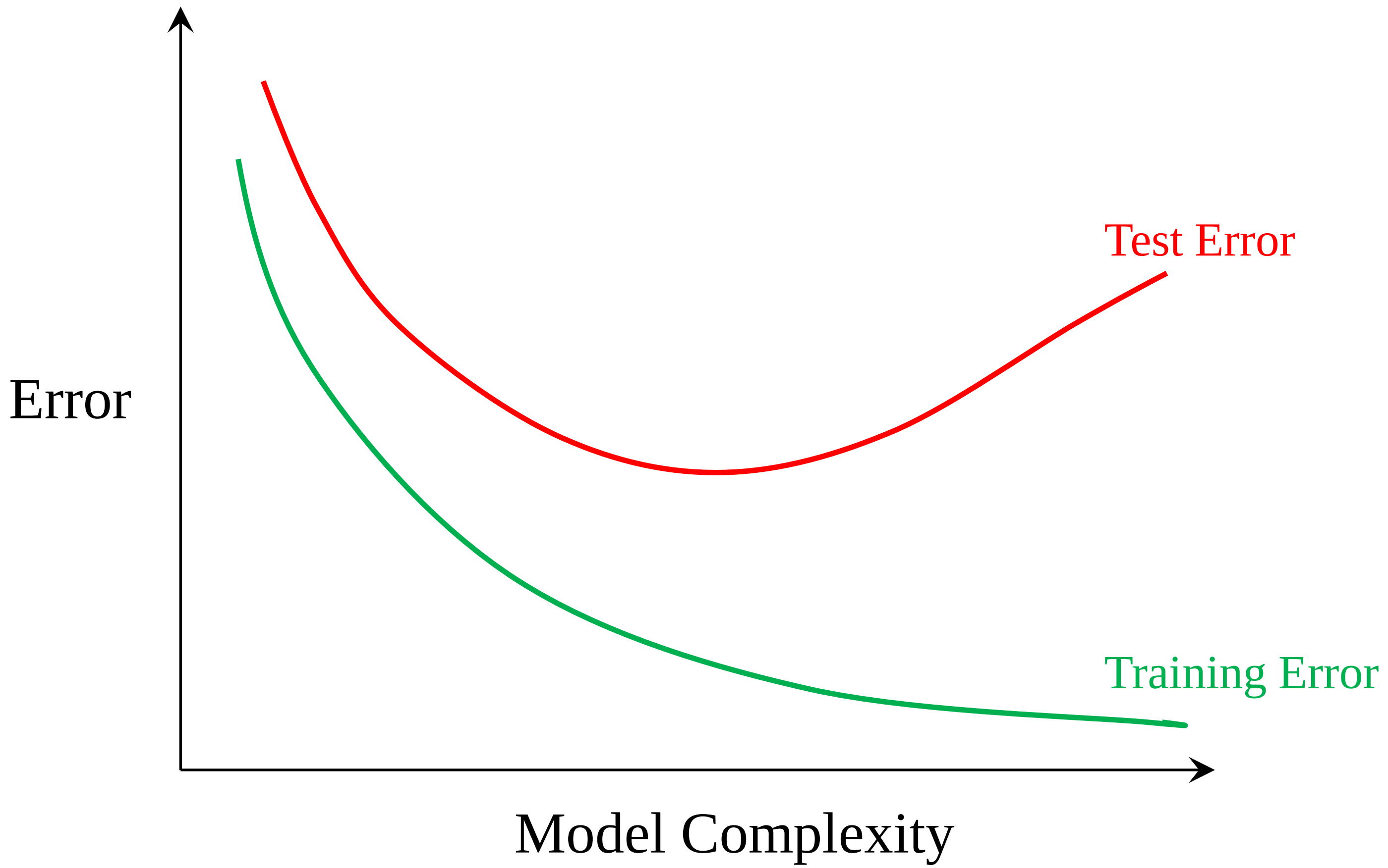
Bias-Variance trade-off

Dripta Mj
Department of Mathematics
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, India

Error vs complexity



Bias-variance decomposition

- Dataset: $\mathcal{D} = \left\{ \left(\mathbf{x}^{(1)}, y^{(1)}\right), \left(\mathbf{x}^{(2)}, y^{(2)}\right), ..., \left(\mathbf{x}^{(N)}, y^{(N)}\right) \right\}$
- Let $g_{\mathcal{D}}$ be the hypothesis which is fit to a particular training dataset $\mathcal{D}$
- Want to compute the expected prediction error at an arbitrary test point with input $\mathbf{x}$ and output y: $\mathbb{E}_{\mathbf{x},y,\mathcal{D}}\left[(g_{\mathcal{D}}(\mathbf{x}) - y)^2\right]$.
- Mean prediction of the machine learning algorithm:

$$\overline{g}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}\left[g_{\mathcal{D}}(\mathbf{x})\right]$$

- So determining the value of $\overline{g}(\mathbf{x})$ involve
  - generating different training datasets ($\mathcal{D}$),
  - training separate functions ($g_{\mathcal{D}}$) for every generated dataset,
  - making predictions at an arbitrary test point $\mathbf{x}$ with all trained functions,
  - and finally, averaging over all the predictions.
- Let $\overline{y}(\mathbf{x})$ be the expected value of the output at $\mathbf{x}$, i.e. $\overline{y}(\mathbf{x}) = \mathbb{E}_{y|\mathbf{x}}[y]$.

Bias-variance decomposition

- The expected error can be simplified as:

$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}\left[(g_{\mathcal{D}}(\mathbf{x}) - y)^2\right] = \mathbb{E}_{\mathbf{x},y,\mathcal{D}}\left[\left((g_{\mathcal{D}}(\mathbf{x}) - \overline{g}(\mathbf{x})) + (\overline{g}(\mathbf{x}) - y)\right)^2\right]$$

$$= \mathbb{E}_{\mathbf{x},\mathcal{D}}\left[\left(g_{\mathcal{D}}(\mathbf{x}) - \overline{g}(\mathbf{x})\right)^2\right] + 2\mathbb{E}_{\mathbf{x},y,\mathcal{D}}\left[\left(g_{\mathcal{D}}(\mathbf{x}) - \overline{g}(\mathbf{x})\right)\left(\overline{g}(\mathbf{x}) - y\right)\right]$$

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

The second term on simplification yields

$$2\mathbb{E}_{\mathbf{x},y,\mathcal{D}}\left[\left(g_{\mathcal{D}}(\mathbf{x}) - \overline{g}(\mathbf{x})\right)\left(\overline{g}(\mathbf{x}) - y\right)\right] = 2\mathbb{E}_{\mathbf{x},y}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g_{\mathcal{D}}(\mathbf{x}) - \overline{g}(\mathbf{x})\right)\right]\left(\overline{g}(\mathbf{x}) - y\right)\right]$$

$$= 2\mathbb{E}_{\mathbf{x},y}\left[\left(\mathbb{E}_{\mathcal{D}}\left[g_{\mathcal{D}}(\mathbf{x})\right] - \overline{g}(\mathbf{x})\right)\left(\overline{g}(\mathbf{x}) - y\right)\right]$$

$$= 2\mathbb{E}_{\mathbf{x},y}\left[\left(\overline{g}(\mathbf{x}) - \overline{g}(\mathbf{x})\right)\left(\overline{g}(\mathbf{x}) - y\right)\right]$$

$$= 2\mathbb{E}_{\mathbf{x},y}\left[0\right]$$

$$= 0$$

Bias-variance decomposition

The third term can be simplified as

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

where $\overline{y}(\mathbf{x})=\mathbb{E}_{y|\mathbf{x}}\left[y\right]$ and the last term can be simplified as:

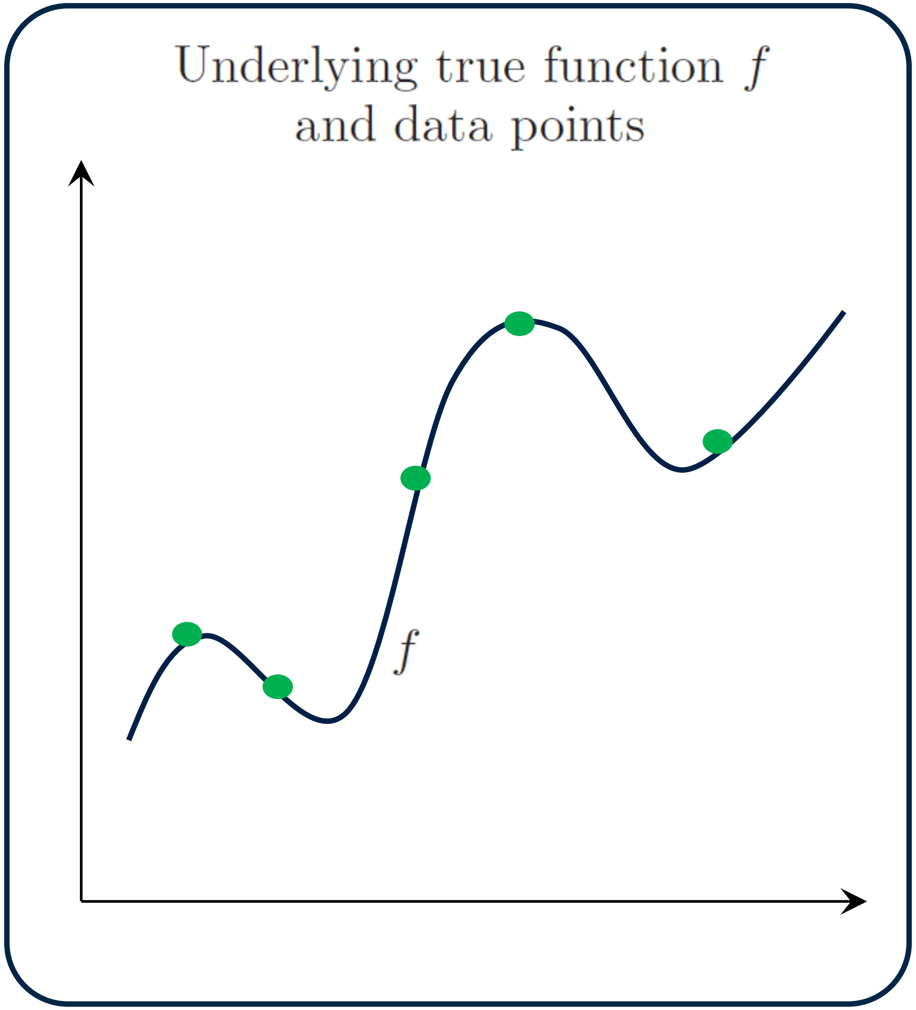
$$\begin{aligned}
2\mathbb{E}_{\mathbf{x},y}\left[\left(\overline{g}(\mathbf{x})-\overline{y}(\mathbf{x})\right)\left(\overline{y}(\mathbf{x})-y\right)\right] &= 2\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{y|\mathbf{x}}\left[\left(\overline{y}(\mathbf{x})-y\right)\right]\left(\overline{g}(\mathbf{x})-\overline{y}(\mathbf{x})\right)\right]\\
&= 2\mathbb{E}_{\mathbf{x}}\left[\left(\overline{y}(\mathbf{x})-\mathbb{E}_{y|\mathbf{x}}\left[y\right]\right)\left(\overline{g}(\mathbf{x})-\overline{y}(\mathbf{x})\right)\right]\\
&= 2\mathbb{E}_{\mathbf{x}}\left[\left(\overline{y}(\mathbf{x})-\overline{y}(\mathbf{x})\right)\left(\overline{g}(\mathbf{x})-\overline{y}(\mathbf{x})\right)\right]\\
&= 0
\end{aligned}$$

Bias-variance decomposition

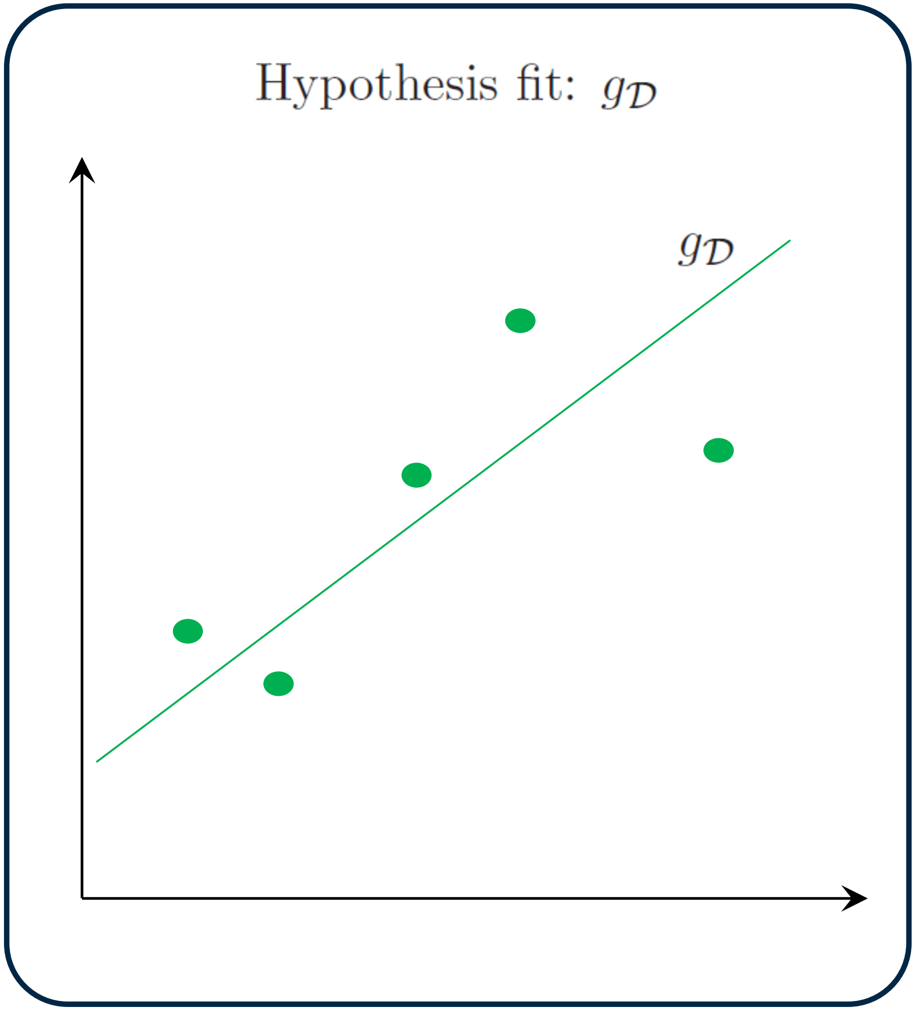
$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}\left[(g_{\mathcal{D}}(\mathbf{x})-y)^2\right] = \underbrace{\mathbb{E}_{\mathbf{x},\mathcal{D}}\left[\left(g_{\mathcal{D}}(\mathbf{x})-\overline{g}(\mathbf{x})\right)^2\right]}_{\text{Variance}} + \underbrace{\mathbb{E}_{\mathbf{x}}\left[\left(\overline{g}(\mathbf{x})-\overline{y}(\mathbf{x})\right)^2\right]}_{\text{Bias}^2} + \underbrace{\mathbb{E}_{\mathbf{x},y}\left[\left(\overline{y}(\mathbf{x})-y\right)^2\right]}_{\text{Noise}}$$

- Variance: It expresses the sensitivity of the solution on the particular choice of dataset $\mathcal{D}$.
- Bias: Difference between the expected prediction (averaged over different datasets) and the expected output value. This is the inherent error arising from the choice of model.
- Noise: Expresses the noise in the data.

Example

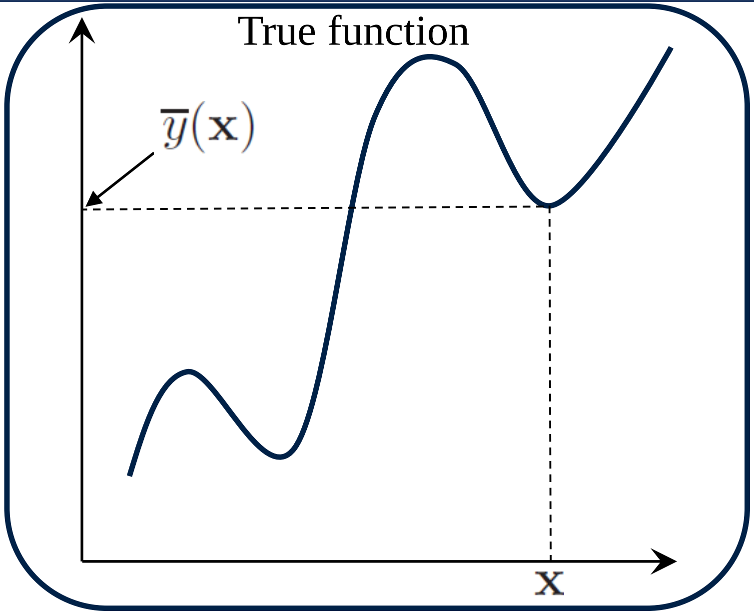


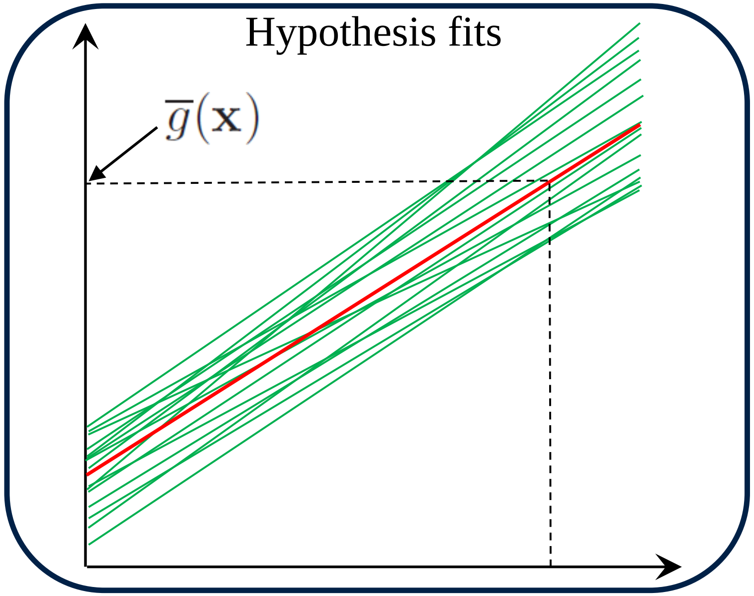
Underlying true function f and data points
f

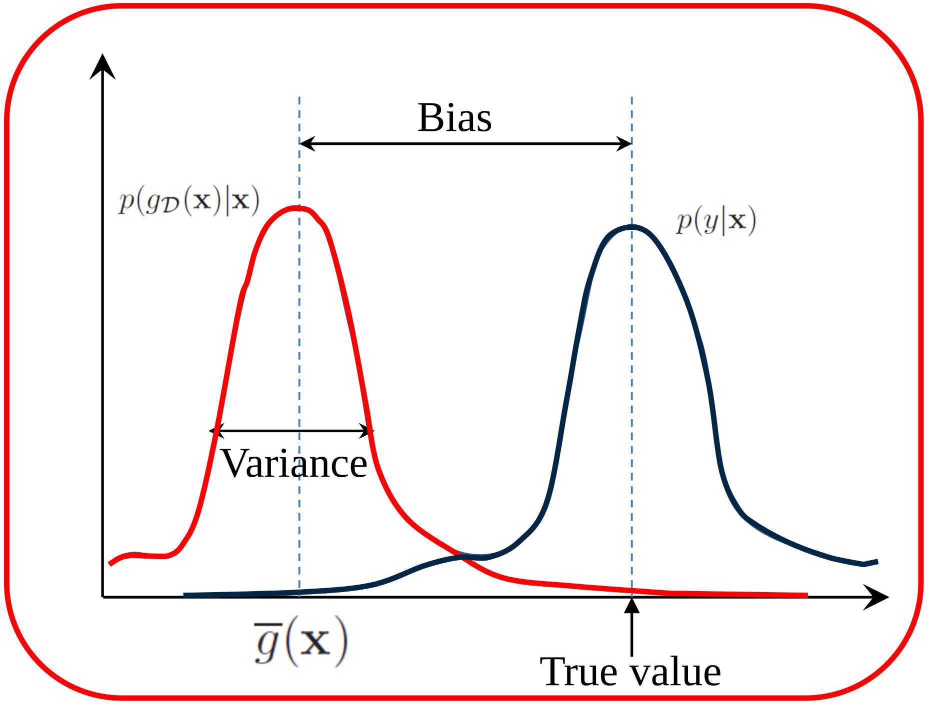


Hypothesis fit: $g_{\mathcal{D}}$
$g_{\mathcal{D}}$

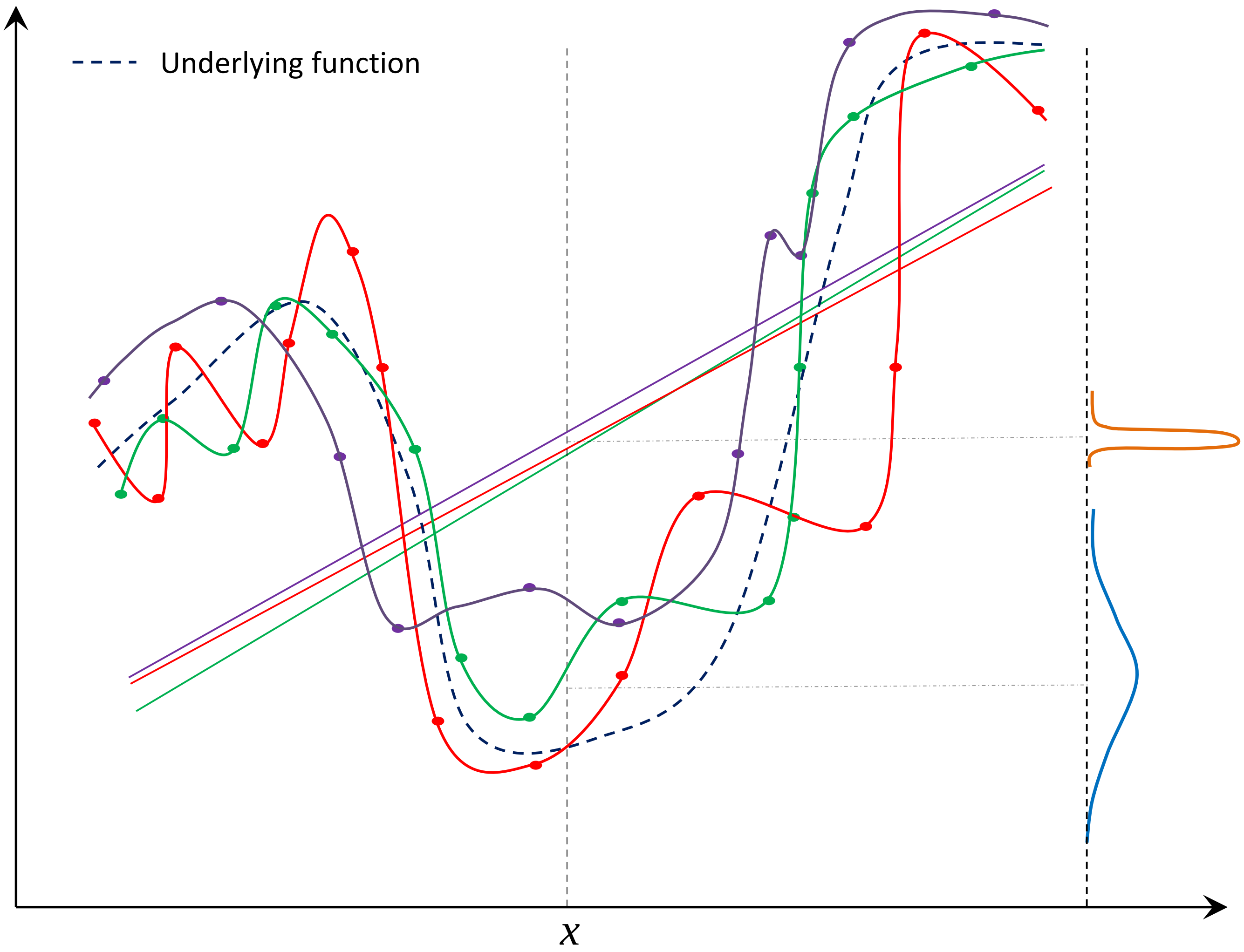
Visualization



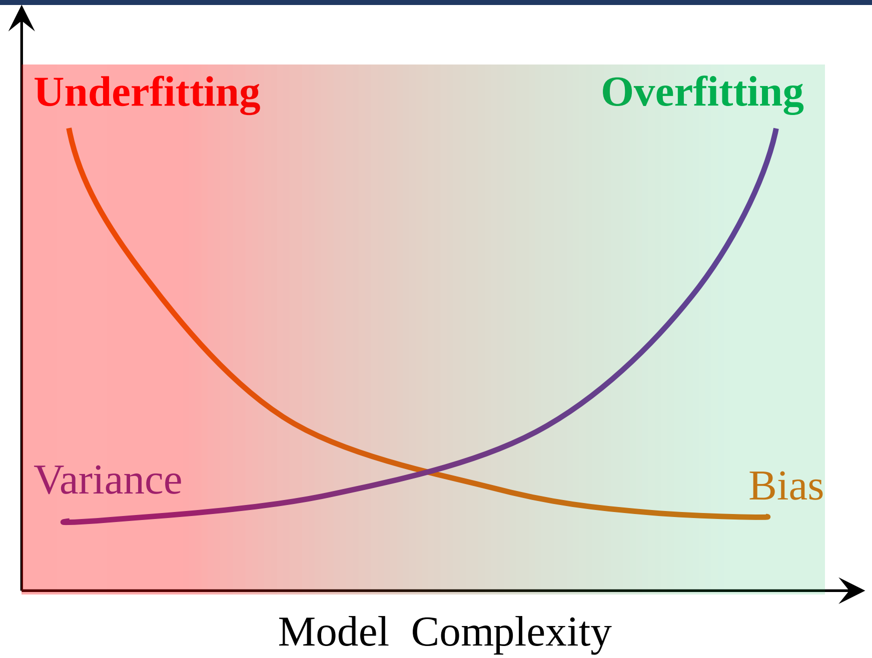




Another example

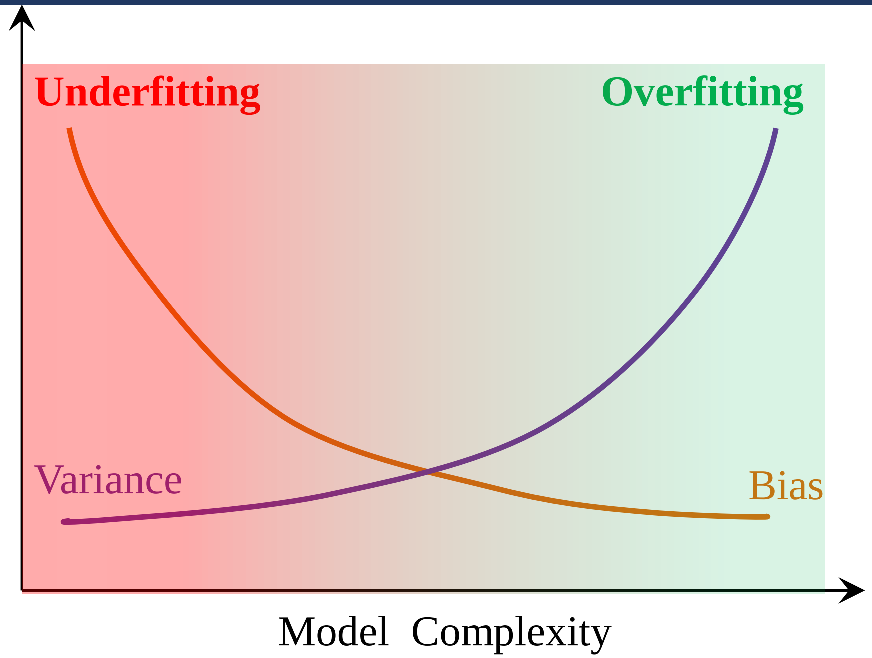


Bias, variance vs model complexity



- **High Bias**: Model is too simple, and so unable to fit the data properly.
  - Results in underfitting.
  - Training and test errors are both large.
- **High Variance**: Model is too complex, and so small changes in the data produce significant changes in the solution.
  - Results in overfitting.
  - Test Error $\gg$ Training Error

Underfitting & Overfitting



- **Underfitting** can be addressed by
  - Increasing the complexity of the model.
  - Minimizing the cost function properly in the training stage.
- **Overfitting** can be addressed by
  - Reducing the complexity of the model.
  - Incorporating some form of regularization inside the cost function.