REGULARISATION

Dr. Brian Mc Ginley

REGULARISATION

- Another term that applies in general to ”most” Machine Learning systems is the idea of Regularisation.
- Definition:
  - Regularisation is a technique used in machine learning models to reduce overfitting, improve model generalization, and manage model complexity. It achieves this by adding an additional penalty term in the error function. The additional term controls for excessively fluctuating model parameters so that the coefficients don’t take extreme values.
- This technique of keeping a check or reducing the value of coefficients are called shrinkage methods or weight decay in case of neural networks.

REGULARISATION

- Some Points:
  - Regularisation is a more general way of avoiding overfitting.
  - Overfitting means your model has high-variance
    - Variance is the amount your model will change if you change the training data
    - More parameters and flexibility cause high-variance
  - Regularisation works be adding a term to the loss function which adds loss if the model is not of a preferred type.
    - In particular, we don’t want any weights to become very large.
    - A large weight can change by a large amount very quickly (very sensitive to small changes in the associated feature).
    - As differences in the magnitude of weights can quickly spiral out of control and individual weights influence can quickly exceed their importance.
    - For this reason, we add a penalty to the loss function that is related to the size of the weights.

MEAN SQUARED ERROR

$$MSE(\boldsymbol{w}) = \frac{1}{2m}\sum_{i=1}^{m}(y^i - f_{\boldsymbol{w}}(\boldsymbol{x}^i))^2$$

- It can happen that when optimising this Loss function, some of the weights may become very large.
- That means the features corresponding to these large weights will increase in their importance
- This can cause overfitting.
- So, what can we do with these large weights?

L1 AND L2 NORM

MEAN SQUARED ERROR - L2 (RIDGE REGRESSION)

- We add a regularisation term (λ) to our Loss function that penalises large weights

$$L(\boldsymbol{w}) = MSE(\boldsymbol{w}) + \lambda \sum_{j=1}^{m} w_j^2$$

- This means that if any particular w_j is large, the overall Loss will be large too.
- This is known as the L2 norm.
- It does not matter if we are using MSE or the logistic regression loss function, we can add on the regularisation term to any loss metric.
- Here is the MSE with regularisation fully written out:

$$L(\boldsymbol{w}) = \frac{1}{2m} \sum_{i=1}^{m} (y^i - f_{\boldsymbol{w}}(\boldsymbol{x}^i))^2 + \lambda \sum_{j=1}^{n} w_j^2$$

- where n is the number of features and m is sample size. Notice we do not include the bias term (w_0)

L2 REGULARISATION

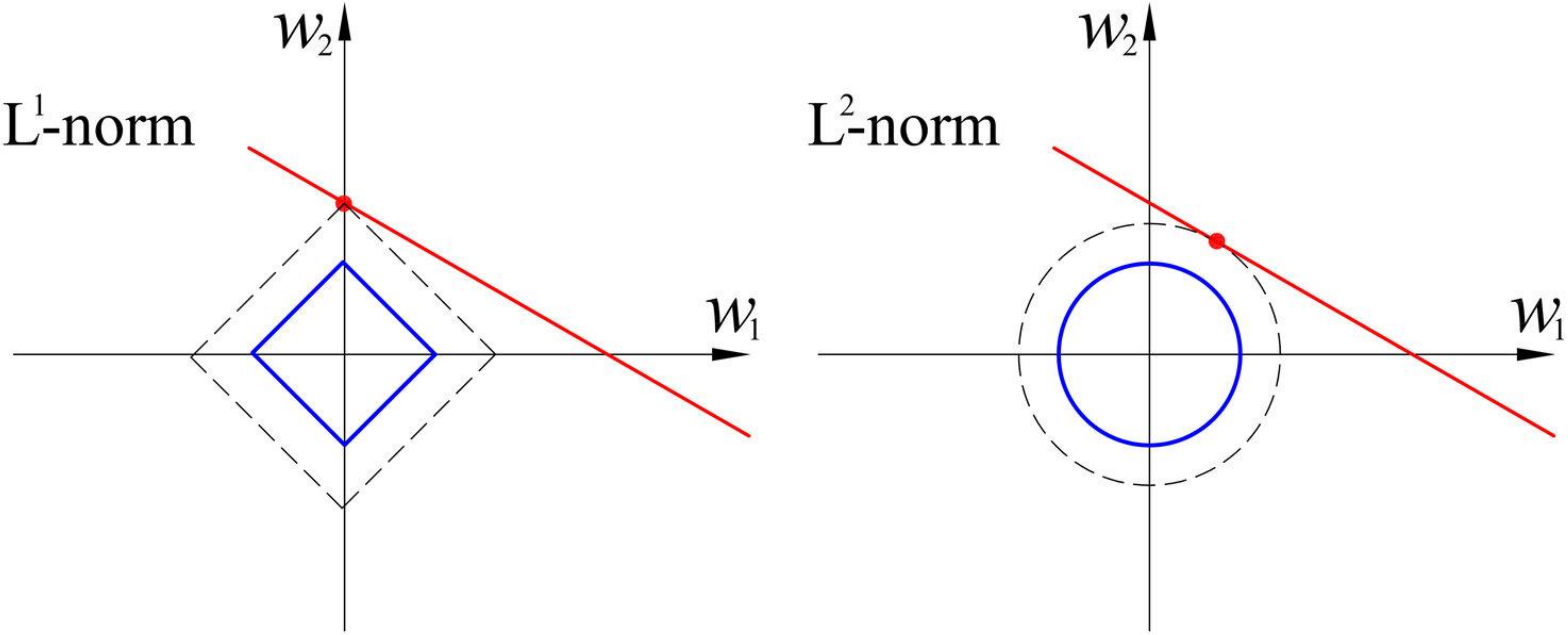
- The effect of regularisation is to make it so the model prefers to learn small weights, all other things being equal.
- This means that the trained model will be a slightly worse fit for the training data but will avoid overtraining (adds bias (penalty), reduces variance)
- Large weights will only be allowed if they considerably improve the first part of the cost function.
- The relative importance of the two elements of the compromise depends on the value of λ:
  - when λ is small, we prefer to minimize the original cost function when λ is large we prefer small weights.
  - Large λ moves in the direction of underfitting, and small λ moves in the direction of overfitting.
- So it is another hyperparameter that we have to choose - perhaps via cross-validation.
- Additionally Gradient Descent applies just as it does without regularisation, it just has an extra term when calculating the updates/gradients.

LASSO REGRESSION

- LASSO Regression is very similar to Ridge Regression except that it takes the absolute value of the coefficients/weights (rather than squaring them).
- This is known as the L1 norm (Manhattan Distance) vs. L2 Norm (Euclidean Distance)

$$L(\boldsymbol{w}) = \frac{1}{2m}\sum_{i=1}^{m}(y^i - f_{\boldsymbol{w}}(\boldsymbol{x}^i))^2 + \lambda\sum_{j=1}^{m}|w_j|$$

- The main difference is that Lasso Regression can exclude non-relevant features from model (increasing sparsity) (e.g. biomedical dataset with irrelevant biomarkers).
- Ridge regression performs better when the variables are known to be useful.


w_2
L^1-norm
w_1
w_2
L^2-norm
w_1

REGULARIZATION ADVANTAGES

- Regularization usually performs better when:
  - The data has multicollinearity (e.g. in a housing dataset): Features are highly correlated, and Regularization can prevent overfitting.
  - The dataset has a large number of features (especially when there are more features than observations) (e.g. gene expression datasets): Regularization can prevent overfitting by shrinking the coefficients and selecting the most relevant features. So, if you have limited data, Regularization helps keep model variance low.
  - The dataset has some irrelevant features. Regularization helps by forcing these coefficients towards zero
  - The dataset has noise: Regularization can reduce the model's sensitivity to noise. In contrast, Linear Regression does not have this penalty mechanism and will try to fit all features, including noisy ones (which might cause worse performance).