M6. Regression: Linear Models -

[a popular (supervised ML) appn. of linear algebra]

Manikandan Narayanan

Week 10 (Sep 29-, 2025)

PRML Jul-Nov 2025 (Grads Section)

Acknowledgment of Sources

- Slides based on content from related
  - Courses:
    - IITM – Profs. Arun/Harish/Chandra's PRML offerings (slides, quizzes, notes, etc.), Prof. Ravi's "Intro to ML" slides – cited respectively as [AR], [HR]/[HG], [CC], [BR] in the bottom right of a slide.
    - India – NPTEL PR course by IISc Prof. PS. Sastry (slides, etc.) – cited as [PSS] in the bottom right of a slide.
  - Books:
    - PRML by **Bishop**. (content, figures, slides, etc.) – cited as **[CMB]**
    - Pattern Classification by Duda, Hart and Stork. (content, figures, etc.) – [DHS]
    - Mathematics for ML by Deisenroth, Faisal and Ong. (content, figures, etc.) – [DFO]
    - Information Theory, Inference and Learning Algorithms by David JC MacKay – [DJM]

Outline for Module M5

- M6. Regression (Linear models)
  - M6.0 Introduction/context
    - Problem formulation
    - Motivating examples, incl. basis functions
  - M6.1 Linear regression approaches
    - Direct approach I: least-squares (error) (w_{LS})
    - Direct approach II: least-squares (error) with regularization (w_{RLS})
    - Discriminative model approach I: MLE ($w_{ML} = w_{LS}$)
    - Discriminative model approach IIa: Bayesian linear regression ($w_{MAP} = w_{RLS}$)
    - Discriminative model approach IIb: Bayesian linear regression (posterior & predictions)
  - M6.2 Model Complexity/Selection
    - Motivation (hyperparameter tuning to avoid overfitting)
    - Frequentist view (bias-variance decomposition)
  - M6.3 Concluding thoughts

Outline for Module M5

- M6. Regression (Linear models)
  - **M6.0 Introduction/context**
    - **Problem formulation**
    - **Motivating examples, incl. basis functions**
  - M6.1 Linear regression approaches
    - Direct approach I: least-squares (error) (w_{LS})
    - Direct approach II: least-squares (error) with regularization (w_{RLS})
    - Discriminative model approach I: MLE ($w_{ML} = w_{LS}$)
    - Discriminative model approach IIa: Bayesian linear regression ($w_{MAP} = w_{RLS}$)
    - Discriminative model approach IIb: Bayesian linear regression (posterior & predictions)
  - M6.2 Model Complexity/Selection
    - Motivation (hyperparameter tuning to avoid overfitting)
    - Frequentist view (bias-variance decomposition)
  - M6.3 Concluding thoughts

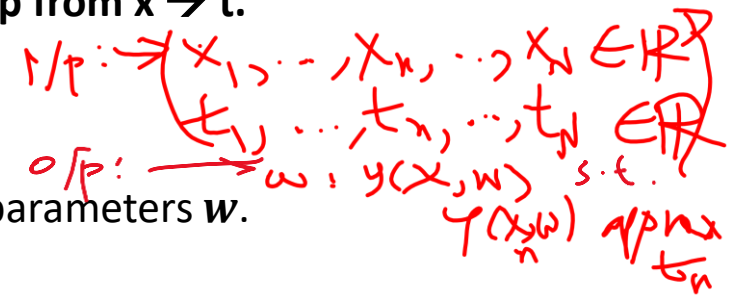
Context: ML Paradigms

- Unsupervised Learning (informally aka “learning patterns from (unlabelled) data”)
    - Density estimation
    - Clustering
    - Dimensionality reduction
- Supervised Learning (informally aka curve-fitting or function approximation or “function learning from (labelled) data”)
    - Learn an input and output map (features **x** to target **t**)
        - **Regression: continuous output/target t**
        - Classification: categorical output

[BR,HR]

Regression: the problem

- Learn a map from input variables (features **x**) to a continuous output variable (target **t**).
  Informally, known as function approximation/learning or curve fitting, since
  Given $\{\boldsymbol{x_n}, \boldsymbol{t_n}\}_{n=1\ldots N}$ pairs, we seek a function $\boldsymbol{y_w}(x)$ that "approximates" the map from x → t.



  Linear regression assumes $\mathrm{y_w(x)} \coloneqq y(x, w)$ is a **linear function of** the adjustable parameters $\boldsymbol{w}$.
  It could be linear or non-linear in x.

- A foundational supervised learning problem/algorithm:
  - Practical limitations for complex data, but sets analytical foundation for other sophisticated learning algorithms.
  - Due to its simplicity, first and predominant choice of statistical model in many applied areas with moderate sample sizes:

  ((E.g., in bioinformatics: to adjust for known confounding factors (covariates) in Disease Genomics and Genome-wide Association (GWAS) Studies, Causal inference such as in Mendelian Randomization, etc.))

- Our approach in this lecture: Mostly [CBM, Chapters 1,3].

[CMB]

Regression: what does **"approximate"** mean? recall three approaches (from Decision Theory)

- Generative model approach:
  - (I) Model $p(t, \mathbf{x})$
  - (I) Infer $p(t|\mathbf{x})$ from $p(t, \mathbf{x})$
  - (D) Take conditional mean/median/mode/any other optimal decision outcome as y(x)

- **Discriminative model** approach:
  - (I) Model $p(t|\mathbf{x})$ directly
  - (D) Take conditional mean/median/mode/any other optimal decision outcome as y(x)

- **Direct regression** approach:
  - (D) Learn a regression function y(x) directly from training data

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

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

[CMB]

Linear regression: from linear combination of input variables ($\mathrm{x} \in \mathbb{R}^D$) to that of basis functions ($\boldsymbol{\phi}(\mathrm{x}) \in \mathbb{R}^M$)

- Simplest model of linear regn. involving D input vars.:

$$y(\mathrm{x}, w) = w_0 + w_1 x_{(1)} + \ \ldots + w_D x_{(D)}$$

$$= w_0 \,.\, 1 + \sum_{j=1}^{D} w_j x_j \quad = \quad [w_0 \ w_1 \ \ldots w_D] \begin{bmatrix} 1 \\ \mathrm{x} \end{bmatrix} = \ w^T \begin{bmatrix} 1 \\ \mathrm{x} \end{bmatrix}$$

- Model of linear regn. involving M basis fns. (**fixed** non-linear fns. of the input vars.):

$$y(\mathrm{x}, w) \ = w_0 + w_1 \phi_1(\mathrm{x}) + \ \ldots + w_{M-1} \phi_{M-1}(\mathrm{x})$$

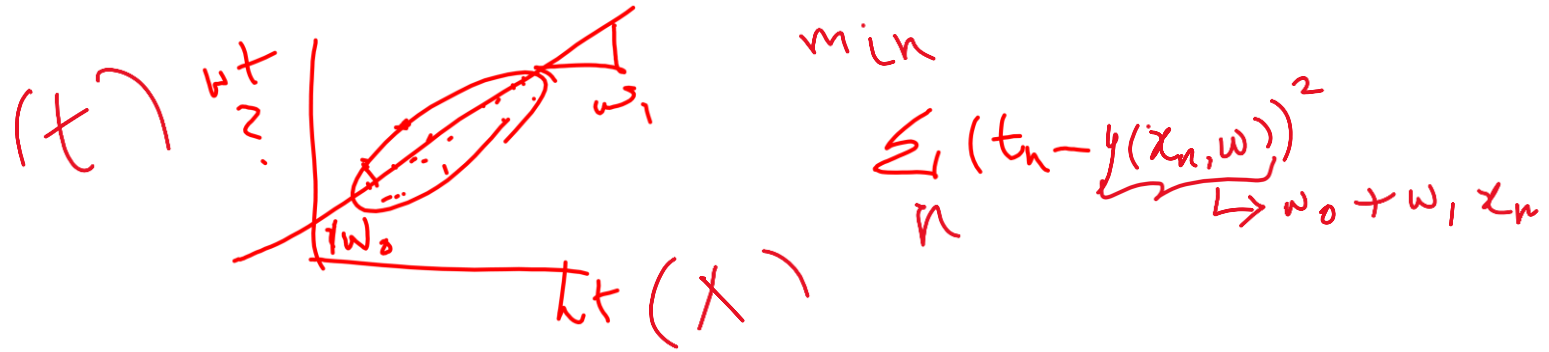
$$= w_0 . 1 + \sum_{j=1}^{M-1} w_j \phi_j(\mathrm{x})$$

$$= w^T \boldsymbol{\phi}(\mathrm{x})$$

($\boldsymbol{\phi}$: $\mathbb{R}^D \rightarrow \mathbb{R}^M$, with convention $\phi_0(\mathrm{x}) = 1$)

Linear regression: recall standard examples

- Predicting weight t from height x: $\boldsymbol{y(x, w)}$ **linear in both $\boldsymbol{x}$ and $\boldsymbol{w}$.**



- Estimation of fetal weight t (actually $\log_{10}$t) from ultrasound measurements: $\boldsymbol{y(x, w)}$ **linear in $\boldsymbol{w}$, not $\boldsymbol{x}$.**

Author	Components	Formula
Ferrero	AC, FL	10^(0.77125 + 0.13244 · AC – 0.12996 · FL – 1.73588 · AC^2/1,000 + 3.09212 · FL · AC/1,000 + 2.18984 · FL/AC) (g, cm)
Hadlock I	BPD, HC, AC, FL	10^(1.3596 + 0.0064 · HC + 0.0424 · AC + 0.174 · FL + 0.00061 · BPD · AC – 0.00386 · AC · FL) (g, cm)
Hadlock III	BPD, AC, FL	10^(1.335 – 0.0034 · AC · FL + 0.0316 · BPD + 0.0457 · AC + 0.1623 · FL) (g, cm)
Hadlock V	BPD, AC	10^(1.1134 + 0.05845 · AC – 0.000604 · AC^2 – 0.007365 · BPD^2 + 0.000595 · BPD · AC + 0.1694 · BPD) (g, cm)
Shepard	BPD, AC	10^(–1.7492 + 0.166 · BPD + 0.046 · AC – 0.002546 · AC · BPD) (kg, cm)

[From Hoopmann et al. *Fetal Diagn Ther*. 2011;30(1):29-34. doi:10.1159/000323586]

More examples of lin. regn. of basis functions

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_{(1)} + \ldots + w_D x_{(D)}$$

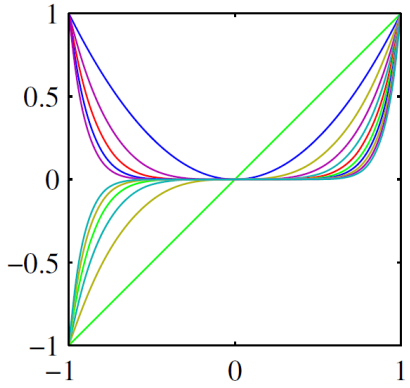
$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_{(1)} + w_2 \sqrt{x_{(2)}} + w_3 x_{(1)} \sqrt{x_{(2)}} + w_4 \log(x_{(1)})$$

$\phi : \mathbb{R}^2 \to \mathbb{R}^5 \qquad x = (x_{(1)}, x_{(2)}) \to (1, x_{(1)}, \sqrt{x_{(2)}}, \ldots)$

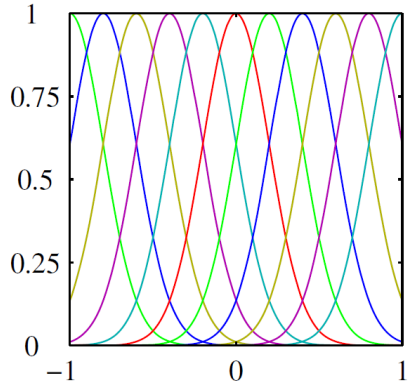
$\phi : \mathbb{R}^D \to \mathbb{R}^{D+1} \qquad \phi : x \to (1, x)$

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{M-1} w_j \phi_j(\mathbf{x})$$

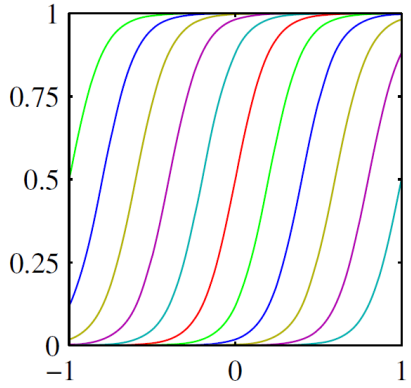
$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$



Polynomials (& extn. to spatially restricted splines)



$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$



$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$$

$$\sigma(a) = \frac{1}{1+\exp(-a)}.$$

$(x \in \mathbb{R}^D)$

$n \rightsquigarrow \phi(x_n) \in \mathbb{R}^M$

Fourier basis (sinusoidal fns.), Wavelets, etc.

[CMB]

Outline for Module M5

- M6. Regression (Linear models)
  - M6.0 Introduction/context
    - Problem formulation
    - Motivating examples, incl. basis functions
  - **M6.1 Linear regression approaches**
    - **Direct approach I: least-squares (error) (w_{LS})**
    - Direct approach II: least-squares (error) with regularization (w_{RLS})
    - Discriminative model approach I: MLE ($w_{ML} = w_{LS}$)
    - Discriminative model approach IIa: Bayesian linear regression ($w_{MAP} = w_{RLS}$)
    - Discriminative model approach IIb: Bayesian linear regression (posterior & predictions)
  - M6.2 Model Complexity/Selection
    - Motivation (hyperparameter tuning to avoid overfitting)
    - Frequentist view (bias-variance decomposition)
  - M6.3 Concluding thoughts

Linear regression: a direct approach

Approach: minimize sum-of-squares error; aka least- squares solution/approach

$$\min_{\mathbf{w}} \left(E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \right)$$

$\mathcal{D}_N = \{x_n, t_n\}_{n=1}^{N}$

$\downarrow$

$\phi(x_n) \in \mathbb{R}^M$

where $y(x, \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{\phi}(x)$

Solution: w_{LS} that minimizes $E_D(w)$ (via matrix notation)

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2. = \frac{1}{2}\|\Phi w - \vec{t}\|^2$$

$w \in \mathbb{R}^M$

$$\vec{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}$$

$\vec{y} = \Phi w$

basis fns

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

data pts

$N \times M$

Normal equations from setting gradient to zero, using N x M design matrix:

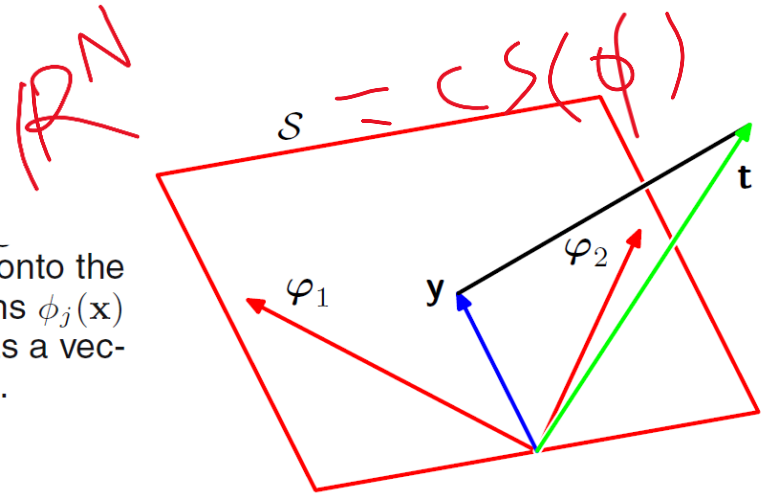
$$\Phi^T \Phi w_{LS} = \Phi^T \vec{t} \quad (1)$$

$$\Rightarrow w_{LS} = (\Phi^T \Phi)^{-1} \Phi^T \vec{t}$$

Ex: Prove (a) (1) always has a soln.
(b) If Φ has linearly indep. cols, then $\Phi^T\Phi$ is invertible.
(c) What if $\Phi^T\Phi$ is not invertible?

[CMB, HR]

Soln.: Geometry of (least-squares) sol.



$y = \Phi w$

Orthogonal projection of the data vector **t** onto the subspace spanned by the basis functions $\phi_j(\mathbf{x})$ in which each basis function is viewed as a vector φ_j of length N with elements $\phi_j(\mathbf{x}_n)$.

[CMB, Figure 3.2]

$$\therefore \min_{w \in \mathbb{R}^m} \|\Phi w - t\|^2 \quad -(A)$$

$$\min_{y \in CS(\Phi)} \|y - t\|^2 \quad -(B)$$

[CMB, HR]

LA Refresher

- See Appendix for LA refresher on
  - related topic of Ax=b when no solution is possible, and
  - some matrix-vector gradient formulas/tricks.

Recall LA: To solve $Ax = b$, we premult. by A^T, and simply solve $A^T Ax = A^T b$.

Linear Equation Solving

$Ax = b$

(X) → No solution if $b \notin$ columnspace (A)

→ Unique solution if $b \in$ columnspace (A) & A has lin. ind columns

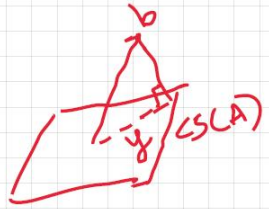
→ Infinite solutions if $b \in CS(A)$ & A has lin. dep. columns.

$$\begin{bmatrix} 1 & 3 \\ 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

(X) Least-squares soln possible. (related to lin. regn)

$\min_x \|Ax - b\|^2 == \min_{y \in CS(A)} \|y - b\|^2$

$(Ax^* = y^*)$ (y^* is Proj of b onto CS)



$$\min_{x \in \mathbb{R}^d} \underbrace{\left[(Ax-b)^T(Ax-b)\right]}_{f(x)}$$

$$\nabla f(x) = 2A^T(Ax-b) \overset{\text{set to}}{=} 0$$

$$\Rightarrow A^T(Ax^*-b) = 0 \quad \text{(normal eqn.)}$$

$$\Rightarrow A^T A x^* = A^T b$$

$$\Rightarrow x^* = (A^T A)^{-1} A^T b$$

Ex.: Prove:

1) at least one soln. x* exists for the normal eqn.
2) soln. x* unique if $(A^T A)$ is invertible ($\Leftrightarrow A$ has lin. indep. cols.)
3) **infinite** solns. x* if $(A^T A)$ is **non**-invertible ($\Leftrightarrow A$ has lin. **dep.** cols.)

Ex.:

i) Prove $NS(A) = NS(A^T A)$.

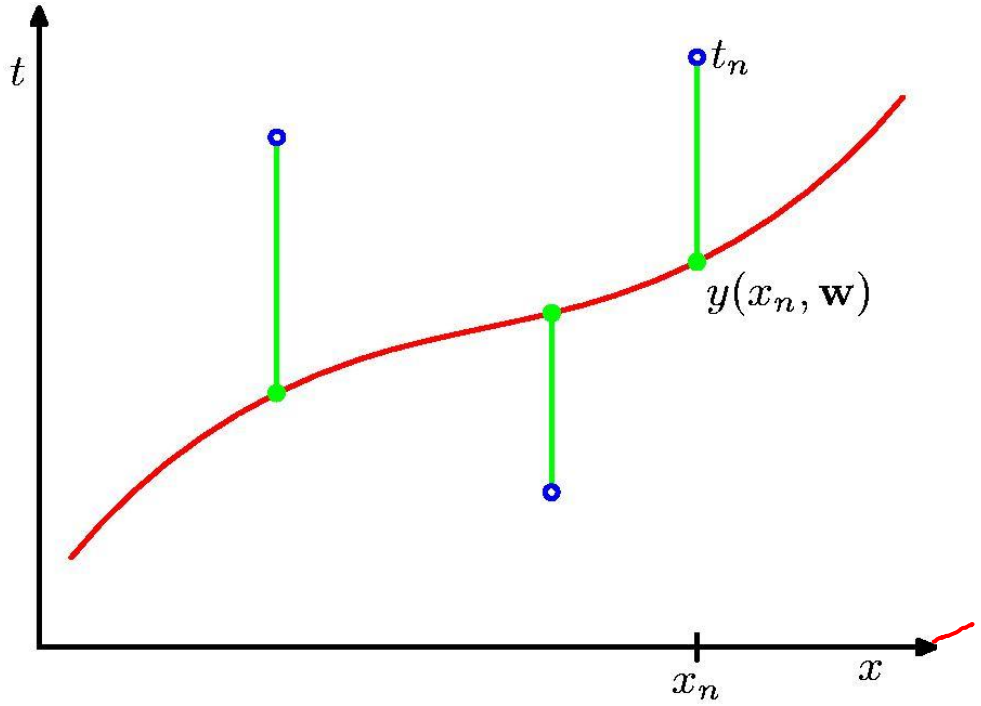
ii) Use orthog. complementarity of $NS(A^T), CS(A)$ to derive normal eqns.

Outline for Module M5

- M6. Regression (Linear models)
  - M6.0 Introduction/context
    - Problem formulation
    - Motivating examples, incl. basis functions
  - **M6.1 Linear regression approaches**
    - Direct approach I: least-squares (error) (w_{LS})
    - **Direct approach II: least-squares (error) with regularization ($\boldsymbol{w_{RLS}}$)**
    - Discriminative model approach I: MLE ($w_{ML} = w_{LS}$)
    - Discriminative model approach IIa: Bayesian linear regression ($w_{MAP} = w_{RLS}$)
    - Discriminative model approach IIb: Bayesian linear regression (posterior & predictions)
  - M6.2 Model Complexity/Selection
    - Motivation (hyperparameter tuning to avoid overfitting)
    - Frequentist view (bias-variance decomposition)
  - M6.3 Concluding thoughts

From sum-of-squares to regularized error!

Running example: polynomial curve fitting ((via sum-of-squares error function / least squares approach))

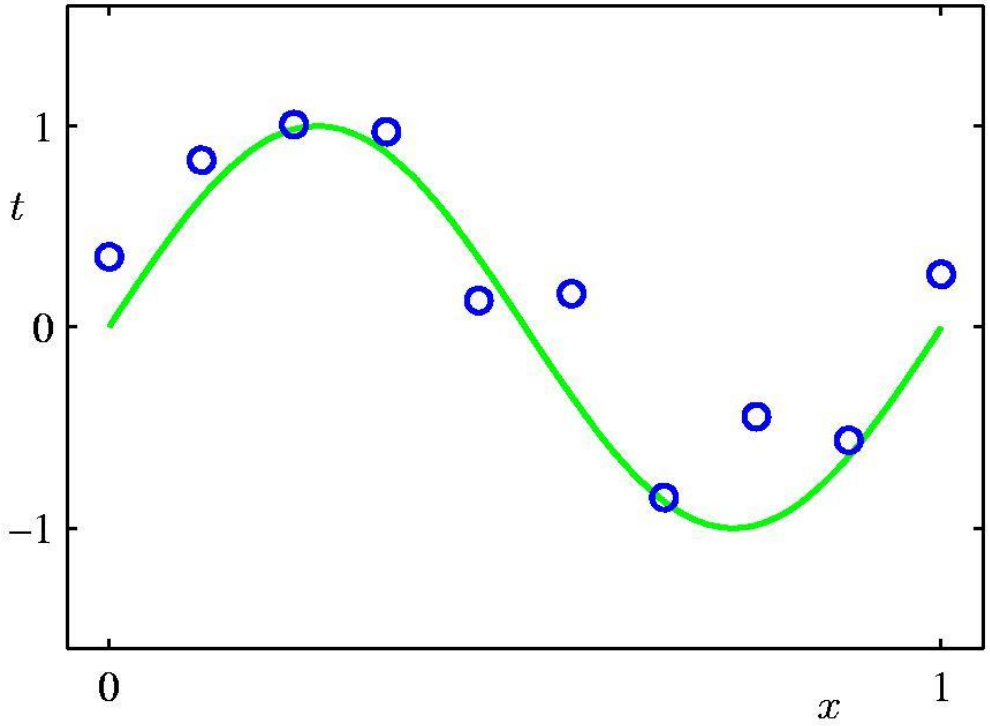


$$\min_{\mathbf{w}} \left(E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \right)$$

where $y(x, \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{\phi}(x)$

[CMB]

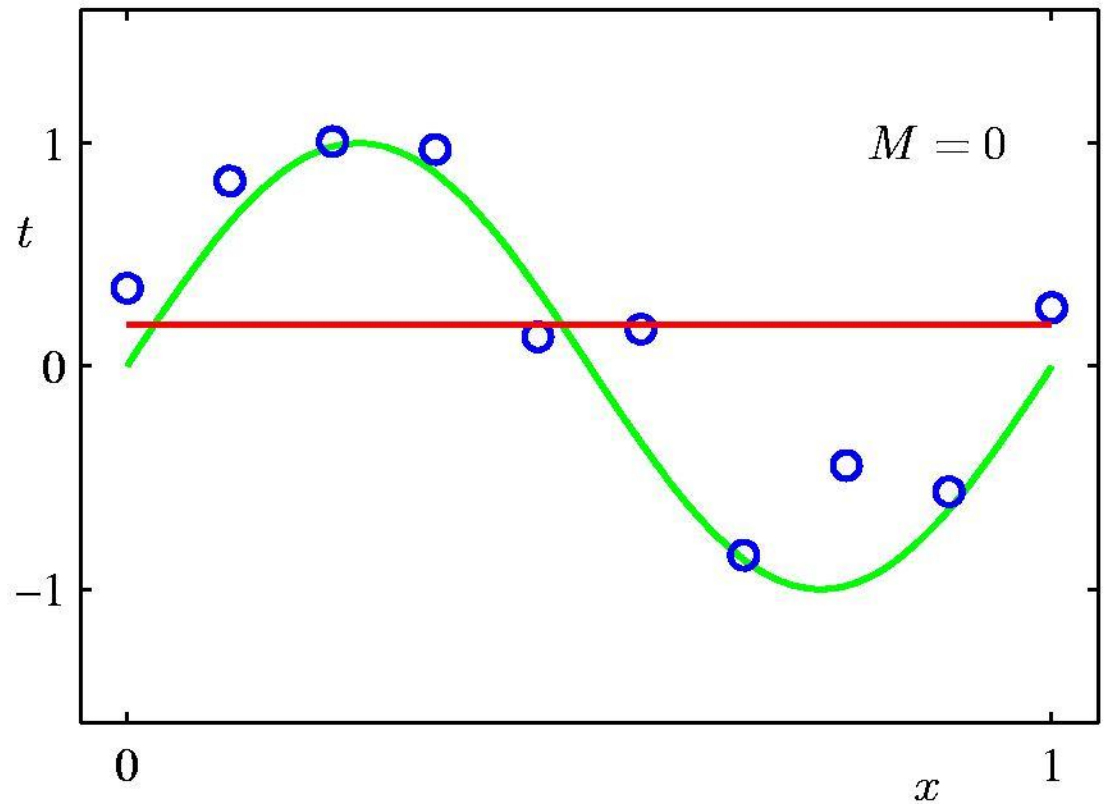
Polynomial Curve Fitting



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

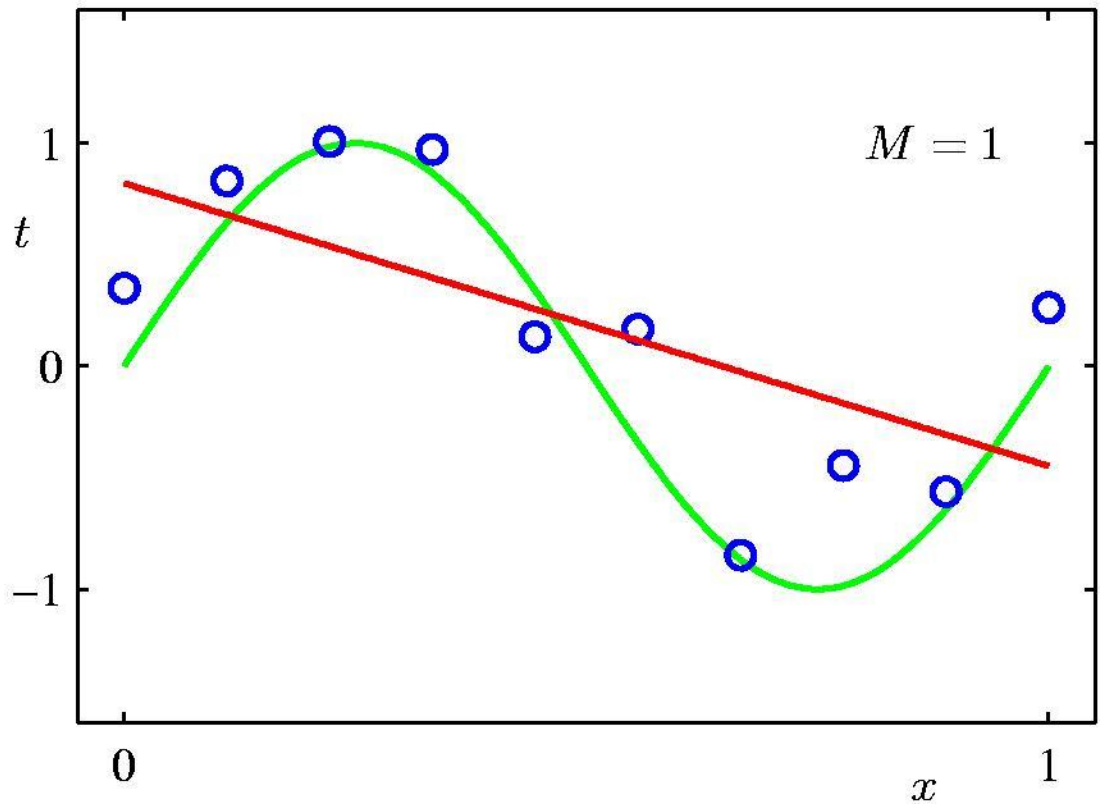
[CMB]

0^{th} Order Polynomial



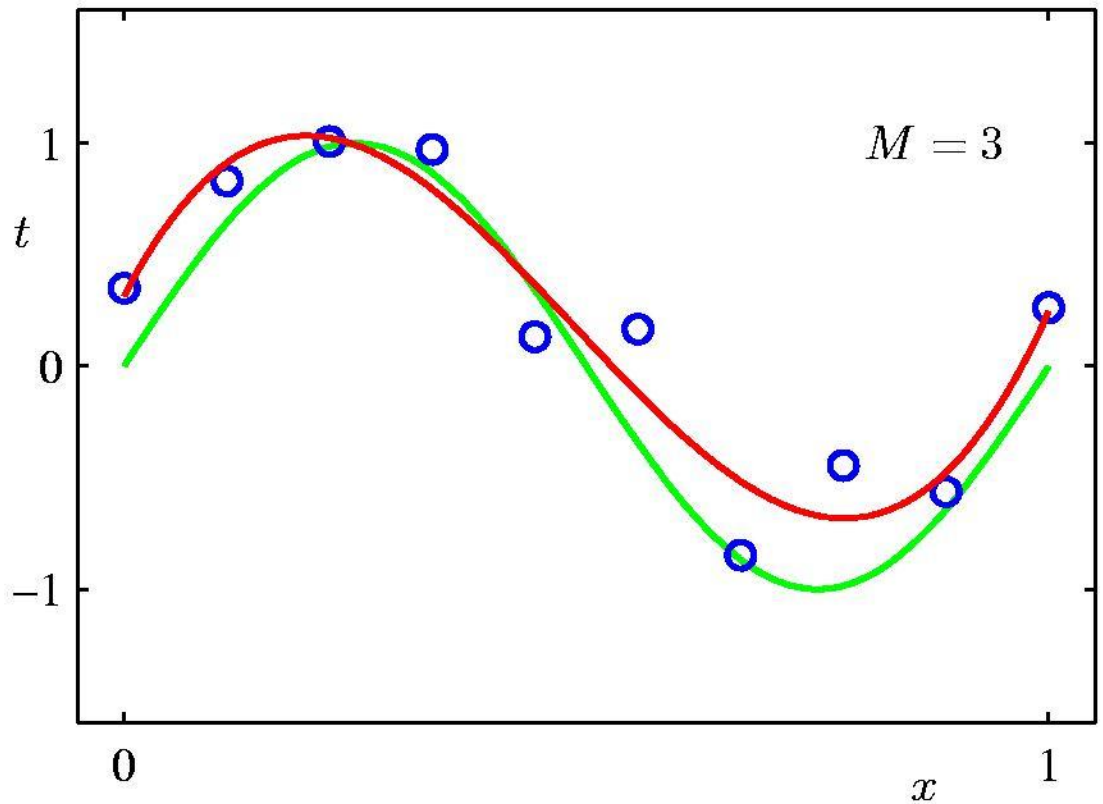
[CMB]

1st Order Polynomial



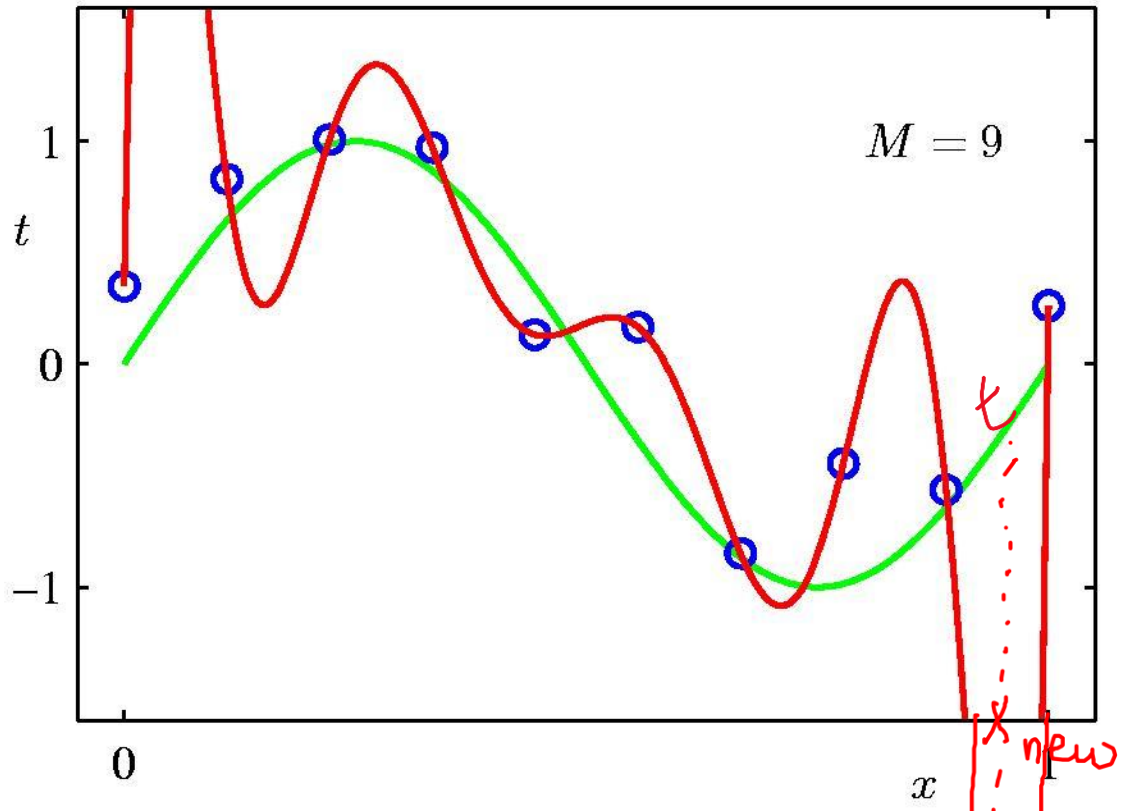
[CMB]

3rd Order Polynomial



[CMB]

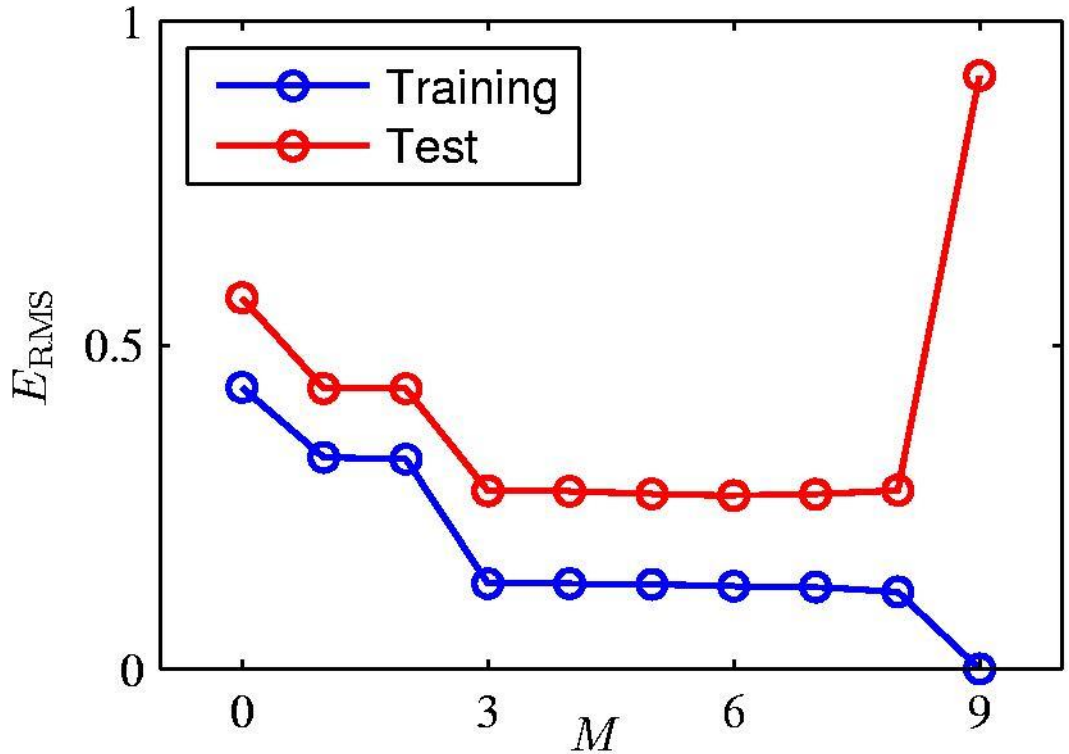
9th Order Polynomial



[CMB]

Brainstorm: What degree polynomial would you choose?

Over-fitting



Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^{\star})/N}$

[CMB]

Polynomial Coefficients

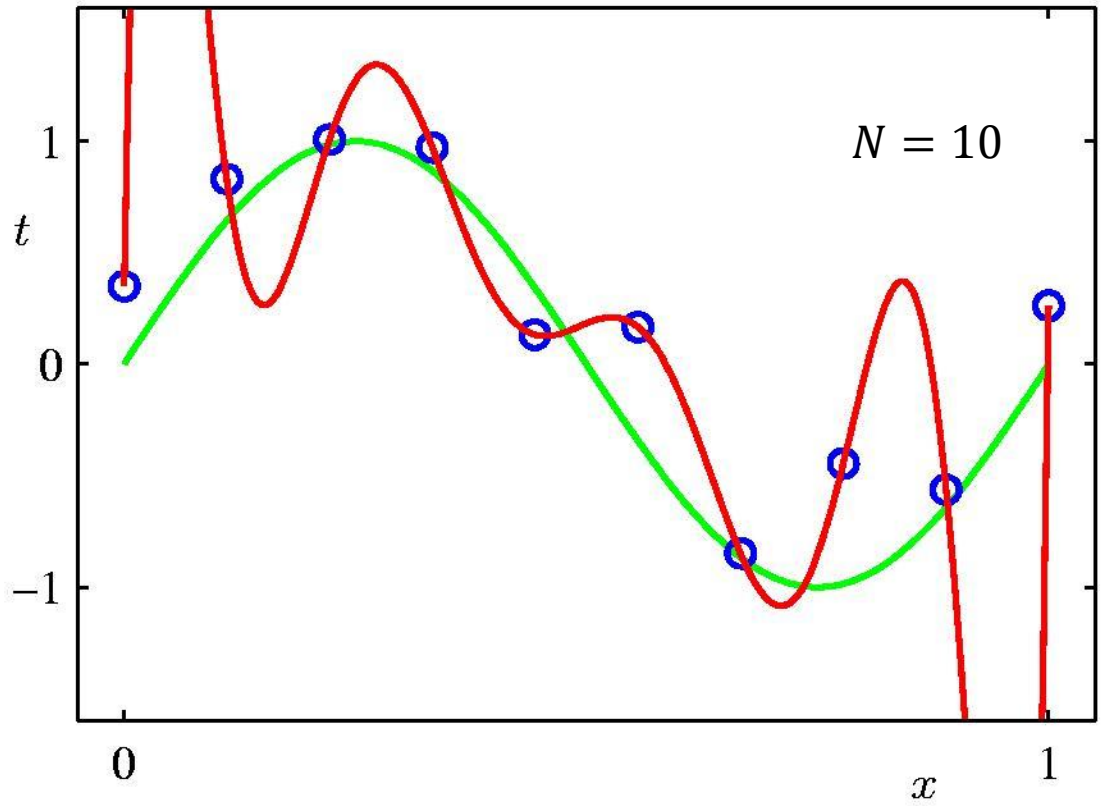
	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

[CMB]

The role of data set size N?

Data Set Size: $N = 10$

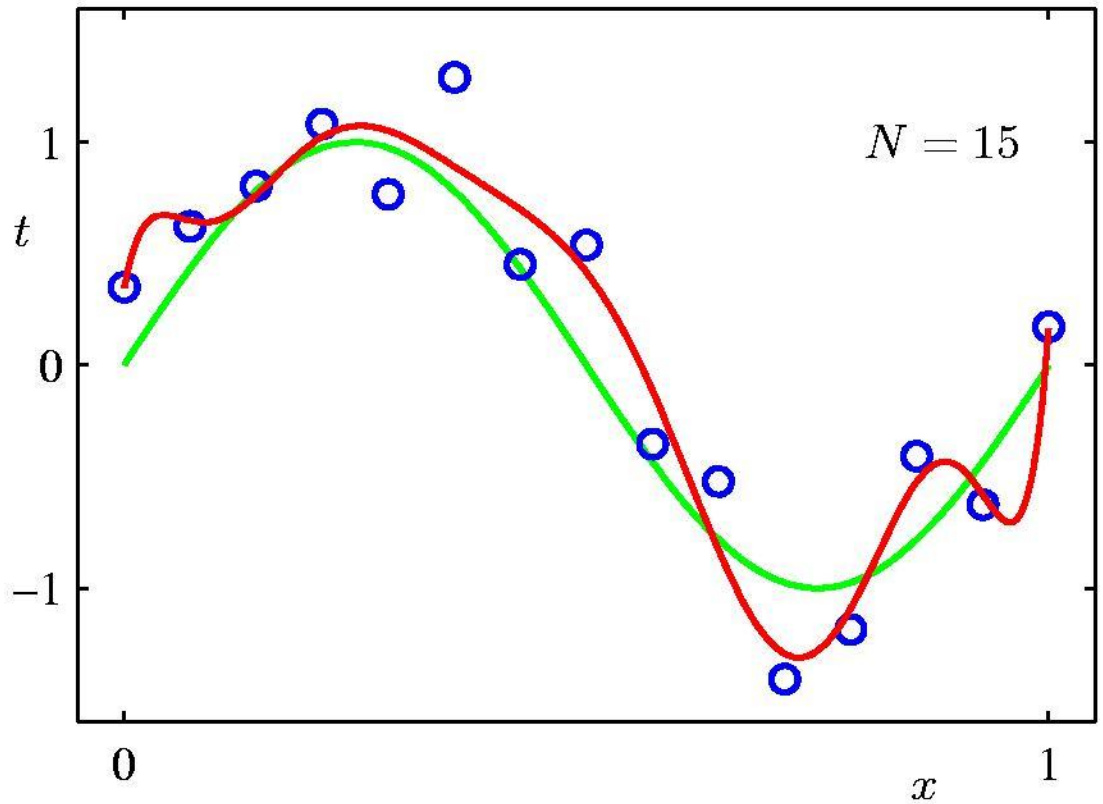
9th Order Polynomial



[CMB]

Data Set Size: $N = 15$

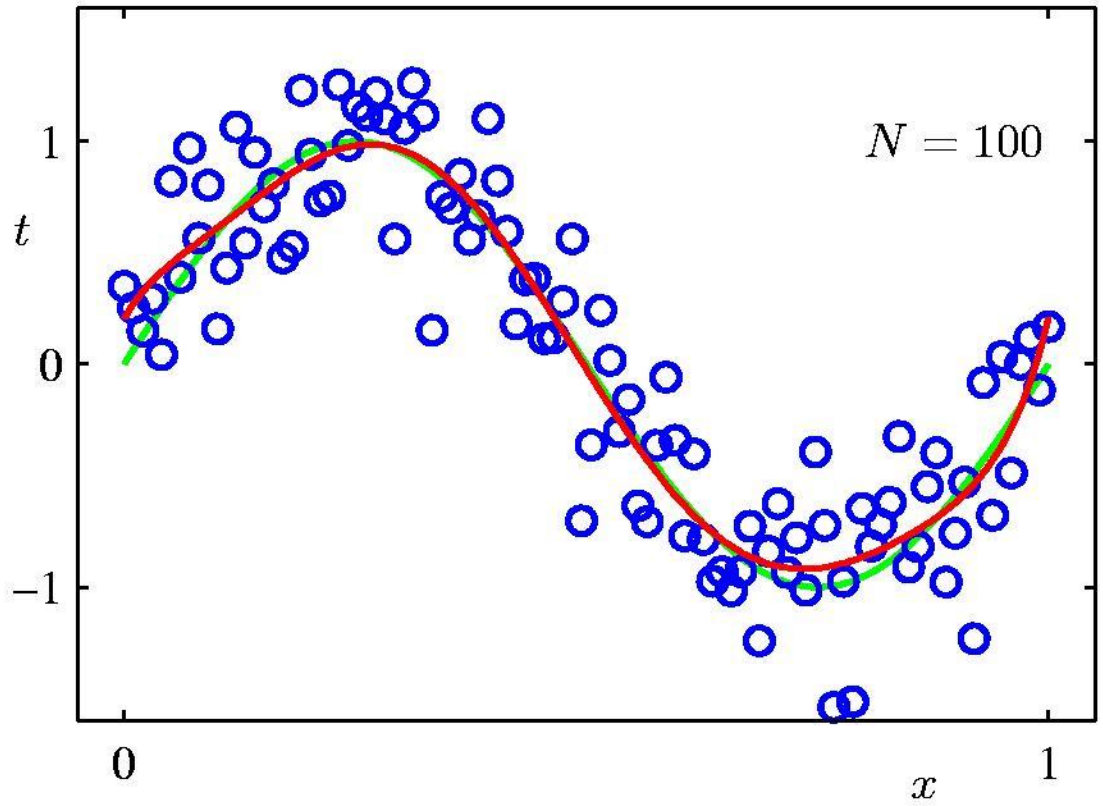
9th Order Polynomial



[CMB]

Data Set Size: $N = 100$

9th Order Polynomial



[CMB]

How to take care of both data set size and model complexity tradeoffs?

Regularization

- Penalize large coefficient values

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

favors complex models

penalizes complex models

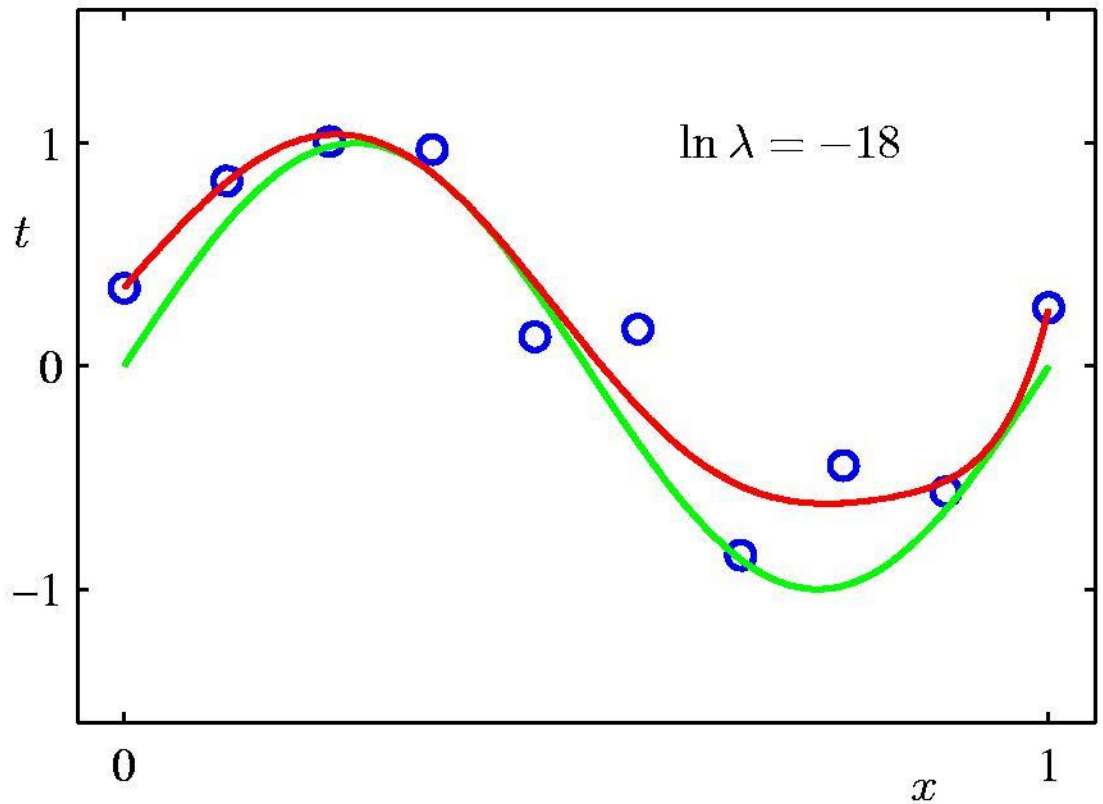
[CMB]

Polynomial Coefficients

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	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

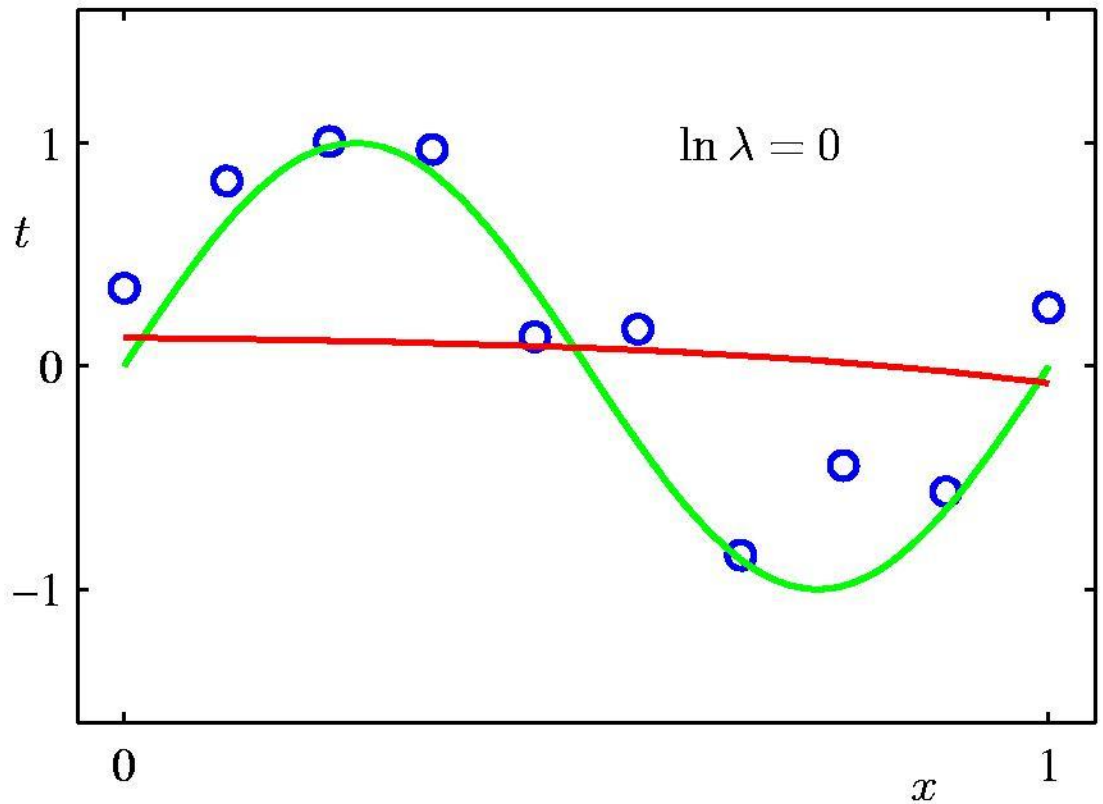
[CMB]

Regularization: $\ln \lambda = -18$



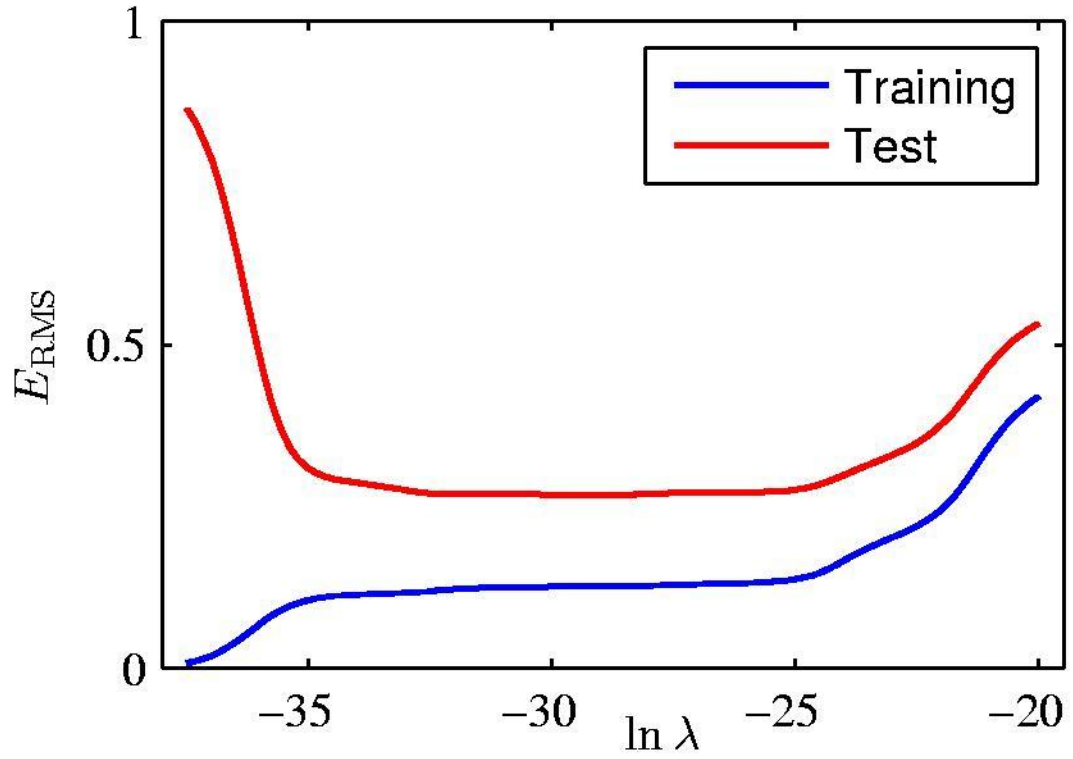
[CMB]

Regularization: $\ln \lambda = 0$



[CMB]

Regularization: E_{RMS} vs. $\ln \lambda$



more complex model ⟷ less complex model

[CMB]

Now, let's see how to solve the minimization problem!

$$\min_{w} \quad \widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Matrix notation gradient formula (chain rule product "builds towards the left"):

$f(x) = g(Ax)$

$\nabla f(x) = A^T \nabla g(Ax)$

$f: \mathbb{R}^d \to \mathbb{R}$, $g: \mathbb{R}^n \to \mathbb{R}$

$A \in \mathbb{R}^{n \times d}$

Regularized Least Squares (1): Solution for ridge regression

- Consider the error function:

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

Data term + Regularization term

λ is called the regularization coefficient.

- With the sum-of-squares error function and a quadratic regularizer, we get

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

- which is minimized by

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

RLS

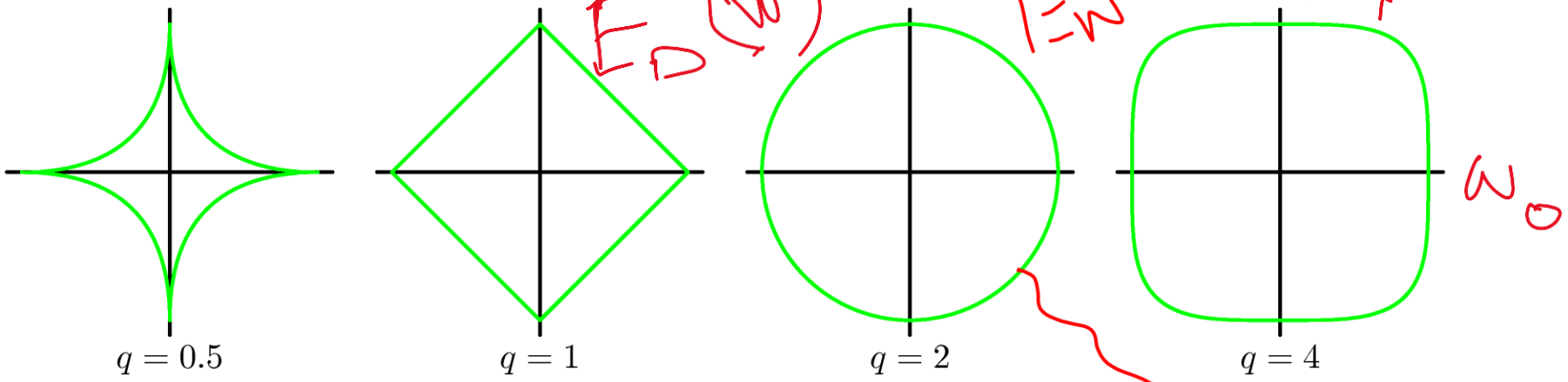
Ex: ST $\lambda I + \Phi^T\Phi$ is invertible for $\lambda > 0$

[CMB]

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

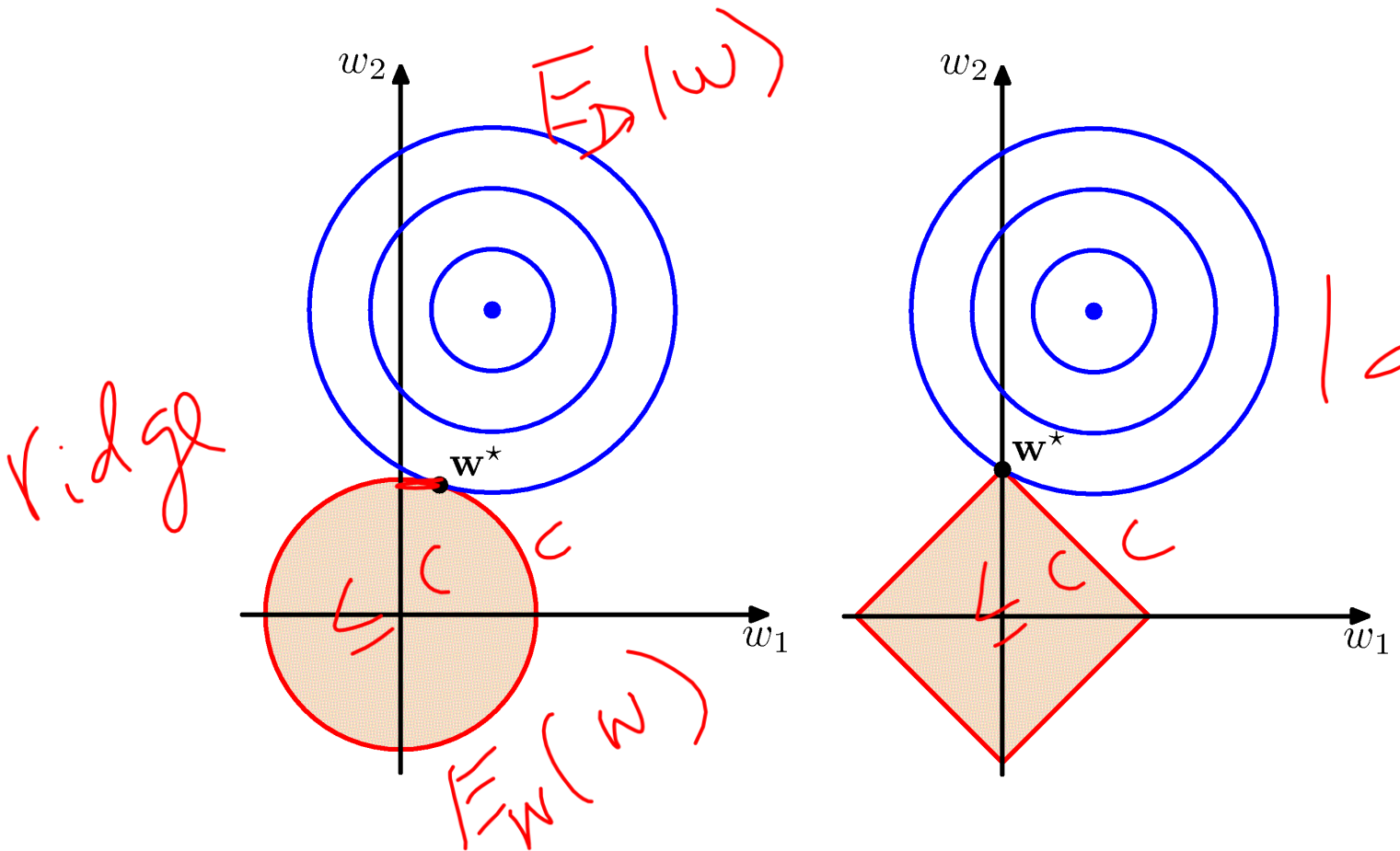
Quadratic

ridge

[CMB]

Regularized Least Squares (3)

- Lasso tends to generate sparser solutions than a ridge (quadratic) regularizer.
- Regularization aka penalization/weight-decay in ML or parameter shrinkage in statistics literature.



trad. $N \gg M$

hi dim $N \ll M$

200 20k genes

[CMB]

Outline for Module M5

- M6. Regression (Linear models)
  - M6.0 Introduction/context
    - Problem formulation
    - Motivating examples, incl. basis functions
  - **M6.1 Linear regression approaches**
    - Direct approach I: least-squares (error) (w_{LS})
    - Direct approach II: least-squares (error) with regularization (w_{RLS})
    - **Discriminative model approach I: MLE ($\boldsymbol{w_{ML} = w_{LS}}$)**
    - Discriminative model approach IIa: Bayesian linear regression ($w_{MAP} = w_{RLS}$)
    - Discriminative model approach IIb: Bayesian linear regression (posterior & predictions)
  - M6.2 Model Complexity/Selection
    - Motivation (hyperparameter tuning to avoid overfitting)
    - Frequentist view (bias-variance decomposition)
  - M6.3 Concluding thoughts

A different view of min. E(w) or its regularized version?

Q: How do you convert this intuition/empirical-art into science, and derive E(w) or its (other) regularized versions systematically?

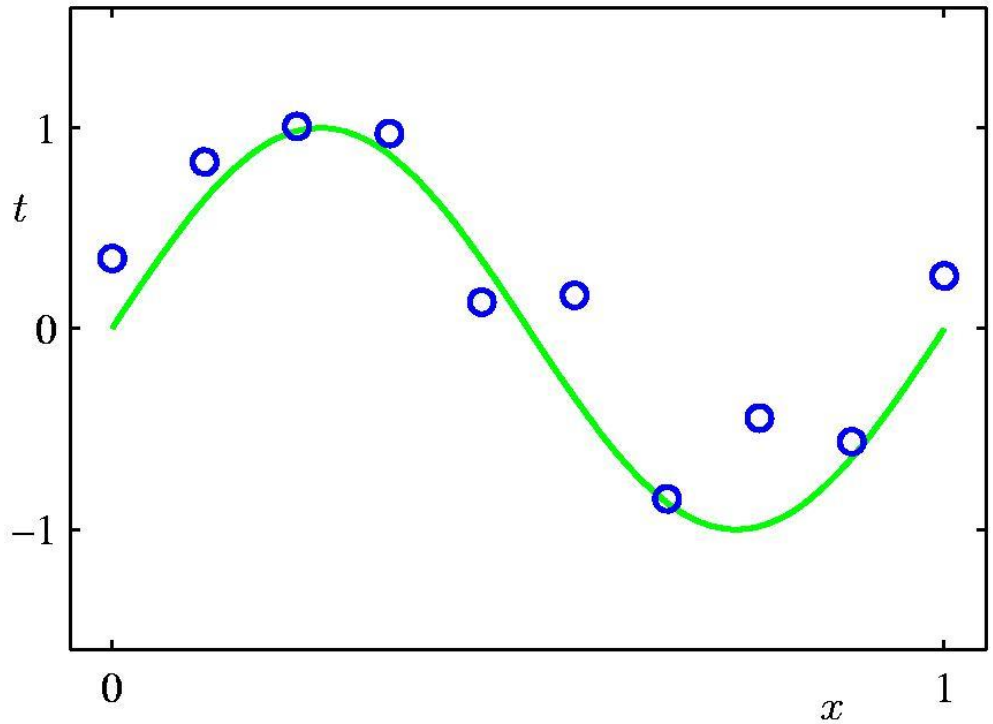
A: Probabilistic view helps. Discriminative Approach: Model $p(t|x)$

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

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

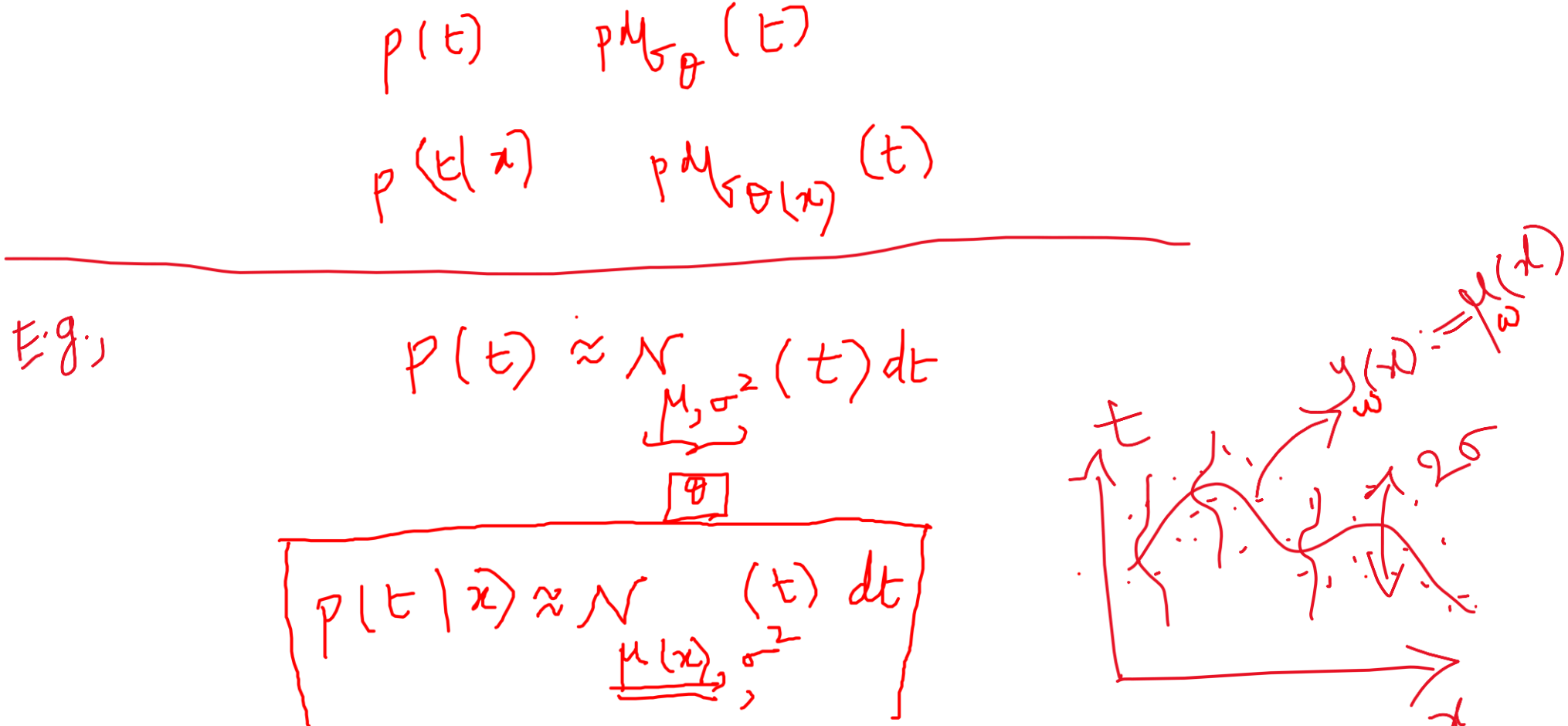
Q: What has regression got to do with our previous topic: density estimation?

- Brainstorm: how to model P(t|x)?



Q: What has regression got to do with our previous topic: density estimation?

- Brainstorm: how to model P(t|x)?



Q: What has regression got to do with our previous topics?

A: **P(t|x)** captures the input-output map. Steps involved are:

(1) Model/estimate **P(t|x)**

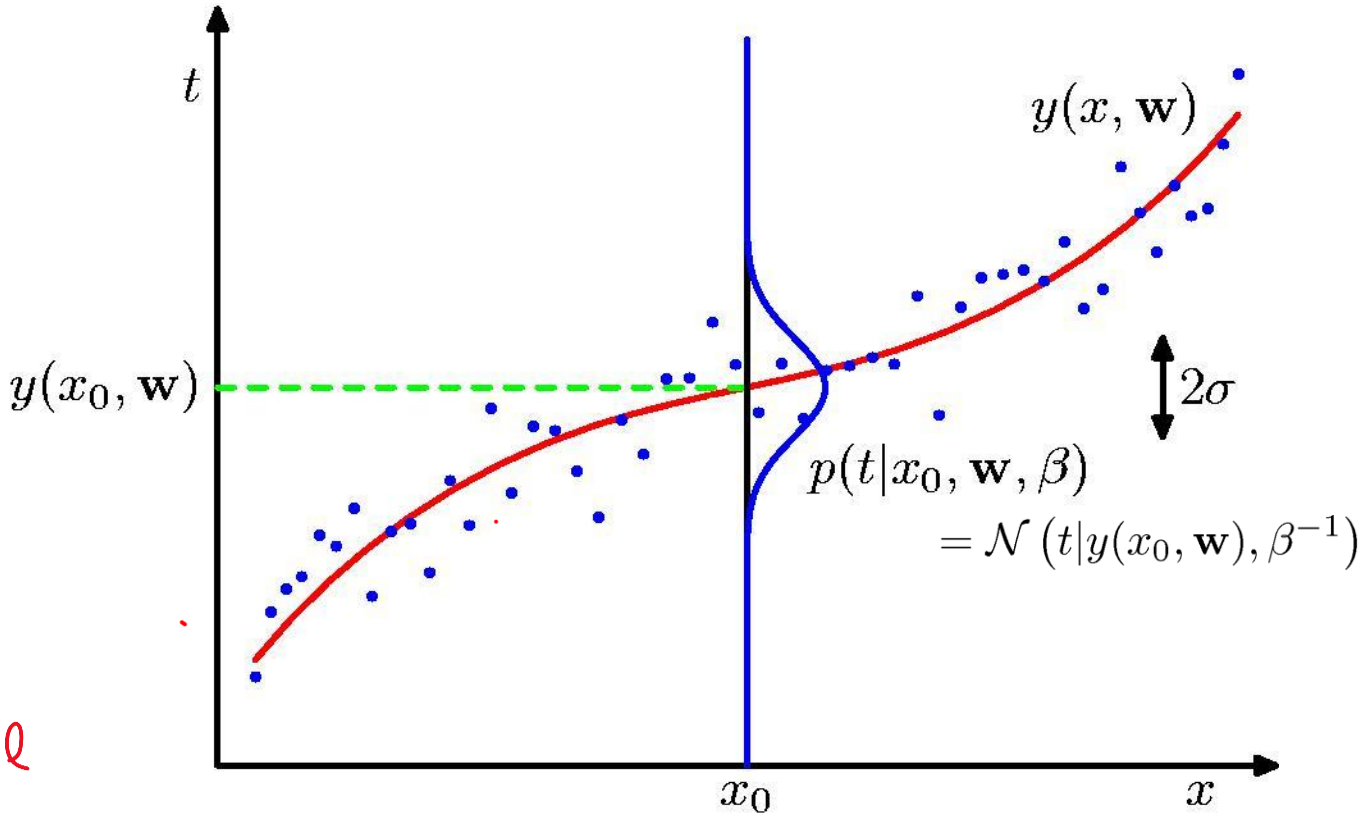
(how? *Density Estimation*; MLE/Bayesian Inference)

(2) Predict **t** for a new **x** from estimated P(t|x)

(how? *Decision Theory*; e.g., $y(\mathbf{x}_{new}) = E[\boldsymbol{t}|\mathbf{x} = \mathbf{x}_{new}]$)

Curve Fitting: Going to the basics!

using a **Probabilistic Model** & its **Density Estimation** (**MLE**/Bayesian)



Discriminative

model $t|X: t = y(X, w) + \epsilon$, $\epsilon \sim N(0, \beta^{-1})$, $\beta^{-1} = \sigma^2$

[CMB]

ML estimation

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \beta^{-1}\right) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

$$\begin{aligned}
\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \\
&= \underbrace{-\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2}_{\beta E(\mathbf{w})} + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) \\
&= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})
\end{aligned}$$

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

Determine $\mathbf{w}_{\mathrm{ML}}$ by minimizing sum-of-squares error $E_D(w)$.

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

[CMB]

Summary: Linear model for regression - $w_{ML} == w_{LS}$

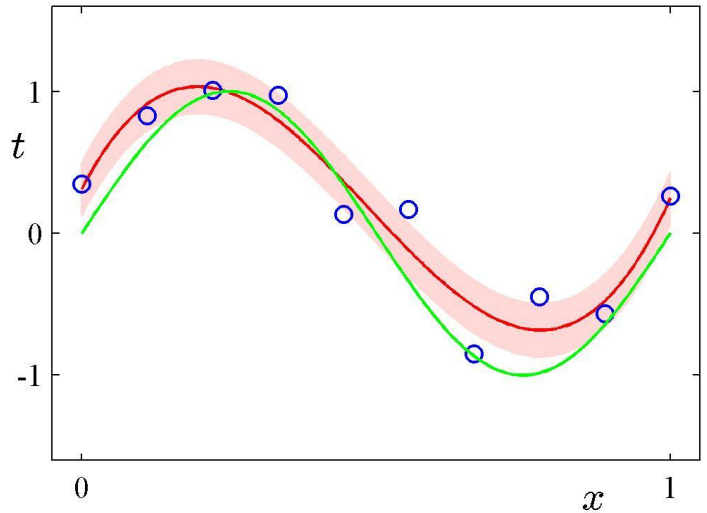
$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

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

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$$

$t = y(x, w) + \epsilon$

mle: $p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)$

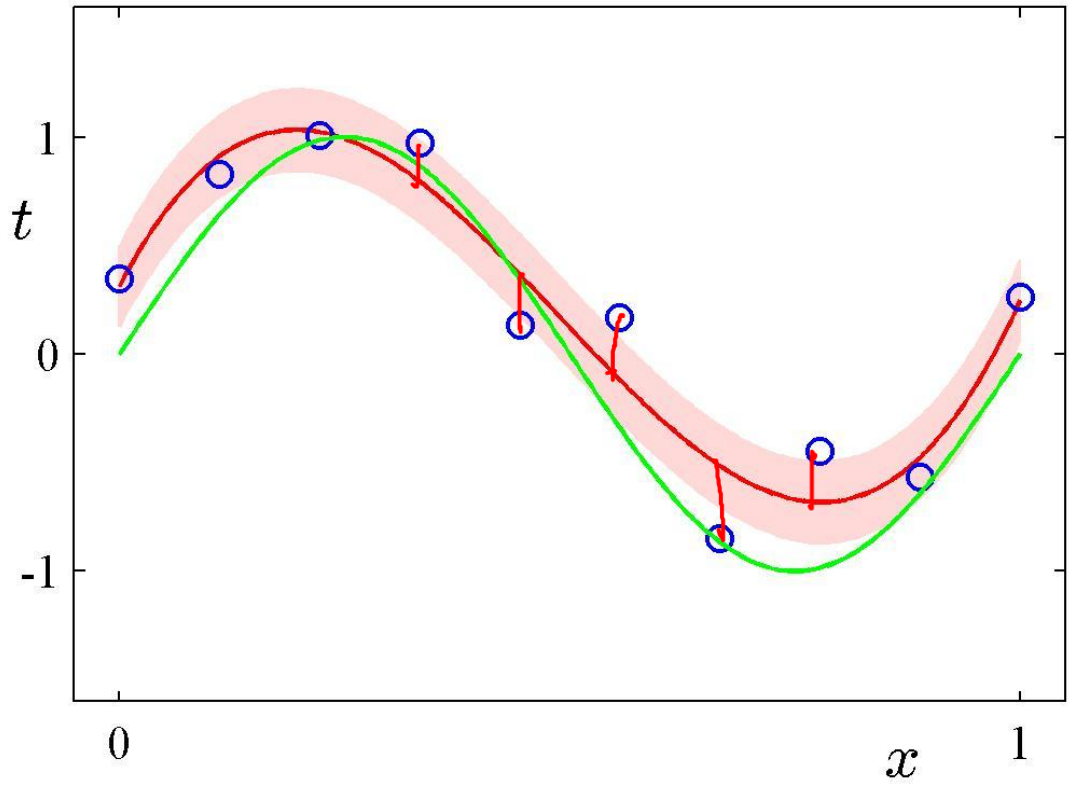


$\epsilon \sim N(0, \frac{1}{\beta})$

[CMB]

Addtnl. Advantage: ML Predictive Distribution

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)$$

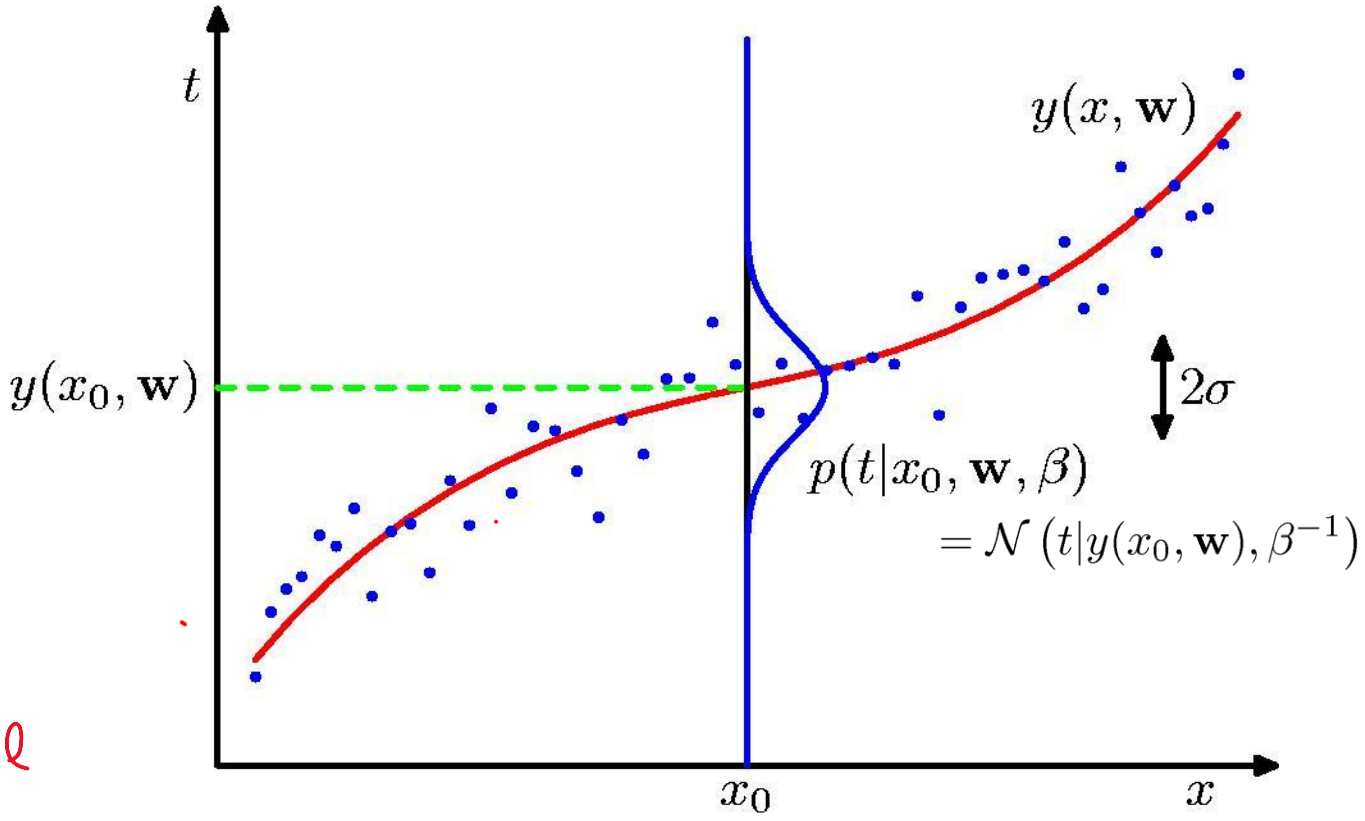


[CMB]

Outline for Module M5

- M6. Regression (Linear models)
  - M6.0 Introduction/context
    - Problem formulation
    - Motivating examples, incl. basis functions
  - **M6.1 Linear regression approaches**
    - Direct approach I: least-squares (error) (w_{LS})
    - Direct approach II: least-squares (error) with regularization (w_{RLS})
    - Discriminative model approach I: MLE ($w_{ML} = w_{LS}$)
    - **Discriminative model approach IIa: Bayesian linear regression ($\boldsymbol{w_{MAP} = w_{RLS}}$)**
    - Discriminative model approach IIb: Bayesian linear regression (posterior & predictions)
  - M6.2 Model Complexity/Selection
    - Motivation (hyperparameter tuning to avoid overfitting)
    - Frequentist view (bias-variance decomposition)
  - M6.3 Concluding thoughts

Recall: Probab. Model & its Density Estimation (MLE)



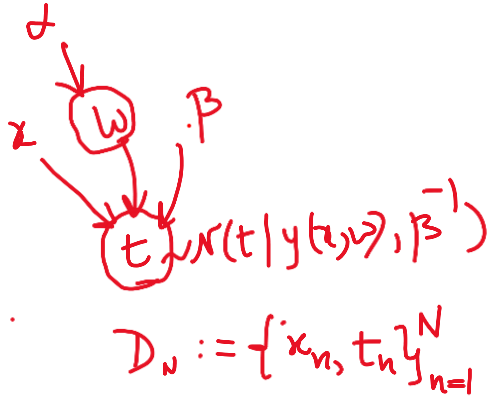
Discriminative model $t|X: t = y(X, w) + \epsilon$, $\epsilon \sim N(0, \beta^{-1})$, $\beta^{-1} = \sigma^2$

[CMB]

Bayesian inference: what would you model as a rv instead of a fixed value?

- Brainstorm
  - What would you model?
  - What would you infer?

Bayesian inference: what would you model as a rv instead of a fixed value?



- Brainstorm
  - What would you model? $P(w)$ prior, besides the $P(t|x,w)$ model.
  - What would you infer? (1) $P(w|D_N)$ = post.

(2a) MAP-based predn. (or) (2b) $P(t_{new}|x_{new}, D_N)$ post. pred.

Compare with MLE: (I/p is same: D_N) O/p desired: (1) w^* that max. $LL(w)$ or min $E(w)$
(2) Pred. $t_{new} = y(x_{new}, w^*)$

Relation between Bayesian MAP and Regularized linear regression

Let look at a simpler problem first – mode of the posterior of w

$$w_{MAP} = argmax_w \, P(w \mid \boldsymbol{D_N} \coloneqq \{\mathbf{x}, \boldsymbol{t}\} \coloneqq \{\boldsymbol{x_n}, \boldsymbol{t_n}\}_{n=1 \; to \; N})$$

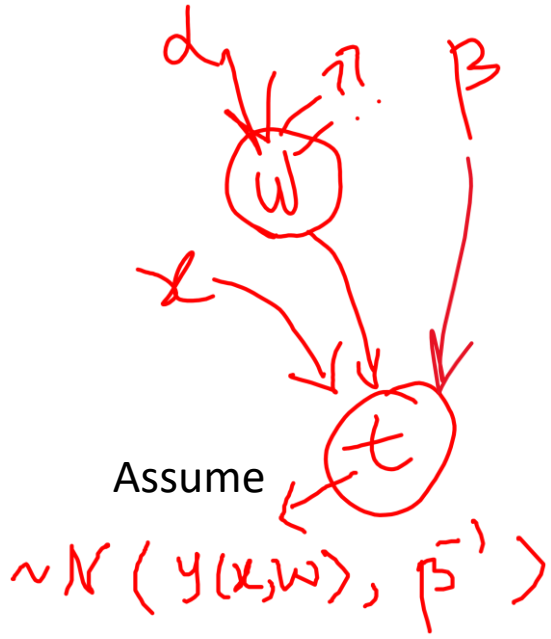
We will actually show that $w_{MAP} = w_{RLS}$!!

Bayesian inference: a first step via MAP

Assume $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$

Then, $p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$

D_N



Determine maximizer of this posterior, i.e., w_{MAP}, by minimizing regularized sum-of-squares error $\tilde{E}(w)$, because:

$$\beta\widetilde{E}(\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

$\beta E_D(w)$ — contribns. likhd.

$\beta E_W(w)$ $\lambda \triangleq \frac{\alpha}{\beta}$ $\cdot \left(\frac{\lambda}{2}\|w\|^2\right)$ — penalty/regularizer/prior (shrinks w towards 0)

[cf. full details of proof in next slide]

[CMB]

Full details of w_{MAP} derivation, & $w_{MAP} = w_{RLS}$ proof!

likhd. $p(\underline{t}|\underline{x}, w, \beta) = \prod_{n=1}^{N} p(t_n | x_n, w, \beta)$

$= \prod_{n=1}^{N} \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\{-\frac{\beta}{2}(y(x_n, w) - t)^2\}$

$\Rightarrow \ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})$

prior. $p(w|\alpha) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$

$\ln p(\boldsymbol{w}|\alpha) = \frac{M}{2}\ln\alpha - \frac{M}{2}\ln(2\pi) - \frac{\alpha}{2}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w}$

---

post. $p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$

$\Rightarrow \ln p(\boldsymbol{w}|\mathbf{x}, \boldsymbol{t}, \alpha, \beta) = \ln p(\boldsymbol{t}|\mathbf{x}, \boldsymbol{w}, \alpha, \beta) + \ln p(\boldsymbol{w}|\alpha) + const.$ (assume α, β are both known hyperparams.)

$= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\boldsymbol{w}) + \frac{M}{2}\ln\alpha - \frac{M}{2}\ln(2\pi) - \frac{\alpha}{2}w^T w + const.$

So maximizing this $\ln(posterior)$ is the same as max. $-\beta E_D(\boldsymbol{w}) - \frac{\alpha}{2}||\boldsymbol{w}||^2$ (ignoring terms indept. of $\boldsymbol{w}$) , or equivalently

min. $\beta E_D(\boldsymbol{w}) + \alpha\frac{1}{2}||\boldsymbol{w}||^2 = \beta\left(E_D(\boldsymbol{w}) + \frac{\alpha}{\beta}E_W(\boldsymbol{w})\right) = \beta\tilde{E}(\boldsymbol{w})$, or equivalently min. $\tilde{E}(\boldsymbol{w})$. ((here, we set $\lambda := \frac{\alpha}{\beta}$))

Changing the prior from Gaussian to Laplacian!

- Prior: $\mathrm{p}(w) = p(w_0)p(w_1) \ldots p(w_{M-1})$
  - What if $p(w_i)$ changed from

    $\left(\frac{\alpha}{2\pi}\right)^{\frac{1}{2}} \exp\{-\frac{\alpha}{2}\boldsymbol{w}_i^2\}$ → $\left(\frac{\alpha}{4}\right) \exp\{-\frac{\alpha}{2}|\boldsymbol{w}_i|\}$?

  - Then, $p(w)$ changes from

    $\left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \exp(-\frac{\alpha}{2}||\boldsymbol{w}||_{\mathbf{2}}^{\mathbf{2}})$ → $\left(\frac{\alpha}{4}\right)^{M} \exp(-\frac{\alpha}{2}||\boldsymbol{w}||_1)$

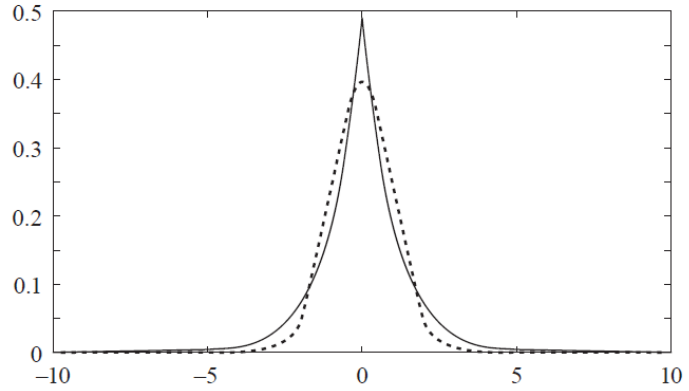


Figure 20.3 Laplacian distribution ($\beta = 1$) and Gaussian distribution ($\sigma^2 = 1$)

- Then regularization term $E_w(w)$ in regularized error $\tilde{E}(w) = E_D(w) + \lambda E_w(w)$ changes from:
  - $\frac{1}{2}\sum_i w_i^2$ (ridge regn. / L_2 regularization) → $\frac{1}{2}\sum_i |w_i|$ (lasso regn. / L_1 regularization)

Recall: Our original ridge regn. obj. fn: $\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$

General prior $\mathbf{p}(\boldsymbol{w})$ that generalizes both Gaussian and Laplacian prior

$$p(\mathbf{w}|\alpha) = \left[\frac{q}{2}\left(\frac{\alpha}{2}\right)^{1/q}\frac{1}{\Gamma(1/q)}\right]^{M}\exp\left(-\frac{\alpha}{2}\sum_{j=1}^{M}|w_j|^q\right)$$

Outline for Module M5

- M6. Regression (Linear models)
  - M6.0 Introduction/context
    - Problem formulation
    - Motivating examples, incl. basis functions
  - **M6.1 Linear regression approaches**
    - Direct approach I: least-squares (error) (w_{LS})
    - Direct approach II: least-squares (error) with regularization (w_{RLS})
    - Discriminative model approach I: MLE ($w_{ML} = w_{LS}$)
    - Discriminative model approach IIa: Bayesian linear regression ($w_{MAP} = w_{RLS}$)
    - **Discriminative model approach IIb: Bayesian linear regression (posterior & predictions)**
  - M6.2 Model Complexity/Selection
    - Motivation (hyperparameter tuning to avoid overfitting)
    - Frequentist view (bias-variance decomposition)
  - M6.3 Concluding thoughts

Bayesian inference: second and third steps

- Assume Gaussian prior for $\boldsymbol{w}$ going forward.
- We don't want just a single-point estimate (MAP) of $\boldsymbol{w}$; we want

Step 2) the full posterior of $\boldsymbol{w}$, and in turn use it to

Step 3) predict t for new x (via model-averaging…

- …wherein each model is given by a particular possible value of w and the averaging weight is given by the model' s posterior)

Step 2) Let's see an example of full posterior…

$$t \cong y(x, w) = w_0 + w_1 x$$

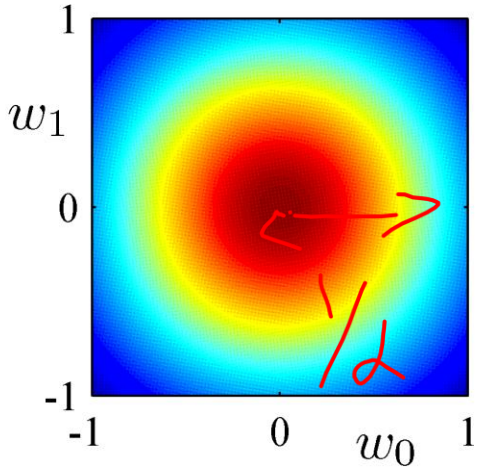
Use a given set of training data points to not just infer one optimal model specified by w_{MAP}, but all possible w models and the training dataset's support (posterior probab.) for each such model w.

Bayesian Linear Regression Example (1)

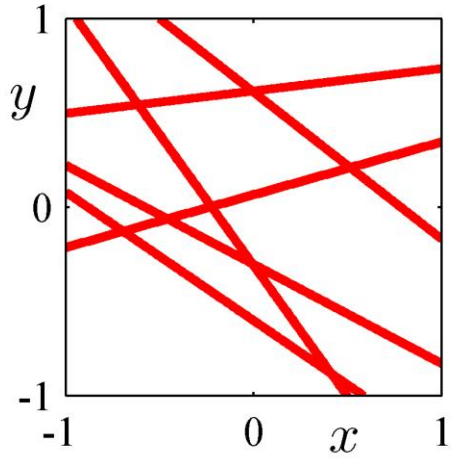
0 data points observed

$\mathcal{N}(0, \alpha^{-1} I)$

Prior



Data Space

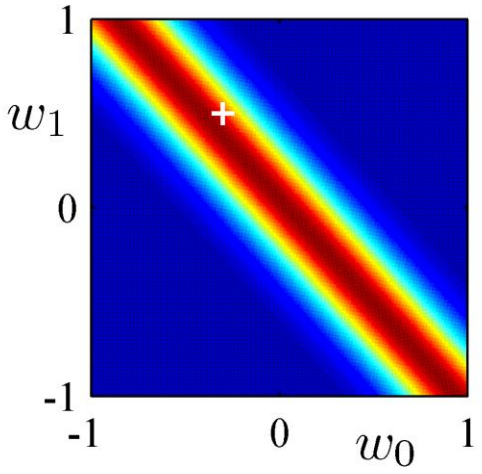


[CMB]

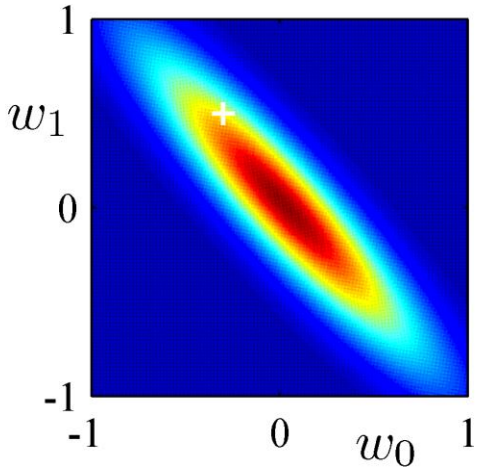
Bayesian Linear Regression Example (2)

1 data point observed

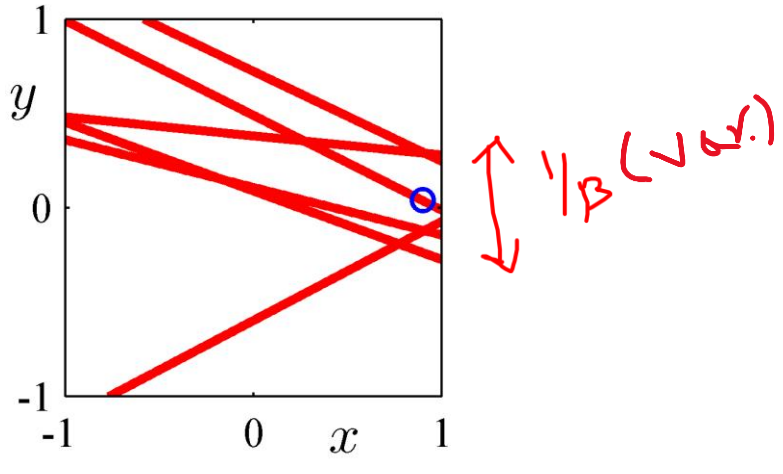
Likelihood



Posterior



Data Space



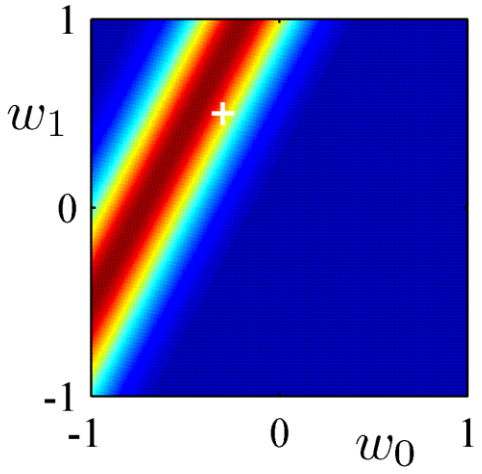
$p(w|t) \propto p(t|w)\, p(w)$

[CMB]

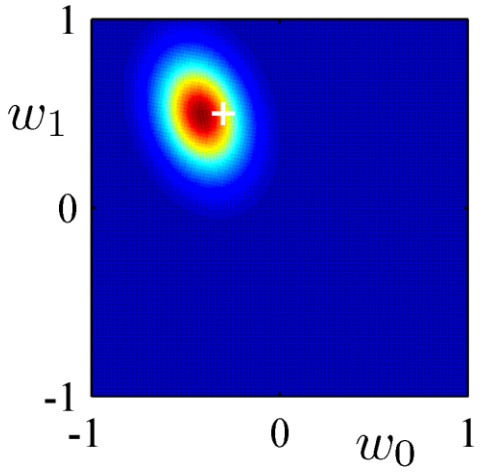
Bayesian Linear Regression Example (3)

2 data points observed

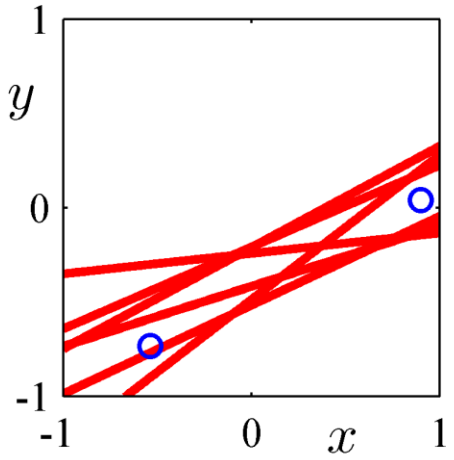
Likelihood



Posterior



Data Space

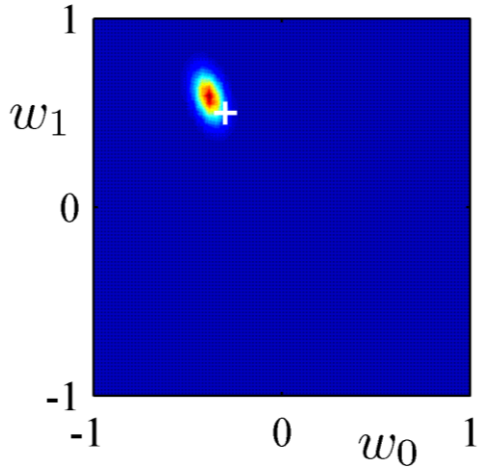


[CMB]

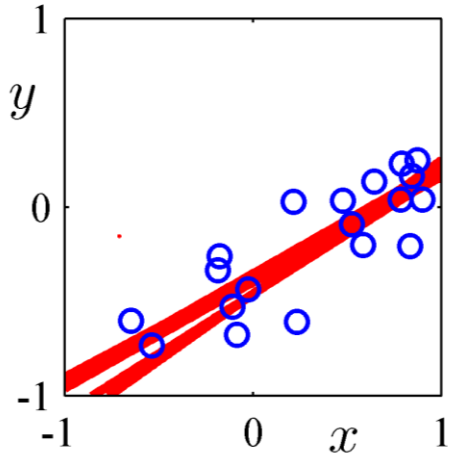
Bayesian Linear Regression Example (4)

20 data points observed

Posterior



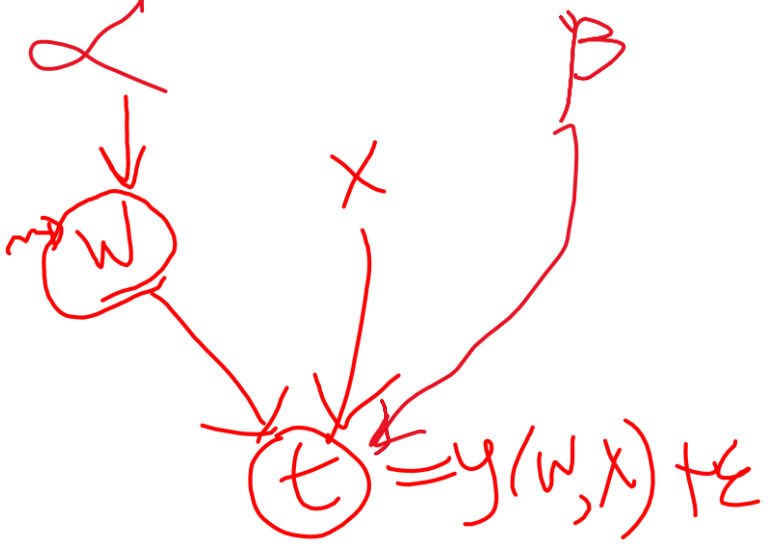
Data Space



[CMB]

From example posterior plots to Full posterior in the general case:

Bayesian Linear Regression



- Define a conjugate **prior** over $\mathbf{w}$. A common choice is

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

- Combining this with the **likelihood** function and using results for marginal and conditional Gaussian distributions, gives the **posterior**

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \mathbf{\Phi}^{\mathrm{T}} \mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}. \end{aligned}$$

[Qn. What is m_N?]

Marginal and Conditional Gaussians

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}^{-1}) \quad (2.113)$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \quad (2.114)$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \quad (2.115)$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\mathbf{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda}\boldsymbol{\mu}\}, \mathbf{\Sigma}) \quad (2.116)$$

where

$$\mathbf{\Sigma} = (\mathbf{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \quad (2.117)$$

[CMB]

Recall: MVG Handy Results (cheat-sheet)

Partitioned Gaussians

Given a joint Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ and

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{2.94}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}. \tag{2.95}$$

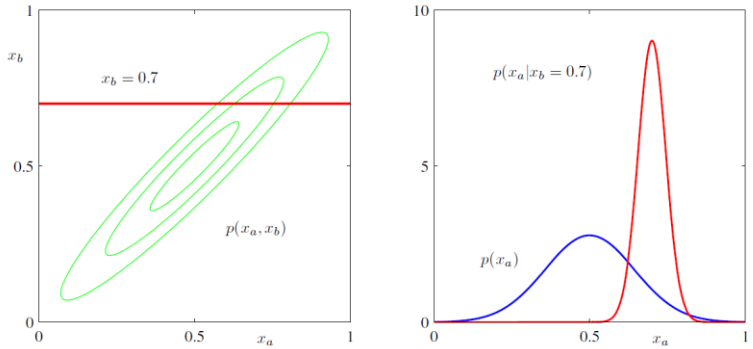
Conditional distribution:

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \tag{2.96}$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b). \tag{2.97}$$

Marginal distribution:

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}). \tag{2.98}$$



Marginal and Conditional Gaussians

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2.114}$$

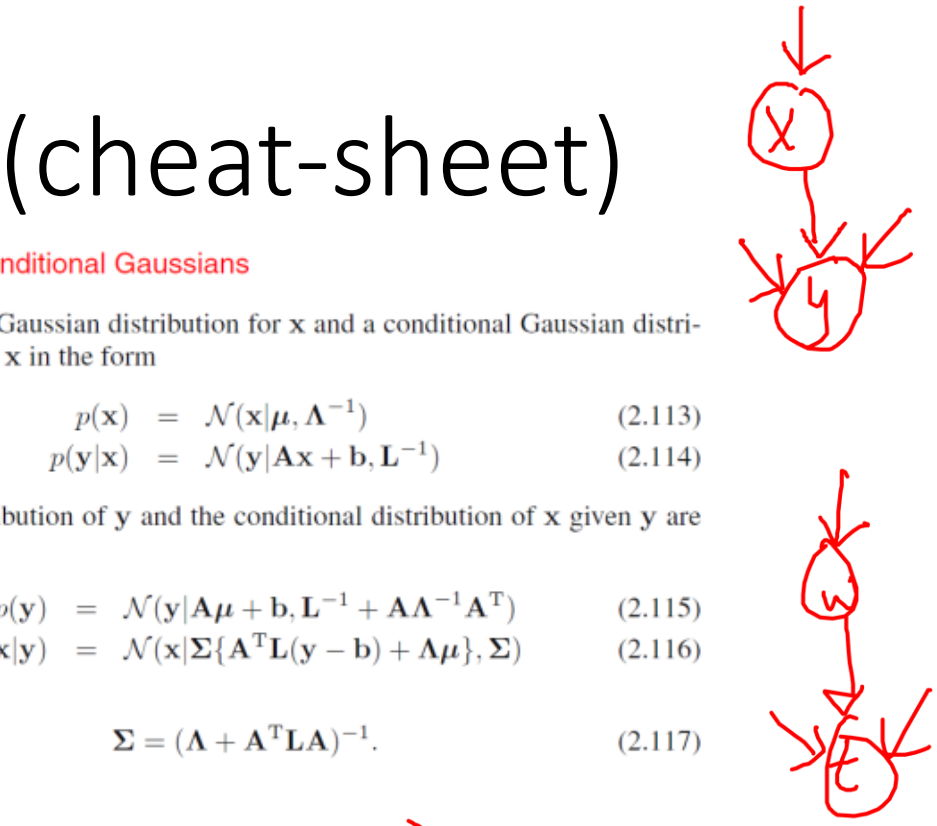
the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \tag{2.115}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \tag{2.116}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \tag{2.117}$$



Bayesian Network.

[CMB: Bishop, Chapter 2]

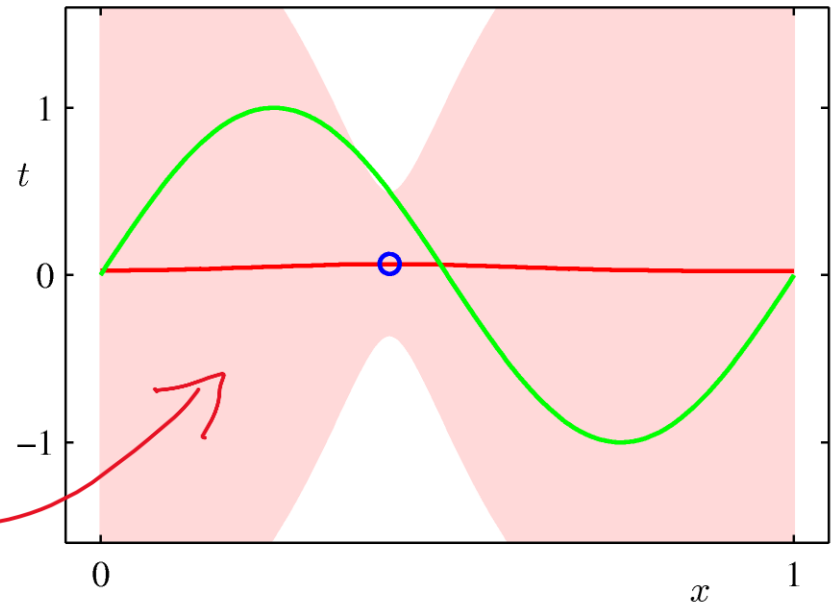
Finally, Step 3: What about predicting t for a new point x?

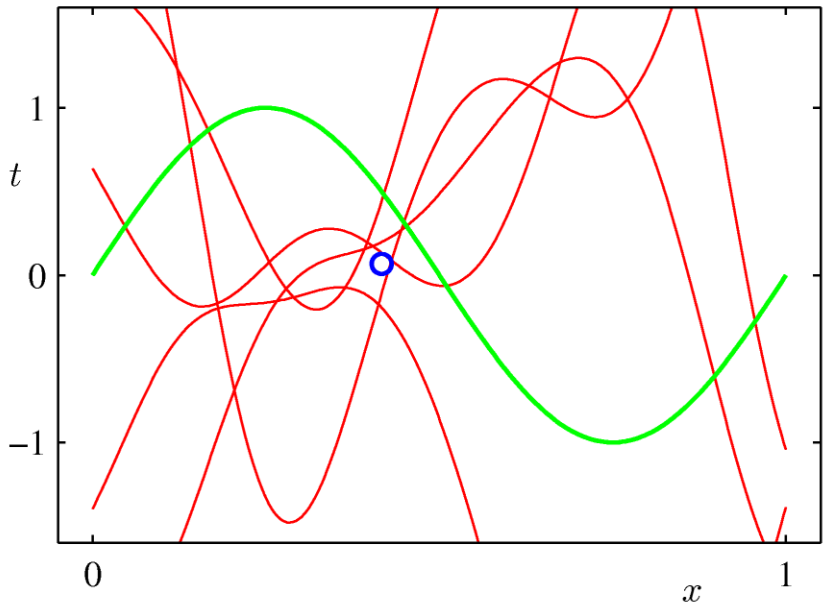
- ((we don’t want just an estimate or posterior of w; we want to use it to predict t for new x))

((again example plots first, and post. predictive distbn. form for the general case after that.))

Example Predictive Distribution (1)

- Example: Sinusoidal data, 9 Gaussian basis fns. ($\phi: \mathbb{R} \to \mathbb{R}^9$), 1 data point

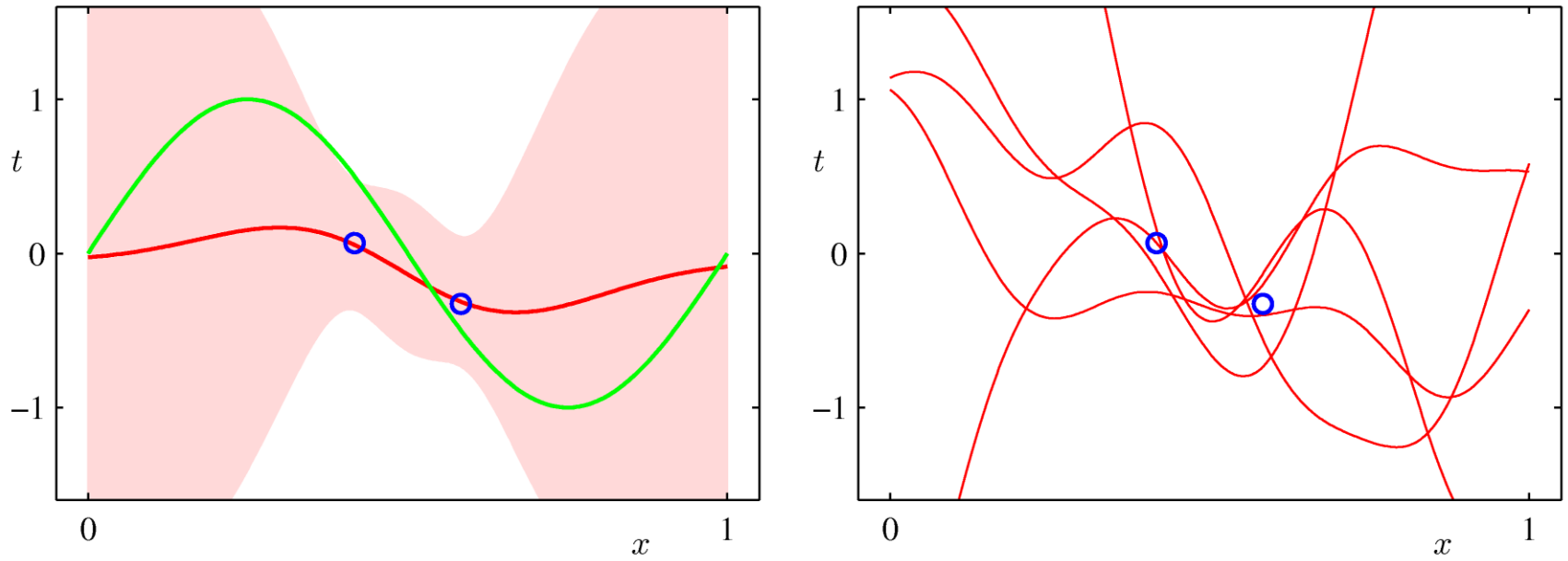




[CMB]

Example Predictive Distribution (2)

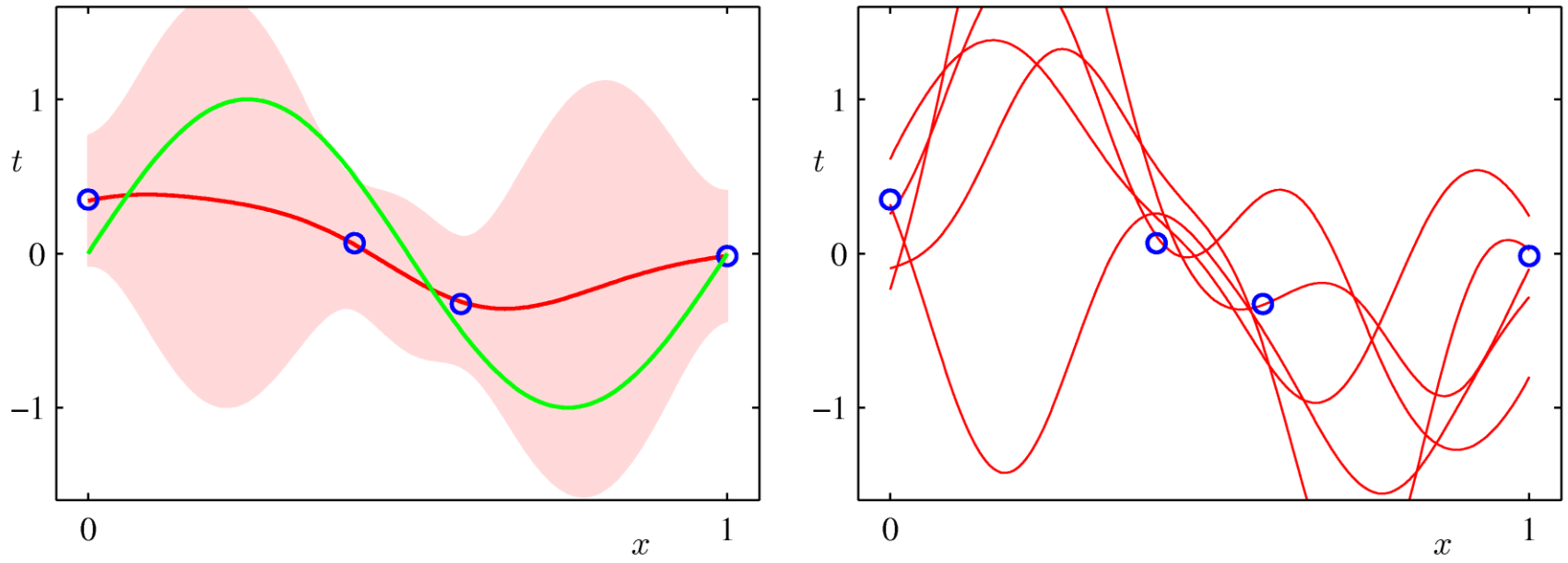
- Example: Sinusoidal data, 9 Gaussian basis functions, 2 data points



[CMB]

Example Predictive Distribution (3)

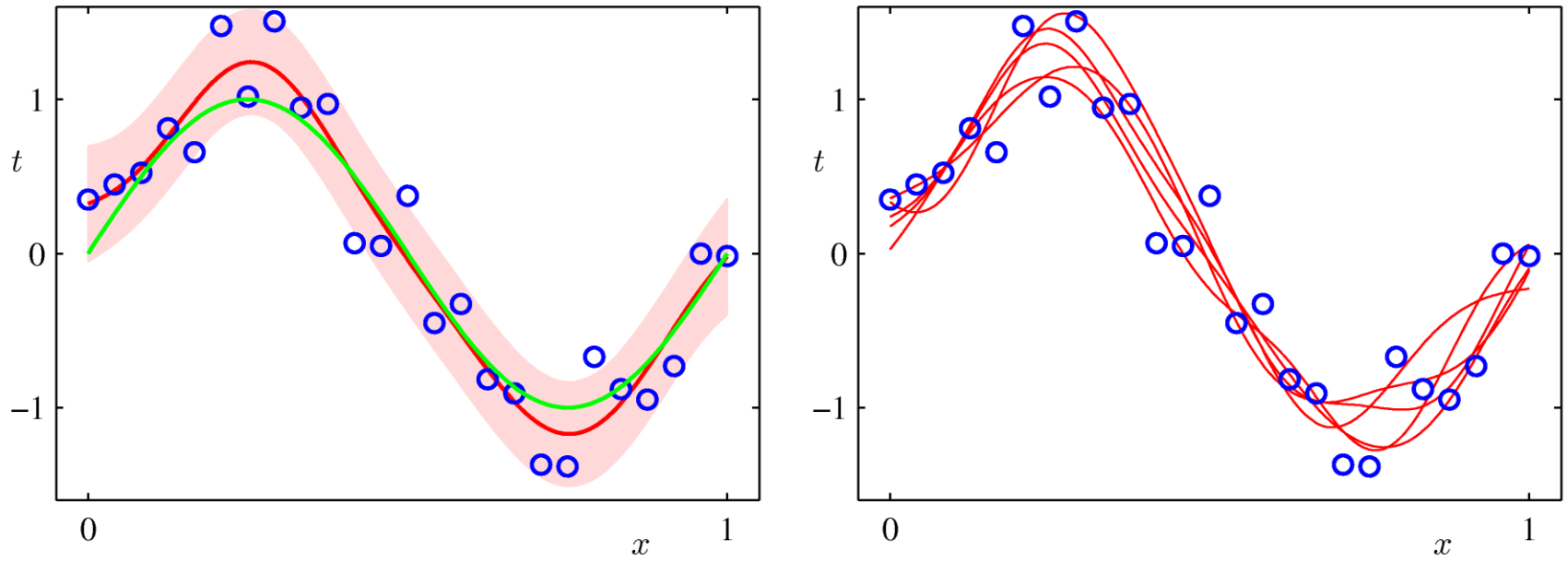
- Example: Sinusoidal data, 9 Gaussian basis functions, 4 data points



[CMB]

Example Predictive Distribution (4)

- Example: Sinusoidal data, 9 Gaussian basis functions, 25 data points



[CMB]

(Bayesian/Posterior) Predictive Distribution (1)

- Predict t for new values of x by integrating over w (model-averaging)

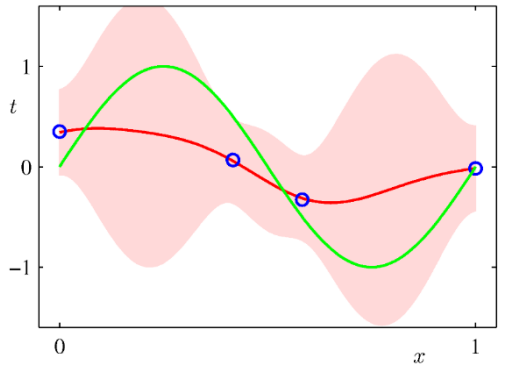
$$p(t \mid x, \mathbf{x}, \mathbf{t}, \alpha, \beta) = \int p(t \mid x, \mathbf{w}, \beta)\, p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)\, \mathrm{d}\mathbf{w}$$
$$= \mathcal{N}\left(t \mid \mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(x), \sigma_N^2(x)\right)$$

post-pred. D_N post. w

- where

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}(x)$$

(ML) (extra due to uncert in w)



$$\mathbf{m}_N = \beta \mathbf{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$$
$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}.$$

Exercise: Prove that $\mathrm{m_N} = w_{RLS} = w_{MAP}$ (recall $\lambda := \alpha/\beta$), and hence that the posterior predictive mean is same as that of the predicted value in the direct RLS approach. [CMB]

Linear regression direct vs. discriminative model approaches: **summary**

Discriminative model-based approaches have two advantages:

1. Convert intuition for obj. fns. to probabilistic model driven motivations:

   (Least-squares or min $E(w)$) $w_{LS} = w_{ML}$ (MLE)

   (Reg. Least-squares or min $\tilde{E}(w)$) $w_{RLS} = w_{MAP}$ (Bayesian)

2. Give additional advantage of capturing the uncertainty over the predicted values, viz., a predictive **distribution**

$$p(t|x) = N(t\,|y(x, w_*) \coloneqq w_*^T \phi(x), \mathbf{var}),$$ (and **not just a single predicted value** $y(x, w_*)$ as in the direct approach).

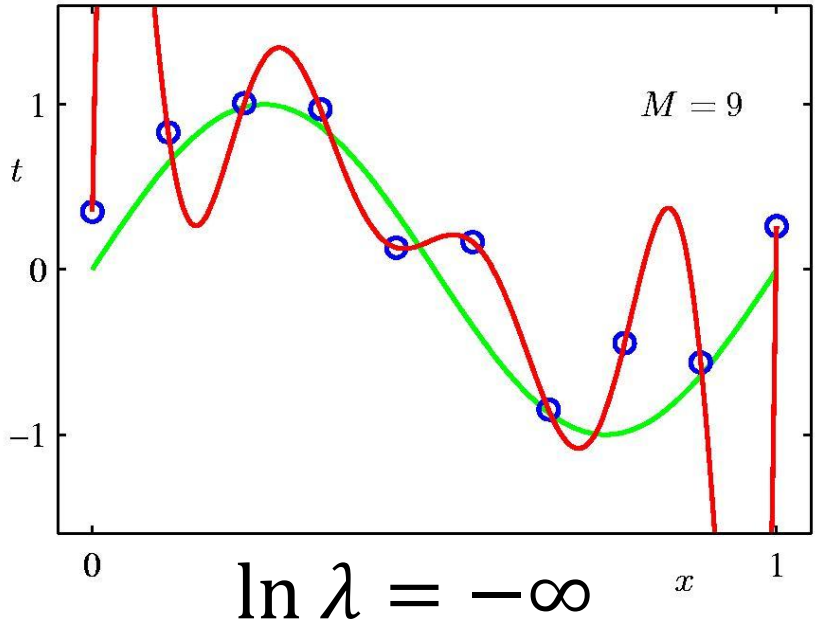
- In MLE, (pred.) **var** is a dataset-wide single variance ($\sigma^2 = \beta^{-1}$)
- In Bayesian, (post. pred.) **var** is datapoint-specific ($\sigma^2(x) = \beta^{-1} + \phi(x)^T S_N \phi(x)$)

Outline for Module M5

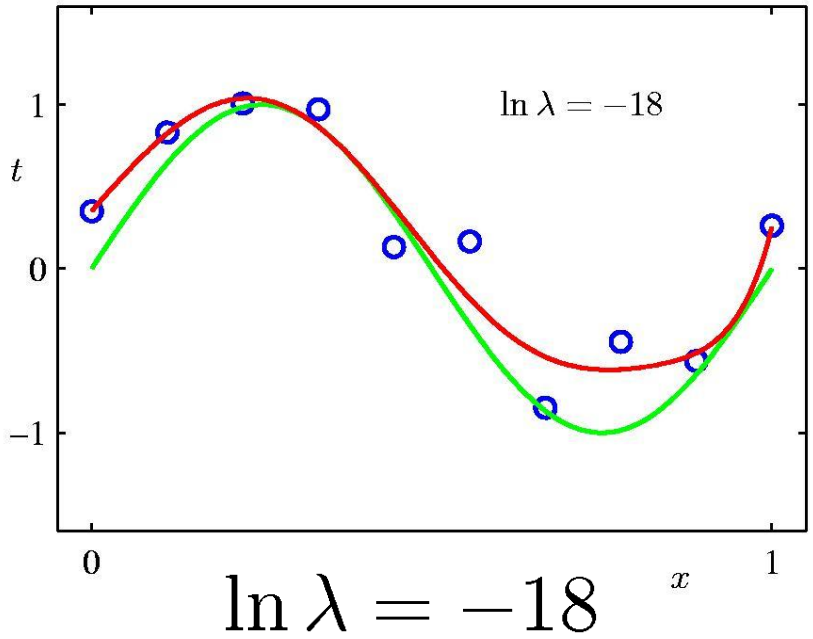
- M6. Regression (Linear models)
  - M6.0 Introduction/context
    - Problem formulation
    - Motivating examples, incl. basis functions
  - M6.1 Linear regression approaches
    - Direct approach I: least-squares (error) (w_{LS})
    - Direct approach II: least-squares (error) with regularization (w_{RLS})
    - Discriminative model approach I: MLE ($w_{ML} = w_{LS}$)
    - Discriminative model approach IIa: Bayesian linear regression ($w_{MAP} = w_{RLS}$)
    - Discriminative model approach IIb: Bayesian linear regression (posterior & predictions)
  - **M6.2 Model Complexity/Selection**
    - **Motivation (hyperparameter tuning to avoid overfitting)**
    - Frequentist view (bias-variance decomposition)
  - M6.3 Concluding thoughts

Recall: Motivating example for Regularization, and Hyperparam. Tuning

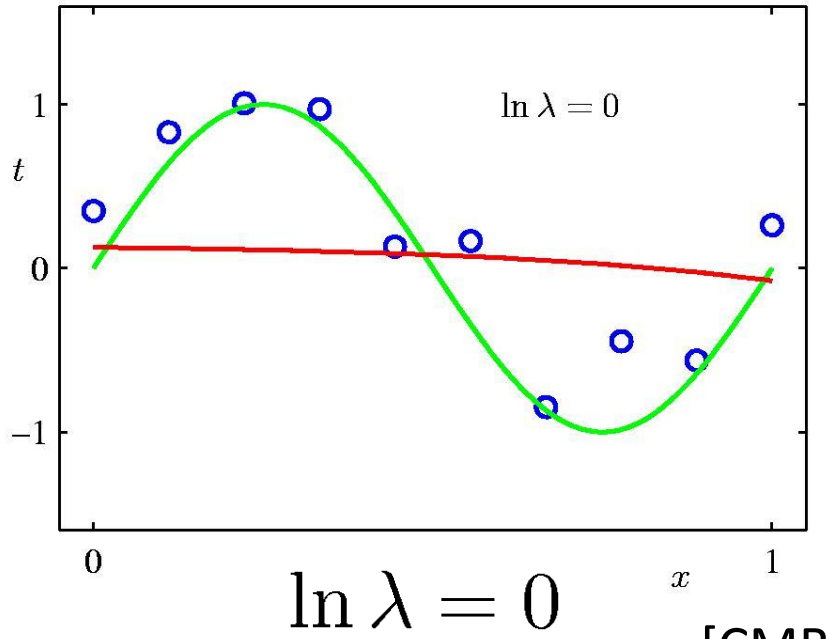
$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$



$\ln \lambda = -\infty$



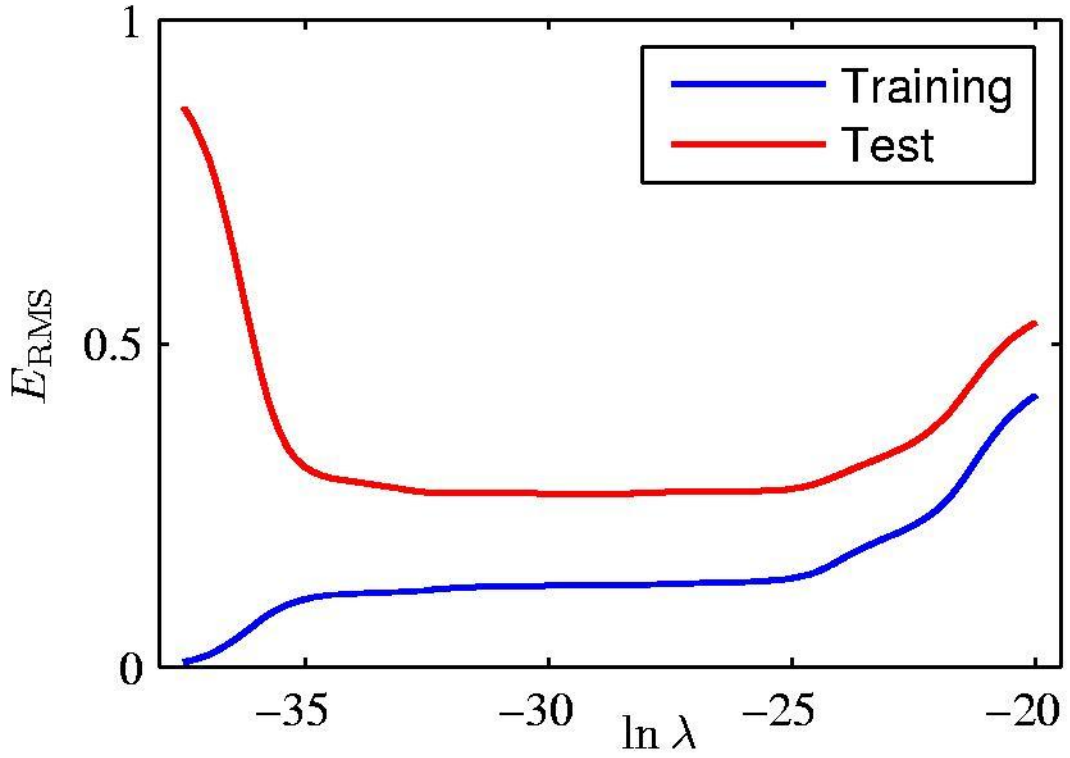
$\ln \lambda = -18$



$\ln \lambda = 0$

[CMB]

Regularization, and Hyperparam. tuning



Regularization penalizes large coefs.
of 9th order polynomial fit of the data.

more complex model ← less complex model

[CMB]

Some Motivating Questions

- Can we understand the error in our predictions better?
  - That is, can we identify the different components of the error in our predictions?
  - How are these different components related to the complexity of our model, and to overfitting?
- Can we use above knowledge to better tune model complexity (hyperparameters) to avoid overfitting?
- If the trends in data require fitting of a complex model, then
  - can overfitting be detected by understanding the stability of the optimal (frequentist/ML) model across different training datasets?
  - can a Bayesian model overcome the overfitting "naturally/implicitly" by not settling in on a single optimal model and instead averaging over multiple models?
    - Bayesian view (model averaging & empirical Bayes) possible, but out of scope for this course.

Outline for Module M5

- M6. Regression (Linear models)
  - M6.0 Introduction/context
    - Problem formulation
    - Motivating examples, incl. basis functions
  - M6.1 Linear regression approaches
    - Direct approach I: least-squares (error) (w_{LS})
    - Direct approach II: least-squares (error) with regularization (w_{RLS})
    - Discriminative model approach I: MLE ($w_{ML} = w_{LS}$)
    - Discriminative model approach IIa: Bayesian linear regression ($w_{MAP} = w_{RLS}$)
    - Discriminative model approach IIb: Bayesian linear regression (posterior & predictions)
  - **M6.2 Model Complexity/Selection**
    - Motivation (hyperparameter tuning to avoid overfitting)
    - **Frequentist view (bias-variance decomposition)**
  - **M6.3 Concluding thoughts**

Recall: Decision Theory for Regression min. squared loss (cond. expn. as minimizer)

$$E[L] = E_x\left[E_{t|x}\left[(y(x)-t)^2\right]\right]$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \mathrm{var}\,[t|\mathbf{x}]\, p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$(y(x;\mathcal{D}) - h(x))^2 \qquad \text{noise}$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

[CMB]

Bias-variance analysis proof

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x})\,\mathrm{d}t.$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

Goal: Decompose average error $E_D[E_{x,t}[L]]$ into different terms.

Now, simply view "y(x;D) – h(x)" as a random variable Z; and apply the variance formula: $Var_D(Z) = E_D[Z^2] - (E_D[Z])^2$ to get the bias-variance decomposition of error below:

$$\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}.$$

[CMB]

Bias-variance decomposition in formula

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

where

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

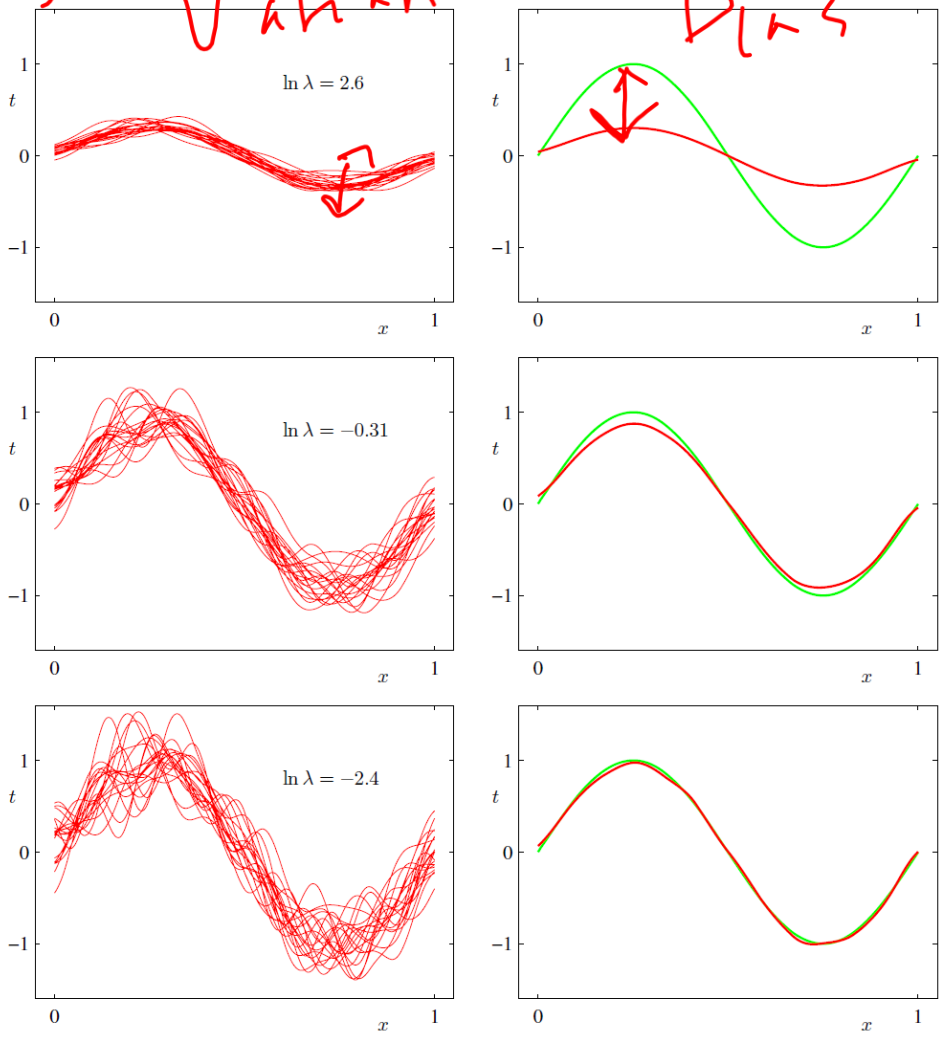
$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x})\,\mathrm{d}t.$$

expected loss = $E_D\left[E_{x,t}[L]\right] = E_D\left[E_{x,t}[(y(x;D) - t)^2]\right]$

Exercise: cf. worksheet for careful understanding of what random variables the expectation above is taken over!

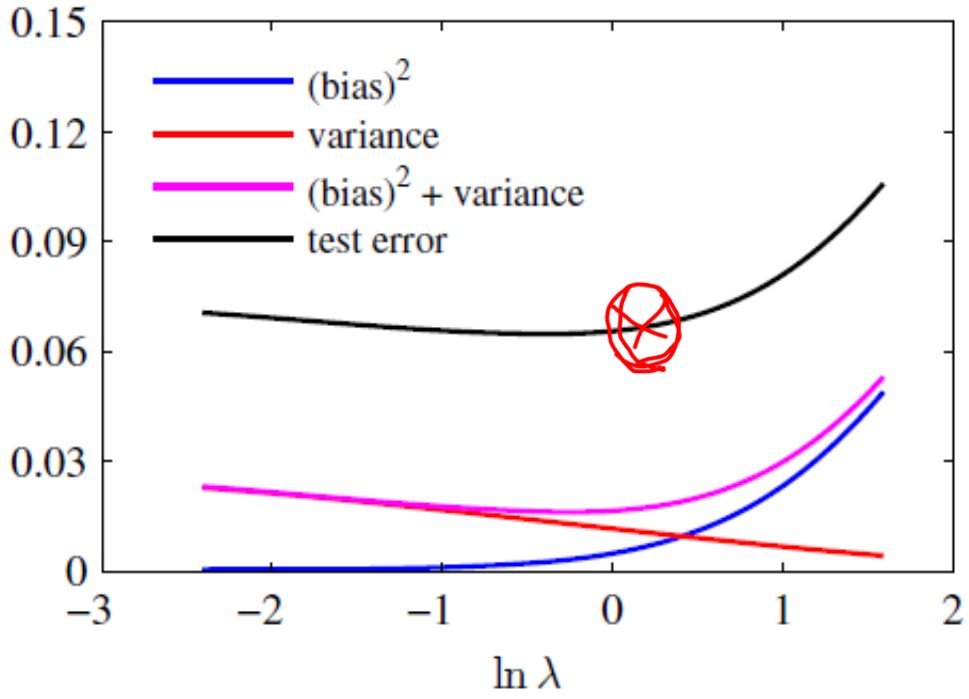
[CMB]

Bias-variance in pictures (for an example)



[CMB]

Bias-variance analysis (for the example)



hi ← model cpt → lo

expected loss $= (\text{bias})^2 + \text{variance} + \text{noise}$

[CMB]

Bias-variance anals.: applicability in practice?

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

where

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

$h(x) = \mathbb{E}[t|x]$??

[CMB]

Bias-variance anals.: applicability in practice? details

$E_D[E_{x,t}[L]]$

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

where

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

$E_X(\cdot)$

$h(x) = E[t|x]$??

$$\overline{y}(x) = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$$

$$(\text{bias})^2 = \frac{1}{N}\sum_{n=1}^{N} \{\overline{y}(x_n) - h(x_n)\}^2$$

$$\text{variance} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L} \{y^{(l)}(x_n) - \overline{y}(x_n)\}^2$$

$\{x_n\}_{n=1}^{N} \overset{iid}{\sim} p(x)$ E_X E_D

[CMB]

Concluding thoughts

- Linear regression forms the foundations of other sophisticated methods, so it is good to invest enough time on it.
  - Two views: direct loss fn. view (**E(w)/regularization**) & probabilistic model view (**MLE/Bayesian**)
  - But lin. regn. has limitations in practice, even with non-linear basis functions, closed-form solutions and other analytical advantages.
  - Mainly because basis functions are fixed before seeing the training data (curse of dimensionality when dimensionality of feature vectors D grows).

- Next steps:
  - linear models for classification, which play similar basic role for other classification methods.
  - Move from fixed basis fns to selection of basis functions or adaptation of basis functions using training data, in later lectures on non-linear models.

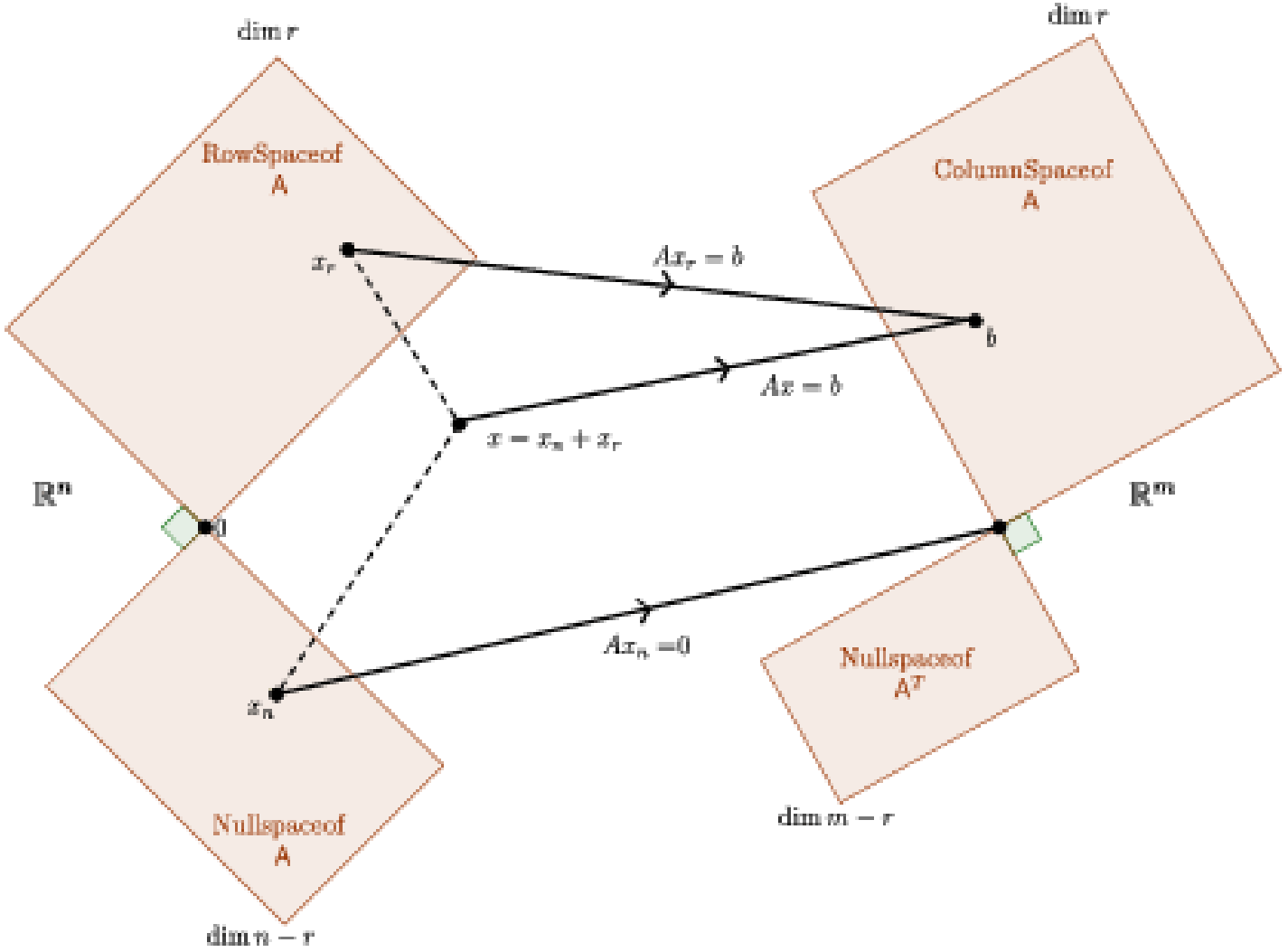
Thank you!

Backup slides

Linear Algebra (LA) Refresher

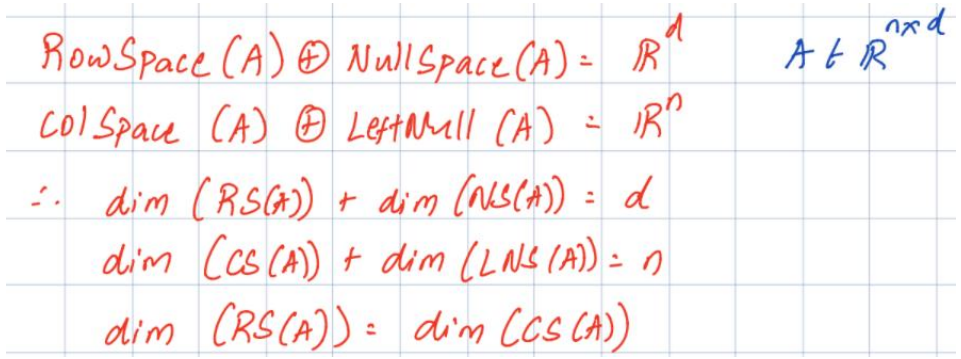
- Switch to LN Pandey’s notes

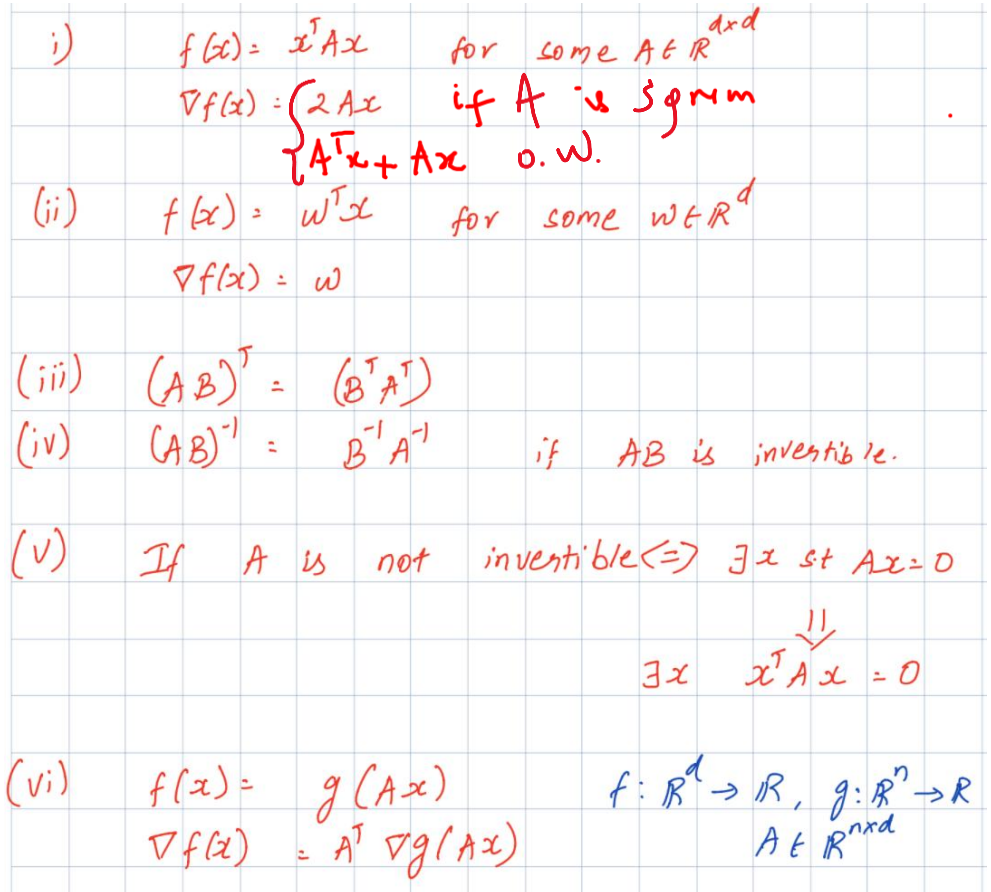
LA Cheat Sheet: The four subspaces of a matrix



[From https://mathworld.wolfram.com/FundamentalTheoremofLinearAlgebra.html, Strang LA book]

LA + Opt. Cheat Sheet

- Real, symmetric matrices S can be diagonalized as $S = Q\Lambda Q^T$
  - S is psd iff all its eigen values are non-negative.
- 

- 

[HR]

Recall LA: To solve $Ax = b$, we premult. by A^T, and simply solve $A^T Ax = A^T b$.

Linear Equation Solving

$Ax = b$

(X) → No solution if $b \notin$ columnspace (A)

→ Unique solution if $b \in$ columnspace (A) & A has lin. ind columns

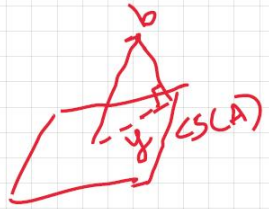
→ Infinite solutions if $b \in CS(A)$ & A has lin. dep. columns.

$$\begin{bmatrix} 1 & 3 \\ 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

(X) Least-squares soln possible. (related to lin. regn)

$\min_x \|Ax - b\|^2 == \min_{y \in CS(A)} \|y - b\|^2$

$(Ax^* = y^*)$ (y^* is Proj of b onto CS)



$$\min_{x \in \mathbb{R}^d} \underbrace{\left[(Ax-b)^T(Ax-b)\right]}_{f(x)}$$

$$\nabla f(x) = 2A^T(Ax-b) \overset{\text{set to}}{=} 0$$

$$\Rightarrow A^T(Ax^*-b) = 0 \quad \text{(normal eqn.)}$$

$$\Rightarrow A^T A x^* = A^T b$$

$$\Rightarrow x^* = (A^T A)^{-1} A^T b$$

Ex.: Prove:

1) at least one soln. x* exists for the normal eqn.
2) soln. x* unique if $(A^T A)$ is invertible ($\Leftrightarrow A$ has lin. indep. cols.)
3) **infinite** solns. x* if $(A^T A)$ is **non**-invertible ($\Leftrightarrow A$ has lin. **dep.** cols.)

Ex.:

i) Prove $NS(A) = NS(A^T A)$.

ii) Use orthog. complementarity of $NS(A^T), CS(A)$ to derive normal eqns.

Choice of Prior

Different priors on parameter $\theta (:= w)$ leads to…

$$P(\boldsymbol{\theta} \mid \sigma^2) = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\theta_i^2}{2\sigma^2}\right\},$$

$$P_{Laplacian}(\theta_i \mid \beta) = \frac{1}{2\beta} \exp\left\{-\frac{|\theta_i|}{\beta}\right\}.$$

$$P(\boldsymbol{\theta} \mid \beta) = \prod_{i=1}^{M} \frac{1}{2\beta} e^{-\frac{|\theta_i|}{\beta}}$$

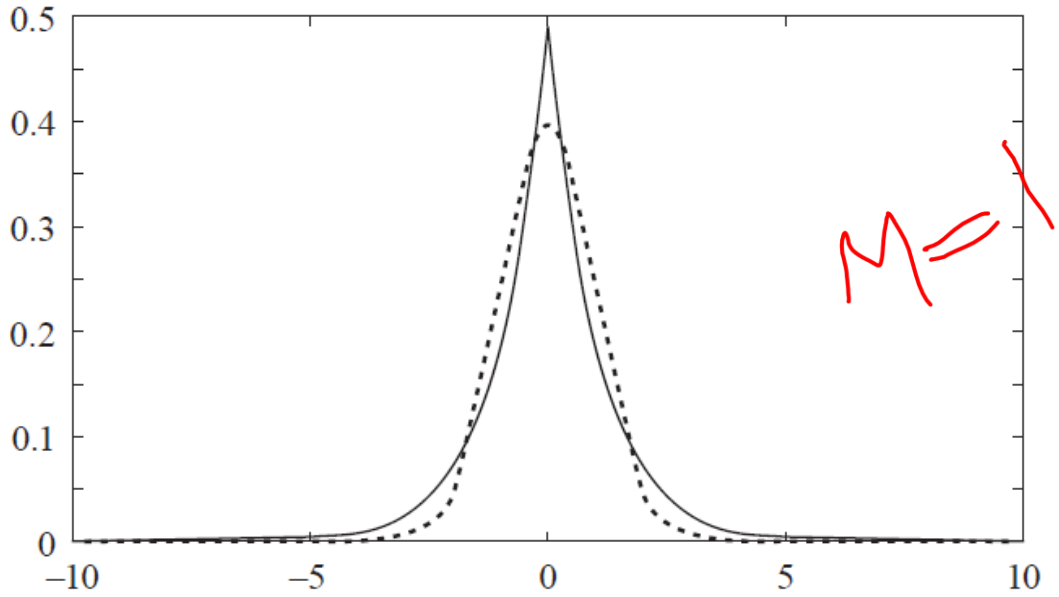


Figure 20.3 Laplacian distribution ($\beta = 1$) **and Gaussian distribution** ($\sigma^2 = 1$)

…different regularizations (ridge vs. lasso regn.)

$$\text{Gauss.} \sum_{i=1}^{N} + \frac{\theta_i^2}{2\sigma^2} \quad \text{min} \quad \text{ridge}$$

$$\text{Lap.} \sum_{i=1}^{N} + \frac{|\theta_i|}{\beta} \quad \text{min.} \quad \text{lasso}$$

Bias-variance analysis (alternate proof)

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x})\, \mathrm{d}t.$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t.$$

$$\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2.$$

$$\begin{aligned}
&\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\
&\quad = \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\
&\qquad +2\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}.
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2\right] \\
&\quad = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2\right]}_{\text{variance}}.
\end{aligned}$$

[CMB]