L1, L2 norm
L1, L2 regularization

陳冠穎
2016/11/25

1.

L1 norm & L2 norm

Vector Norms

來看一個簡單的例子：二維空間

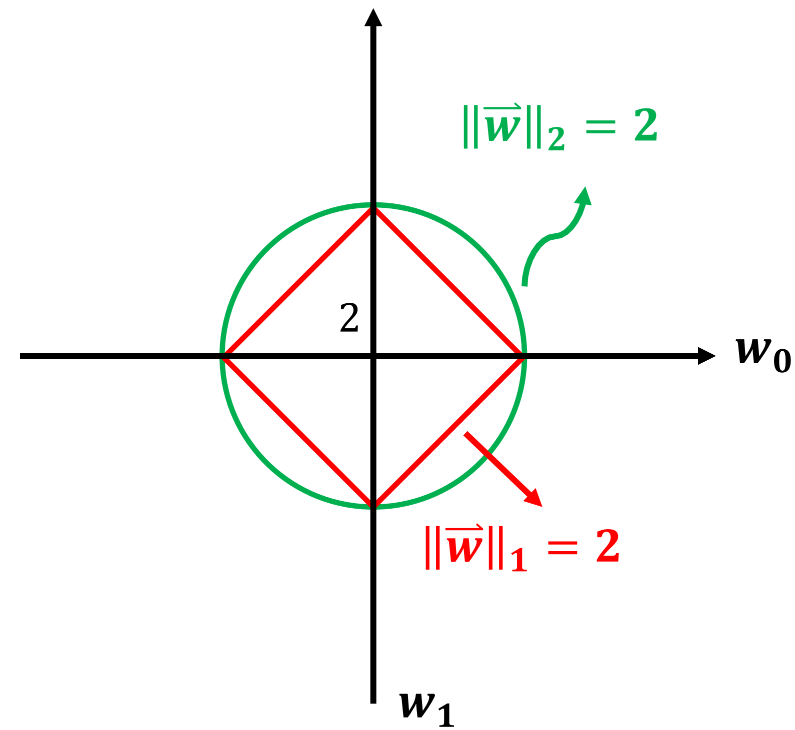
$$\vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

◎L1-Norm:

- $\|\vec{w}\|_1 = |w_0| + |w_1|$

◎L2-Norm:

- $\|\vec{w}\|_2 = \sqrt{w_0{}^2 + w^2}$



簡單來說…

L1 norm：曼哈頓距離

L2 norm：歐氏距離

但在 L2 的實作上，我們通常在比較的是距離大小，因此常常省略開平方的動作，只看平方和，以節省計算。

2.

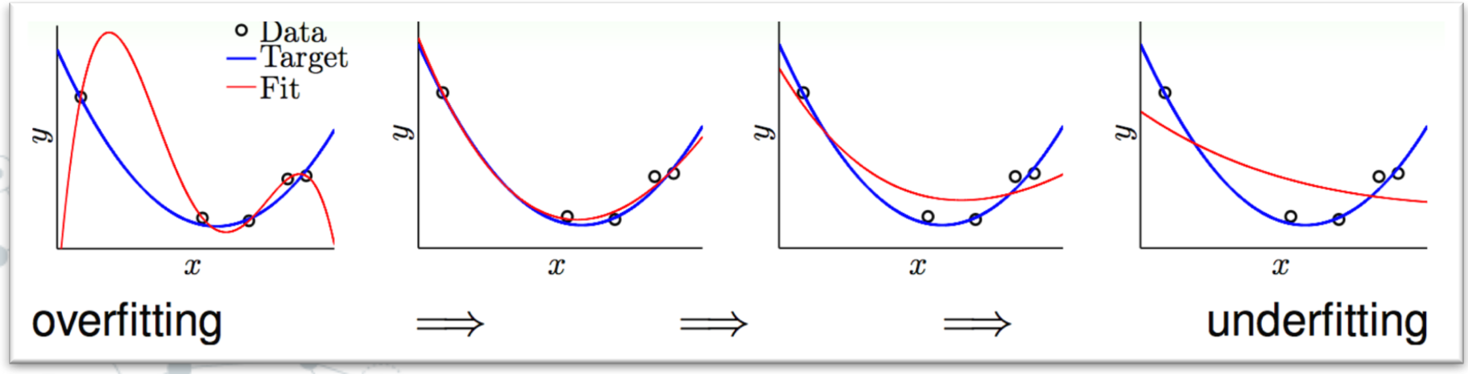
L1 regularization & L2 regularization

Regularization

Regularization 是為了防止模型過度學習、適配（Overfitting）訓練資料。

◎Mathematically speaking, it adds a *regularization term* in order to prevent the coefficients to fit so perfectly to overfit.

- 在最佳化損失函數的問題上，額外加上 *regularization term*，避免模型過於複雜而過度適配訓練資料。
- 加入 *regularization term* 能使學習出來的模型變得平滑、較簡單。



Regularization

Regularization 為何能避免模型過於複雜而過度學習？

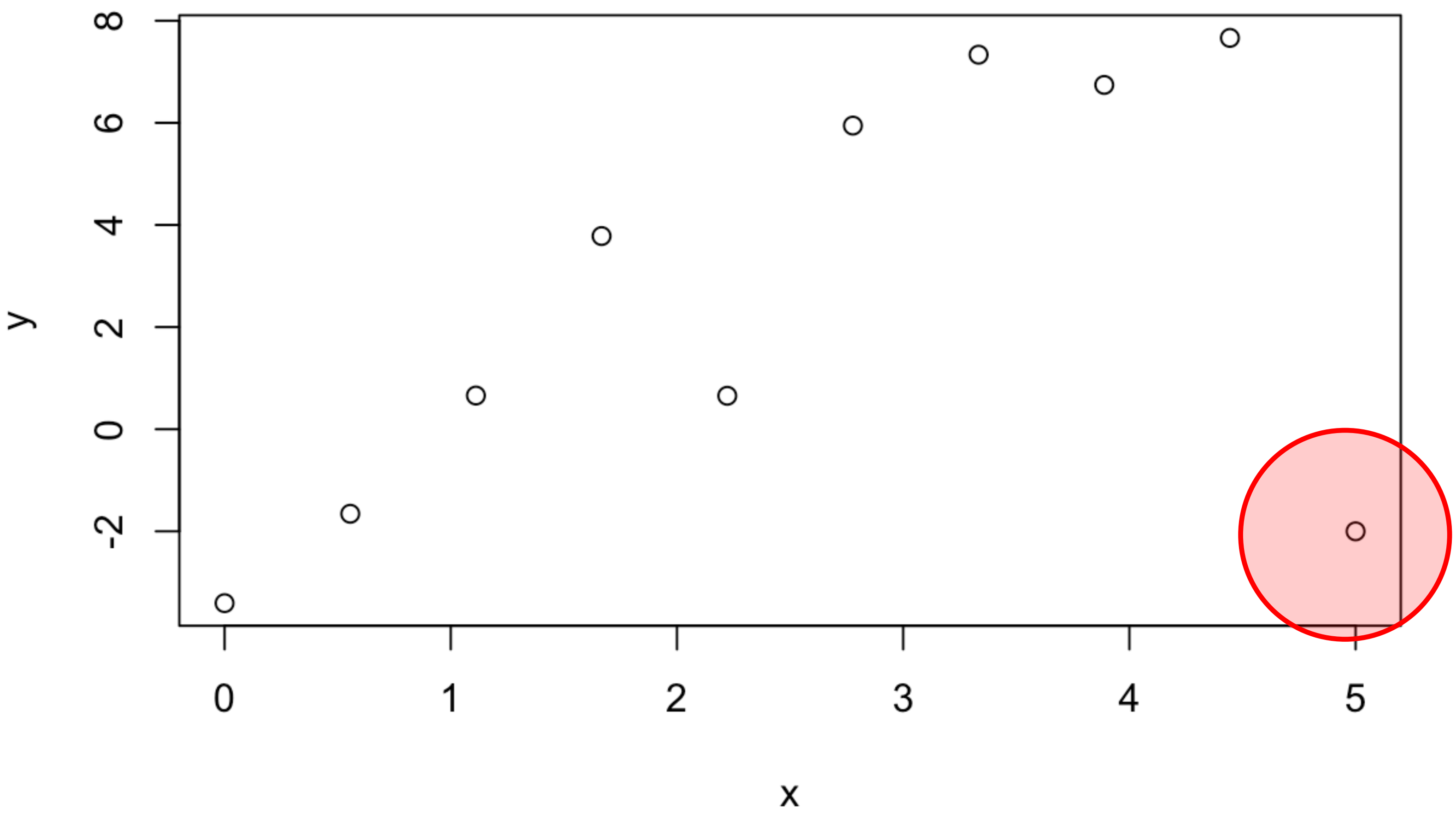
- Hypothesis w in H_{10}: $w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots + w_{10} x^{10}$

  - 簡單來說，高次多項式的模型一定比較複雜，容易過度適配訓練資料。
  - 所以我們希望讓模型變得較簡單、平滑一點。
  - → 那可以額外加上什麼<u>條件限制</u>嗎？

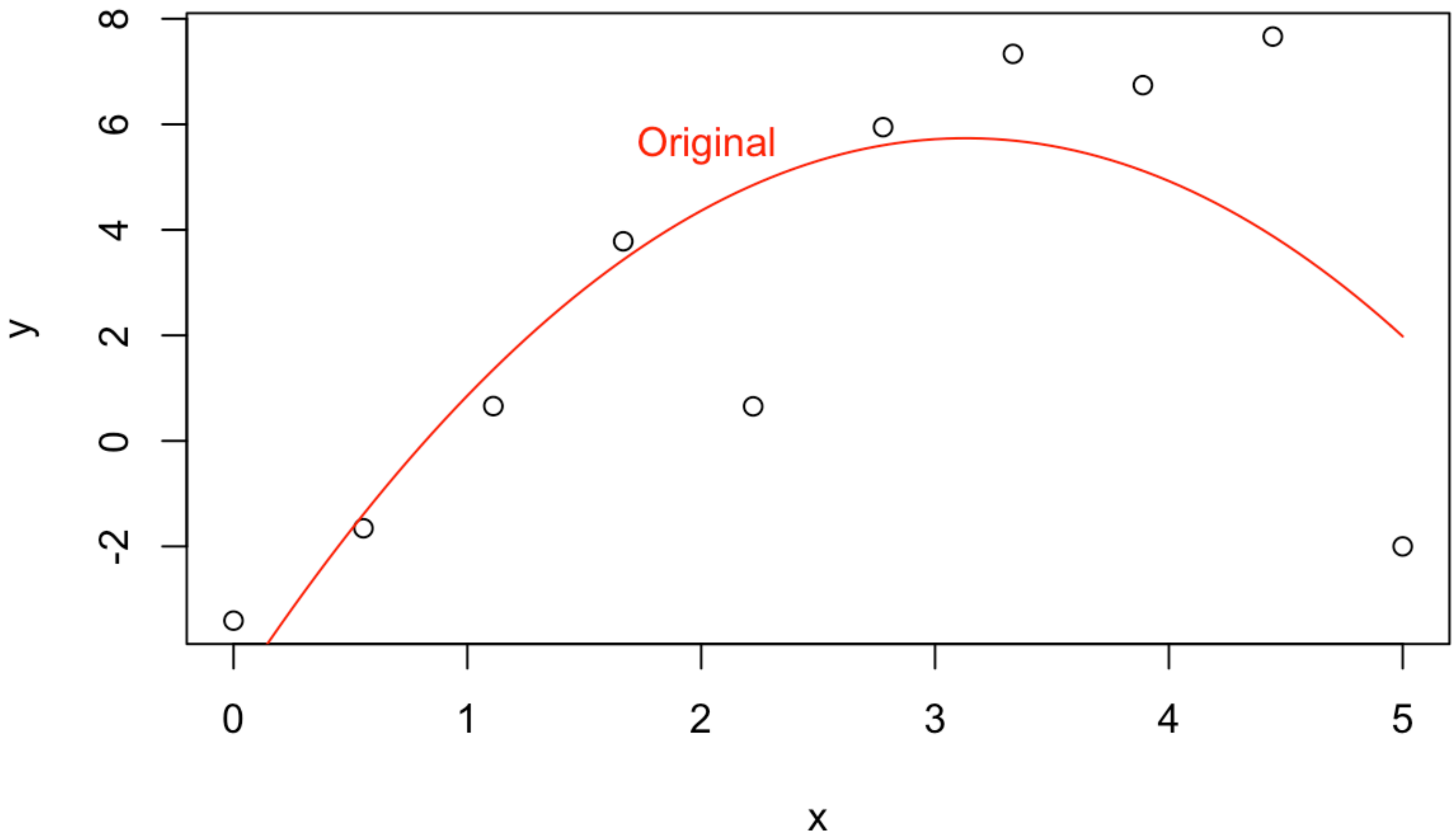
- $H(C) \equiv \{w \in \mathbb{R}^{10+1}, while\ \|w\|^2 \le C\}$
- Regression with $H(C)$: $\min_{w \in \mathbb{R}^{10+1}} E_{in}(w)\ \ s.t. \sum_{q=0}^{10} w_q^2 \le C$

  - <u>條件限制</u>的用意：C 是設定好的上限，我們希望 w_q 權重值可以限制在某範圍內，當 w_q 值接近或等於零時，w_q^2 便會很小。
  - → 多項式函數的次方項影響較輕微 → 多項式函數模型較平滑。

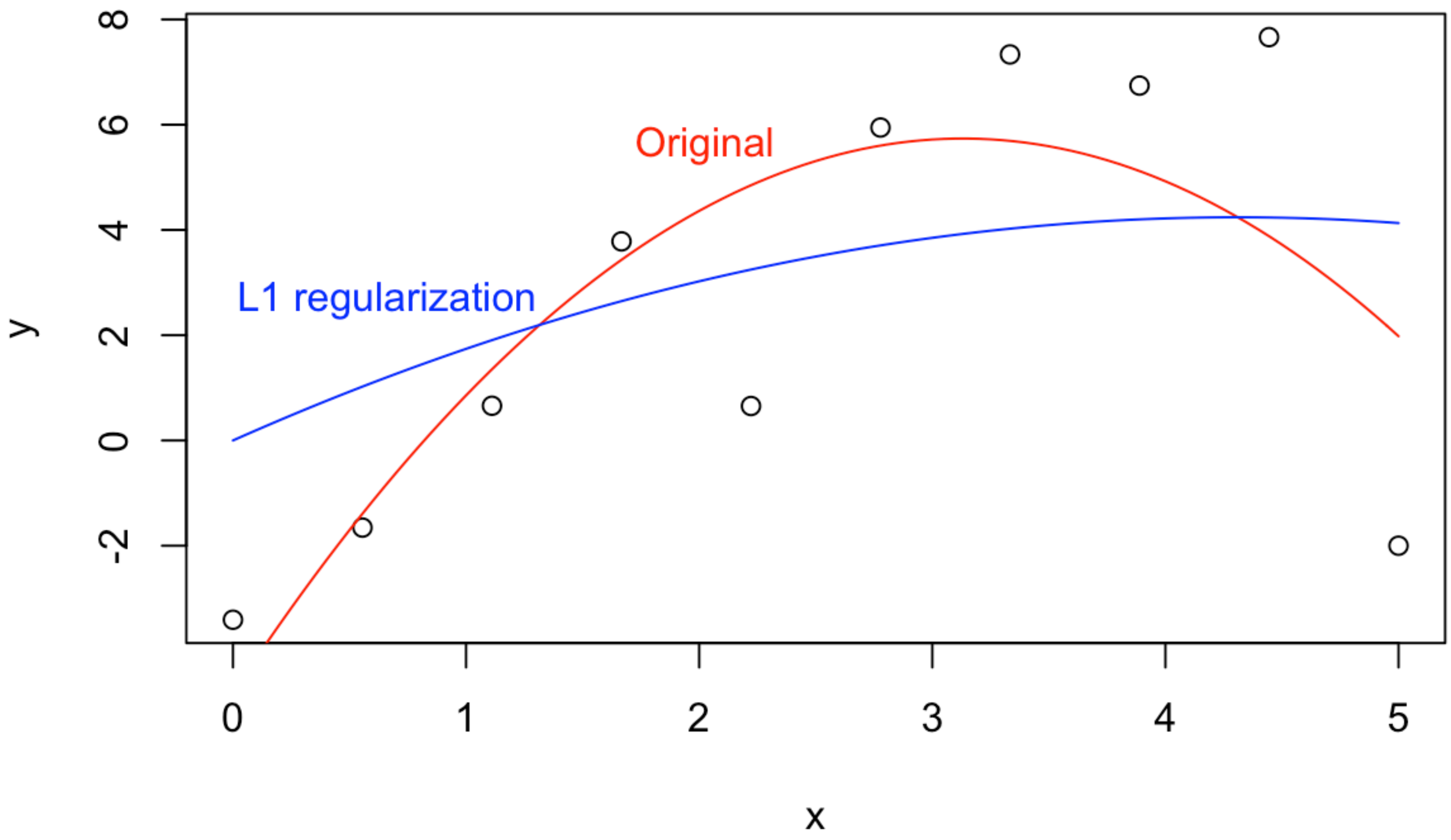
Regularization



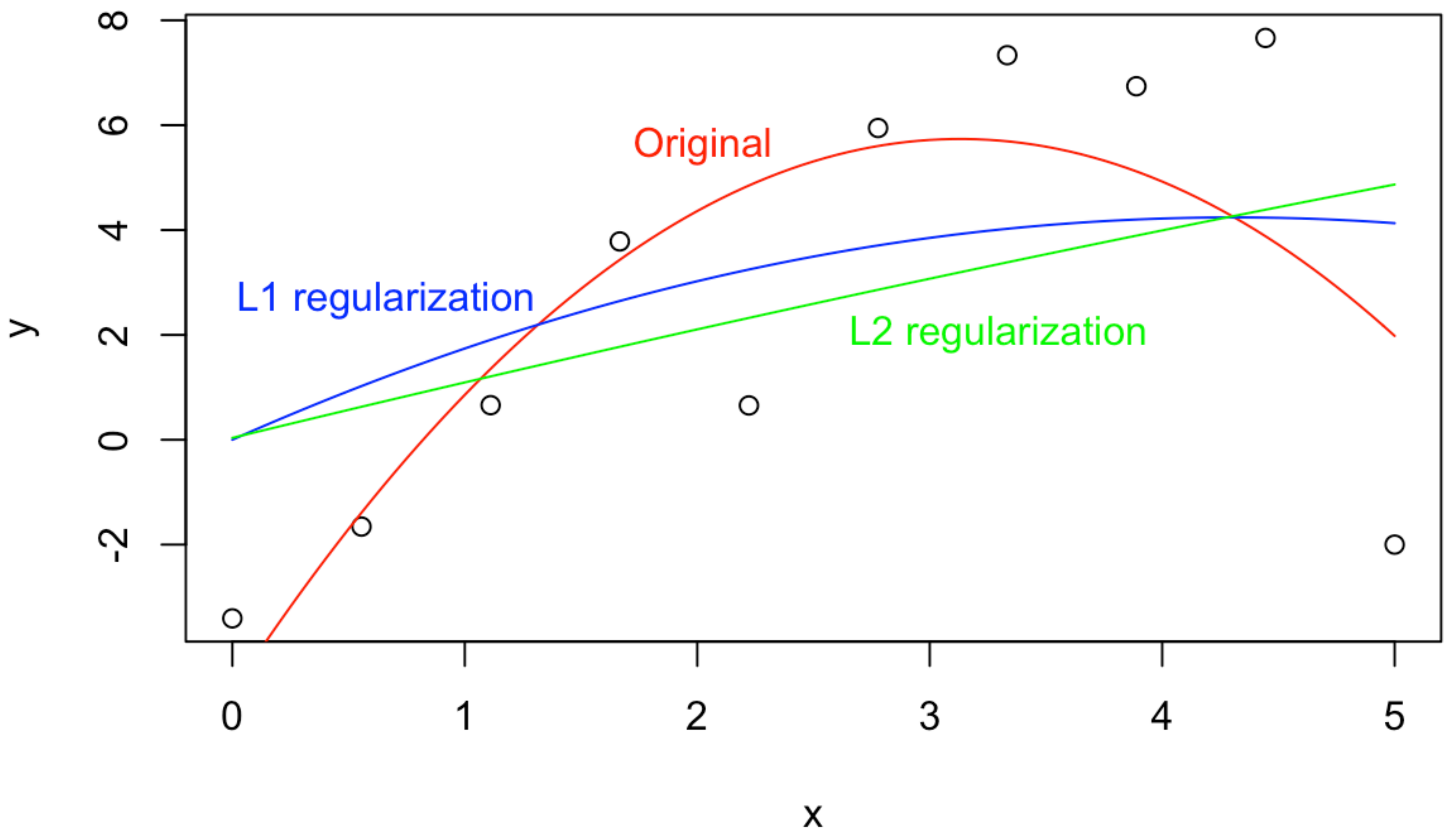
Regularization



Regularization



Regularization



Regularization

用一個簡單的線性迴歸來說明：

最小平方法 $OLS\colon \min_{\vec{w}} \|\vec{y} - A\vec{w}\|_2^2$

◎L1 regularization:

- $\min_{\vec{w}} \|\vec{y} - A\vec{w}\|_2^2 + \lambda \|\vec{w}\|_1$ Lasso Regression

◎L2 regularization:

- $\min_{\vec{w}} \|\vec{y} - A\vec{w}\|_2^2 + \lambda \|\vec{w}\|_2^2$ Ridge Regression

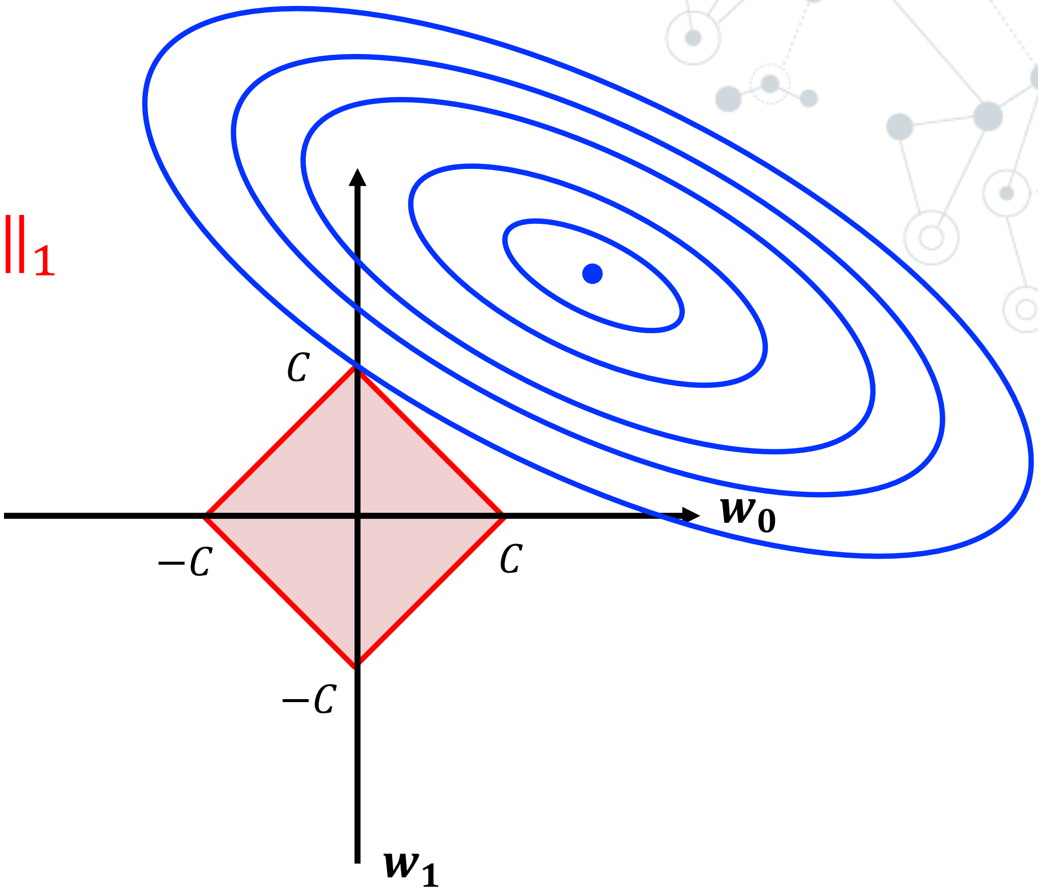
L1 regularization

◎L1 regularization:

- $\min_{\vec{w}} \|\vec{y} - A\vec{w}\|_2^2 + \lambda \|\vec{w}\|_1$

等同於：

- $\min_{\vec{w}} \|\vec{y} - A\vec{w}\|_2^2$,
- $s.t.\ \|\vec{w}\|_1 \leq C$

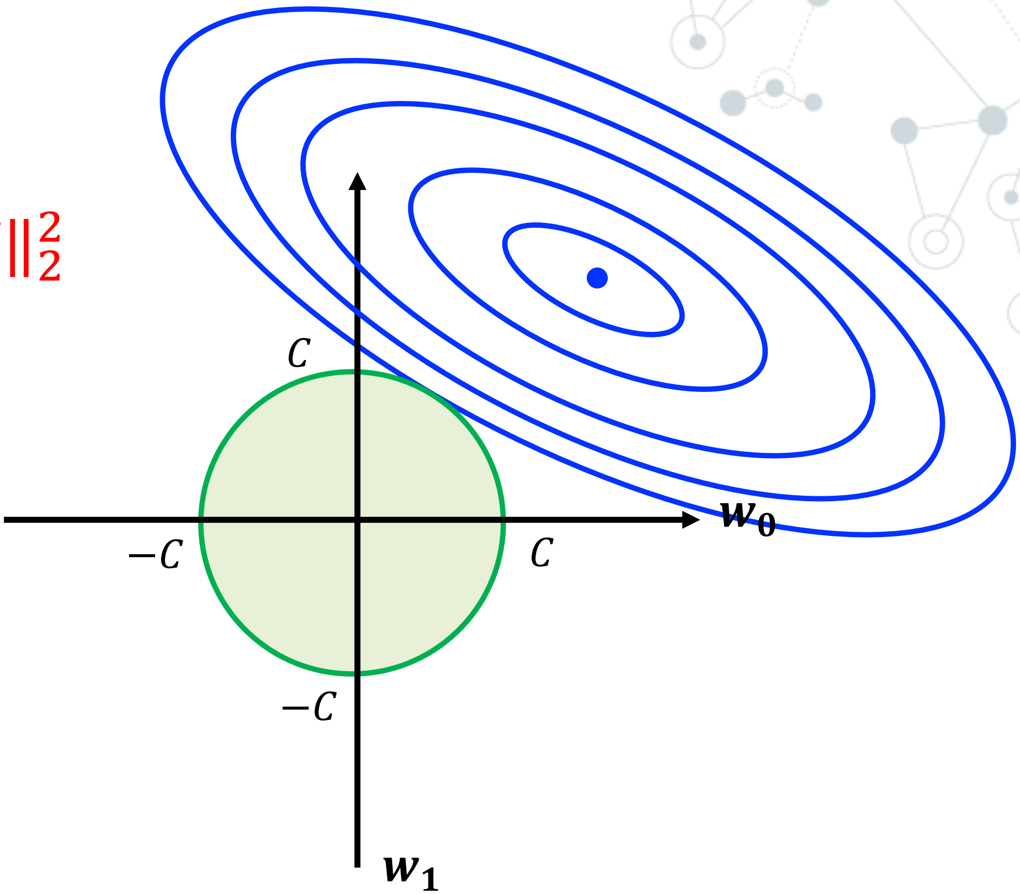


- *L1 regularization* 最佳化問題的解，常常會發生在 $\|\vec{w}\|_1 = C$ 的菱形頂點上。
- → Sparsity in solution: $\vec{w}$ 裡面的分量有很多是0，只有少數有值。

L2 regularization

- L2 regularization:
  - $\min_{\vec{w}} \|\vec{y} - A\vec{w}\|_2^2 + \lambda \|\vec{w}\|_2^2$

等同於：
  - $\min_{\vec{w}} \|\vec{y} - A\vec{w}\|_2^2$,
  - $s.t. \|\vec{w}\|_2^2 \leq C$



- *L1, L2 regularization* 最佳化問題：當 C 越小時（對應於：λ 越大），regularization 越強烈，懲罰越多，模型越平滑。

Comparison

L2 regularization	L1 regularization
L2 norm	L1 norm
differentiable everywhere	not differentiable everywhere
easy to optimize	sparsity in solution (i.e. built-in feature selection)

weight-decay regularization

Interpretability: 模型容易解釋

Summary

- Regularization can be used with <u>any ML classification technique that's based on a mathematical equation</u>.
  - Examples include logistic regression, probability classification and neural networks.

- The major advantage of using regularization is that it often leads to a more accurate model.
  - 避免模型過度學習

- The major disadvantage is that it introduces an additional parameter value that must be determined, the regularization weight.
  - λ 值必須事先給定

簡單來說…

Regularization 就是一種避免模型過度適配（Overfitting）的手段，在 Python 套件中常常以懲罰項（ Penalty ）表示。

Thanks!