small parameter values) but reduces **variance**:



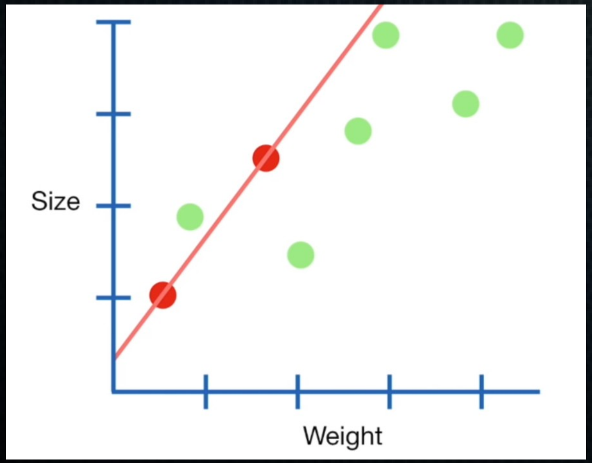
Source: StatQuest. “Regularization Part 1: Ridge (L2) Regression” YouTube, Joshua Starmer, 24 Sep. 2018,https://www.youtube.com/watch?v=Q81RR3yKn30

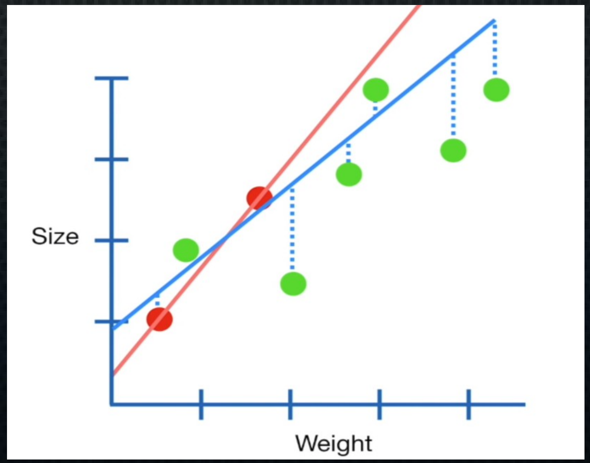
Regularisation: ridge regression

- Change the cost function to apply a cost for complexity

$$J(\theta_0, \ldots, \theta_n) = \boxed{\frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} (\theta_j)^2} \qquad h_\theta(x) = \theta_0 + \theta_1 x$$

- Add some **bias** (constrain hypothesis to a set with small parameter values) but reduces **variance**:





Constrained how much the line may increase with Weight (biased) → generalises better to test set

Source: StatQuest. “Regularization Part 1: Ridge (L2) Regression” YouTube, Joshua Starmer, 24 Sep. 2018,https://www.youtube.com/watch?v=Q81RR3yKn30

Regularisation: LASSO regression

- Same idea, slightly different execution:

$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} (\theta_j)^2$$

Regularisation: LASSO regression

- Same idea, slightly different execution:

$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \cancel{(\theta_j)^2}$$

Regularisation: LASSO regression

- Same idea, slightly different execution:

$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |\theta_j|$$

Regularisation: LASSO regression

- Same idea, slightly different execution:

$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} |\theta_j|$$

- Ridge regression *shrinks* parameters/weights to 0, LASSO can make them 0 outright, i.e. simply removes uninformative features.

Regularisation: LASSO regression

- Same idea, slightly different execution:

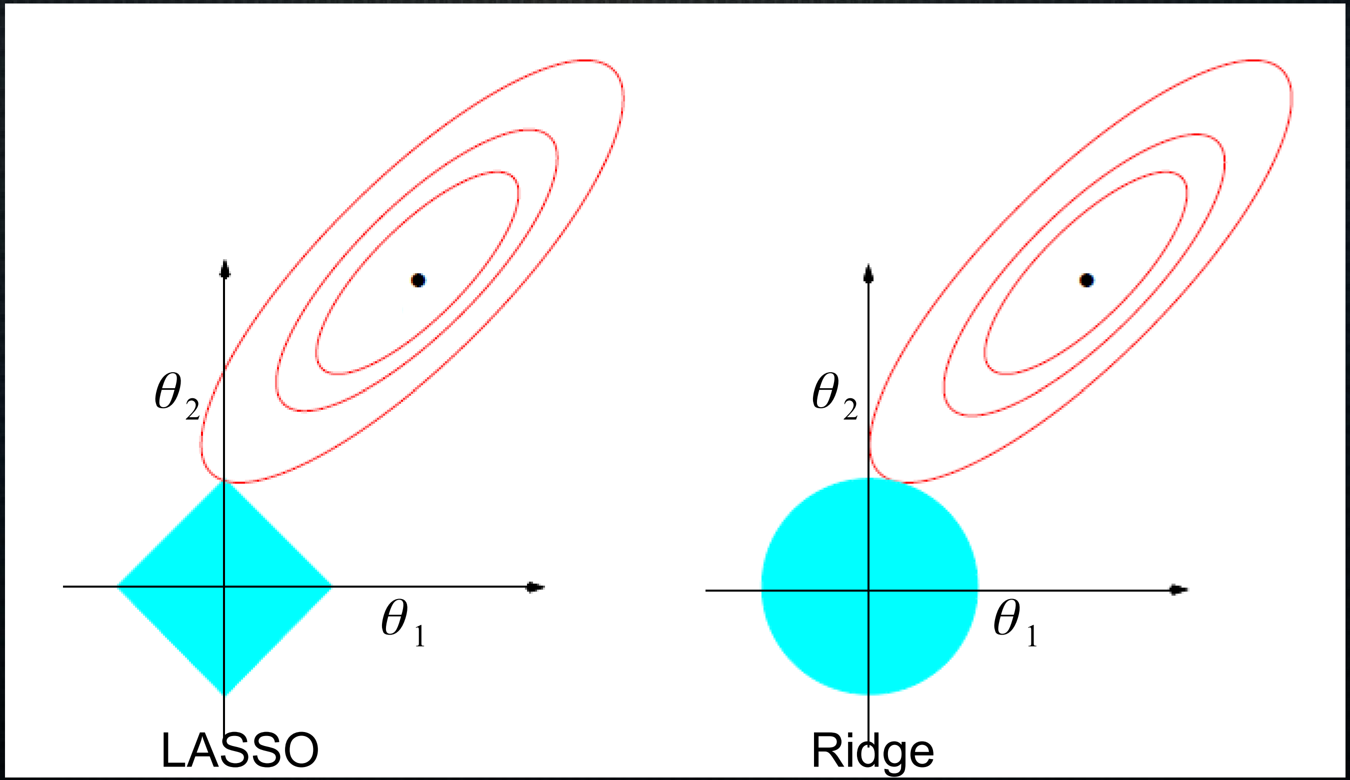
$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |\theta_j|$$

- Ridge regression *shrinks* parameters/weights to 0, LASSO can make them 0 outright, i.e. simply removes uninformative features. → Why?

Regularisation: Ridge vs. LASSO regression

- Ridge regression *shrinks* parameters/weights to 0, LASSO can make them 0 outright, i.e. simply removes uninformative features. → Why?

$$\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x)-y^{(i)}\right)^2$$

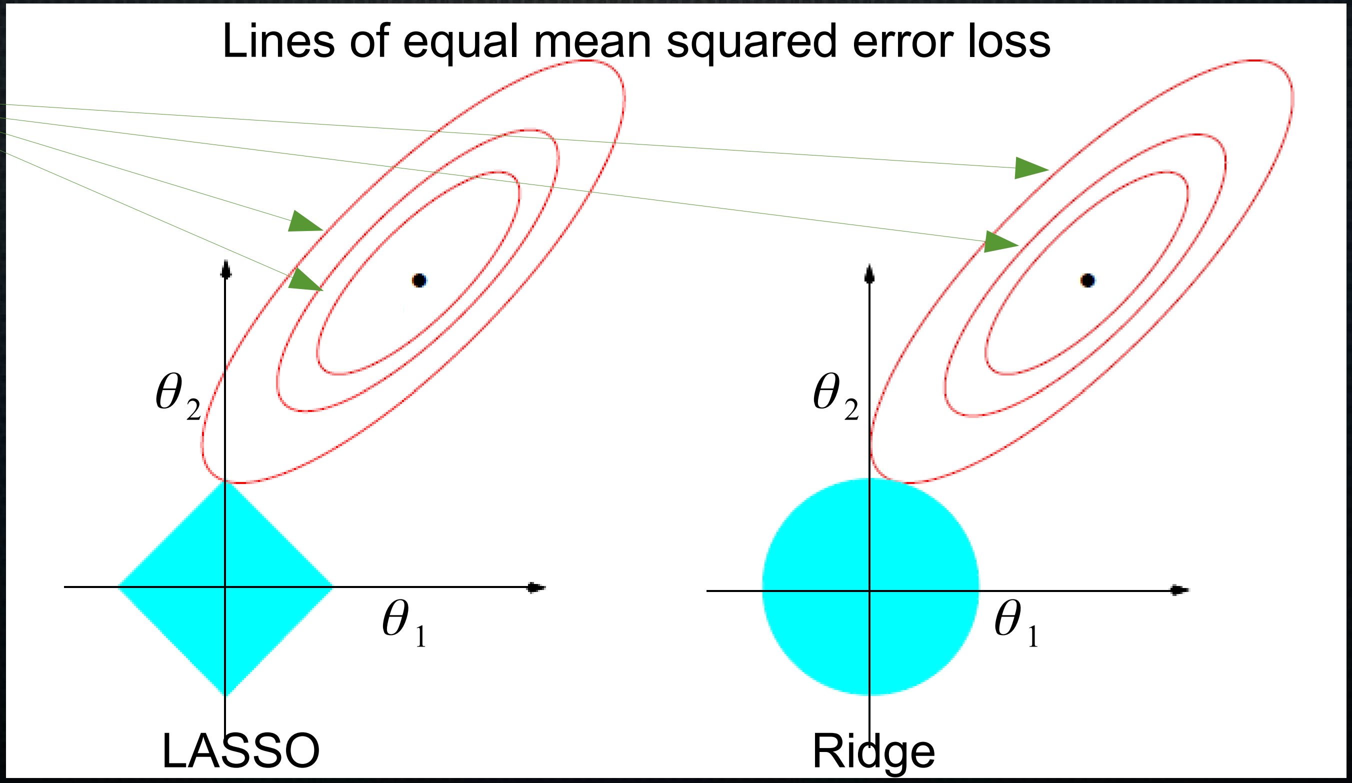


Source: Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Regularisation: Ridge vs. LASSO regression

- Ridge regression *shrinks* parameters/weights to 0, LASSO can make them 0 outright, i.e. simply removes uninformative features. → Why?

$$\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x)-y^{(i)}\right)^2$$

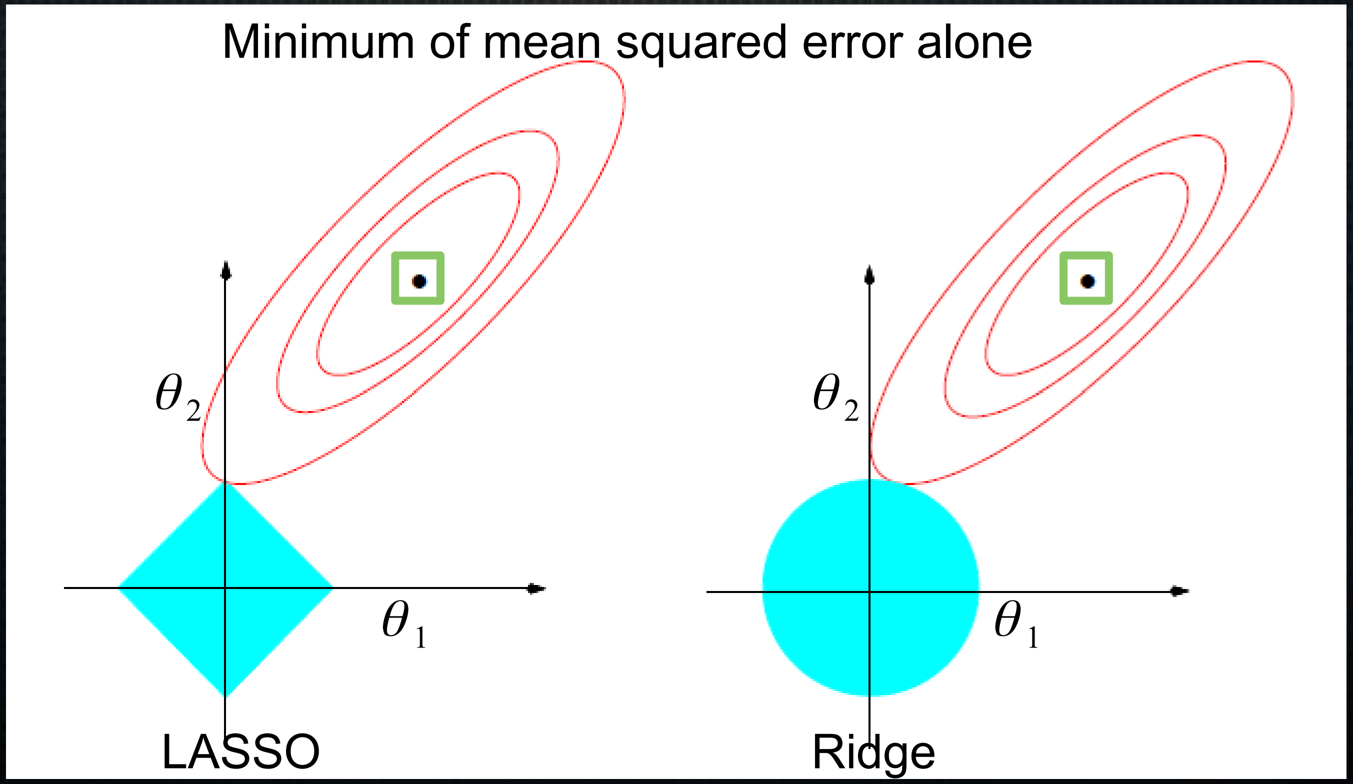


Source: Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Regularisation: Ridge vs. LASSO regression

- Ridge regression *shrinks* parameters/weights to 0, LASSO can make them 0 outright, i.e. simply removes uninformative features. → Why?

$$\boxed{\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x)-y^{(i)}\right)^2}$$



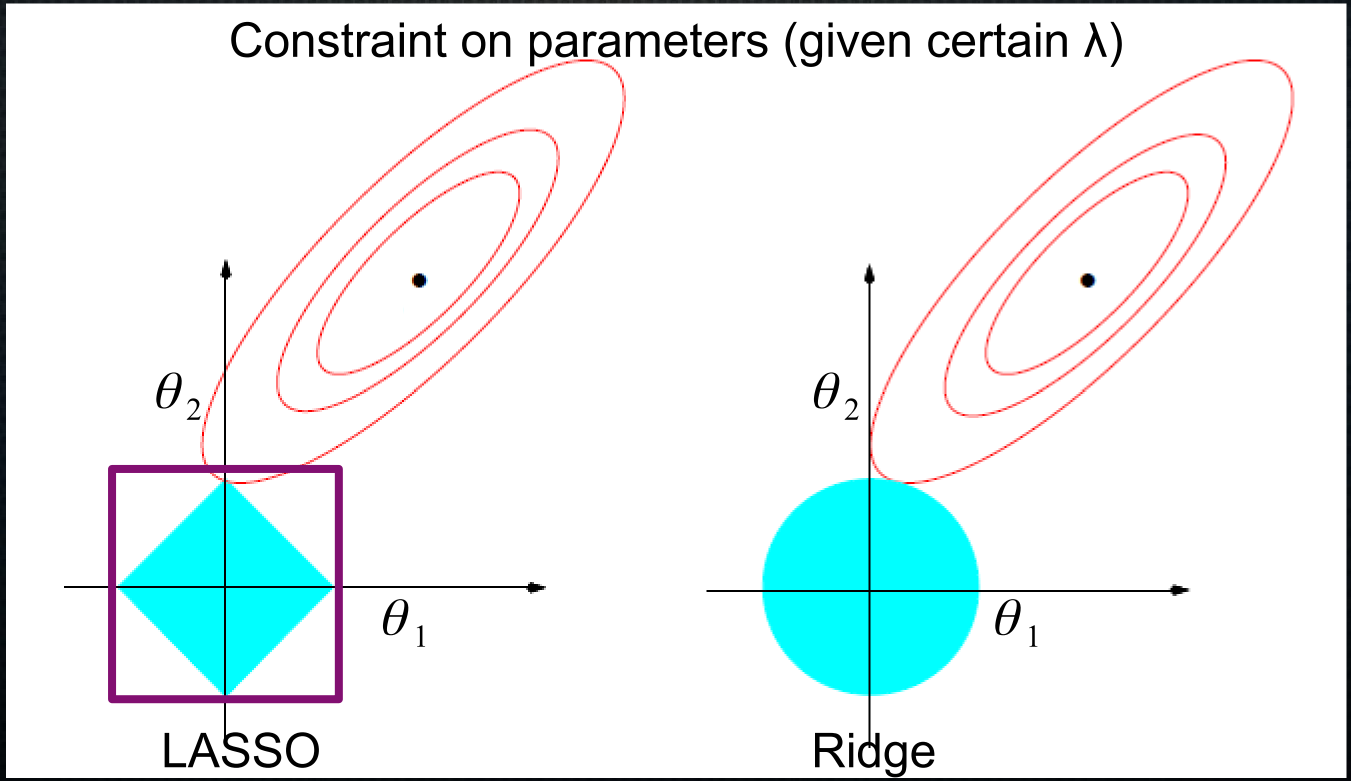
Source: Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Regularisation: Ridge vs. LASSO regression

- Ridge regression *shrinks* parameters/weights to 0, LASSO can make them 0 outright, i.e. simply removes uninformative features. → Why?

$$\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x)-y^{(i)}\right)^2$$

$$\boxed{\lambda\sum_{j=1}^{n}|\theta_j|}$$



Source: Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

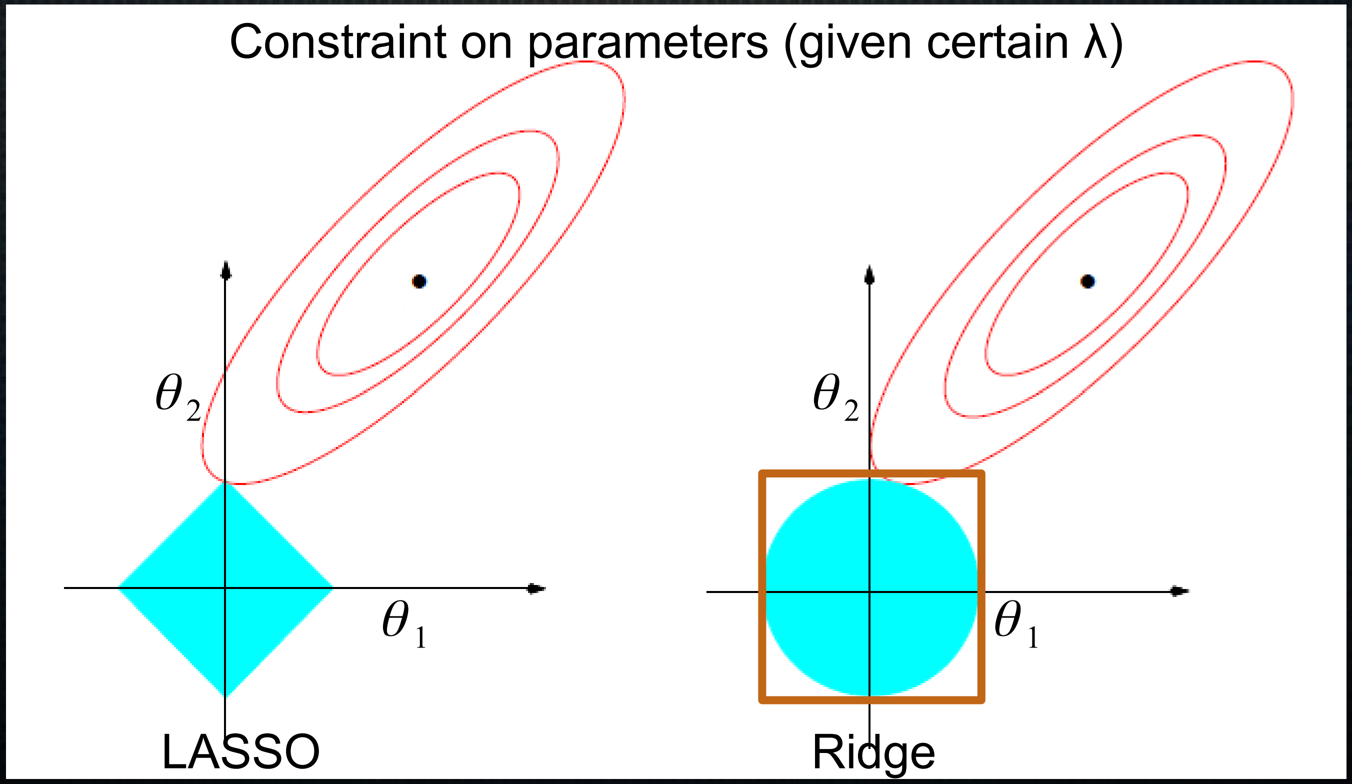
Regularisation: Ridge vs. LASSO regression

- Ridge regression *shrinks* parameters/weights to 0, LASSO can make them 0 outright, i.e. simply removes uninformative features. → Why?

$$\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x)-y^{(i)}\right)^2$$

$$\lambda\sum_{j=1}^{n}|\theta_j|$$

$$\boxed{\lambda\sum_{j=1}^{n}(\theta_j)^2}$$



Source: Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

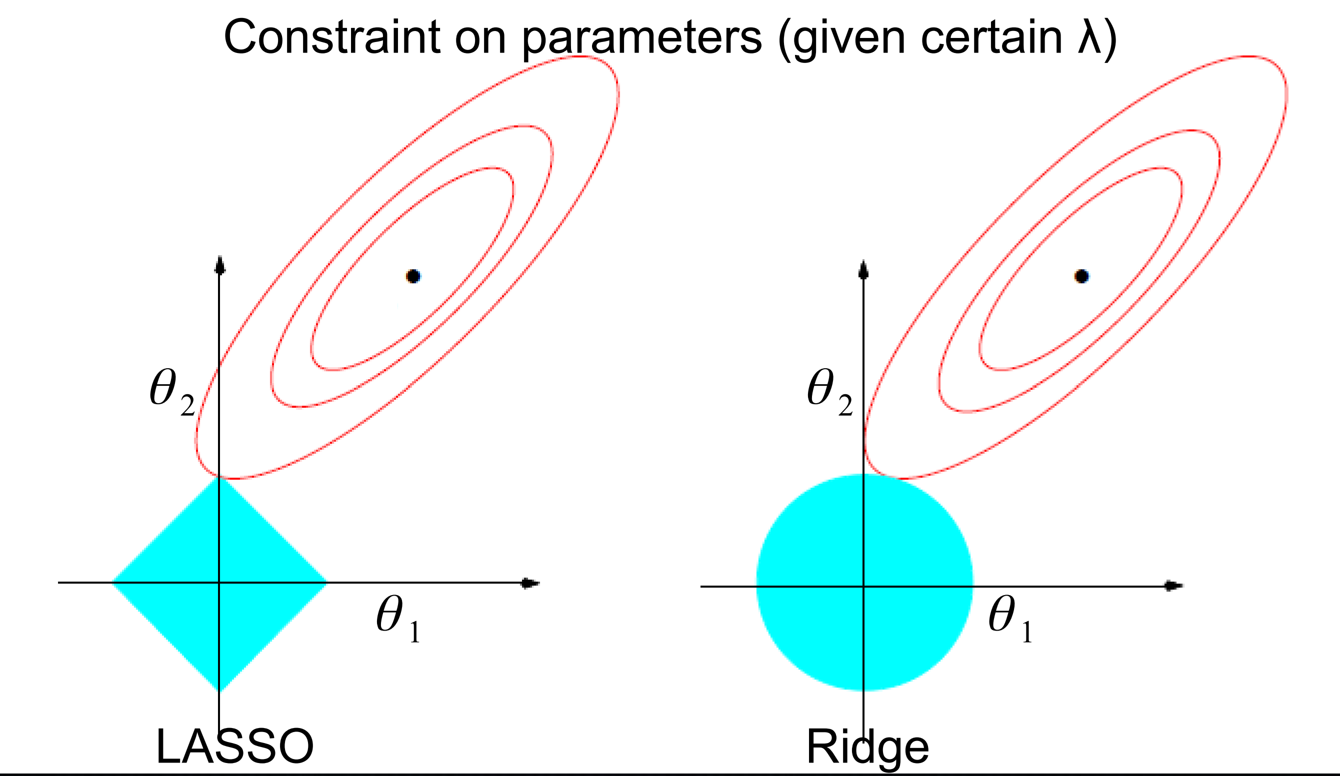
Regularisation: Ridge vs. LASSO regression

- Optimum = best least squares fit given constraint = intersect red lines with blue area.

$$\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x)-y^{(i)}\right)^2$$

$$\lambda\sum_{j=1}^{n}|\theta_j|$$

$$\lambda\sum_{j=1}^{n}(\theta_j)^2$$



Source: Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Regularisation: Ridge vs. LASSO regression

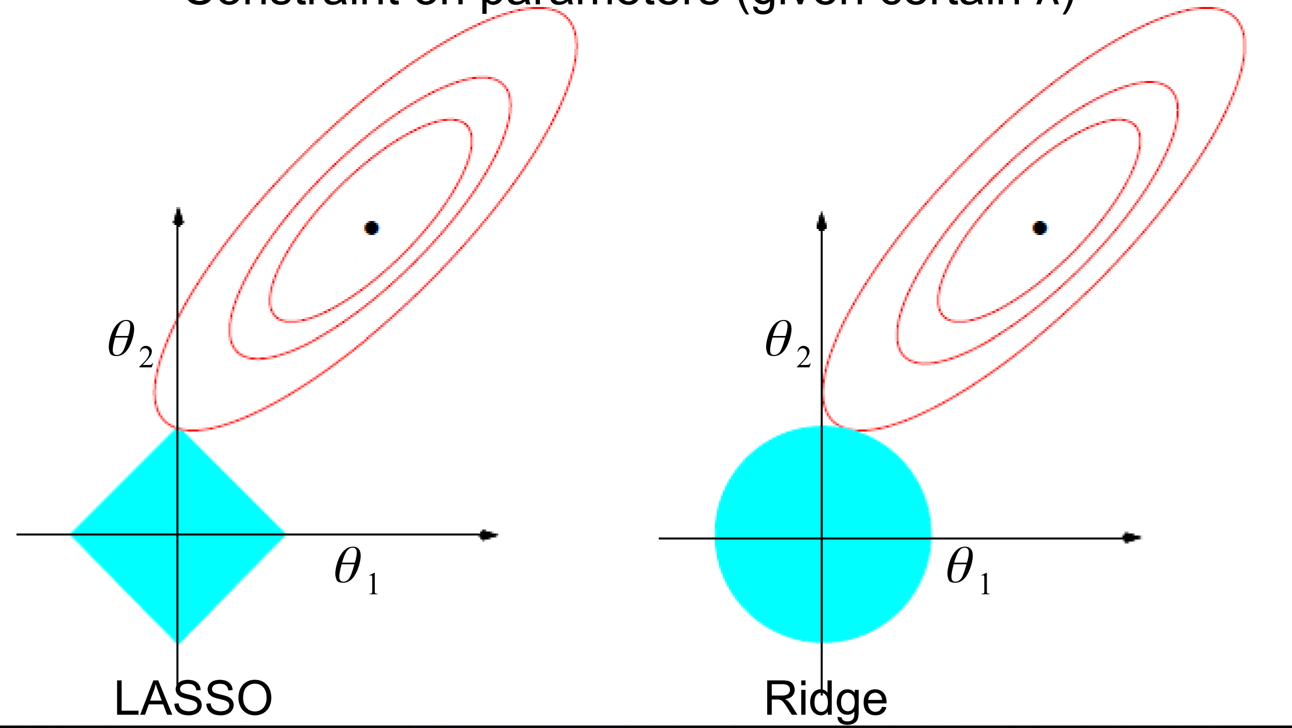
- Optimum = best least squares fit given constraint = intersect red lines with blue area.
- LASSO: can be at tip of rhombus, where theta1 = 0.
- Ridge: intersection ellips with circle never where either = 0

$$\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x)-y^{(i)}\right)^2$$

$$\lambda\sum_{j=1}^{n}|\theta_j|$$

$$\lambda\sum_{j=1}^{n}(\theta_j)^2$$

Constraint on parameters (given certain λ)



Source: Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

What about λ?

$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |\theta_j|$$

- Lambda is another hyperparameter
- Intuition: higher lambda → constrain parameters more, i.e. increase **bias**.
  lower lambda → constrain parameters less, i.e. increase **variance**

What about λ?

$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |\theta_j|$$

- How to pick a good value?

What about λ?

$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |\theta_j|$$

- How to pick a good value?
- Nested cross-validation!
- Will be explained later what this means.