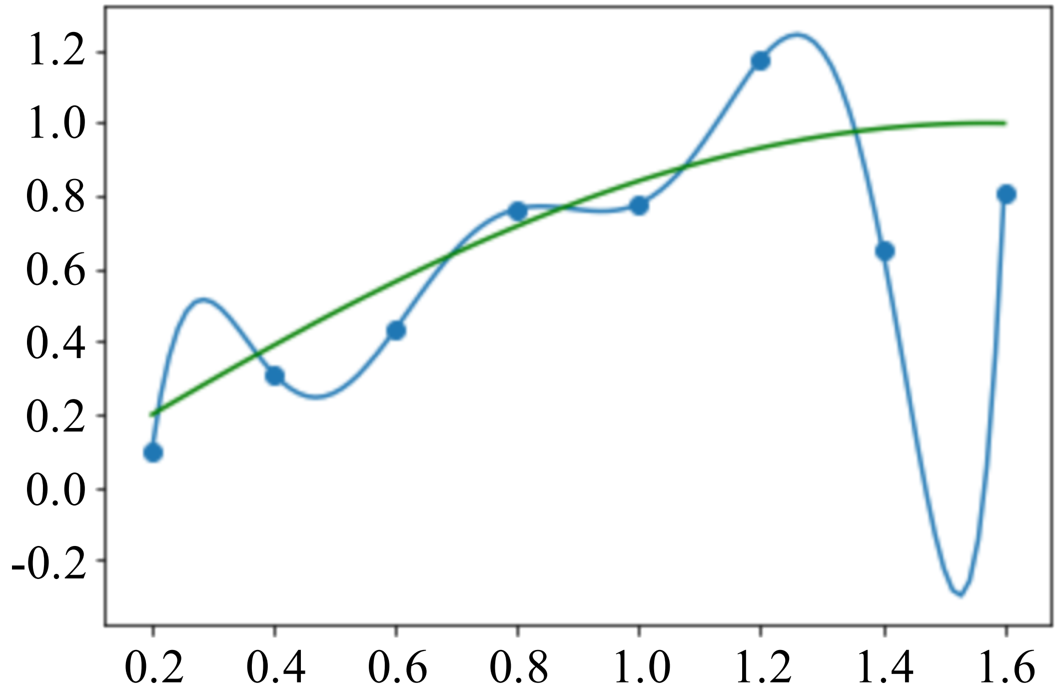
Overfitting example

Training set: $\{0.2, 0.4, \dots, 1.6\}$, $y = \sin(x) + \epsilon$

Model: $a(x) = b + w_1 x + w_2 x^2 + \cdots + w_8 x^8$

Parameters: $(130.0, -525.8, \dots, 102.6)$

Model just incorporates target into parameters!

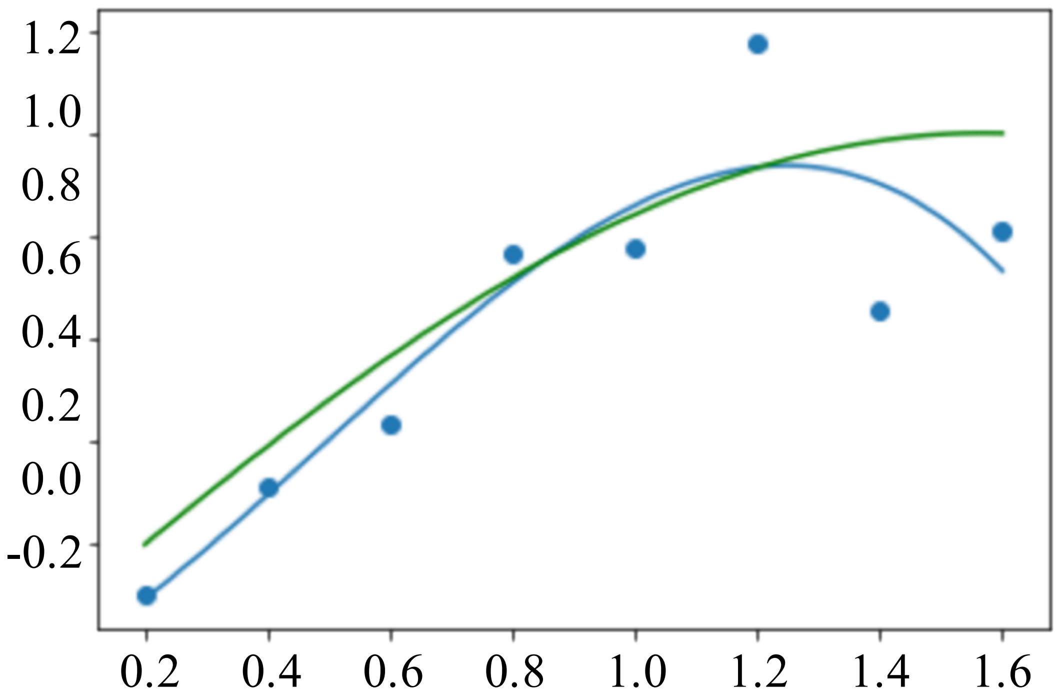


Overfitting example

Training set: $\{0.2, 0.4, \ldots, 1.6\}$, $y = \sin(x) + \epsilon$

Model: $a(x) = b + w_1 x + w_2 x^2 + w_3 x^3$

Parameters: $(0.634, 0.918, -0.626)$



Regularization

Good model weights: $(0.634, 0.918, -0.626)$

Overfitted model weights: $(130.0, -525.8, \dots, 102.6)$

Weight penalty

$$L_{reg}(w) = L(w) + \lambda R(w) \to \min_{w}$$

- $L(w)$ — loss function (MSE, log-loss, etc.)
- $R(w)$ — regularizer (e.g. penalizes large weights)
- λ — regularization strength

L2 penalty

$$L_{reg}(w) = L(w) + \lambda \|w\|^2 \to \min_{w}$$

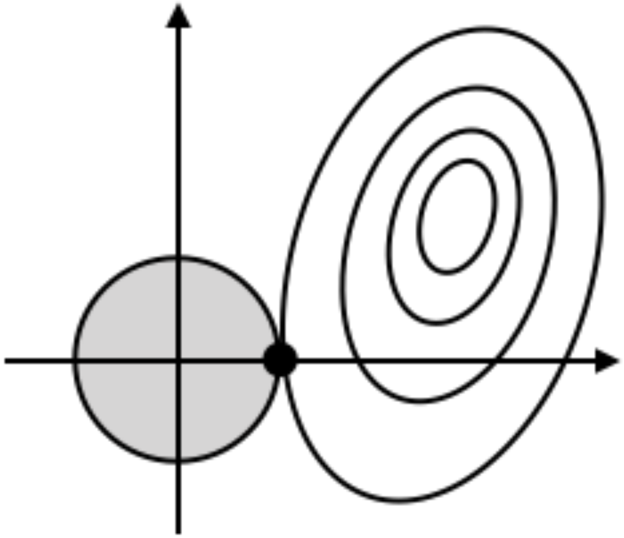
- $\|w\|^2 = \sum_{j=1}^{d} w_j^2$
- Drives all weights **closer** to zero
- Can be optimized with gradient methods

L2 penalty

$$L_{reg}(w) = L(w) + \lambda \|w\|^2 \to \min_w$$

The optimization problem is equivalent to

$$\begin{cases} L(w) \to \min\limits_w \\ \text{ s.t. } \|w\|^2 \leq C \end{cases}$$



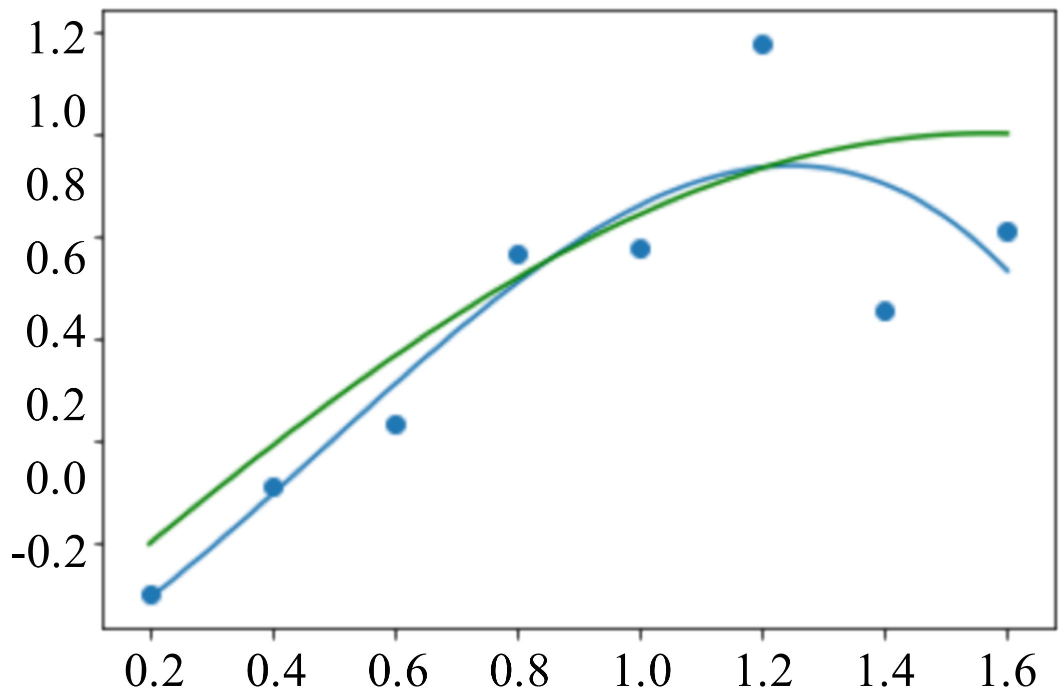
L2 penalty

$$L_{reg}(w) = L(w) + \lambda \|w\|^2 \to \min_w$$

Training set: $\{0.2, 0.4, \dots, 1.6\}$, $y = \sin(x) + \epsilon$

Model: $a(x) = b + w_1 x + w_2 x^2 + \dots + w_8 x^8$

Parameters: $(0.166, 0.168, 0.13, 0.075, 0.014, -0.04, -0.05, 0.018)$



L1 penalty

$$L_{reg}(w) = L(w) + \lambda \|w\|_1 \to \min_w$$

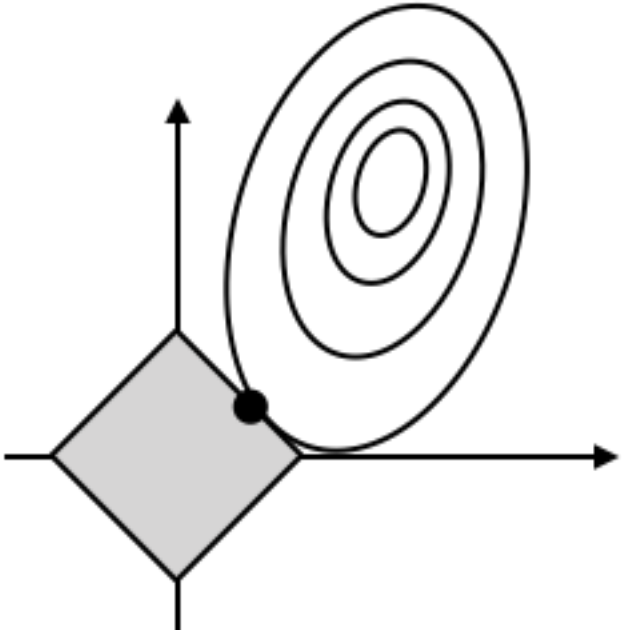
- $\|w\|_1 = \sum_{j=1}^{d} |w_j|$
- Drives some weights **exactly** to zero
- Learns sparse models
- Cannot be optimized with simple gradient methods

L1 penalty

$$L_{reg}(w) = L(w) + \lambda \|w\|_1 \to \min_w$$

The optimization problem is equivalent to

$$\begin{cases} L(w) \to \min\limits_w \\ \text{ s.t. } \|w\|_1 \leq C \end{cases}$$



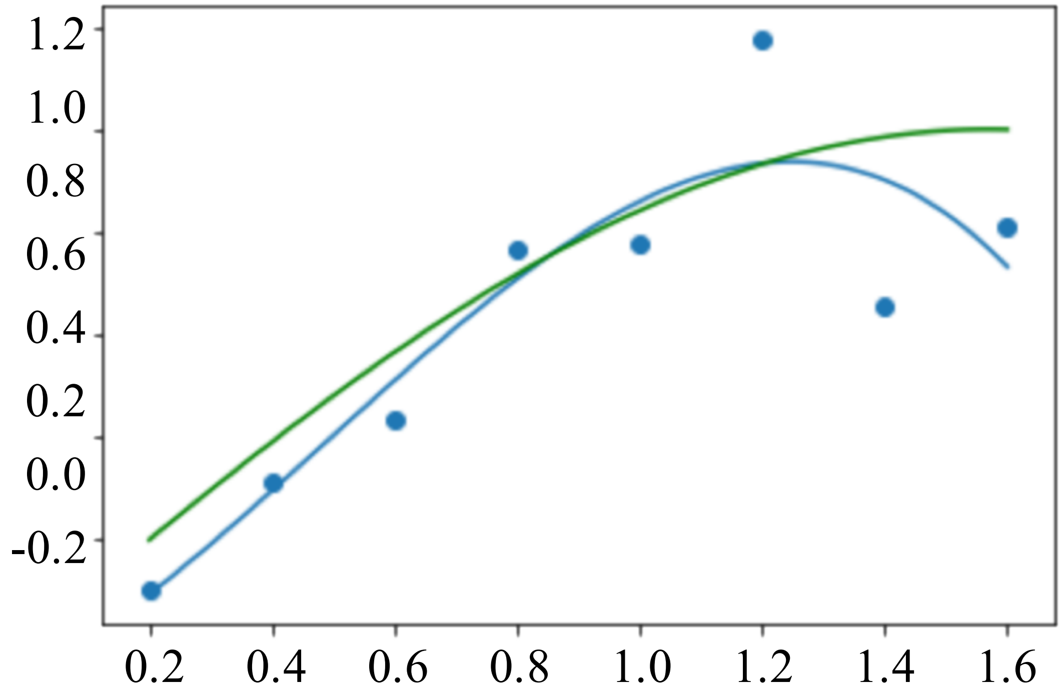
L1 penalty

$$L_{reg}(w) = L(w) + \lambda \|w\|_1 \to \min_w$$

Training set: $\{0.2, 0.4, \dots, 1.6\}$, $y = \sin(x) + \epsilon$

Model: $a(x) = b + w_1 x + w_2 x^2 + \cdots + w_8 x^8$

Parameters: (for $\lambda = 0.01$): (0.78, 0.03, **0**, **0**, **0**, −0.016, −0.01, **0**)



Other regularization techniques

- Dimensionality reduction
- Data augmentation
- Dropout
- Early stopping
- Collect more data

Summary

- One should restrict model complexity to prevent overfitting
- Common approach: penalize large weights
- Other approaches: next modules