Fundamentals of Machine Learning (Part I)

Mohammad Emtiyaz Khan
AIP (RIKEN), Tokyo

http://emtiyaz.github.io

emtiyaz.khan@riken.jp

April 12, 2018

Goals

Understand (some) fundamentals of Machine learning[1].

of part I

Part I : Understand the basic set-up to analyze data under a machine-learning framework.

1. Before Machine Learning.
2. ML Problem: Regression.
3. Model: Linear Regression.
4. Cost Function: MSE.
5. Algorithm 1: Gradient Descent.
6. Algorithm 2: Least Squares.

of part II

Part II : Understand what can go wrong when learning from data and how to correct it.

6. Challenge: Overfitting.
7. Solutions: Regularization.
8. Bias-Variance Decomposition.
9. Recent Advances.

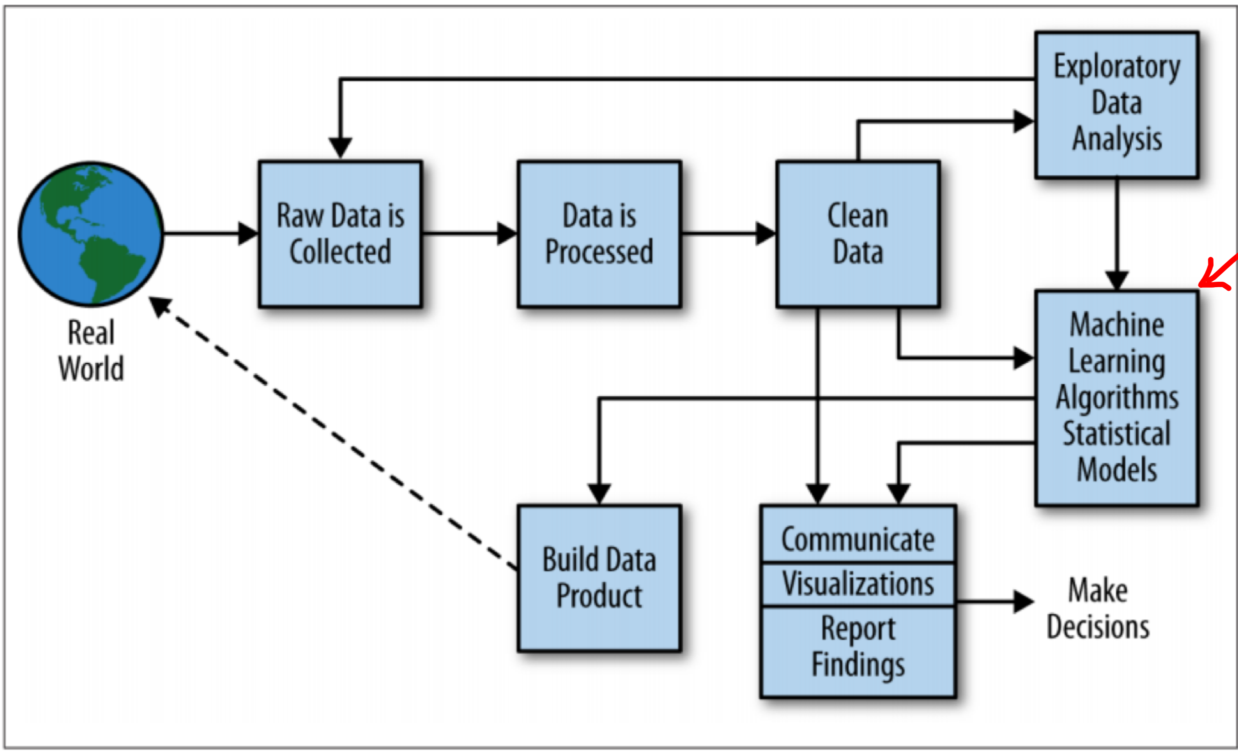
[1]Some figures are taken from Hastie, Tibshirani, and Friedman's book on statistical learning and also from Chris Bishop's Machine learning book

1 Before Machine Learning

Acquiring Data

Data is the most important component of modern Machine Learning. There are many important steps that can have a huge impact on the performance of a machine-learning system. To name a few: data collection, cleaning, validation, pre-processing, and storage.

In this course, You will learn this.



Picture taken from “Doing data science”.

Defining an ML problem

e.g.

Who is going to win in the Next election?

Once we have some data, the next step is to re-define the real-world problem in the context of data, and then to convert it to a machine-learning problem.

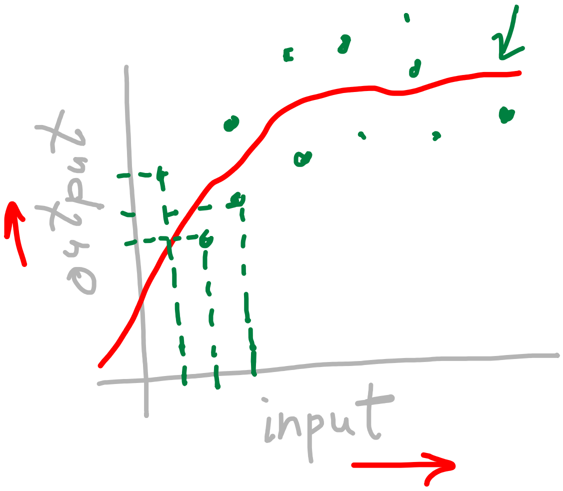
ML problems can be categorized into 3 main types: supervised, unsupervised, and reinforcement learning. In practice, a successful end-to-end system might require a combination of these problems.

In this course, we will focus on supervised learning.

Goal: Define regression and its two goals.

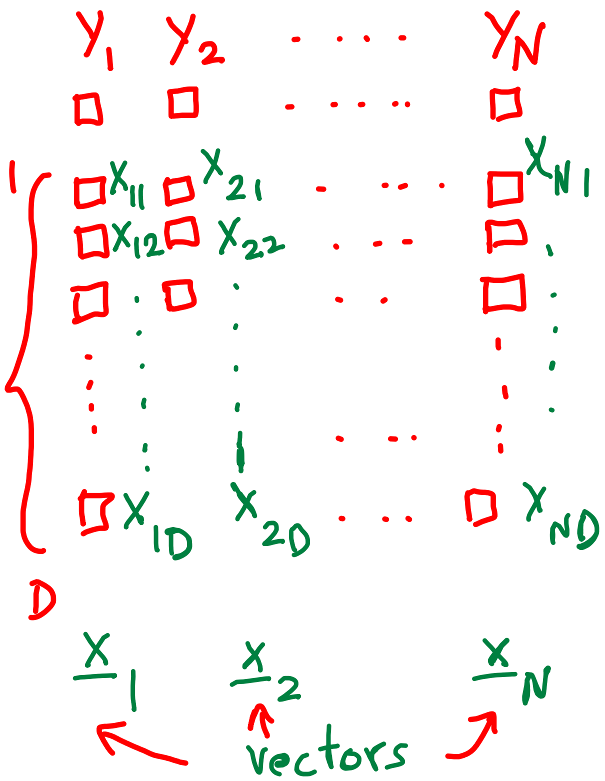
2 ML Problem: Regression

What is regression?



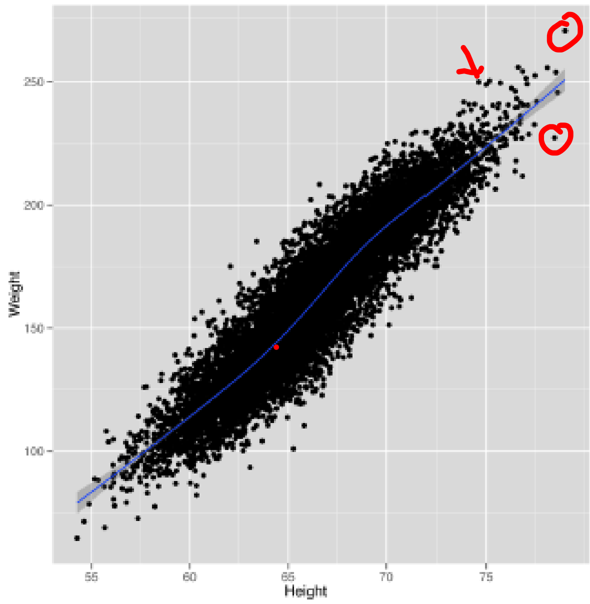
Regression is to relate input variables to the output variable, to either predict outputs for new inputs and/or to understand the effect of the input on the output.

Dataset for regression



In regression, data consists of pairs $(y_n, \mathbf{x}_n)$, where y_n is the n'th output and $\mathbf{x}_n$ is a vector of D inputs. Number of pairs N is the data-size and D is the dimensionality.

Examples of regression



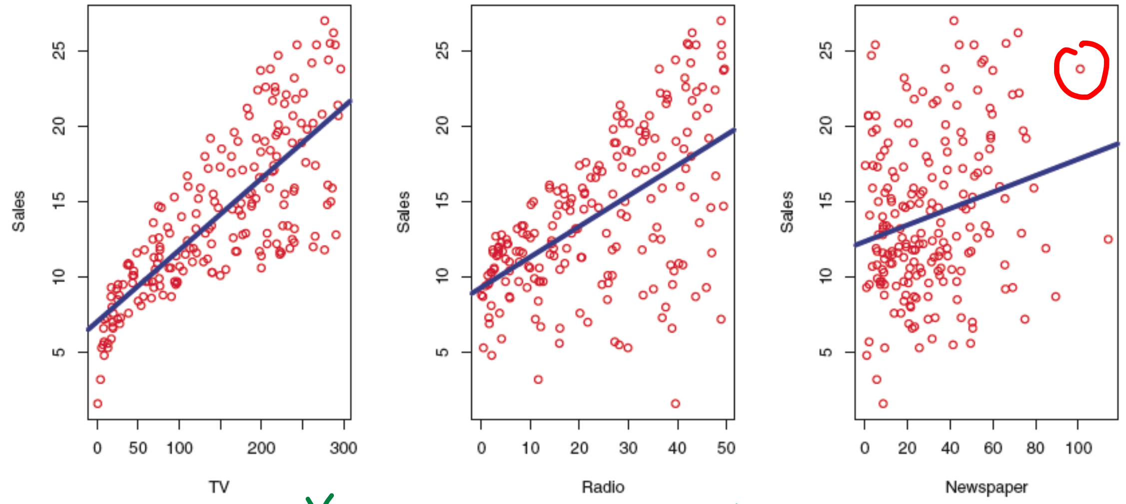
(x_n, y_n)

$N = 10{,}000$

$D = 1$ input

(a) Height is correlated with weight. Taken from "Machine Learning for Hackers"

Y_n



X_{n1} X_{n2} X_{n3}

$D = 3$

(b) How does advertisement in TV, radio, and newspaper affect sales? Taken from the book "An Introduction to statistical learning"

Two goals of regression

In prediction, we wish to predict the output for a new input vector, e.g. what is the weight of a person who is 170 cm tall?

In interpretation, we wish to understand the effect of inputs on output, e.g. are taller people heavier too?

The regression function

For both the goals, we need to find a function that approximates the output "well enough" given inputs.

$$y_n \approx f(\mathbf{x}_n), \text{ for all } n$$

Additional Notes

Prediction vs Interpretation

Some questions to think about: are these prediction tasks or interpretation task?

1. What is the life-expectancy of a person who has been smoking for 10 years? Prediction
2. P Does smoking cause cancer?
3. I When the number of packs a smoker smokes per day doubles, their predicted life span gets cut in half?
4. P A massive scale earthquake will occur in California within next 30 years.
5. P More than 300 bird species in north America could reduce their habitat by half or more by 2080.

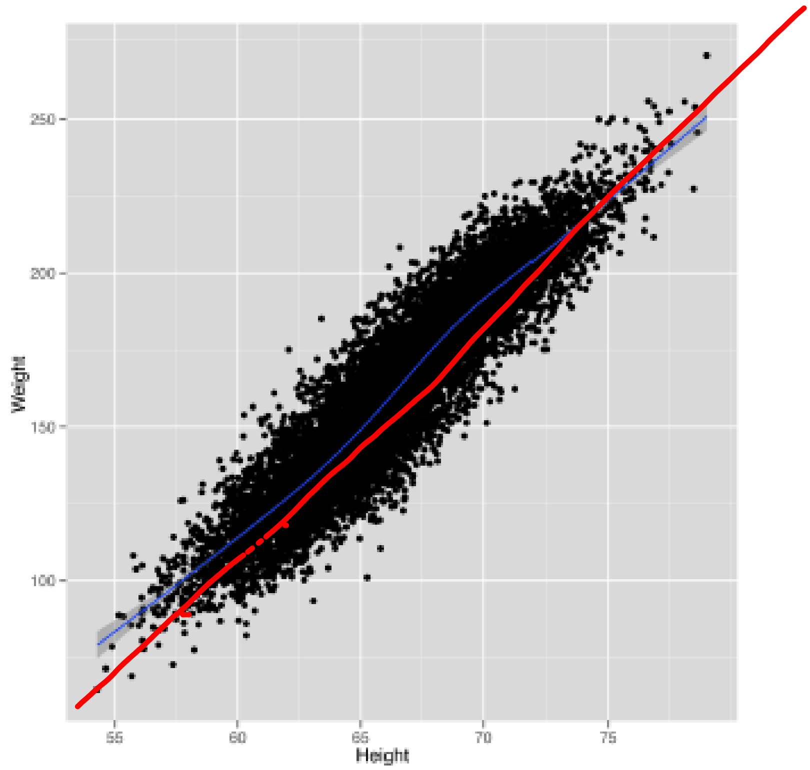
Goals: Define regression and its two goals.

Goal: Define and describe linear regression.

3 Model: Linear Regression

What is it?

Linear regression is a model that assumes a linear relationship between inputs and the output.



Why learn about *linear* regression?

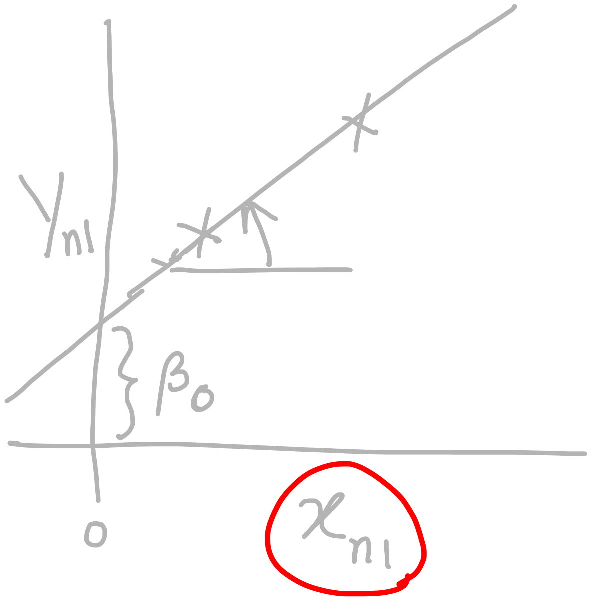
Plenty of reasons: simple, easy to understand, most widely used, easily generalized to non-linear models. Most importantly, you can learn almost all fundamental concepts of ML with regression alone.

Simple linear regression

With only one input dimension, it is simple linear regression.

$$y_n \approx f(\mathbf{x}_n) := \beta_0 + \beta_1 x_{n1}$$

Here, β_0 and β_1 are parameters of the model.



Multiple linear regression

With multiple input dimension, it is multiple linear regression.

$$\begin{aligned} y_n &\approx f(\mathbf{x}_n) \\ &:= \beta_0 + \beta_1 x_{n1} + \ldots + \beta_D x_{nD} \\ &= \widetilde{\mathbf{x}}_n^T \boldsymbol{\beta} \end{aligned} \tag{1}$$

$$\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \vdots\\ \beta_D\end{bmatrix} \begin{bmatrix}1\\ x_{n1}\\ x_{n2}\\ \vdots\\ x_{nD}\end{bmatrix}$$

$$\underline{\beta}^T \quad \underline{\tilde{x}}_n$$

"inner product of two vectors"

Learning/estimation/fitting

Given data, we would like to find $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_D]$. This is called learning or estimating the parameters or fitting the model.

Example

True weight	$\approx$	Predicted weight	=	$\beta_0 + \beta_1 \times$ height
100 kg	$\approx$	100	=	$0 + \frac{1}{2} \times 200$ cm
70 "	$\approx$	85	=	$0 + \frac{1}{2} \times 170$ cm
90 "	$\approx$	105	=	$0 + \frac{1}{2} \times 210$
60 "	$\approx$	80	=	$0 + \frac{1}{2} \times 160$

with
$\beta_0 = 0$
$\beta_1 = 1/2$

Prediction → $95 = 0 + \frac{1}{2} \times 190$ cm ← New person

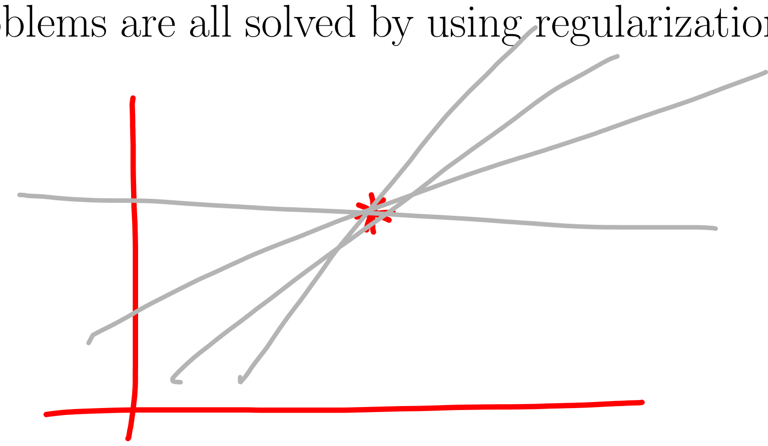
Additional Notes

$p > n$ Problem

Consider the following simple situation: You have $N = 1$ and you want to fit $y_1 \approx \beta_0 + \beta_1 x_{11}$, i.e. you want to find β_0 and β_1 given one pair (y_1, x_{11}). Is it possible to find such a line?

This problem is related to something called $p > n$ problem. In our notation, this will be called $D > N$ problem, i.e. the number of parameters exceeds number of data examples.

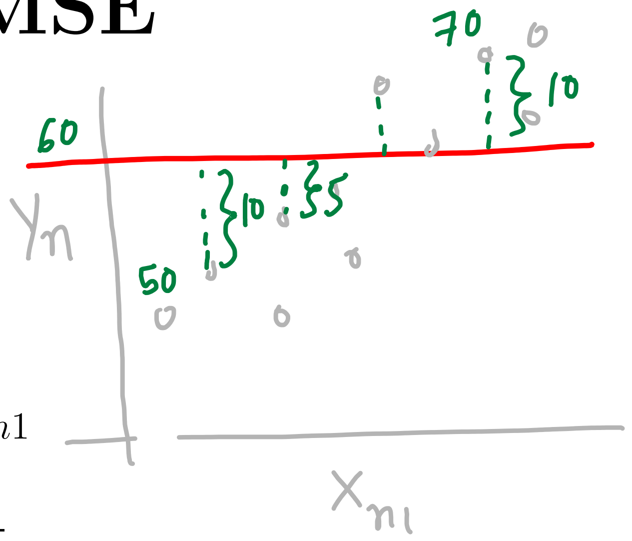
Similar issues will arise when we use gradient descent or least-squares to fit a linear model. These problems are all solved by using regularization, which we will learn later.



Goal: Define and describe linear regression.

Goal: What is MSE and what could go wrong with it.

4 Cost Function: MSE



Motivation

Consider the following models.

1-parameter model: $y_n \approx \beta_0$ ✓

2-parameter model: $y_n \approx \beta_0 + \beta_1 x_{n1}$

How can we estimate (or guess) learn values of $\boldsymbol{\beta}$ given the data $\mathcal{D}$?

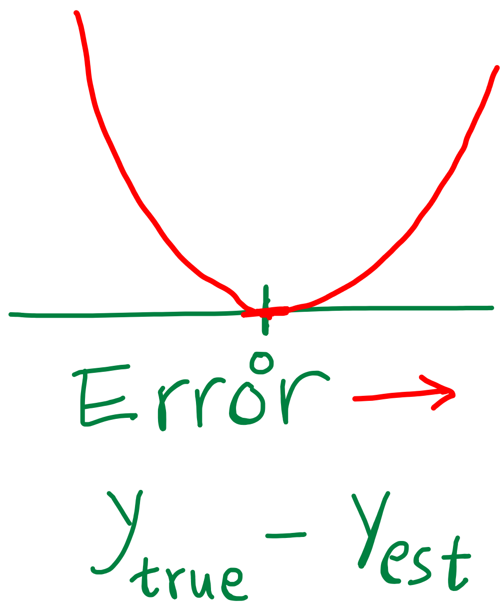
$\mathcal{D} = \{y_1, y_2, \dots y_N, \underline{x}_1, \underline{x}_2, \dots \underline{x}_N\}$

What is a cost function?

Cost functions (or utilities or energy) are used to learn parameters that explain the data well. They define how costly our mistakes are.

Two desirable properties of cost functions

When y is real-valued, it is desirable that the cost is symmetric around 0, since both +ve and -ve errors should be penalized equally.



~~Also, our cost function should penalize "large" mistakes and "very-large" mistakes almost equally.~~

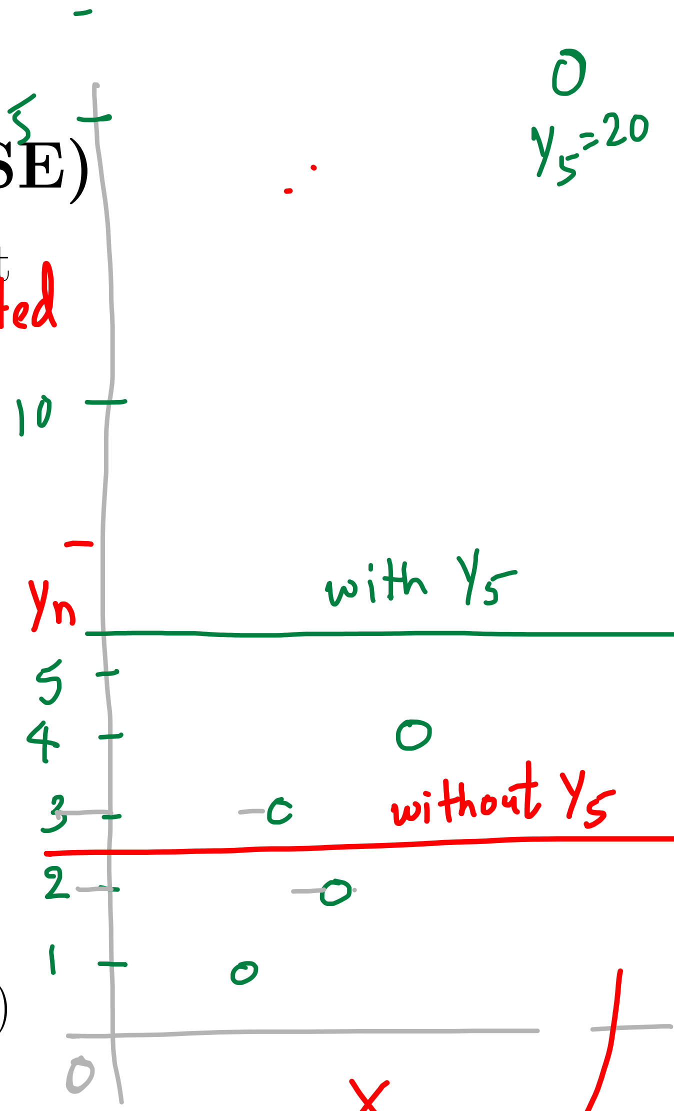
Mean Square Error (MSE)

MSE is one of the most popular cost function.

True Predicted

$$\color{blue}{MSE}(\boldsymbol{\beta}) := \frac{1}{2N}\sum_{n=1}^{N}[y_n - f(\mathbf{x}_n)]^2$$

Does it have ~~both~~ the properties? Yes



An exercise for MSE

Compute MSE for 1-param model:

$$\mathcal{L}(\beta_0) := \frac{1}{2N}\sum_{n=1}^{N}[y_n - \beta_0]^2 \qquad (2)$$

Each row contains a y_n and column is β_0. First, compute MSE for for $y_n = \{1, 2, 3, 4\}$ and draw MSE as a function of β_0 (by adding the first four rows). Then add $y_n = 20$ to it, and redraw MSE. What do you observe and why?

β_0 →

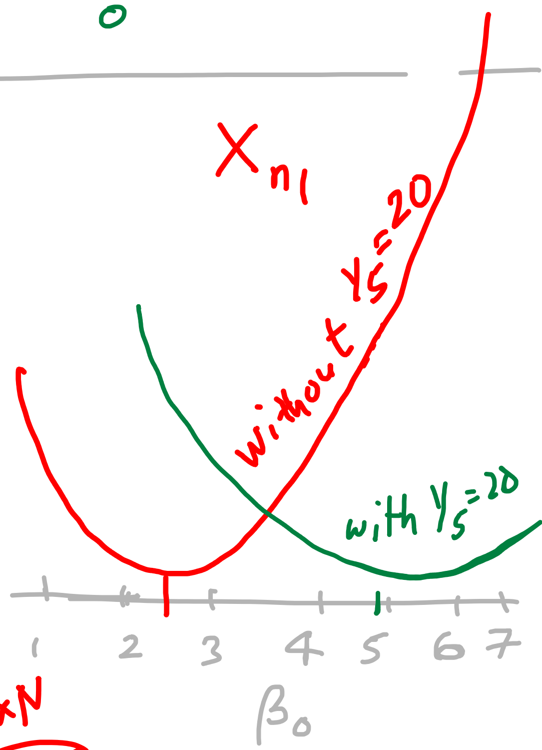
M →

N

$y_n \backslash \beta_0$	1	2	3	4	5	6	7
y_1 1	0	1					
y_2 2	1	0					
y_3 3	2	1					
y_4 4	3	2					
MSE	14	6	6	14	30	54	86
y_5 20	$+19^2$	18^2	..	...	..	.	13^2
MSE							

M×N

4×7



minimum at $\beta_0 = 2.5$

fill this.

where is the minimum?

Additional Notes

A question for cost functions

Is there an automatic way to define loss functions?

Nasty cost functions: Visualization

See Andrej Karpathy Tumblr post for many cost functions gone "wrong" for neural network. http://lossfunctions.tumblr.com/.

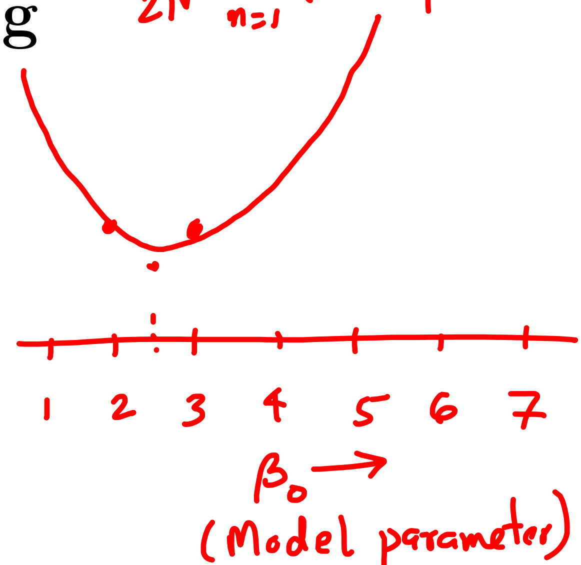
Goal: What is MSE and what could go
wrong with it.

Goal: Derive GD and SGD for linear regression

5 Algorithm 1: Gradient Descent

$\mathcal{L}(\beta_0) = \frac{1}{2N}\sum_{n=1}^{N}(y_n - \beta_0)^2$

Learning/estimation/fitting



Given a cost function $\mathcal{L}(\boldsymbol{\beta})$, we wish to find $\boldsymbol{\beta}^*$ that minimizes the cost:

$$\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}), \qquad \text{subject to } \boldsymbol{\beta} \in \mathbb{R}^{D+1}$$

This is learning posed as an optimization problem. We will use an algorithm to solve the problem.

Computation Complexity: Number of computation

$D=0$

$\approx O(M^D N D)$ (big O notation)

Grid search

Grid search is one of the simplest algorithms where we compute cost over a grid (of say M points) to find the minimum. This is extremely simple and works for any kind of loss when we have very few parameters and the loss is easy to compute.

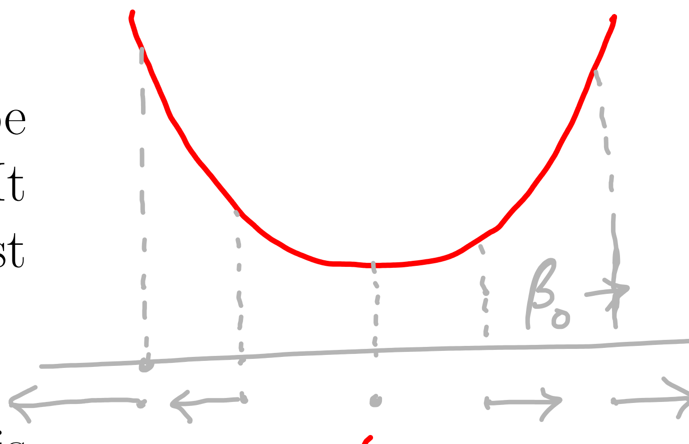
for $\beta_1 = 1, 2, \dots 100$
for $\beta_0 = 1, 2, \dots 100$
for $n = 1, 2, \dots, 4$
compute cost $O(D)$
$y_n - \tilde{x}_n^T \beta$
end
end

For a large number of parameters, however, grid search has too many "for-loops", resulting in exponential computational complexity. Choosing a good range of values is another problem.

Are there any other issues? You might miss the minimum.

Follow the gradient

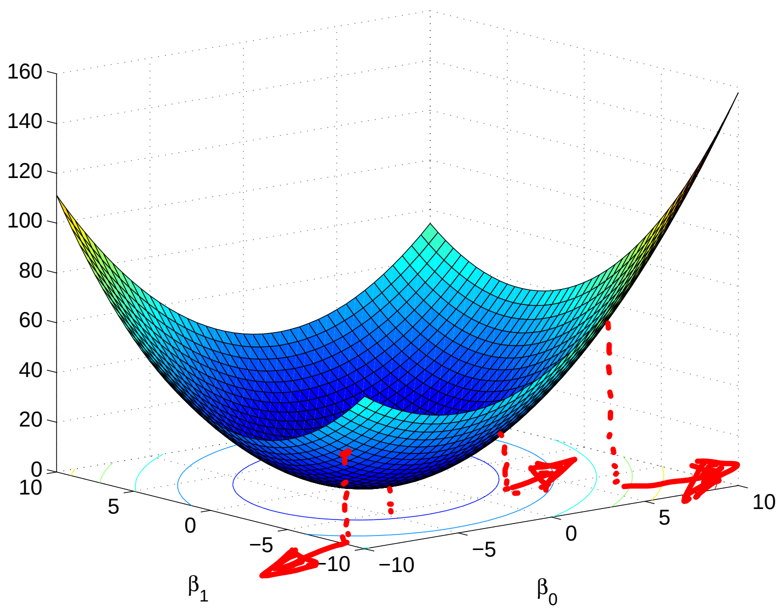
A gradient (at a point) is the slope of the tangent (at that point). It points to the direction of largest increase of the function.



stationary pt

For 2-parameter model, MSE is shown below.

(I used $\mathbf{y}^T = [2, -1, 1.5]$ and $\mathbf{x}^T = [-1, 1, -1]$).



Math

$$\mathcal{L}(\beta_0) = \frac{1}{2N}\sum_{n=1}^{N}(y_n - \beta_0)^2$$

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = -\frac{1}{N}\sum_{n=1}^{N}(y_n - \beta_0)$$

$$= -\underbrace{\frac{1}{N}\sum_{n=1}^{N} y_n}_{\bar{y}} + \beta_0$$

(Average of y_n)

$$= \beta_0 - \bar{y}$$

This is the Gradient for Lin Reg. →

$$\frac{\partial \mathcal{L}}{\partial \underline{\beta}} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial \beta_0} \\ \frac{\partial \mathcal{L}}{\partial \beta_1} \\ \frac{\partial \mathcal{L}}{\partial \beta_2} \end{bmatrix}$$

Batch "gradient descent"

To minimize the function, take a step in the (opposite) direction of the gradient

$$\boldsymbol{\beta}^{(k+1)} \leftarrow \boldsymbol{\beta}^{(k)} - \alpha \frac{\partial \mathcal{L}(\boldsymbol{\beta}^{(k)})}{\partial \boldsymbol{\beta}}$$

where $\alpha > 0$ is the step-size (or learning rate).

$$\beta_0 \leftarrow \beta_0 - \alpha \underbrace{\frac{\partial \mathcal{L}(\beta_0)}{\partial \beta_0}}$$

$$\beta_0^{(k+1)} \leftarrow \beta_0^{(k)} - \alpha \underbrace{(\beta_0^{(k)} - \bar{y})}_{\text{error}}$$

step-size

Q: What is the best α for 1-parameter model?

Gradient descent for 1-parameter model to minimize MSE:

A: $\alpha = 1$

$$\beta_0^{(k+1)} = (1-\alpha)\beta_0^{(k)} + \alpha\bar{y}$$

Minimum at $\beta_0 - \bar{y} = 0$

$\beta_0^* = \bar{y}$

Where $\bar{y} = \sum_n y_n / N$. When is this sequence guaranteed to converge?

Answer this on your own.

Gradients for MSE

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2N} \sum_{n=1}^{N} (y_n - \widetilde{\mathbf{x}}_n^T \boldsymbol{\beta})^2 \qquad (3)$$

then the gradient is given by,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = -\frac{1}{N} \sum_{n=1}^{N} \underbrace{(y_n - \widetilde{\mathbf{x}}_n^T \boldsymbol{\beta})}_{\text{error}} \widetilde{\mathbf{x}}_n \qquad (4)$$

Please derive this on your own.

What is the computational complexity of batch gradient descent?

Answer this on your own.

Stochastic gradient descent (SGD)

When N is large, choose a random pair $(\mathbf{x}_i, y_i)$ in the training set and approximate the gradient:

This is "Unbiased" gradient, meaning that it's correct on average.

Multiply with N

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} \approx -\frac{1}{N}\left[N\underbrace{(y_i - \widetilde{\mathbf{x}}_i^T \boldsymbol{\beta})}_{\text{error}}\widetilde{\mathbf{x}}_i\right] \quad (5)$$

Using the above "stochastic" gradient, take a step:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \alpha^{(k)}\underbrace{(y_i - \widetilde{\mathbf{x}}_i^T \boldsymbol{\beta}^{(k)})\widetilde{\mathbf{x}}_i}_{\text{gradient of } i\text{'th data point}}$$

What is the computational complexity? → Derive on your own.

For convergence, $\alpha^k \to 0$ "appropriately". One such condition called Robbins-Monroe condition suggests to take α^k such that:

"Don't drink and drive" Walk slowly.

$$\sum_{k=1}^{\infty} \alpha^{(k)} = \infty, \qquad \sum_{k=1}^{\infty} (\alpha^{(k)})^2 < \infty \quad (6)$$

One way to obtain such sequence is $\alpha^{(k)} = 1/(1+k)^r$ where $r \in (0.5, 1)$.

Regression → Model → Cost → Algorithm

Model: Lin Reg; Cost: MSE; Algorithm: Gradient descent

Whes is the problem difficult to optimize?

5.1 Algorithm 2: Least Squares

“Ill-conditioned problems”

In rare cases, we can compute the minimum of the cost function analytically. Linear regression using MSE is one such case. The solution is obtained using normal equations. This is called least squares.

$$\frac{\partial \mathcal{L}}{\partial \underline{\beta}} \approx \sum_n \underbrace{\left(y_n - \tilde{\underline{x}}_n^T \underline{\beta}\right)}_{\text{error}} \tilde{\underline{x}}_n$$

$e_n * \tilde{\underline{x}}_n$

scalar

$$\underbrace{\begin{bmatrix} \frac{\partial \mathcal{L}}{\partial \beta_0} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial \beta_D} \end{bmatrix}}_{D+1} \approx \square * \underbrace{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}_{D+1}$$

To derive the equation, we use the optimality conditions. See the lecture notes for Gradient Descent.

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}} = 0$$

Using this, derive the normal equation for 1-parameter model.

When x_n's are similar, gradients are not informative, which will make the problem harder.

This is related to ill-conditioning, whis is related to

$$\begin{array}{c} \xleftarrow{\quad N \quad} \\ D \updownarrow \begin{bmatrix} \underline{x}_1 & \underline{x}_2 & \underline{x}_2 & \cdots & \underline{x}_N \end{bmatrix} \end{array}$$

$D \times N$

Linearly-dependent columns

lead to “rank deficient” matrix.

e.g. when an x_n is approximately equal to another x_m..

This could happen easily, e.g. two people of same height-weight.

Normal equations

Recall the expression of the gradient for multiple linear regression:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = -\frac{1}{N}\widetilde{\mathbf{X}}^T \mathbf{e} = -\frac{1}{N}\widetilde{\mathbf{X}}^T (\mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta})$$

Set it to zero to get the normal equations for linear regression.

$$\widetilde{\mathbf{X}}^T \mathbf{e} = \widetilde{\mathbf{X}}^T (\mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) = 0$$

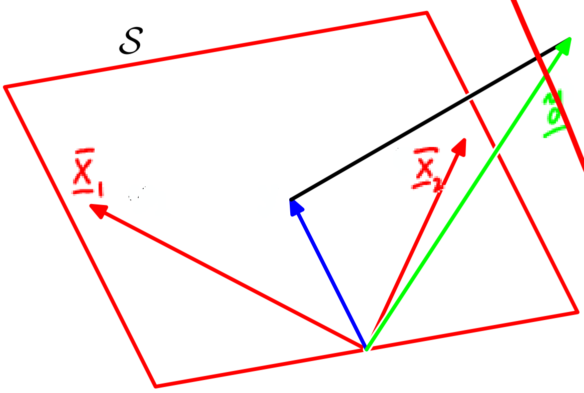
implying that the error is orthogonal to rows of $\widetilde{\mathbf{X}}^T$ and columns of $\widetilde{\mathbf{X}}$.

Geometric Interpretation

Denote the d'th column of $\widetilde{\mathbf{X}}$ by $\bar{\mathbf{x}}_d$.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \widetilde{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1D} \\ 1 & x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{ND} \end{bmatrix}$$

The normal equations suggest to choose a vector in the span of $\widetilde{\mathbf{X}}$. The following figure illustrates this (taken from Bishop's book).



Least-squares

When $\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$ is invertible, we have a closed-form expression for the minimum.

$$\boldsymbol{\beta}^* = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{y}$$

We can predict values for a new $\mathbf{x}_*$.

$$\hat{y}_* = \widetilde{\mathbf{x}}_*^T\boldsymbol{\beta}^* = \widetilde{\mathbf{x}}_*^T(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{y}$$

Invertibility and uniqueness

The Gram matrix $\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$ is invertible iff $\widetilde{\mathbf{X}}$ has full column rank.

Proof: Assume $N > D$. The fundamental theorem of linear algebra states that the dimensionality of null space is zero for full column rank. This implies that the Gram matrix is positive definite, which implies invertibility.

Important: It will be in the exam!

Rank deficiency and ill-conditioning

Unfortunately, $\widetilde{\mathbf{X}}$ could often be rank deficient in practice, e.g. when $D > N$, or when the columns $\bar{\mathbf{x}}_d$ are (nearly) collinear. In the later case, the matrix is ill-conditioned, leading to numerical issues.

Important

Summary of linear regression

We have studied three methods:

1. Grid search
2. (Stochastic) gradient descent
3. ~~Least squares~~

Additional Notes

Implementation

There are many ways to implement matrix inversion, but using QR decomposition is one of the most robust ways. Matlab's backslash operator implements this (and much more) in just one line.

```
beta = inv(X'*X) * (X'*y)
beta = pinv(X'*X) * (X'*y)
beta = (X'*X) \ (X'*y)
```

For robust implementation, see Sec. 7.5.2 of Kevin Murphy's book. e.g. here

To do

1. Revise linear algebra to understand why $\widetilde{\mathbf{X}}$ needs to have full rank. Read the Wikipedia page on rank of a matrix.
2. For details on the geometrical interpretation, see Bishop 3.1.2. However, better to read this after the lecture on "basis-function expansion". Also, note that notation in the book is different. This might make the reading difficult.
3. Understand matrix inversion robust implementation and play with it using the code for labs. Read Kevin Murphy's section 7.5.2 for details.

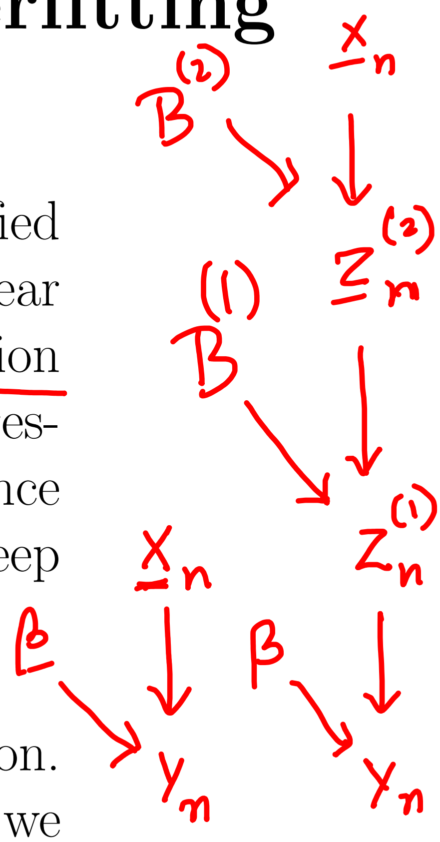
Read this!

4. Understand ill-conditioning. Reading about the "condition number" in Wikipedia will help. Also, understanding SVD is essential. Here is another link provided by Dana Kianfar (EPFL) http://www.cs.uleth.ca/~holzmann/notes/illconditioned.pdf.
5. Work out the computational complexity of least-squares (use the Wikipedia page on computational complexity).

a small change in the input leads to a large change in the output.

6 Challenge: Overfitting

Motivation



$z_n^{(1)} = B^{(1)} \phi(\underline{x}_n)$

$y_n = \beta^T \phi(\underline{x}_n)$

Linear model can be easily modified to obtain more powerful non-linear model. We can use basis function expansion to get a non-linear regression model, and then use a sequence of these models to construct a deep model.

Consider simple linear regression. Given one-dimensional input x_n, we can generate a polynomial basis.

$$\phi(x_n) = [1, x_n, x_n^2, x_n^3, \ldots, x_n^M]$$

Feature vector

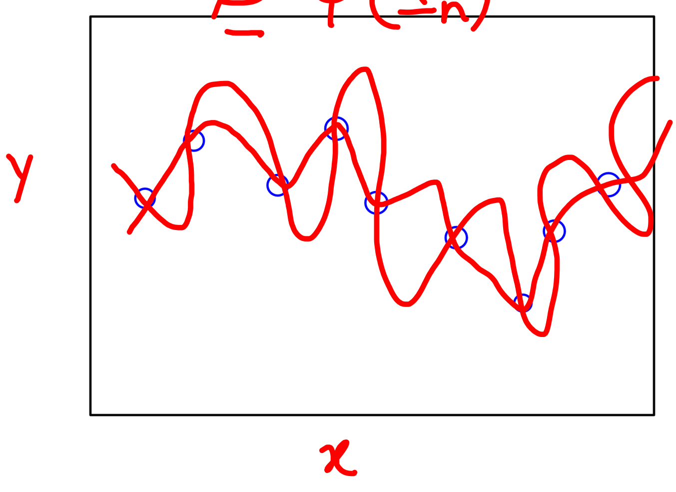
$\tilde{\phi}(x_n) = [1, x_n, x_n^2, x_n^3$

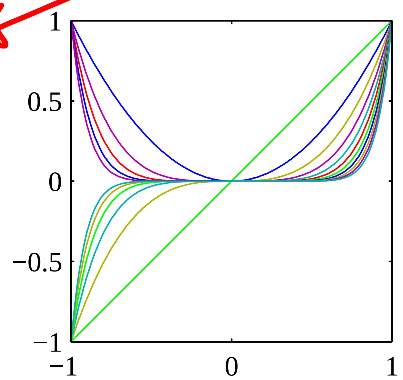
Then we fit a linear model using the original *and* the generated features:

$$y_n \approx \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \ldots + \beta_M x_n^M.$$

$= \underline{\beta}^T \tilde{\phi}(\underline{x}_n)$

Is this good?





$\mathcal{L}(\beta^*) = 0$

minimum

2 layer NN can approximate any function!

Overfitting and Underfitting

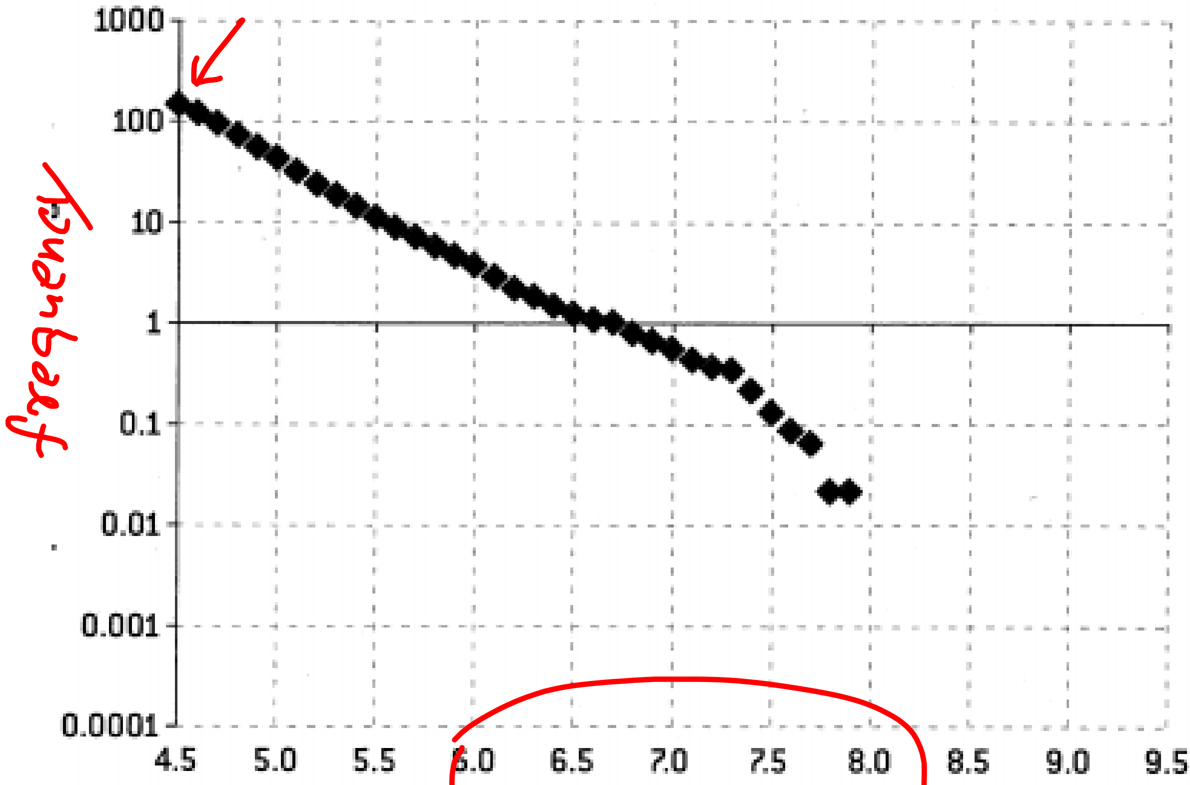
Overfitting is fitting the noise in addition to the signal. Underfitting is not fitting the signal well. In reality, it is very difficult to be able to tell the signal from the noise.

"With great power comes great responsibility"

– Spiderman's Uncle

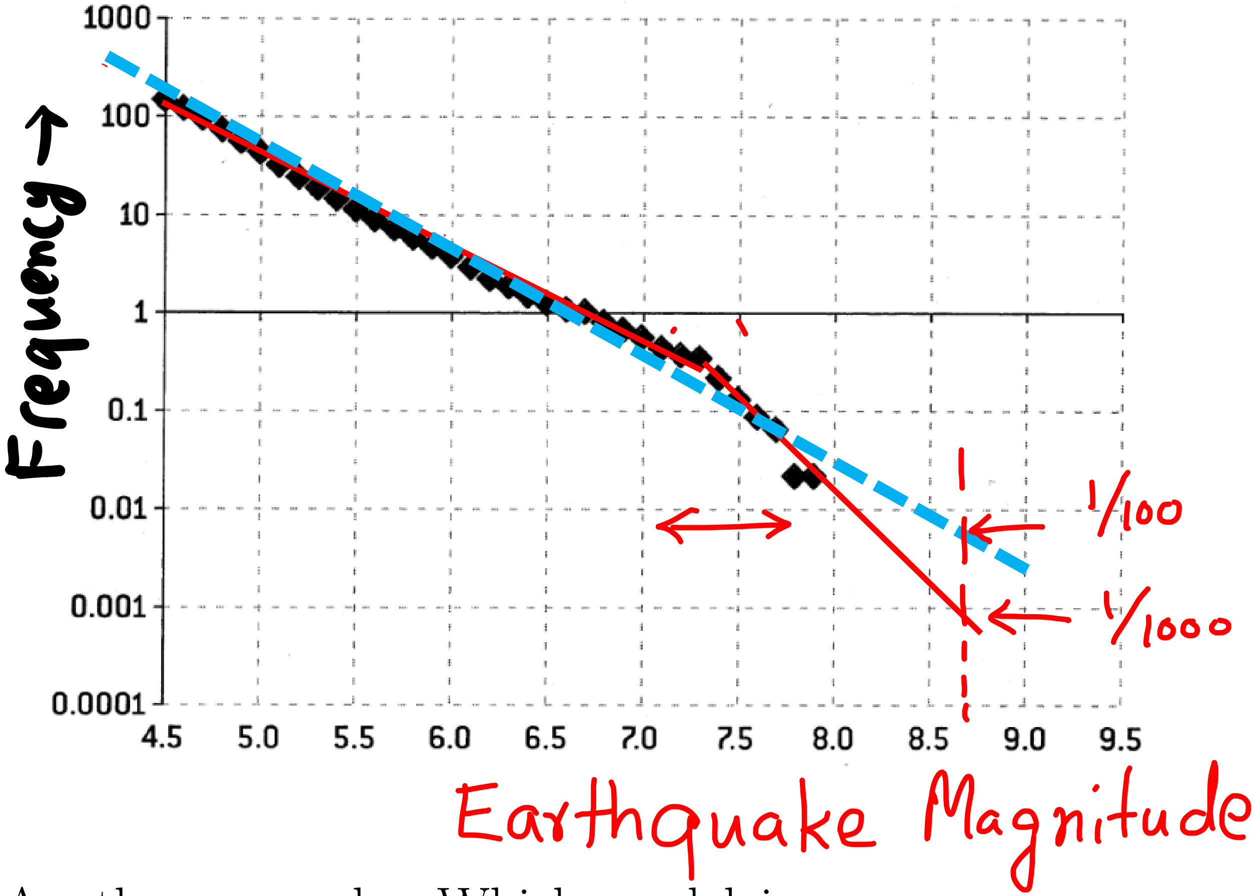
Which is a better fit?

Try a real situation. Below, y-axis is the frequency of an event and x-axis is the magnitude. It is clear that as magnitude increases, frequency decreases.

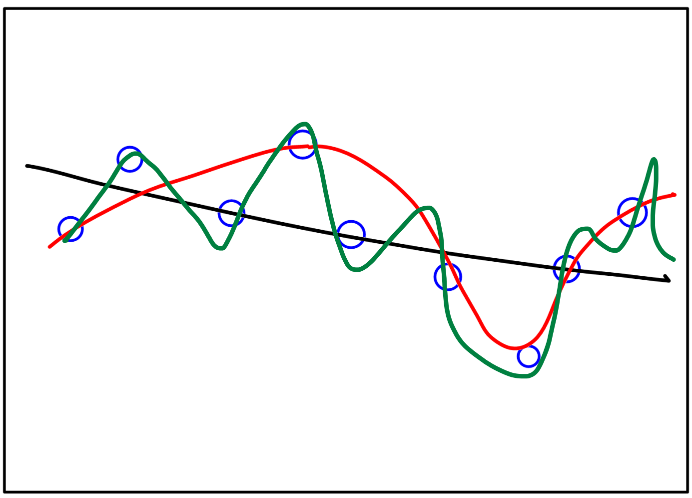


This example is taken from Nate Silver's book.

Which model is a better fit? blue or red?



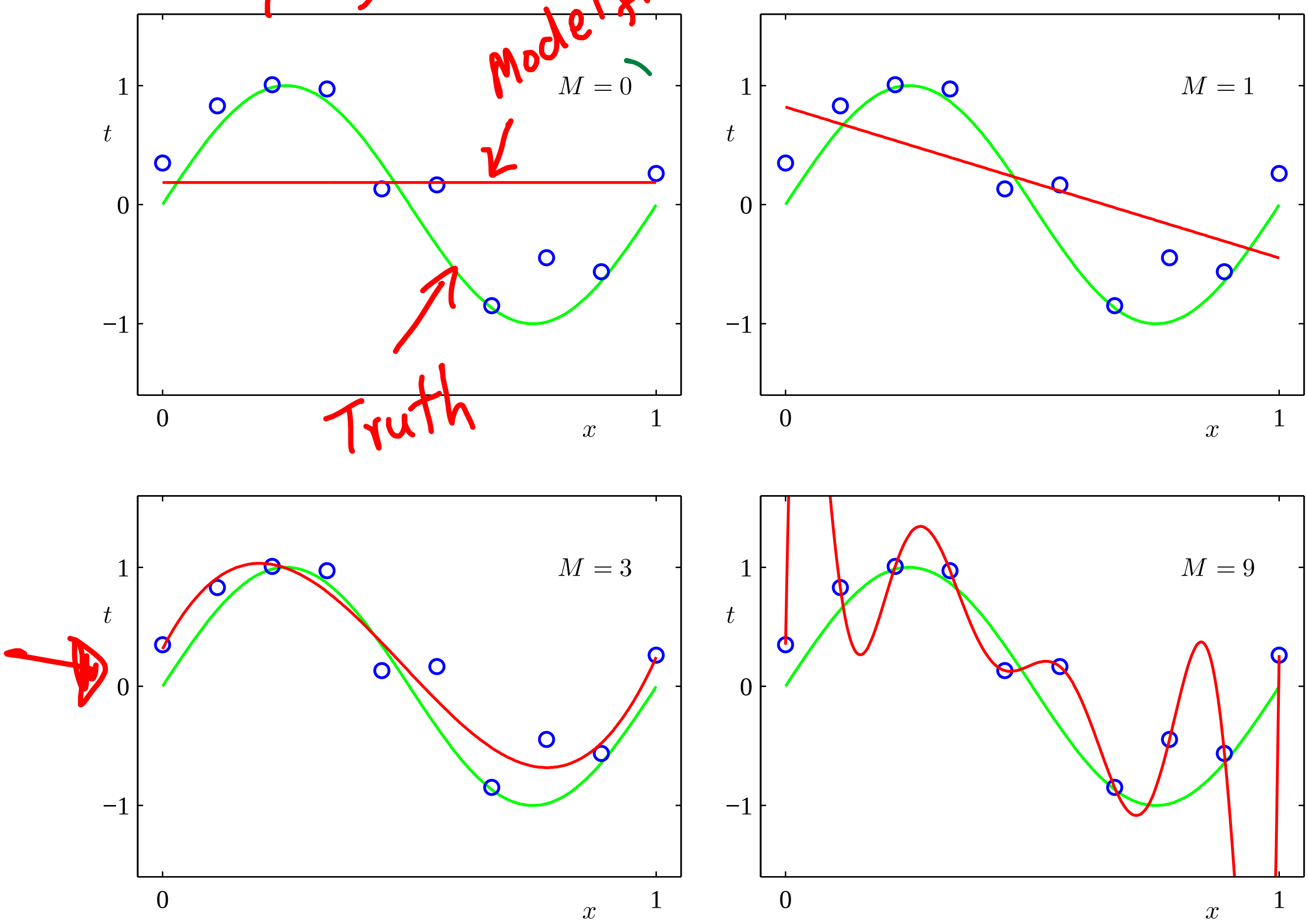
Another example: Which model is a better fit? black or red? Data is denoted by circle.



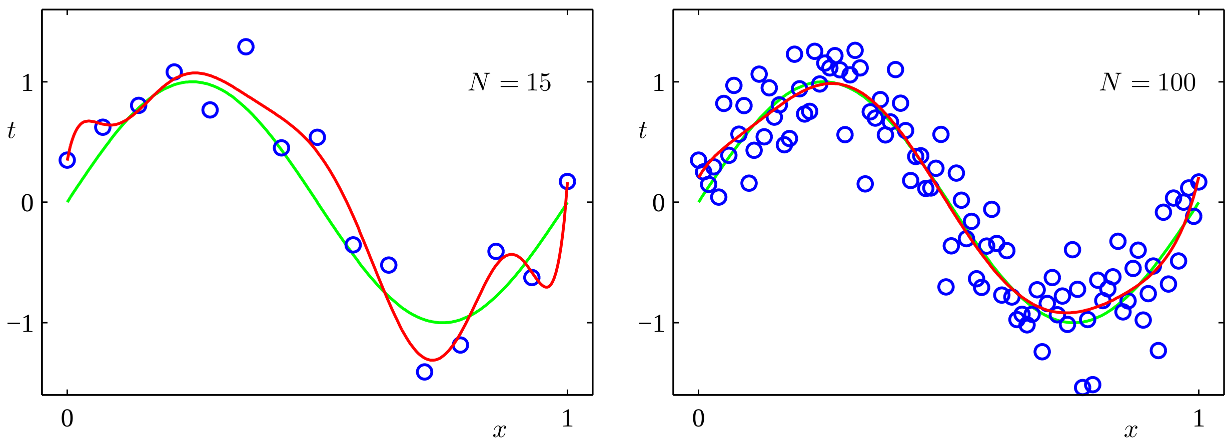
Complex models overfit easily

Circles are data points, green line is the truth & red line is the model fit. M is the maximum degree in the generated polynomial basis.

$\phi(x) = [1]$



If you increase the amount of data, overfitting *might* reduce.



Occam's razor

One solution is dictated by Occam's razor which states that "Simpler models are better – in absence of certainty."

Sometimes, if you increase the amount of data, you might reduce overfitting. But, when unsure, choose a simple model over a complicated one.

"Why you don't need to be so deep all the time"

Additional Notes

Read about overfitting in the paper by Pedro Domingos (section 3 and 5 of "A few useful things to know about machine learning"). You can also read Nate Silver's book on "The signal and the noise" (the earthquake example is taken from this book).

7 Solutions: Regularization

What is regularization?

Through regularization, we can penalize complex models and favor simpler ones:

β_0 is not in here. What could be the reason?

Q: $\lambda \to \infty$
$\lambda \to 0$

$$\min_{\beta} \mathcal{L}(\boldsymbol{\beta}) + \frac{\lambda}{\cancel{2N}} \sum_{j=1}^{M} \beta_j^2$$

The second term is a regularizer (with $\lambda > 0$). The main point here is that an input variable weighted by a small β_j will have less influence on the output.

λ is high, Simple
λ is low, Complex

Regularization Parameter

The parameter λ can be tuned to reduce overfitting. But, how do you choose λ?

"Don't minimize $\mathcal{L}$"

The generalization error

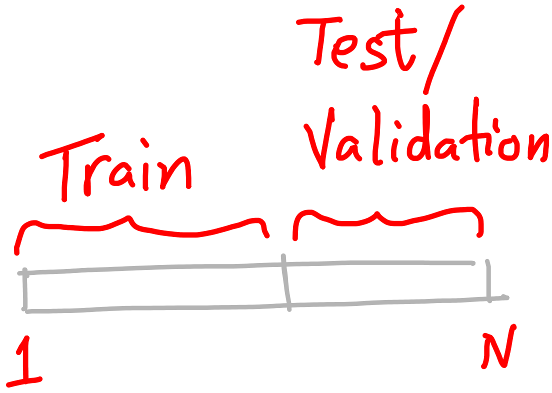
A/B testing

The generalization error of a learning method is the expected prediction error for *unseen* data, i.e. mistakes made on the data that we are going to see in the future. This quantifies how well the method *generalizes*.

It is like your "Exam"

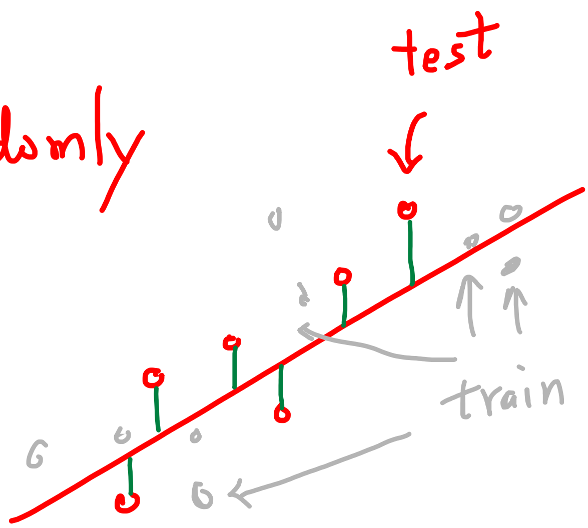
Simulating the future

Ideally, we should choose λ to minimize the mistakes that will be made in the future. Obviously, we do not have the future data, but we can always *simulate the future* using the data in hand.



Splitting the data

Randomly



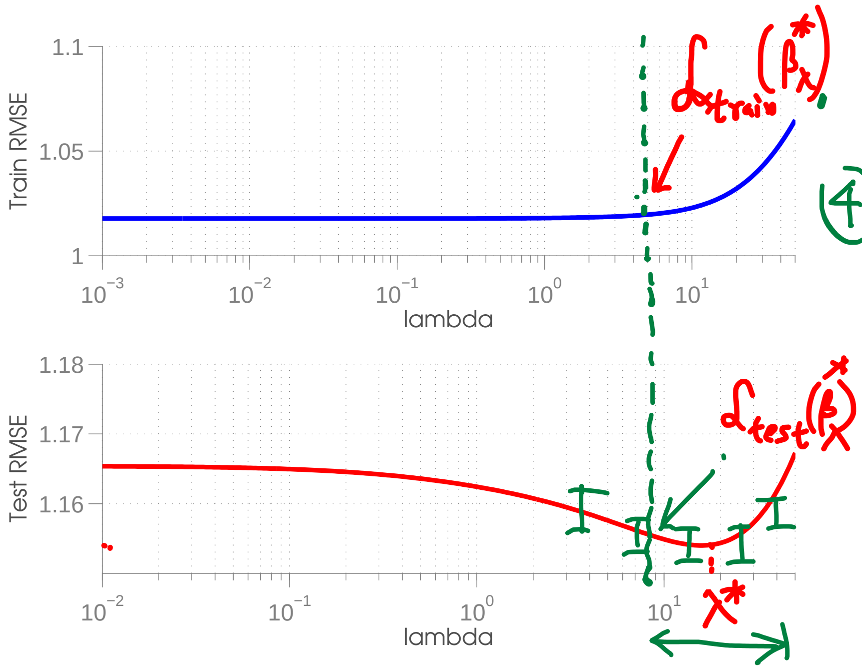
For this purpose, we split the data into train and validation sets, e.g. 80% as training data and 20% as validation data. We pretend that the validation set is the future data. We fit our model on the training set and compute a prediction-error on the validation set. This gives us an *estimate* of the generalization error (one instant of the future).

1. Choose a λ
2. Split data
3. Fit training data

$$\beta^*_\lambda = \arg\min_\beta \underbrace{\mathcal{L}_{train}(\beta)}_{\text{on training data}} + \lambda \sum_j \beta_j^2$$

4. Compute $\mathcal{L}_{test}(\beta^*_\lambda)$ on test points.

$$\mathcal{L}_{train}(\beta) = \sum_{i \in \text{Training Set}} \dots$$

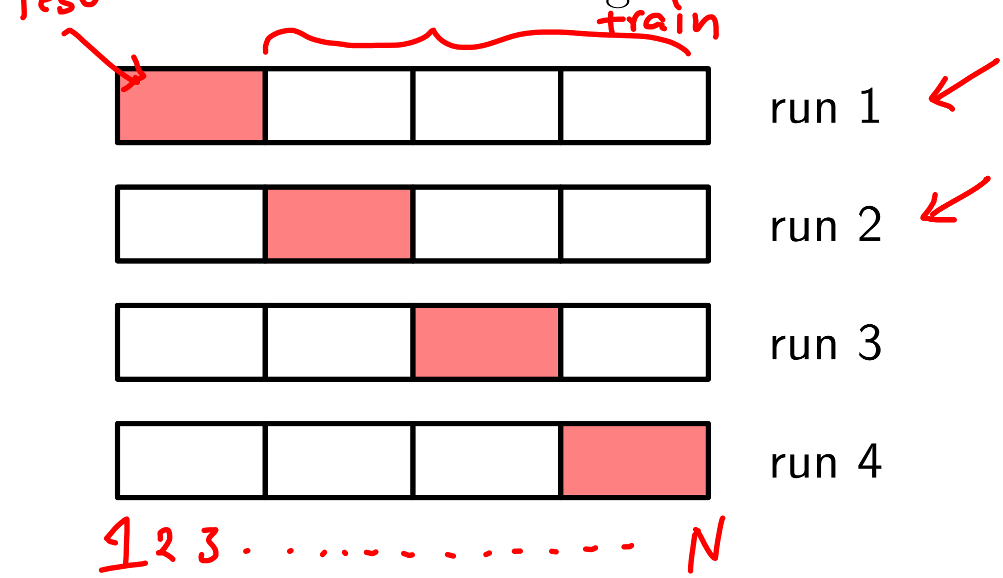


Cross-validation

Random splitting (aka bootstrap) is not an efficient method.

Q. where does it fail?

K-fold cross-validation allows us to do this efficiently. We randomly partition the data into K groups. We train on $K - 1$ groups and test on the remaining group. We repeat this until we have tested on all K sets. We then average the results.



Cross-validation returns an estimate of the *generalization error.*

Additional Notes

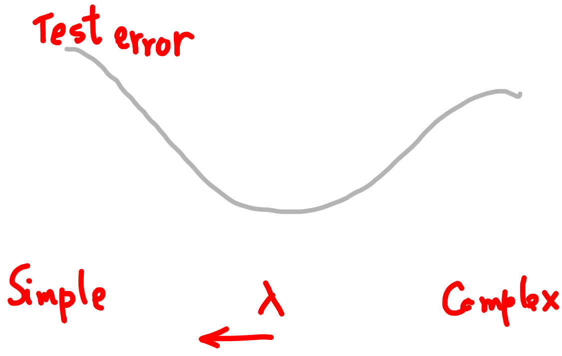
Details on cross-validation are in Chapter 7 in the book by Hastie, Tibshirani, and Friedman (HTF). You can also read about bootstrap in Section 7.11 in HTF book. This method is related to random splitting and is a very popular method.

8 Bias-Variance Decomposition

What is bias-variance?

"choosing λ"

One natural question is how does the test error vary wrt λ? When λ is high, the model underfits, while when λ is small, the model overfits. Therefore, a good value is somewhere in between.



Bias-variance decomposition explains the shape of this curve.

Generalization error

Given training data $\mathcal{D}_{tr}$ of size N, we would like to estimate the expected error made in future prediction. This error is the generalization error. Below is a definition suppose that we have infinite test data $\mathcal{D}_{te}$,

$\mathcal{D}_{test} = \{ \dots \}$ ∞

$$teErr(\mathcal{D}_{tr}) := \mathbb{E}_{\mathcal{D}_{te}}[\{y - f(\mathbf{x})\}^2]$$ ← Unknown

all of the future

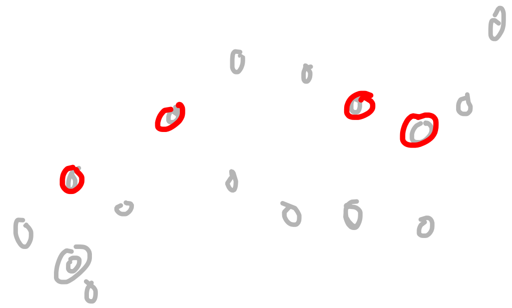
Generalization error is different from the training error which measures how well you fit the data.

$$trErr(\mathcal{D}_{tr}) := \sum_{n=1}^{N}[\{y_n - f(\mathbf{x}_n)\}^2]$$ ← Unknown when N is large

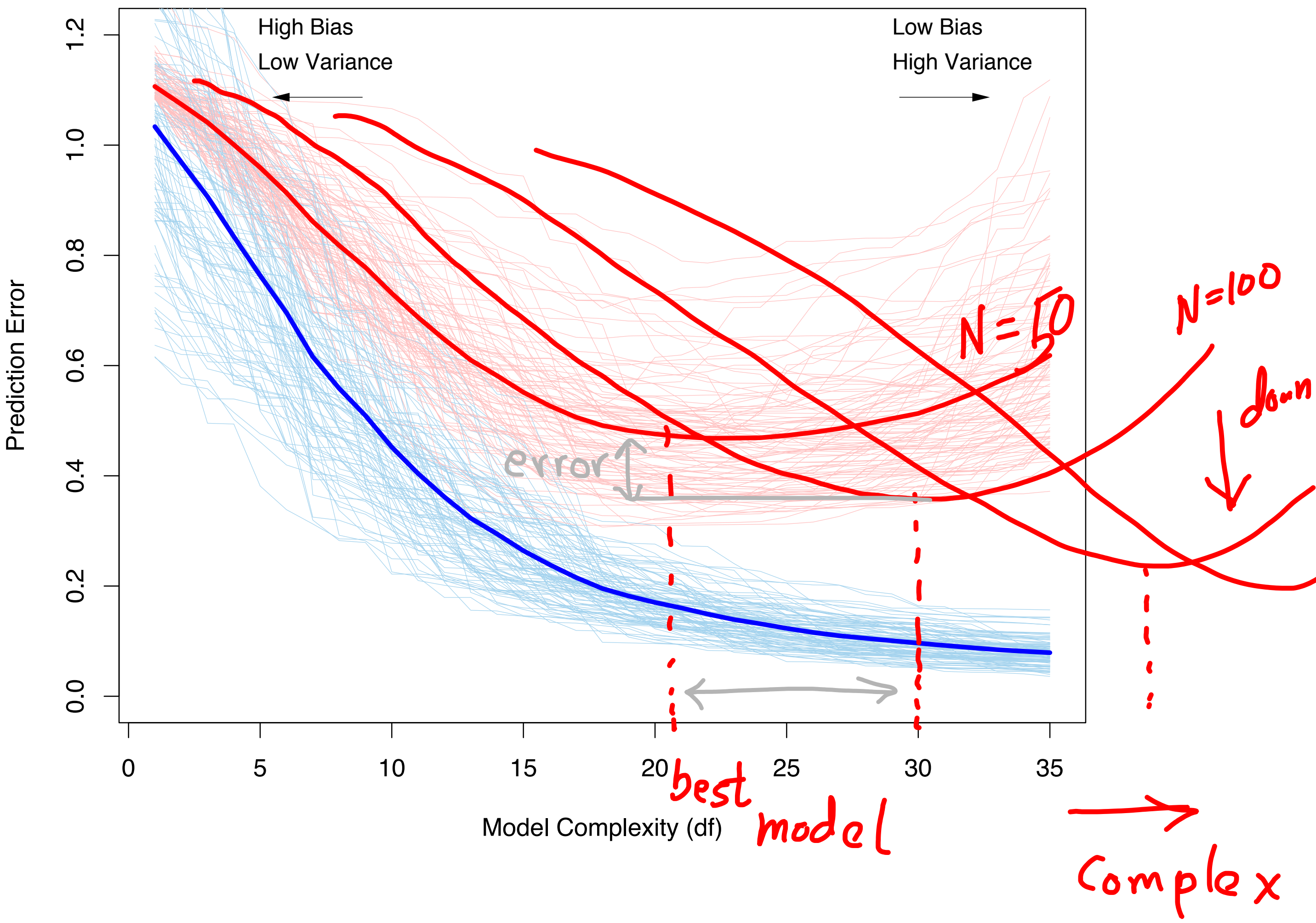
Errors vs model complexity

As we increase the model complexity, how do these errors vary? The blue line shows training error for a dataset with $N = 50$, while the red line shows the generalization error for that dataset.

"If you were God"



Simple model have high train and generalization error since they have a high bias, while complex model have low train but high generalization error because they have high variance.



Bias-variance decomposition

The shape of these curves can be explained using bias-variance decomposition. The following four points can be explained by using the decomposition:

1. both bias and variance contribute to generalization error.

2. For bias, both model-bias and estimation-bias are important. When we increase model complexity, we increase generalization error due to increased variance.

3. ~~Regularization increases estimation bias while reducing variance.~~

9 Recent Advances

Deep Learning & Overfitting

Deep learning has shown a new (but old) way to combat overfitting. For many applications, more data and deep architecture combined with stochastic gradient-descent is able to get us to a good minimum which generalizes well.

Challenges

There are many challenges ahead. Learning from nasty, unreliable data still remains a challenge (e.g. small sample size, redundant data, non-stationary data, sequential learning).

On the other hand, living beings - even young ones - are very good in dealing with such data. How do they do it, and how can we design ML methods that can learn like them?