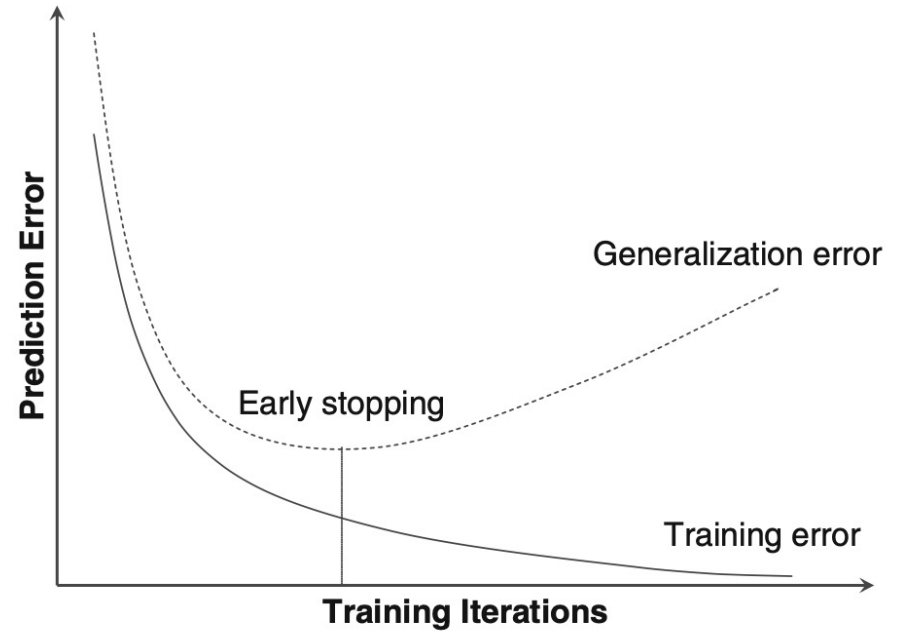
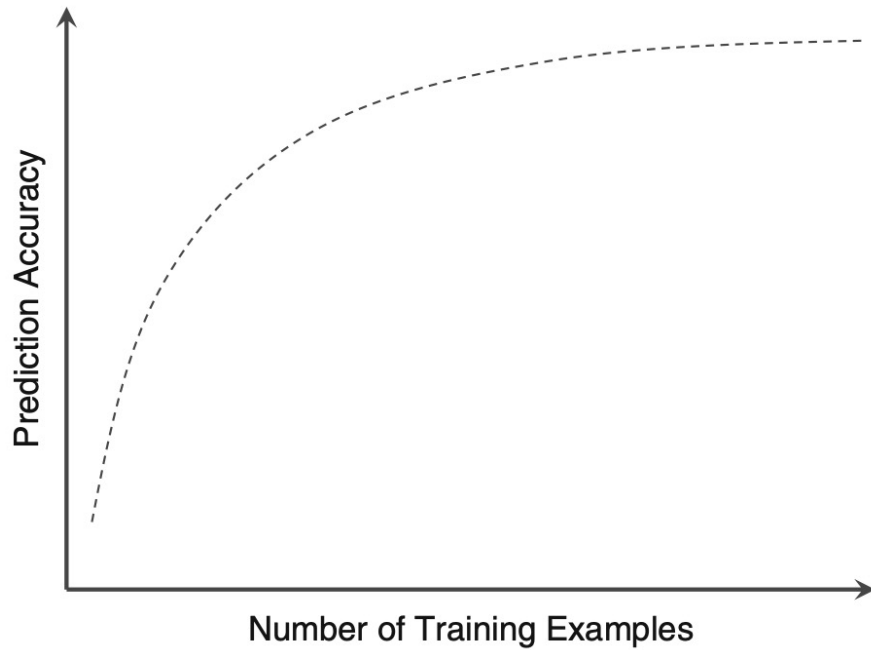
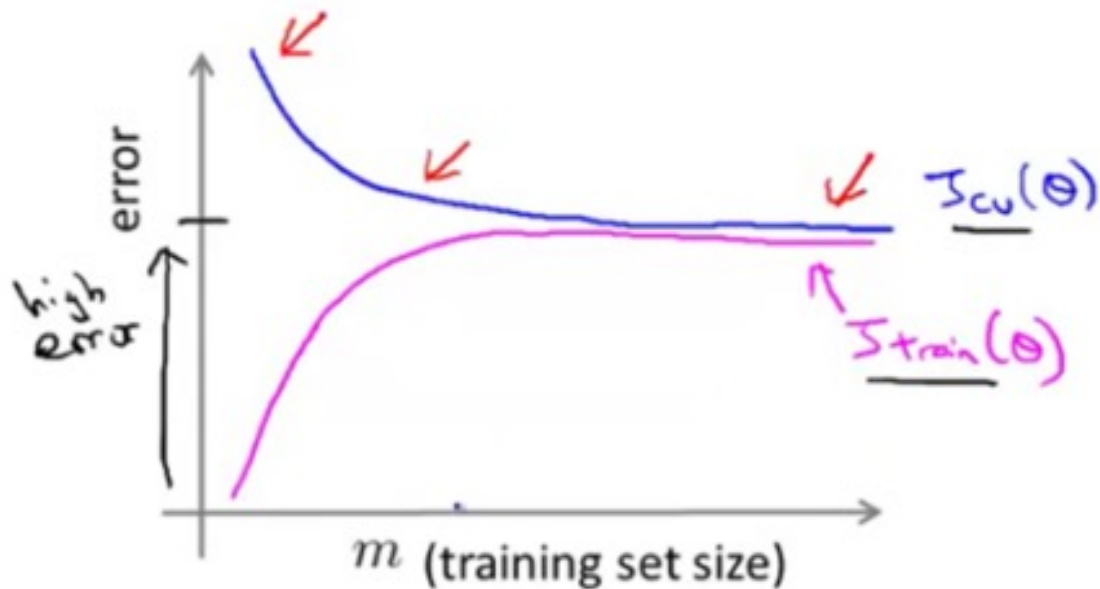


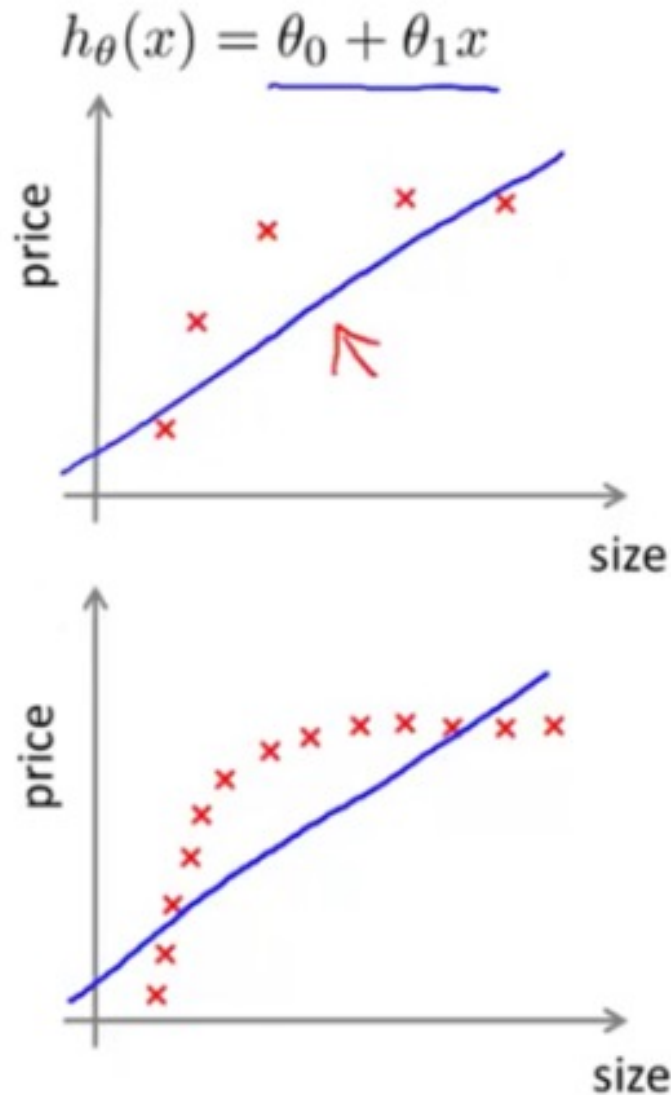
Encyclopedia of Machine Learning and Data Mining - 2nd



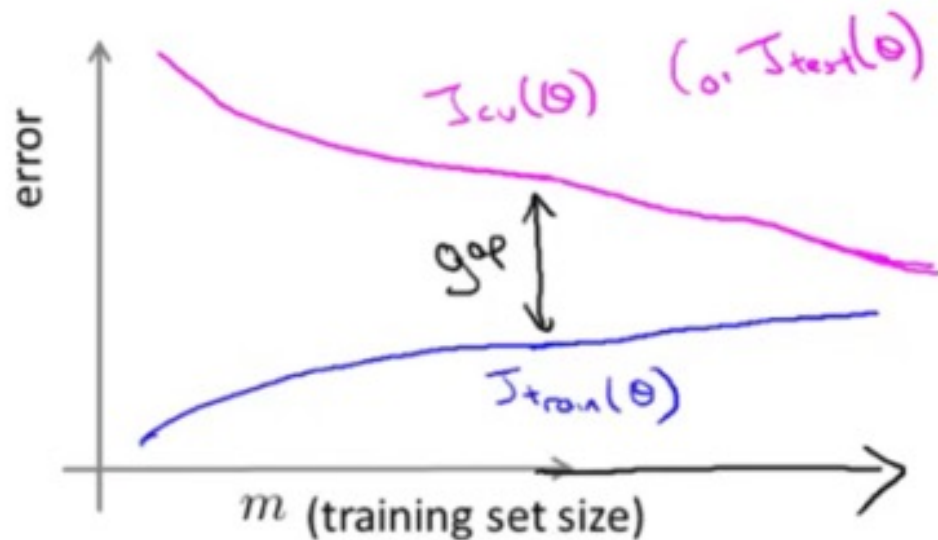
High bias



If a learning algorithm is suffering from high bias, getting more training data will not (by itself) help much.

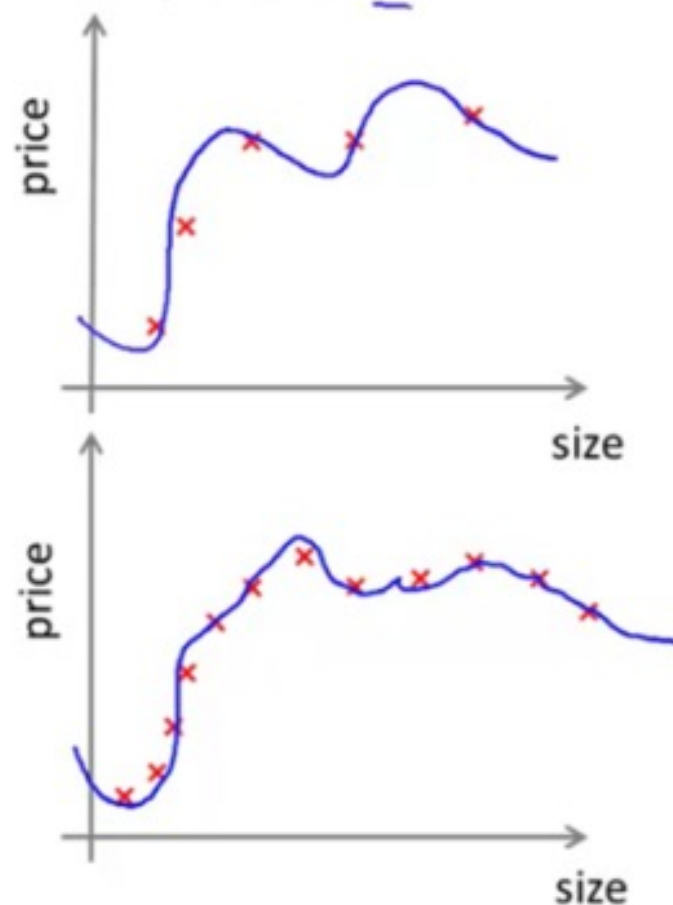


High variance



If a learning algorithm is suffering from high variance, getting more training data is likely to help. \leftarrow

$$h_{\theta}(x) = \theta_0 + \theta_1 x + \dots + \theta_{100} x^{100} \quad (\text{and small } \lambda) \quad \nwarrow$$



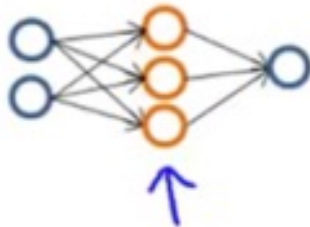
Debugging a learning algorithm:

Suppose you have implemented regularized linear regression to predict housing prices. However, when you test your hypothesis in a new set of houses, you find that it makes unacceptably large errors in its prediction. What should you try next?

- Get more training examples \rightarrow fixes high variance
- Try smaller sets of features \rightarrow fixes high variance
- Try getting additional features \rightarrow fixes high bias
- Try adding polynomial features (x_1^2, x_2^2, x_1x_2 , etc) \rightarrow fixes high bias.
- Try decreasing λ \rightarrow fixes high bias
- Try increasing λ \rightarrow fixes high variance

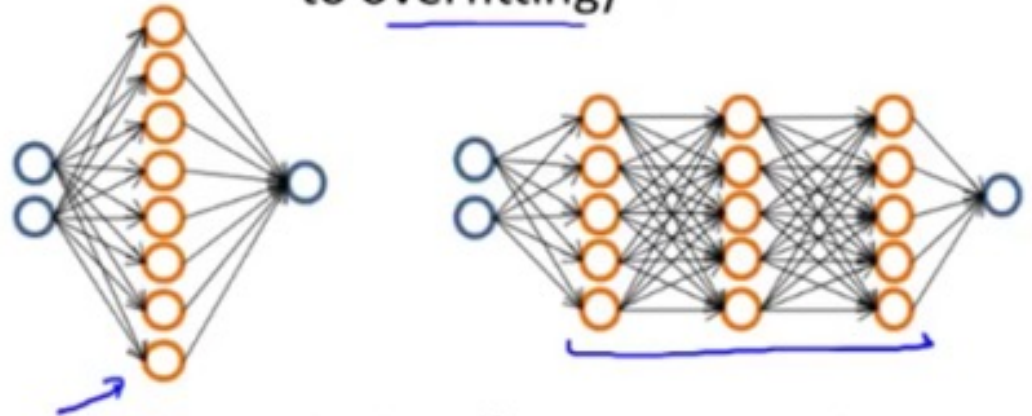
Neural networks and overfitting

→ “Small” neural network
(fewer parameters; more prone to underfitting)



Computationally cheaper

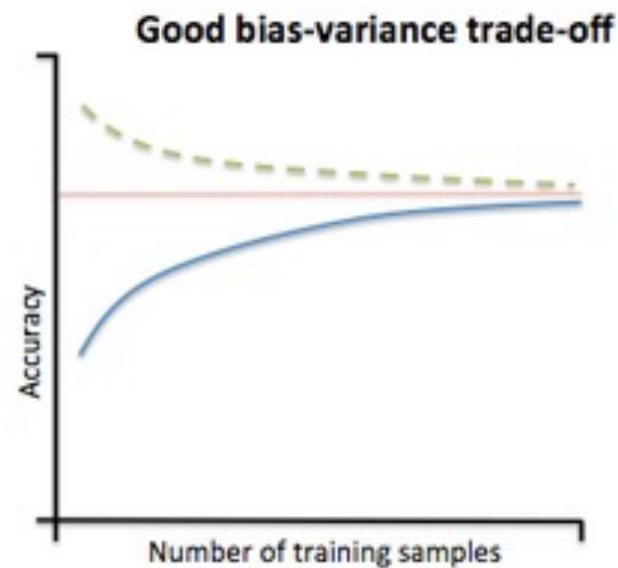
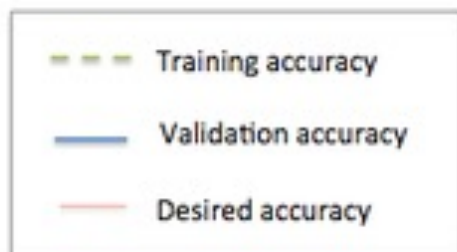
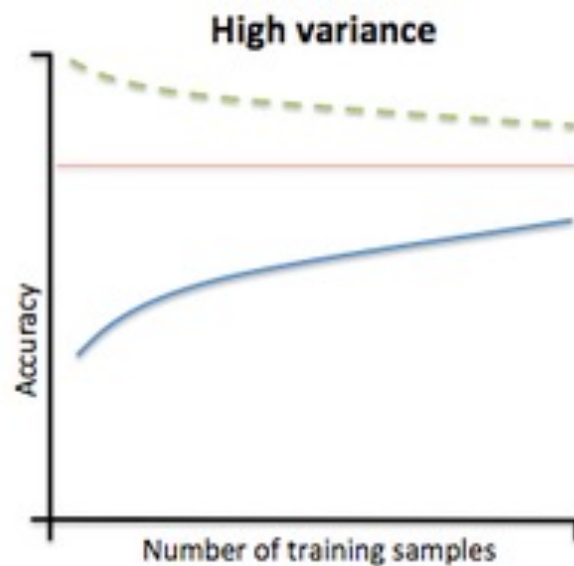
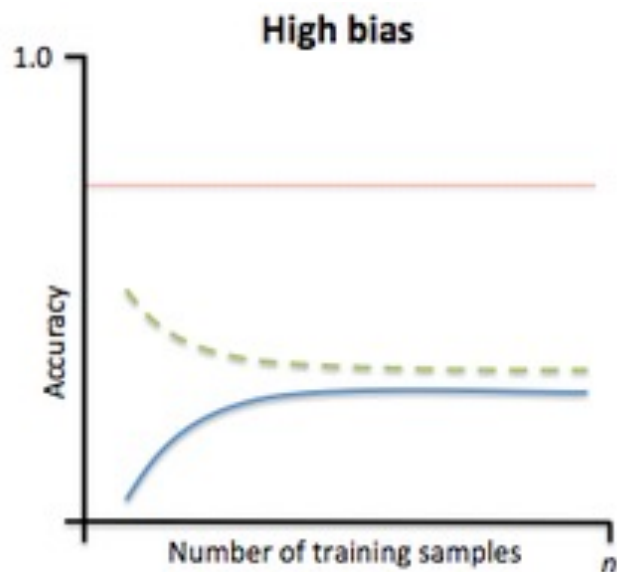
→ “Large” neural network
(more parameters; more prone to overfitting)



Computationally more expensive.

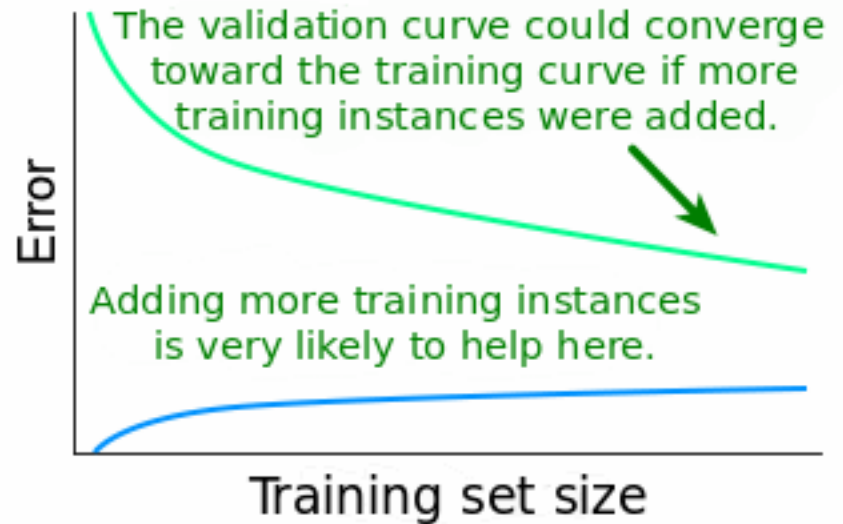
Use regularization (λ) to address overfitting.
↑

<https://sebastianraschka.com/faq/docs/ml-solvable.html>

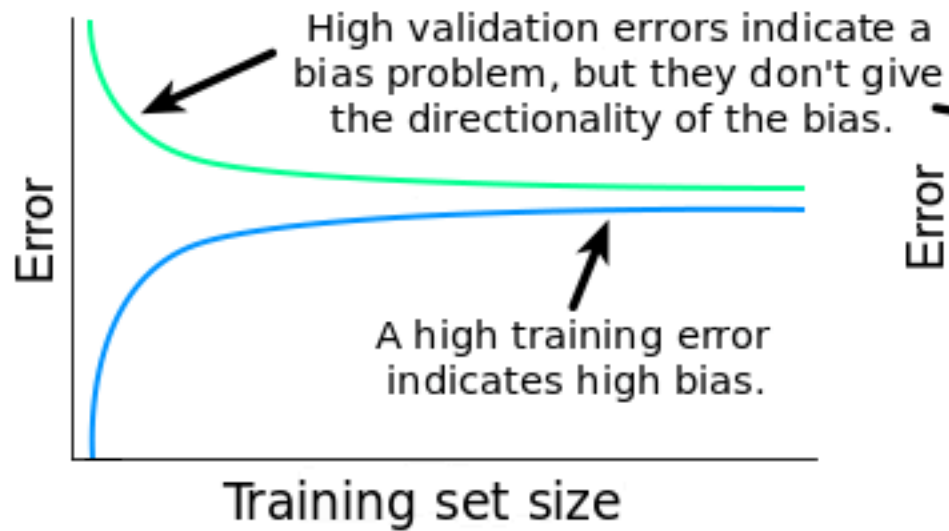


- our data is skewed
- there is a lot of noise
- there are many outliers
- our features are not informative enough
- we don't have enough training samples

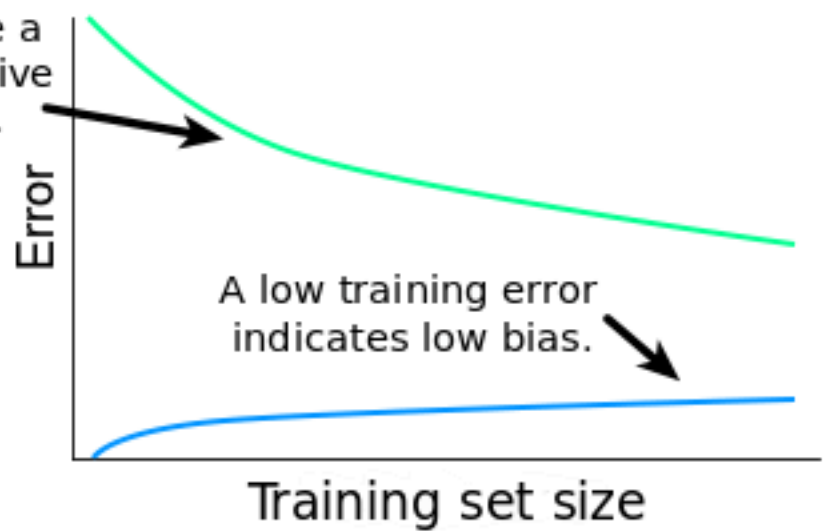
<https://www.dataquest.io/blog/learning-curves-machine-learning/>



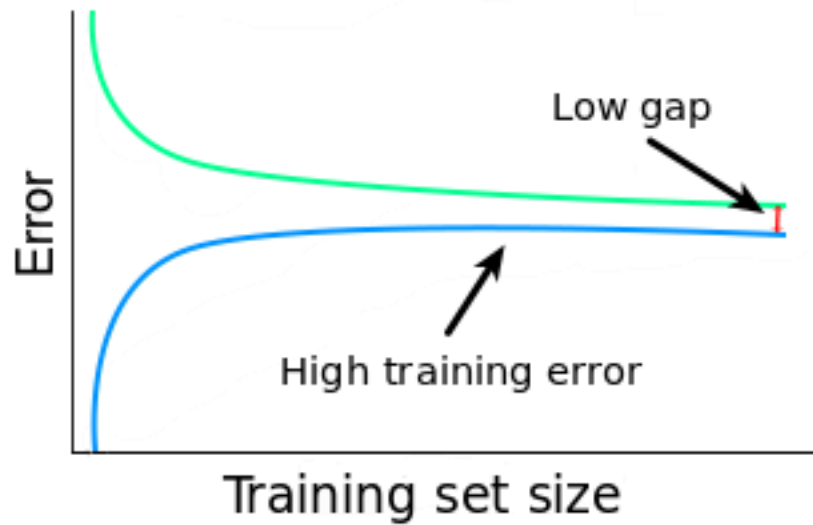
High base case



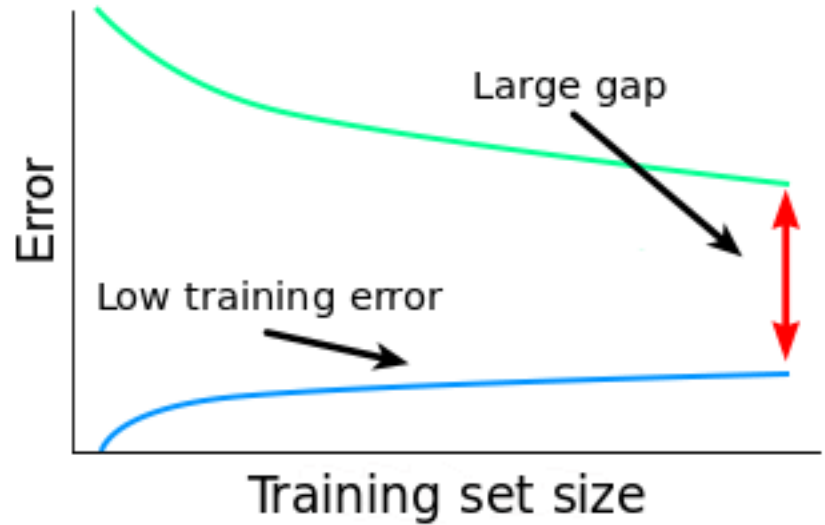
Low bias case

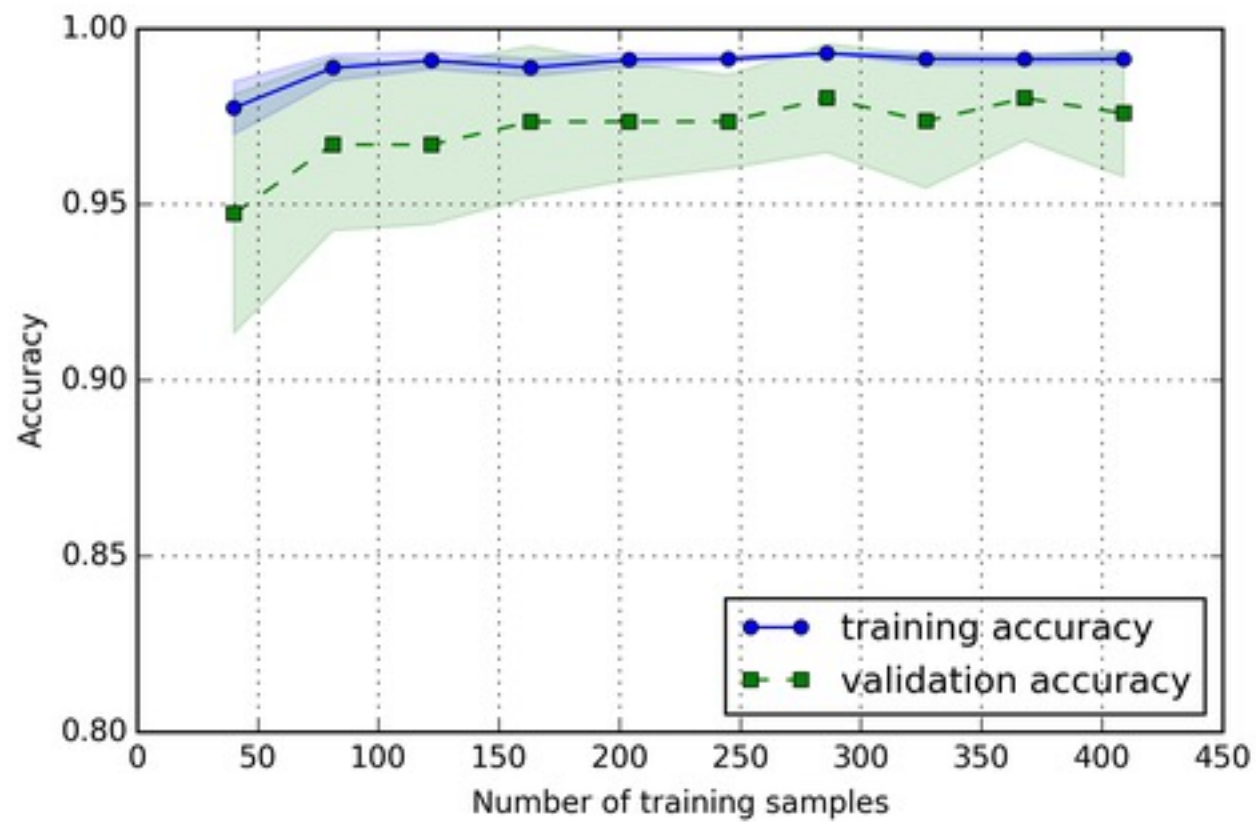


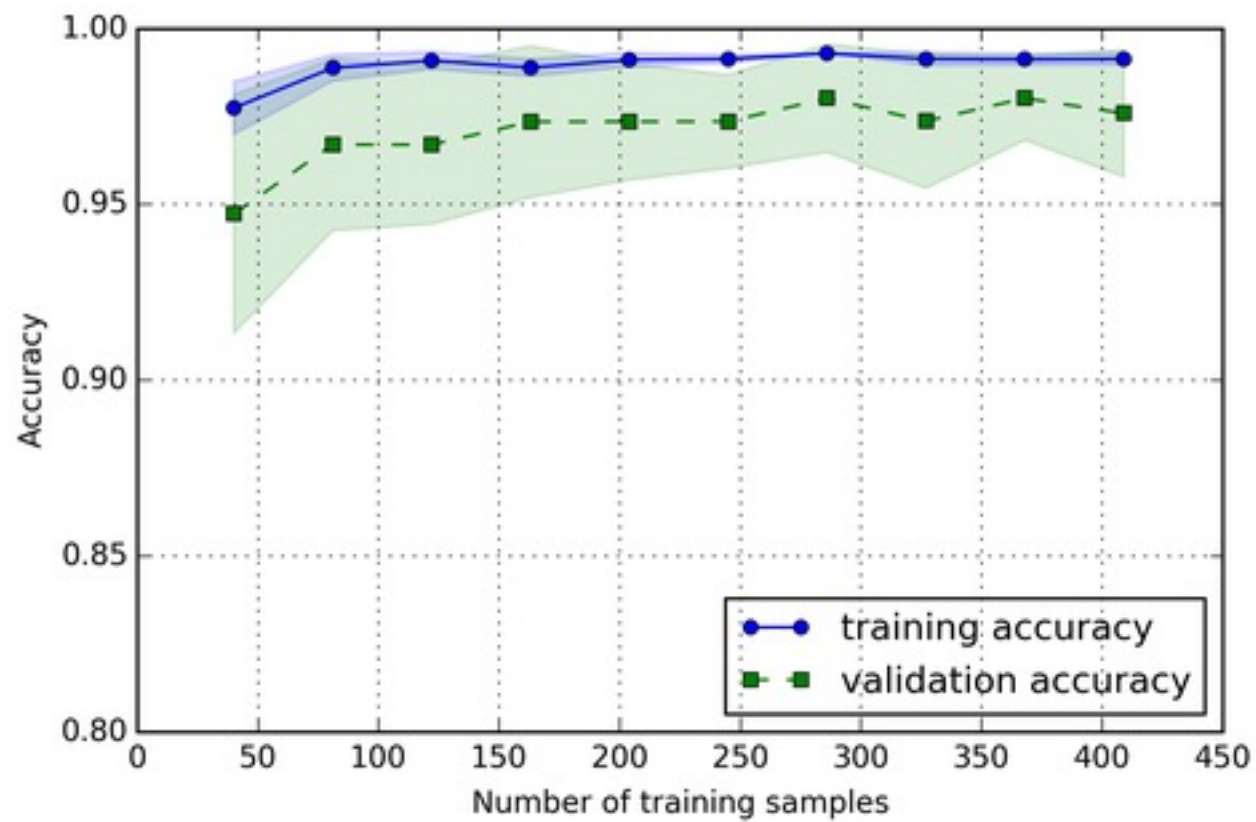
Low variance case



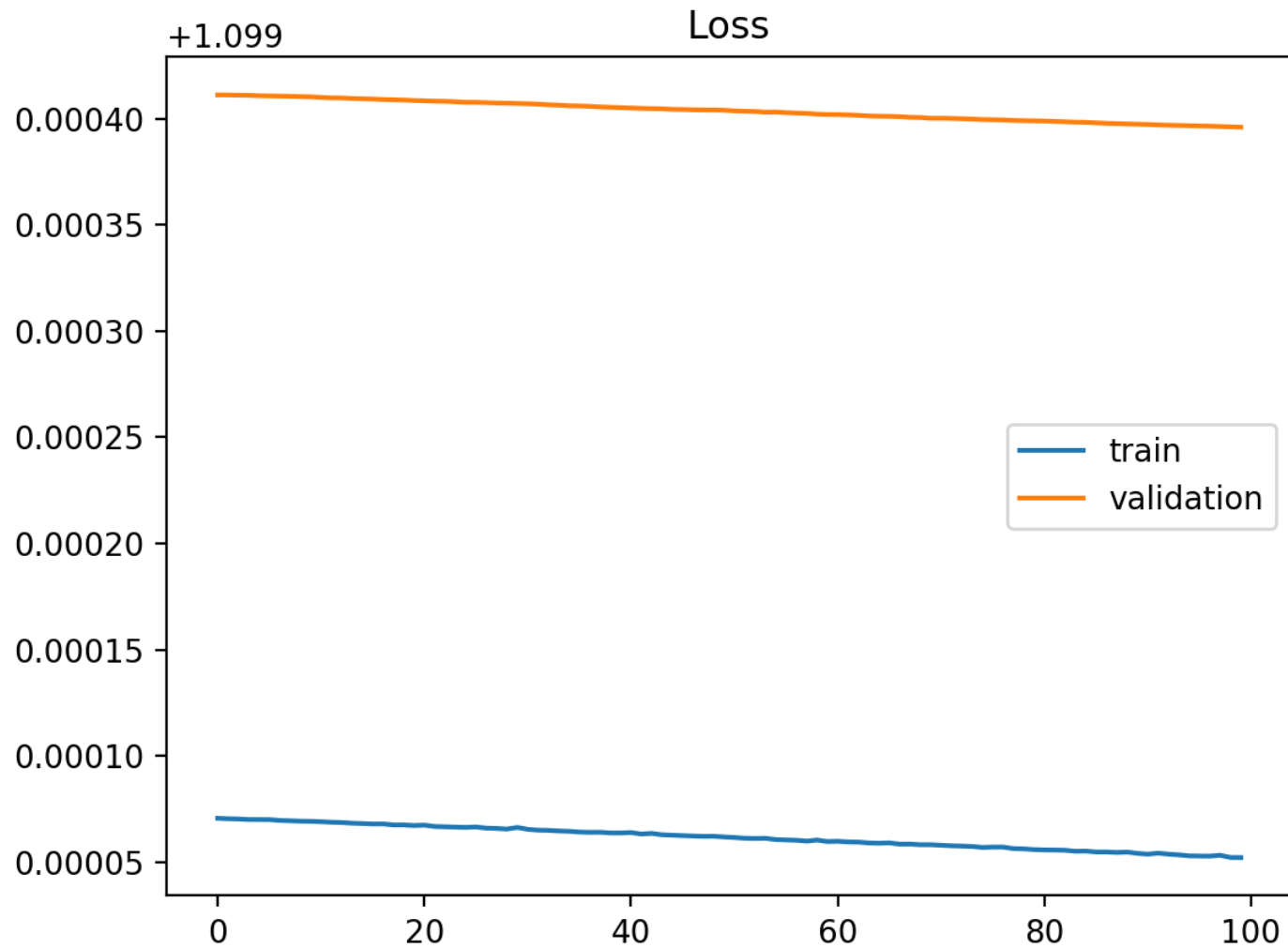
High variance case



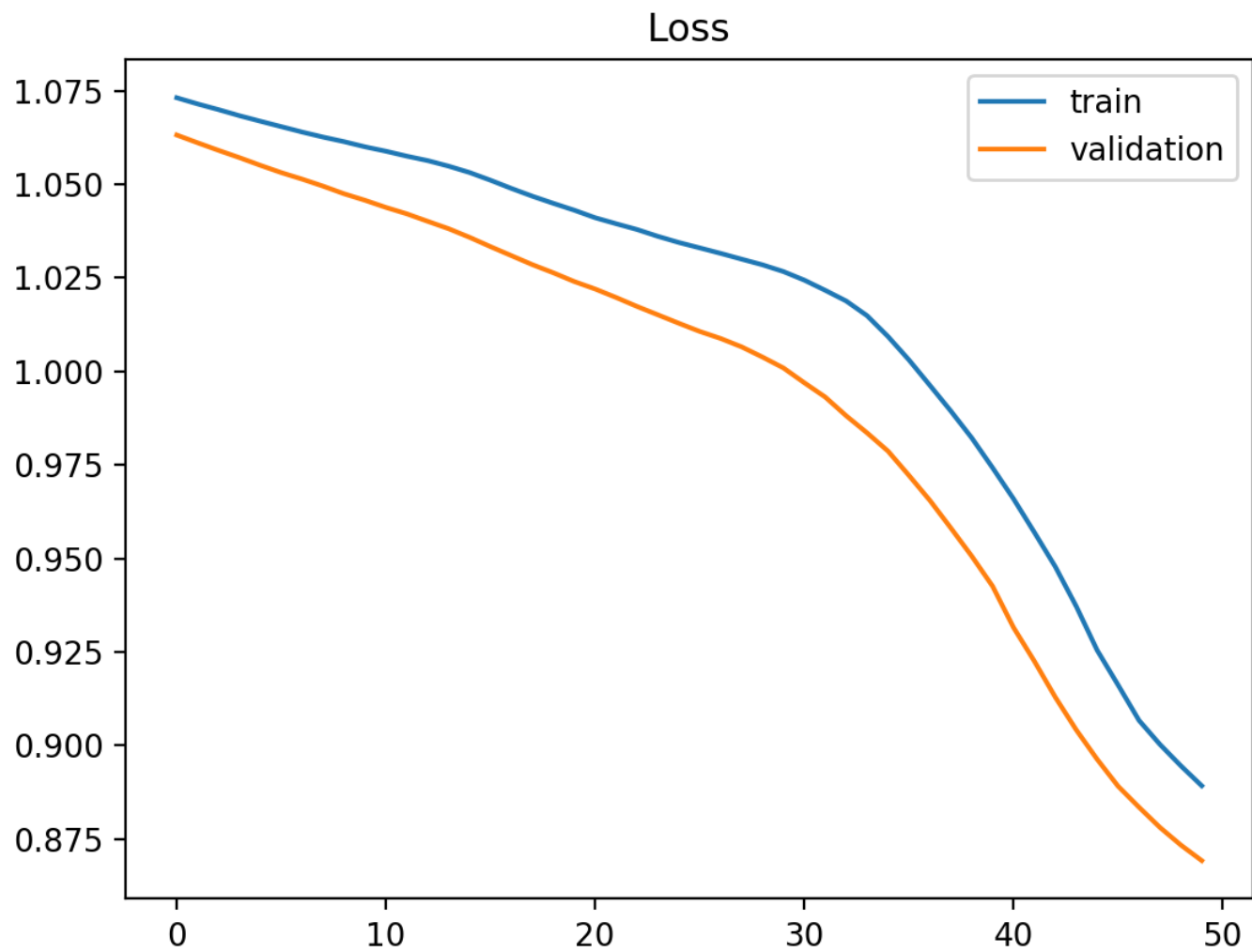




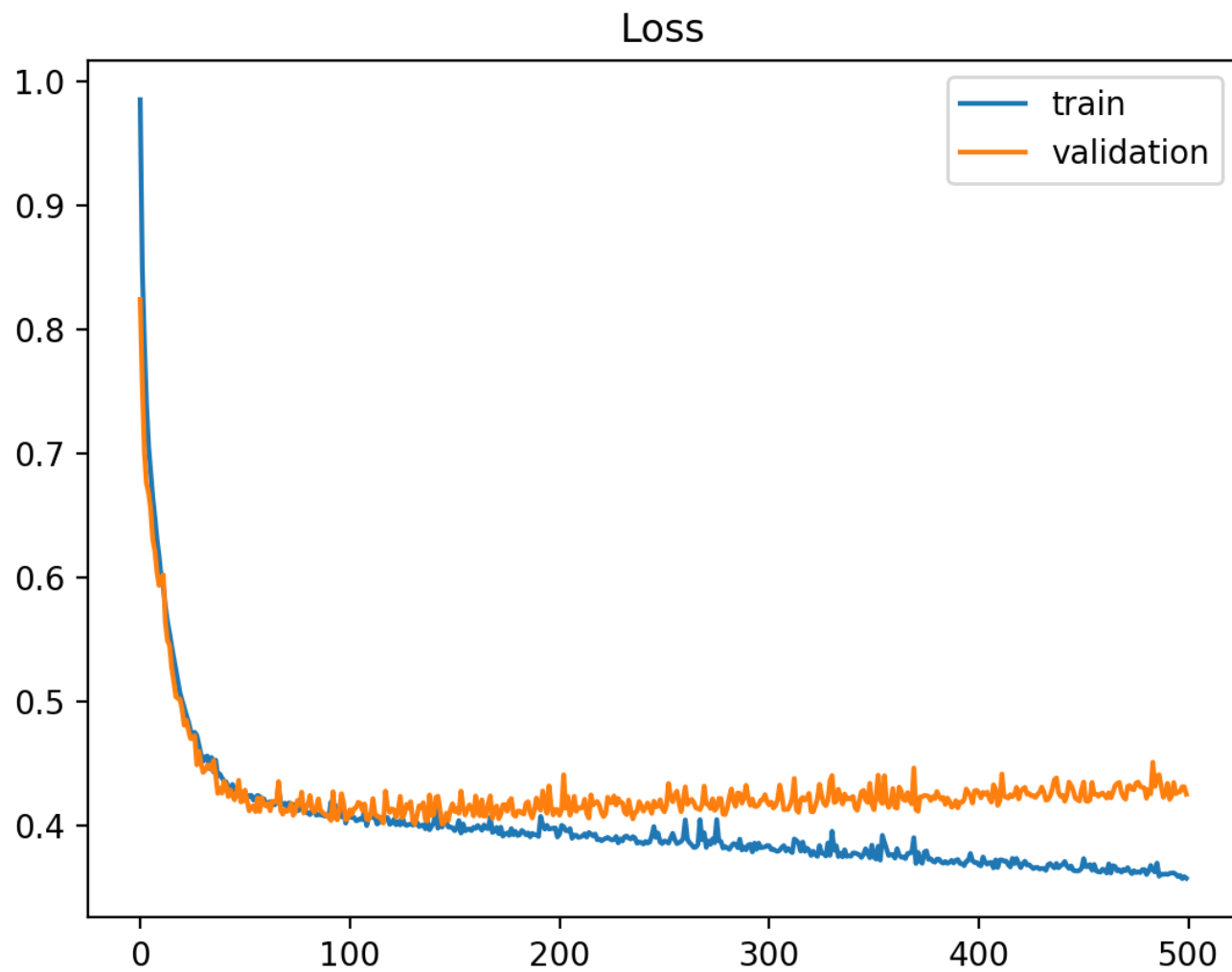
Underfit (1)



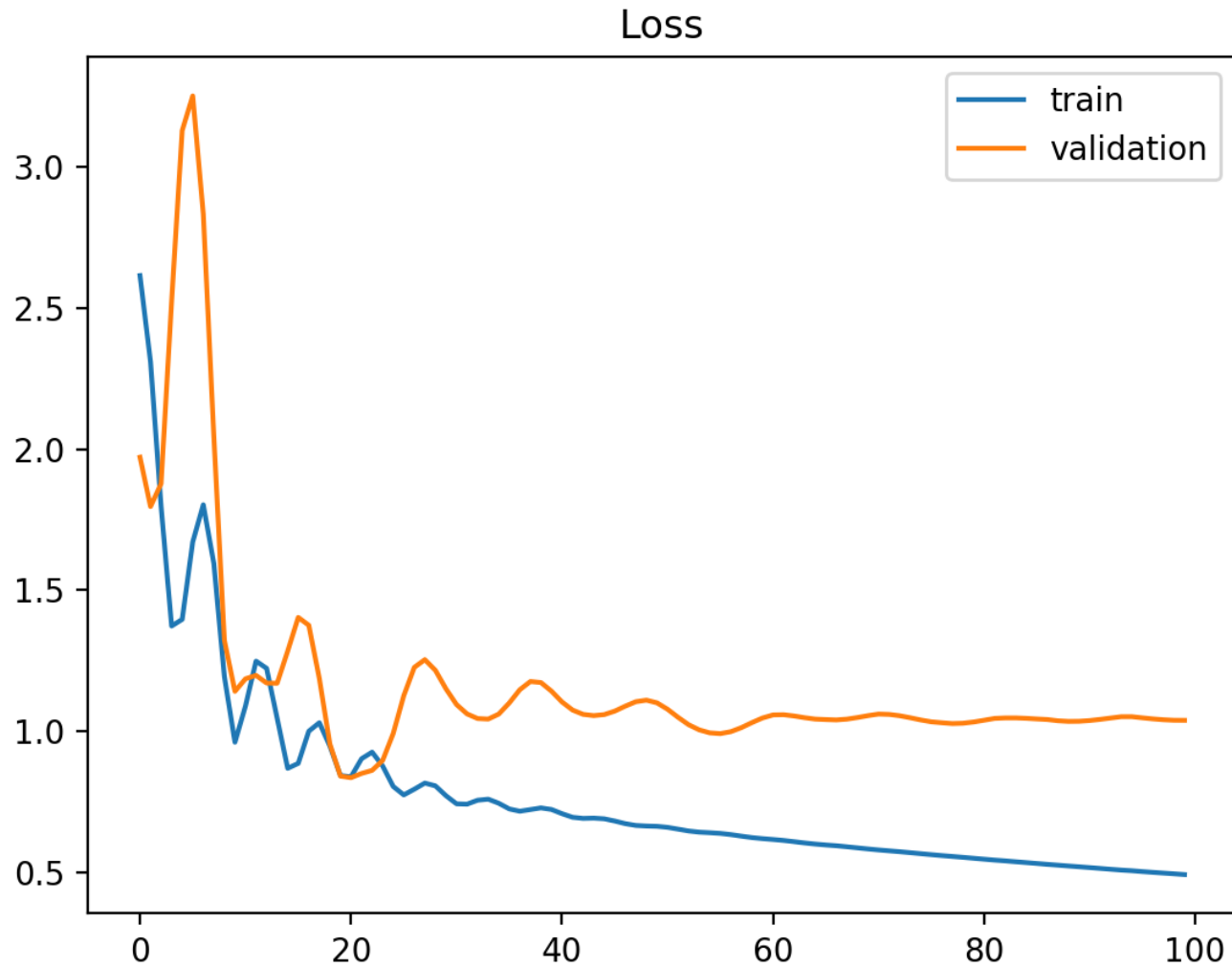
Underfit (2)



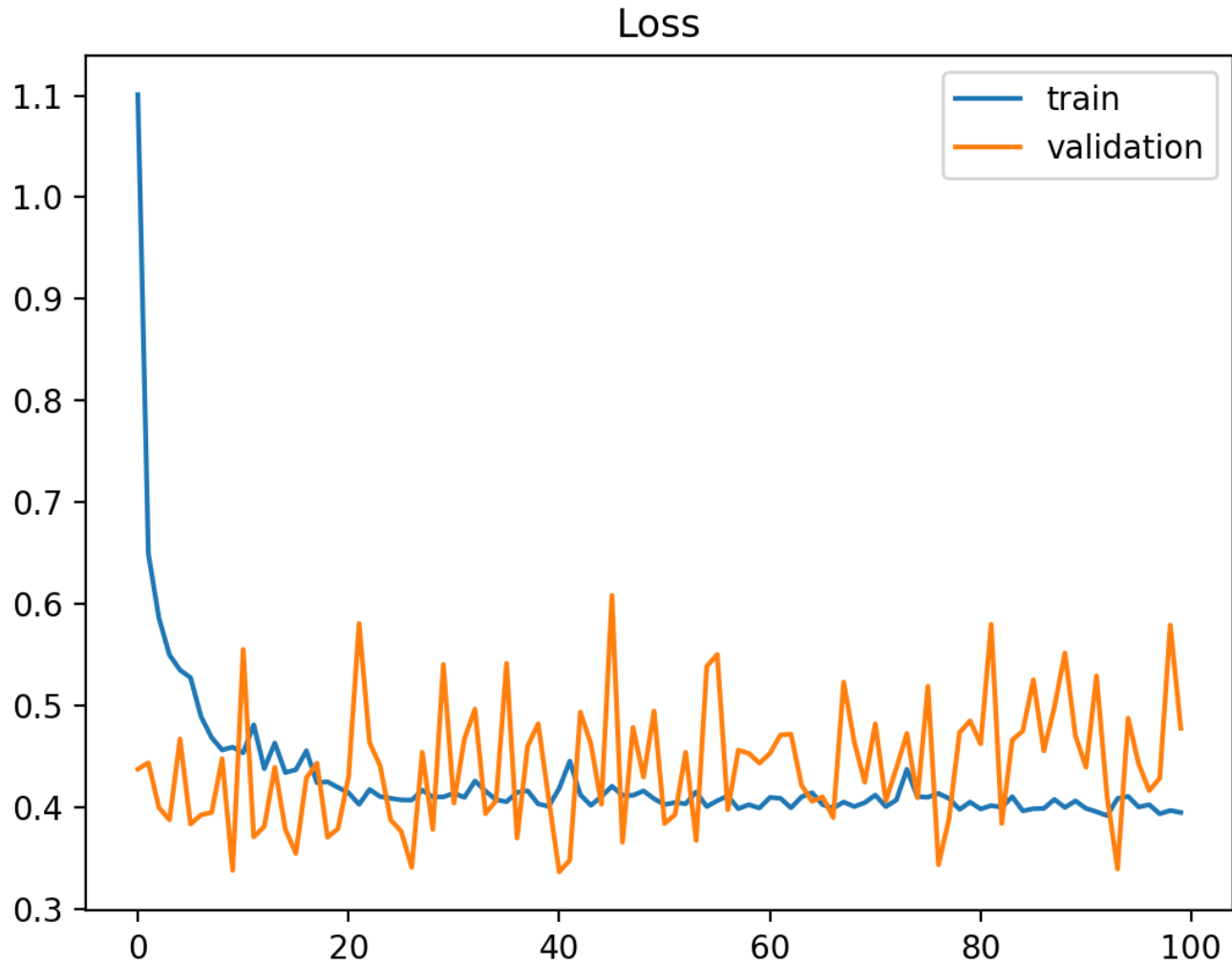
Overfit (1)



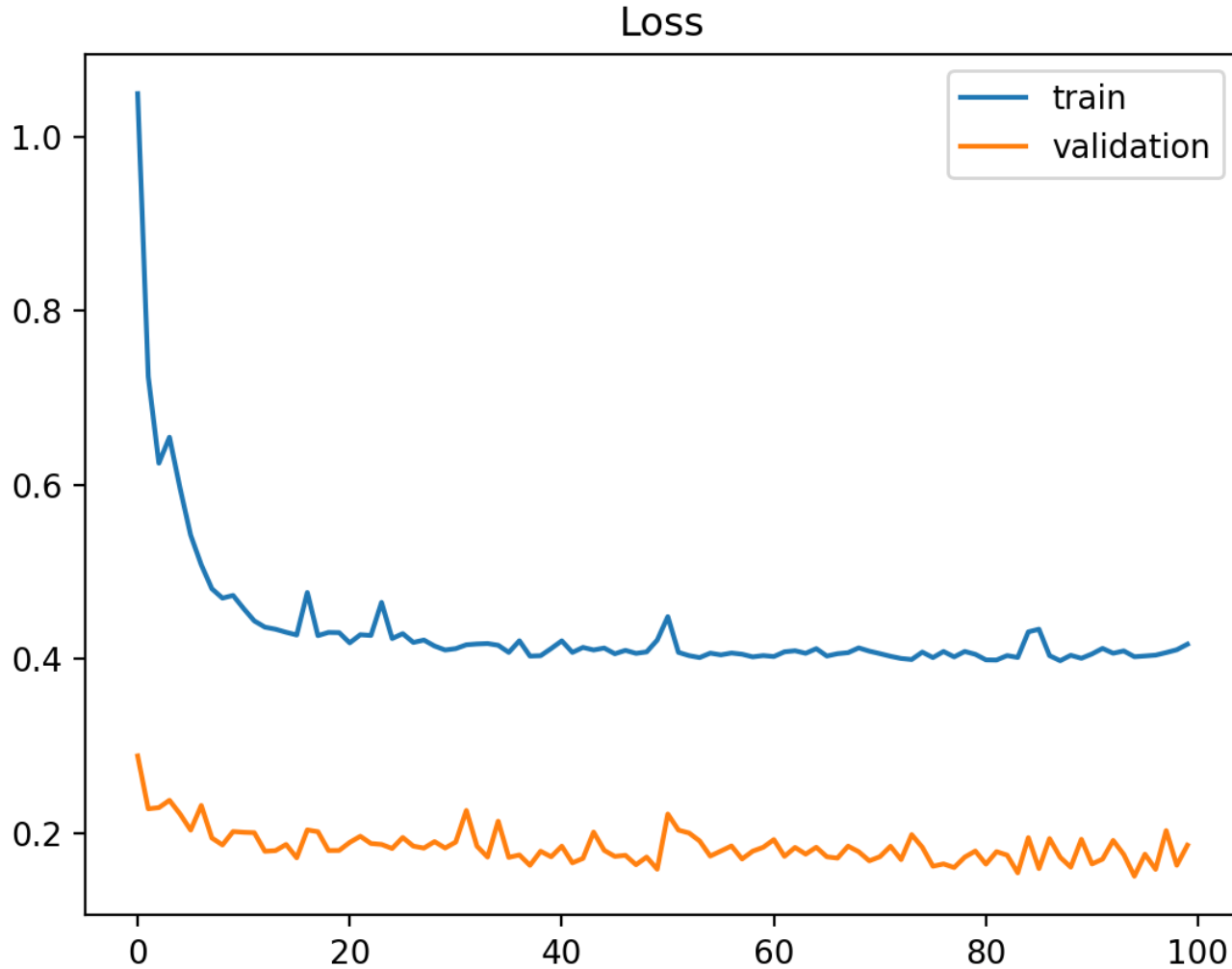
Unrepresentative Train Dataset



Unrepresentative Validation Dataset (1)



Unrepresentative Validation Dataset (2)



the validation dataset may be easier for the model to predict than the training dataset

