Ensemble Methods

Bias and Variance

- Bias measures the difference between expected value of an estimator (E[Y]) and the "true" value (p).

$$\text{Bias} = E[Y] - p$$

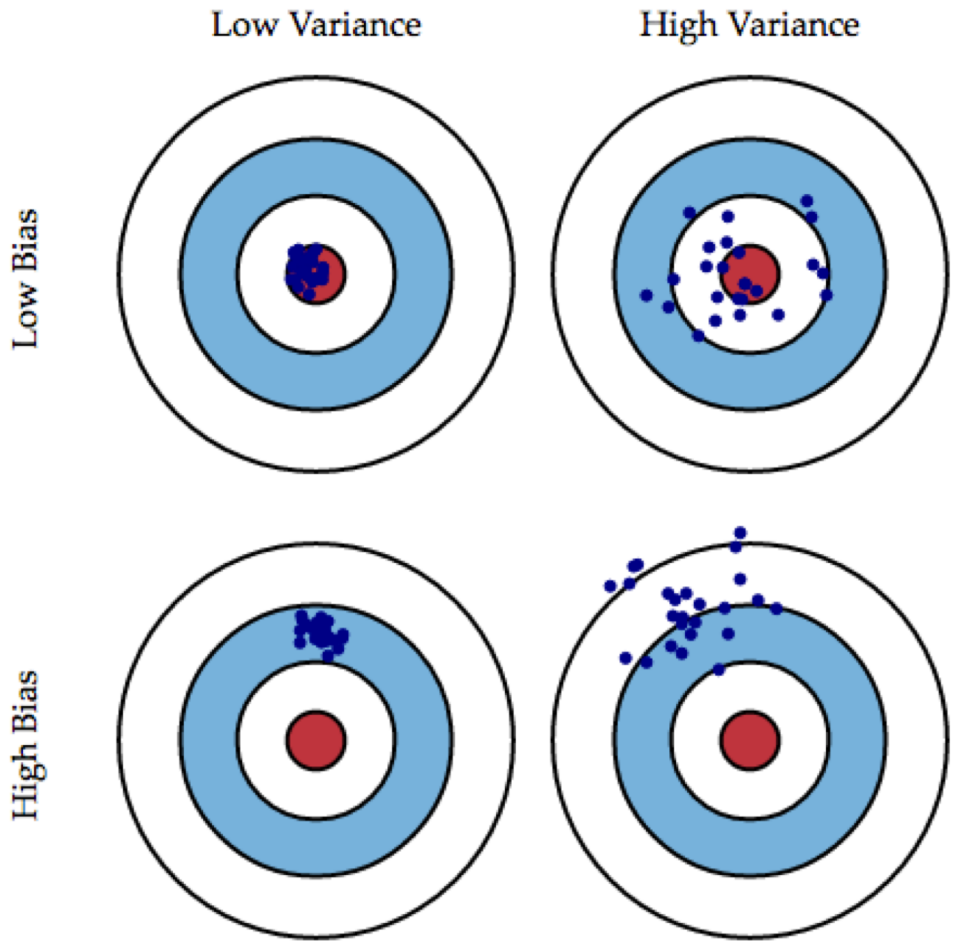
A lower bias is preferred

- Variance measures how much the estimate varies between samples. A low variance is generally preferred.

Intuition

- Bias:
  - measures accuracy of quality of an algorithm.
  - high bias means a poor match.
- Variance:
  - measures the precision or specificity of a match
  - Did the algorithm just get lucky on a few samples? Or is it really a good classifier?
  - a high variance means a weak match.
- Ideally, we would like to minimize both of them. Is that possible?

Bias-Variance



Mathematical Derivation

- Let's consider the case of regression
- True function is:

$$y = f(x) + \epsilon$$

where

y is the output,
f(x) is the "best" possible function
ϵ is a term that accounts for the noise in data

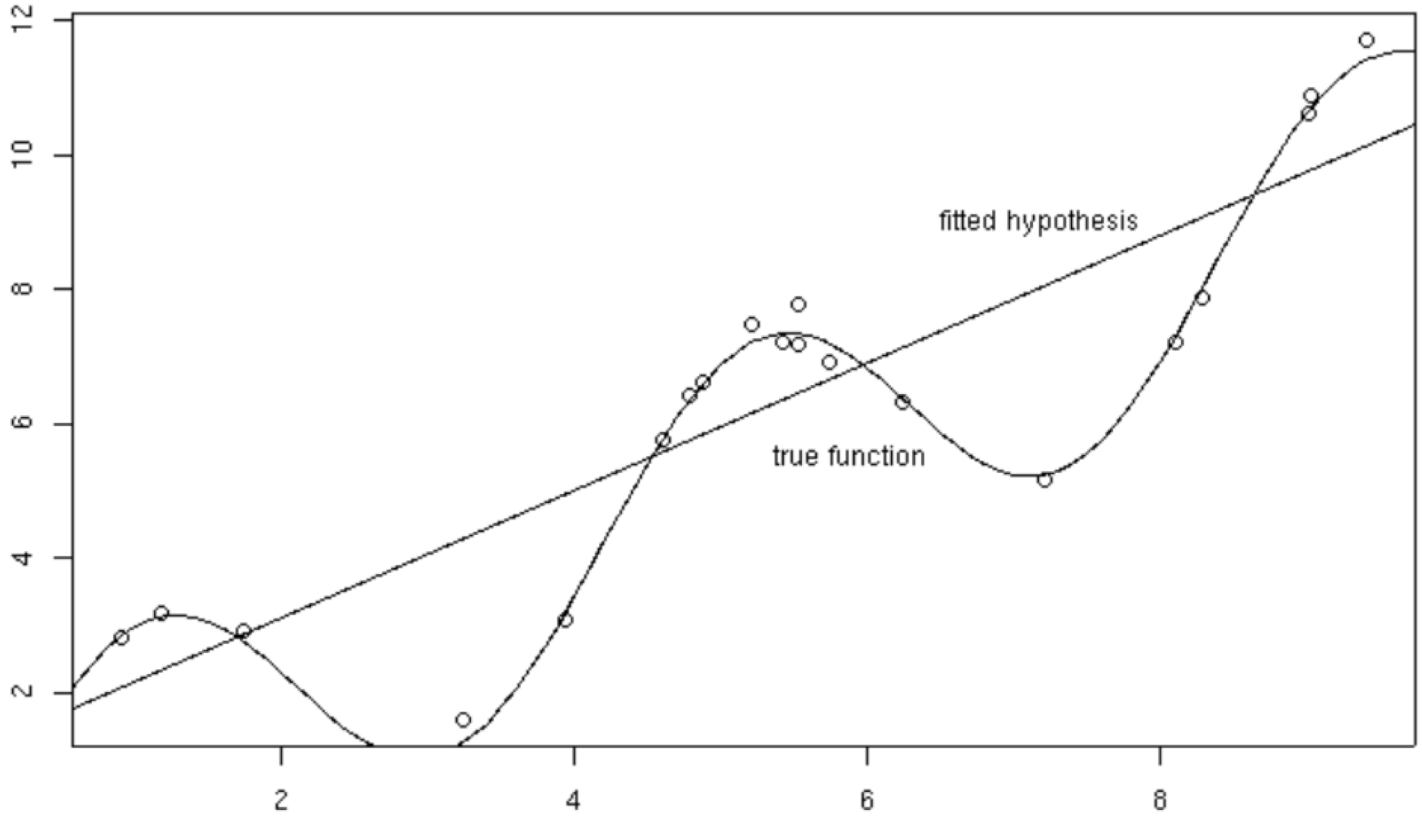
We are assuming that there is some noise in the data.

- Our hypothesis is $h(x) = w^T x + b$
- We seek to minimize squared error:

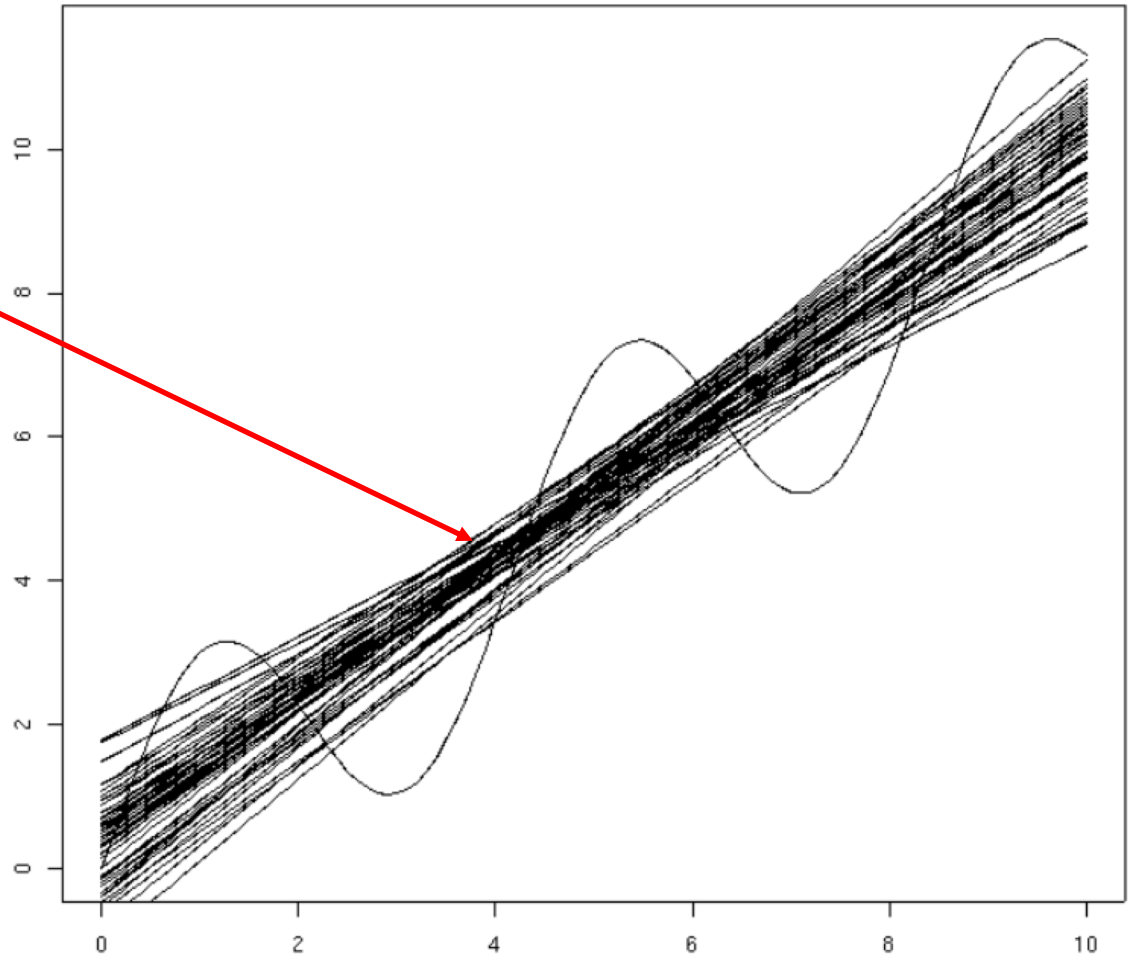
$$E = \sum_i [y_i - h(x_i)]^2$$

Example: 20 points

$y = x + 2 \sin(1.5x) + N(0,0.2)$ Noise term



50 samples of 20 instances each are sampled and hypothesis is tested on them.



Bias – Variance Analysis

- Now, given a new data point x*, with observed value $y^* = f(x^*) + \epsilon$
- Prediction of our hypothesis = h(x*)
- Let's examine expected value of squared error:

$$E[(h(x^*) - y^*)^2]$$

- Before going there, let's look at a lemma that will help us.

Bias – Variance Analysis

- Let Z be a random variable with distribution P(Z) and average value defined as:

$$\bar{Z} = E_P[Z]$$

- Lemma: $E[(Z - \bar{Z})^2] = E[Z^2] - \bar{Z}^2$

- Proof: $E[(Z - \bar{Z})^2] = E[Z^2 + \bar{Z}^2 - 2Z\bar{Z}]$

$$= E[Z^2] + \bar{Z}^2 - 2\bar{Z}E[Z]$$

$$= E[Z^2] - \bar{Z}^2$$

Using linearity property of E and the fact that
E[Z] = $\bar{Z}$; also the fact that
E[C] = C where C is a constant.

- Corollary: $E[Z^2] = E[(Z - \bar{Z})^2] + \bar{Z}^2$

Bias Variance Analysis

- Let's go back to what we were trying to solve:

$$E[(h(x^*) - y^*)^2] = E\left[h(x^*)^2 - 2h(x^*)y^* + y^{*2}\right]$$

$$= E[h(x^*)^2] - 2E[h(x^*)]E[y^*] + E[y^{*2}]$$

$$= \left\{E\left[\left(h(x^*) - \overline{h(x^*)}\right)^2\right] + \overline{h(x^*)}^2\right\} - 2\overline{h(x^*)}f(x^*) +$$

$$\{E\left[\left(y^* - f(x^*)\right)^2\right] + f(x^*)^2\}$$

$$= E\left[\left(h(x^*) - \overline{h(x^*)}\right)^2\right] + \left(\overline{h(x^*)} - f(x^*)\right)^2 + E\left[\left(y^* - f(x^*)\right)^2\right]$$

$$= var\left(h(x^*)\right) + Bias\left(h(x^*)\right)^2 + Noise^2$$

We have used the fact that $E[y^*] = E[f(x^*) + \epsilon] = f(x^*)$ as $E[\epsilon] = 0$ and f is a function output [not random variable]

This is the noise term as $y^* - f(x^*) = \epsilon$

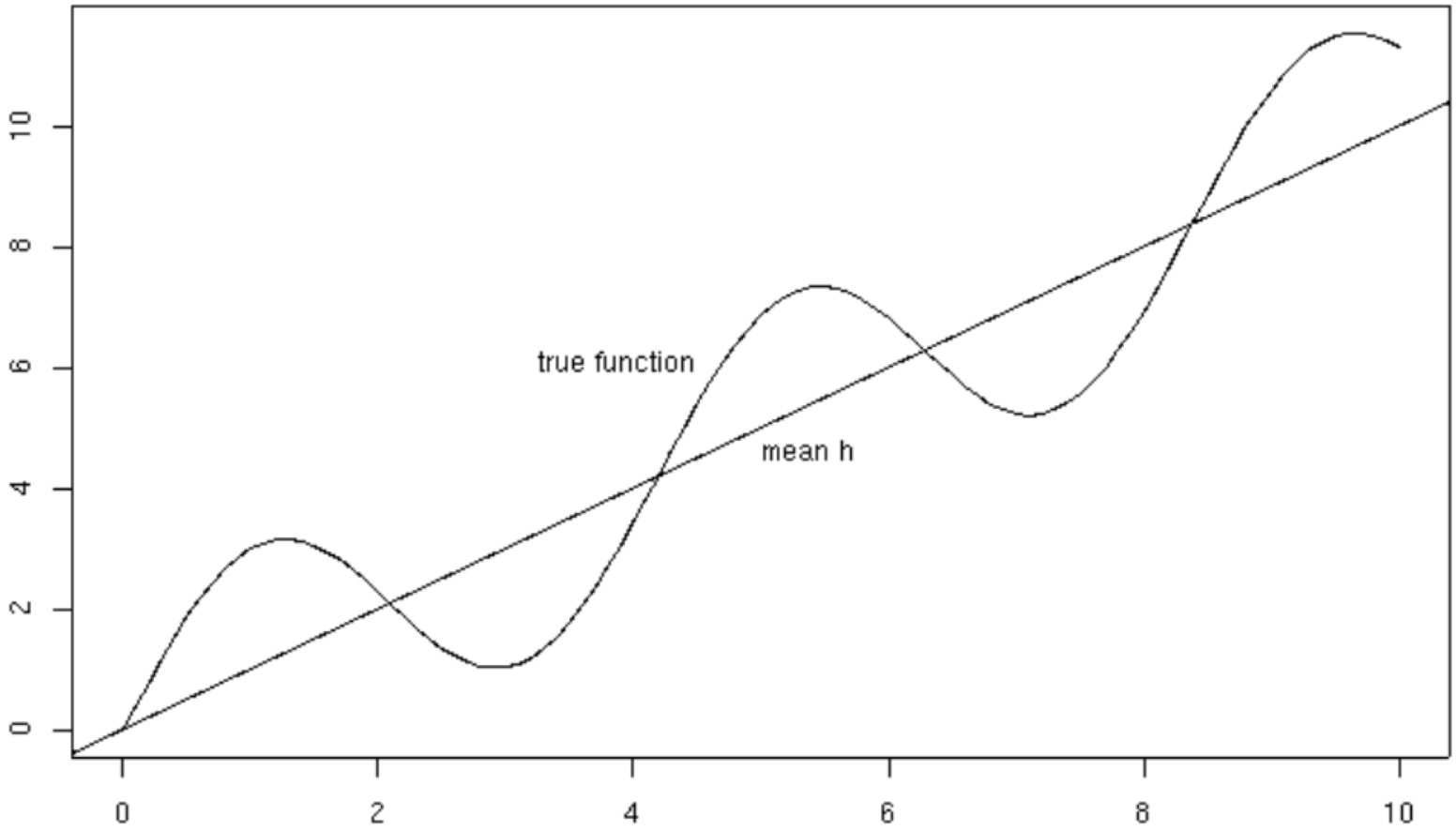
Bias Variance Tradeoff

- Summarizing,

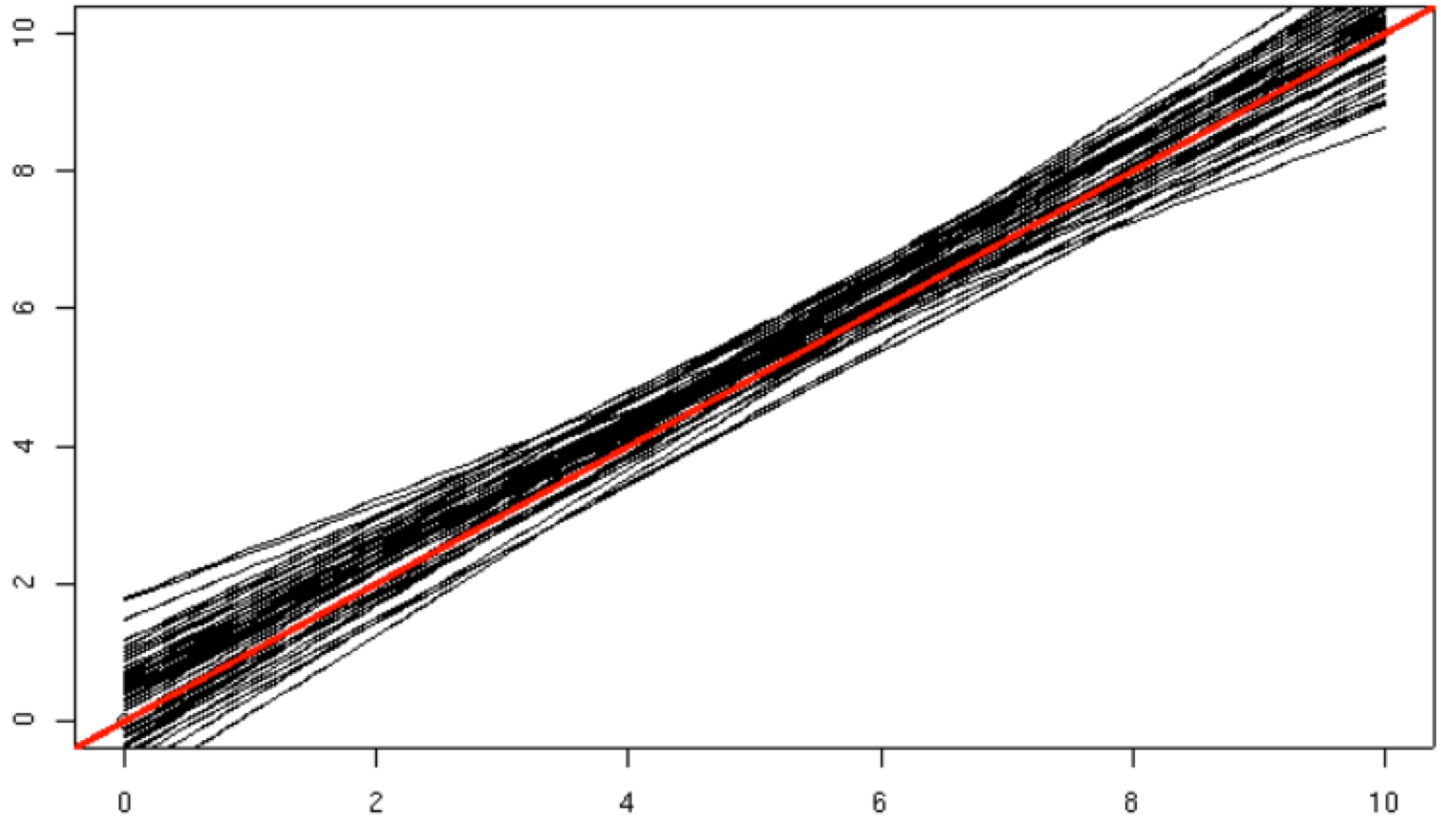
$$E[(h(x^*) - y^*)^2] = var\big(h(x^*)\big) + Bias\big(h(x^*)\big)^2 + Noise^2$$

- Variance = How much h(x*) varies from its average value over multiple samples.
- Bias = Describes average error of h i.e. how far is average error from actual expected value
- Noise = Describes how much the best function differs from actual value

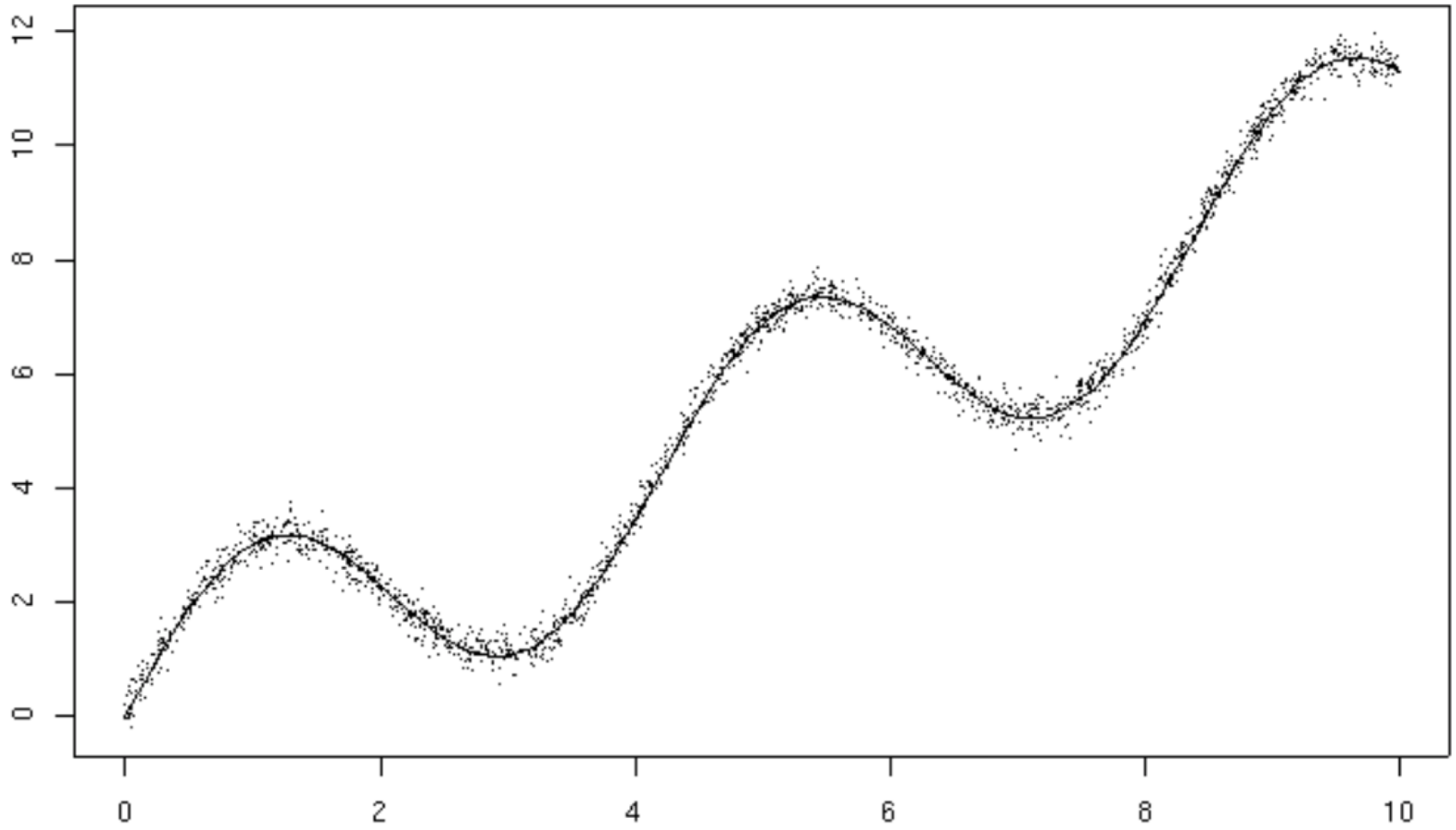
Bias



Variance

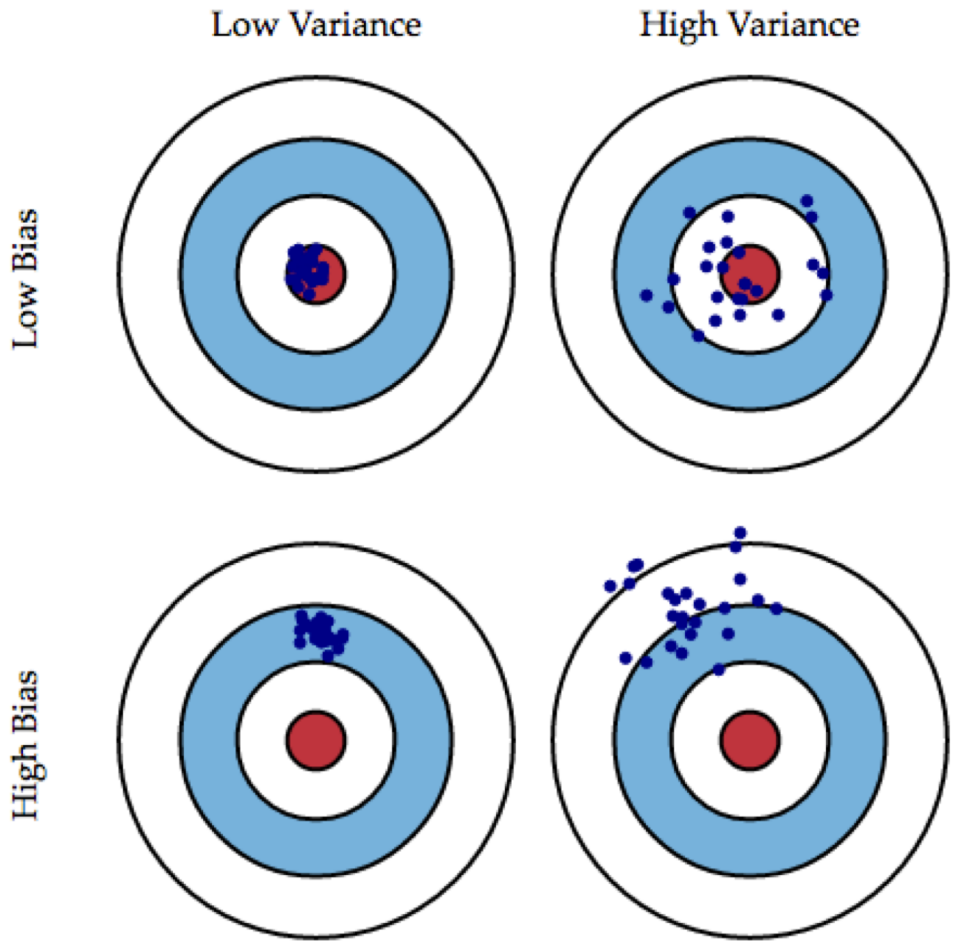


Noise



Bias-Variance

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$



Bias Variance Tradeoff

- For machine learning models, there is always a tradeoff between bias and variance.
- Generally, we can't have low bias and low variance for real world data.
- The next slides explain this concept.

Bias – Variance Tradeoff

- Low Bias – model gets close to actual value
  - linear regression applied to linear data
  - 2^{nd} degree polynomial applied to quadratic data
- High Bias – model is a poor approximator
  - constant function applied to highly non-linear data
  - linear regression applied to non-linear data
- Low Variance – for different samples, output of model doesn't change much
  - constant function always
  - simple model independent of data
- High Variance – for different samples, output of the model changes a lot
  - complex model e.g. neural net
  - high degree polynomial function

Bias Variance Tradeoff

- Can I get both low bias and low variance?
- To get low bias
  -> you need to get very close to the real function.
  -> need a complex model that has a number of parameters.
  -> parameters depend on data
  -> so for each sample, you get a different output from model.
- If you get different output from the model for different samples, you have a high variance.
- So, I get a low bias but ended up with high variance.

Bias Variance Tradeoff

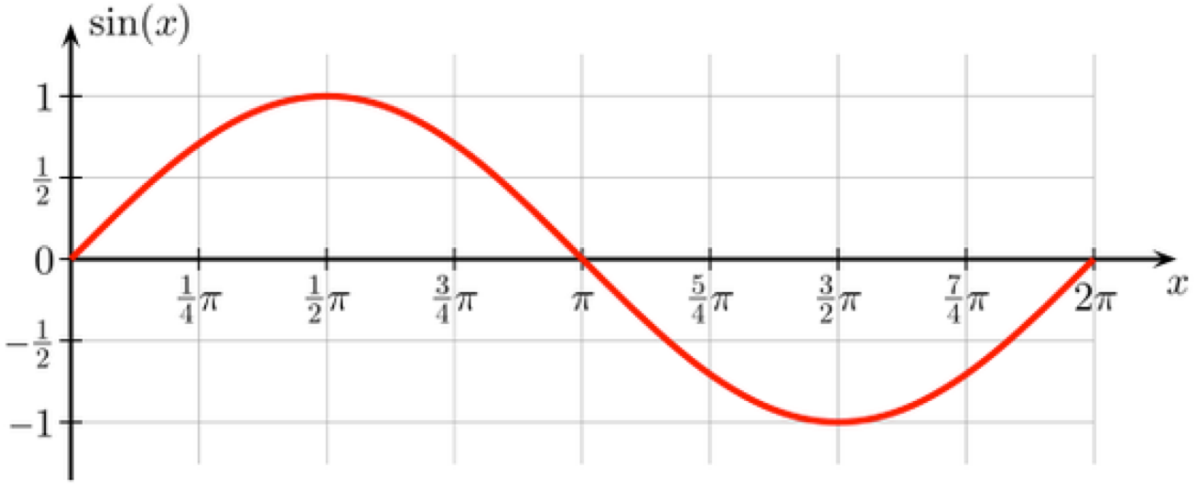
- Can I get both low bias and low variance?
- To get low variance
  -> you need a model whose output doesn't change across multiple samples
  -> Suppose you choose a constant function. It doesn't change its output for different data samples.
- But, the output will be very different from the real function that you are trying to approximate => High Bias
- So, I get a low variance but ended up with high bias.

Why can't we reduce both error and variance

- If we make the mode more complex, bias goes down but variance goes up.
- If we create a very simple model, variance is low but bias goes up.

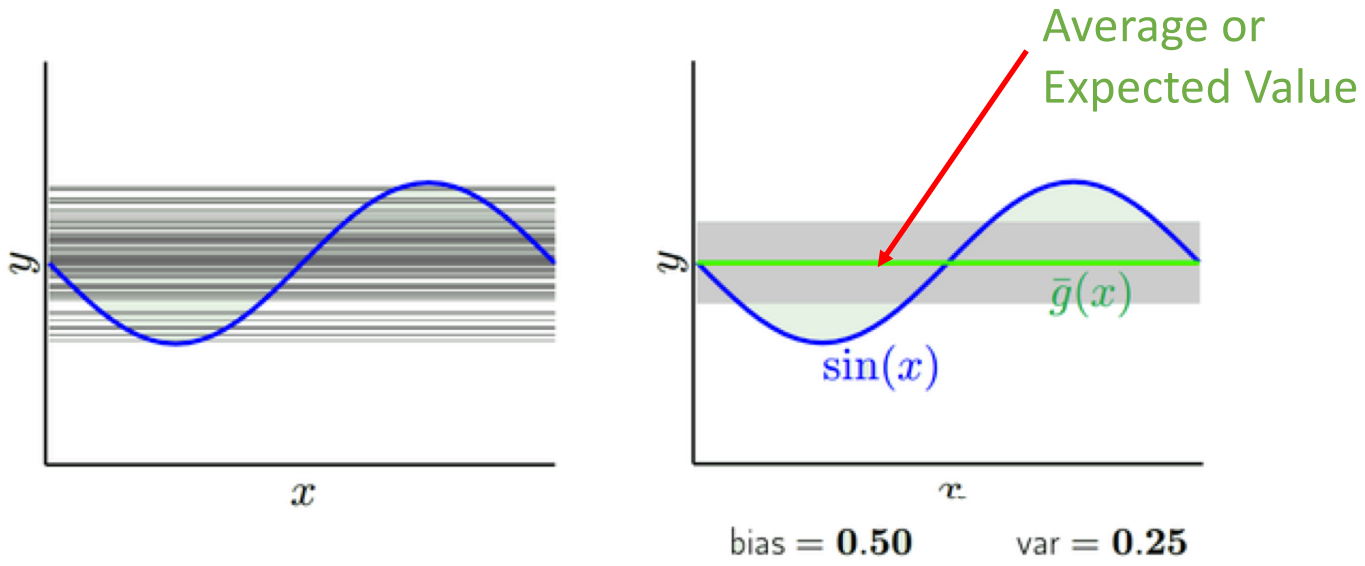
Bias / Variance

- Suppose, we are trying to learn the sine function:



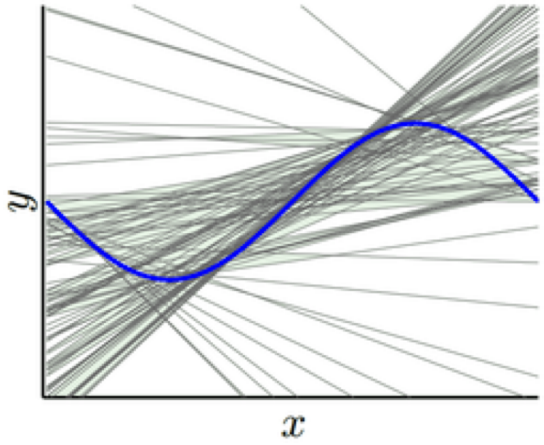
Bias/Variance

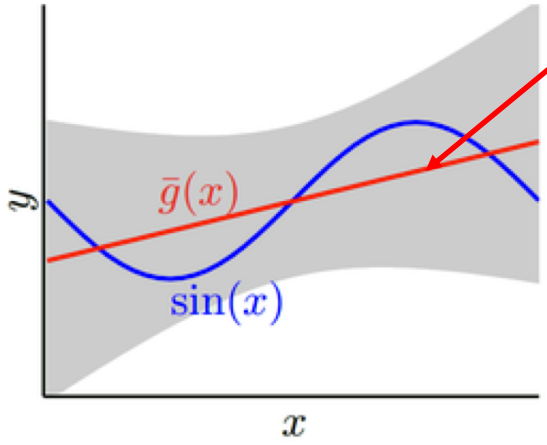
- We have a small sample dataset e.g. only 2 training points.
- Case I:
  Model: $h_0 = b$
  where b is a parameter that we optimize



Bias/Variance

- Case II :
  Model: $h_1(x) = a\,x + b$
  where b is a parameter that we optimize

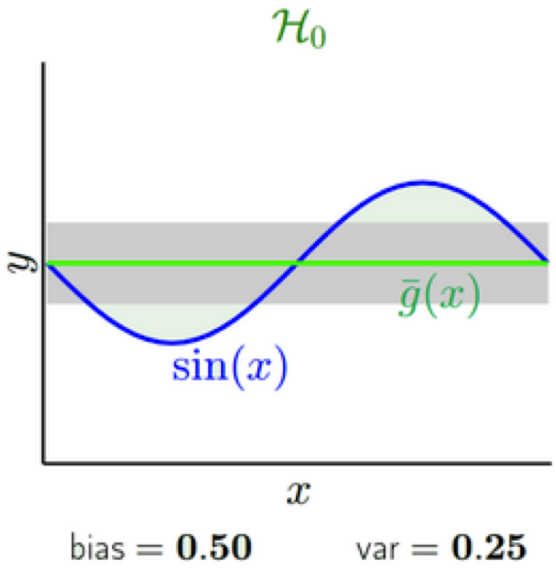


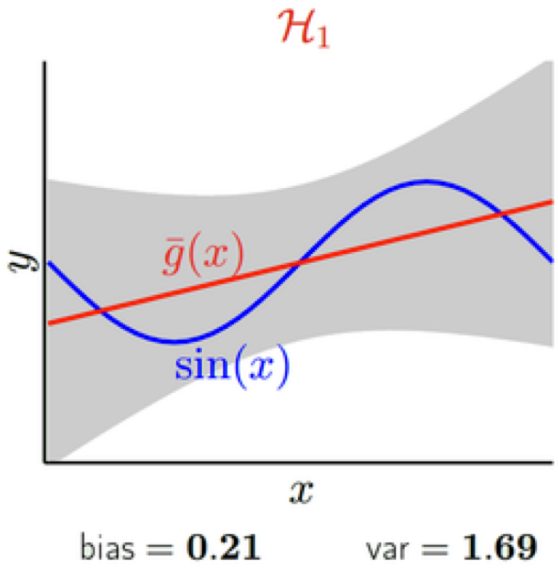


Average or
Expected Value

bias = **0.21** var = **1.69**

Comparison

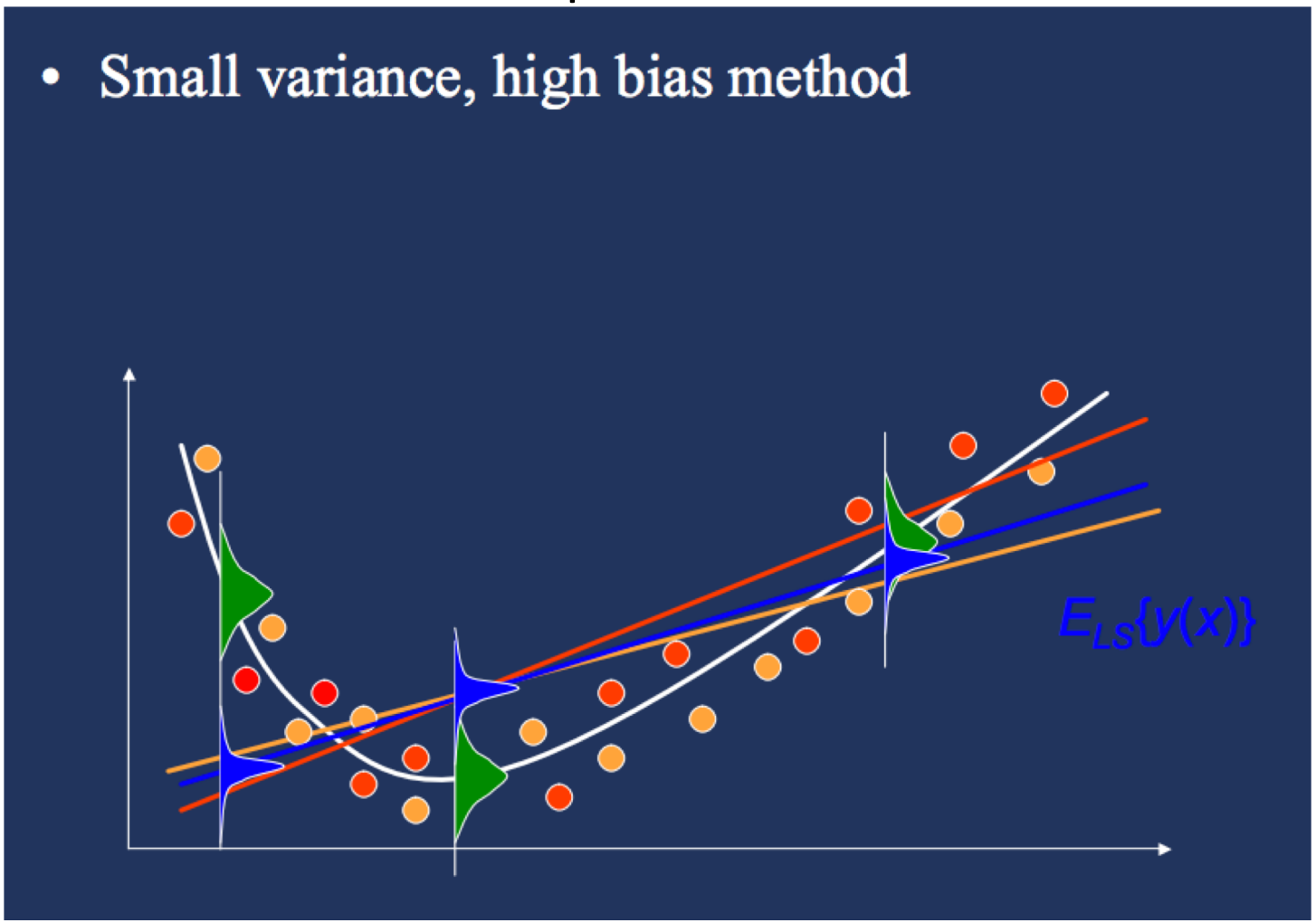




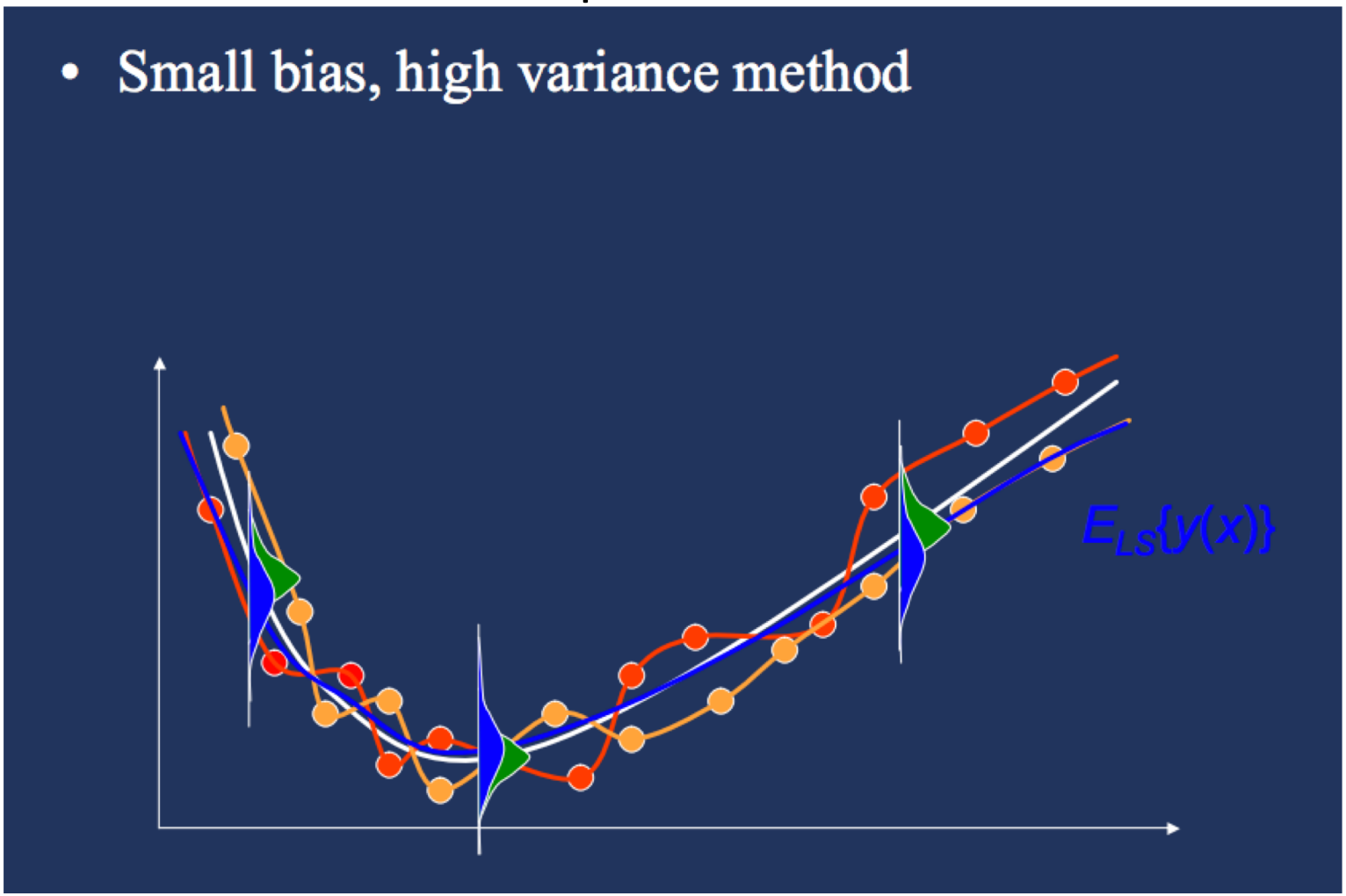
Observations:

- By making the model more complex, we have reduced bias but increase variance.
- A simple model has low variance, but high bias.
- Bias Variance Tradeoff

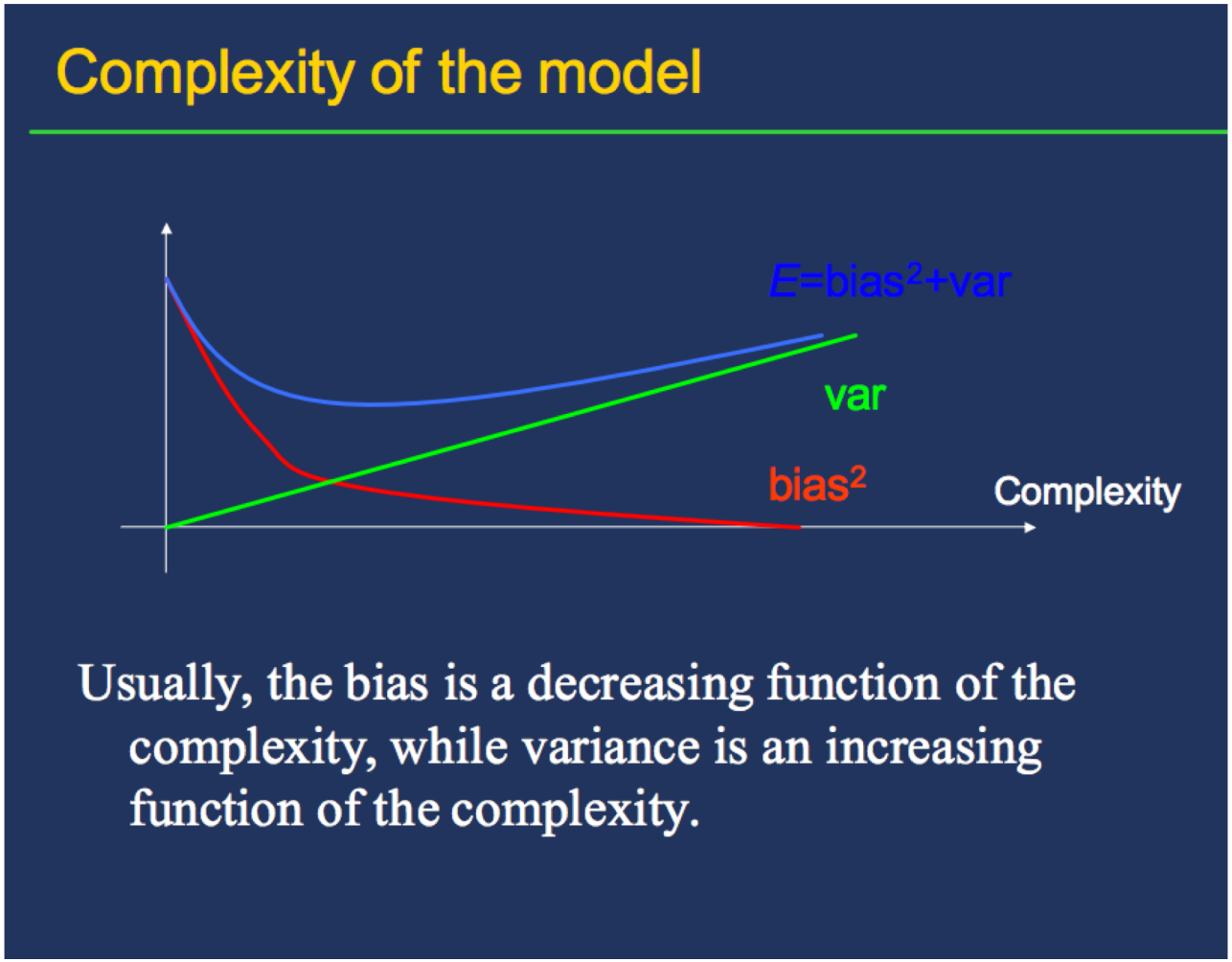
Bias-Variance Example 2



Bias-Variance Example 2



Model Complexity



How does all of this help me in ML?

- You begin to ask – do I really need to understand all of this?
- Keep in mind that my wish list is:
  - low bias
  - low variance
  - reasonable model complexity
  - avoid overfitting

How does all of this help me in ML?

- You begin to ask – do I really need to understand all of this?
- Keep in mind that my wish list is:
  - low bias
  - low variance
  - reasonable model complexity
  - avoid overfitting
- Can I get all of the above?

Classifiers based on idea of bias and variance

Background - Variance

- Property of Variance:

$$var(aX + b) = a^2 var(X)$$

- Let's compute variance of the mean of a random variable:

$$var(\bar{X}) = var(\frac{1}{N}\sum_i X_i) = \frac{1}{N^2}(\sum_i var(X_i)) = \frac{1}{N}\frac{\sum_i var(X_i)}{N}$$

- Suppose each X_i represents the output of a model on a different dataset.
  If we take the variance of the average of these outputs, it will be lower than the average of the variances by a factor of 1/N

Background - Variance

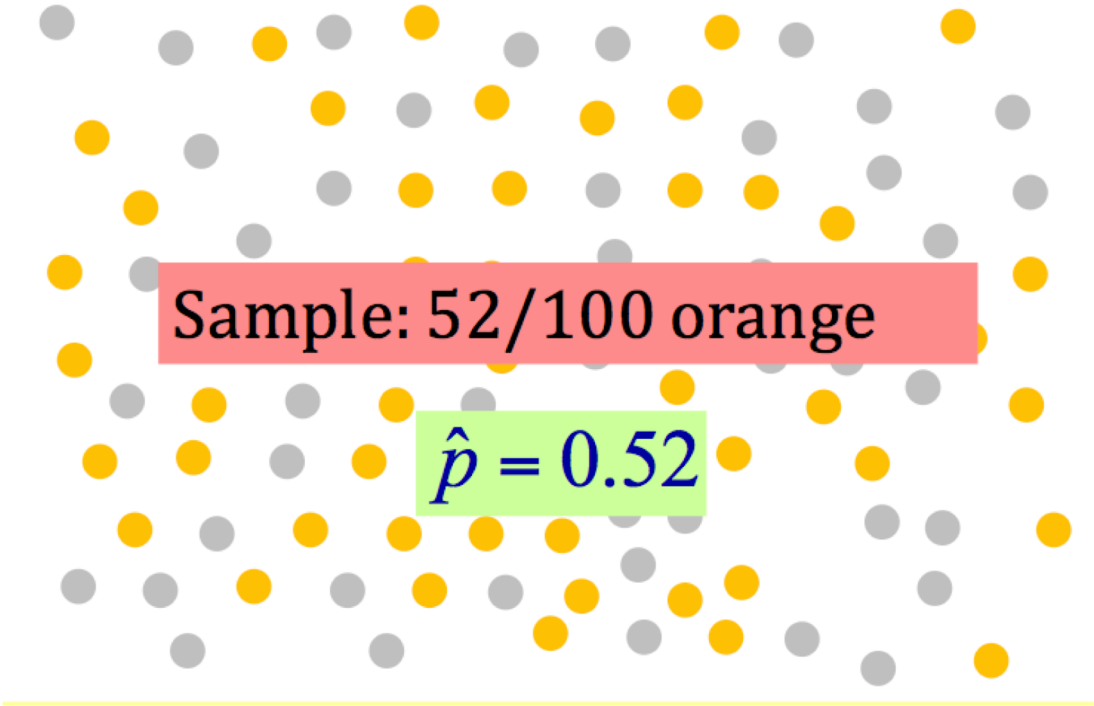
- How do we get different datasets?
- In reality, you have only 1 dataset available and don't have the power to sample over and over from a population.
- How do we simulate random sampling using only the data available?

BOOTSTRAP

- Bootstrap – create new samples of the same size from original dataset by sampling with replacement.

Bootstrap - Example

- I have one data sample.
- I can't sample again from the population.
- Can I simulate new samples from this sample?
- Sample with replacement from the dataset available.

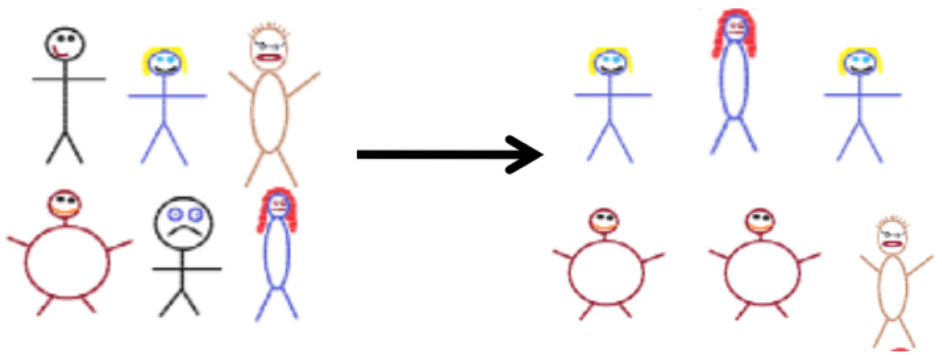


Bootstrap - Example

- Suppose I have a sample of 6 people.
- That's all the data that you have to work with.
- I'd like another sample of same size.

Bootstrap - Example

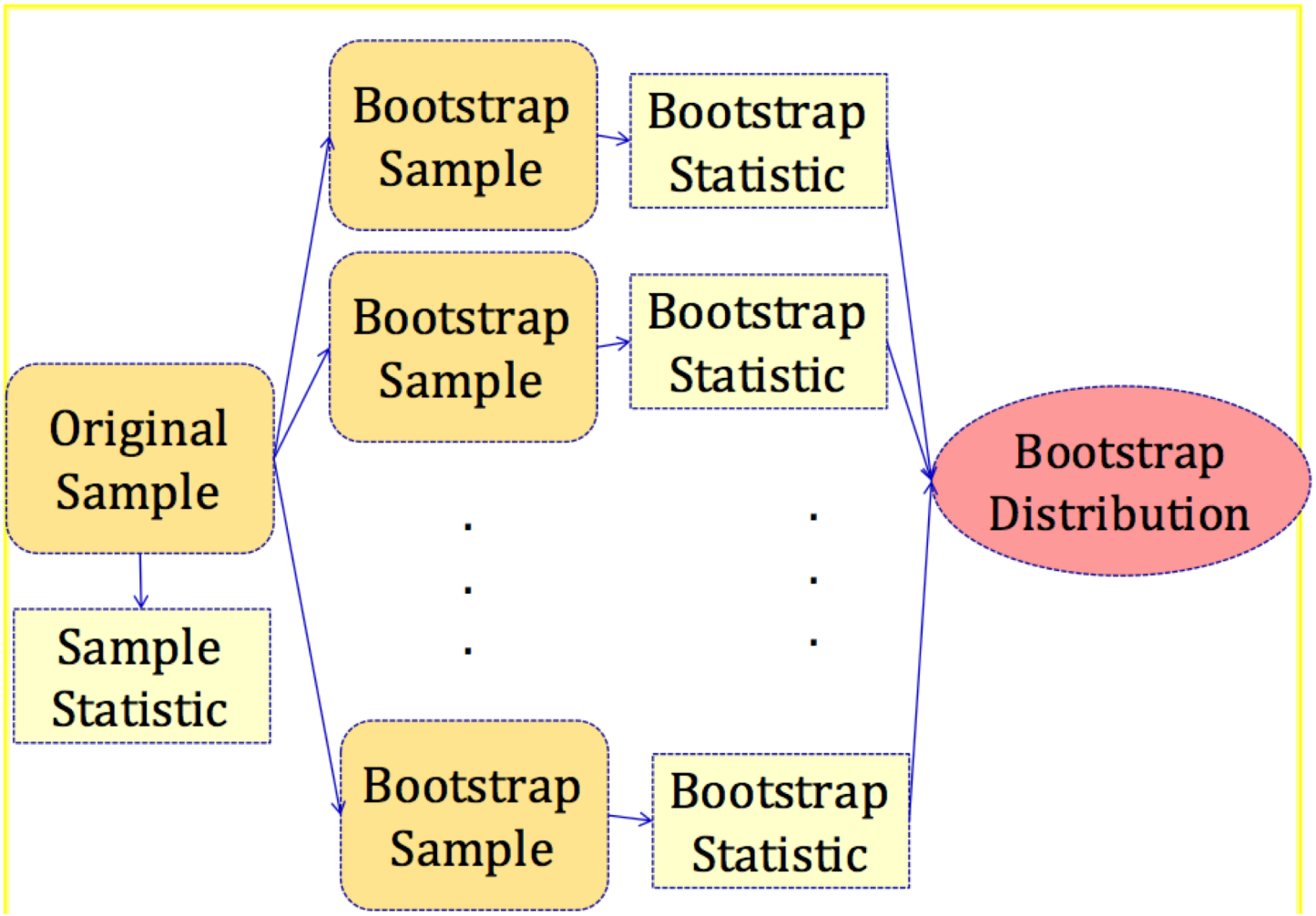
- Suppose I have a sample of 6 people.
- That's all the data that you have to work with.
- I'd like another sample.
- Idea of sampling with replacement: For each slot, I have all n choices.



Original Sample

Bootstrap Sample

Bootstrap Distribution



Bootstrap

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21, 22

Bootstrap

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21, 22

a) Yes
b) No

22 is not a value from the original sample

Bootstrap

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21

Bootstrap

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21

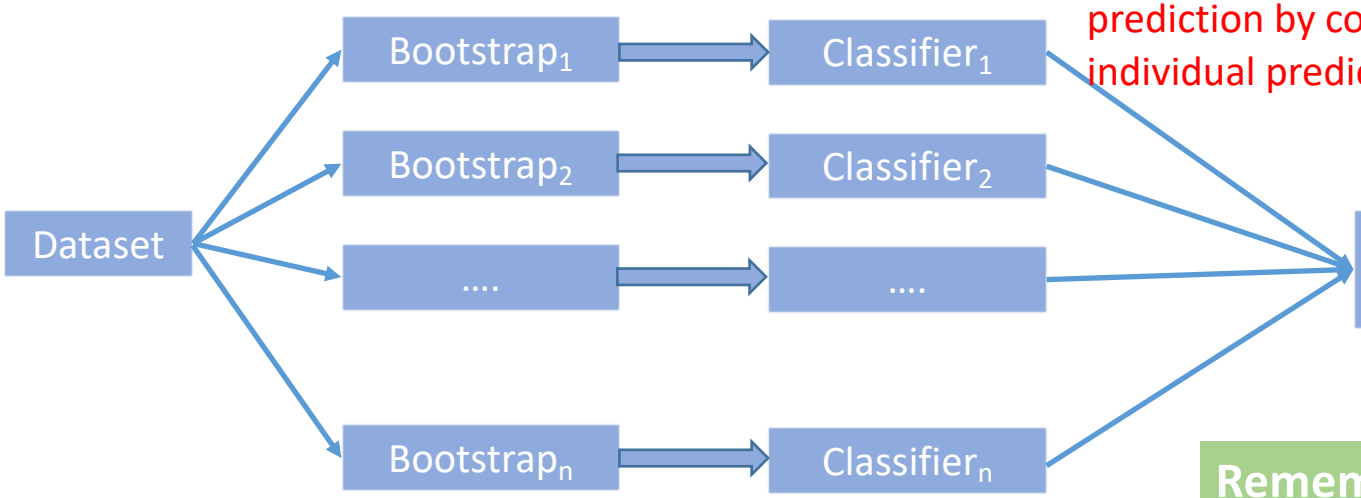
a) Yes

b) No

Bootstrap samples must be the same size as the original sample

Bagging – Bootstrap Aggregation

- A set of classifiers based on the idea of bootstrap + aggregating the output of individual classifiers.



How to combine?
Classification = majority label
Regression = average of predictions

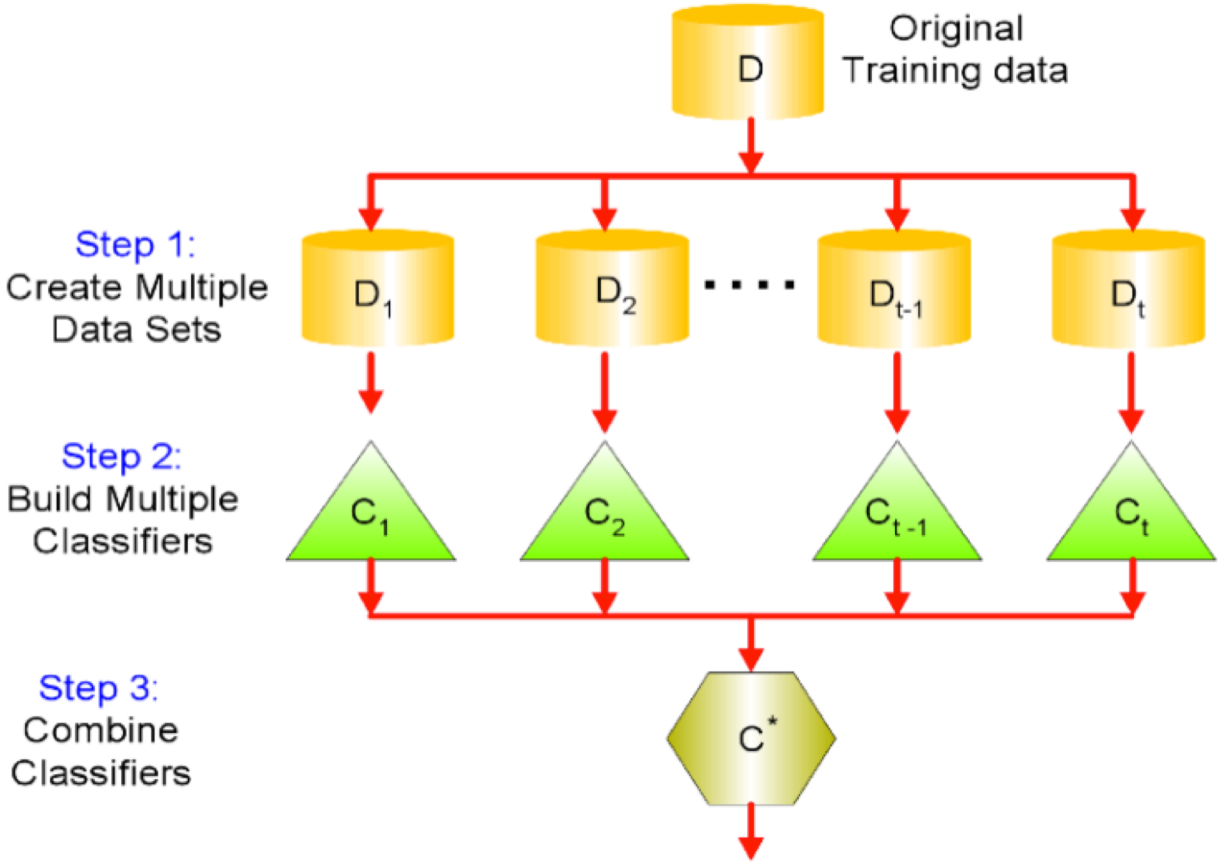
Remember: Averaging reduces variance

Bagging Algorithm

- Input: Dataset D containing N training examples.
- Steps:

1. Create k bootstrap samples D_1, D_2, …, D_k
2. Train distinct classifiers C_1, C_2, …, C_k on each of these datasets
3. Create an aggregate prediction by taking the majority / average of these individual classifiers.

Bagging Algorithm



Analysis

- In a any bootstrap sample of size N, a given data point has probability of not being selected as: $\left(1 - \frac{1}{N}\right)^N$
- So, probability of being selected as (p): $1 - \left(1 - \frac{1}{N}\right)^N$
- Think of it like a binomial distribution with p as being probability of heads.
- Expected number of points that are selected in any given bootstrap sample = Np

  $= N[1 - \left(1 - \frac{1}{N}\right)^N]$

Analysis

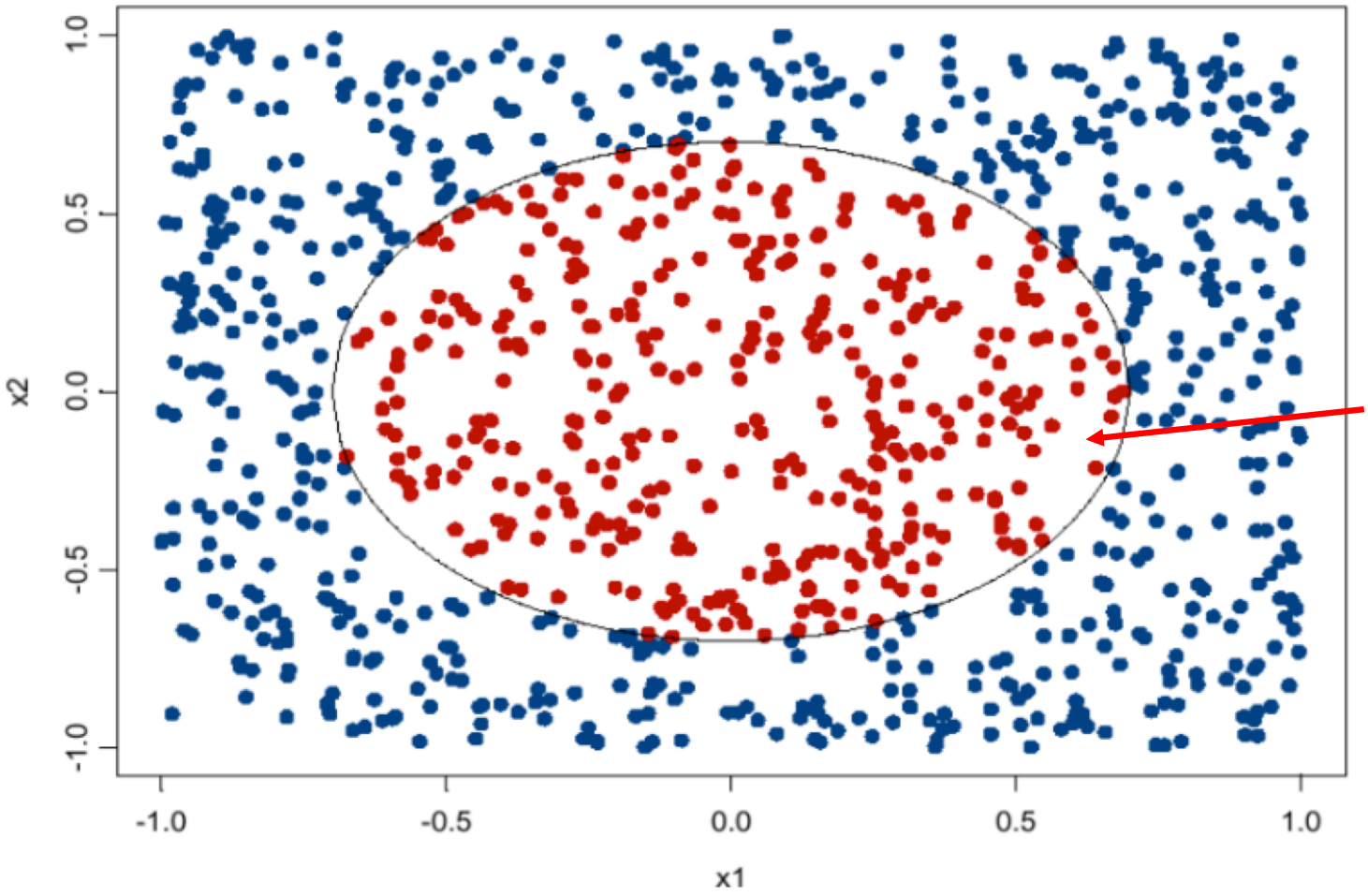
- Remember, that e^x can be written as:

$$\exp(x) = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n$$

- So, for large enough values of N, the expected number of points in a bootstrap sample becomes:

$$\begin{aligned} \mathrm{E} &= \mathrm{N}[1 - \left(1 - \tfrac{1}{N}\right)^N] \\ &= N[1 - \exp(-1)) \\ &= 0.632\ N \end{aligned}$$

Analysis



Original Sample

Bootstrap Sample

Analysis

- So, what does Bagging model do for me?
  Creates an aggregated model with less variance.
- What is the reduction?

  Ideally, $\mathrm{var}(\mathrm{Bagging}(\mathrm{C}(\mathrm{x}, \mathrm{D})) = \frac{\mathrm{var}(\mathrm{C}(\mathrm{x}, \mathrm{D}))}{\mathrm{N}}$

- But, in reality the models are not independent (since their datasets are correlated), so reduction is less than (1/N) times.
- Does Bagging also lead to an increase in accuracy?
  Yes, it has been empirically shown to

Bagging Results

Data Set	$\bar{e}_S$	$\bar{e}_B$	Decrease
waveform	29.1	19.3	34%
heart	4.9	2.8	43%
breast cancer	5.9	3.7	37%
ionosphere	11.2	7.9	29%
diabetes	25.3	23.9	6%
glass	30.4	23.6	22%
soybean	8.6	6.8	21%

Breiman "Bagging Predictors" Berkeley Statistics Department TR#421, 1994

When Will Bagging Improve Accuracy?

- Depends on the stability of the base-level classifiers.
- A learner is unstable if a small change to the training set *D* causes a large change in the output hypothesis φ.
  - If small changes in *D* causes large changes φ in then there will be an improvement in performance.
- Bagging helps unstable procedures, but could hurt the performance of stable procedures.
- Neural nets and decision trees are unstable.
- k-nn and naïve Bayes classifiers are stable.

Random Forests

- In Bagging classifiers, we created new datasets by varying the instances from the original dataset.
- What if we vary the attributes also?
- This is the idea behind Random Forests (RF). The base classifiers are Decision Trees.
- Idea: Create Decision Tree (DT) models from the bootstrap samples, but in each DT limit the splitting criteria to a random subsample of the attributes.

Random Forest Algorithm

Input:

Dataset D containing N instances and M attributes

Training Phase:

For b = 1 to B

Create a bootstrap sample of size N from original sample
Grow tree T_b using the bootstrap sample as follows:
Create a random sample of m attributes (m < M)
Use these attributes to construct a decision tree as usual

Test Phase:

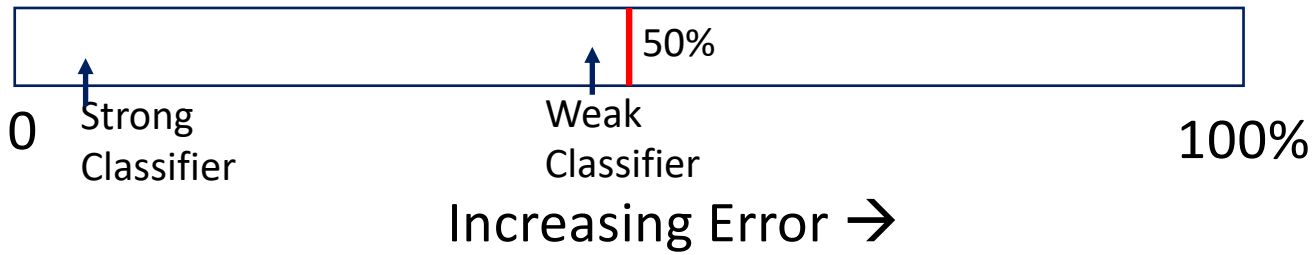
For a new data point X, take the aggregated prediction of models
For classification: aggregate = majority
For regression: aggregate = average

Best of both worlds

- Can we get lower variance and lower bias at the same time?
- We studied the bias-variance tradeoff.
- But, is there a way around it?
- Can we average models ***and*** reduce bias?
- Boosting

Boosting

- Boosting – combining a set of "weak" classifiers to produce a strong classifier.
- What is a "weak classifier"?
- In machine learning terminology, even a random predictor can predict with 50% error.
- If there exists a classifier that generates slightly better results than a random predictor, it's known as a weak classifier.
- Error of a weak classifier = (50 - γ) %, where γ is a small positive number.



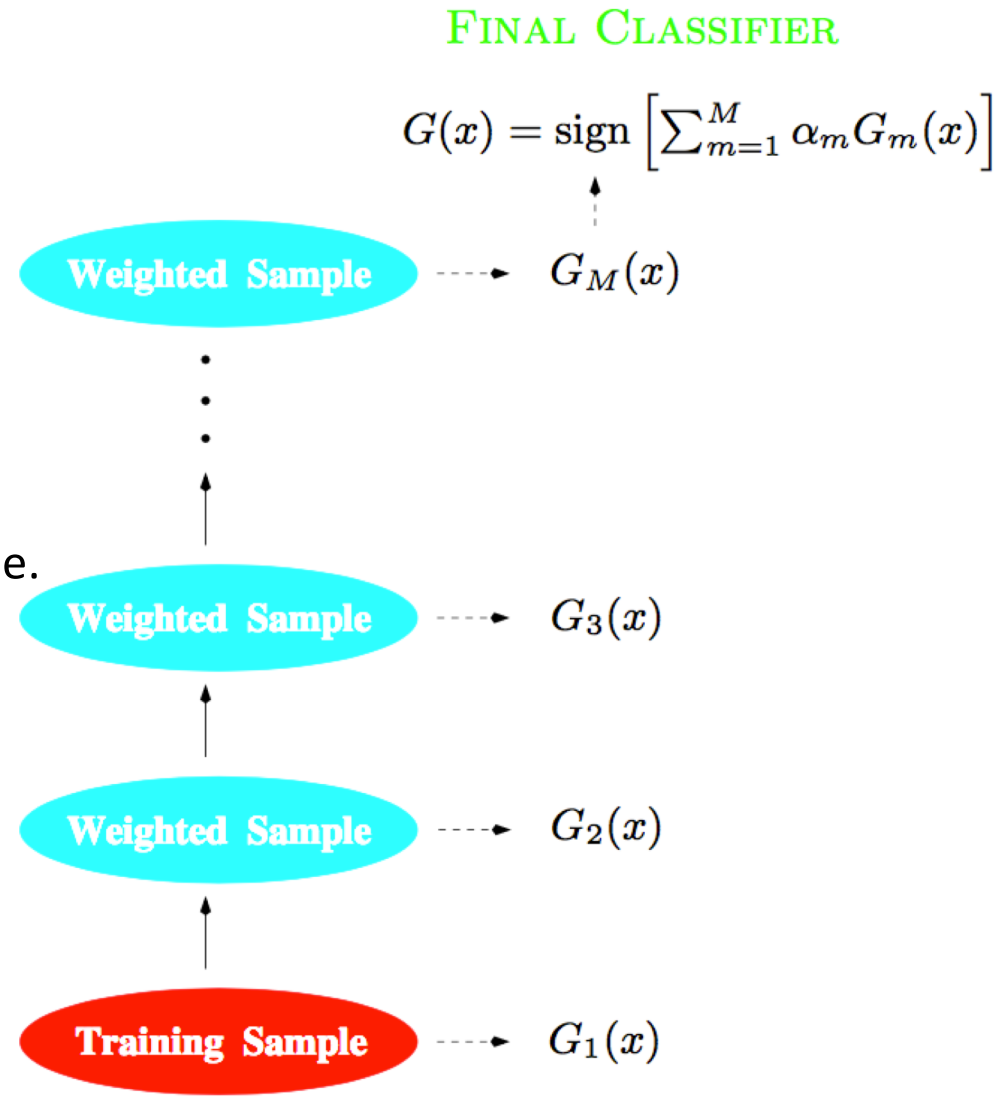
Boosting - Idea

- Start with equal weights given to all training instances.
- Run 1st rule of thumb (hypothesis G_1) => Keep track of the instances where G_1 makes an error. Increase the weight of those instances.
- Repeat this process for 2nd, 3rd, …, kth hypothesis.
- Create a final hypothesis by taking a weighted average of these k hypotheses.

$$G_{final} = \sum_{i=1}^{k} \alpha_i G_i$$

Boosting

Key points to note:
- Each classifier gets a differently weighted sample.
- The classifiers work in a serial way i.e. one after the other.
- The classifiers are not independent as the output of one is input for the other



Boosting

- You want to build a spam filter based on N training examples $(\boldsymbol{X}^i, y^i)$ where y^i is {Non-Spam (0), Spam(1)}, and i = 1 to N.
- Approach:

  Start with equal weight for each example: $w_i = \frac{1}{N}$

  The initial weight distribution $D_0 = \{w_0 = \frac{1}{N}, \dots, w_N = \frac{1}{N}\}$

  for i = 1 to T:

  - Apply a rule of thumb (hypothesis G_i) to the data with weight distribution D_i
  - Keep track of examples where the hypothesis makes an error and increase their weights to create a new distribution D_{i+1}

  end for

Adaboost

- One of the most popular implementations of Boosting was proposed by Freund & Schapire in 1996.
- Simple to implement and combines many weak classifier to produce a single strong hypothesis with excellent results.
- Algorithm presented on next slide.

Adaboost

- Given: Set of examples $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ where $x_i \in X\ (Domain)$, $y_i \in Y = \{-1, +1\}$
- Algorithm:
  Initialize $D_1(i) = 1/m$
  For t = 1 to T:
  - Train weak learner h_t using distribution D_i
  - Get weak hypothesis $h_t: X \to \{-1, +1\}$ with error:

$$\epsilon_t = \frac{\sum_{i=1}^{m} w_i\, I[h_t(x_i) \neq y_i]}{\sum_{i=1}^{m} w_i}$$

  - Choose $\alpha_t = \frac{1}{2} \ln(\frac{1-\epsilon_t}{\epsilon_t})$
  - Update weights as: $D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$

  where Z_t is a normalization factor so that D_{t+1} sums to 1.

I is the identity function:
$I(1) = 1$
$I(0) = 0$

Note:
For correct classification
$y_i h_t(x_i) > 0$
For incorrect classification
$y_i h_t(x_i) < 0$

Adaboost

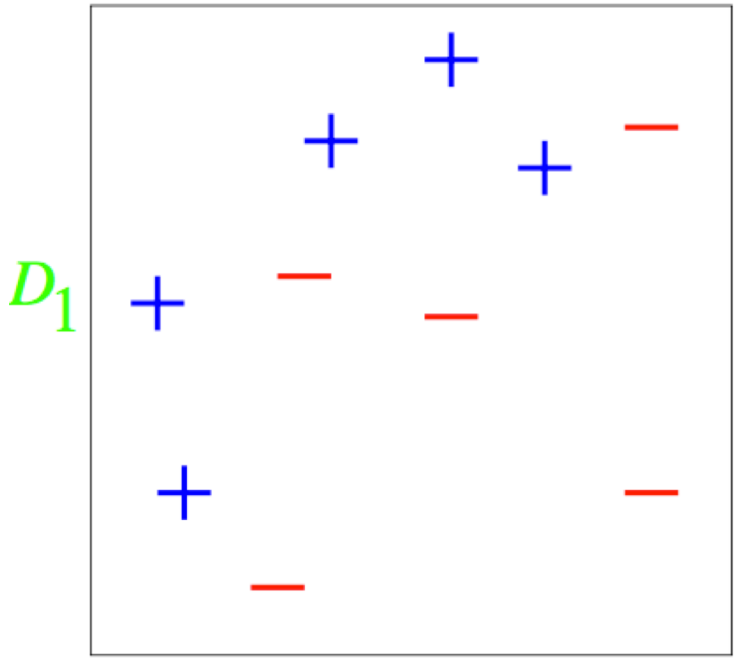
- The final hypothesis is obtained as:

$$H_{final}(x) = sign(\sum_{t} \alpha_t h_t(x))$$

Toy Example:

- Let's say we have the training data as shown on the right.
- There are 10 data points – 5 of each class.
- Initial weight distribution:

$D_1 = \{w_i = \frac{1}{10}\}$ for each i = 1 to 10.



Toy Example – Round 1

- Our first weak hypothesis is shown on right.
- 3 points are misclassified.

$$\epsilon_1 = \frac{3 \times 0.1}{1} = 0.30$$

$$\alpha_1 = \frac{1}{2} \ln\left(\frac{1 - 0.30}{0.30}\right) = 0.424$$

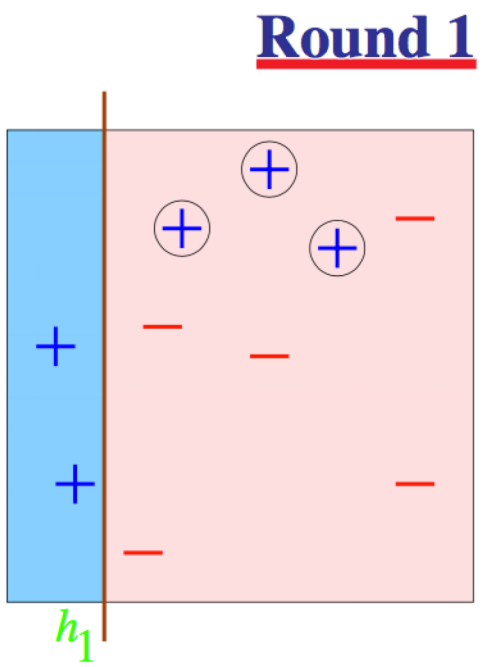
The weight of misclassified points:

$$D_2 = 0.1 \times \exp(0.424) = 0.153$$

The weight of correctly classified points:

$$D_2 = 0.1 \times \exp(-0.424) = 0.065$$

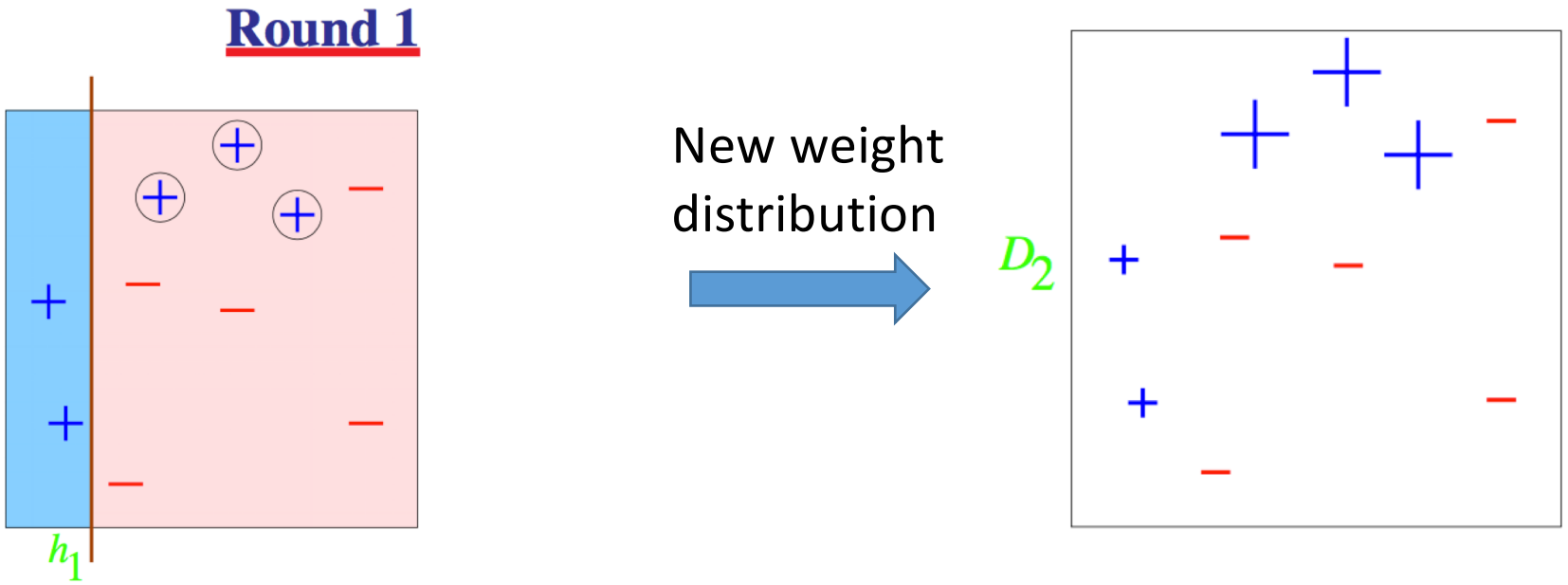
These are not normalized. After normalizing – dividing each value by total of weight of 0.914 -



Weight of misclassified:
0.167 each

Weight of correctly classified:
0.071 each

Toy Example:



- After round 1, we get a new distribution D2.
  Note that the 3 misclassified points get a higher weight

Toy Example – Round 2

- Our second weak hypothesis is shown on right.
- 3 points are misclassified.

$$\epsilon_2 = \frac{3 \times 0.071}{1} = 0.213$$

$$\alpha_2 = \frac{1}{2}\ln\left(\frac{1-0.213}{0.213}\right) = 0.653$$

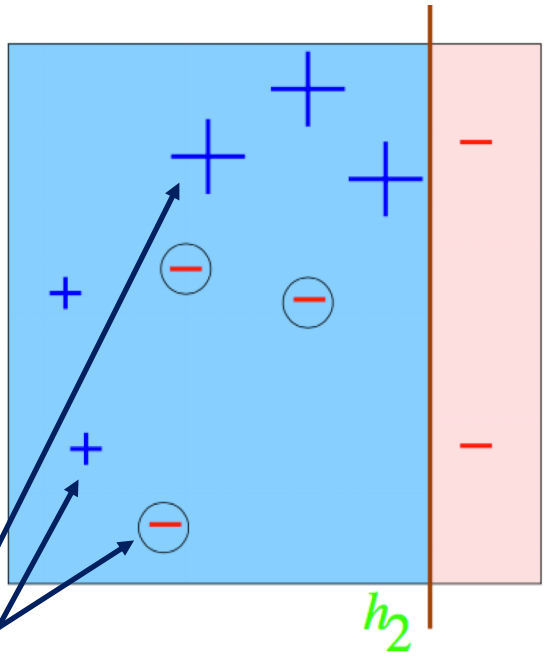
The weight of misclassified points:

$$D_2 = 0.071 \times \exp(0.653) = 0.136$$

The weight of correctly classified points:

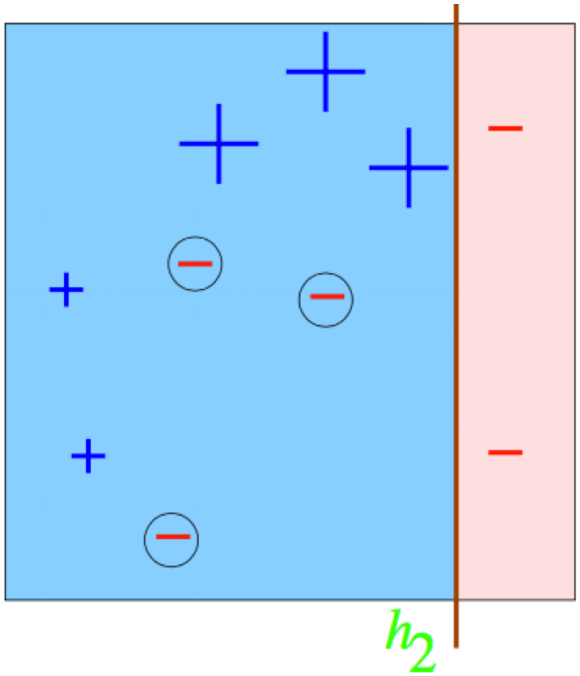
$$D_2 = 0.167 \times \exp(-0.653) = 0.087$$

$$D_2 = 0.071 \times \exp(-0.653) = 0.037$$

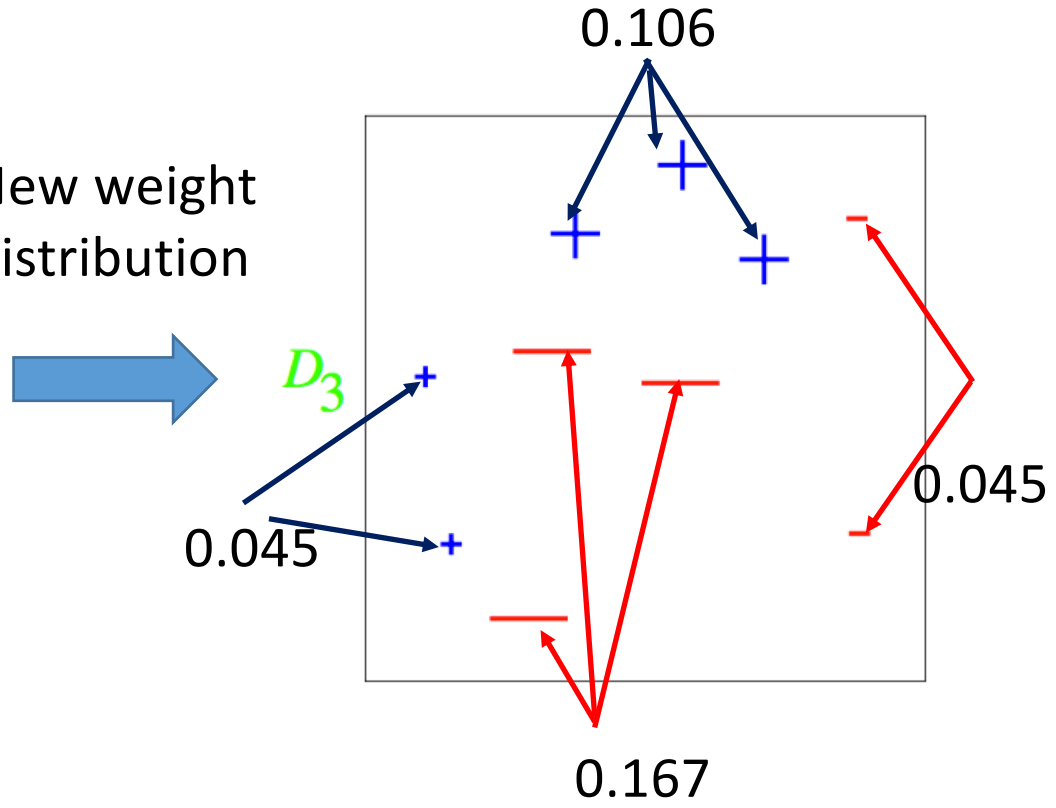


Total weight = 3 * 0.136 + 3 * 0.087 + 4 * 0.037 = 0.817

Toy Example



New weight distribution

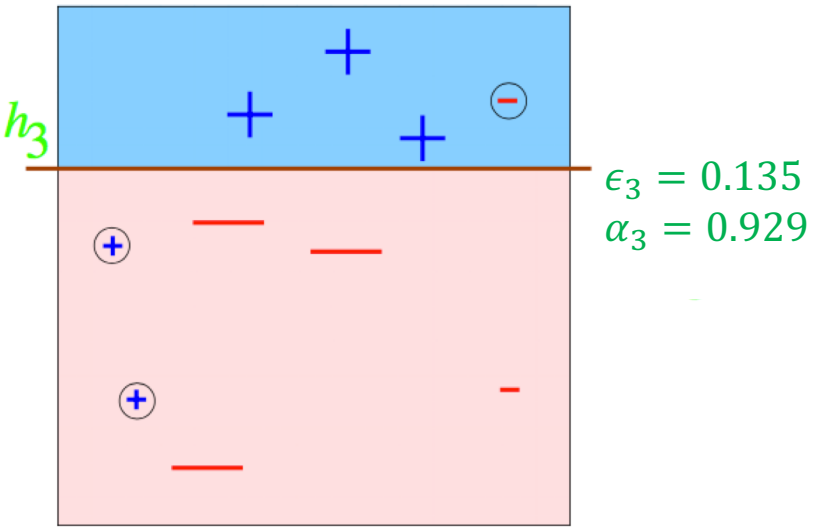


Toy Example – Round 3

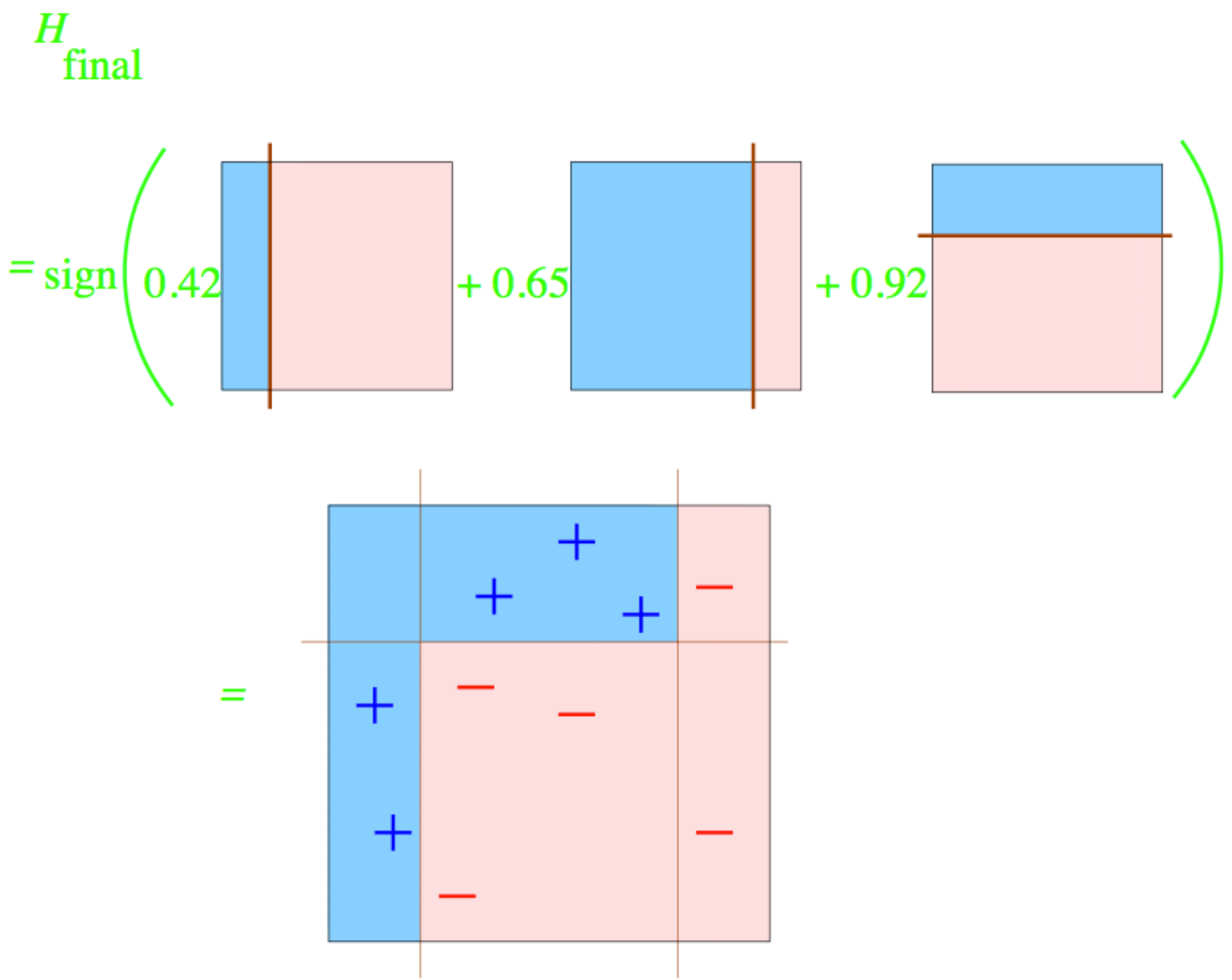
- Our third weak hypothesis is shown on right.
- 3 points are misclassified.

$$\epsilon_3 = \frac{3 \times 0.045}{1} = 0.135$$

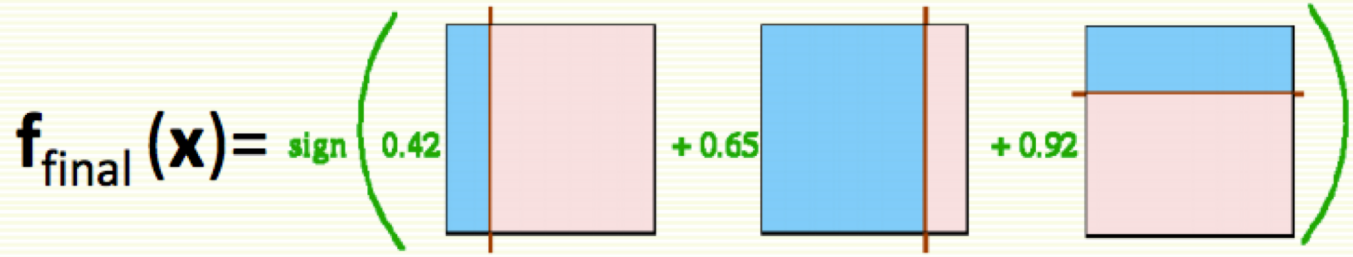
$$\alpha_3 = \frac{1}{2}\ln\left(\frac{1-0.135}{0.135}\right) = 0.929$$

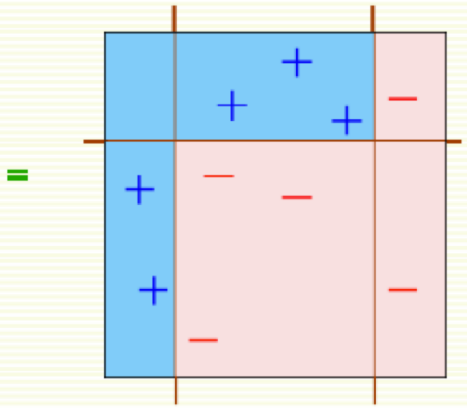


Final Hypothesis



How to use this hypothesis





$$f_{final}(x) = \text{sign}(0.42\,\text{sign}(3 - x_1) + 0.65\,\text{sign}(7 - x_1) + 0.92\,\text{sign}(x_2 - 4))$$

- note non-linear decision boundary

AdaBoost Analysis

- It is shown in the paper that the training error drops exponentially fast.

$$Error_{train} \leq \exp(-2\sum_{t}\gamma_t^2)$$

- Remember the definition of $\boldsymbol{\gamma}$: $\gamma_t = \epsilon_t - \frac{1}{2}$, where ϵ_t is the error at round t of the Boosting algorithm.
- Example: : Let errors for the first four rounds be, 0.3, 0.14, 0.06, 0.03, 0.01 respectively. Then

$$Error_{train} \leq \exp[-2(0.2^2 + 0.36^2 + 0.44^2 + 0.47^2 + 0.49^2)]$$

$$\approx 0.19$$

Another Toy Example

MIT Admissions Training Data

ID	Name	**Admit/Deny**	# of High School Detentions	SAT
1	Andrew	Deny	3	2050
2	Burt	Admit	1	2200
3	Charlie	Admit	2	2090
4	Derek	Deny	4	2230
5	Erica	Admit	5	2330
6	Faye	Deny	6	2220
7	Greg	Admit	6	2390
8	Helga	Admit	7	2320
9	Ivana	Deny	8	2330
10	Jan	Deny	8	2090

Another Toy Example

- Propose two simple hypothesis:
  h1: If # of High School Detentions < 3, then ADMIT
  h2: If SAT > 2200, then ADMIT
- Can you combine these two weak hypothesis to produce a strong hypothesis?

Summary

- Ensemble methods combine individual classifiers to achieve better results.
- Bagging doesn’t work so well with stable models. Boosting might still help.
- Boosting might hurt performance on noisy datasets. Bagging doesn’t have this problem.
- On average, boosting helps more than bagging, but it is also more common for boosting to hurt performance.
- Bagging is easier to parallelize.