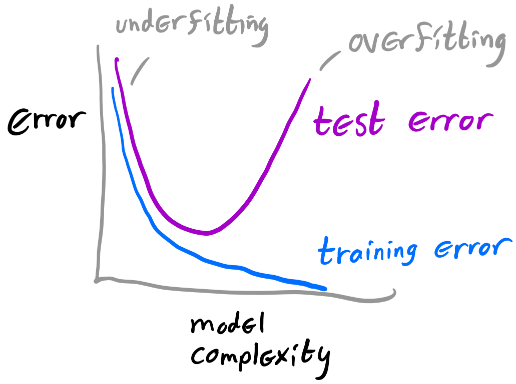
Day 11 - Ridge Regression

Agenda:

- Review - Bias Variance Tradeoff
- Ridge Regression
- Analytical Formula for Solution to Ridge Regression
- Background - Singular Value Decompositions
- Ridge Regression and Bias Variance Tradeoff

Bias-Variance Tradeoff

Standard Statistical ML story:

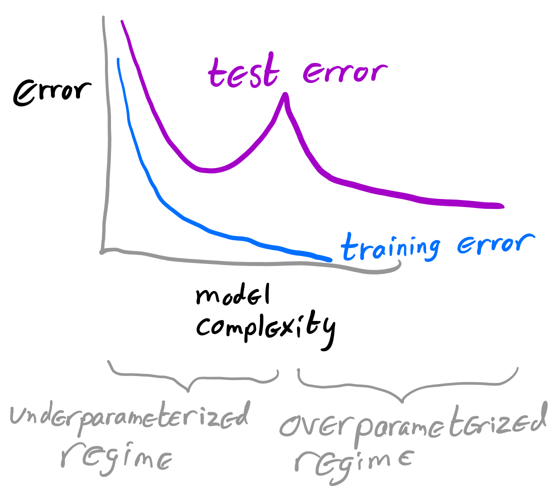


higher complexity models have lower bias but higher variance

If complexity is too high, it overfits data, variance term dominates test error

after a certain threshold, "larger models are worse"

Modern Story based on Neural Nets:



Test error can decrease as model complexity continues increasing.

And it can be lower than in underparameterized regime

Phenomenon: double descent

"larger models are better"

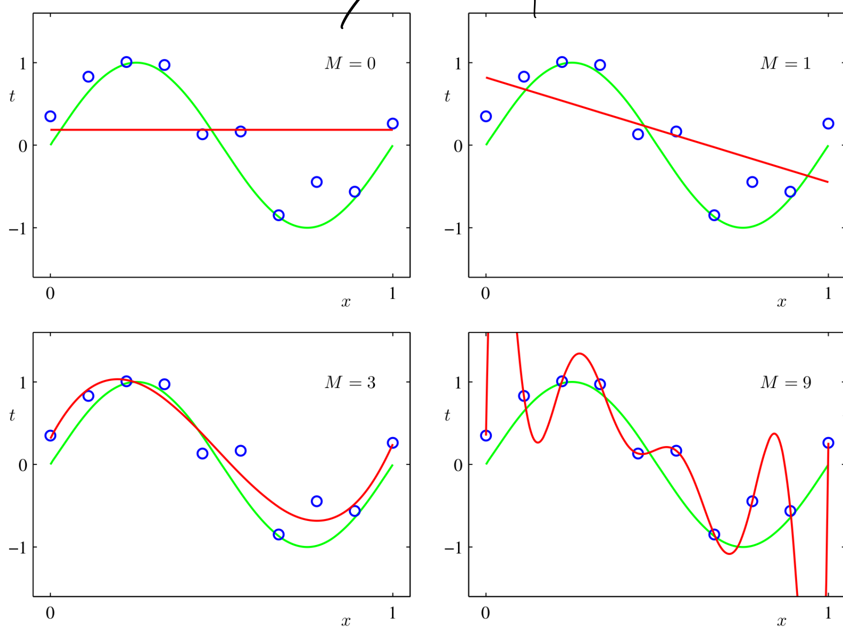
Ridge Regression

So far, we have used MLE to estimate model parameters from data

Concern: overfitting

Example: Fitting data w/ a degree M polynomial

underfitting



o data

— noiseless function

— best fit degree M polynomial

Overfitting

Figure 1.4 Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

One way to reduce overfitting,

use a hypothesis class with lower complexity (fewer unknown parameters)

Another way,

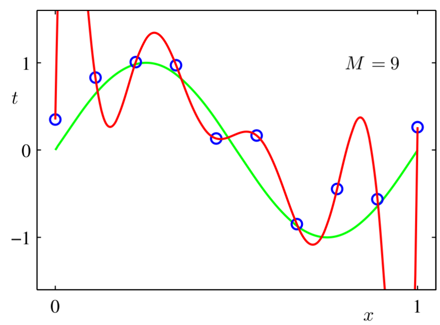
add regularization

A possible indication of overparameterization is having very large learned parameters.

This often happens when features are highly correlated

Table 1.2 Table of the coefficients $\mathbf{w}^\star$ for $M = 9$ polynomials with various values for the regularization parameter λ. Note that $\ln\lambda = -\infty$ corresponds to a model with no regularization, i.e., to the graph at the bottom right in Figure 1.4. We see that, as the value of λ increases, the typical magnitude of the coefficients gets smaller.

	$\ln\lambda = -\infty$
$w_0^\star$	0.35
$w_1^\star$	232.37
$w_2^\star$	-5321.83
$w_3^\star$	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



Idea: penalize predictors that have large values of unknown parameters

New formulation for least squares:

Given data $\{(X_i, y_i)\}_{i=1\cdots n}$ w/ $X_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

where $y = X\theta + \varepsilon$ w/ $\varepsilon \in \mathbb{R}^n$ has $\mathcal{N}(0, \sigma^2)$ entries

Estimate θ by solving — ridge regression problem

$$\min_\theta \; \|y - X\theta\|^2 + \underbrace{\lambda \|\theta\|^2}_{\ell_2 \text{ penalization} / \ell_2 \text{ regularization} / \text{weight decay}}$$

Solution is given by

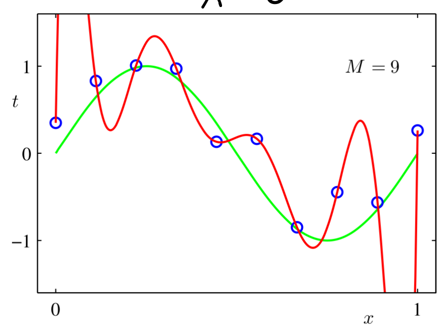
$$\hat{\theta}_{ridge} = (X^t X + \lambda I_{d\times d})^{-1} X^t y$$

w/ $I_{d\times d}$ = $d\times d$ Identity matrix $= \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$

What do solutions look like?

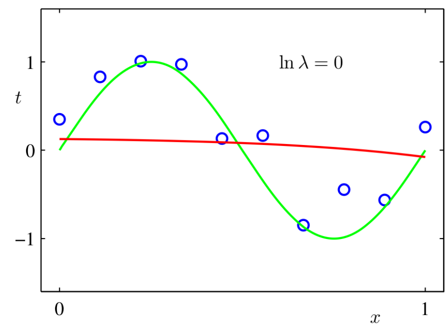
$\lambda = 0$

λ is too low



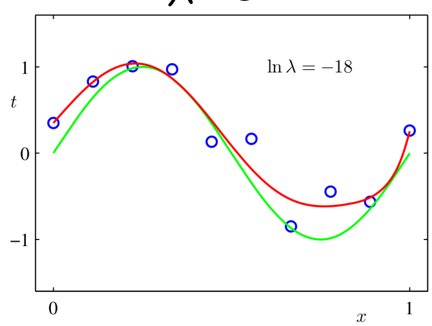
$\lambda = 1$

λ is too high



$\lambda = e^{-18}$

λ is about right



$$\min_{\theta} \|y - X\theta\|^2 + \lambda \|\theta\|^2$$

Solution to ridge regression problem

Let $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$

The unique solution to

$$\min_{\theta \in \mathbb{R}^d} \|y - X\theta\|^2 + \lambda \|\theta\|^2$$

is $$\hat{\theta}_{ridge} = (X^t X + \lambda I_{d \times d})^{-1} X^t y$$

Proof: Let $f(\theta) = \|X\theta - y\|^2 + \lambda \|\theta\|^2$

$$\nabla f(\theta) = 2X^t(X\theta - y) + 2\lambda\theta$$

Set $\nabla f(\theta) = 0$

$$\Rightarrow 2X^t(X\theta - y) + 2\lambda\theta = 0$$

$$\Rightarrow X^tX\theta - X^ty + \lambda\theta = 0$$

$$\Rightarrow (X^tX + \lambda I_{d\times d})\theta = X^ty$$

$$\Rightarrow \theta = (\underbrace{X^tX + \lambda I_{d\times d}}_{})^{-1}X^ty.$$

Note: this matrix is always invertible if $\lambda > 0$

why?

Background in Linear Algebra - Singular Value Decomposition

SVD of a square matrix:

Suppose $A \in \mathbb{R}^{n \times n}$. An SVD of A is given by

$$A = U \Sigma V^t$$

where U is $d \times d$ matrix w/ orthonormal columns
V is $d \times d$ matrix w/ orthonormal columns
Σ is diagonal w/ nonnegative entries $\sigma_1, \sigma_2, \cdots \sigma_d$
where $\sigma_i \geq \sigma_{i+1} \geq 0$

The columns of U are the left singular vectors of A
— — — V — — right singular vectors — —
The diagonal entries of Σ are the singular values of A

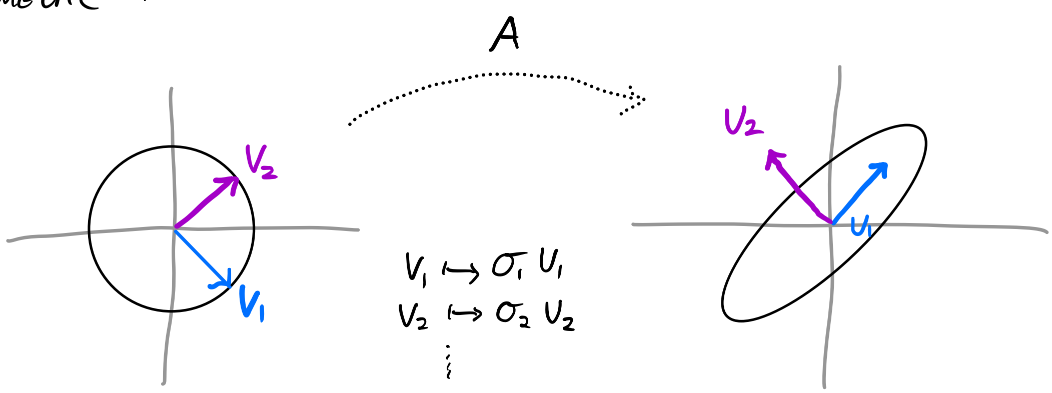
$$A = \begin{pmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_n \end{pmatrix} \begin{pmatrix} - v_1^t - \\ - v_2^t - \\ \vdots \\ - v_n^t - \end{pmatrix}$$

Note: A set $\{u_1 \cdots u_n\}$ is orthonormal if
- $\|u_i\|^2 = 1$ for all i
- $u_i \cdot u_j = 0$ if $i \neq j$

The ij entry of $U^t U = u_i^t u_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$

So U has orthonormal columns if $\underbrace{U^t U}_{n \times n} = I_{n \times n}$

Geometric picture of SVD:



Linear operators map the unit circle to an ellipsoid. The left singular vectors provide the principal axes of the ellipsoid.

Alternatively, any A is a diagonal matrix if the domain & range spaces use the right bases.

Given a basis $\{v_1 \cdots v_d\}$ of $\mathbb{R}^d$, if $V = \begin{pmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_d \\ | & | & & | \end{pmatrix}$ then the coefficients of x in the basis $\{v_1 \cdots v_d\}$ is given by

$$V^t x.$$

So, SVD can be interpreted as

$$A = \underbrace{U}_{\text{convert from basis given by } U}\ \underbrace{\Sigma}_{\text{diagonal operator}}\ \underbrace{V^t}_{\text{put input vector in basis given by } V}$$

Example

You can use SVD to manipulate matrices easily

Show that if $A \in \mathbb{R}^{n \times n}$ is invertible,

and $A = U \Sigma V^t$ is SVD of A, then

$$A^{-1} = V \Sigma^{-1} U^t$$

Proof: If Σ is invertible, $\sigma_d > 0$.

otherwise V_d would be in null space of A, and hence A isn't invertible.

We will show $A(V \Sigma^{-1} U^t) = I_n$.

$$A V \Sigma^{-1} U^t = U \Sigma \underbrace{V^t V}_{I_n} \Sigma^{-1} U^t$$

$$= U \underbrace{\Sigma \Sigma^{-1}}_{I_n} V^t$$

$$= U U^t$$

$$= I_n \quad \blacksquare$$

SVD of a tall rectangular matrix

Let $A \in \mathbb{R}^{n \times d}$ w/ $n \geq d$.

An SVD of A is given by

$$A = U \Sigma V^t$$

w/ U – $n \times d$ matrix w/ orthonormal columns

Σ – $d \times d$ diagonal nonnegative matrix w/ decreasing values along diagonal

V – $d \times d$ matrix w/ orthonormal columns

$$A = \begin{pmatrix} | & & | \\ a_1 & \cdots & a_d \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_2 \\ | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & \ddots & 0 \\ 0 & & \sigma_d \end{pmatrix} \begin{pmatrix} — & v_1^t & — \\ — & v_2^t & — \\ & \vdots & \\ — & v_d^t & — \end{pmatrix}$$

Note $U^t U = I_d$ but $UU^t \neq I_n$ (if $d > n$)

$V^t V = I_d$ & $VV^t = I_d$

Ridge Regression and the Bias Variance Tradeoff

Suppose data $\{(x_i, y_i)\}_{i=1\cdots n}$ follows the distribution

$$y_i = x_i^t \theta^* + \varepsilon_i \quad w/ \quad \varepsilon_i \sim N(0, \sigma^2)$$

That is,

$$y = X\theta^* + \varepsilon$$

Let $X = U\Sigma V^t$ be the SVD of X, where $\Sigma = diag(\sigma_1 \cdots \sigma_d)$

The ridge regression estimate of θ^* is

$$\hat{\theta}_{ridge} = \underbrace{V\, diag\left(\frac{\sigma_1^2}{\sigma_1^2+\lambda}, \cdots, \frac{\sigma_d^2}{\sigma_d^2+\lambda}\right) V^t \theta^*}_{\text{Signal} \quad \hat{\theta}^{signal}_{ridge}} + \underbrace{V\, diag\left(\frac{\sigma_1}{\sigma_1^2+\lambda}, \cdots, \frac{\sigma_d}{\sigma_d^2+\lambda}\right) U^t \varepsilon}_{\text{noise} \quad \hat{\theta}^{noise}_{ridge}}.$$

Let's analyze bias and variance of $\hat{\theta}_{ridge}$.

Note: – $\mathbb{E}\, \hat{\theta}^{noise}_{ridge} = 0$. So first term controls bias

– first term doesn't depend on ε. So second term controls variance

Analyze $\hat{\theta}^{signal}_{ridge}$ – if $\lambda = 0$ $\quad \hat{\theta}^{signal}_{ridge} = VV^t\theta^* = \theta^*$

Unbiased

if $\lambda = \infty$ $\quad \hat{\theta}^{signal}_{ridge} = 0$ $\quad$ biased

Bias increases with λ.

Analyze $\hat{\theta}^{noise}_{ridge}$ – if $\lambda = \infty$ $\hat{\theta}^{noise}_{ridge} = 0$ low variance

if $\lambda = 0$ $\hat{\theta}^{noise}_{ridge} = V \operatorname{diag}\left(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_d}\right) V^t \varepsilon$

high variance

$$\mathbb{E}_{\varepsilon} \|\hat{\theta}^{noise}_{ridge}\|^2 = \sum_{j=1}^{d} \left(\frac{\sigma_i}{\sigma_j^2 + \lambda}\right)^2 \sigma^2$$

Variance decreases with λ.

Observe: λ trades off between bias & variance

Justification of ridge regression estimate $\hat{\theta}_{ridge}$:

Let $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$.

By formula above

$$\hat{\theta}_{ridge} = (X^t X + \lambda I_d)^{-1} X^t y = (X^t X + \lambda I_d)^{-1} X^t (X\theta^* + \varepsilon)$$

Let $X = U \Sigma V^t$ be the SVD of X, where

U – $n \times d$ matrix with orthonormal columns

V – $d \times d$ matrix with orthonormal columns

Σ – $d \times d$ diagonal matrix $= \operatorname{diag}(\sigma_1, \dots, \sigma_d)$ w/ $\sigma_i \geq \sigma_{i+1} \geq 0$

Note $X^t X = V \Sigma^t \underbrace{U^t U}_{U^t U = I_d} \Sigma V^t = V \underbrace{\Sigma^t I_d \Sigma}_{\Sigma^2} V^t = V \Sigma^2 V^t$

So

$$\hat{\theta}_{ridge} = (V\Sigma^2V^t + \lambda I)^{-1}\left[X^tX\,\theta^* + X^t\varepsilon\right]$$

$$= (V\Sigma^2V^t + \lambda I)^{-1}\left[V\Sigma^2V^t\theta^* + V\Sigma^tU^t\varepsilon\right]$$

$$= \left(V(\Sigma^2+\lambda I)V^t\right)^{-1}\left[V\Sigma^2V^t\theta^* + V\Sigma^tU^t\varepsilon\right]$$

$$= V(\Sigma^2+\lambda I)^{-1}V^t\left[V\Sigma^2V^t\theta^* + V\Sigma^tU^t\varepsilon\right]$$

$$= V(\Sigma^2+\lambda I)^{-1}\left[\Sigma^2V^t\theta^* + \Sigma U^t\varepsilon\right]$$

$$= V(\Sigma^2+\lambda I)^{-1}\Sigma^2 V^t\theta^*$$
$$+ V(\Sigma^2+\lambda I)^{-1}\Sigma U^t\varepsilon$$

Note $(\Sigma^2+\lambda I)^{-1} = \mathrm{diag}\left(\frac{1}{\sigma_1^2+\lambda}, \dots, \frac{1}{\sigma_d^2+\lambda}\right)$

So

$$\boxed{\begin{aligned}\hat{\theta}_{ridge} &= V\,\mathrm{diag}\left(\frac{\sigma_1^2}{\sigma_1^2+\lambda}, \dots, \frac{\sigma_d^2}{\sigma_d^2+\lambda}\right)V^t\theta^* \\ &\quad + V\,\mathrm{diag}\left(\frac{\sigma_1}{\sigma_1^2+\lambda}, \dots, \frac{\sigma_d}{\sigma_d^2+\lambda}\right)U^t\varepsilon\end{aligned}}$$