Bias-Variance

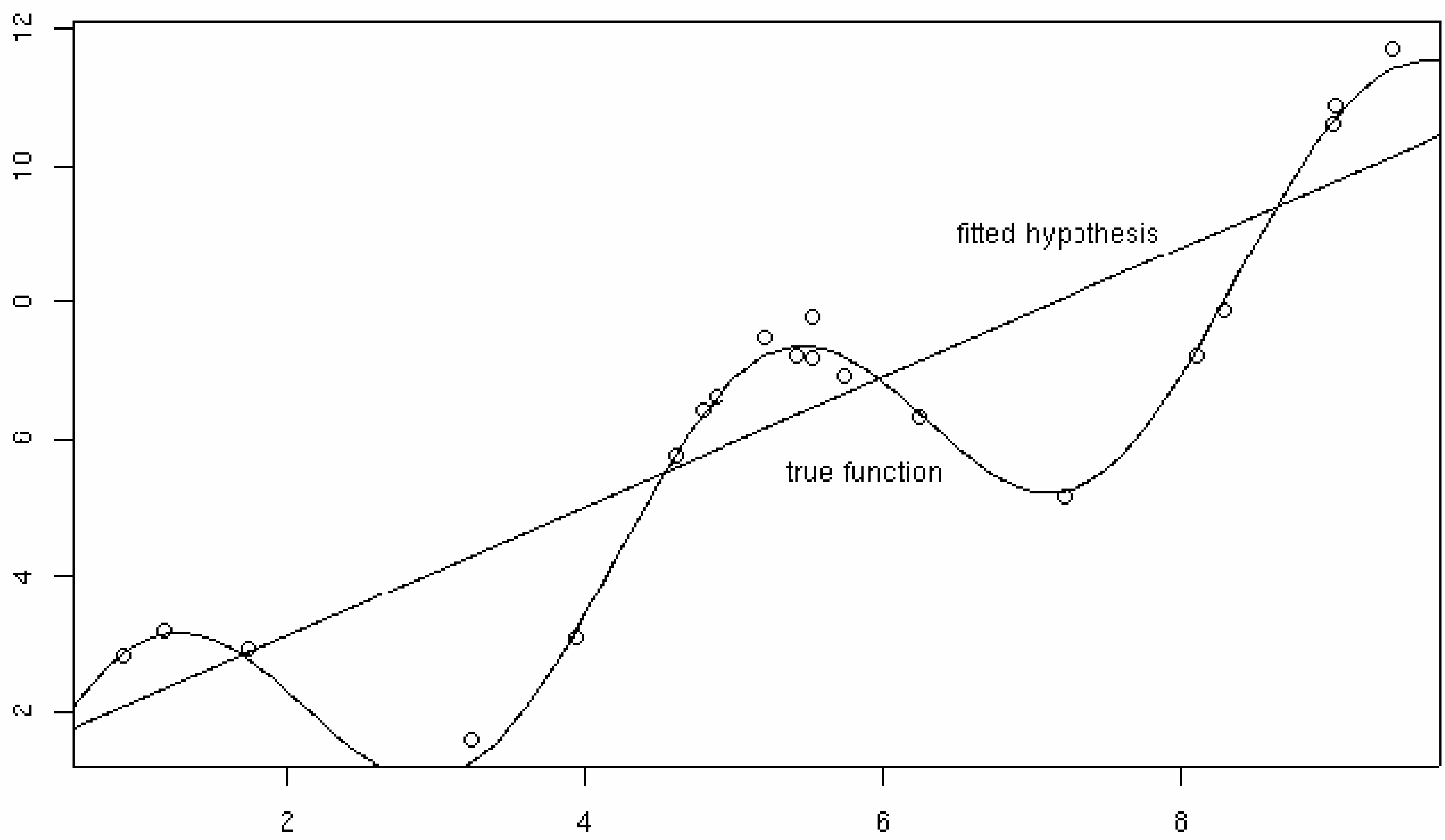
Slides by Tom Dietterich

Bias-Variance in Regression

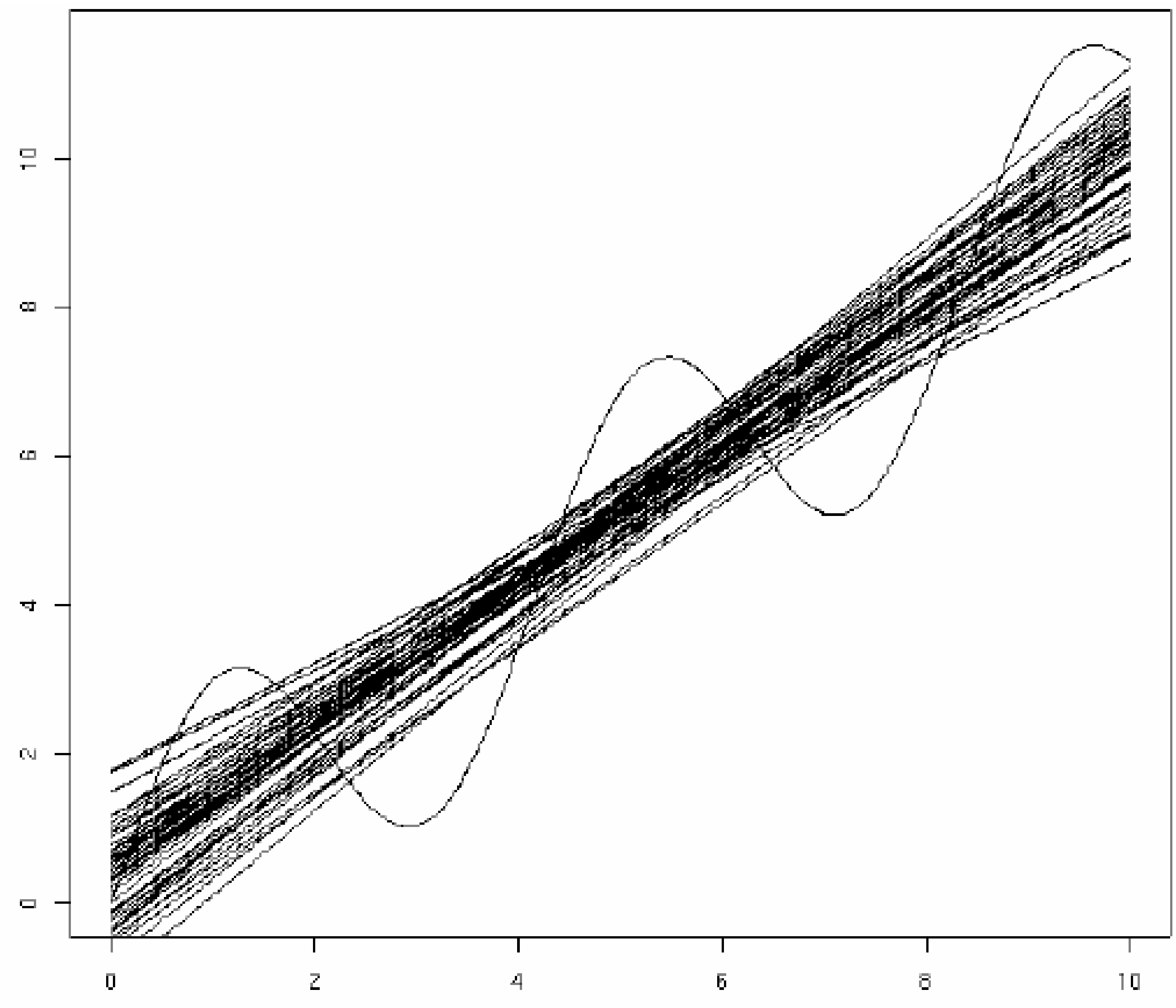
- True function is $y = f(x) + \varepsilon$, where ε is normally distributed with zero mean and standard deviation σ.
- Given a set of training examples, {(xi, yi)}, we fit an hypothesis $h(x) = w \cdot x + b$ to the data to minimize the squared error

$$\Sigma_i \, [y_i - h(x_i)]^2$$

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



50 fits (of 20 examples each)



Bias-Variance Analysis

- Now, given a new data point x^* (with observed value $y^* = f(x^*) + \varepsilon$), we would like to understand the expected prediction error

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

Statistical Analysis

- Imagine that our particular training sample S is drawn from some population of possible training samples according to P(S).
- Compute $E_P [(y^* - h(x^*))^2]$
- Decompose this into "bias", "variance", and "noise"

A Side Note

Let Z be a random variable with probability distribution P(Z)

Let $\underline{Z}$ = E_P[Z] be the average value of Z.

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

$E[\, (Z - \underline{Z})^2 \,] = E[\, Z^2 - 2\, Z\, \underline{Z} + \underline{Z}^2 \,]$

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

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

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

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

Bias Variance Noise

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

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

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

$- 2\ \underline{h(x^*)}\ f(x^*)$

$+ E[\ (y^* - f(x^*))^2\ ] + f(x^*)^2$ (lemma)

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

$(\underline{h(x^*)} - f(x^*))^2 +$ [bias^2]

$E[\ (y^* - f(x^*))^2\ ]$ [noise]

Error = Variance + Bias^2 + Noise^2

Bias Variance and Noise

- Variance: $E[\ (h(x^*) - \underline{h(x^*)})^2\ ]$

Describes how much $h(x^*)$ varies from one training set S to another

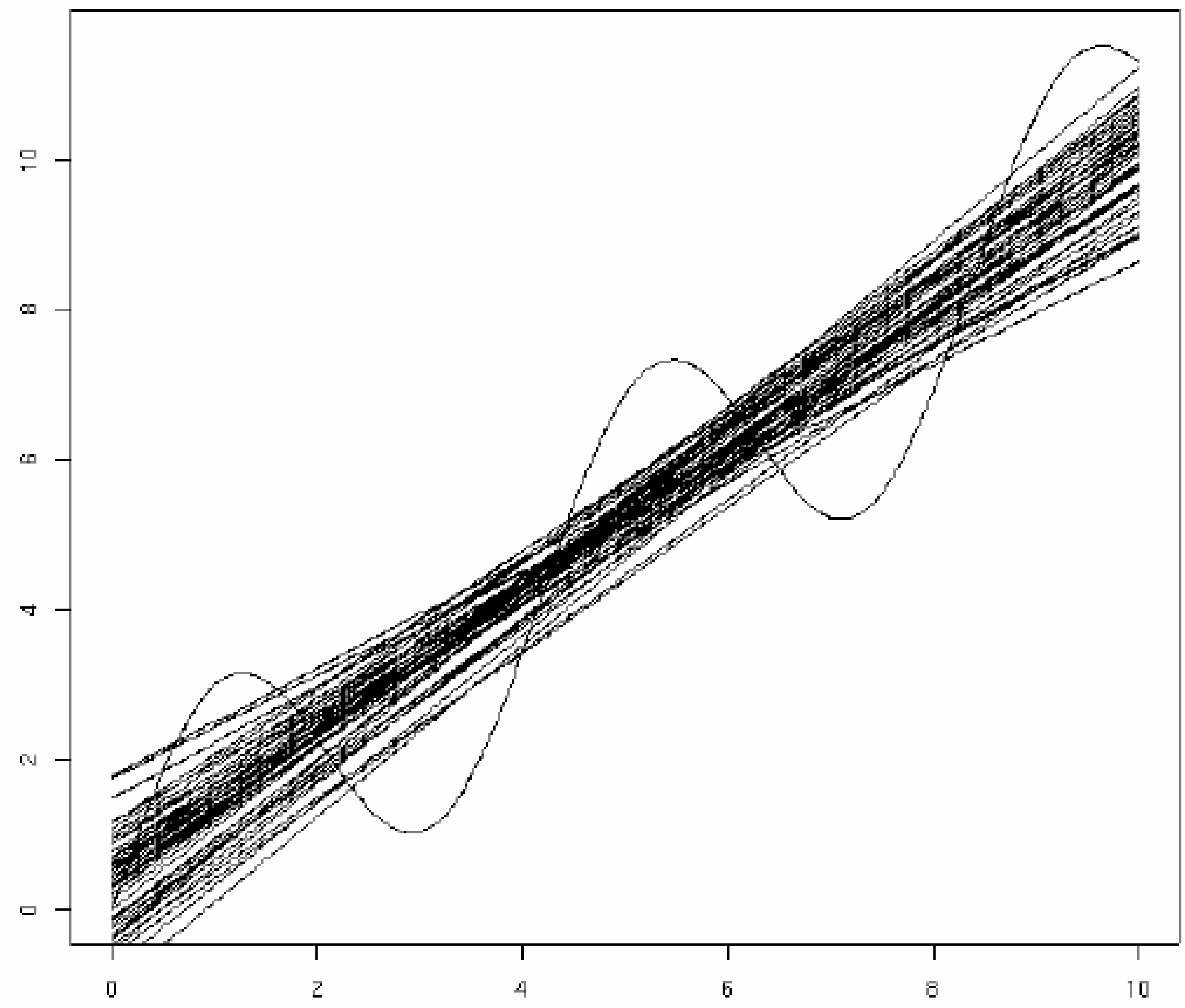
- Bias: $[\underline{h(x^*)} - f(x^*)]$

Describes the average error of $h(x^*)$.

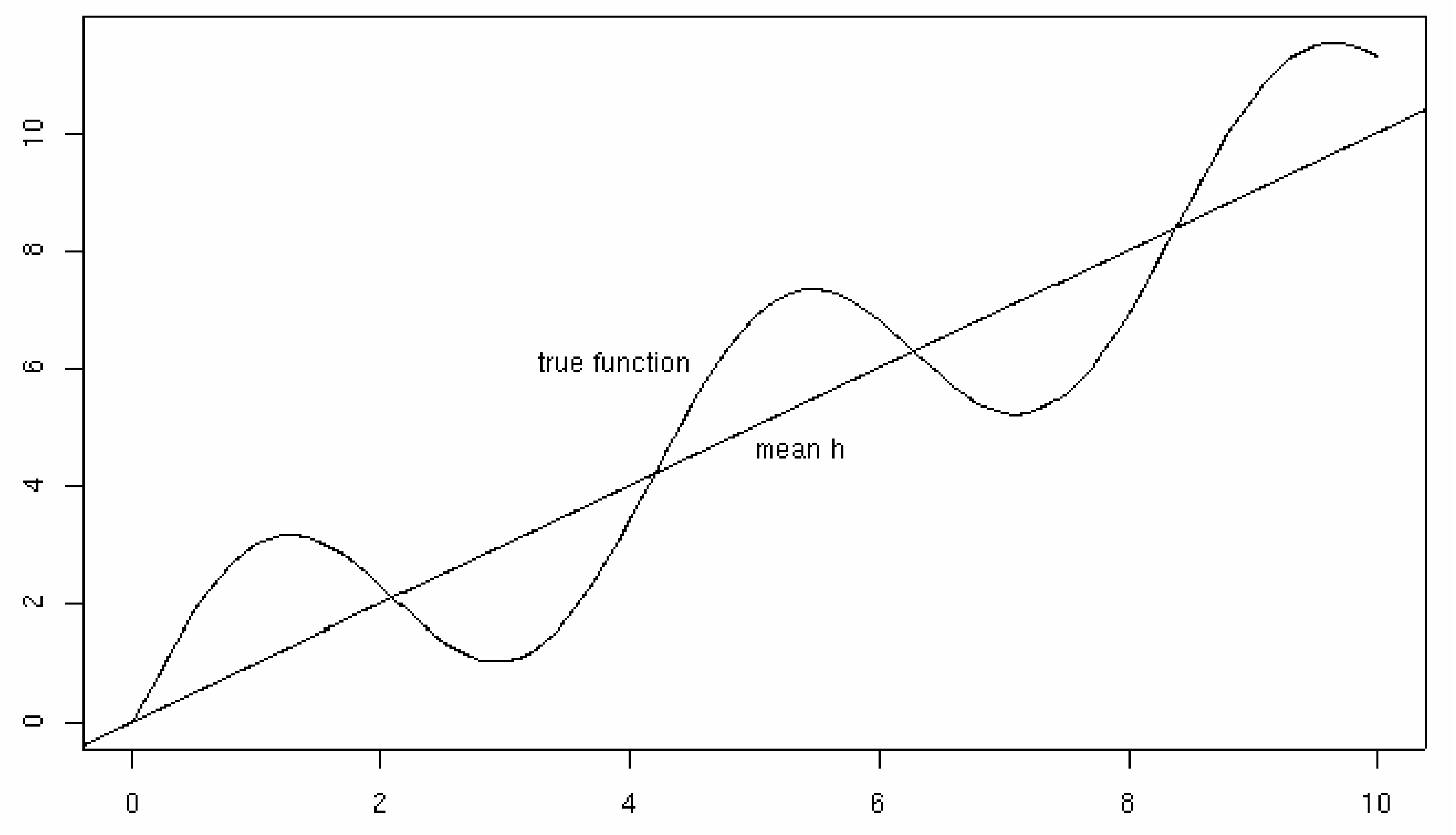
- Noise: $E[\ (y^* - f(x^*))^2\ ] = E[\varepsilon^2] = \sigma^2$

Describes how much y^* varies from $f(x^*)$

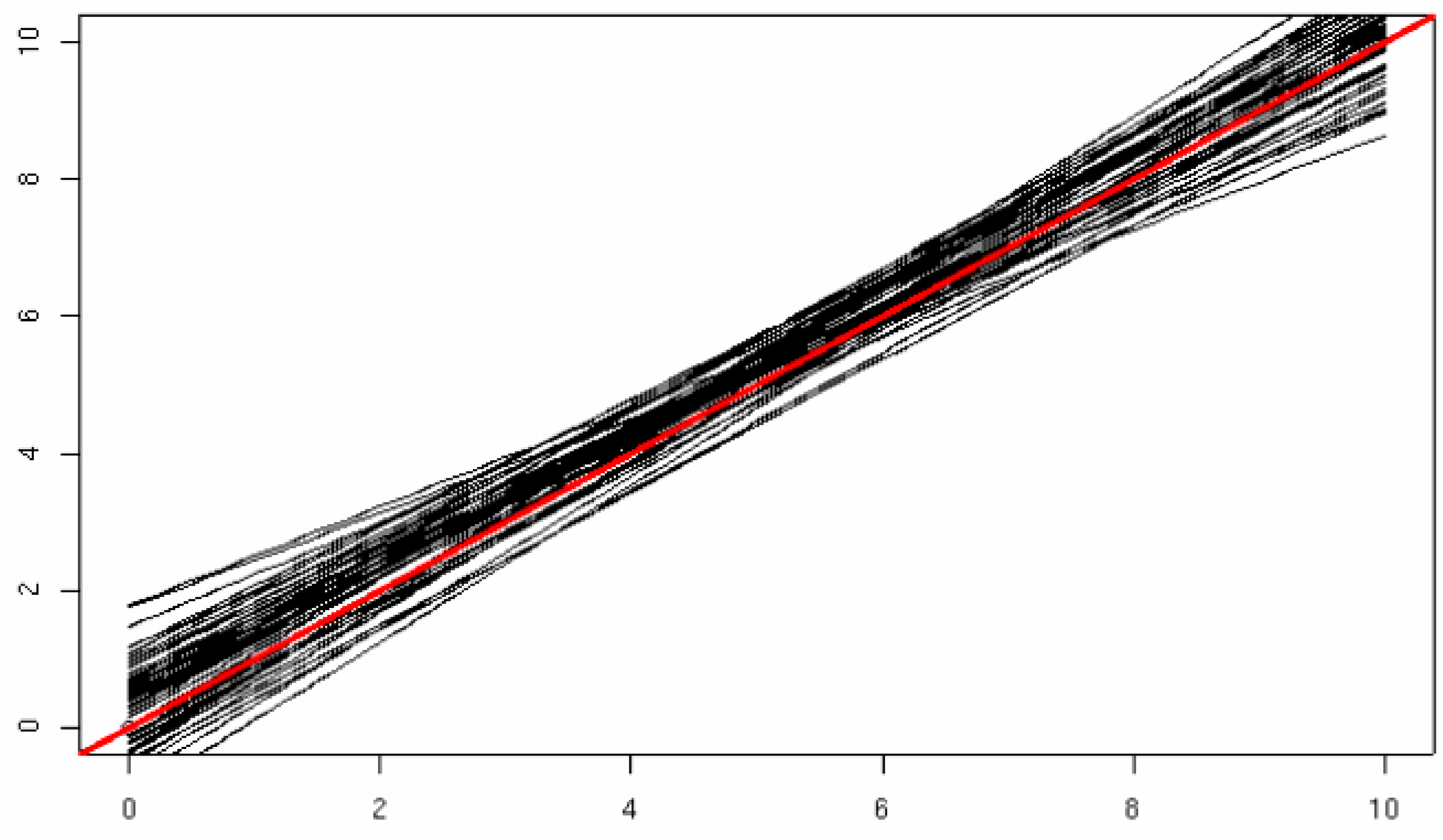
50 fits (of 20 examples each)



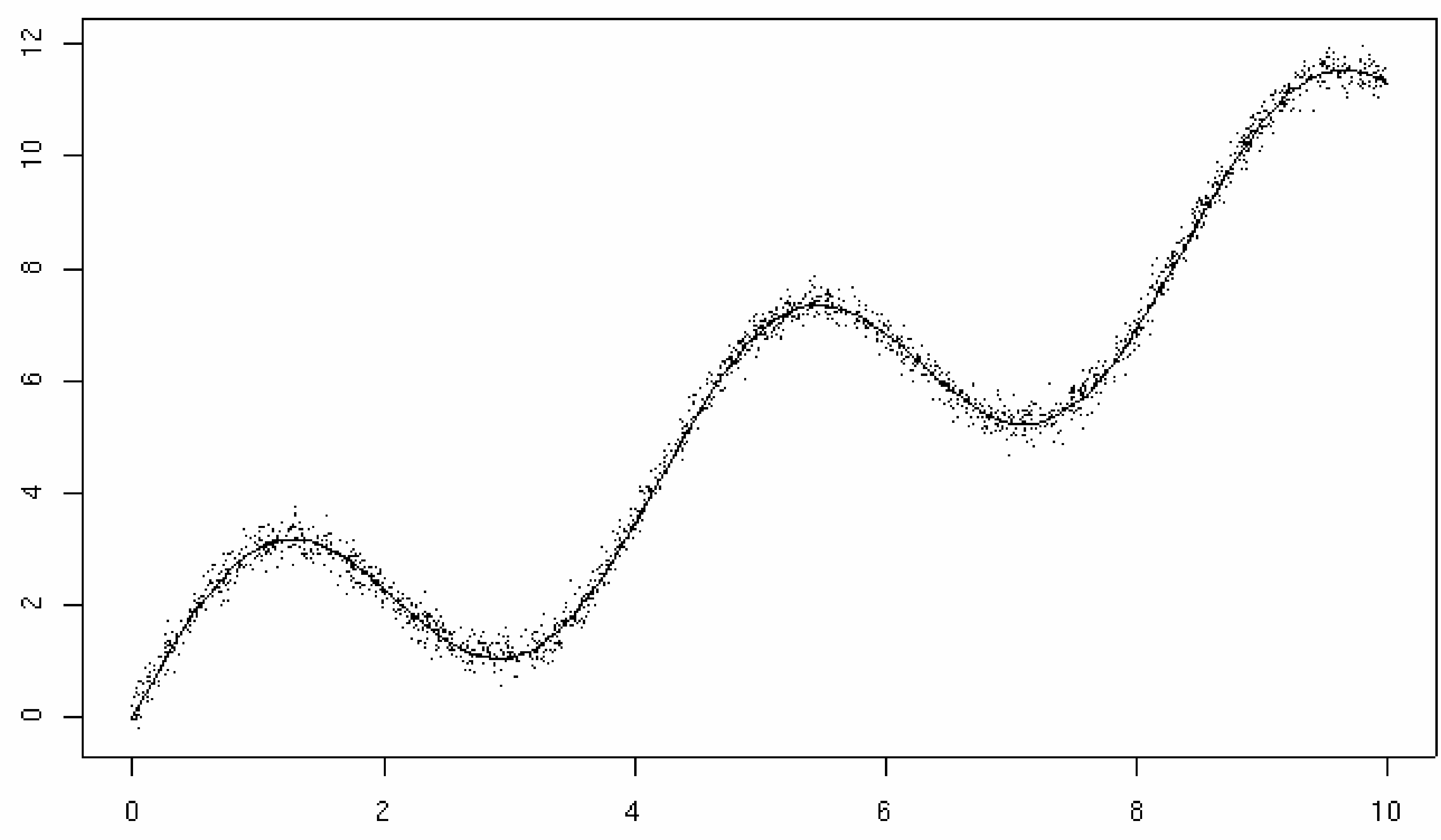
Bias



Variance



Noise



Measuring Bias and Variance

- In practice (unlike in theory), we have only ONE training set S.
- We can simulate multiple training sets by bootstrap replicates

  S’ = {x | x is drawn at random with replacement from S} and |S’| = |S|.

Bias and Variance Measurement Procedure

- Construct B bootstrap replicates of S (e.g., B = 200): S1, …, S_B
- Apply learning algorithm to each replicate S_b to obtain hypothesis h_b
- Let $T_b = S \setminus S_b$ be the data points that do not appear in S_b (out of bag points)
- Compute predicted value $h_b(x)$ for each x in T_b

Estimating B/V/N

- For each data point x, we will now have the observed corresponding value y and several predictions $y_1, \ldots, y_K$

- Compute the average prediction $\underline{h}$

- Estimate bias as $(\underline{h} - y)$

- Estimate variance as $\Sigma_k (y_k - h)^2/(K - 1)$

- Assume noise is 0

Approximations

- Bootstrap replicates are not real data
- We ignore the noise
  - If we have multiple data points with the same x value, then we can estimate the noise
  - We can also estimate noise by pooling y values from nearby x values

This naturally leads us to Ensemble methods – Bagging and Boosting