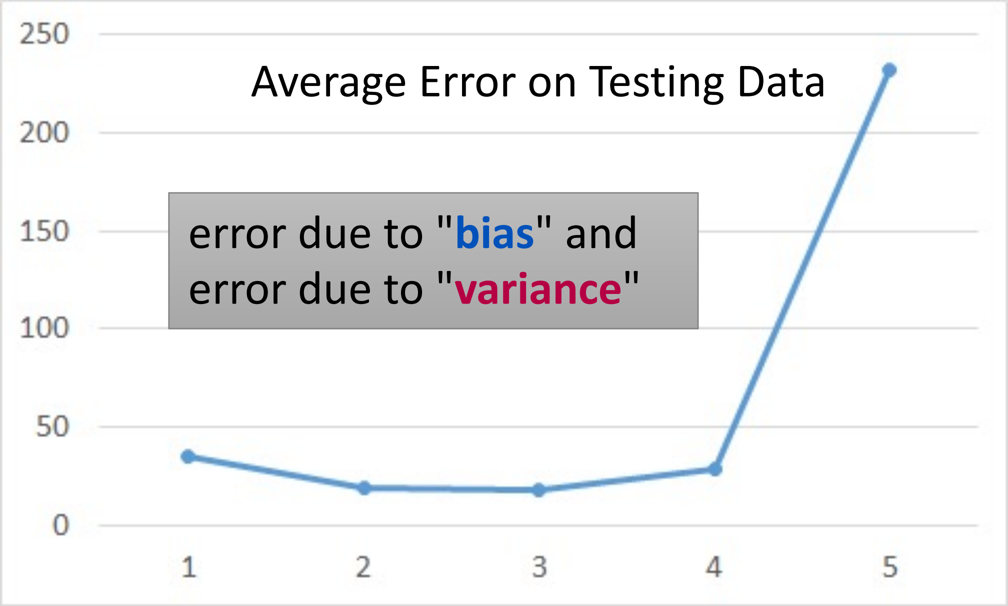
Bias and Variance

Part I

Review



A more complex model does not always lead to better performance on **testing data**.

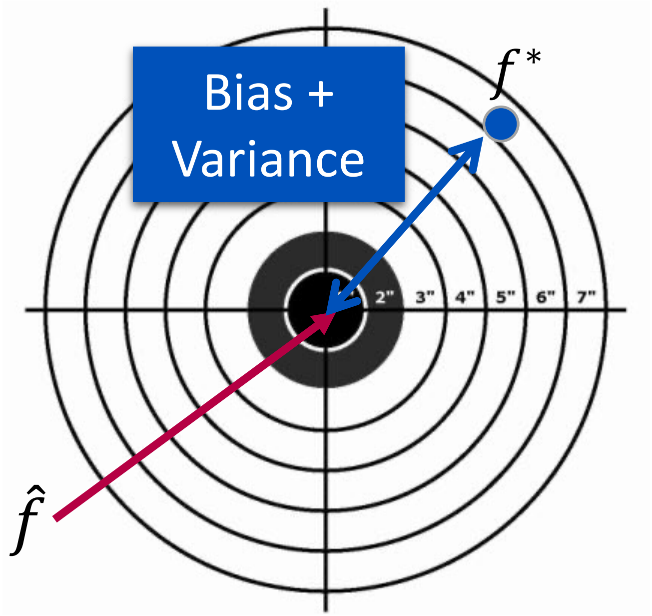
Estimator

$\hat{y} = \hat{f}($  $)$

Only Niantic knows $\hat{f}$

From training data, we find f^*

f^* is an estimator of $\hat{f}$



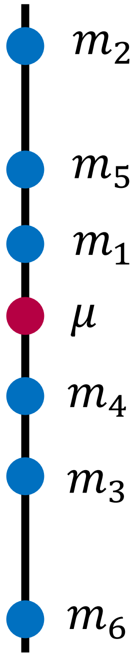
Bias and Variance of Estimator: Part I

- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of mean μ
  - Sample N points: $\{x^1, x^2, \dots, x^N\}$

$$m = \frac{1}{N}\sum_n x^n \neq \mu$$

$$E[m] = E\left[\frac{1}{N}\sum_n x^n\right] = \frac{1}{N}\sum_n E[x^n] = \mu$$

unbiased



Bias and Variance of Estimator: Part II

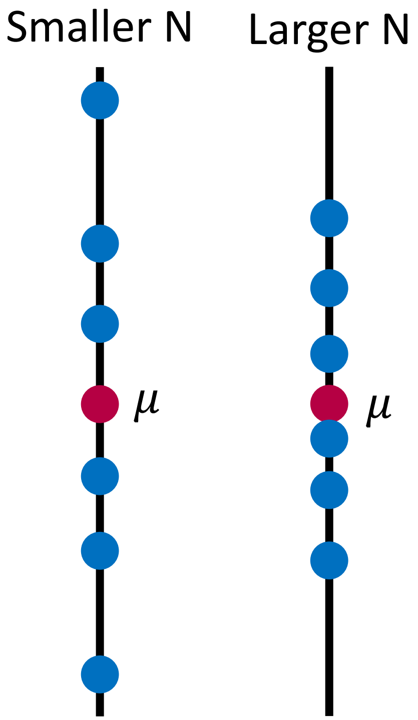
- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of mean μ
  - Sample N points: $\{x^1, x^2, \dots, x^N\}$

$$m = \frac{1}{N}\sum_n x^n \neq \mu$$

$$Var[m] = \frac{\sigma^2}{N}$$

Variance depends on the number of samples

unbiased



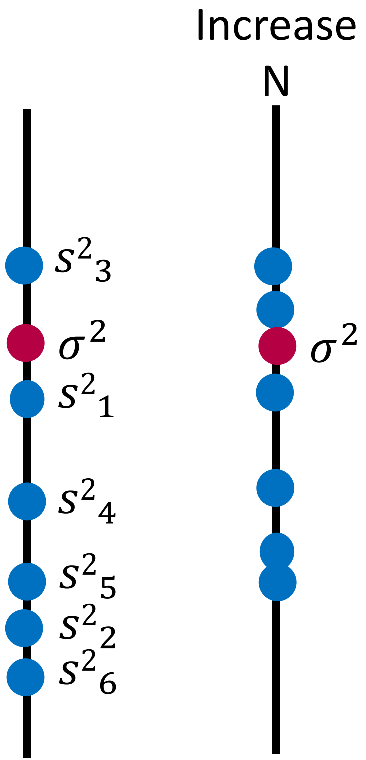
Bias and Variance of Estimator: Part III

- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of variance σ^2
  - Sample N points: $\{x^1, x^2, \ldots, x^N\}$

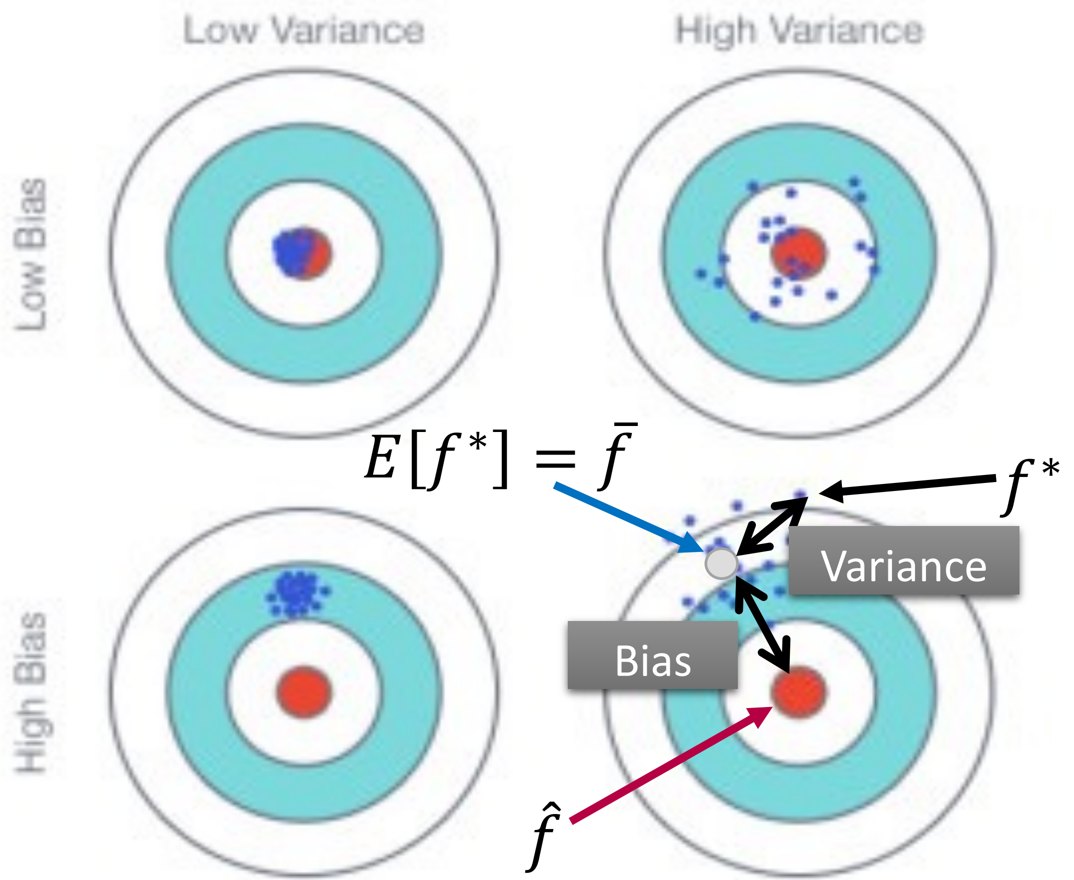
$$m = \frac{1}{N}\sum_n x^n \quad s^2 = \frac{1}{N}\sum_n (x^n - m)^2$$

Biased estimator

$$E[s^2] = \frac{N-1}{N}\sigma^2 \neq \sigma^2$$



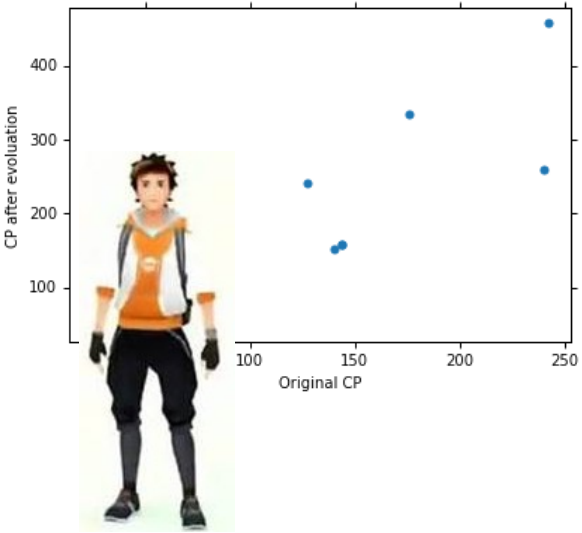
Bias and Variance: Part IV



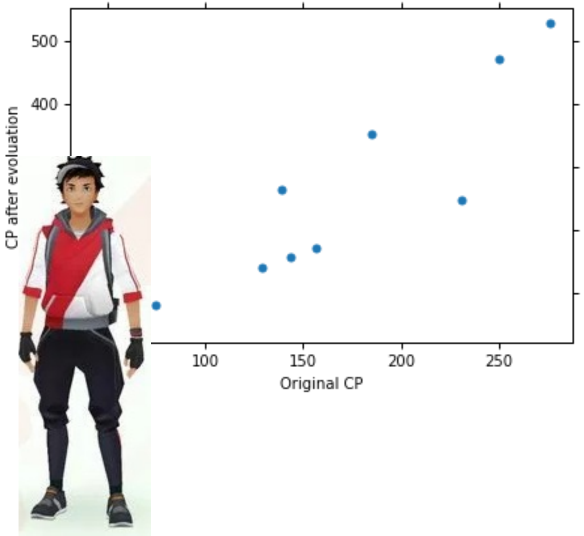
Parallel Universes

In all the universes, we are collecting (catching) 10 Pokémon as training data to find f^*

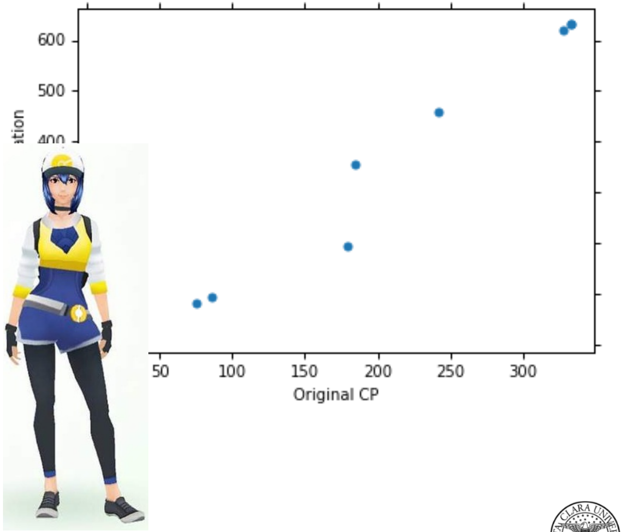
Universe 1



Universe 2



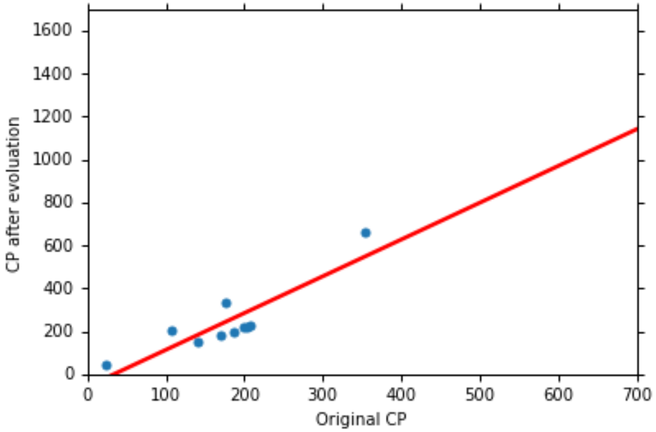
Universe 3



Parallel Universes, Cont’d

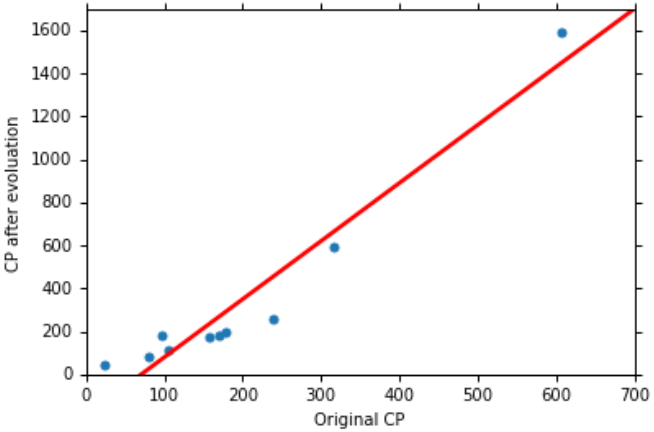
In different universes, we use the same model, but obtain different f^*

Universe 123



$y = b + w \cdot x_{cp}$

Universe 345

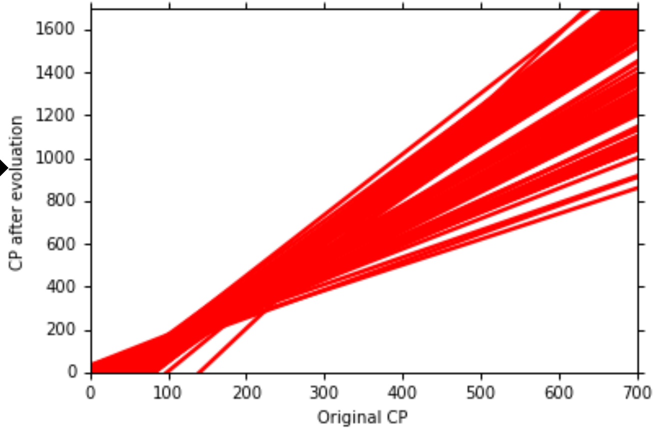


$y = b + w \cdot x_{cp}$

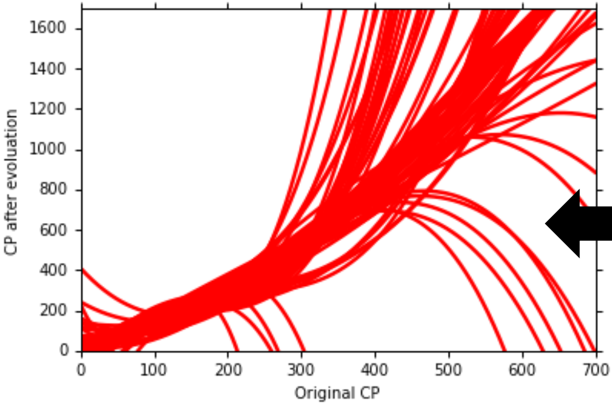
Variance

f^* in 100 Universes

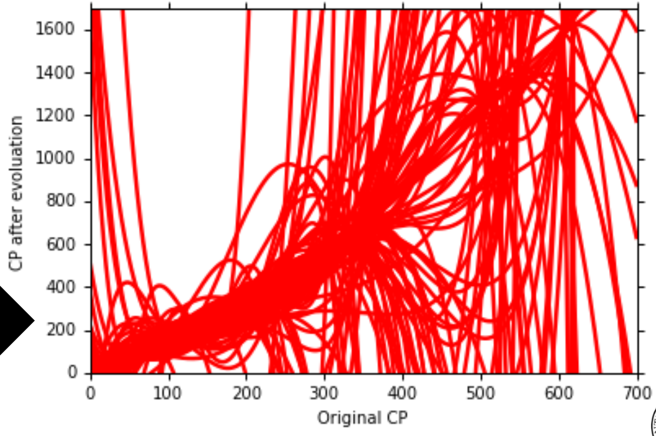
$y = b + w \cdot x_{cp}$



$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3$

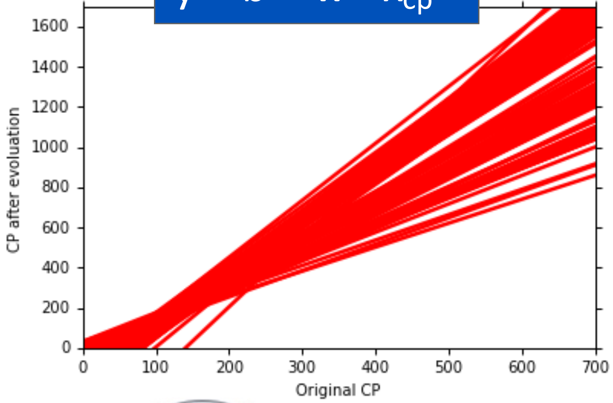


$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$



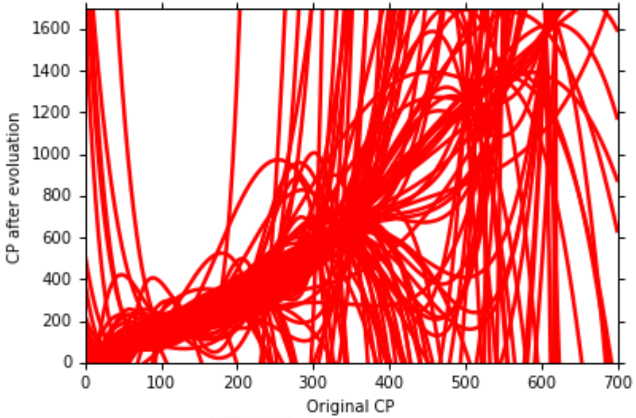
Variance, Cont'd

$$y = b + w \cdot x_{cp}$$



Small Variance

$$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$$



Large Variance

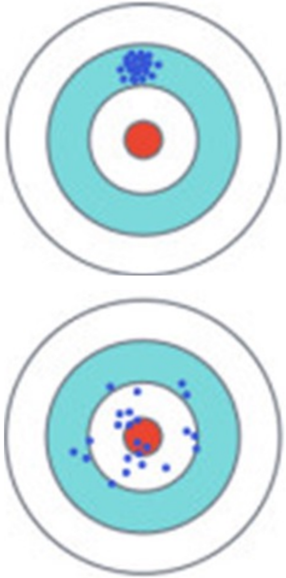
Simpler model is less influenced by the sampled data

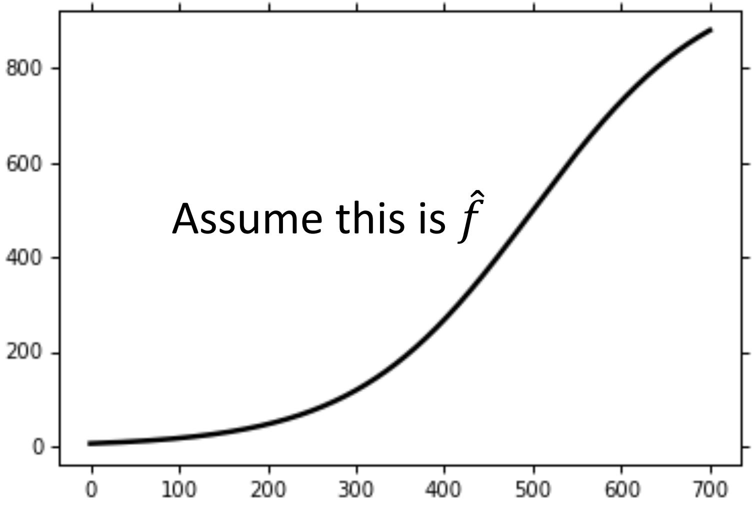
Consider the extreme case $f(x) = 5$

Bias: Part I

$$E[f^*] = \bar{f}$$

- Bias: If we average all the f^*, is it close to $\hat{f}$?





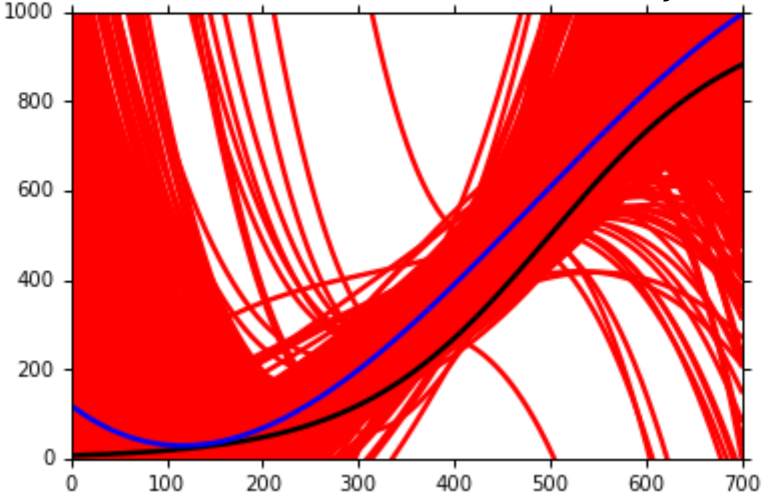
Bias: Part II

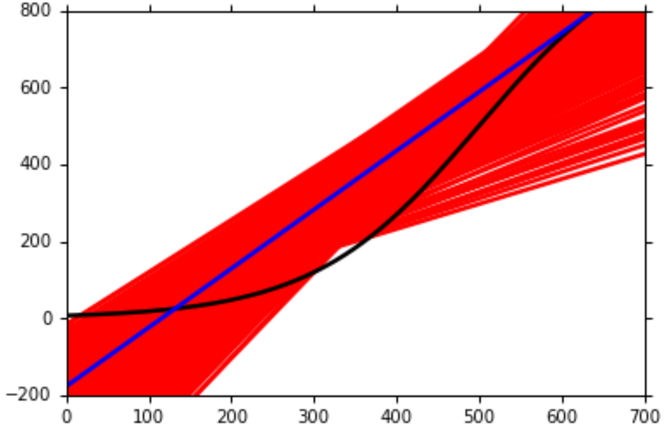
Black curve: the true function $\hat{f}$

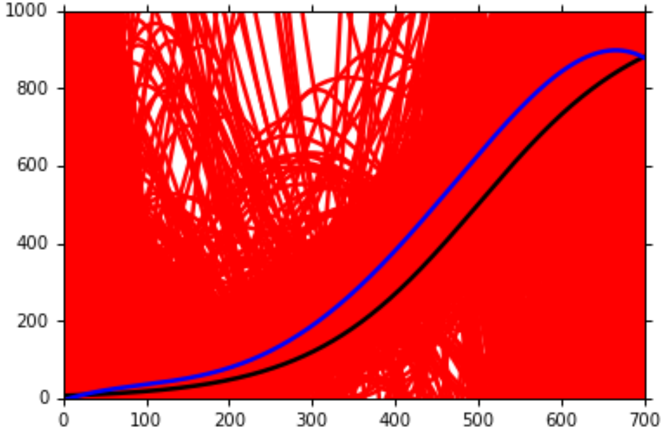
Red curves: 5000 f^*

Blue curve: the average of 5000 f^*

$= \bar{f}$



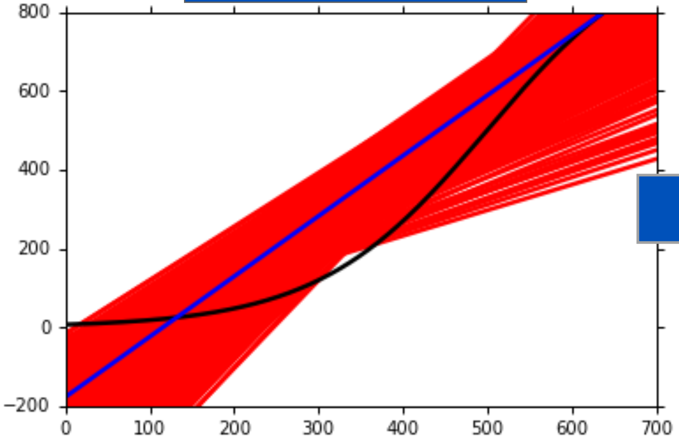


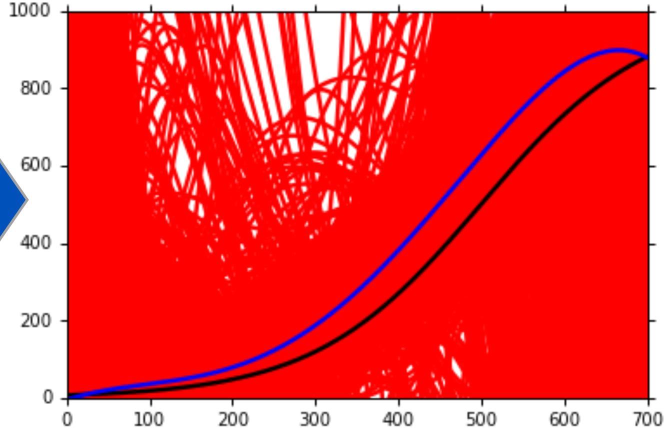


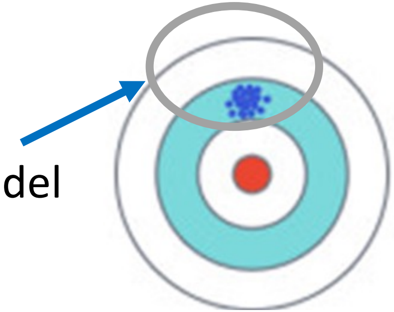
Bias: Part III

$y = b + w \cdot x_{cp}$

$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$

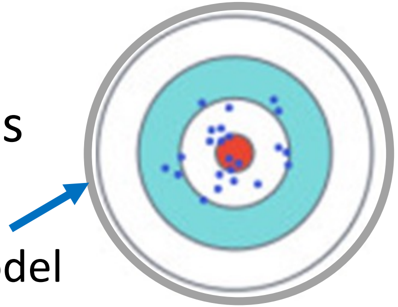






model

Large Bias



model

Small Bias

Bias Versus Variance