「パターン認識と機械学習」
1-1 例：多項式曲線フィッティング

新納浩幸

回帰を例に機械学習を説明

$$y = f(\mathbf{x})$$

推定したい

N 個の観測データがある

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \cdots, (\mathbf{x}_N, y_N)\}$$

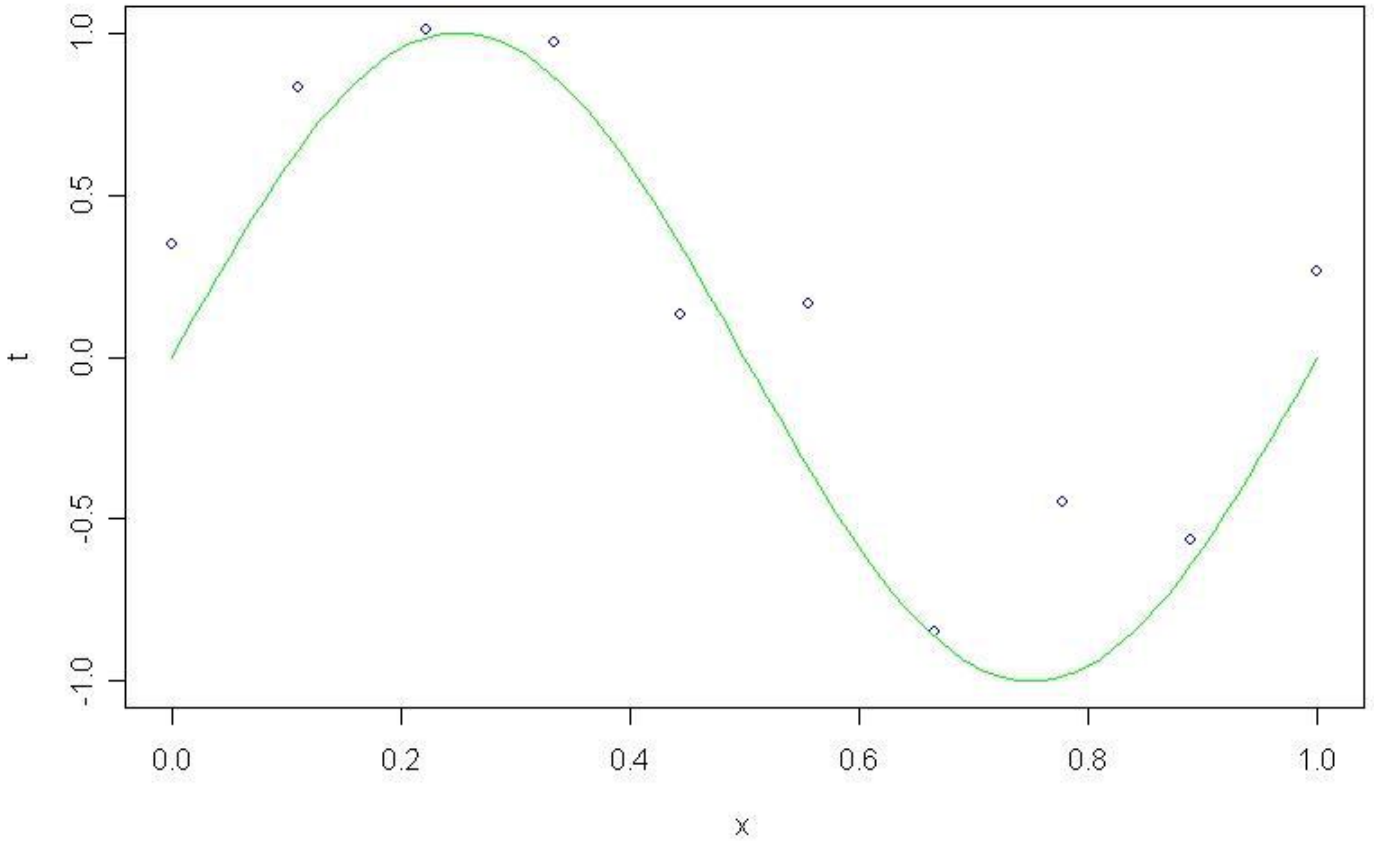
例

$$\sin(2\pi x) + N(0, 0.03)$$

真の関数　　ノイズ

確率的に生じる誤差

以下の 10 点で
真の関数を推定できるか



多項式モデル

$$f(x, w) = w_0 + w_1 x^1 + w_2 x^2 + \cdots + w_M x^M$$

$$= \sum_{i=0}^{M} w_i x^i$$

M は ？ とりあえずいくつかに固定する

つまり、以下が推定対象

$$\mathbf{w} = (w_0, w_1, \cdots, w_M)^t$$

最小2乗法

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{f(x_n, \mathbf{w}) - y_n\}^2$$

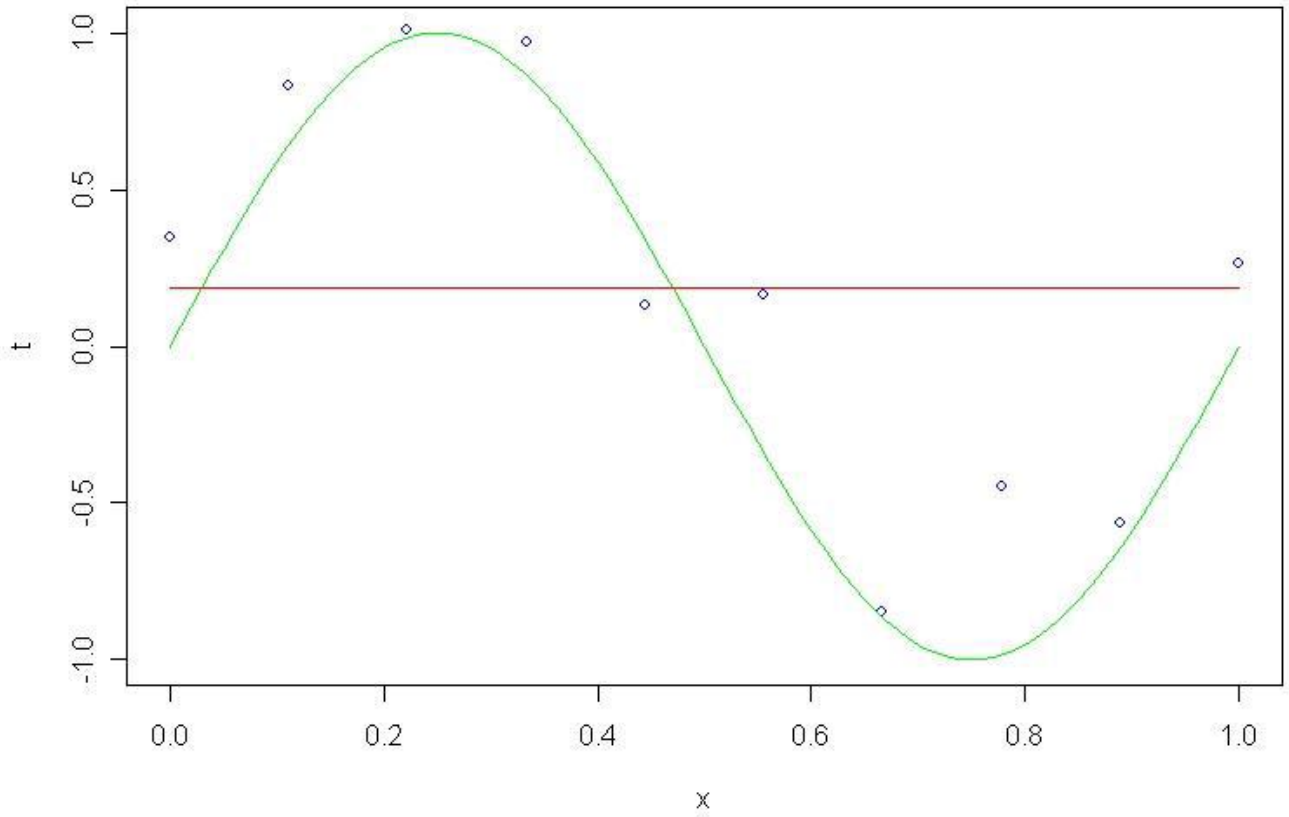
設定したモデル上の
理論値

実際の観測値

これを最小化する ｗ を求めればよい・・・

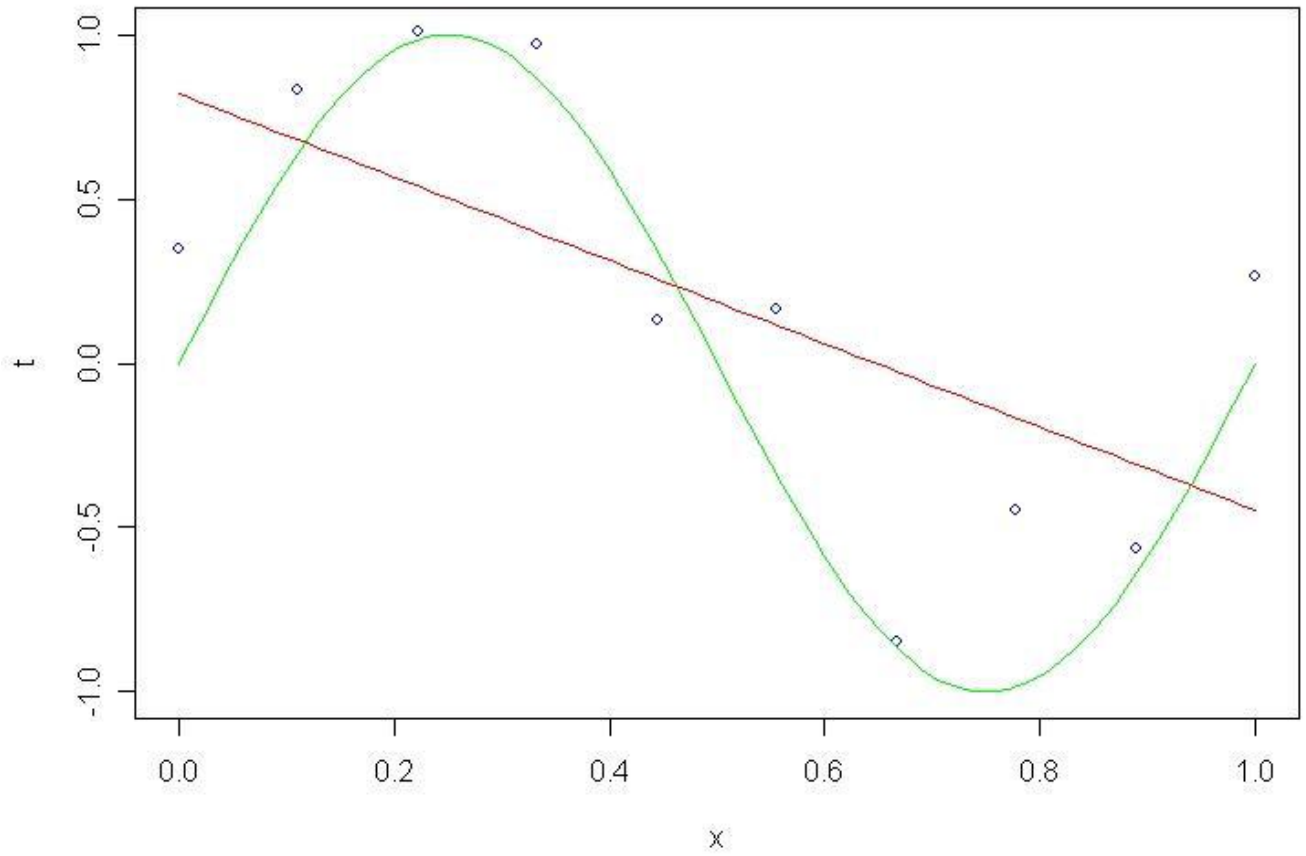
これは極値条件から簡単に求まる

M（次元）は？ M=0



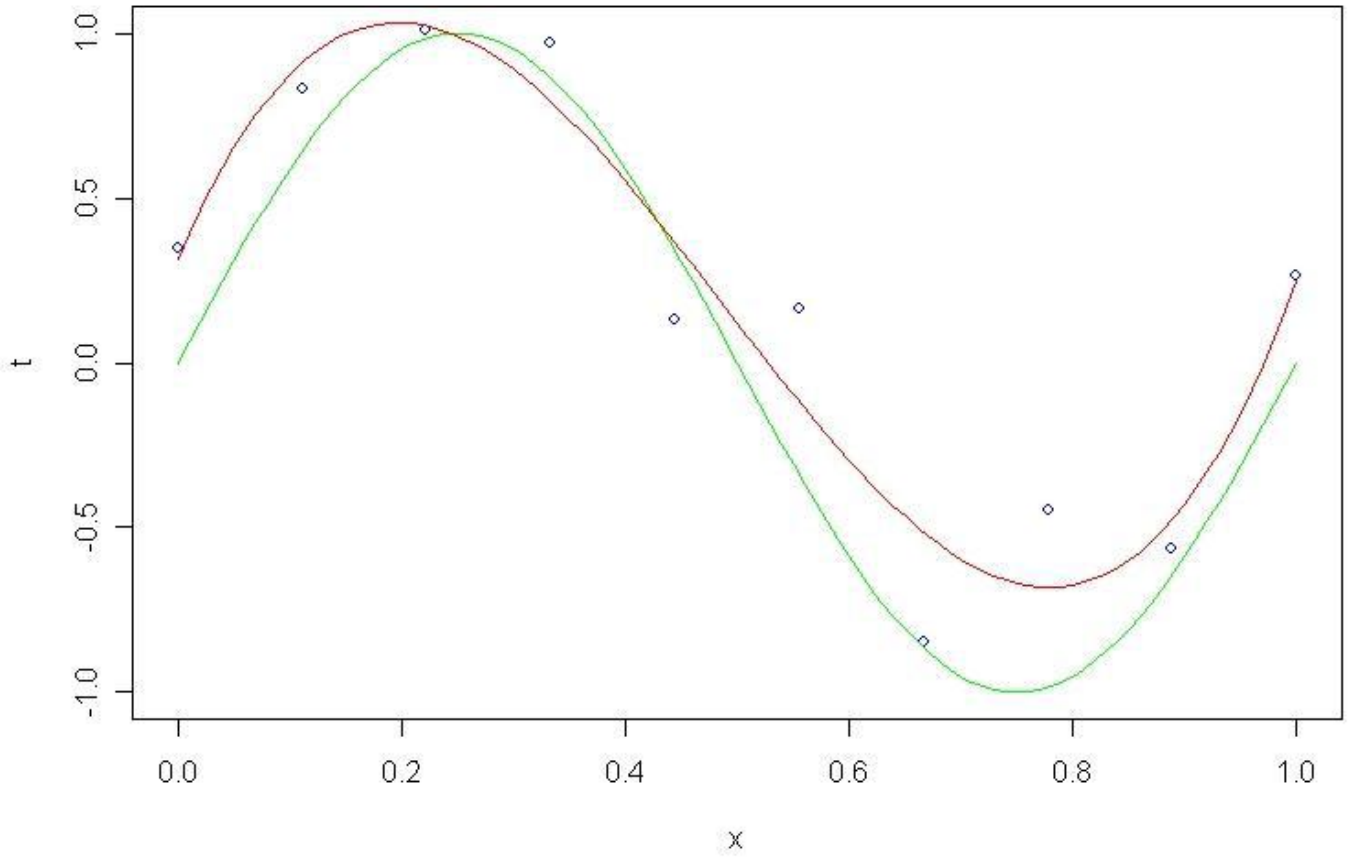
ダメ

M（次元）は？　M=1



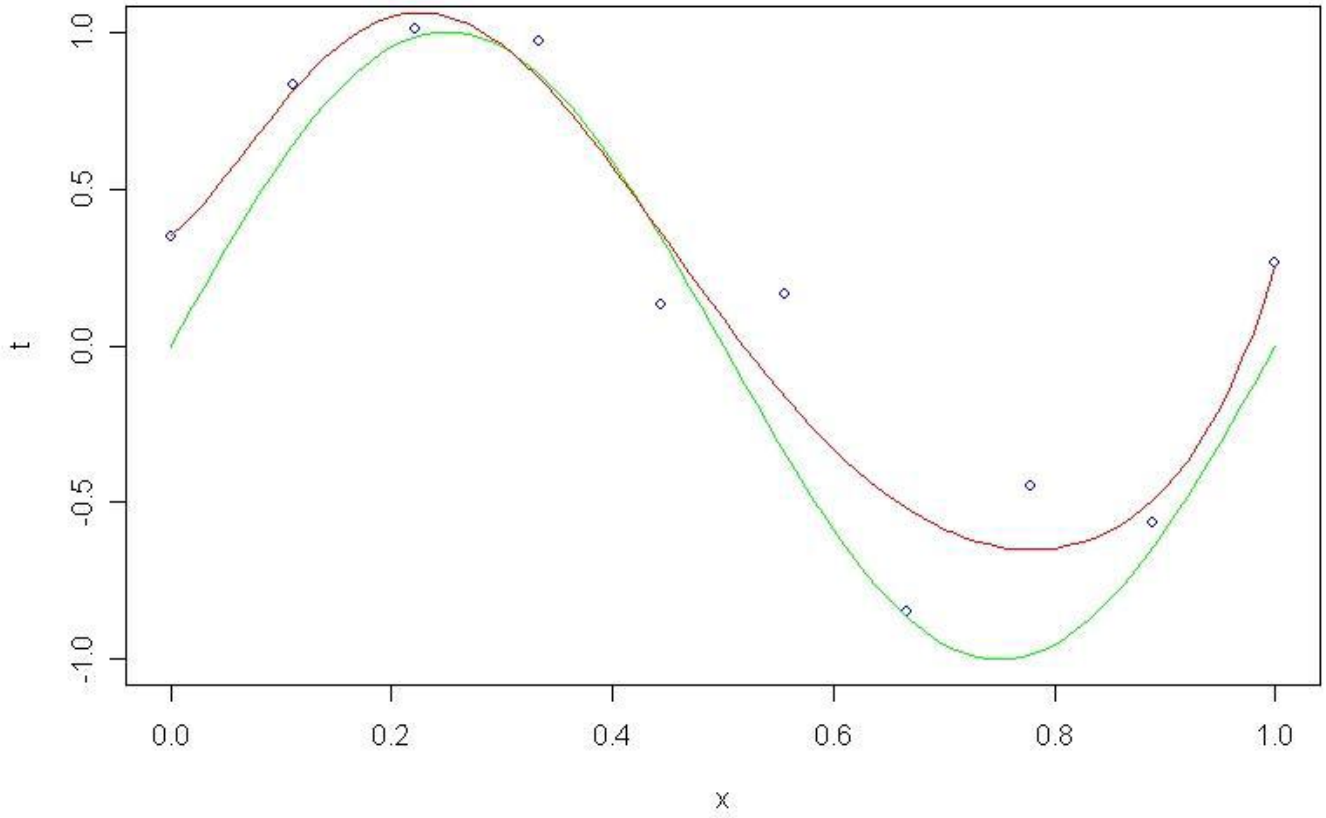
ダメ

M（次元）は？ M=3



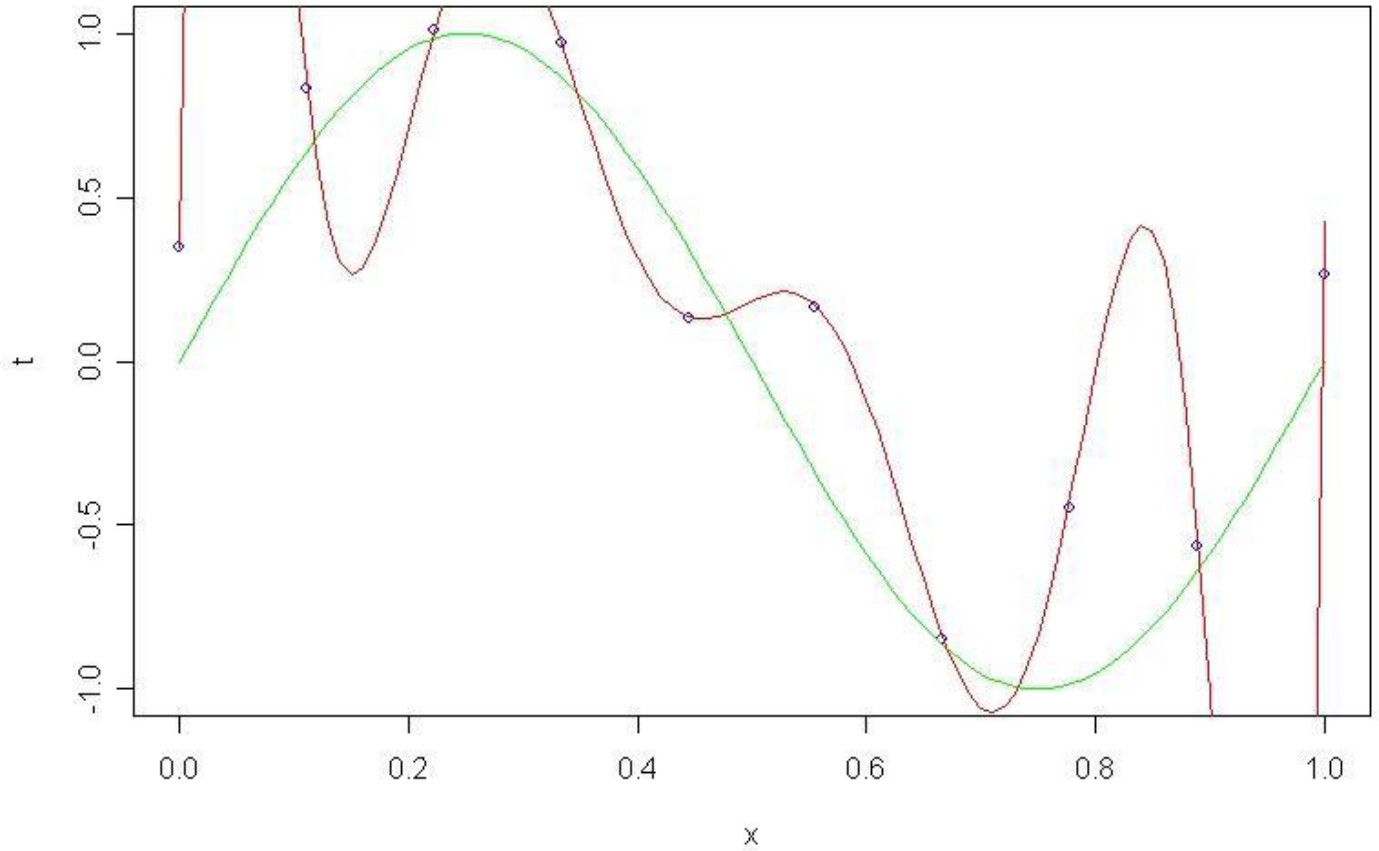
けっこういい？

M（次元）は？ M=6



これもまあまあ良い

M（次元）は？　M=9



ひどすぎる

多項式モデルのポイント

何次元で近似するか？

モデル選択の問題

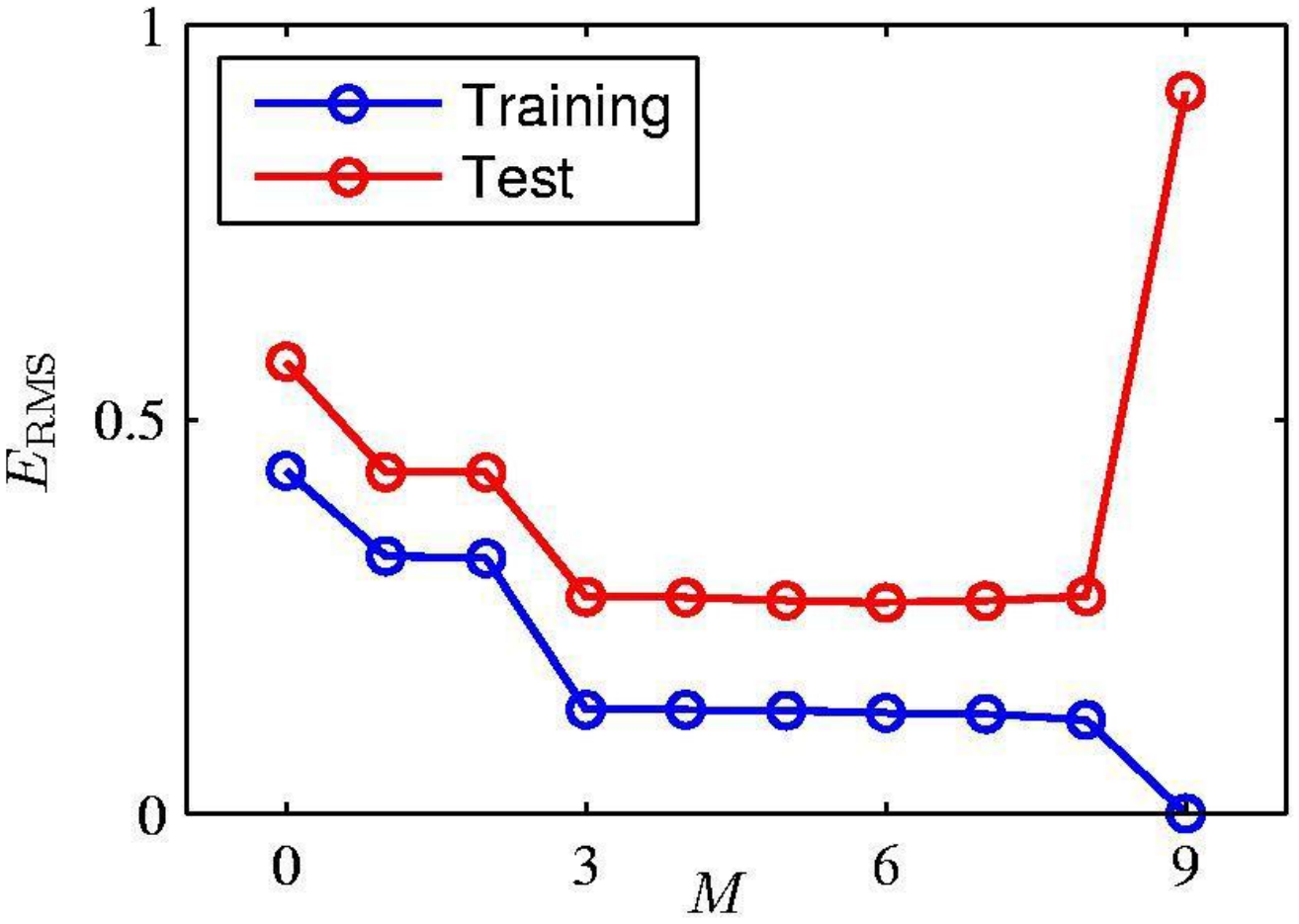
次元が高いと表現力が豊かなので良さそうだけ、
推定すべきパラメータ数が多くなって、簡単にはいかない

平均2乗平方根誤差

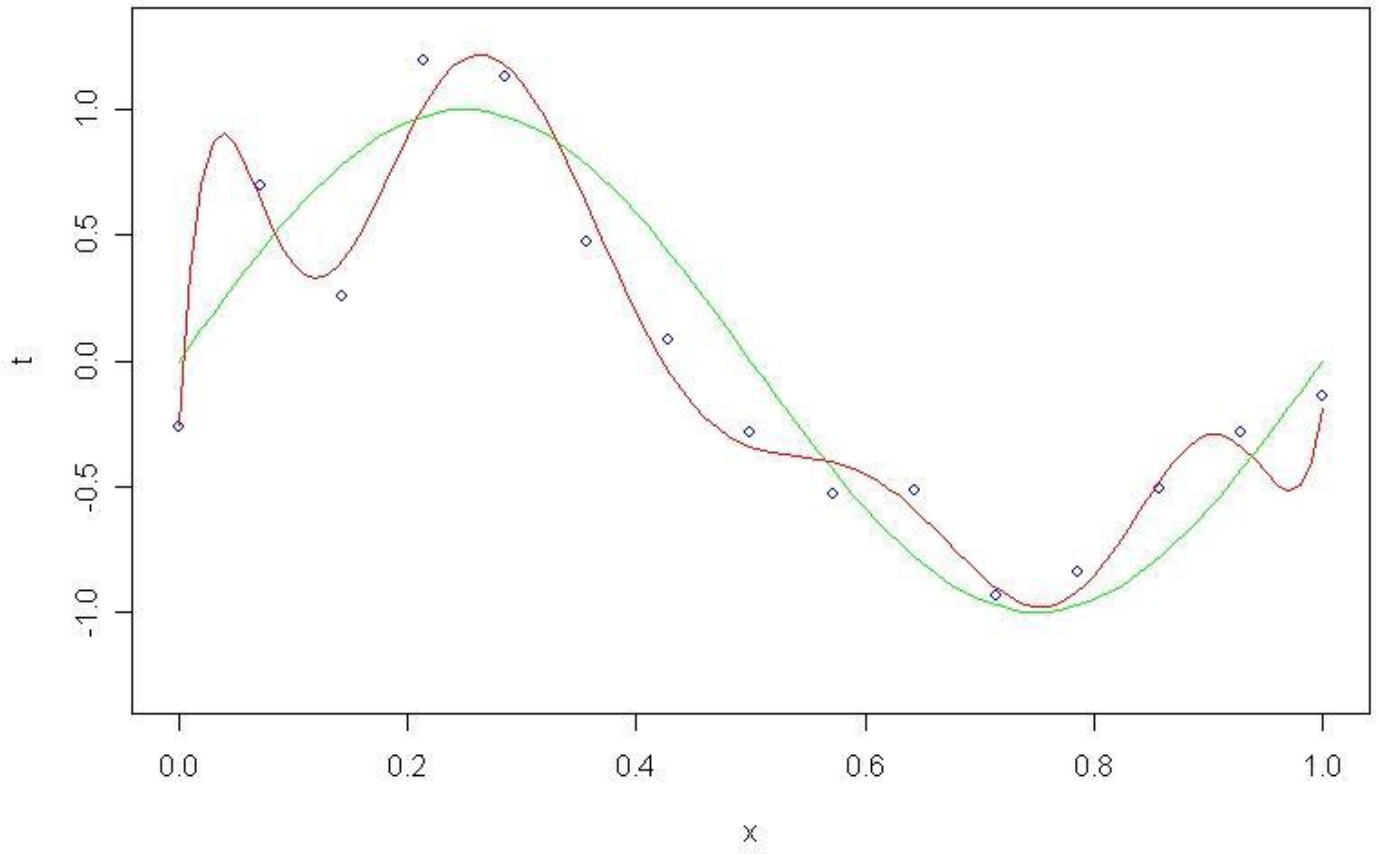
w の良さを測る尺度、小さいほど良い

$$E_{\mathrm{RMS}} = \sqrt{\frac{2E\left(w^{*}\right)}{N}}$$

RMS の結果

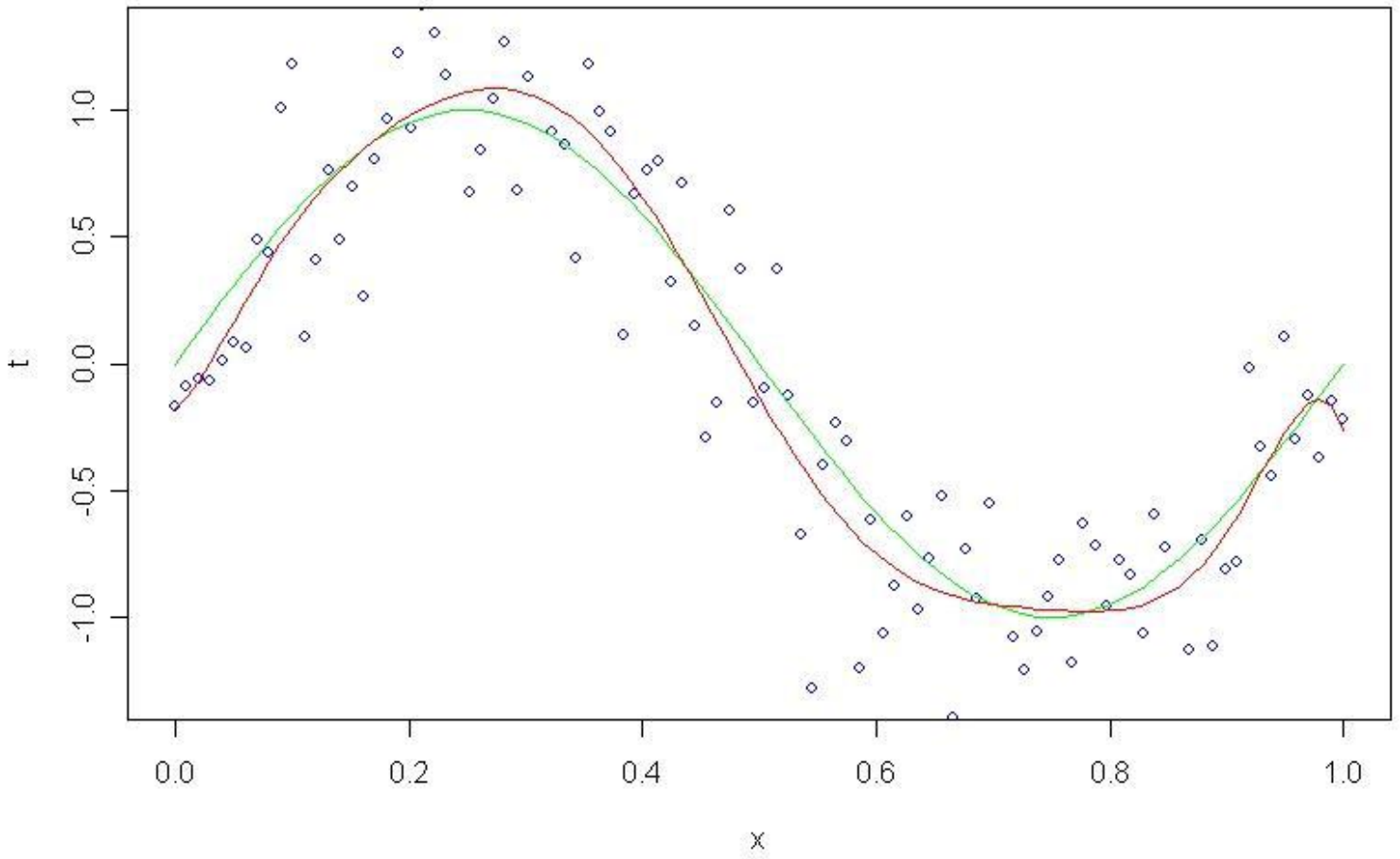


訓練データ数との関係（１）



M=9, N = 15

訓練データ数との関係（2）



M=9, N = 100

モデルとデータ数

適切なモデルの複雑さはデータ数とも関連する

訓練データの個数はパラメータ数の５倍が目安

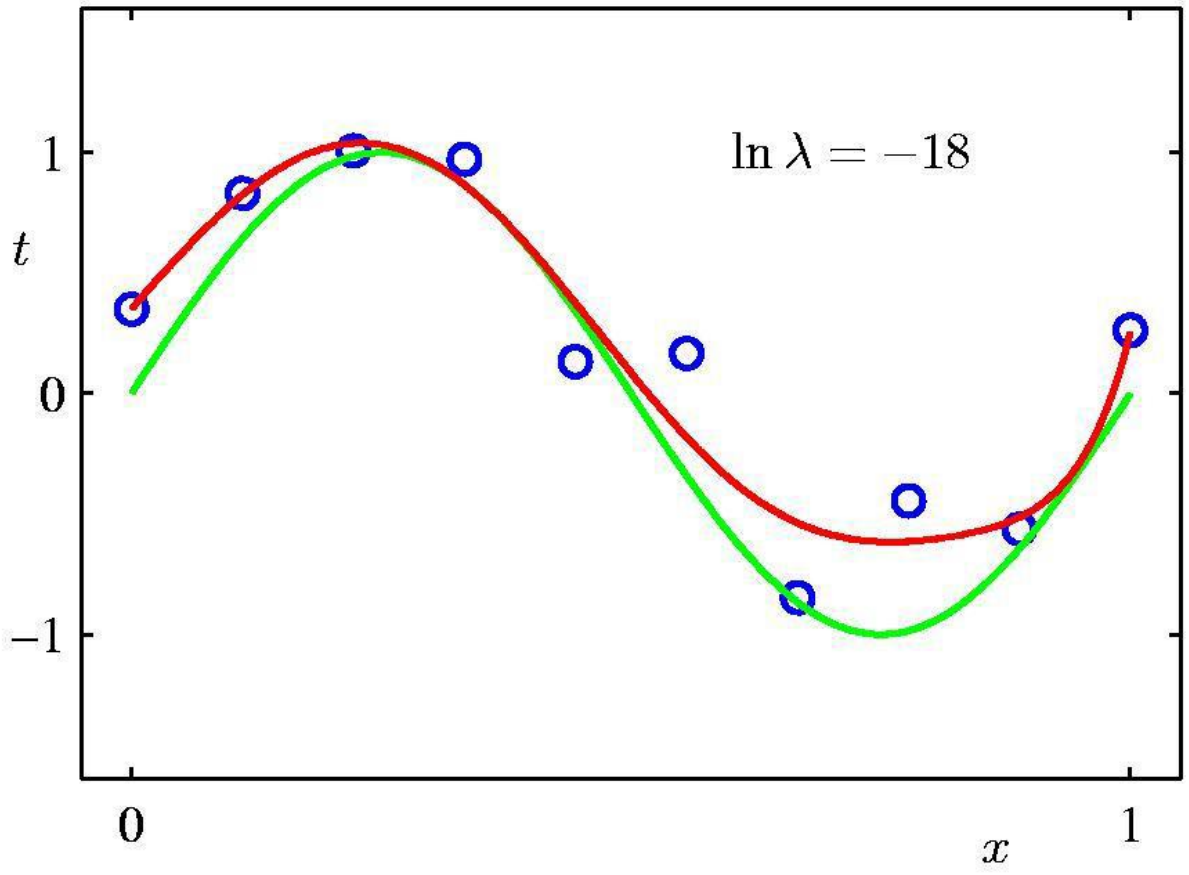
正則化

単純な最小2乗法ではなく、そこにペナルティ項をつけて学習させる

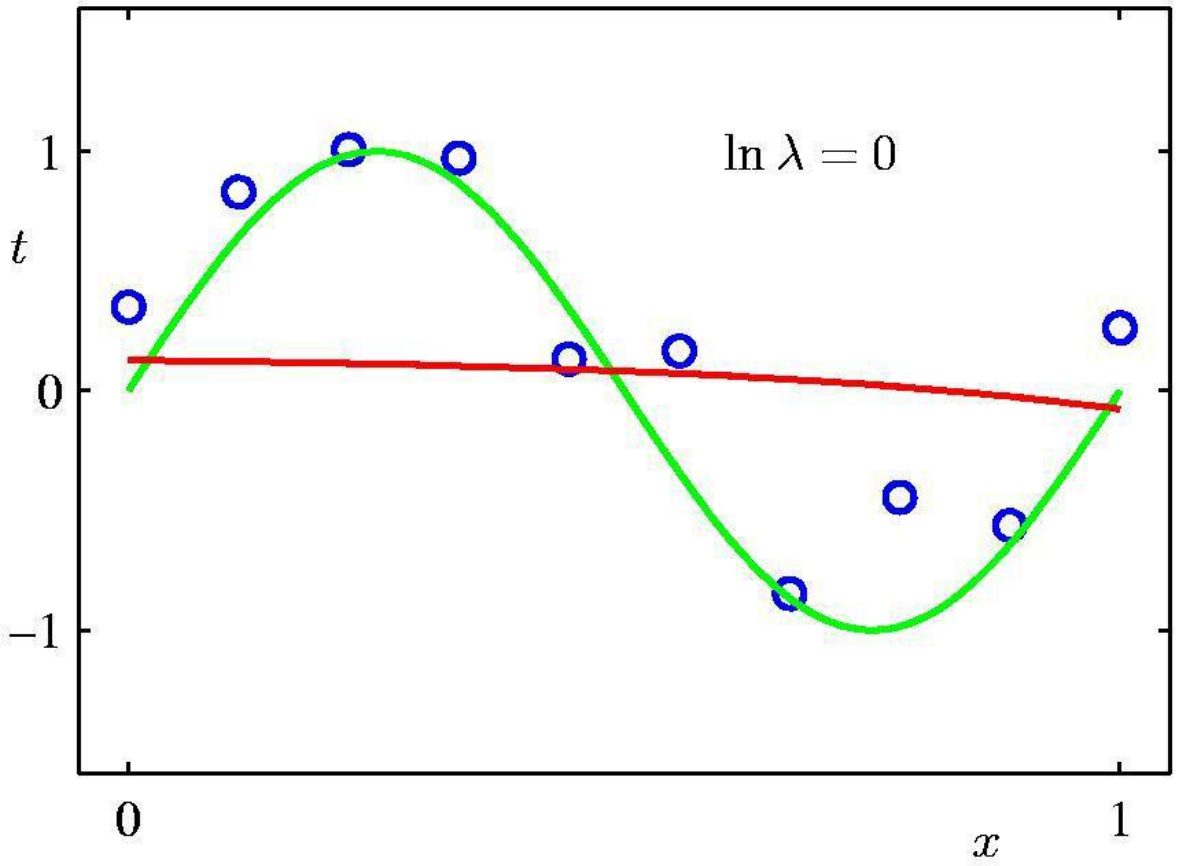
$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{f(x_n, \mathbf{w}) - y_n\}^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

$$||\mathbf{w}||^2 = w_0^2 + w_1^2 + \cdots + w_M^2$$

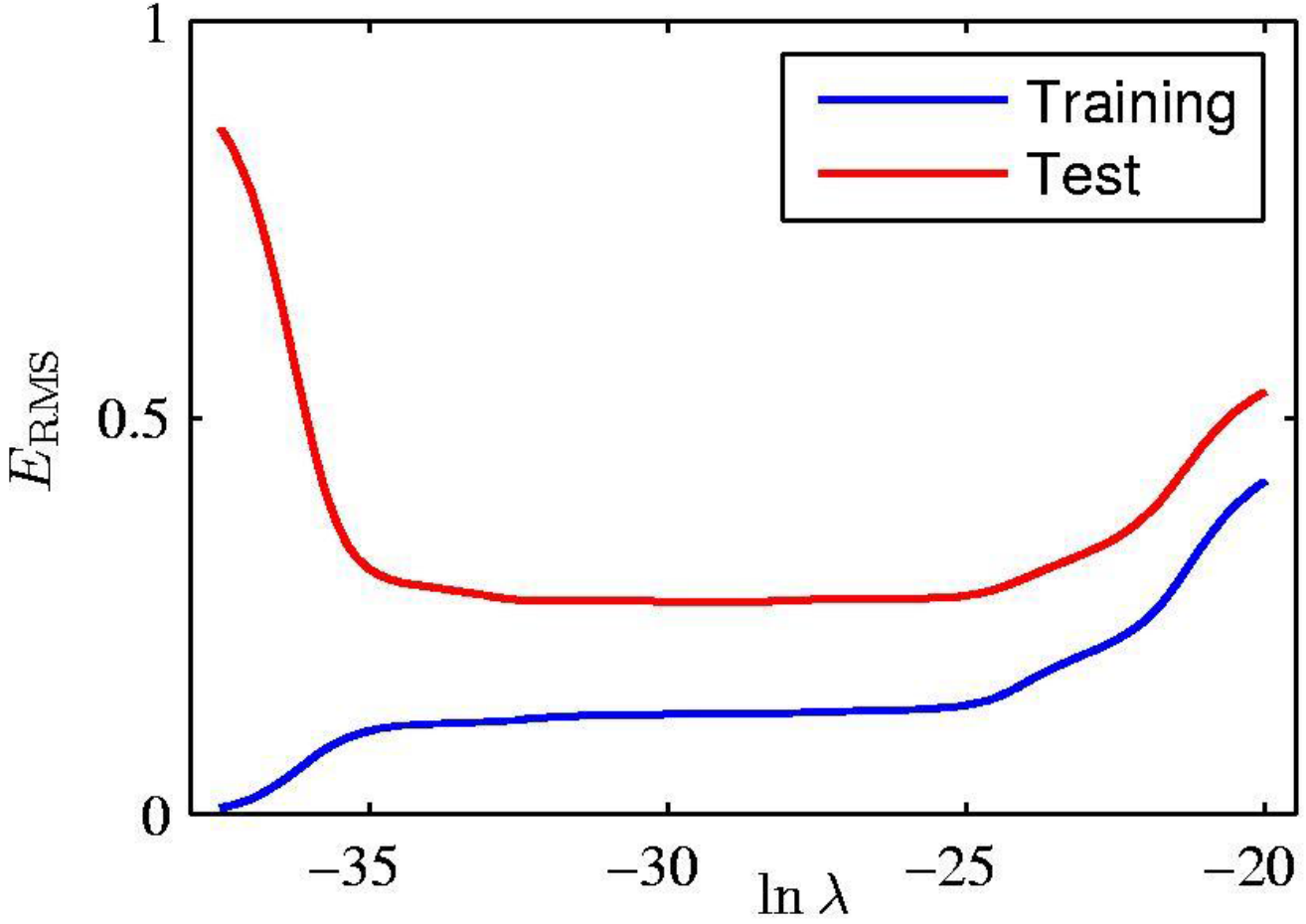
M=9, ln λ = -18



M=9, ln λ = 0



正則化項による誤差



結論

最小2乗法による多項式関数への回帰

- 次数を大きくすると振動が大きくなる
- 別のテストデータでパラメータの評価が可能
- データ数の増加で振動が少なくなる
- 正則化項をつけると振動が少なくなる
- モデルの複雑さを決めるのは難しい問題