Machine Learning

CMPT 726

Mo Chen
SFU School of Computing Science
2022-10-13

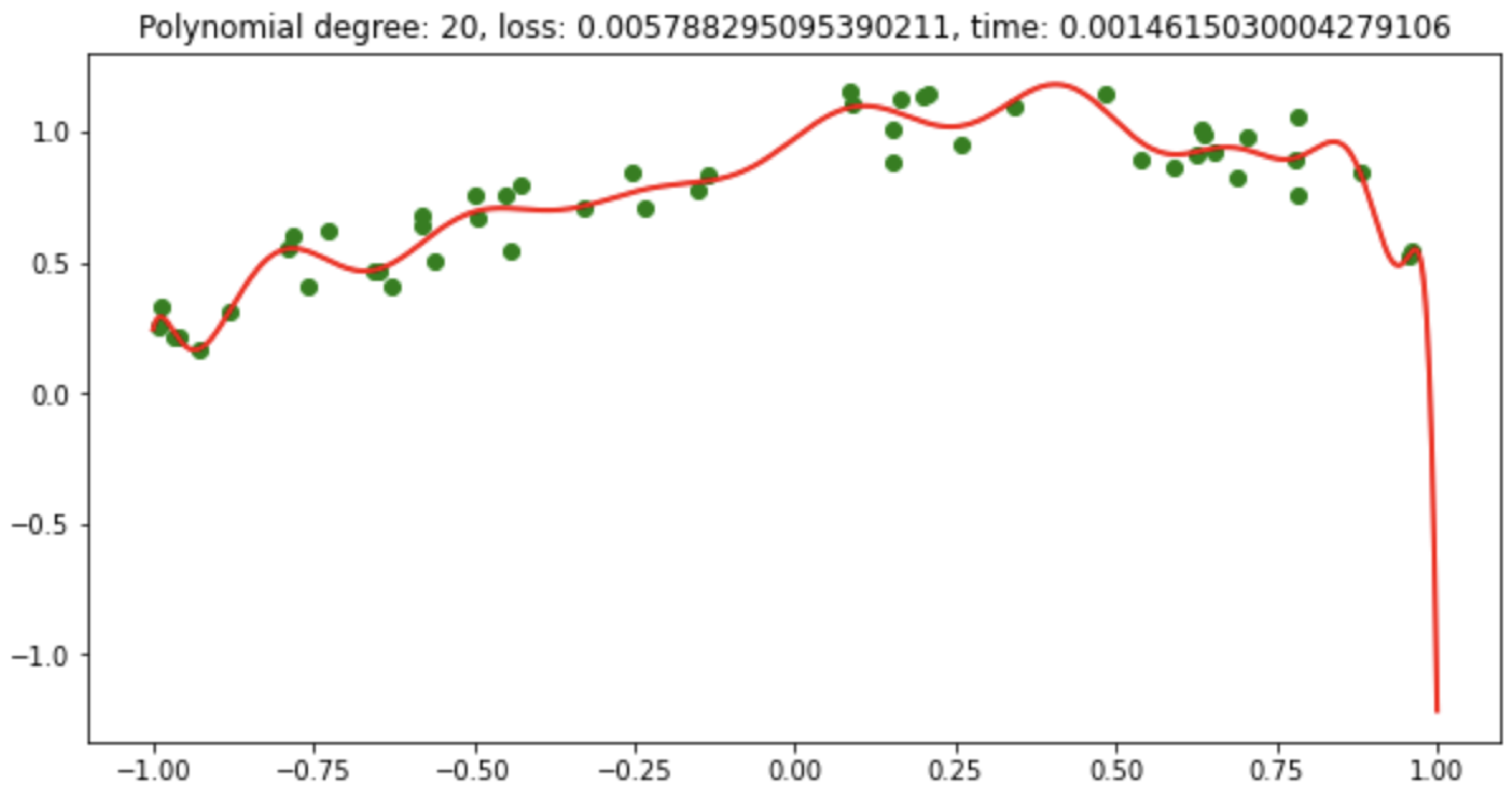
Bias-Variance Tradeoff

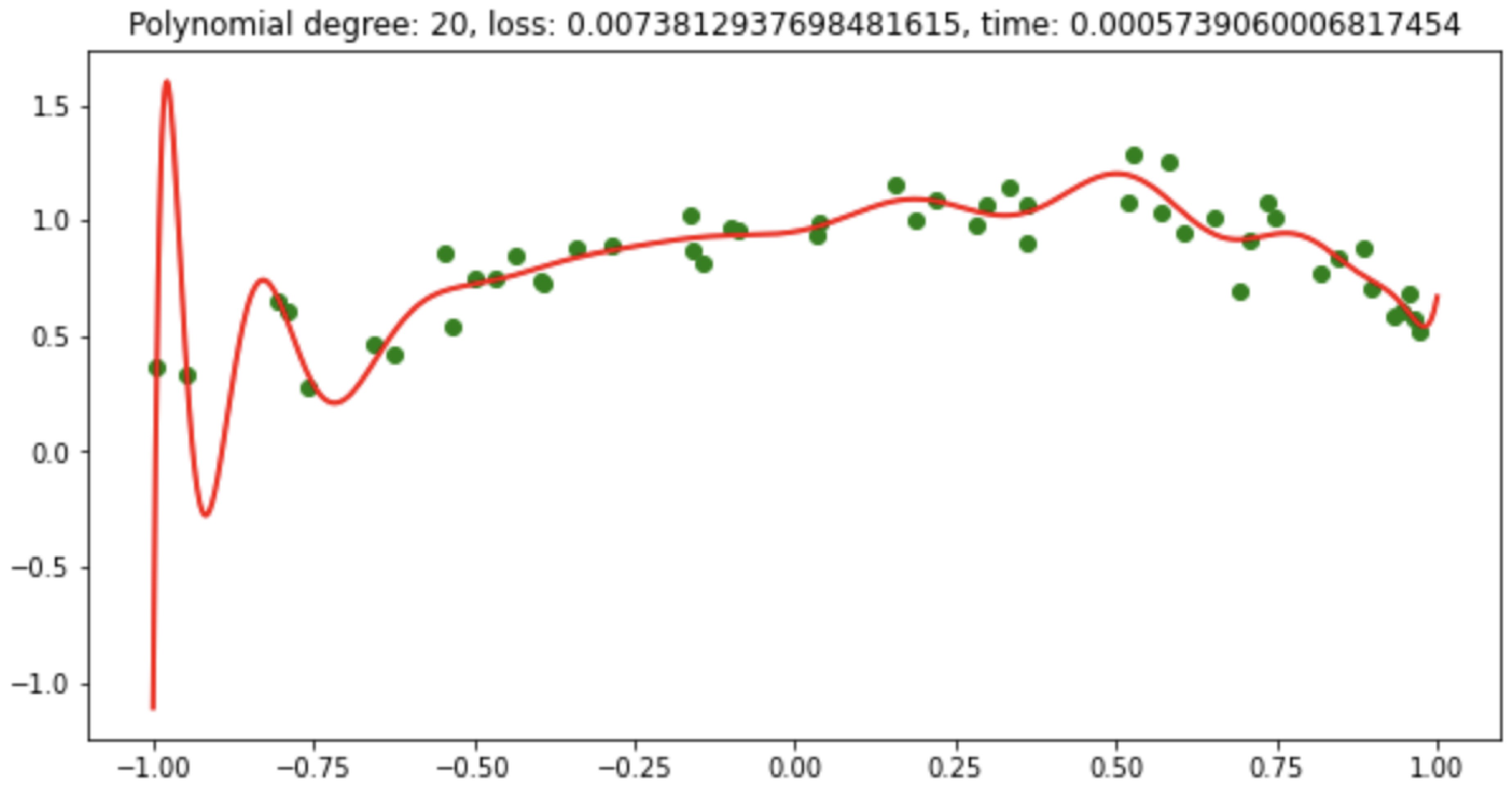
Understanding Overfitting

Recall: Overfitting happens when the model is fitting to the noise in the training data.

So, when overfitting happens, different training datasets can result in very different optimal parameter values, which often result in wrong predictions for some inputs. (See previous Colab notebook)

Ideally, we would like to use a family of models that would typically produce accurate predictions on unseen testing data, regardless of the particular training dataset.





Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x}; \theta^*(\mathcal{D}, L)) - y)^2|\vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L))|\vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\underbrace{\vec{x}}_{\text{Unseen Testing Example}}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x};\theta^*(\mathcal{D},L)) - y)^2|\vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x};\theta^*(\mathcal{D},L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x};\theta^*(\mathcal{D},L))|\vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, \underbrace{f}_{\text{Model}}, L) = E_{y,\mathcal{D}}[(f(\vec{x}; \theta^*(\mathcal{D}, L)) - y)^2 | \vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right) \middle| \vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L)) | \vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, \underbrace{L}_{\text{Loss Function}}) = E_{y,\mathcal{D}}[(f(\vec{x};\theta^*(\mathcal{D},L)) - y)^2|\vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x};\theta^*(\mathcal{D},L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x};\theta^*(\mathcal{D},L))|\vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x}; \theta^*(\underbrace{\mathcal{D}}_{\text{Training Dataset}}, L)) - y)^2 | \vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L))|\vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x}; \underbrace{\theta^*(\mathcal{D}, L)}_{\text{Optimal parameters on training data}}) - y)^2 | \vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L))|\vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(\underbrace{f(\vec{x}; \theta^*(\mathcal{D}, L))}_{\text{Prediction of trained model on new testing example}} - y)^2 | \vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L))|\vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x}; \theta^*(\mathcal{D}, L)) - \underbrace{y}_{\text{(Possibly noisy) label corresponding to testing example}})^2 | \vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L))|\vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[\underbrace{(f(\vec{x};\theta^*(\mathcal{D},L)) - y)^2}_{\text{Mean squared error (MSE) on testing example}}|\vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x};\theta^*(\mathcal{D},L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x};\theta^*(\mathcal{D},L))|\vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, L) = \underbrace{E_{y,\mathcal{D}}[(f(\vec{x}; \theta^*(\mathcal{D}, L)) - y)^2 | \vec{x}]}_{\text{Testing error averaged over training datasets compared to noisy versions of the label}}$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L))|\vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x};\theta^*(\mathcal{D},L)) - y)^2|\vec{x}]$$ Expected Error

$$= \Big(\underbrace{E_{\mathcal{D}}\big[f\big(\vec{x};\theta^*(\mathcal{D},L)\big)\big|\vec{x}\big]}_{\text{Prediction of the trained model on testing example, averaged over training datasets}} - E_y[y|\vec{x}]\Big)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x};\theta^*(\mathcal{D},L))|\vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x}; \theta^*(\mathcal{D}, L)) - y)^2 | \vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right)\middle|\vec{x}\right] - \underbrace{E_y[y|\vec{x}]}_{\text{Average of noisy versions of the label for the testing example}}\right)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L))|\vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x}; \theta^*(\mathcal{D}, L)) - y)^2 | \vec{x}]$$ Expected Error

$$= \underbrace{\left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right) \middle| \vec{x}\right] - E_y[y|\vec{x}]\right)^2}_{\text{Squared difference between average prediction and average label}}$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L)) | \vec{x})$$ Variance

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x}; \theta^*(\mathcal{D}, L)) - y)^2 | \vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared

$$+\underbrace{\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L))|\vec{x})}$$ Variance

Variance of the prediction on the testing example over different training datasets

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error

Bias-Variance Decomposition

$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x}; \theta^*(\mathcal{D}, L)) - y)^2 | \vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared

$$+\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L))|\vec{x})$$ Variance

$$+\underbrace{\mathrm{Var}(y|\vec{x})}_{\text{Variance of different noisy versions of the label of the testing example}}$$ Irreducible Error

Bias-Variance Tradeoff

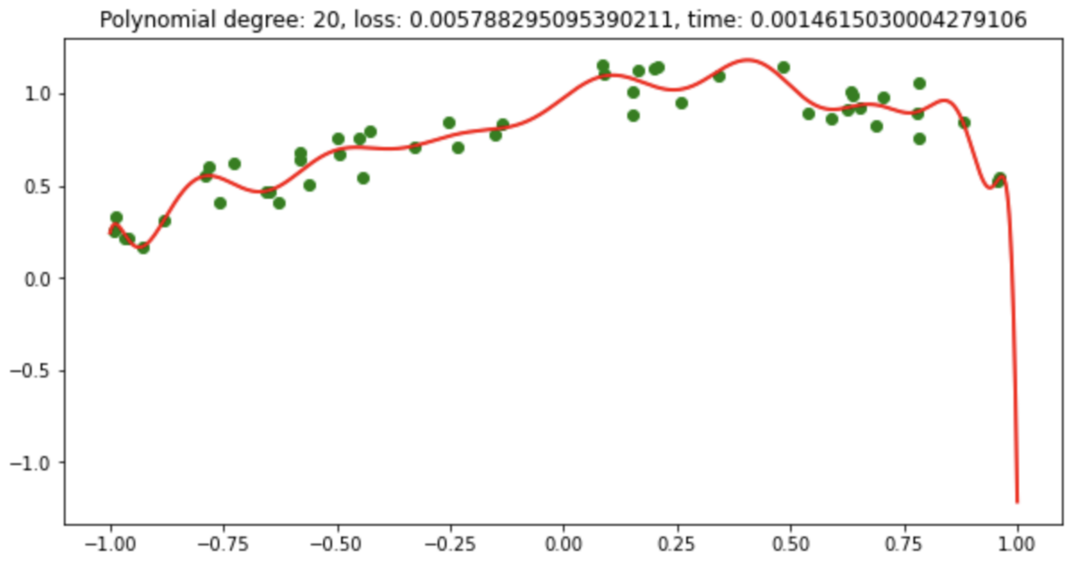
When overfitting:

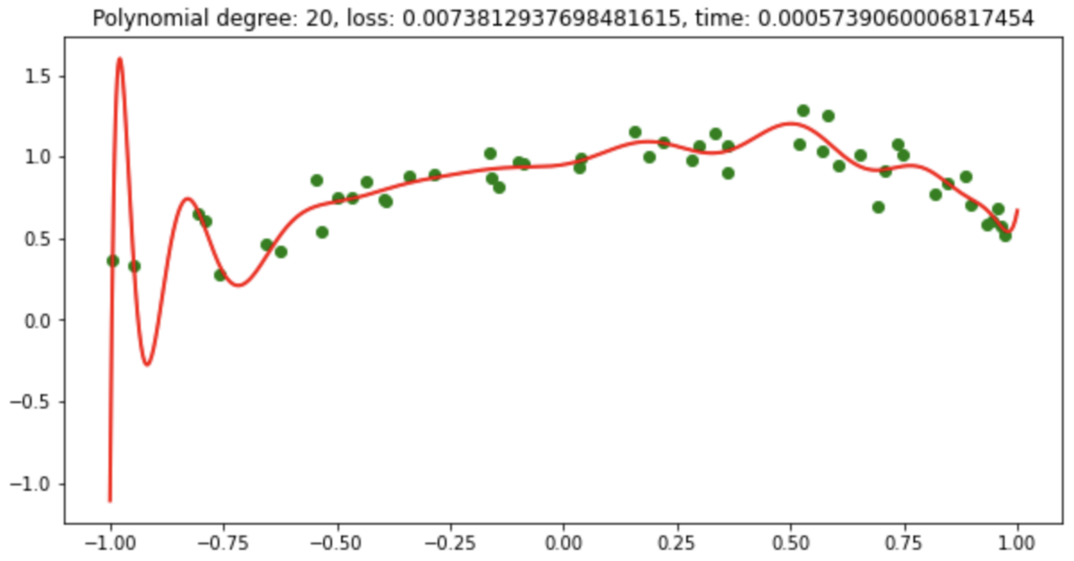
$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x};\theta^*(\mathcal{D},L)) - y)^2|\vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x};\theta^*(\mathcal{D},L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared: Small

$$+\mathrm{Var}(f(\vec{x};\theta^*(\mathcal{D},L))|\vec{x})$$ Variance: Large

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error





Bias-Variance Tradeoff

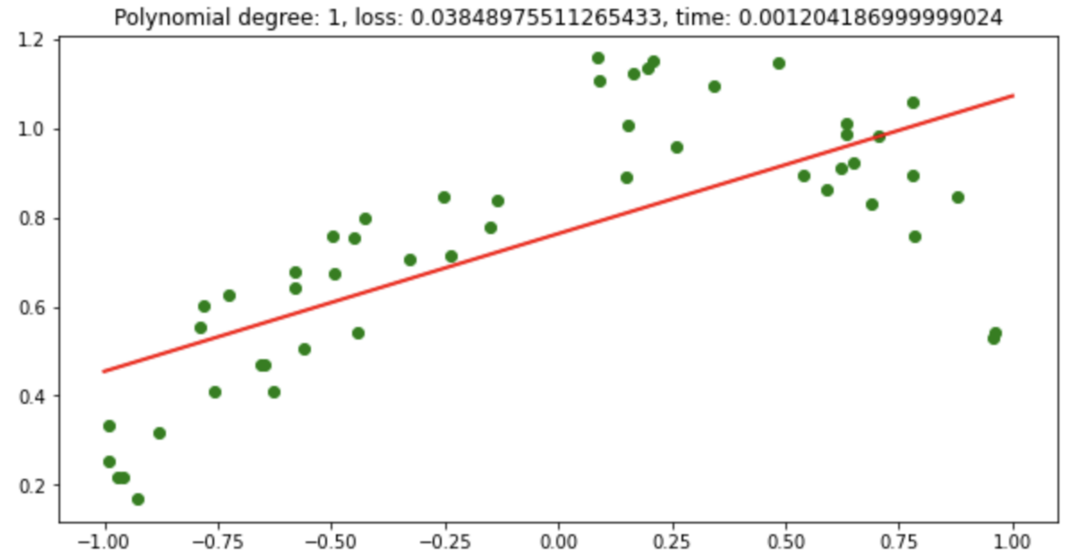
When underfitting:

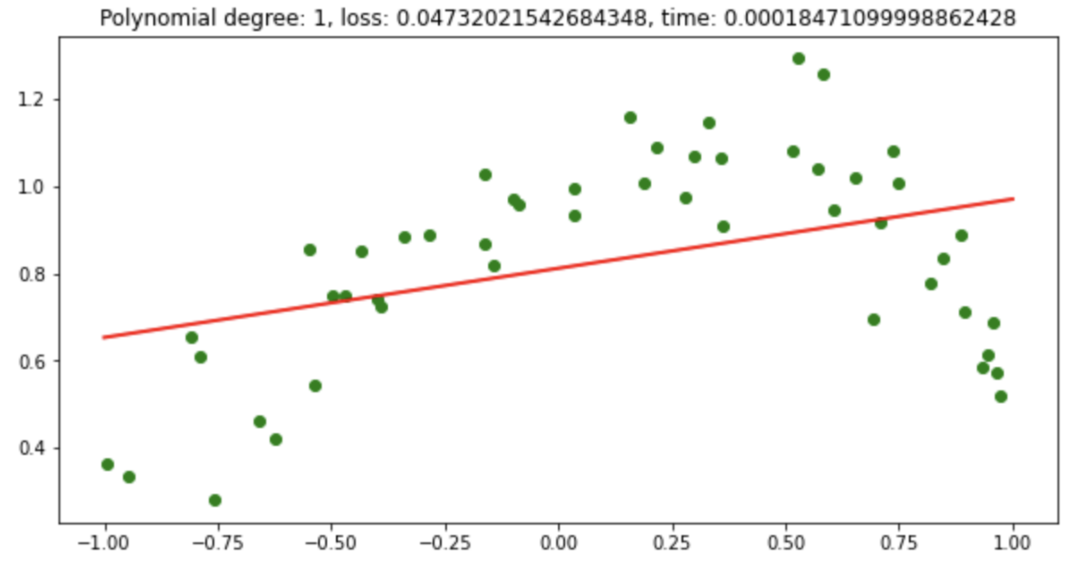
$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x}; \theta^*(\mathcal{D}, L)) - y)^2 | \vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared: Large

$$+\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L))|\vec{x})$$ Variance: Small

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error





Bias-Variance Tradeoff

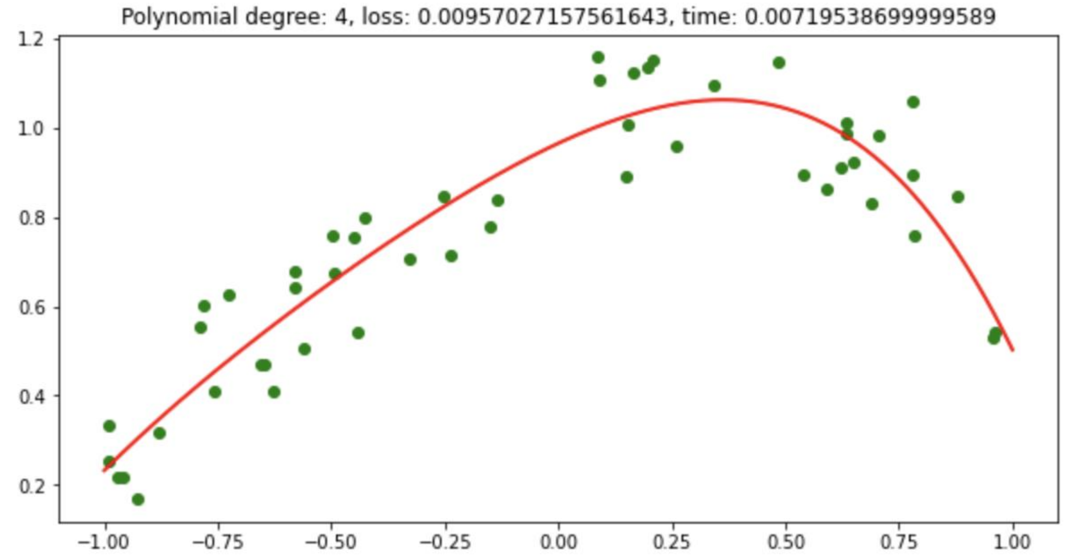
When fitted just right:

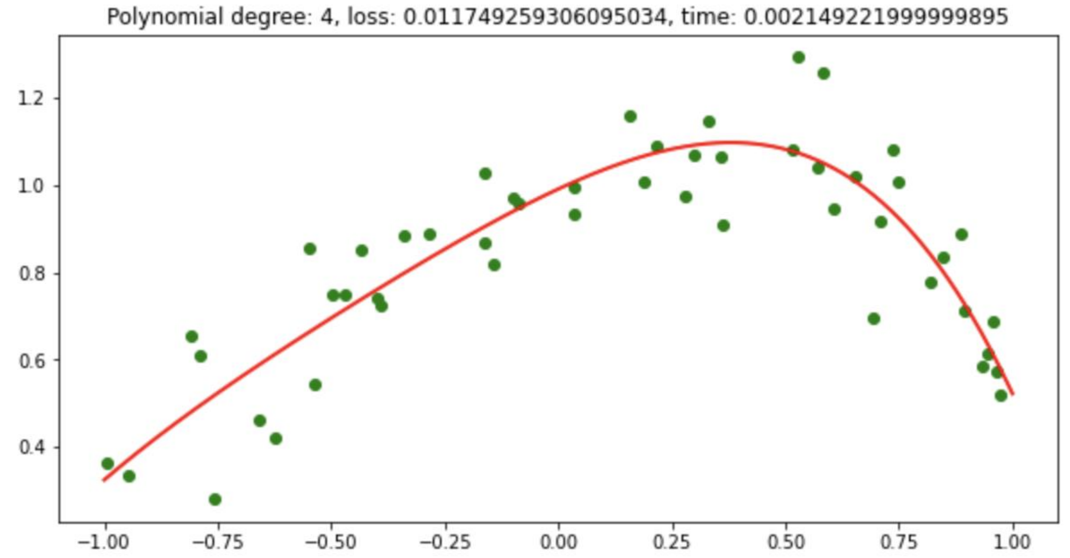
$$\epsilon(\vec{x}, f, L) = E_{y,\mathcal{D}}[(f(\vec{x}; \theta^*(\mathcal{D}, L)) - y)^2 | \vec{x}]$$ Expected Error

$$= \left(E_{\mathcal{D}}\left[f\left(\vec{x}; \theta^*(\mathcal{D}, L)\right)\middle|\vec{x}\right] - E_y[y|\vec{x}]\right)^2$$ Bias Squared: Small

$$+\mathrm{Var}(f(\vec{x}; \theta^*(\mathcal{D}, L))|\vec{x})$$ Variance: Small

$$+\mathrm{Var}(y|\vec{x})$$ Irreducible Error





Takeaways

Expected Error on Testing Example = Bias^2 + Variance + Irreducible Error

Bias: Difference between average prediction and average label of the testing example

Variance: Variance of the prediction on the testing example over different training datasets

Both bias and variance must be low in order for the expected error to be low

When overfitting, bias is low and variance is high

When underfitting, bias is high and variance is low

Simply achieving low bias (which happens when the model is very expressive) isn’t enough!