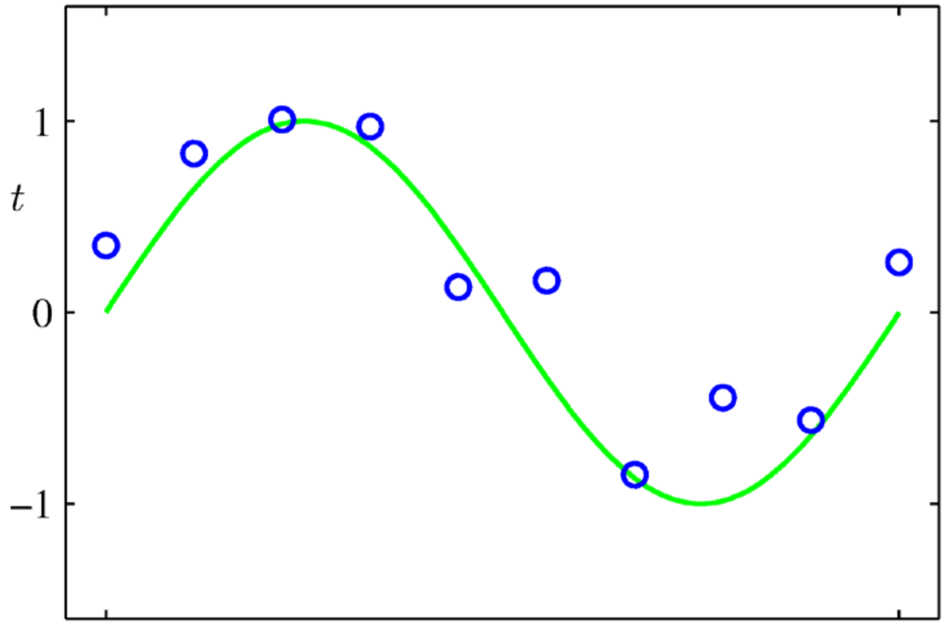
Key concept: Regularization

CS434

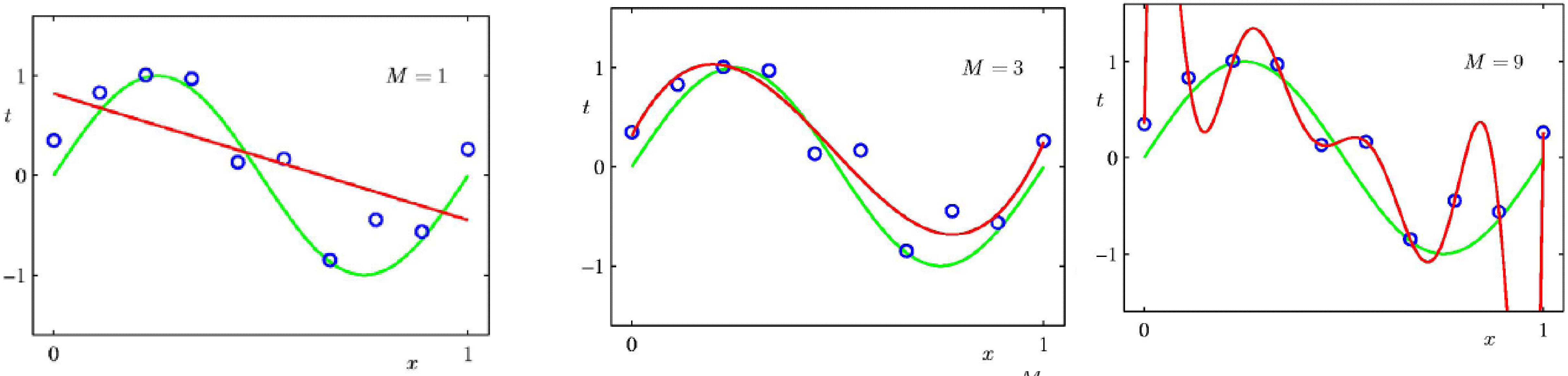
A regression example: Polynomial Curve Fitting



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

- In this example, there is only one real feature x. We learn a function of M-order polynomial
- Achieved by learning a linear regression using $(1, x, x^2, \ldots, x^M)$ as the features.
- Note that this new feature space is derived from the original input x
- Such derived features are often referred to as the basis functions

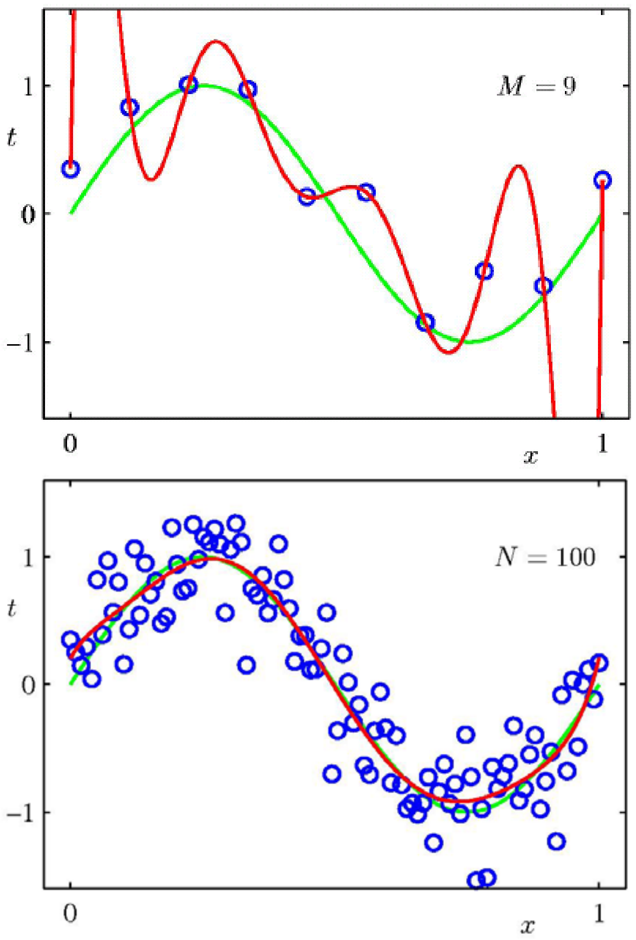
Consider different choices for M



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

- Larger M leads to higher model complexity
- Given 10 data points, if M=9, we can fit the training data perfectly – severely <u>overfitting</u>
    - We fit the training data perfectly but perform terribly for inputs we have not seen in training

Over-fitting issue



- What can we do to curb over-fitting
  - Use less complex model
  - Use more training examples
  - **Regularization**

In linear regression, overfitting can often be characterized by large weights

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

Regularized Linear Regression

- Consider the following loss function:

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

Data term + Regularization term (penalize complex models)

$$\sum_{i=1}^{N} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2 + \lambda \sum_{j=0}^{M} w_j^2$$

Encourage small weight values

	M = 0	M = 1	M = 3	M = 9
w_0	0.19	0.82	0.31	0.35
w_1		-1.27	7.99	232.37
w_2			-25.43	-5321.83
w_3			17.37	48568.31
w_4				-231639.30
w_5				640042.26
w_6				-1061800.52
w_7				1042400.18

L2 Regularized Linear Regression

The new objective combines the SSE loss with a **quadratic regularizer**

$$\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \lambda\sum_{j=0}^{M} w_j^2$$

Or equivalently

$$(X\boldsymbol{w} - \boldsymbol{y})^T(X\boldsymbol{w} - \boldsymbol{y}) + \lambda\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w}$$

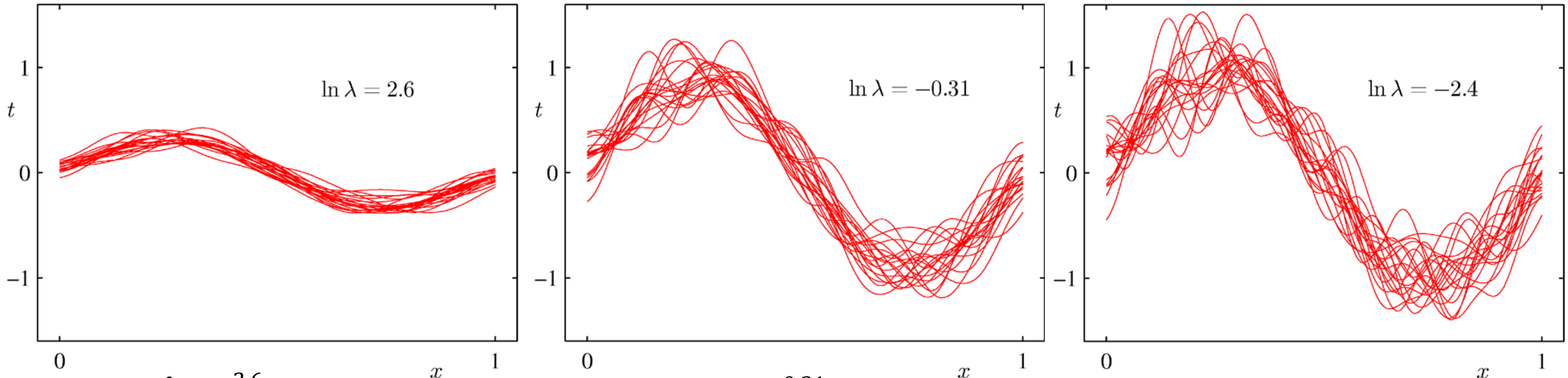
which is minimized by $\boldsymbol{w} = (\lambda I + \mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\boldsymbol{y}$

λ: the regularization coefficient controls the trade-off between model complexity and the fit to the data

- Larger λ encourages simple model (driving more elements of $\mathbf{w}$ to 0)
- Small λ encourages better fit of the data (driving SSE to zero)

Effect of regularization

Fitted curves from 10 random points with M=9. Each curve is fitted with one set of 10 random points.



$\lambda = e^{2.6} \approx 13.46$ $\lambda = e^{-0.31} \approx 0.73$ $\lambda = e^{-2.4} \approx 0.09$

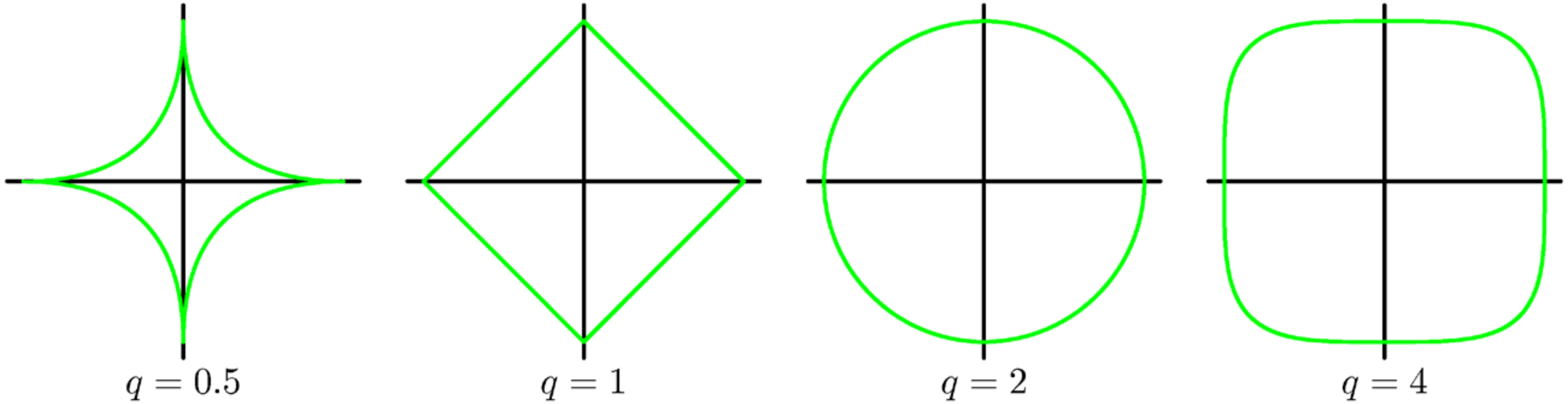
Smaller $\lambda \rightarrow$ more complex curves with achieve closer fit for each set but more overfitting

More Regularization functions

$$\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \lambda\sum_{j=0}^{M}|w_j|^q$$

Equivalent to minimizing SSE subject to $\sum_{i=0}^{M}|w_i|^q \leq \epsilon$

A good explanation of this equivalence is provided here:
http://math.stackexchange.com/questions/335306/why-are-additional-constraint-and-penalty-term-equivalent-in-ridge-regression



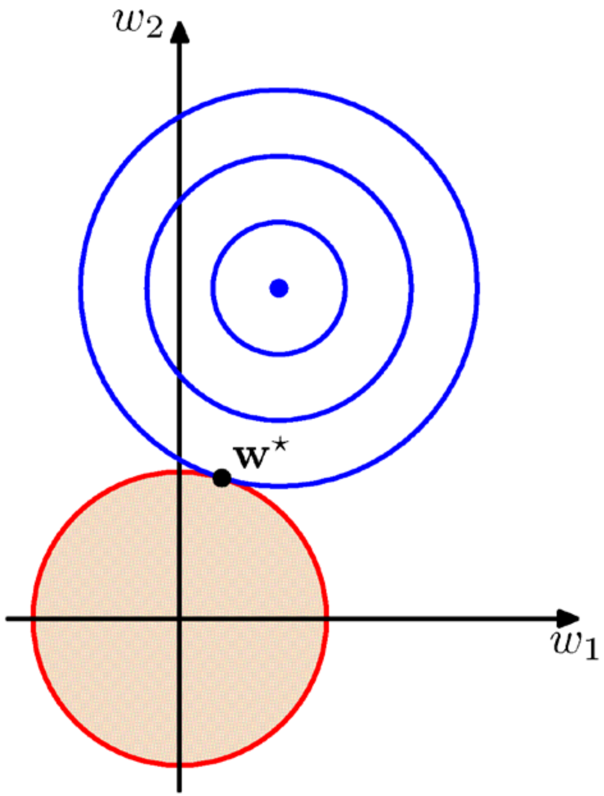
Lasso, L-1

L-2

Shape is determined by q, size determined by λ

Regularized Linear Regression

- Lasso ($q = 1$) tends to generate sparser solutions (majority of the weights shrink to zero) than a quadratic regularizer ($q = 2$, often called ridge regression).



Commonly used regularizers

- L-2 regularization $$\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \lambda \sum_{j=0}^{M} {w_j}^2$$

  Poly-time close-form solution
  Curbs overfitting but does not produce sparse solution

- L-1 regularization $$\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \sum_{j=0}^{M} |w_j|$$

  Poly-time approximation algorithm
  Sparse solution – potentially many zeros in $\mathbf{w}$

- L-0 regularization $$\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \sum_{j=0}^{M} I(w_j \neq 0)$$

  Seek to identify optimal feature subset
  NP-complete problem!

More general use of regularization

- More generally, for a learning task, lets say our parameter is $\boldsymbol{w}$, and the objective is to minimize a loss function $L(\boldsymbol{w})$
- Adding regularization:

$$\min L(w) + \lambda \ \cdot \text{regularizer}$$

- Most commonly used regularizer are norm-based: L_2 and L_1 norm of the weight vector
- Similar trend with changing λ
  - Larger λ leads to simpler model and reduced fit to the training data
  - Smaller λ leads to more complex model and improved fit to training but increase chance of overfitting