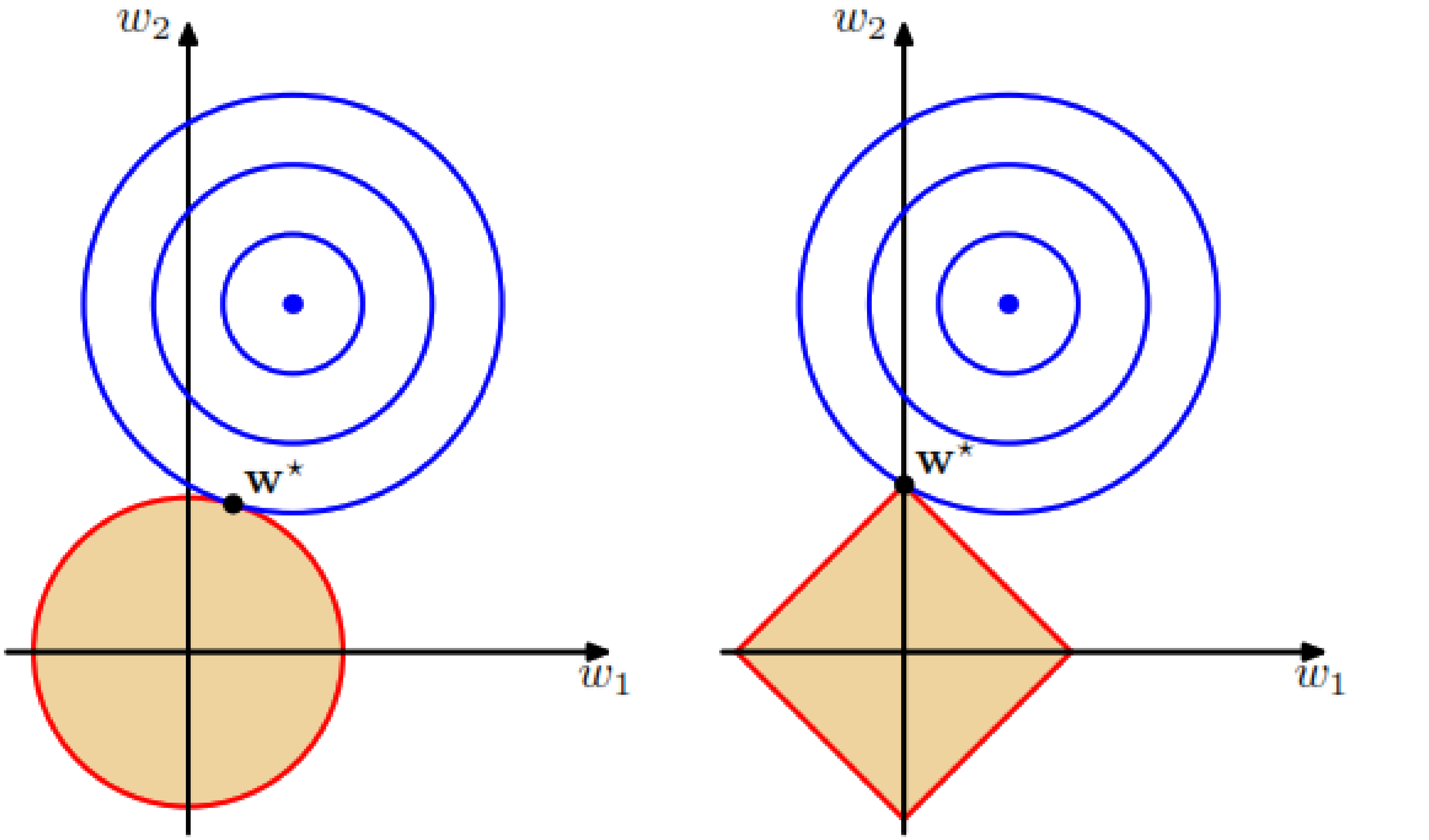
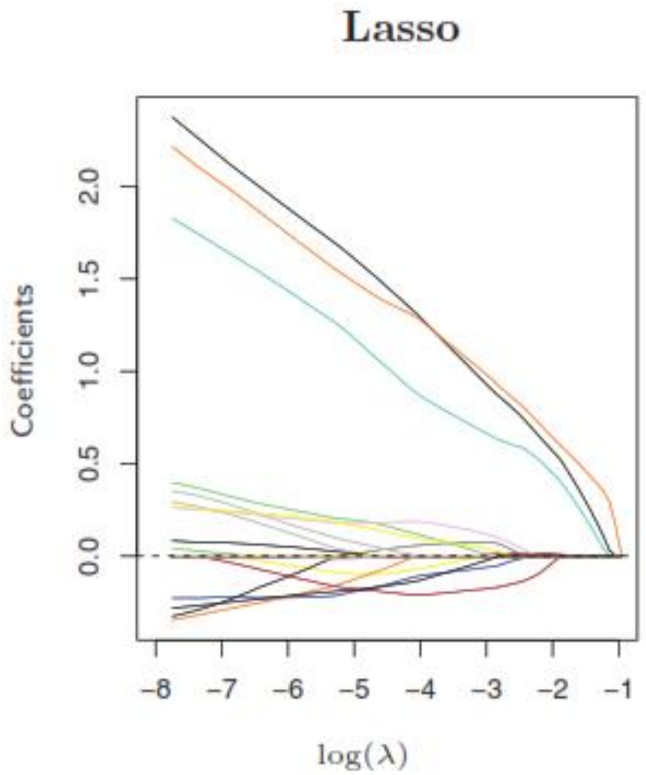
Contours of unregularized error function



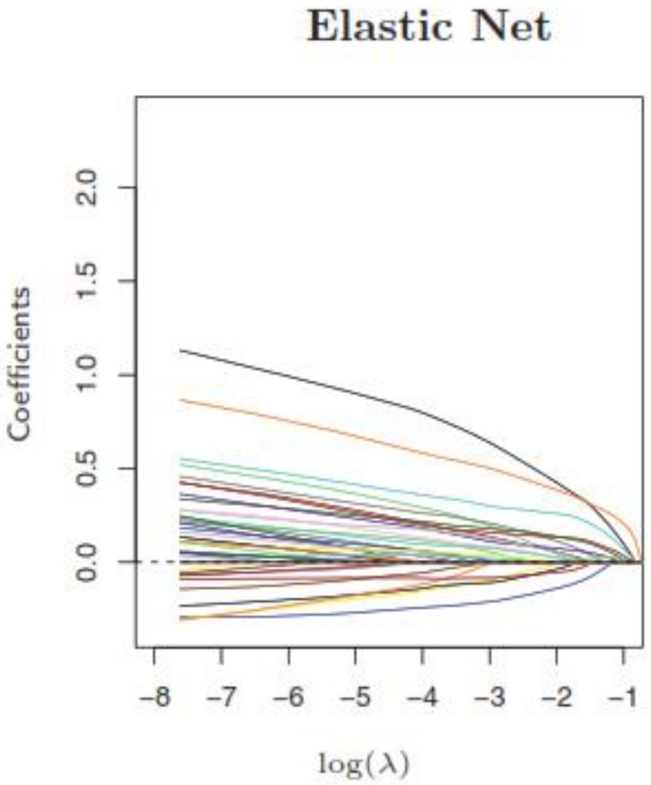
[Bishop]

Elastic net

$$\hat{\beta} = \arg\min_{\beta} (Y - X\beta)^T (Y - XB) + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2)$$



19 non-zero coefficients



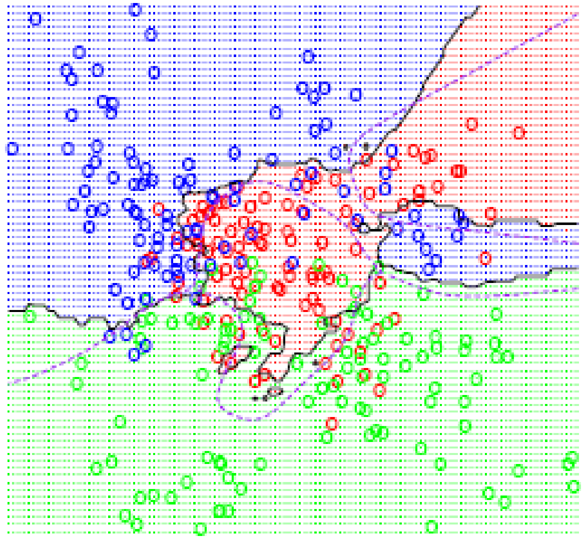
39 non-zero coefficients, but with smaller magnitudes

ESL, fig.18.5

Machine Learning

Lecture 5: Regularization; Bias-Variance tradeoff

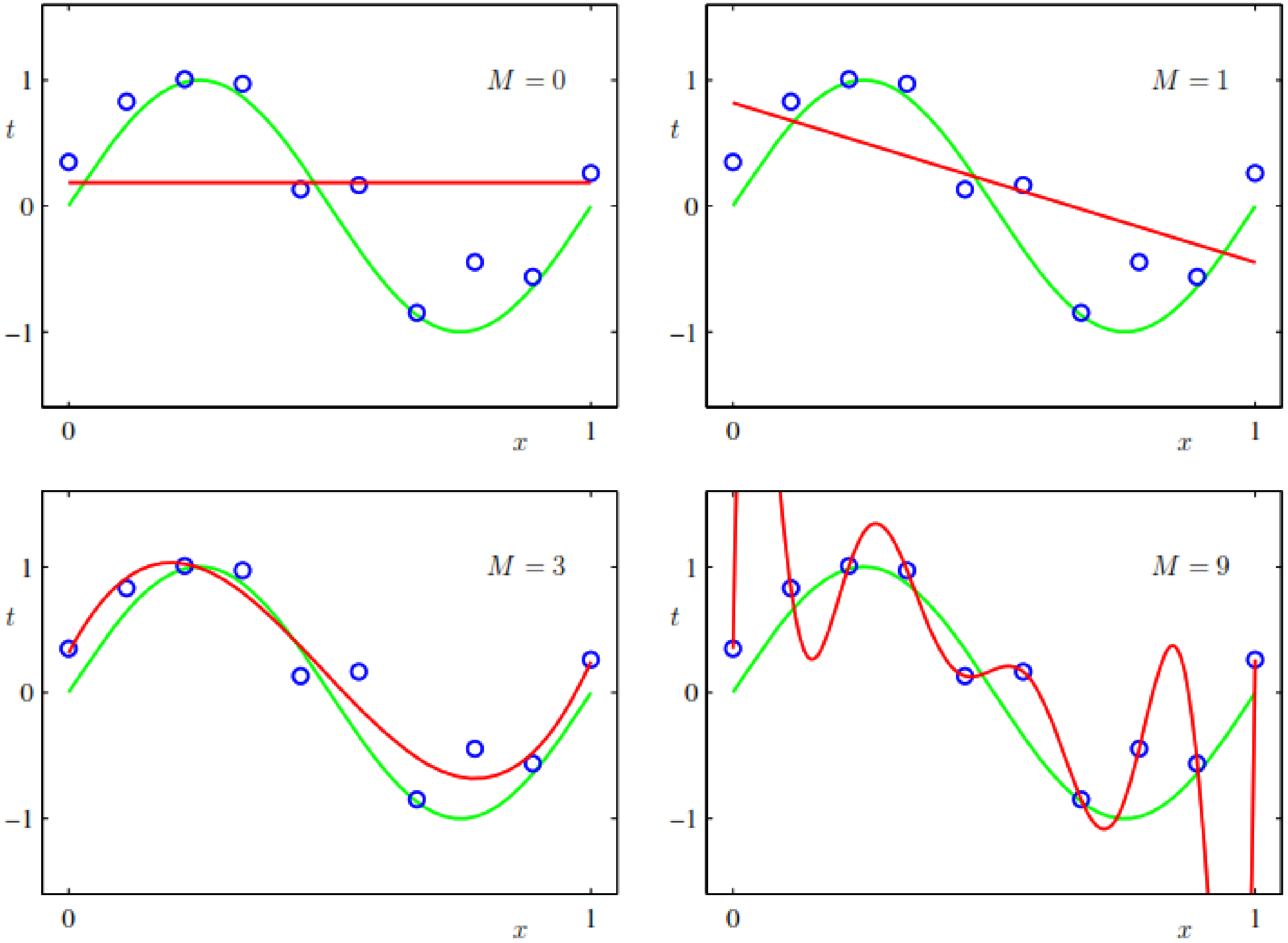
The lectures are mainly offered on white board accompanied by some slides.

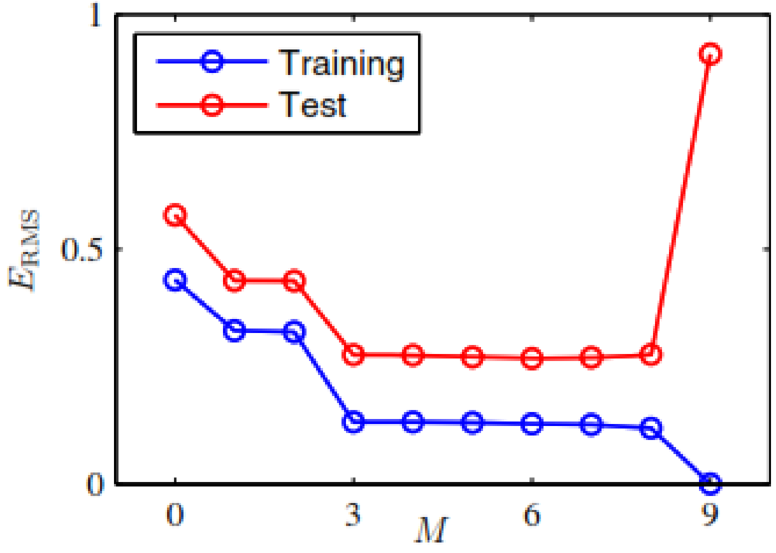


Hesam Montazeri
Department of Bioinformatics, IBB, University of Tehran

Mehr 14, 1398

Polynomials having various orders M





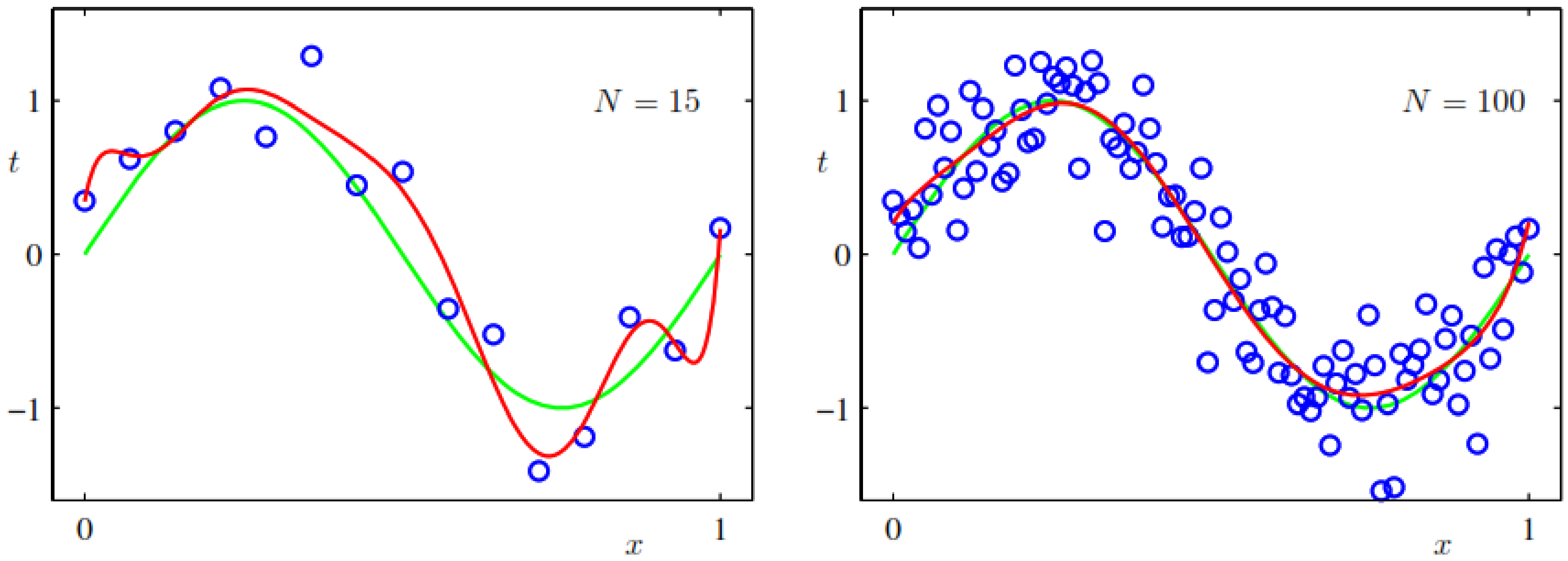
[Bishop]

Magnitude of the coefficients increases with p

	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^\star$	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

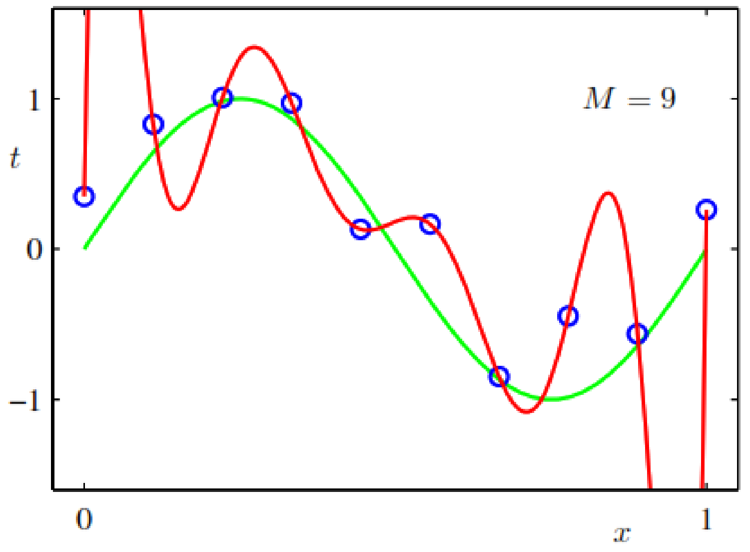
[Bishop]

The increasing size of the data set reduces the over-fitting problem

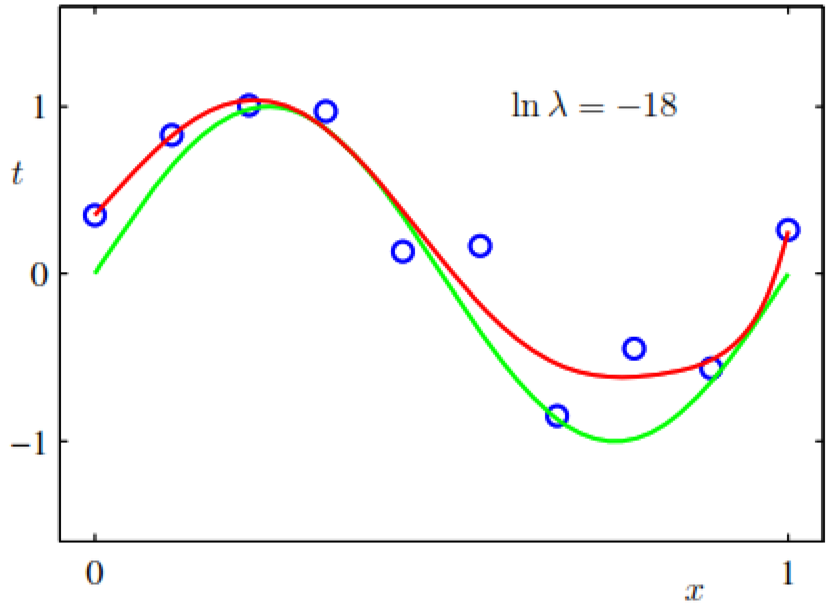


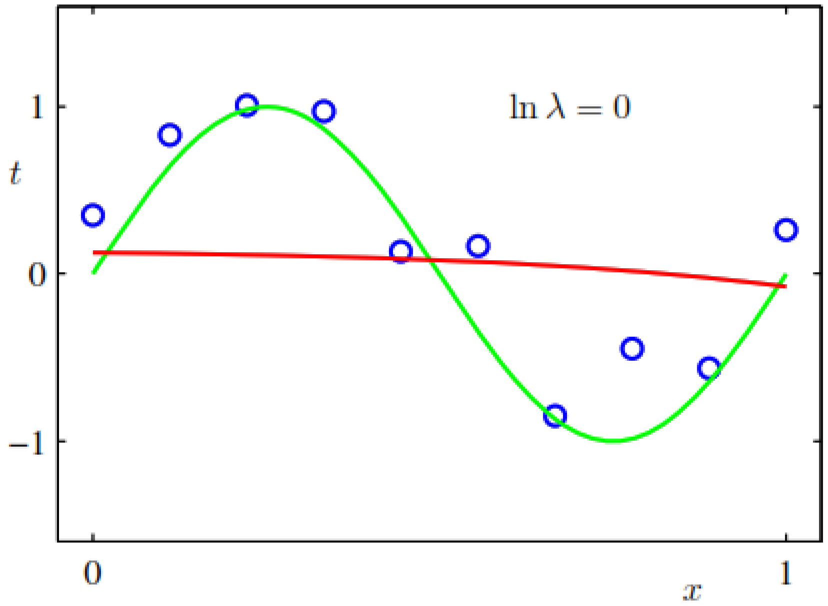
[Bishop]

Regularized error function

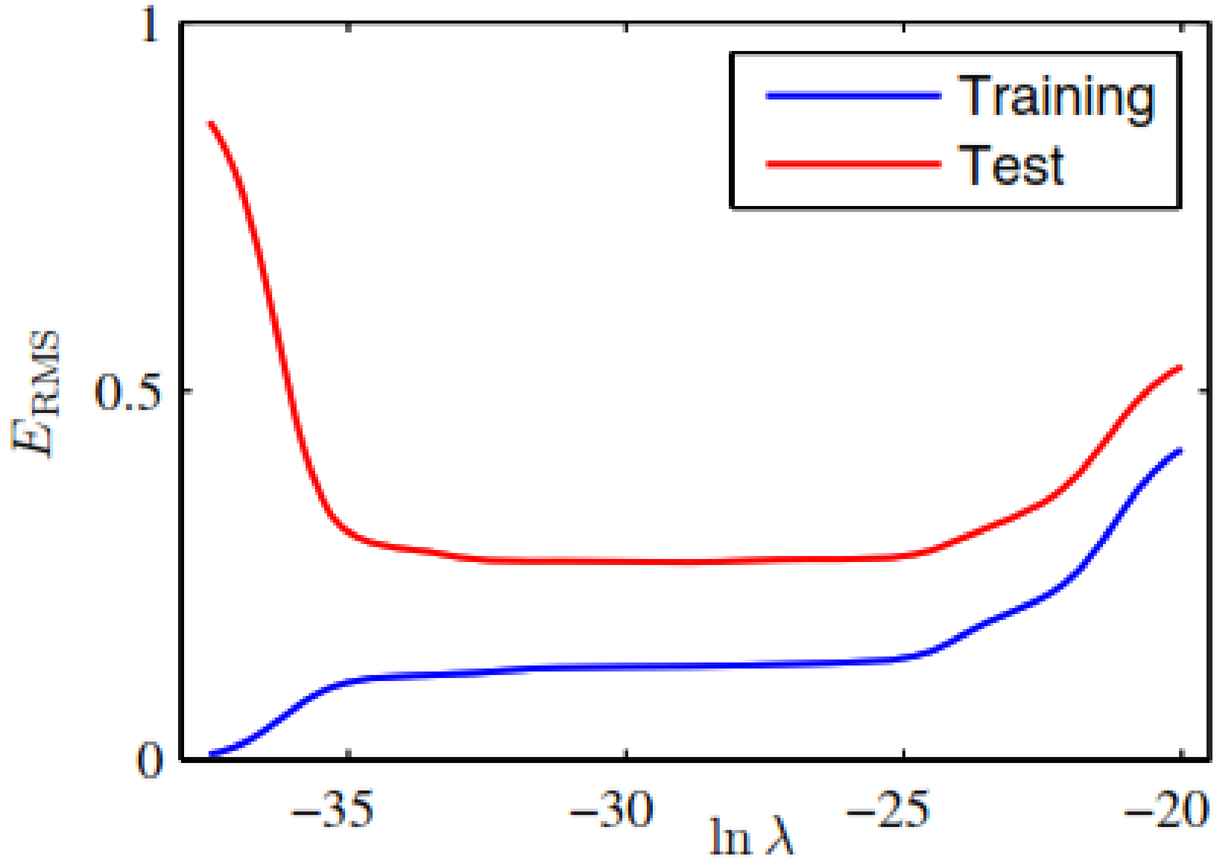


	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01





[Bishop]



[Bishop]

Bias-variance decomposition

- Whiteboard notes

Example: sine target

$f : [-1, 1] \to \mathbb{R}$ $\qquad f(x) = \sin(\pi x)$

Only two training examples! $\quad N = 2$

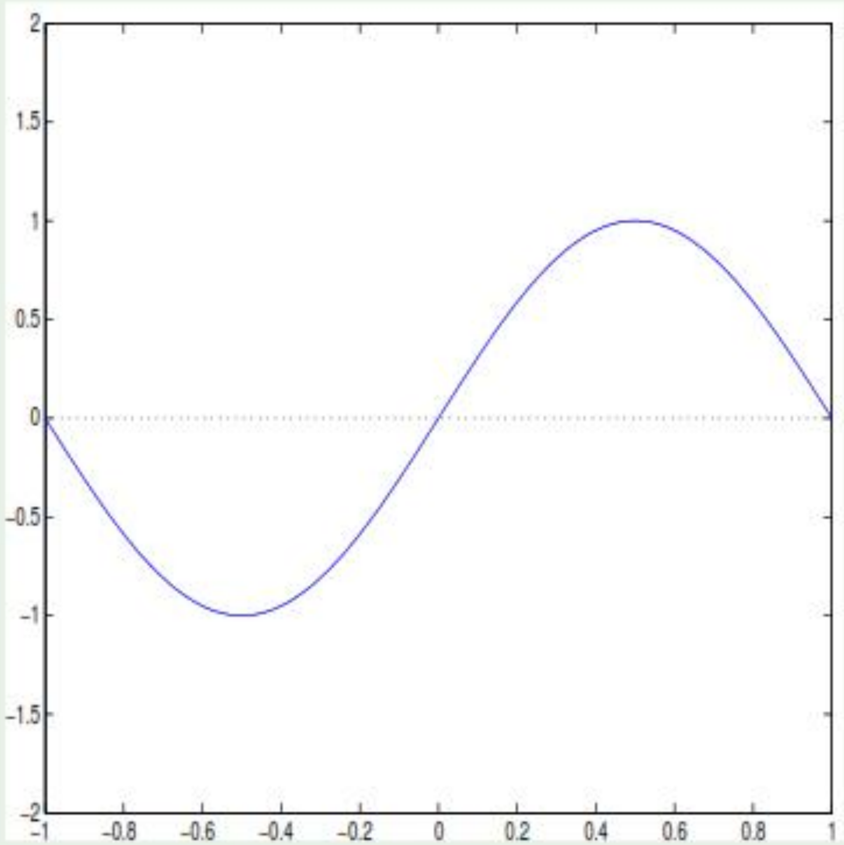
Two models used for learning:

$$\mathcal{H}_0: \quad h(x) = b$$

$$\mathcal{H}_1: \quad h(x) = ax + b$$

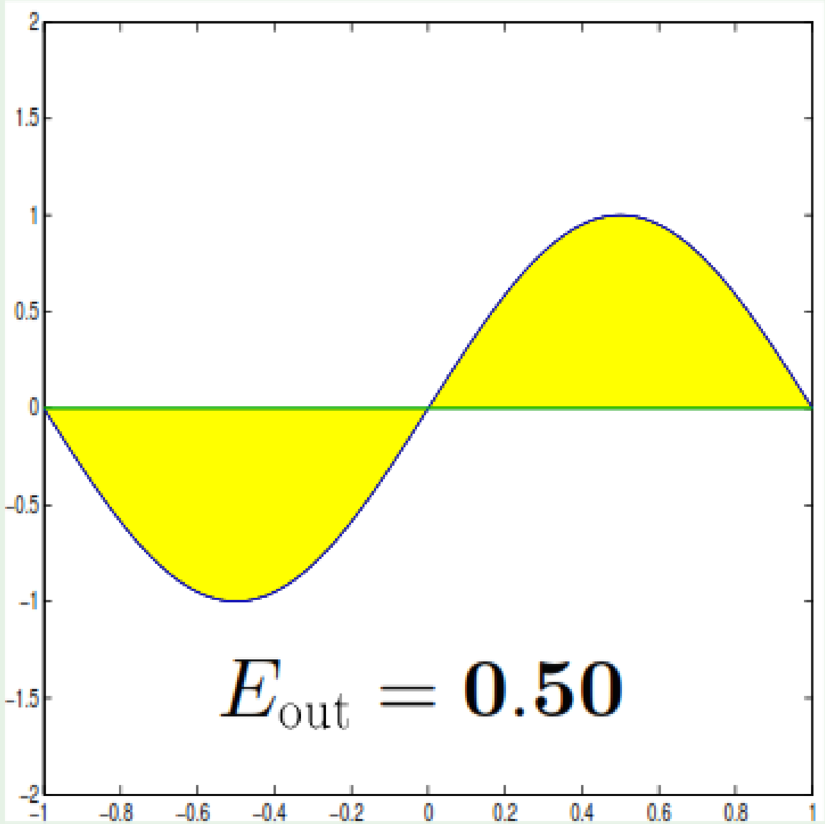
Which is better, $\mathcal{H}_0$ or $\mathcal{H}_1$?

f

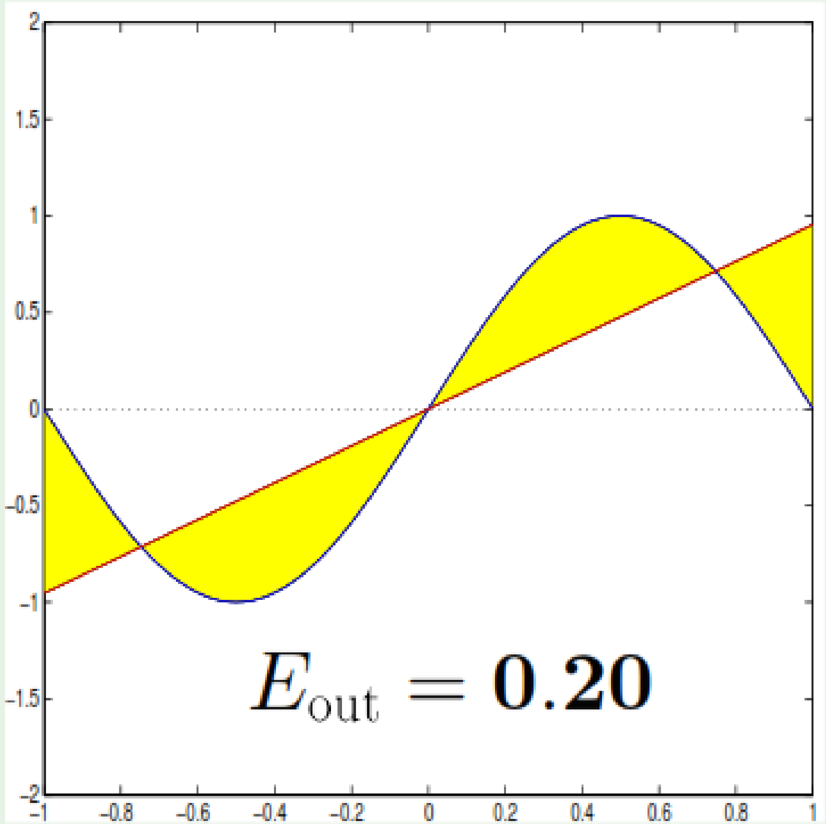


Approximation - $\mathcal{H}_0$ versus $\mathcal{H}_1$

$\mathcal{H}_0$



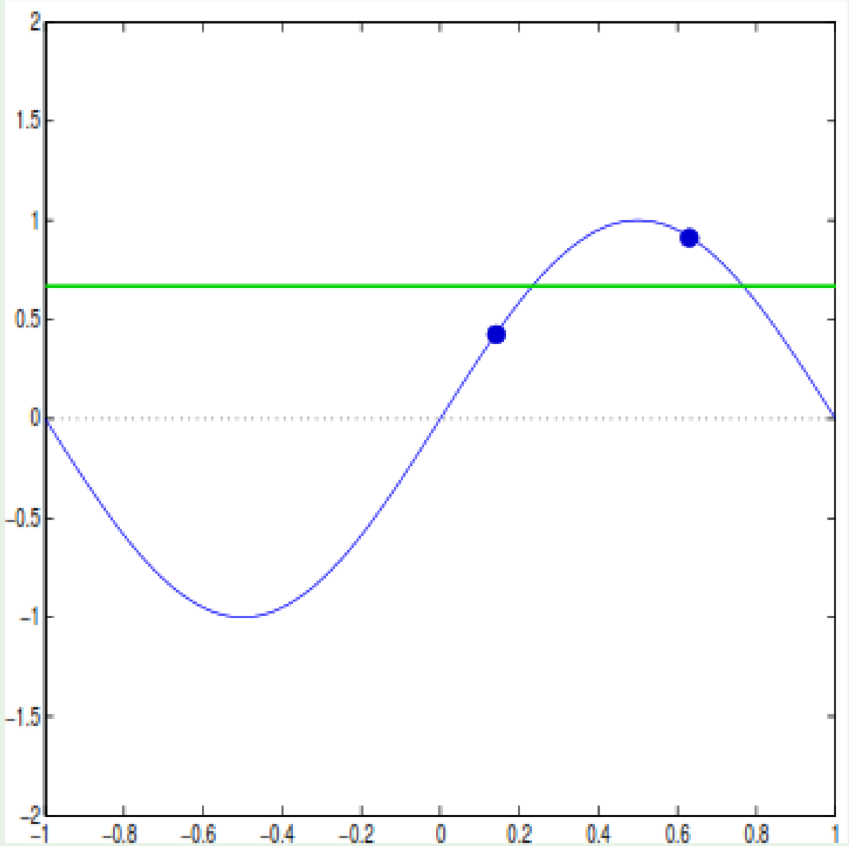
$\mathcal{H}_1$



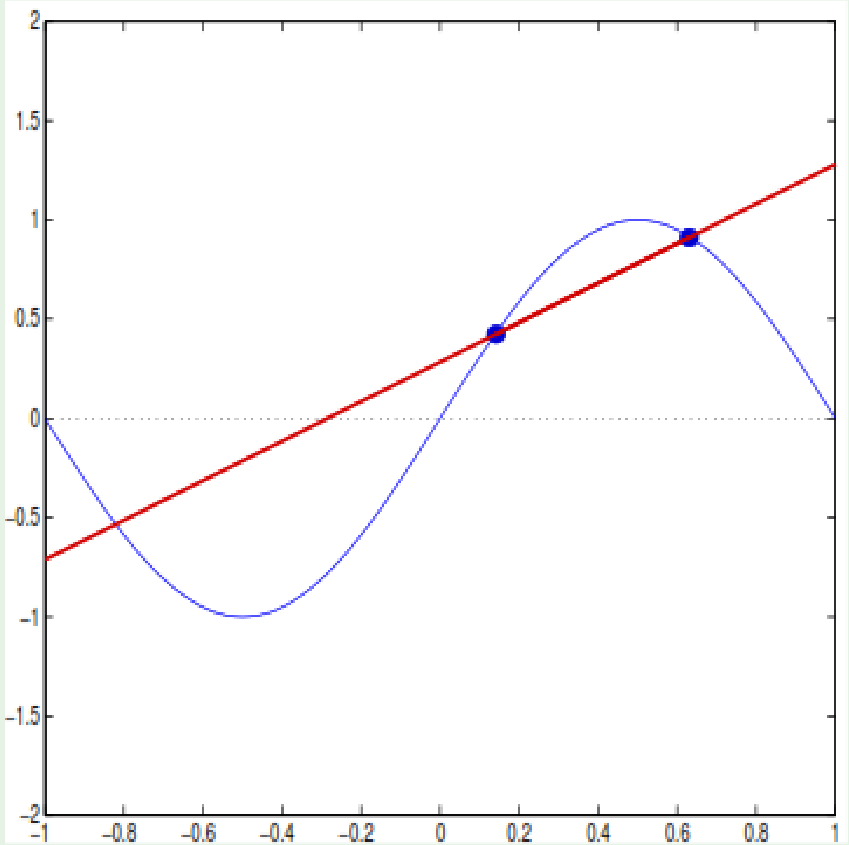
Slides from https://work.caltech.edu/lectures.html

Learning - $\mathcal{H}_0$ versus $\mathcal{H}_1$

$\mathcal{H}_0$

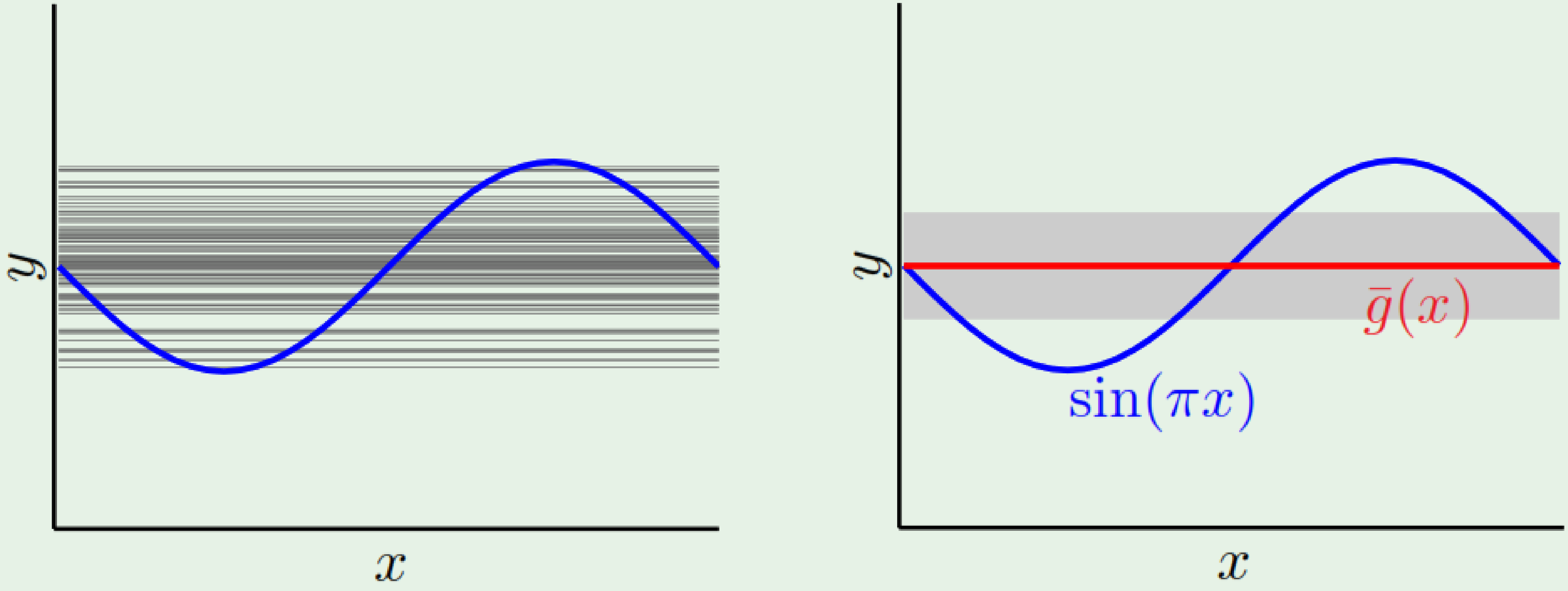


$\mathcal{H}_1$

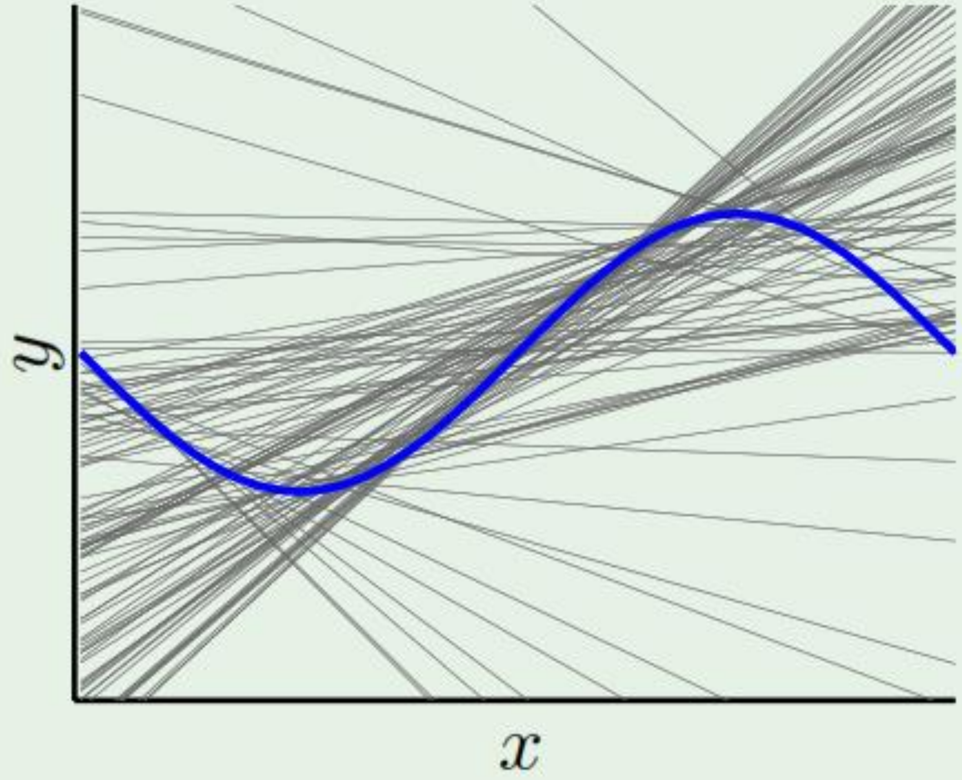


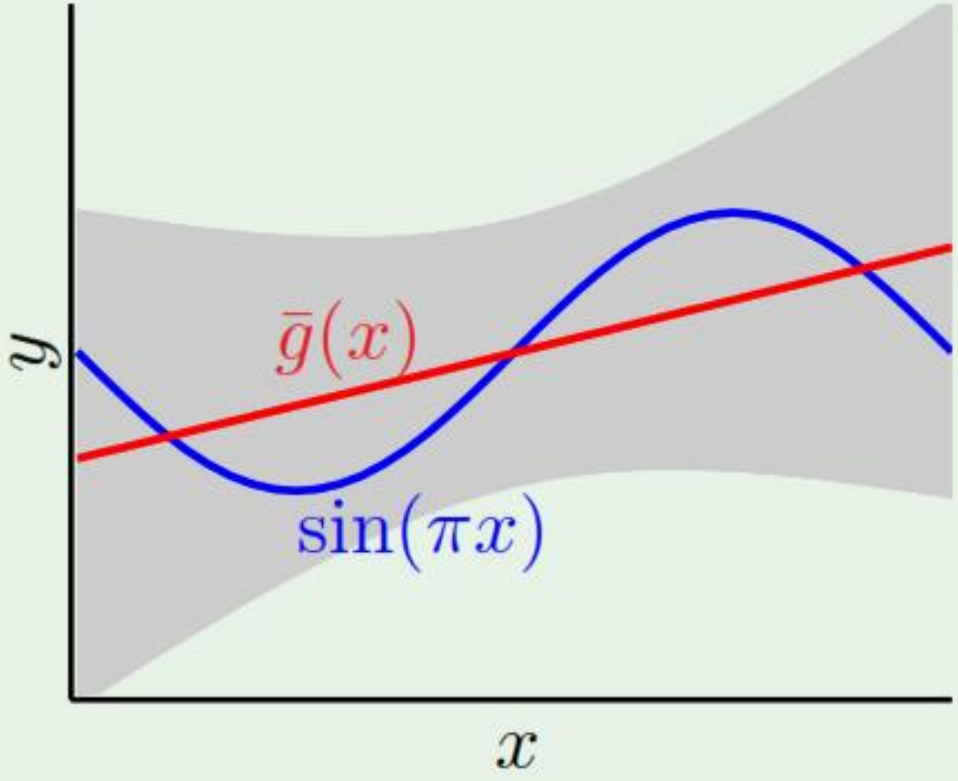
Slides from https://work.caltech.edu/lectures.html

Bias and variance - $\mathcal{H}_0$

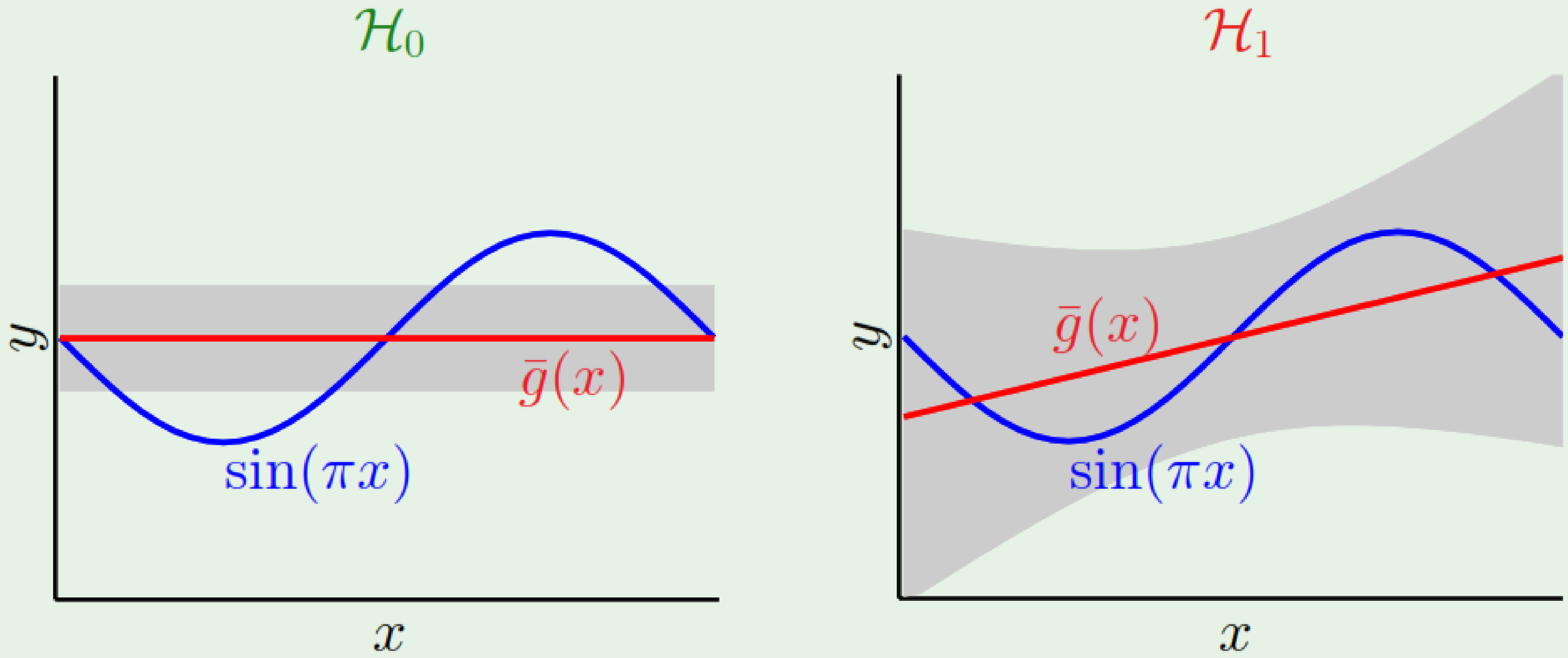


Bias and variance - $\mathcal{H}_1$





and the winner is ...

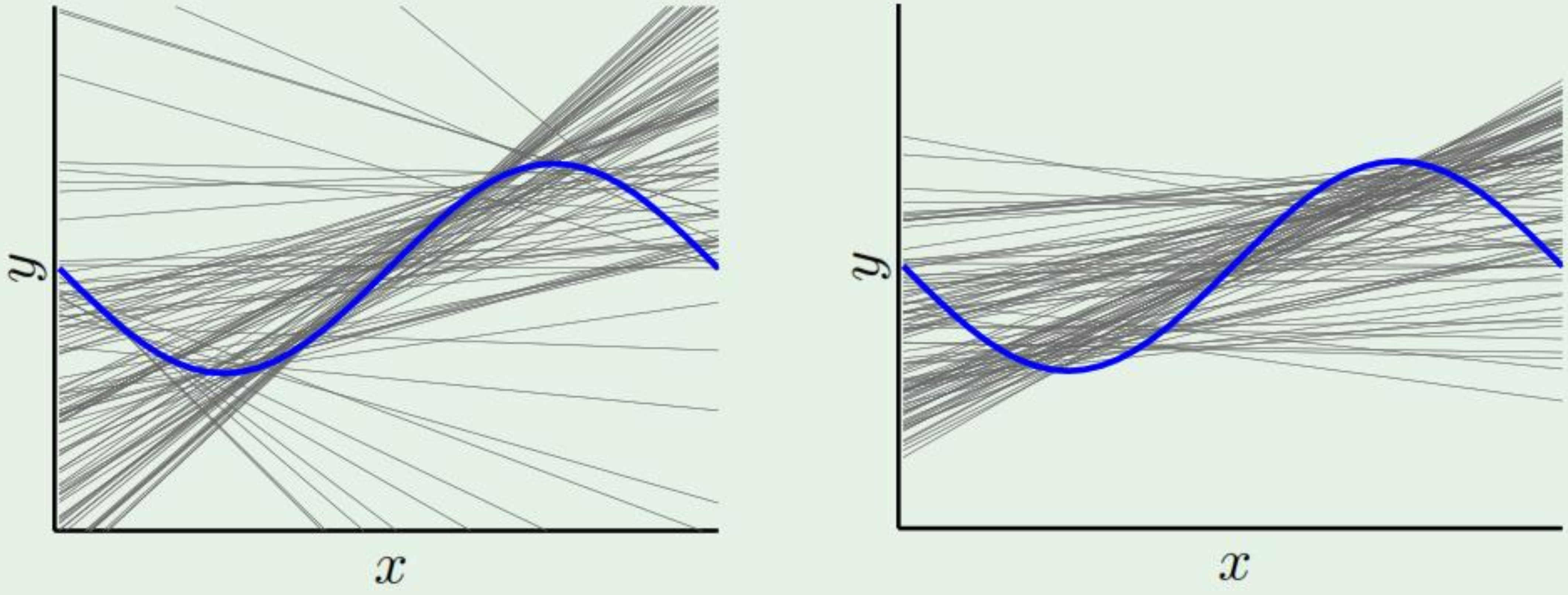


bias $= \mathbf{0.50}$ var $= \mathbf{0.25}$

bias $= \mathbf{0.21}$ var $= \mathbf{1.69}$

Slides from https://work.caltech.edu/lectures.html

A familiar example

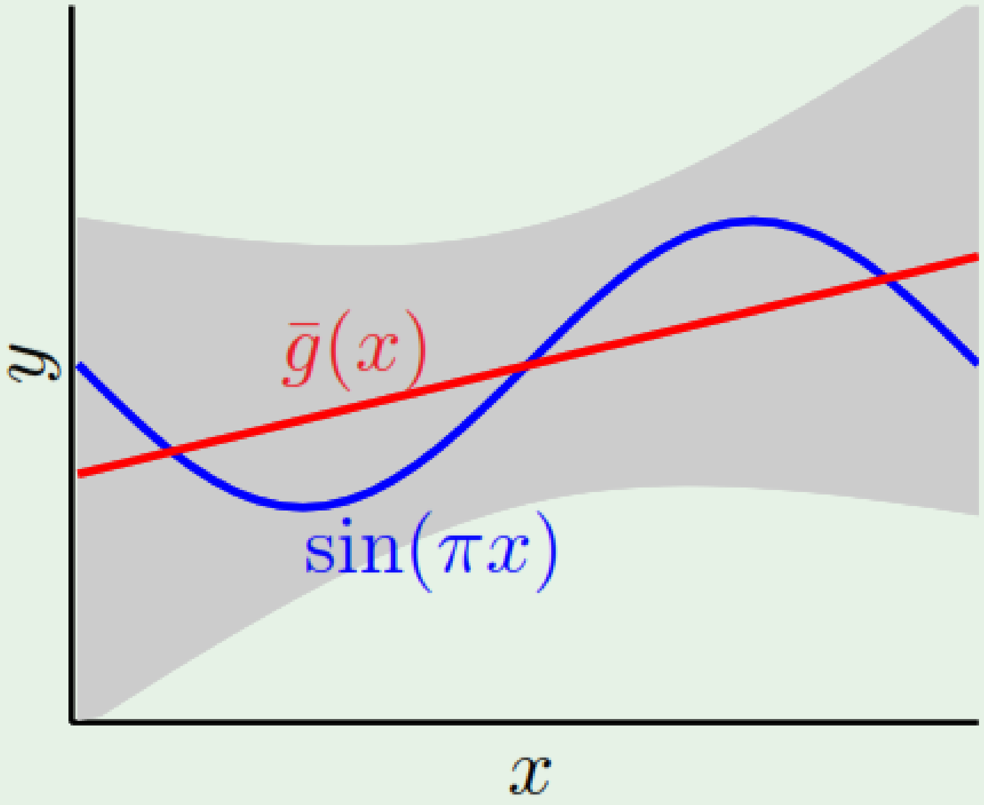


without regularization

with regularization

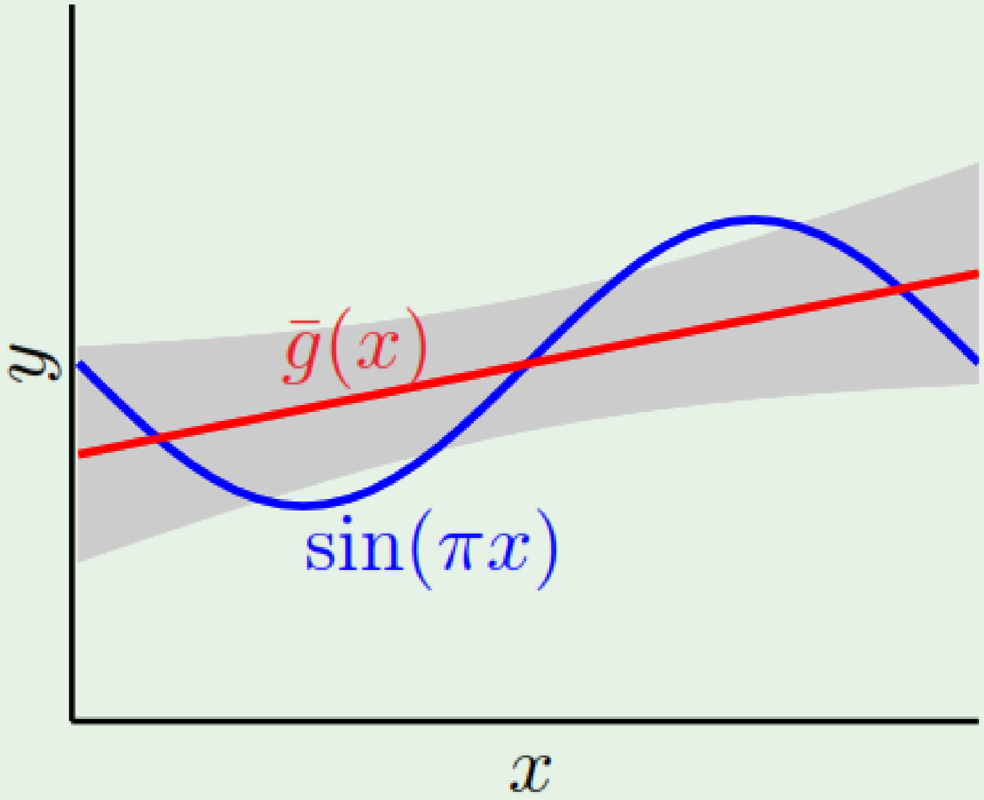
and the winner is . . .

without regularization



bias = $\mathbf{0.21}$ var = $\mathbf{1.69}$

with regularization



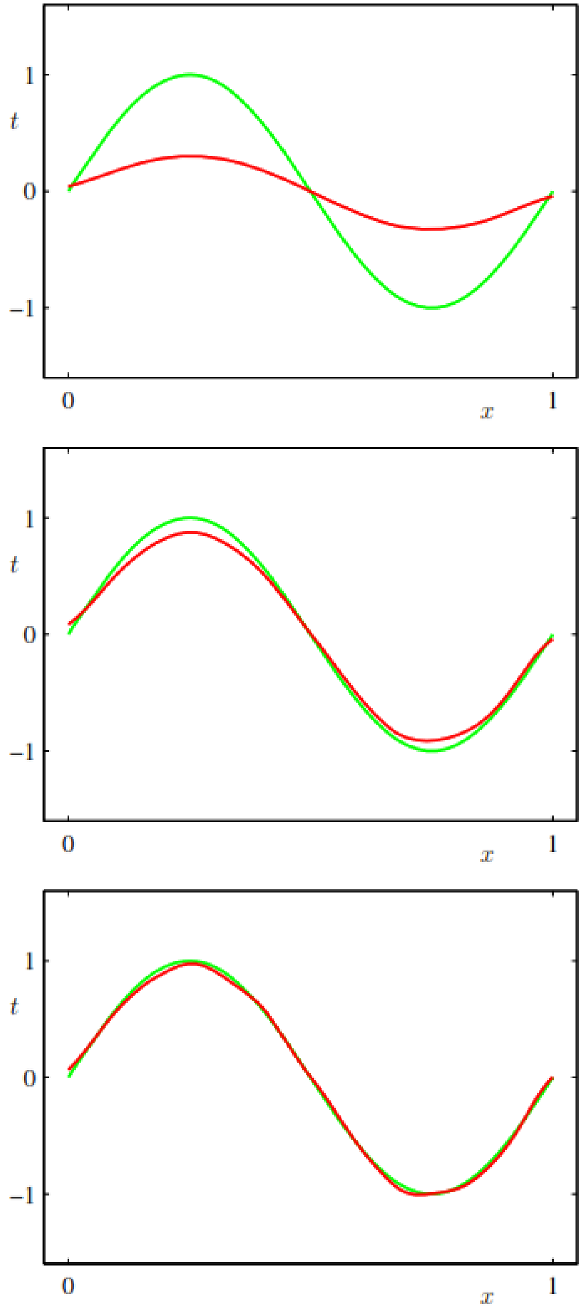
bias = $\mathbf{0.23}$ var = $\mathbf{0.33}$

Slides from https://work.caltech.edu/lectures.html

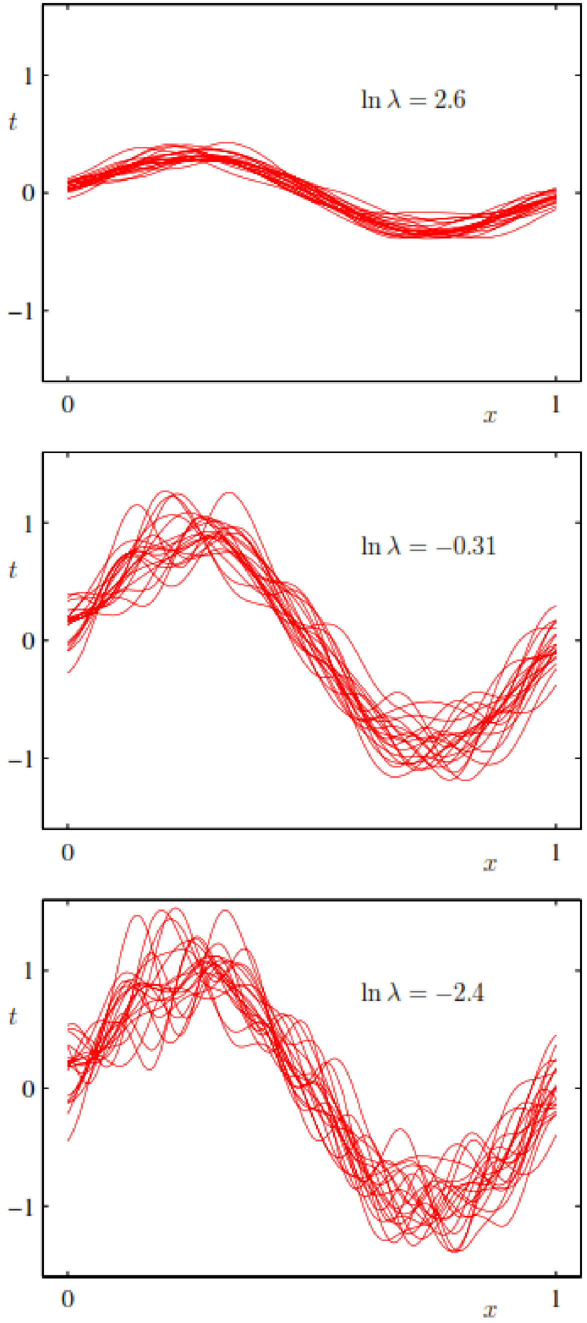
100 datasets

n=100

P=25



Average of 100 fits



Results of all fits

[Bishop]

References

- References
  - Pattern Recognition and Machine Learning by Christopher Bishop
  - Learning from data by Abu-Mostafa, Y.S., Magdon-Ismail, M. and Lin, H.T
    - Slides 10-18 are from the lectures 8 and 12 of *Learning from data* course at Caltech
    - https://work.caltech.edu/lectures.html