CONJUGATE PRIORS AND LEAST SQUARES

Sergey Nikolenko

Harbour Space University, Barcelona, Spain
March 14, 2017

CONJUGATE PRIORS

REMINDER

- Recall that we are trying to learn the parameters of a distribution and/or predict the next points by the data we have.
- Bayesian inference includes:
    - $p(x \mid \theta)$ – likelihood of the data;
    - $p(\theta)$ – prior distribution;
    - $p(x) = \int_\Theta p(x \mid \theta)p(\theta)d\theta$ – marginal likelihood;
    - $p(\theta \mid x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ – posterior distribution;
    - $p(x' \mid x) = \int_\Theta p(x' \mid \theta)p(\theta \mid x)d\theta$ – predictive distribution.
- The problem is usually to find $p(\theta \mid x)$ and/or $p(x' \mid x)$.
- How do we choose $p(\theta)$?

- Reasonable idea: let's choose prior distributions in such a way that they would have the same form *a posteriori*.
- Before the inference we have a prior distribution $p(\theta)$.
- After, we have a new posterior distribution $p(\theta \mid x)$.
- Let us try to get $p(\theta \mid x)$ to have the same form as $p(\theta)$, just with other parameters.

- A not quite formal definition: a family of distributions $p(\theta \mid \alpha)$ is called a family of *conjugate priors* for a family of likelihoods $p(x \mid \theta)$, if after multiplication by a likelihood the posterior distribution $p(\theta \mid x, \alpha)$ remains in the same family: $p(\theta \mid x, \alpha) = p(\theta \mid \alpha')$.
- α are called *hyperparameters*, "parameters of the distribution of parameters".
- Trivial example: the family of all distributions will be conjugate to anything.

- Naturally, the form of a good conjugate prior depends on the form of the likelihood $p(x \mid \theta)$.
- Conjugate priors are known for many distributions.

- What is the conjugate prior for tossing an unfair coin (Bernoulli priors)?

- What is the conjugate prior for tossing an unfair coin (Bernoulli priors)?
- It is the *beta distribution*; the density of the distribution on θ is

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

- The distribution density for the coin parameter θ is

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

- Then, if we sample the coin and get s heads and f tails, we get

$$p(s, f \mid \theta) = \binom{s+f}{s}\theta^{s}(1-\theta)^{f}, \text{ so}$$

$$\begin{aligned} p(\theta|s, f) &= \frac{\binom{s+f}{s}\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}/B(\alpha, \beta)}{\int_0^1 \binom{s+f}{s}x^{s+\alpha-1}(1-x)^{f+\beta-1}/B(\alpha, \beta)dx} = \\ &= \frac{\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}}{B(s+\alpha, f+\beta)}. \end{aligned}$$

- Thus, we get that the conjugate prior for the parameter of an unfair coin θ is

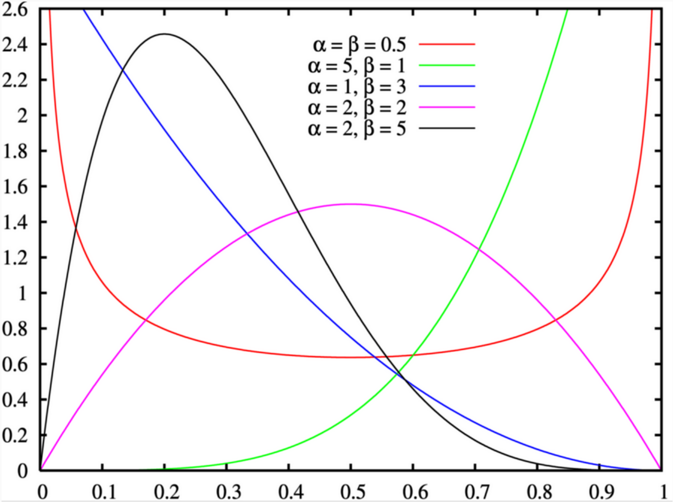
$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- After getting new data with s heads and f tails, the hyperparameters change to

$$p(\theta \mid s+\alpha, f+\beta) \propto \theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}.$$

- At this stage, we can forget about complicated formulas, we have found a very simple learning rule.

BETA DISTRIBUTION



- Simple generalizatoin: consider the multinomial distribution with n trials, k categories, and suppose that x_i of experiments fell into category i.
- Parameters θ_i show the probability of getting into category i:

$$p(x \mid \theta) = \binom{n}{x_1, \dots, x_n} \theta_1^{x_1} \theta_2^{x_2} \dots \theta_k^{x_k}.$$

- The conjugate prior here is the Dirichlet distribution:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1}.$$

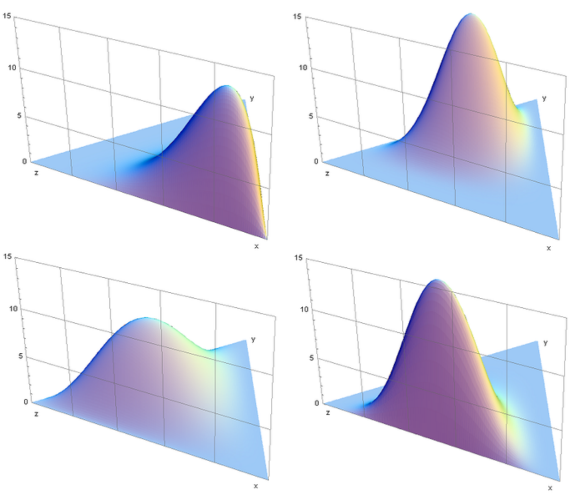
- The conjugate prior here is the Dirichlet distribution:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1}.$$

Exercise. Prove that after getting the data $x_1, \dots, x_k$ hyperparameters change into

$$p(\theta \mid x, \alpha) = p(\theta \mid x + \alpha) \propto \theta_1^{x_1 + \alpha_1 - 1} \theta_2^{x_2 + \alpha_2 - 1} \dots \theta_k^{x_k + \alpha_k - 1}.$$

DIRICHLET DISTRIBUTION



LEAST SQUARES ESTIMATION

- Linear model: consider a linear function

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{p} x_j w_j = \mathbf{x}^\top \mathbf{w}, \quad \mathbf{x} = (1, x_1, \dots, x_p).$$

- For a vector of inputs $\mathbf{x}^\top = (x_1, \dots, x_p)$ we will predict the output y as

$$\hat{y}(\mathbf{x}) = \hat{w}_0 + \sum_{j=1}^{p} x_j \hat{w}_j = \mathbf{x}^\top \hat{\mathbf{w}}.$$

- How do we find optimal parameters $\hat{\mathbf{w}}$ by training data of the form $(\mathbf{x}_i, y_i)_{i=1}^{N}$?
- Least squares estimation: let us minimize

$$\mathrm{RSS}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \mathbf{w})^2.$$

- How would you minimize this function?

- Actually, we can do it exactly:

$$\mathrm{RSS}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}),$$

  where $\mathbf{X}$ is an $N \times p$ matrix, differentiate w.r.t. $\mathbf{w}$, get

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

  if $\mathbf{X}^\top \mathbf{X}$ is nondegenerate.
- $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top$ is called the *Moore–Penrose pseudo-inverse* of matrix $\mathbf{X}$; the correct generalization of the notion of inverse to non-square matrices.
- By the way, how do you take derivatives (gradients) with respect to vectors?
- How many points do we need to train this model?

- Let us now try to formalize linear regression in the framework of Bayesian inference.
- Main assumption: the noise (error in the data) is distributed normally, i.e., variable t that we observe is

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

  In other words,

$$p(t \mid \mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(t \mid y(\mathbf{x}, \mathbf{w}), \sigma^2).$$

- Here y can be an arbitrary function.

- Btw, a natural generalization (not even a generalization) is to consider linear regression with feature functions:

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

(M parameters, $M-1$ feature functions, $\phi_0(\mathbf{x}) = 1$).

- Feature functions ϕ_i can be
  - the result of some separate feature extraction process;
  - extension of the linear model to nonlinear dependencies (e.g., $\phi_j(x) = x^j$);
  - local functions that are significantly nonzero only in a small region, e.g., Gaussian feature functions $\phi_j(\mathbf{x}) = e^{-\frac{(x-\mu_j)^2}{2s^2}}$);
  - ...

- Consider a dataset $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ with correct answers $\mathbf{t} = \{t_1, \dots, t_N\}$.
- We assume that the data points are independent identically distributed:

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \mathbf{w}^\top \phi(\mathbf{x}_n), \sigma^2\right).$$

- We take the logarithm (we omit $\mathbf{X}$ below for brevity):

$$\ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\right)^2.$$

- We take the logarithm (we omit $\mathbf{X}$ below for brevity):

$$\ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\top\phi(\mathbf{x}_n)\right)^2 .$$

- And we see that to maximize the likelihood w.r.t. $\mathbf{w}$ we need to minimze mean squared error!

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = \frac{1}{\sigma^2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\top\phi(\mathbf{x}_n)\right)\phi(\mathbf{x}_n).$$

- Solving the system of equations $\nabla \ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = 0$, we get the same result as above:

$$\mathbf{w}_{ML} = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top \mathbf{t}.$$

- Здесь $\Phi = (\phi_j(\mathbf{x}_i))_{i,j}$.

- Now we can also maximize the likelihood w.r.t. σ^2; we get

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - \mathbf{w}_{ML}^\top \phi(\mathbf{x}_n)\right)^2,$$

  i.e., sample variance of the data around the predicted value.

REGULARIZATION AS A PRIOR

- Bayes theorem:
$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}.$$
- Two main problems of Bayesian inference:
  - find the posterior distribution
  $$p(\theta \mid D) \propto p(D|\theta)p(\theta)$$
  (and/or find the maximal a posteriori hypothesis $\arg\max_\theta p(\theta \mid D)$);
  - find the predictive distribution:
  $$p(x \mid D) \propto \int_{\theta\in\Theta} p(x \mid \theta)p(D|\theta)p(\theta)\mathrm{d}\theta.$$
- We already know that least squares estimation corresponds to maximal likelihood for normally distributed noise.

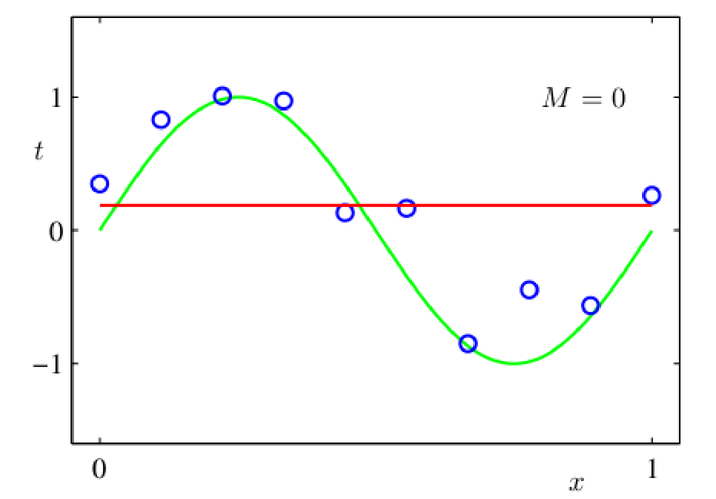
- We considered regression with feature functions:

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}).$$

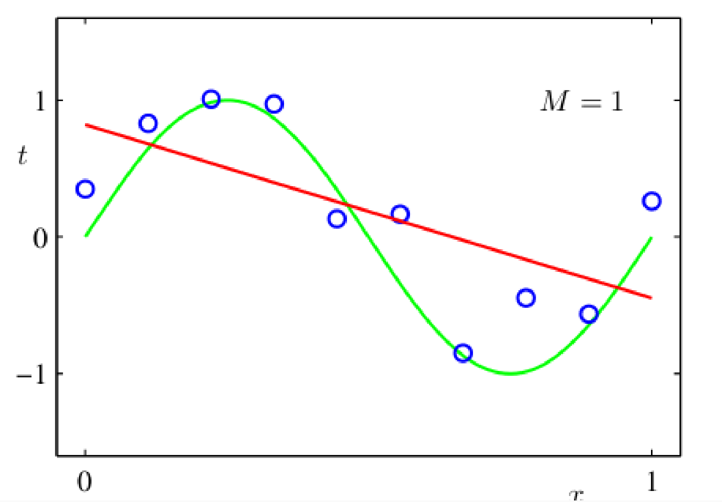
- Let us see an example of such a regression for $\phi_j(x) = x^j$, i.e.,

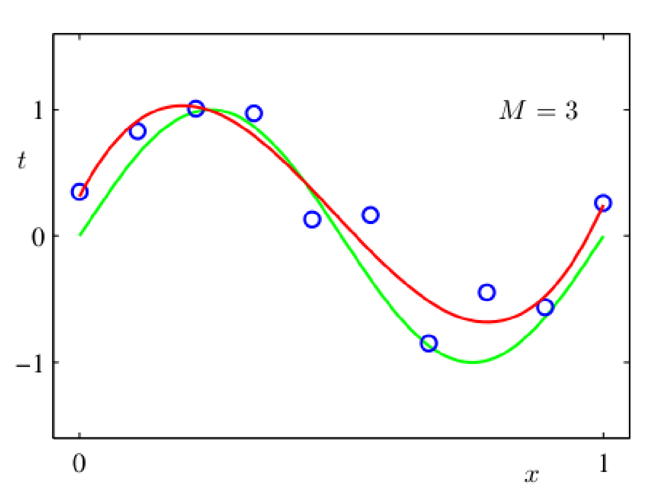
$$f(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M.$$

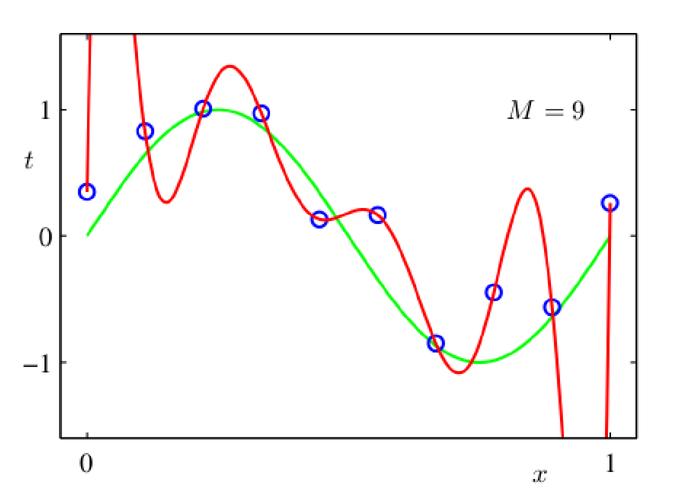
- And we will minimize the mean squared error, as above.



$M = 0$
t
1
0
−1
0
1
x

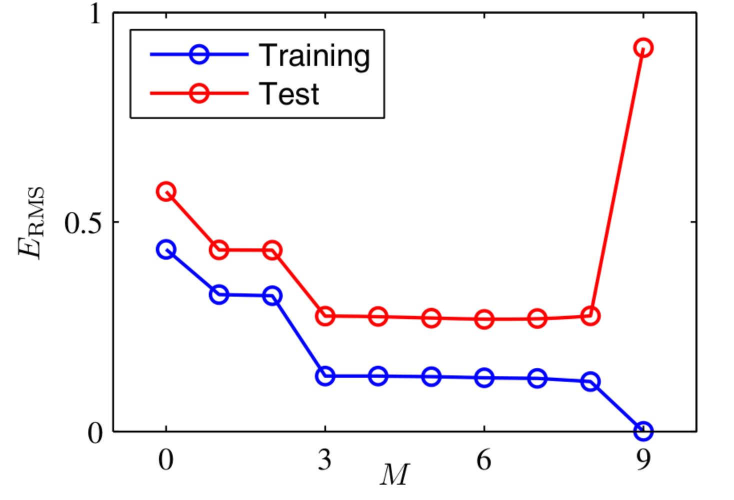

$M = 1$
t
x
1
0
−1
0
1


$M = 3$
1
0
−1
t
0
1
x

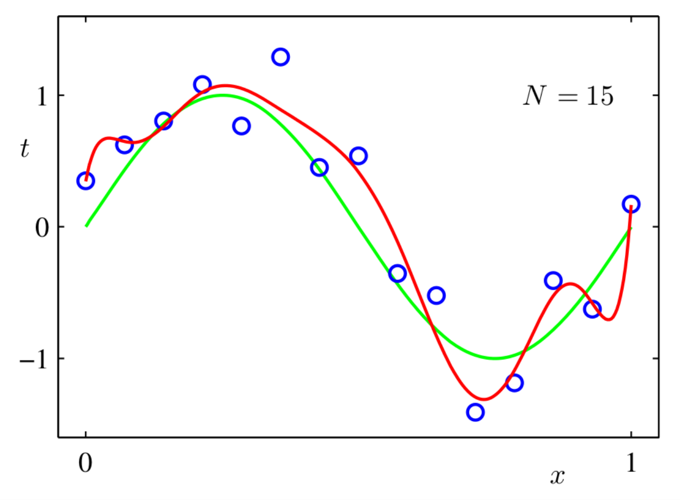


$M = 9$
t
x
1
0
−1
0
1

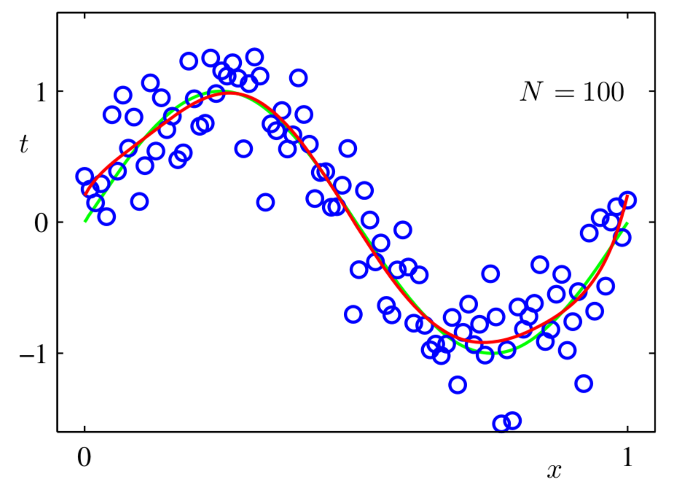
RMS VALUES



IF WE CAN COLLECT MORE DATA...



IF WE CAN COLLECT MORE DATA...



VALUES OF COEFFICIENTS

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

- We see that coefficients grow a lot; this is very improbable.
- Let's try to combat this in a very straightforward way: add the size of the coefficients to the error function.

- Before (for test examples $\{(x_i, y_i)\}_{i=1}^{N}$):

$$\mathrm{RSS}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i, \mathbf{w}) - y_i)^2.$$

- After:

$$\mathrm{RSS}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i, \mathbf{w}) - y_i)^2 + \frac{\alpha}{2}\left\|\mathbf{w}\right\|^2,$$

  where α is the regularization coefficient (we now have to choose it somehow).
- How do we optimize this error function?

- Exactly the same: write

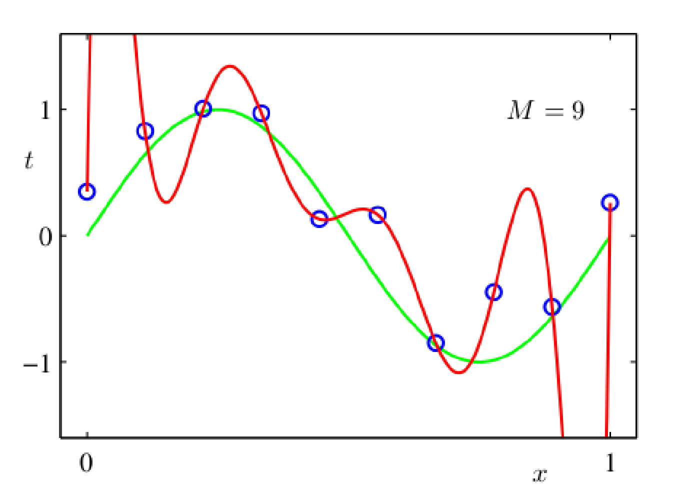
$$\mathrm{RSS}(\mathbf{w}) = \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)^\top \left(\mathbf{y} - \mathbf{X}\mathbf{w}\right) + \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}$$

and take the derivative:

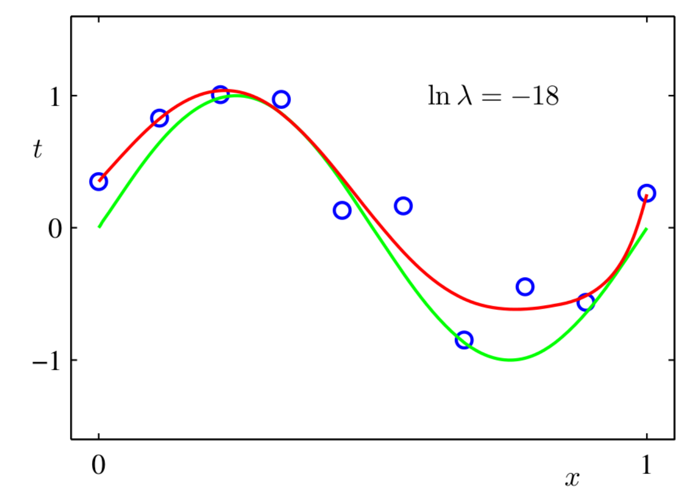
$$\mathbf{w}^* = \left(\mathbf{X}^\top\mathbf{X} + \alpha\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}.$$

- This is called *ridge regression*; by the way, adding $\alpha\mathbf{I}$ to a matrix of incomplete rank makes it invertible; this was the original motivation for ridge regression and for *regularization*.

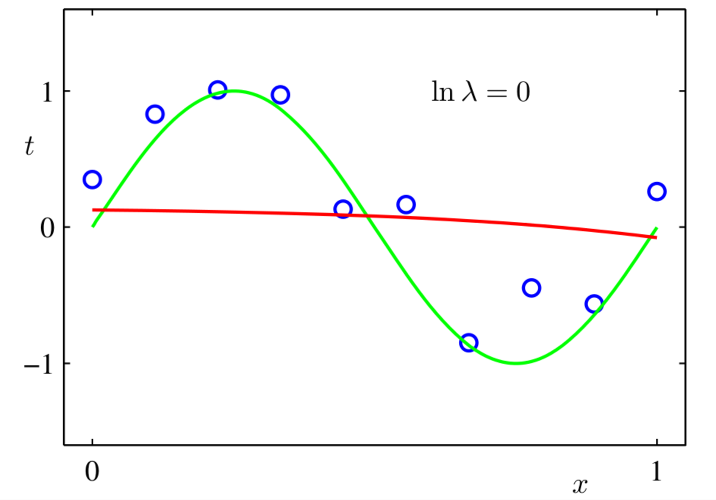
RIDGE REGRESSION: $\ln \alpha = -\infty$



RIDGE REGRESSION: $\ln \alpha = -18$



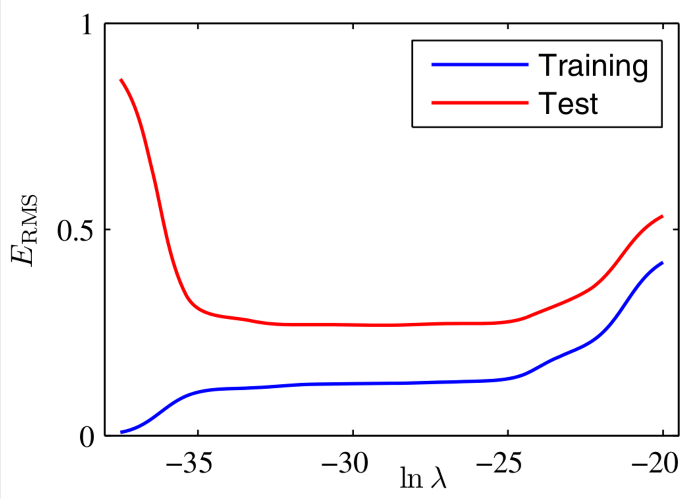
RIDGE REGRESSION: $\ln \alpha = 0$



RIDGE REGRESSION: 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

RIDGE REGRESSION: RMS



- Why exactly $\frac{\alpha}{2}\|\mathbf{w}\|^2$?
- We will see an answer shortly, but in general it's not necessary.
- *Lasso regression* regularizes with L_1 norm rather than L_2:

$$\mathrm{RSS}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i, \mathbf{w}) - y_i)^2 + \alpha \sum_{j=0}^{M} |w_j|.$$

- There are other kinds of regularizers too; more on that later.

THANK YOU!

Thank you for your attention!