Foundations of Machine Learning
AI2000 and AI5000

FoML-09
Underfitting and Overfitting
Regularized Least Squares

Dr. Konda Reddy Mopuri
Department of AI, IIT Hyderabad
July-Nov 2025

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence
& Learning Lab

So far in FoML

- What is ML and the learning paradigms
- Probability refresher
- MLE, MAP, and fully Bayesian treatment
- Linear Regression with basis functions - geometric interpretation

Under and Over fitting

Linear Regression

- Complex functions can be fit to the data
  - Using the basis functions

Linear Regression with basis functions

- Complex functions can be fit to the data
  - Using the basis functions
- There are some choices (hyper parameters) to be made
  - What kind of basis functions?
  - How many of them?

Linear Regression with basis functions

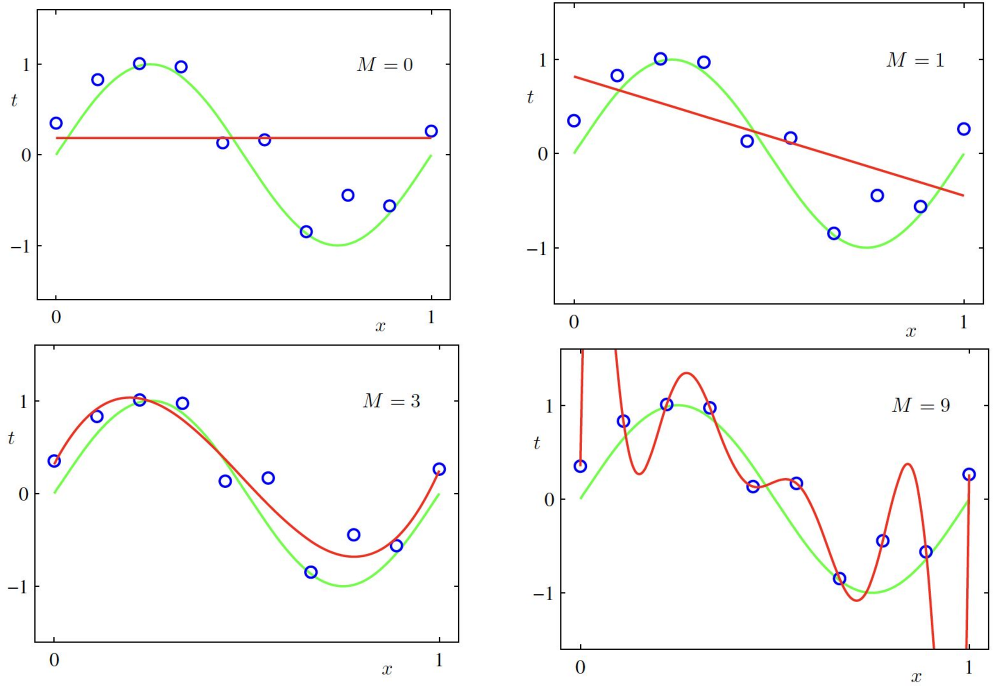
- Complex functions can be fit to the data
    - Using the basis functions
- There are some choices (hyper parameters) to be made
    - What kind of basis functions?
    - How many of them?
- They have consequences
    - over/under fitting

Polynomial basis functions

Target t = sin(2πx) + $\sigma . \epsilon$

Noise ϵ ~ N(0, 1)

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$



How to spot Under/Over fitting?

How to spot Under/Over fitting?

- Can the weight values provide some insight?

How to spot Under/Over fitting?

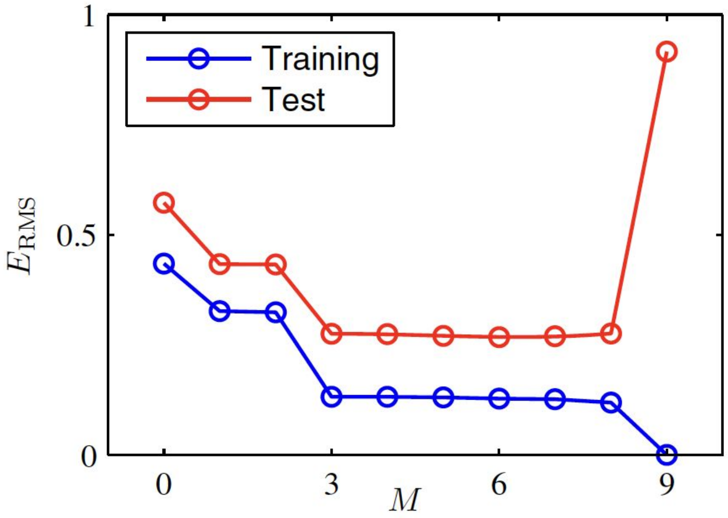
- Can the weight values provide some insight?

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

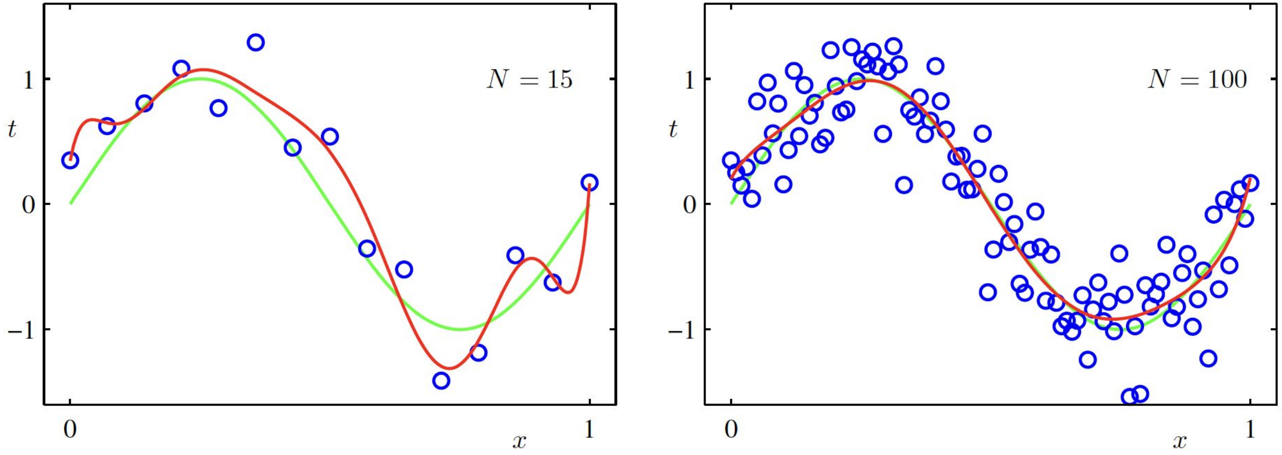
Why didn't our M = 9 model realize a mapping similar to that of M = 3 model?

Better way to spot under/over fitting

- Plot the error (E_{RMS})

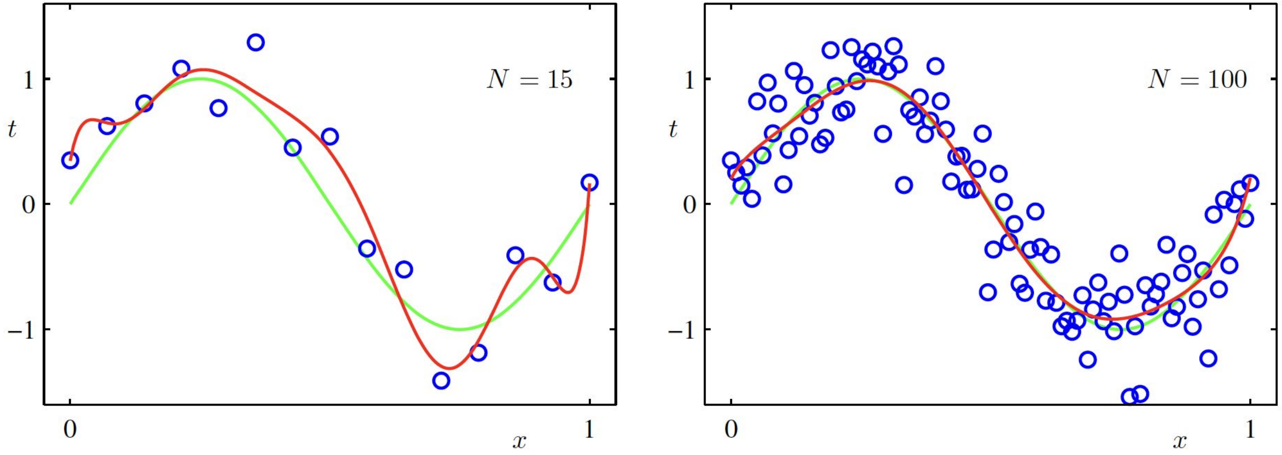


Effect of dataset size on overfitting



M = 9 model fitting to dataset of different sizes

Effect of dataset size on overfitting



M = 9 model fitting to dataset of different sizes

For a given model complexity, the overfitting problem becomes less severe as the dataset size increases

What if it is not easy to collect a lot of data?

- More data helps to avoid overfitting
- But, it may be challenging to collect a lot of data
- Then, what?

Regularization

How to spot Under/Over fitting?

- Large weight values indicate overfitting
- Higher complexity models lead to overfitting

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Regularized least squares

- In case of smaller datasets, instead of manually restricting the number of parameters
    - Add a penalty to avoid large weight values → 'weight decay' regularization

Regularized least squares

- In case of smaller datasets, instead of manually restricting the number of parameters
    - Add a penalty to avoid large weight values → 'weight decay' regularization

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{t_i - \mathbf{w}^T\phi(\mathbf{x}_i)\}^2 +$$

Regularized least squares

- In case of smaller datasets, instead of manually restricting the number of parameters
  - Add a penalty to avoid large weight values

Ridge Regression

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{\mathbf{i}=1}^{\mathbf{N}}\{\mathbf{t_i} - \mathbf{w}^{\mathbf{T}}\phi(\mathbf{x_i})\}^2 + \frac{\lambda}{2}\sum_{\mathbf{i}=1}^{\mathbf{M}-1}\mathbf{w_i}^2$$

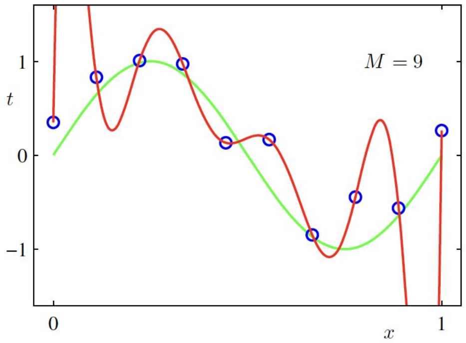
Bias term w_0 may not be included in the regularization

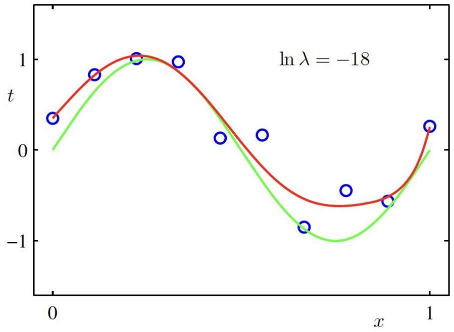
Regularized least squares

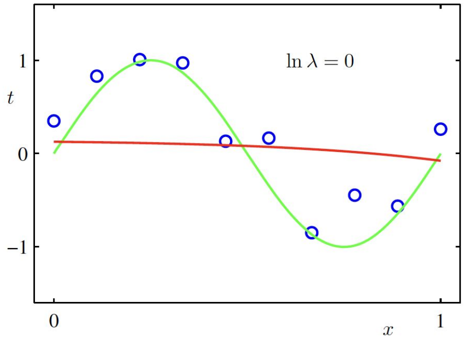
$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{\mathbf{t_i} - \mathbf{w}^{\mathbf{T}}\phi(\mathbf{x_i})\}^2 + \frac{\lambda}{2}\sum_{i=1}^{M-1}\mathbf{w_i}^2$$

- Looks similar to what we saw during the MAP discussion

Regularized Least Squares





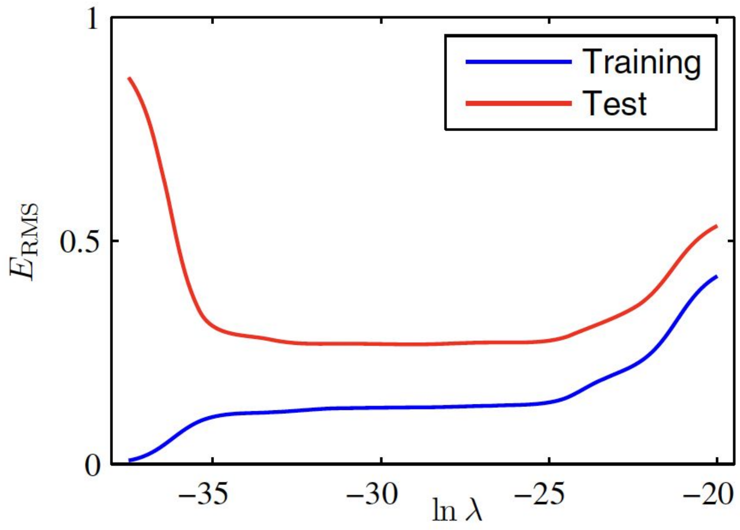


w/o regularization

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Data-driven Intelligence
& Learning Lab

Regularized Least Squares



More general form of the regularization

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{t_i - \mathbf{w}^T\phi(\mathbf{x}_i)\}^2 + \frac{\lambda}{2}\sum_{i=1}^{M-1}|\mathbf{w}_i|^q$$

- When q = 2 → l_2 norm penalty on the parameters

More general form of the regularization

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{t_i - \mathbf{w}^T\phi(\mathbf{x_i})\}^2 + \frac{\lambda}{2}\sum_{i=1}^{M-1}|\mathbf{w_i}|^q$$

- When q = 2 → l_2 norm penalty on the parameters
- When q = 1 → l_1 norm penalty (also, called Lasso)

Geometric interpretation

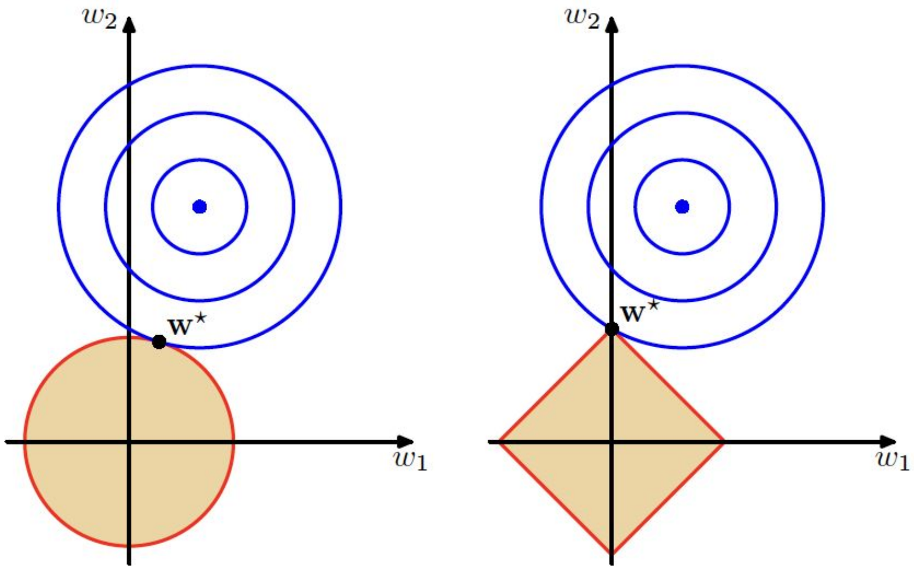
Equivalent to minimizing

$$\frac{1}{2}\sum_{i=1}^{N}\{t_i - \mathbf{w}^T\phi(\mathbf{x}_i)\}^2 \quad \text{with} \quad \sum_{j=1}^{M}|w_j|^q \leq \eta$$

Geometric interpretation

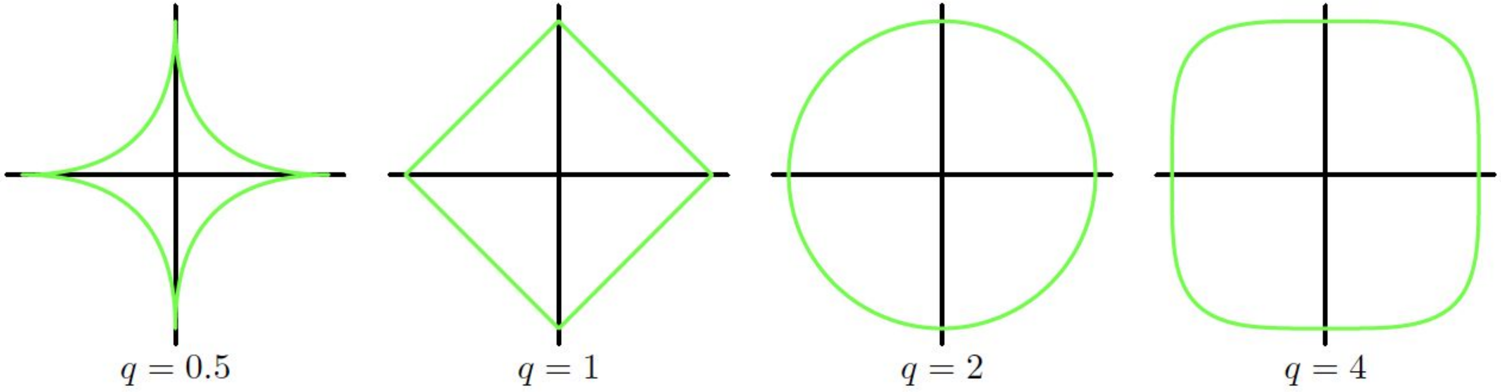
Equivalent to minimizing

$$\frac{1}{2}\sum_{i=1}^{N}\{t_i - \mathbf{w}^T\phi(\mathbf{x}_i)\}^2 \quad \text{with} \quad \sum_{j=1}^{M}|w_j|^q \leq \eta$$



Plots of the 'unregularized' error and the 'regularization' term for q =1 and 2

Geometric interpretation



Contours of regularization term for different values of q

Rough work

Next
Model selection

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Data-driven Intelligence
& Learning Lab