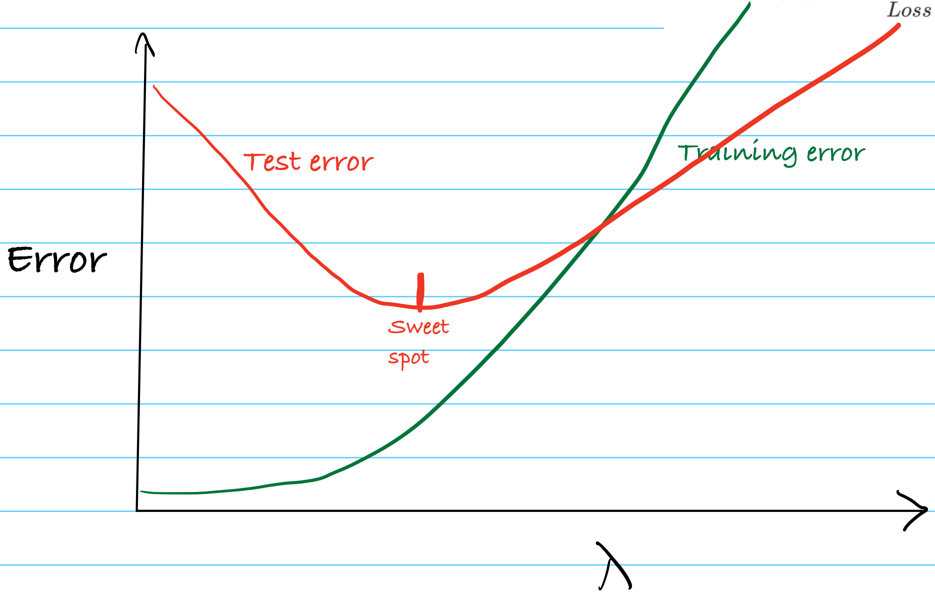
Quiz:

1. Which of the following is a better proxy for generalization error?
   A. Average training loss
   B. Average validation set loss
2. When n is large and hypothesis set size k is small, training error is close to generalization error with high probability (of $1-\delta$), what is this probability over?

How to choose parameters?

Eg: $$\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell(h_{\mathbf{w}}(\mathbf{x}_i), y_i)}_{Loss} + \underbrace{\overset{\lambda}{\lambda} r(w)}_{Regularizer}$$



Idea: Use validation set to pick parameters
(pick one with smallest validation error)

Validation Set

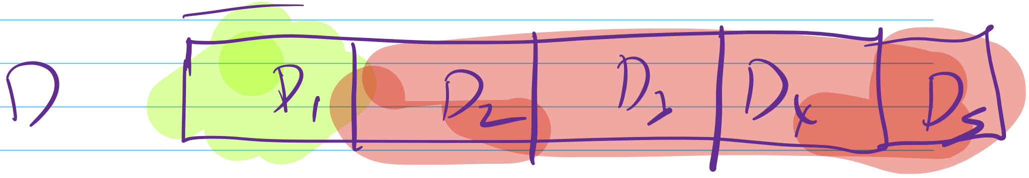
Training	Validation

1. If validation set is too small we dont get good estimate of error
2. But if it is very big then training set size is small

K- fold cross validation

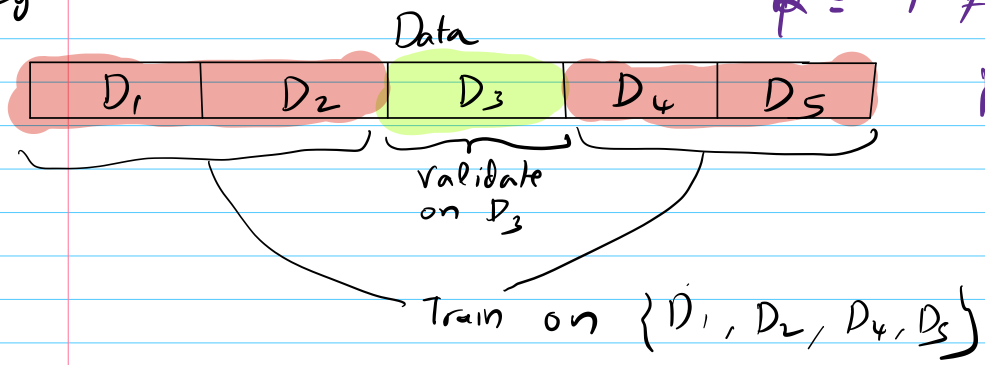
$D_{-k} = D \setminus D_k$

```
Partition data into K folds $D_1, \dots, D_K$
For $\lambda \in \{0.01, 0.02, \dots, 1\}$
    For $k = 1$ to $K$:
        $h_k = \text{Algorithm}(D_{-k}, \lambda)$   % train on all data But $D_k$
        $\varepsilon_{k,\lambda} = \frac{1}{|D_k|} \sum_{(x,y) \in D_k} \ell(h_w(x), y)$   % evaluate on $D_k$
    $\varepsilon_\lambda = \frac{1}{K} \sum_{k=1}^{K} \varepsilon_{k,\lambda}$   % Take average validation error
$\lambda^* = \underset{\lambda}{\text{argmin}}\, \varepsilon_\lambda$   % Pick best $\lambda$
```



$k = 1$ to 5

Eg



$k = 2$

$k = 3$

Search for parameters

1. Grid search 2. Random search
3. Zooming in

Bias Variance Decomposition

Consider the regression problem where given x we want to predict a real valued outcome y. $\ell(h(x), y) = (h(x) - y)^2$

Eg $\quad x = (\text{\#bedrooms, sq footage of house, \# bathrooms}, \ldots)$

$\quad y = \text{house price}$

$$\text{Test error}(h) = \mathbb{E}_{(x,y)\sim p}\left[(h(x) - y)^2\right]$$

Bayes Optimal Predictor = hypothesis with smallest possible test loss

$$\bar{y}(x) = E[y \mid x]$$

is the Bayes optimal predictor

$$\mathbb{E}_{(x,y)\sim p}\left[(\bar{y}(x) - y)^2\right] = \text{inherent (unavoidable) noise}$$

No method with any amount of data can beat the above test loss

Say we have an Algorithm that takes as input dataset D and outputs hypothesis h_D. But $\boxed{h_D}$ depends on sample D which is a random draw. We are interested in understanding the expected test error

$$\boxed{\text{Expected Test Error} = \mathbb{E}_{D\sim p^n}\left[\mathbb{E}_{(x,y)\sim p}\left[(h_D(x) - y)^2\right]\right]}$$

Since h_D is a random let us consider its expected behavior:

$$\bar{h} := \mathbb{E}_{n\sim p^n}[h_D] \qquad \text{ie} \qquad \bar{h}(x) = \mathbb{E}_{D\sim p^n}[h_D(x)]$$

Bias : $\mathbb{E}_{x\sim P}\left[\left(\bar{h}(x) - \bar{y}(x)\right)^2 \right]$

Expected sq. distance between expected model of our Algorithm and the best possible model

$$\mathbb{E}_{D}\, h_D = \bar{h}$$

Variance : $\mathbb{E}_{D\sim P^n}\left[\mathbb{E}_{x\sim P}\left[\left(\bar{h}(x) - h_D(x)\right)^2\right]\right]$

Fluctuation of Algorithm's random model around its mean

We will see that:

Expected test error = Bias + Variance + inherent noise

$$\text{Expected Test Error} = \mathbb{E}_{D\sim P^n}\left[\mathbb{E}_{(x,y)\sim P}\left[\left(h_D(x) - y\right)^2\right]\right]$$

$$= \mathbb{E}_{D\sim P^n}\left[\mathbb{E}_{(x,y)}\left[\left(\underbrace{h_D(x) - \bar{y}(x)}_{A} + \underbrace{\bar{y}(x) - y}_{B}\right)^2\right]\right]$$

$$= \mathbb{E}_{D\sim P^n}\left[\mathbb{E}_{(x,y)}\left[\underbrace{\left(h_D(x) - \bar{y}(x)\right)^2}_{A^2} + \underbrace{2\left(h_D(x) - \bar{y}(x)\right)\left(\bar{y}(x) - y\right)}_{2AB} + \underbrace{\left(\bar{y}(x) - y\right)^2}_{B^2}\right]\right]$$

$$= \mathbb{E}_{D\sim P^n}\,\mathbb{E}_{x,y}\left[\left(h_D(x) - \bar{y}(x)\right)^2\right] + \underbrace{2\,\mathbb{E}_{D\sim P^n}\,\mathbb{E}_{(x,y)}\left[\left(h_D(x) - \bar{y}(x)\right)\left(\bar{y}(x) - y\right)\right]}_{\overset{\|}{0}\ ??}$$

$$+ \mathbb{E}_{(x,y)}\left[\left(\bar{y}(x) - y\right)^2\right]$$

$$= \mathbb{E}_{D\sim P^n}\,\mathbb{E}_{(x,y)}\left[\left(h_D(x) - \bar{y}(x)\right)^2\right] + \mathbb{E}\left[\left(\bar{y}(x) - y\right)^2\right]$$

"inherent noise"

$$= \mathbb{E}_{D\sim P^n}\,\mathbb{E}_{(x,y)}\left[\left(\underbrace{h_D(x) - \bar{h}(x)}_{A} + \underbrace{\bar{h}(x) - \bar{y}(x)}_{B}\right)^2\right] + \mathbb{E}\left[\left(\bar{y}(x) - y\right)^2\right]$$

"inherent noise"

$$= \underset{D}{E}\,\underset{(x,y)}{E}\Big[\underbrace{(h_D(x) - \bar{h}(x))^2}_{A^2} + \underbrace{2(h_D(x) - \bar{h}(x))(\bar{h}(x) - \bar{y}(x))}_{2AB} + \underbrace{(\bar{h}(x) - \bar{y}(x))^2}_{B^2}\Big]$$

$$+ \; E[(\bar{y}(x) - y)^2]$$

"inherent noise"

$$= \underbrace{\underset{D}{E}\,\underset{x}{E}[(h_D(x) - \bar{h}(x))^2]}_{\text{"Variance"}} + \underbrace{2\,\underset{D}{E}\,\underset{x}{E}[(h_D(x) - \bar{h}(x))(\bar{h}(x) - \bar{y}(x))]}_{=0\,??}$$

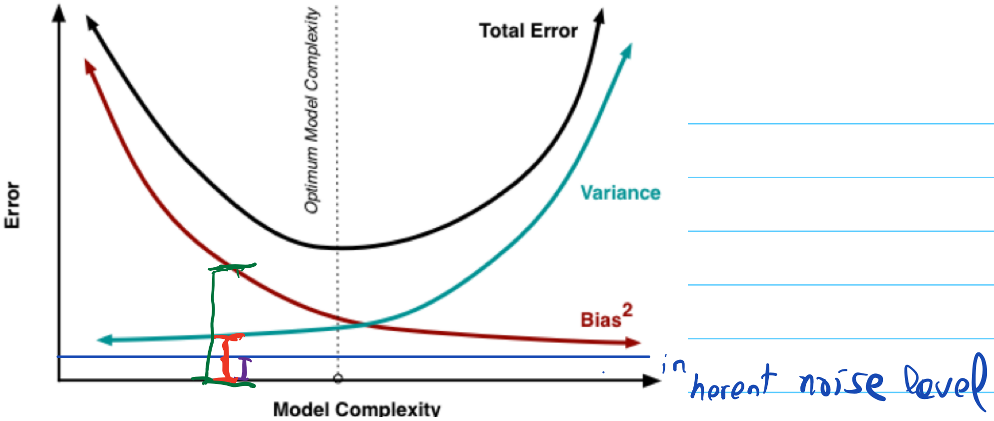
$$+ \underbrace{\underset{x}{E}[(\bar{h}(x) - \bar{y}(x))^2]}_{\text{"Bias"}} + \underbrace{E[(\bar{y}(x) - y)^2]}_{\text{"inherent noise"}}$$

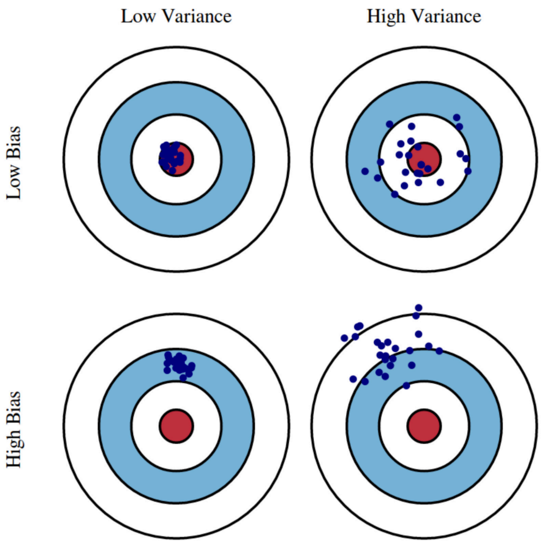
$$= \underbrace{\underset{D}{E}\,\underset{x}{E}[(h_D(x) - \bar{h}(x))^2]}_{\text{"Variance"}} + \underbrace{\underset{x}{E}[(\bar{h}(x) - \bar{y}(x))^2]}_{\text{"Bias"}} + \underbrace{E[(\bar{y}(x) - y)^2]}_{\text{"inherent noise"}}$$

Variance: Captures how much your classifier changes if you train on a different training set. How "over-specialized" is your classifier to a particular training set (overfitting)? If we have the best possible model for our training data, how far off are we from the average classifier?

Bias: What is the inherent error that you obtain from your classifier even with infinite training data? This is due to your classifier being "biased" to a particular kind of solution (e.g. linear classifier). In other words, bias is inherent to your model.

Noise: How big is the data-intrinsic noise? This error measures ambiguity due to your data distribution and feature representation. You can never beat this, it is an aspect of the data.





> Fig 1: Graphical illustration of bias and variance. (Source http://scott.fortmann-roe.com/docs/BiasVariance.html) Fig 2: The variation of Bias and Variance with the model complexity. This is similar to the concept of overfitting and underfitting. More complex models overfit while the simplest models underfit. (Source http://scott.fortmann-roe.com/docs/BiasVariance.html)

Detecting High Bias and High Variance

If a classifier is under-performing (e.g. if the test or training error is too high), there are several ways to improve performance. To find out which of these many techniques is the right one for the situation, the first step is to determine the root of the problem. The graph above plots the training error and the test error and can be divided into two overarching regimes. In the first regime (on the left side of the graph), training error is below the desired error threshold (denoted by ϵ), but test error is significantly higher. In the second regime (on the right side of the graph), test error is remarkably close to training error, but both are above the desired tolerance of ϵ.

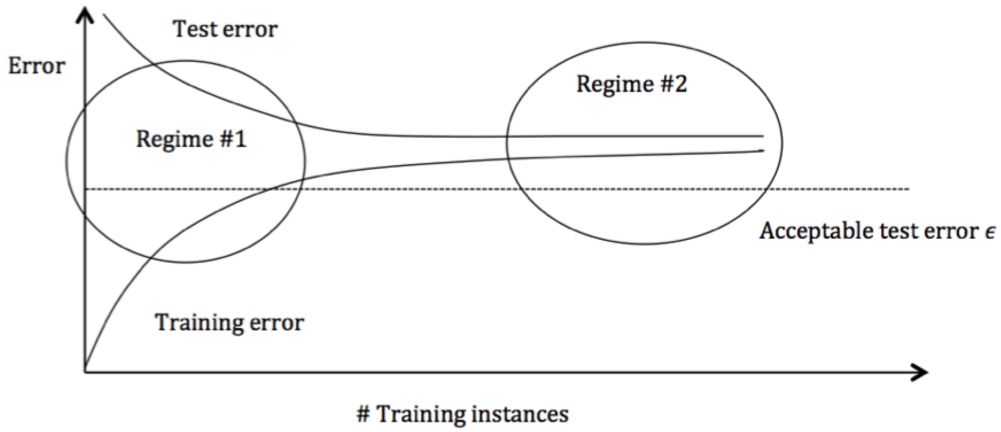


Figure 3: Test and training error as the number of training instances increases.

Regime 1 (High Variance)

In the first regime, the cause of the poor performance is high variance.

Symptoms:

1. Training error is much lower than test error
2. Training error is lower than ϵ
3. Test error is above ϵ

Remedies:

- Add more training data
- Reduce model complexity -- complex models are prone to high variance
- Bagging (will be covered later in the course)

Regime 2 (High Bias)

Unlike the first regime, the second regime indicates high bias: the model being used is not robust enough to produce an accurate prediction.

Symptoms:

1. Training error is higher than ϵ

Remedies:

- Use more complex model (e.g. kernelize, use non-linear models)
- Add features