# Regularized Least Squares (1)

Slides from Bishop
t instead of y,
Φ(x) instead of x

*Regularization* penalizes complexity and reduces variance (but increases bias)

Adds a term to the squared error

With the sum-of-squares error function and a quadratic regularizer, we get (Bishop uses *t* instead of *y*)

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

λ is called the regularization coefficient.

which is minimized by

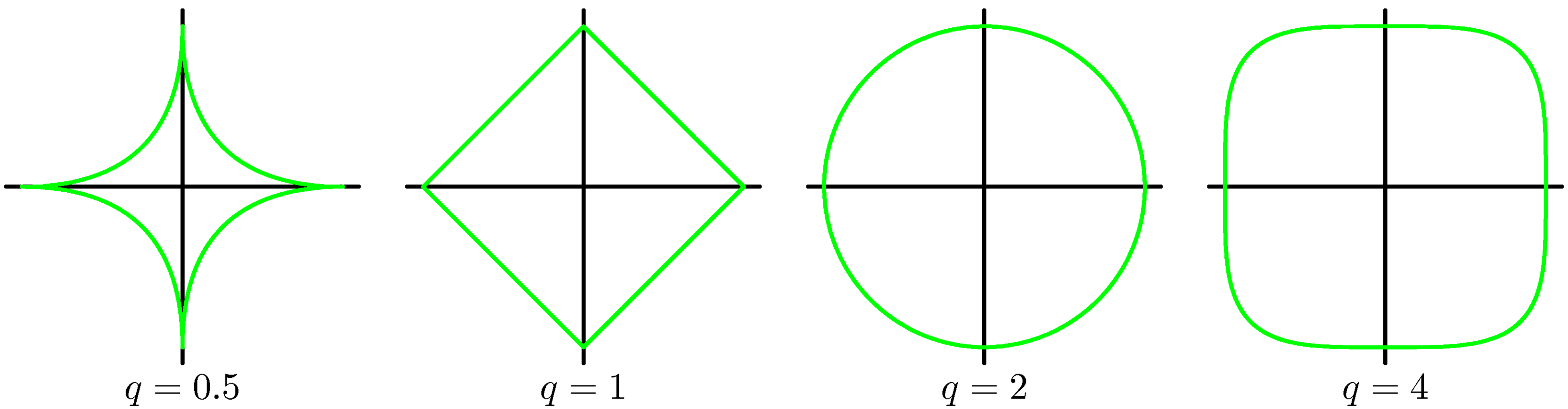
$$\mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}.$$

Some people view regularization as a prior on **w**'s

# Regularized Least Squares (2)

With a more general regularizer, we have

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

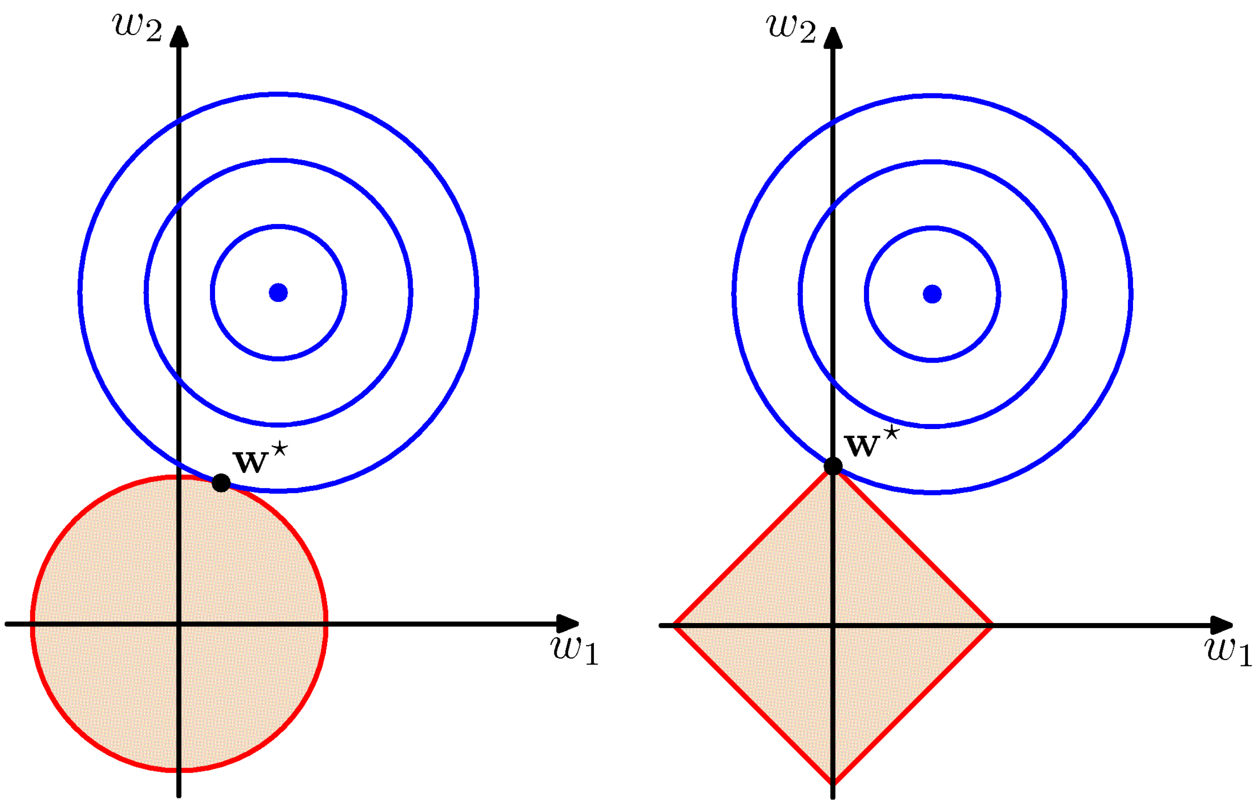



Lasso, $L_1$

Quadratic, $L_2$

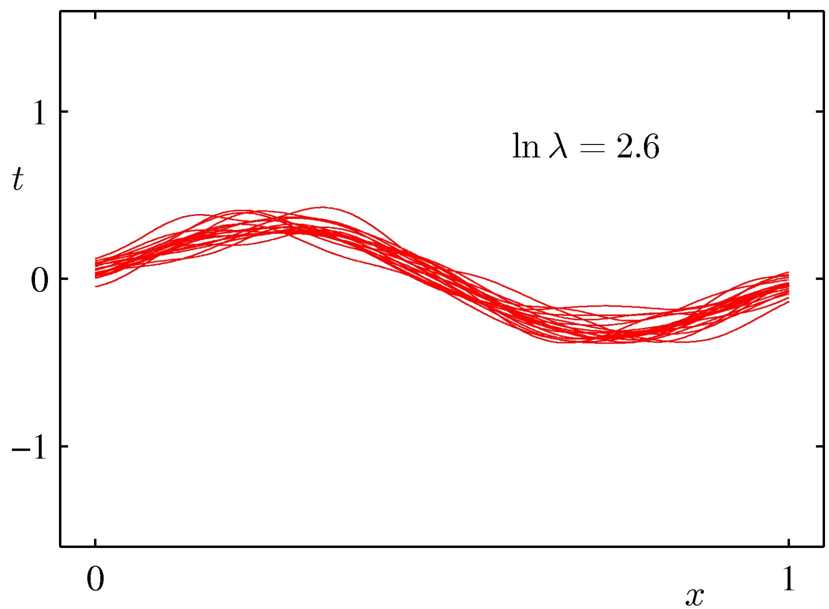
# Regularized Least Squares (3)

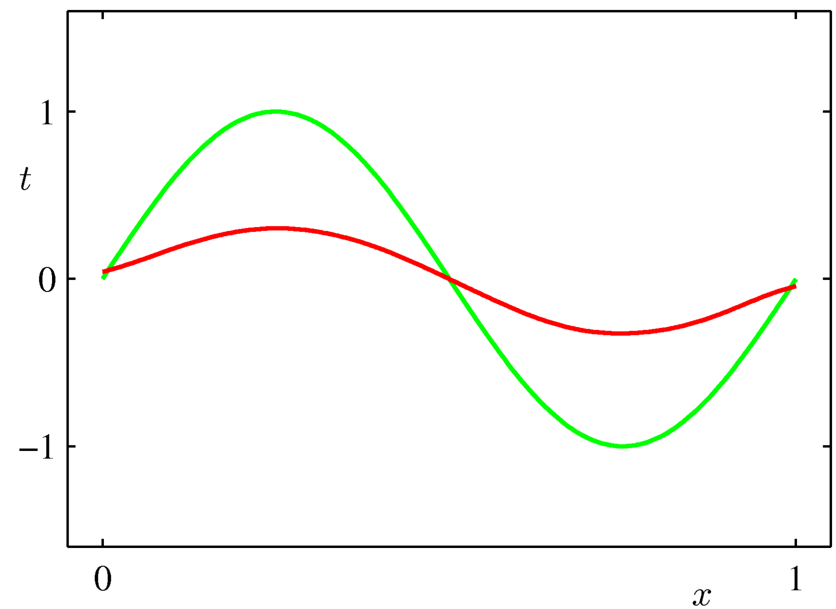
Lasso tends to generate sparser solutions than a quadratic regularizer.



# The Bias-Variance Decomposition (5)

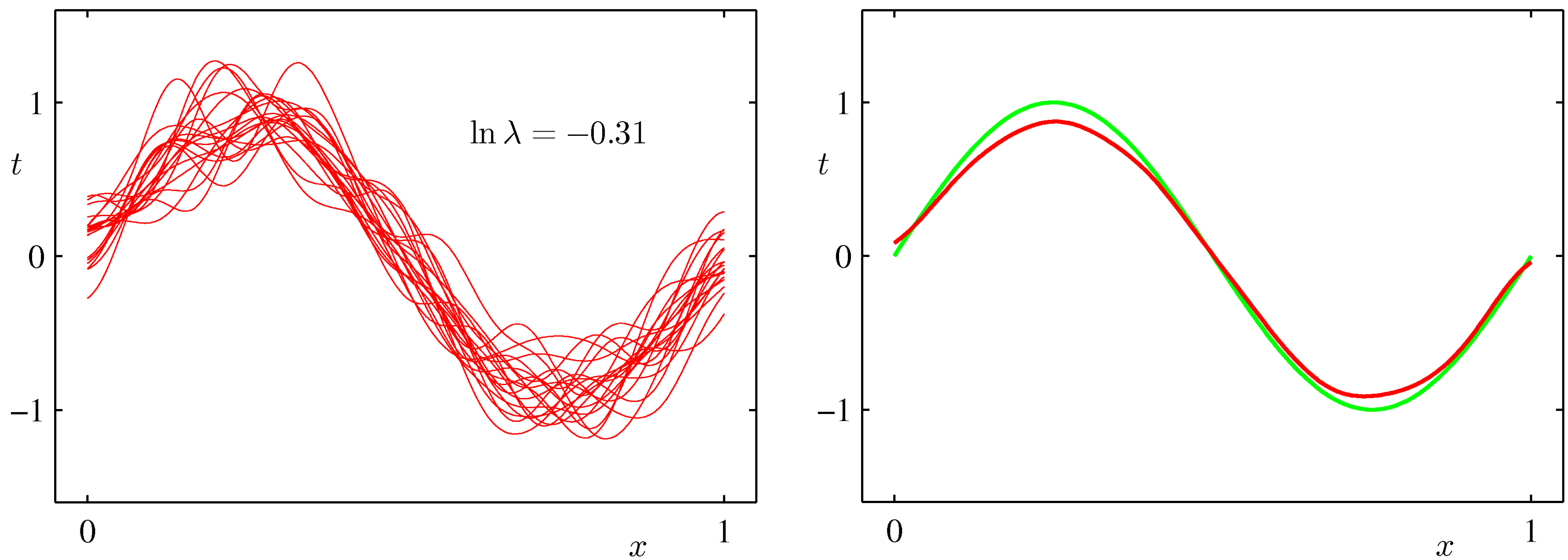
Example: data sets from the sinusoidal, varying the degree of regularization, $\lambda$.



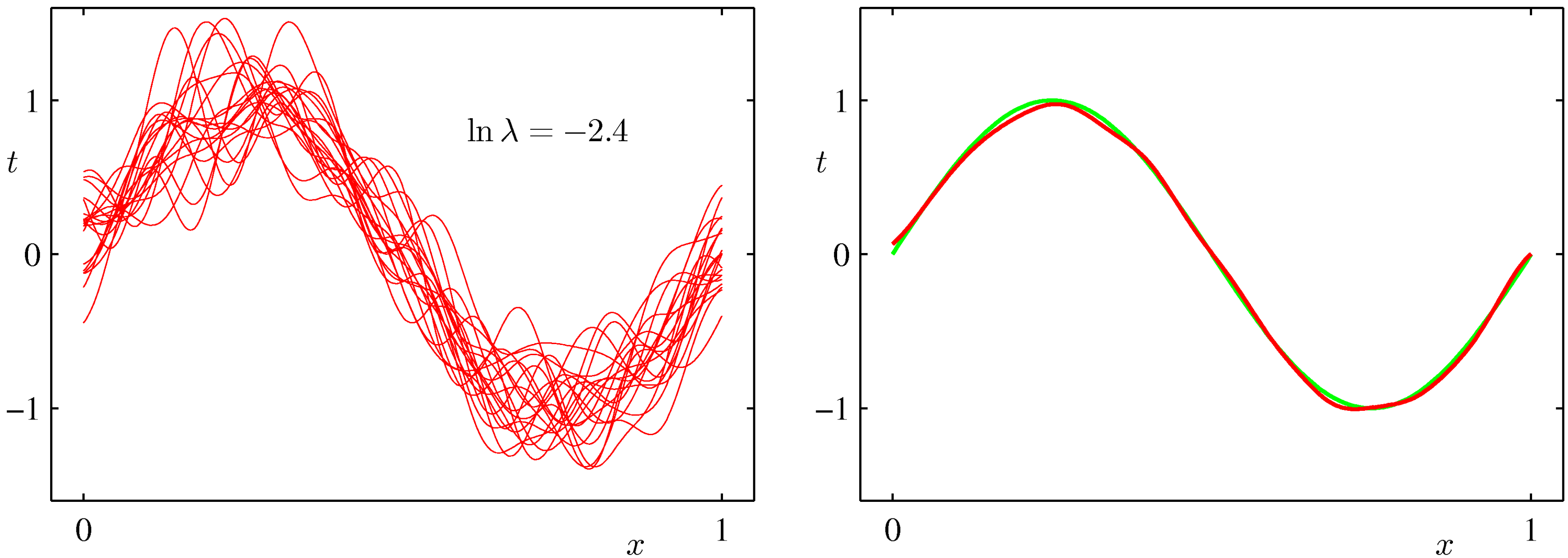
# The Bias-Variance Decomposition (6)

Example: 25 data sets from the sinusoidal, varying the degree of regularization, ¸.



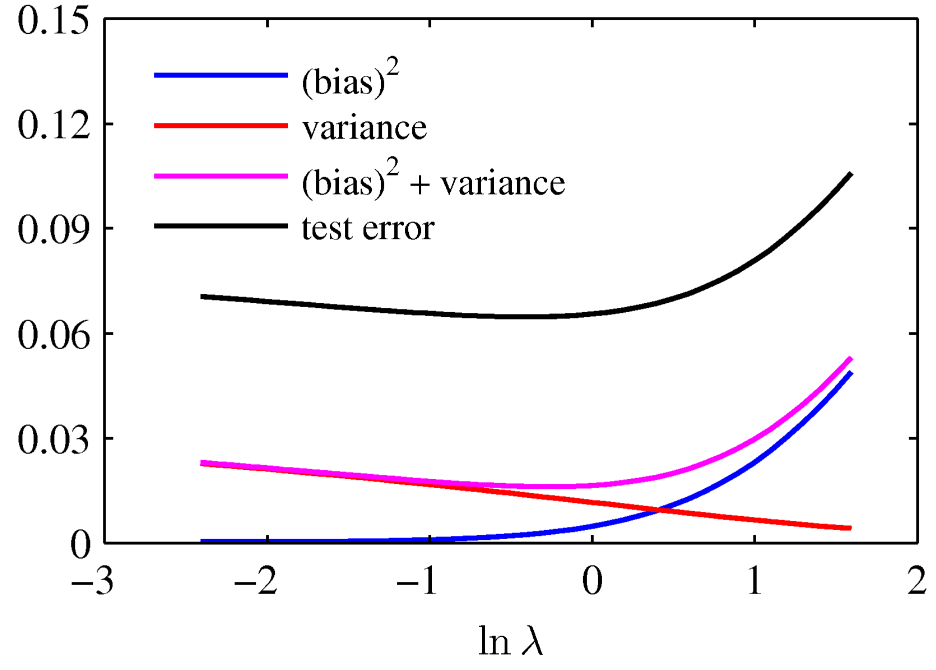
# The Bias-Variance Decomposition (7)

Example: 25 data sets from the sinusoidal, varying the degree of regularization, ¸.



# The Bias-Variance Trade-off

From these plots, we note that an over-regularized model (large $\lambda$) will have a high  bias, while an under-regularized model (small $\lambda$) will have a high variance.



# Main Points

- Least squares is maximum likelihood
- How to finding Maximum-likelihood weights (pseudo-inverse, LMS)
- Regularized Least Squares
- Bias-Variance decomposition
  - Flexibility vs Stability tradeoff