Bias-Variance Theory

- Decompose Error Rate into components, some of which can be measured on unlabeled data

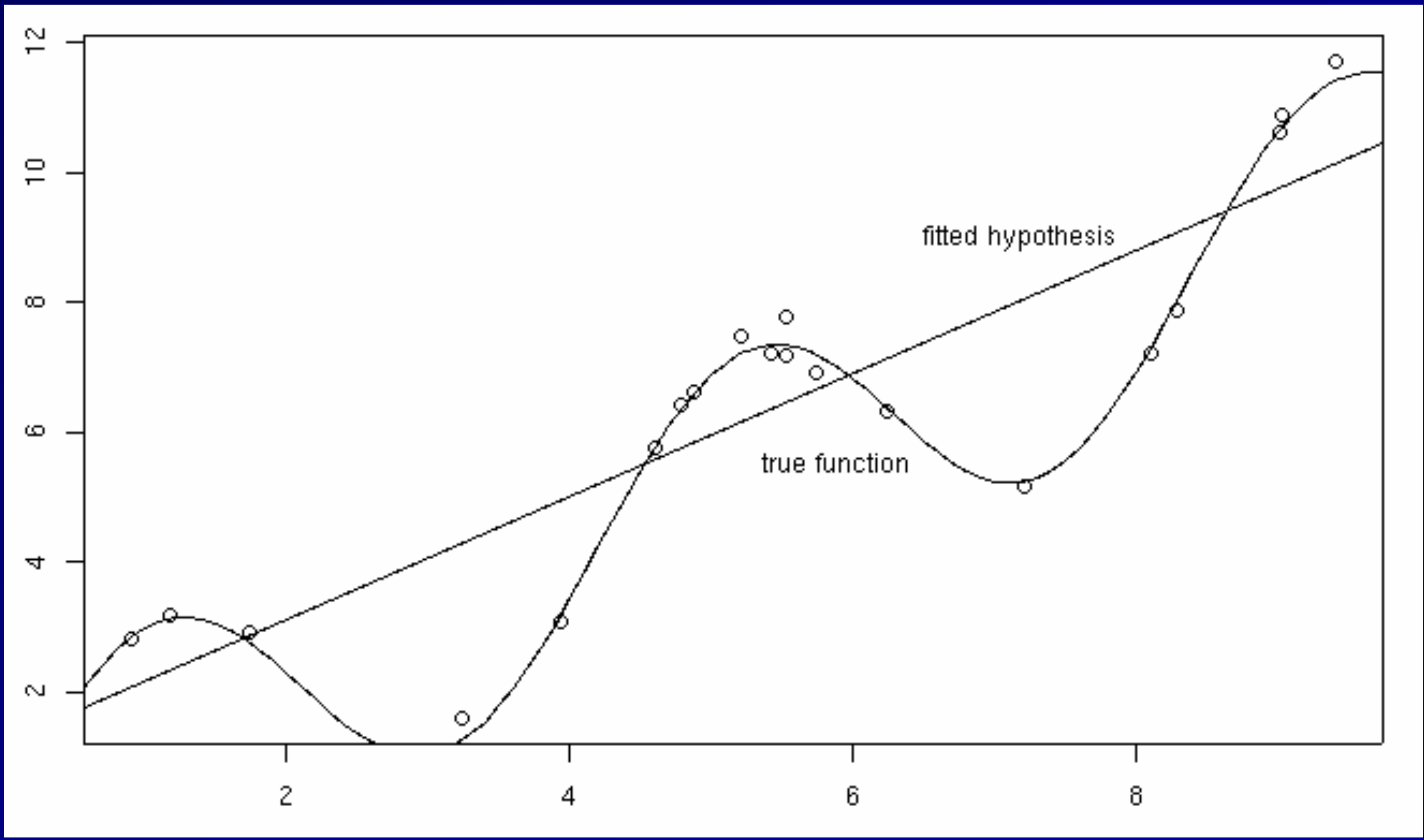
- Bias-Variance Decomposition for Regression
- Bias-Variance Decomposition for Classification
- Bias-Variance Analysis of Learning Algorithms
- Effect of Bagging on Bias and Variance
- Effect of Boosting on Bias and Variance
- Summary and Conclusion

Bias-Variance Analysis in Regression

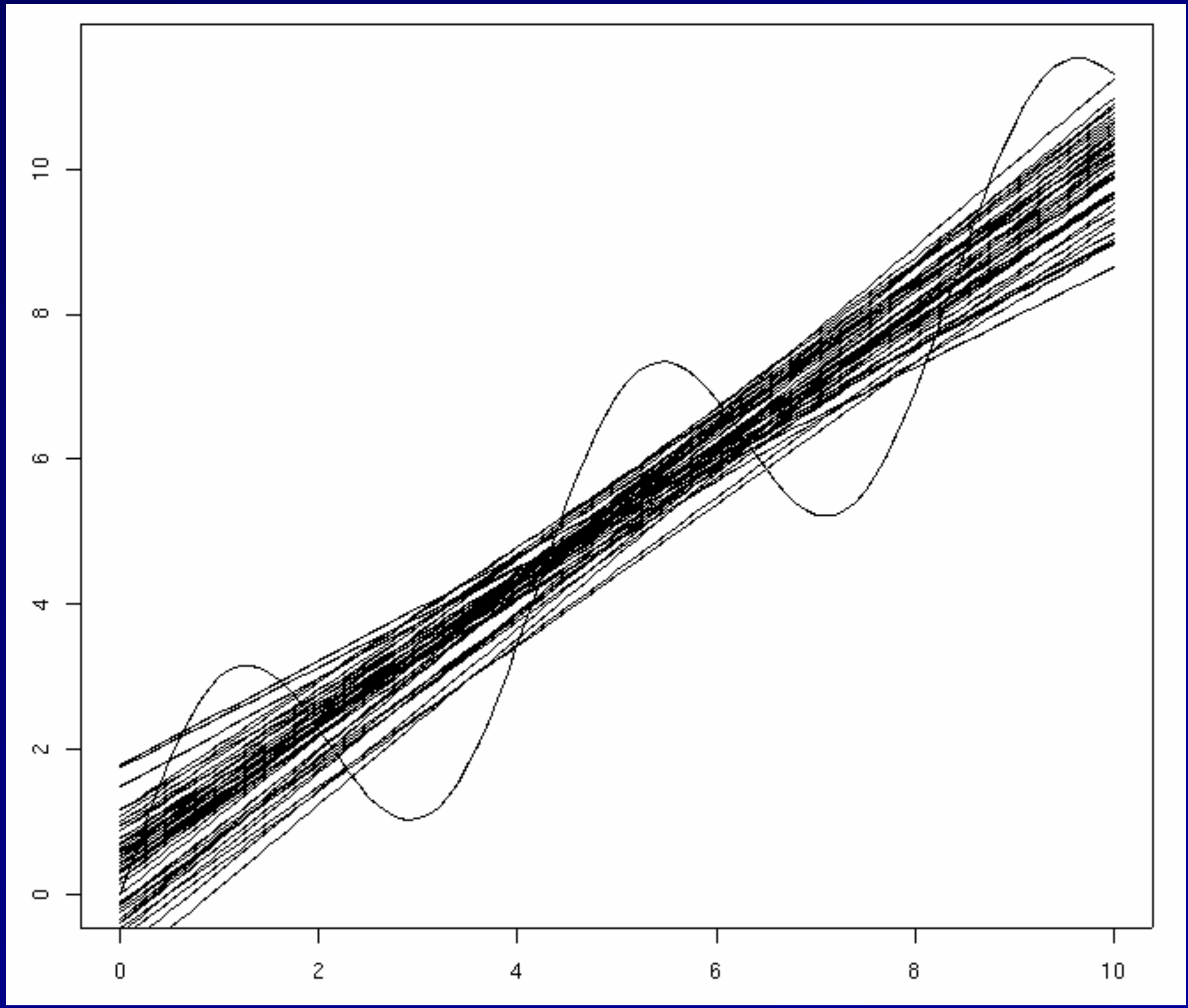
- True function is $y = f(x) + \varepsilon$
  – where ε is normally distributed with zero mean and standard deviation σ.
- Given a set of training examples, $\{(x_i, y_i)\}$, we fit an hypothesis $h(x) = w \cdot x + b$ to the data to minimize the squared error

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

Example: 20 points
$y = x + 2\sin(1.5x) + N(0,0.2)$



50 fits (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

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

Classical 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”

Lemma

- Let Z be a random variable with probability distribution P(Z)
- Let $\underline{Z} = E_P[\, Z \,]$ be the average value of Z.
- Lemma: $E[\,(Z - \underline{Z})^2\,] = E[Z^2] - \underline{Z}^2$

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

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

Bias-Variance-Noise Decomposition

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

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

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

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

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

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

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

$$E[\ (y^* - f(x^*))^2\ ] \qquad \text{[noise]}$$

Derivation (continued)

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

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

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

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

$= Var(h(x^*)) + Bias(h(x^*))^2 + E[\,\varepsilon^2\,]$

$= Var(h(x^*)) + Bias(h(x^*))^2 + \sigma^2$

Expected prediction 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

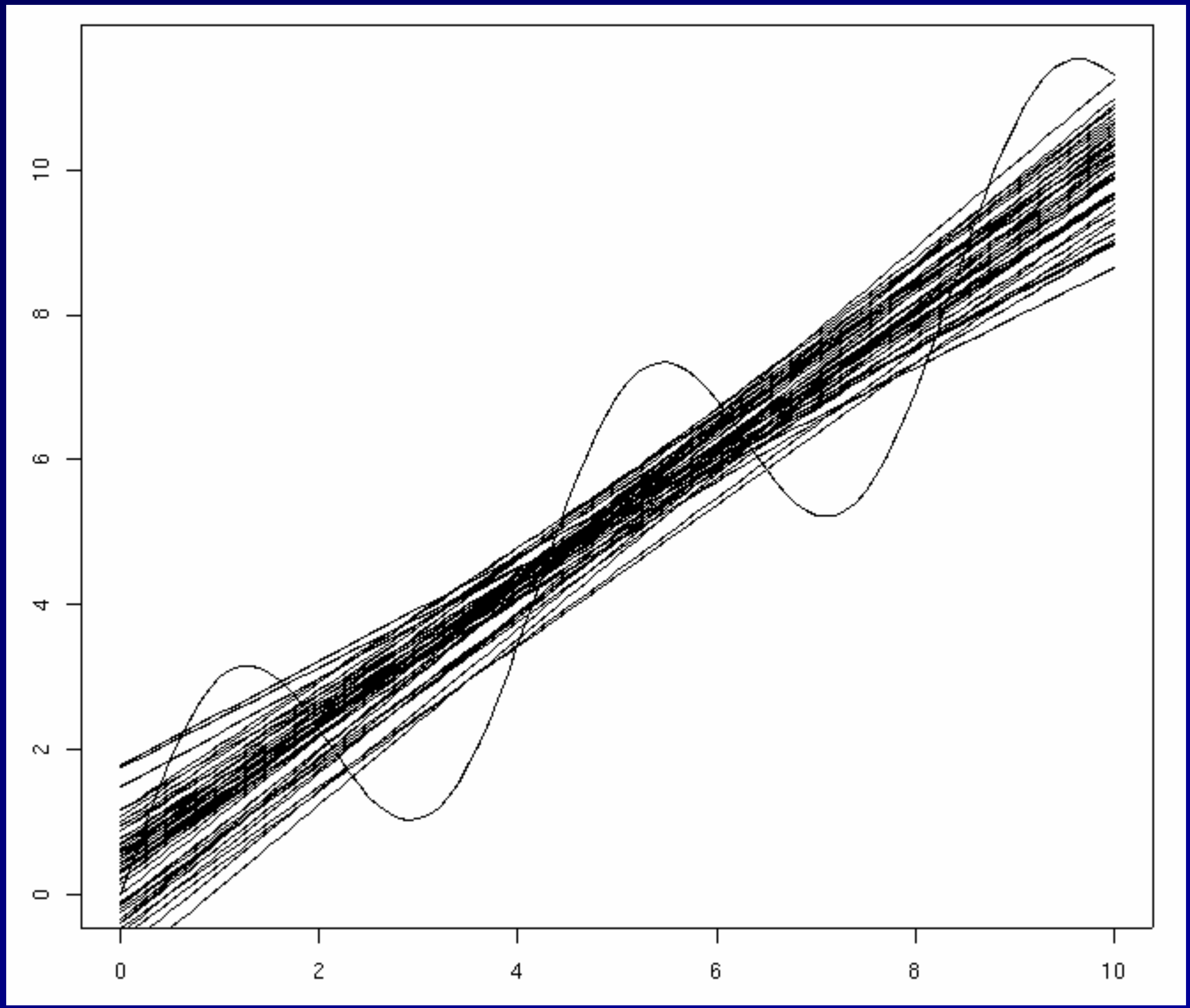
- Bias: $[\underline{h(x^*)} - f(x^*)]$

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

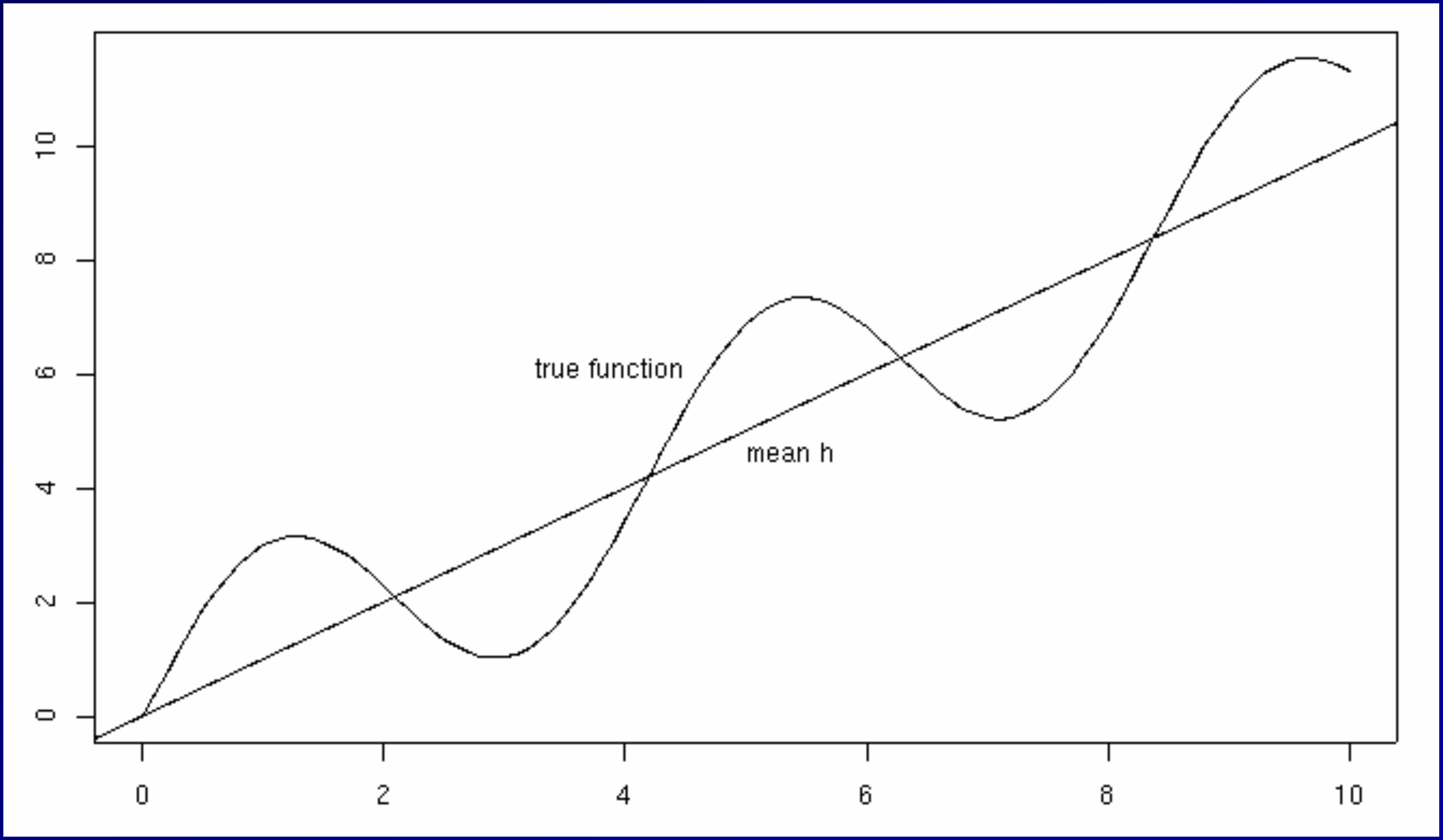
- Noise: $E[\,(y^* - f(x^*))^2\,] = E[\varepsilon^2] = \sigma^2$

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

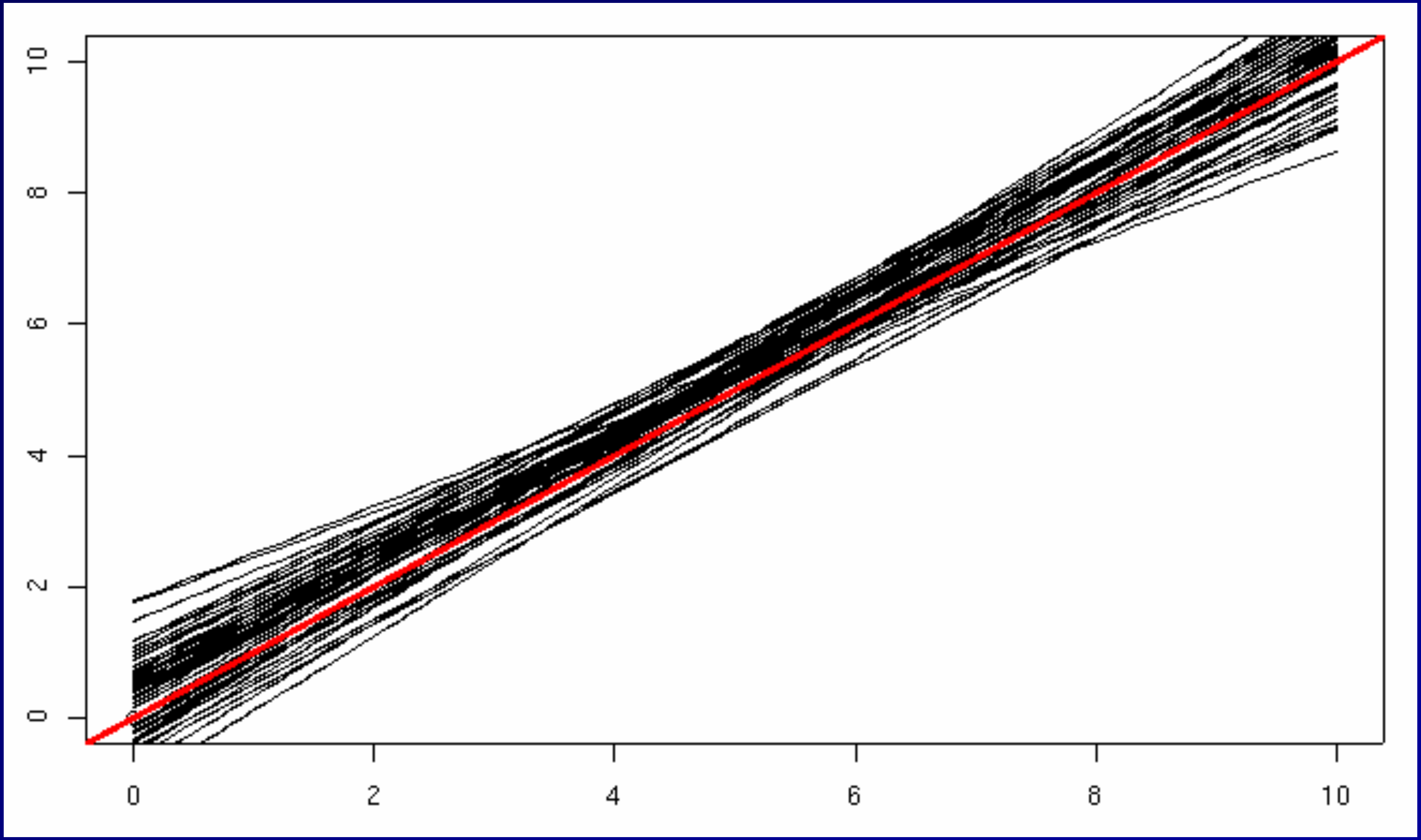
50 fits (20 examples each)



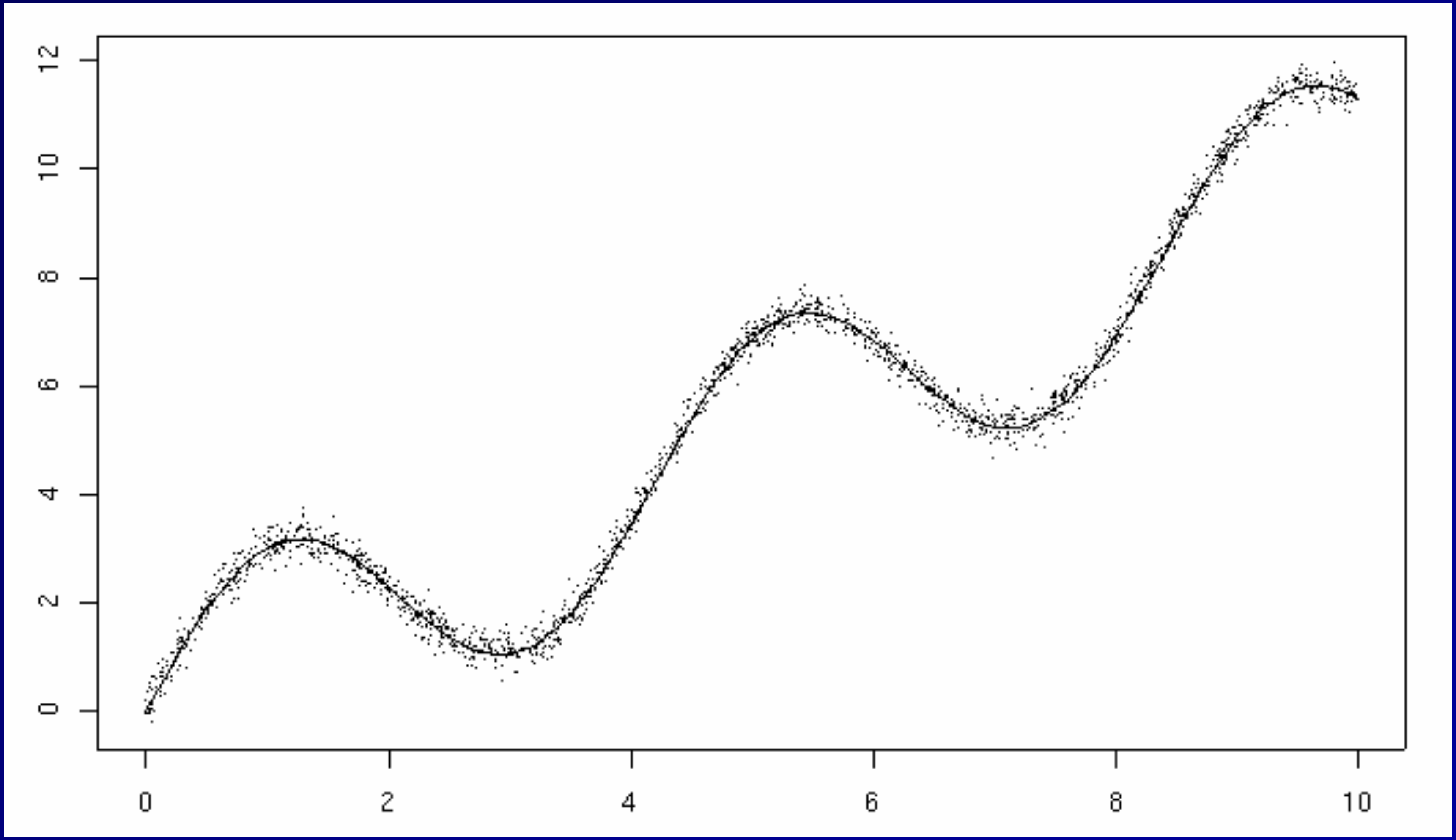
Bias



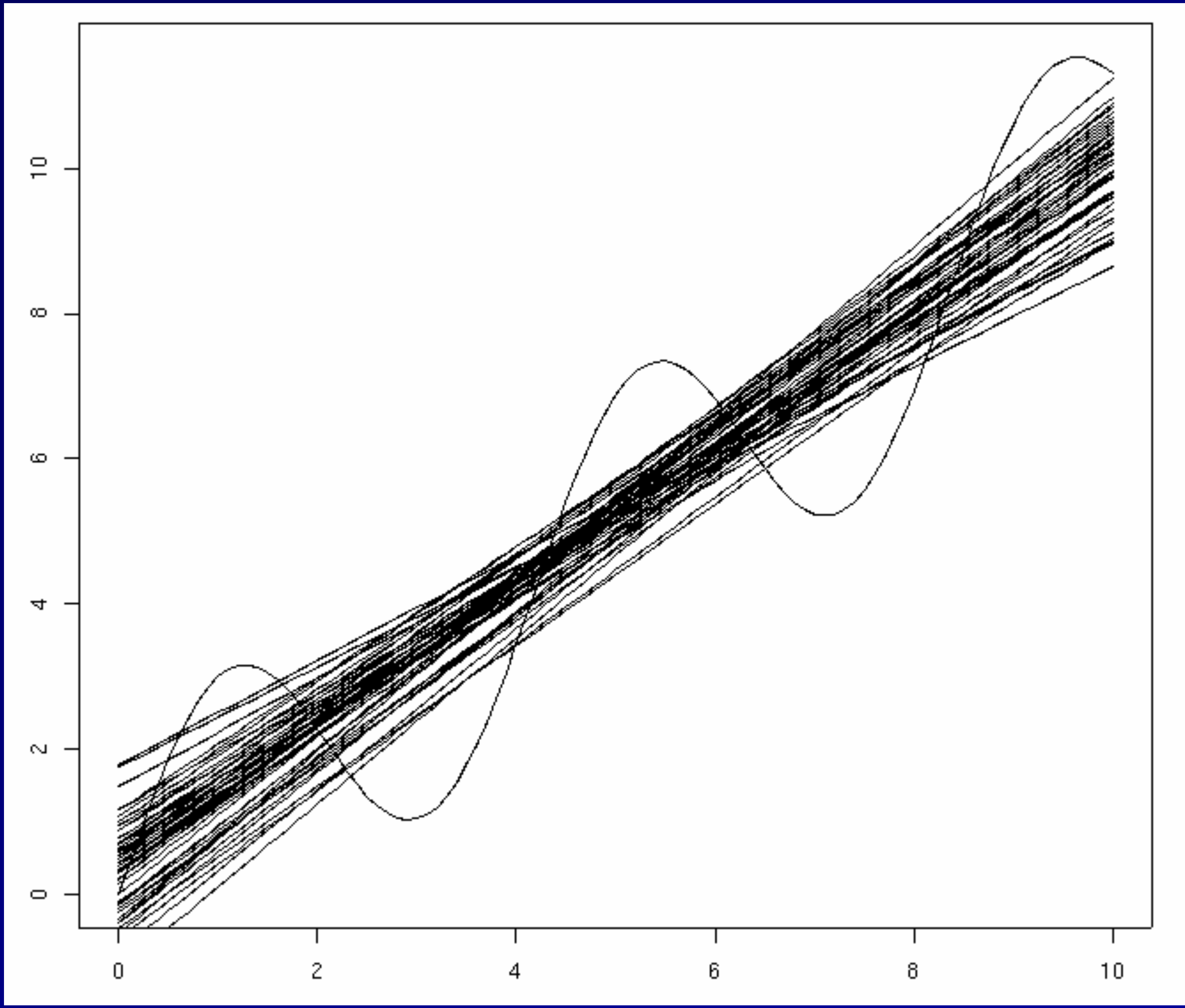
Variance



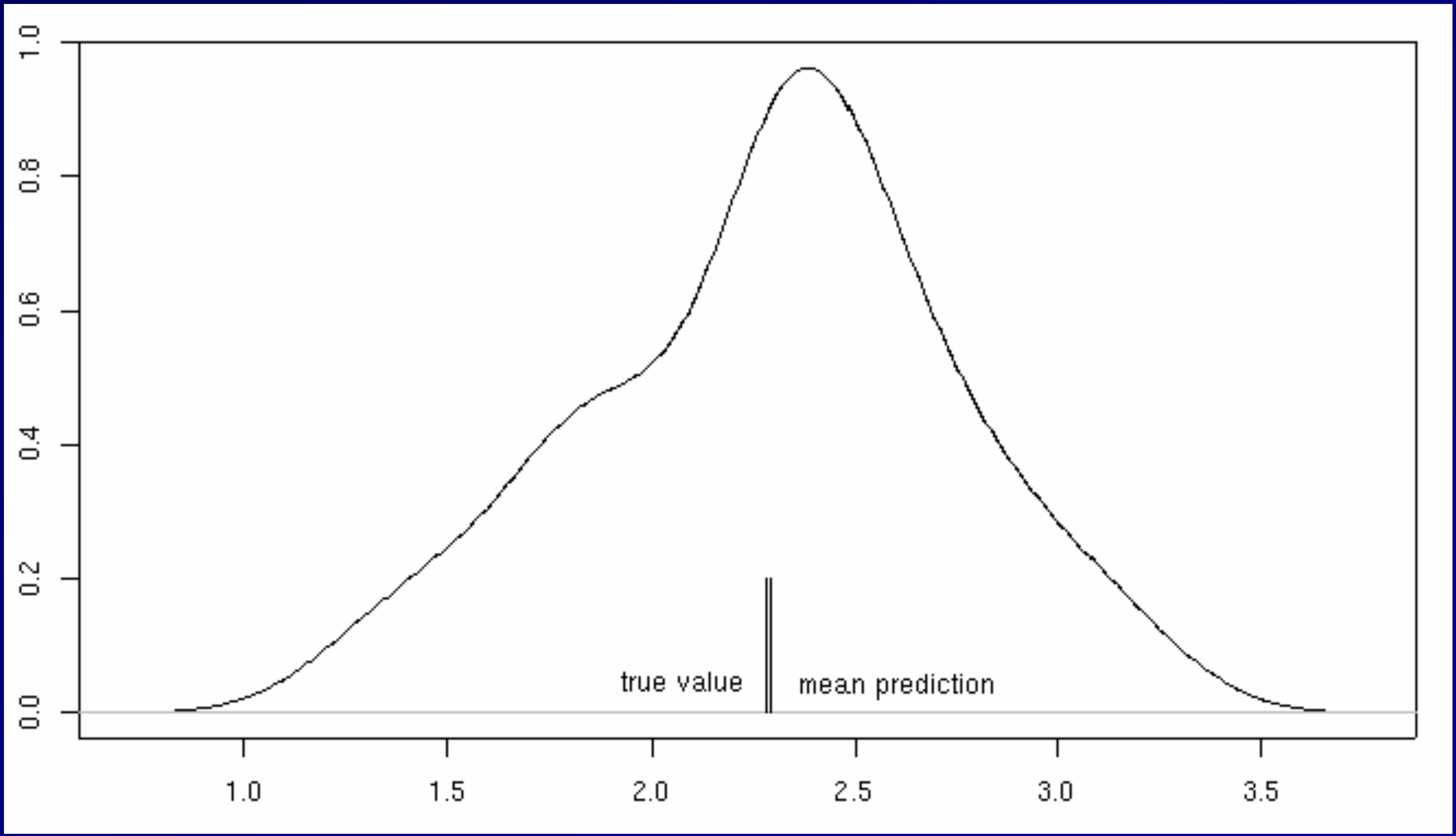
Noise



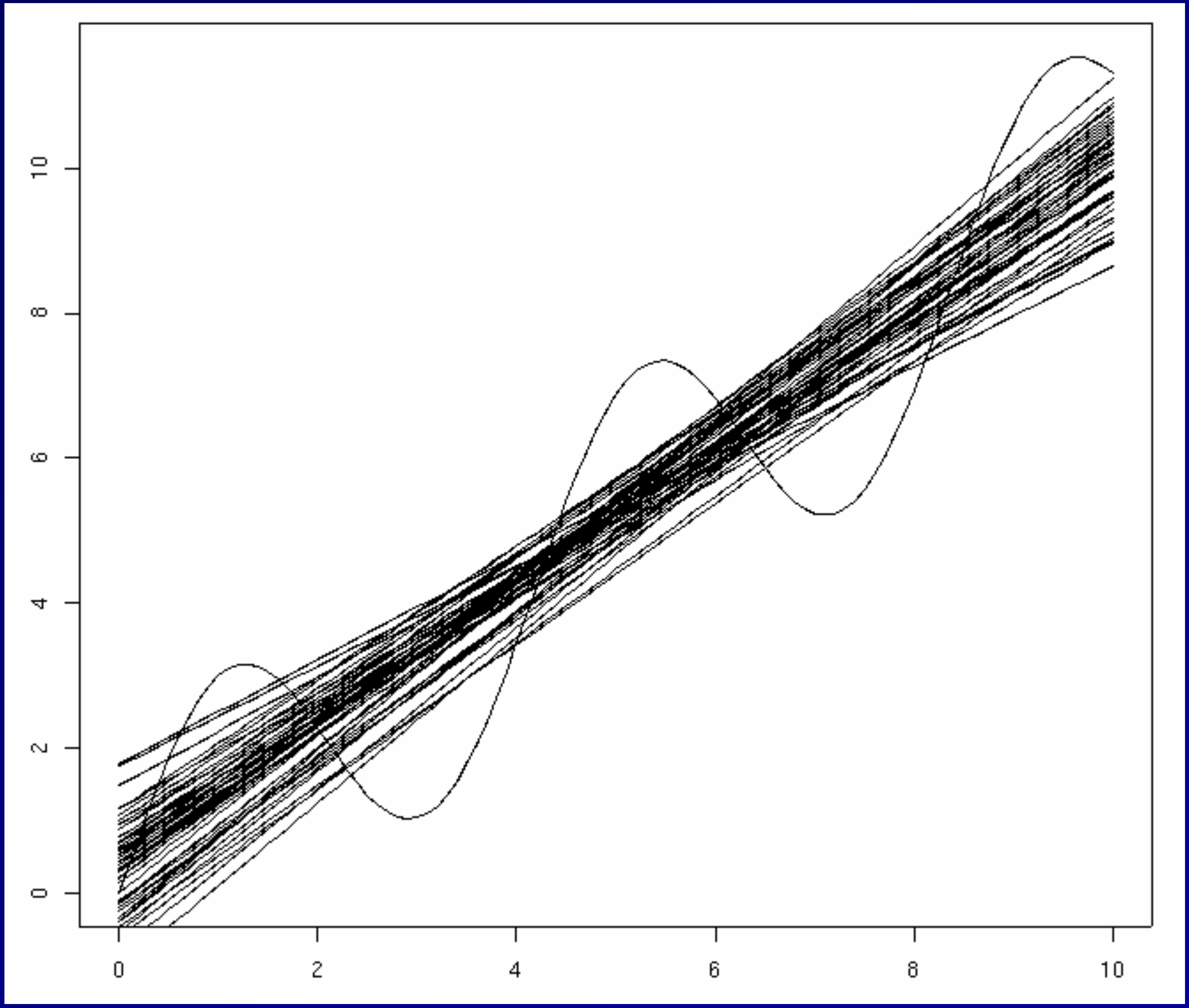
50 fits (20 examples each)



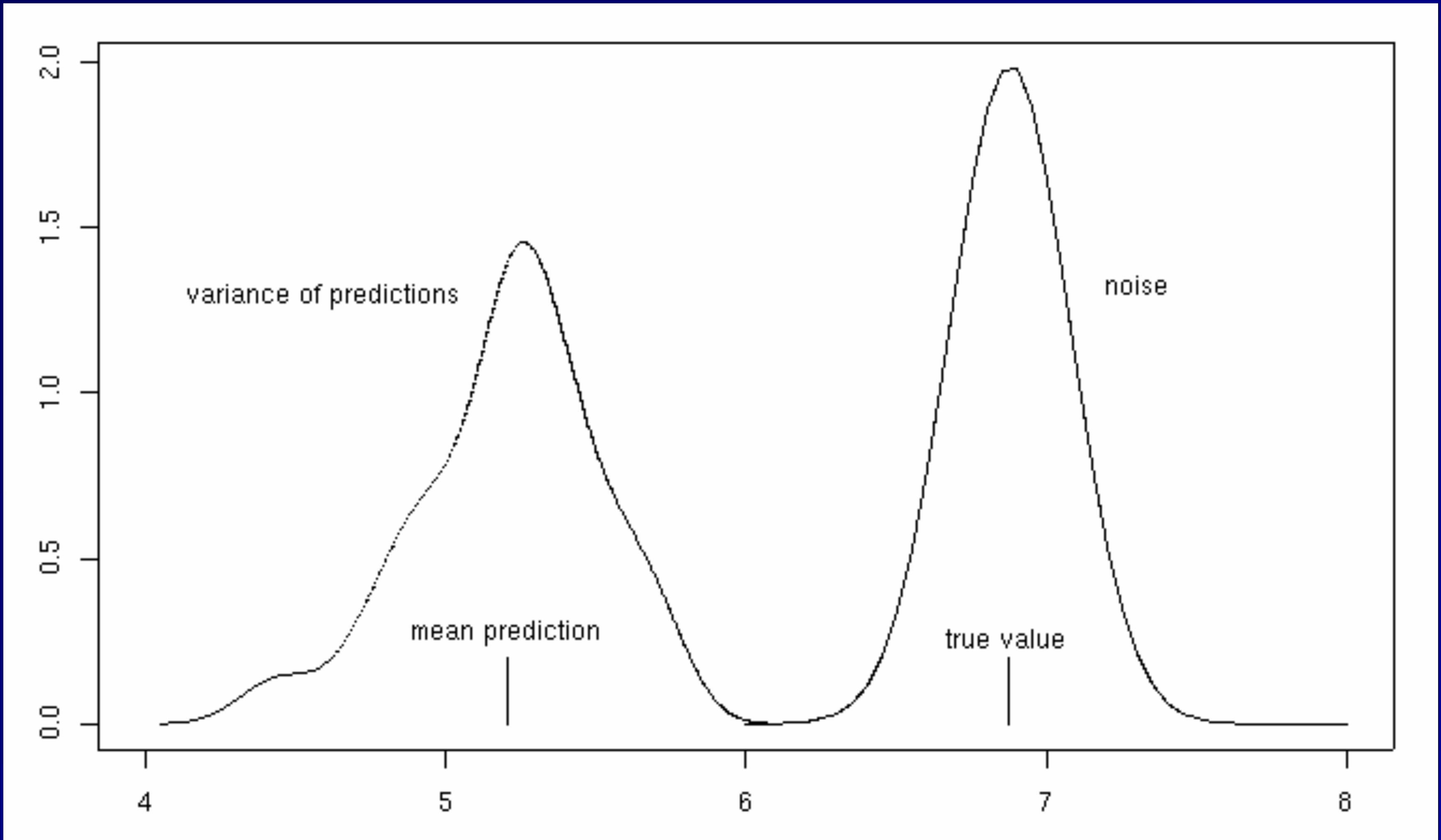
Distribution of predictions at x=2.0



50 fits (20 examples each)



Distribution of predictions at x=5.0



Measuring Bias and Variance

- In practice (unlike in theory), we have only ONE training set S.
- We can simulate multiple training sets by <u>bootstrap replicates</u>
  - S’ = {x | x is drawn at random with replacement from S} and |S’| = |S|.

Procedure for Measuring Bias and Variance

- Construct B bootstrap replicates of S (e.g., B = 200): $S_1, \ldots, S_B$
- 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 Bias and Variance (continued)

- 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 $\Sigma_k (y_k - \underline{h})^2/(K - 1)$
- Assume noise is 0

Approximations in this Procedure

- 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

Ensemble Learning Methods

- Given training sample S
- Generate multiple hypotheses, h_1, h_2, …, h_L.
- Optionally: determining corresponding weights w_1, w_2, …, w_L
- Classify new points according to

$$\sum_l w_l h_l > \theta$$

Bagging: Bootstrap Aggregating

- For b = 1, …, B do
  - S_b = bootstrap replicate of S
  - Apply learning algorithm to S_b to learn h_b
- Classify new points by unweighted vote:
  - $[\sum_b h_b(x)]/B > 0$

Bagging

- Bagging makes predictions according to

  $y = \Sigma_b h_b(x) / B$
- Hence, bagging's predictions are $\underline{h}(x)$

Estimated Bias and Variance of Bagging

- If we estimate bias and variance using the same B bootstrap samples, we will have:
  - Bias = ($\underline{h}$ – y) [same as before]
  - Variance = $\Sigma_k (\underline{h} - \underline{h})^2/(K - 1) = 0$
- Hence, according to this approximate way of estimating variance, bagging removes the variance while leaving bias unchanged.
- In reality, bagging only *reduces* variance and tends to slightly increase bias

Bias/Variance Heuristics

- Models that fit the data poorly have high bias: “inflexible models” such as linear regression, regression stumps
- Models that can fit the data very well have low bias but high variance: “flexible” models such as nearest neighbor regression, regression trees
- This suggests that bagging of a flexible model can reduce the variance while benefiting from the low bias