Regularization

JrPhy

Introduction

- So far we can accept some mistakes, but we still want to the mistakes are the least, it's called regularization. Now we've find a hyperplane and add a constrain $\min_{w} E_{in}(w)$ such that $\sum_{i} {w_i}^2 \le C$

- This problem can be solve by Lagrange undetermined multiplier, here we use square error, let $z_i = f(x_i)$ is our transformation, and the error is
- $E(w_i) = (z_i w^T - y_i)^2$, $E = \sum_{i} E(w_i) = \left(Zw^T - Y\right)\left(Zw^T - Y\right)^T$,

Optimize

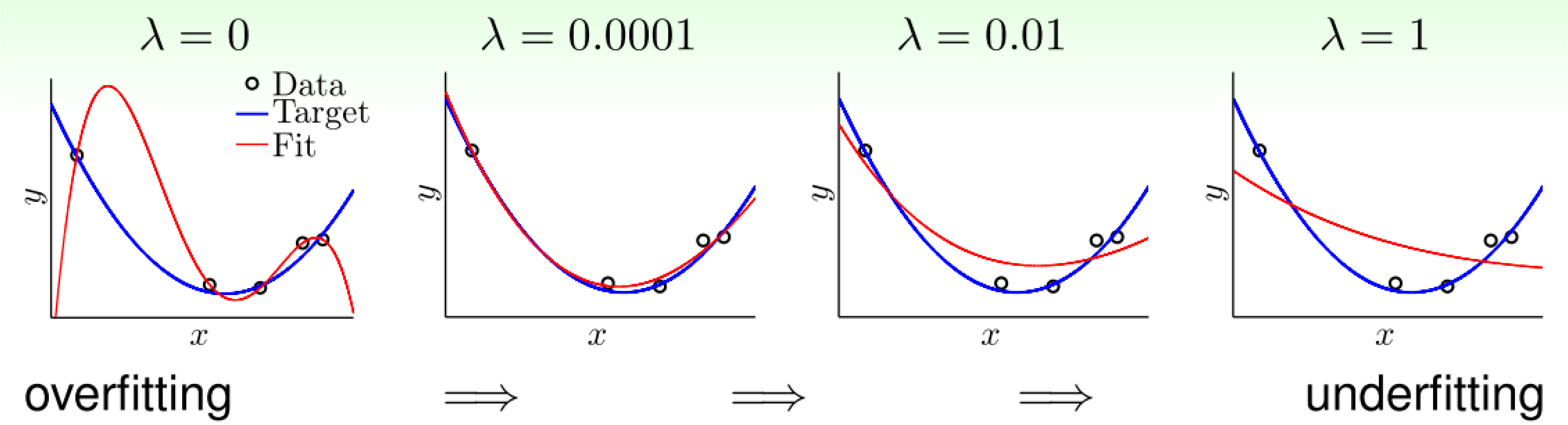
- What we want to solve is $\min_w \left(\left(Zw^T - Y \right)\left(Zw^T - Y \right)^T + \frac{\lambda}{N} ww^T \right) \quad \lambda > 0$

$$\frac{\partial}{\partial w}\left(\left(Zw^T - Y \right)\left(Zw^T - Y \right)^T + \frac{\lambda}{N} ww^T \right) = 0$$

$$\frac{2}{N}\left(ZZ^T w^T - Z^T Y \right) + \frac{2\lambda}{N} w^T = 0 \rightarrow w = \left(ZZ^T + \lambda I \right)^{-1} ZY$$

- As the *x* is bigger, then higher order term increase fast, so its coefficient is usually small. So that λ can suppress higher order term as it increases a small number

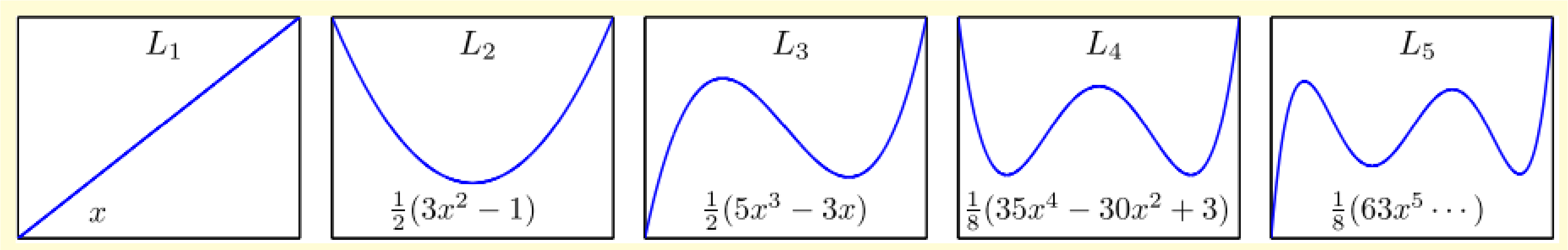
Choosing λ



Hsuan Tien Lin, mltech/214_handout

A tip of polynomial transform

- We‘ve mentioned previous before, if you use polynomial transform, it is recommended to use Legendre transform, it’s also a polynomial but with better numerical property.
- $-1 \le L(x_i) \le 1, \quad \int_{-1}^{1} L(x_i)L(x_j)dx = \delta_{ij}$



Hsuan Tien Lin, mltech/214_handout

L1 and L2 regularization

- So far we use the square error and we want to minimize it, it's called the L2 regularization in machine learning. There is another regularization, L1 regularization, its form is $|\boldsymbol{w}|$, both regularizations can avoid the noise.
- L2 regularization "$\boldsymbol{w}\boldsymbol{w}^{\boldsymbol{T}}$": easy to optimize, differential everywhere
- L1 regularization "$|\boldsymbol{w}|$": hard to optimize, not differential at $|\boldsymbol{w}|$.
- The optimal solution with L1 regularization is near $|\ \boldsymbol{w}\ | = 0$, so the most coefficients are 0, it's called "sparse solution", so it just needs calculate a few terms.

Use your own regularization

- It’s ok that choose your own regularization, there are some criterions of the choices:
- 1. target dependent
- 2. what the user want
- 3. easy to optimize
- If your regularization is bad, λ can protect your choice.

Optimal λ