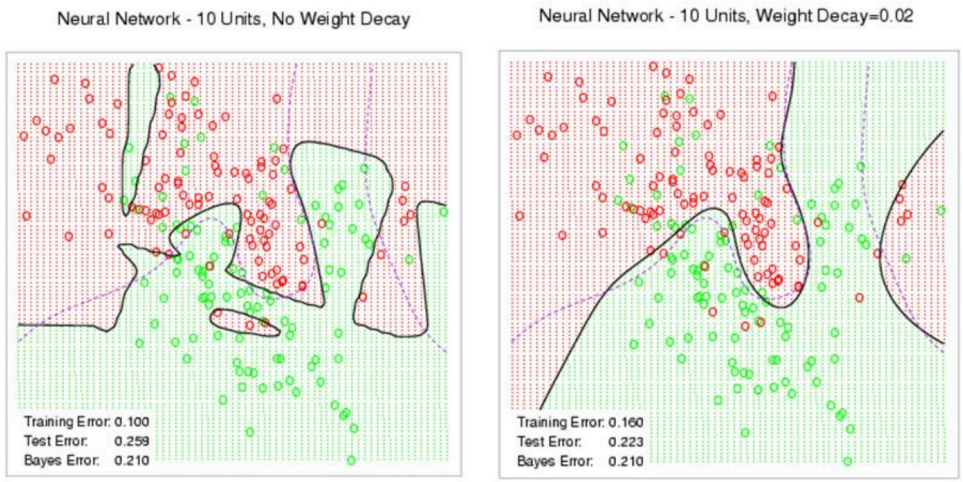
权重衰退

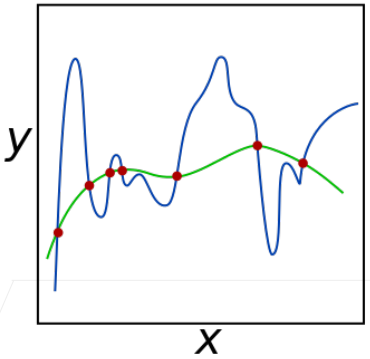

Neural Network - 10 Units, No Weight Decay
Training Error: 0.100
Test Error: 0.259
Bayes Error: 0.210
Neural Network - 10 Units, Weight Decay=0.02
Training Error: 0.160
Test Error: 0.223
Bayes Error: 0.210

使用均方范数作为硬性限制

- 通过限制参数值的选择范围来控制模型容量

$$\min \ \ell(\mathbf{w}, b) \quad \text{subject to} \ \ \|\mathbf{w}\|^2 \leq \theta$$

  - 通常不限制偏移 b （限不限制都差不多）
  - 小的 θ 意味着更强的正则项



使用均方范数作为柔性限制

- 对每个 θ，都可以找到 λ 使得之前的目标函数等价于下面

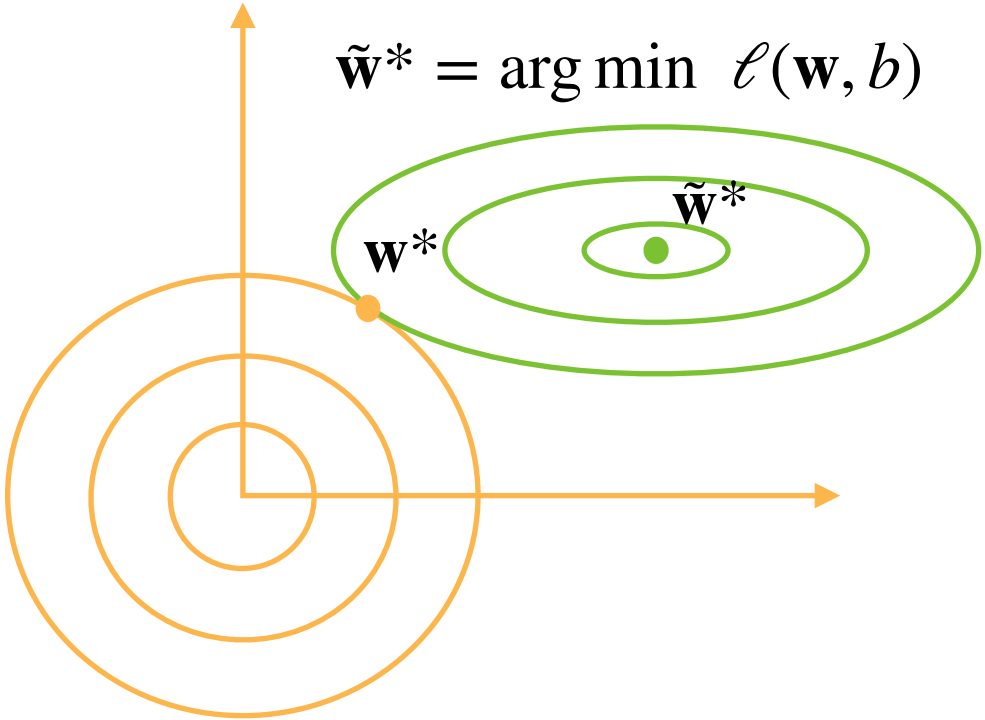
$$\min \ \ell(\mathbf{w}, b) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

  - 可以通过拉格朗日乘子来证明
- 超参数 λ 控制了正则项的重要程度
  - $\lambda = 0$：无作用
  - $\lambda \to \infty, \mathbf{w}^* \to \mathbf{0}$

演示对最优解的影响

$$\mathbf{w}^* = \arg\min \ \ell(\mathbf{w}, b) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

$$\tilde{\mathbf{w}}^* = \arg\min \ \ell(\mathbf{w}, b)$$



参数更新法则

- 计算梯度

$$\frac{\partial}{\partial \mathbf{w}}\left(\ell(\mathbf{w}, b)+\frac{\lambda}{2}\|\mathbf{w}\|^2\right)=\frac{\partial \ell(\mathbf{w}, b)}{\partial \mathbf{w}}+\lambda \mathbf{w}$$

- 时间 t 更新参数

$$\mathbf{w}_{t+1}=(1-\eta \lambda) \mathbf{w}_t-\eta \frac{\partial \ell(\mathbf{w}_t, b_t)}{\partial \mathbf{w}_t}$$

  - 通常 $\eta\lambda < 1$，在深度学习中通常叫做权重衰退

总结

- 权重衰退通过 L2 正则项使得模型参数不会过大，从而控制模型复杂度
- 正则项权重是控制模型复杂度的超参数