## Over-fitting and Regularization

In supervised machine learning, models are trained on a subset of data aka training data. The goal is to compute the target of each training example from the training data.



Now, overfitting happens when model learns signal as well as noise in the training data and wouldn’t perform well on new data on which model wasn’t trained on. In the example below, you can see underfitting in first few steps and overfitting in last few.

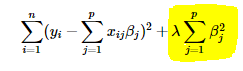
Now, there are few ways you can avoid overfitting your model on training data like cross-validation sampling, reducing number of features, pruning, regularization etc.

Regularization basically adds the penalty as model complexity increases. Regularization parameter (lambda) penalizes all the parameters except intercept so that model generalizes the data and won’t overfit.

*A regression model that uses L1 regularization technique is called****Lasso Regression****and model which uses L2 is called****Ridge Regression****.*

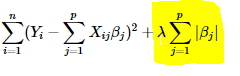
## The key difference between these two is the penalty term.

Ridge regression(L2) adds “squared magnitude” of coefficient as penalty term to the loss function. Here the highlighted part represents L2 regularization element.



Here, if *lambda* is zero then you can imagine we get back OLS. However, if *lambda* is very large then it will add too much weight and it will lead to under-fitting. Having said that it’s important how *lambda* is chosen. This technique works very well to avoid over-fitting issue.

**Lasso Regression(L1)** (Least Absolute Shrinkage and Selection Operator) adds “*absolute value of magnitude*” of coefficient as penalty term to the loss function.



The **key difference** between these techniques is that Lasso shrinks the less important feature’s coefficient to zero thus, removing some feature altogether. So, this works well for **feature selection** in case we have a huge number of features.

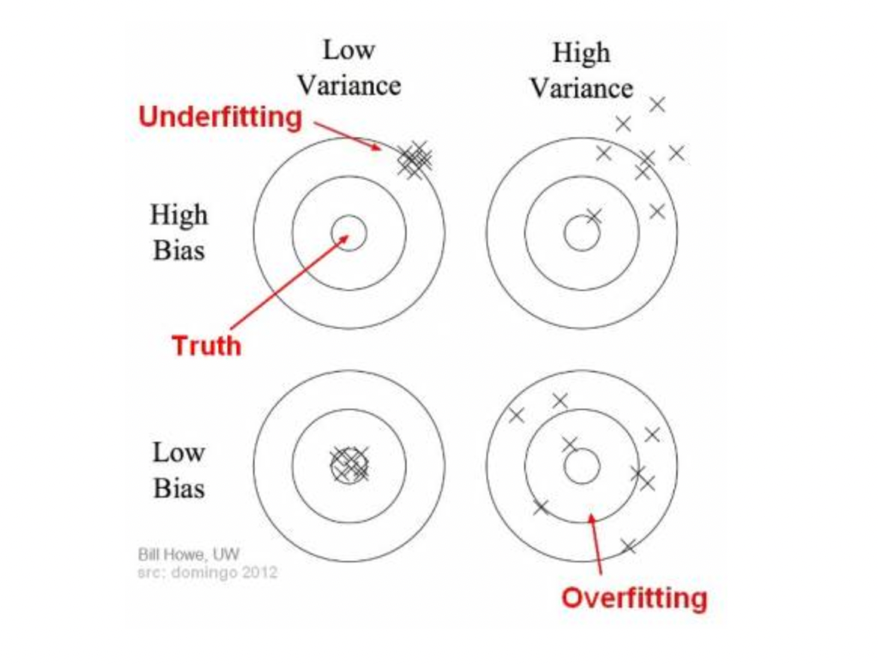


## What is bias?

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

## What is variance?

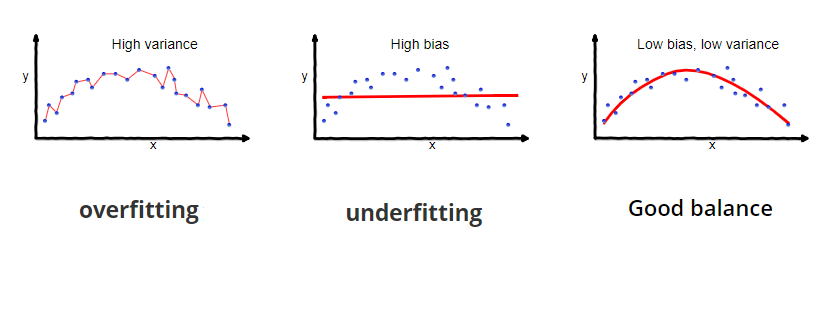
Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn’t seen before. As a result, such models perform very well on training data but has high error rates on test data.



In the above diagram, center of the target is a model that perfectly predicts correct values. As we move away from the bulls-eye our predictions become get worse and worse. We can repeat our process of model building to get separate hits on the target.

In supervised learning, **underfitting** happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance. It happens when we have very less amount of data to build an accurate model or when we try to build a linear model with a nonlinear data. Also, these kind of models are very simple to capture the complex patterns in data like Linear and logistic regression.

In supervised learning, **overfitting** happens when our model captures the noise along with the underlying pattern in data. It happens when we train our model a lot over noisy dataset. These models have low bias and high variance. These models are very complex like Decision trees which are prone to overfitting.



A model that exhibits small variance and high bias will underfit the target, while a model with high variance and little bias will overfit the target.

## Why is Bias Variance Tradeoff?

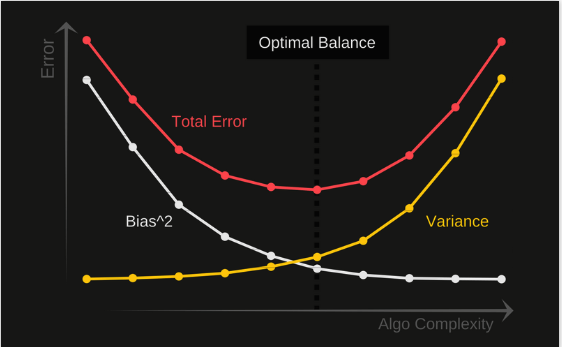
If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it’s going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time.

## Total Error

To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.





An optimal balance of bias and variance would never overfit or underfit the model.

Therefore understanding bias and variance is critical for understanding the behavior of prediction models.

Regularization works by [biasing](https://www.statisticshowto.com/what-is-bias/)data towards particular values (such as small values near zero). The bias is achieved by adding a*tuning parameter* to encourage those values:

* L1 regularization adds an L1 penalty equal to the [absolute value](https://www.statisticshowto.com/integer/#abs) of the magnitude of coefficients. In other words, it limits the size of the coefficients. L1 can yield sparse models (i.e. models with few coefficients); Some coefficients can become zero and eliminated. [Lasso regression](https://www.statisticshowto.com/lasso-regression/) uses this method.
* L2 regularization adds an L2 penalty equal to the square of the magnitude of coefficients. L2 will not yield sparse models and all coefficients are shrunk by the same factor (none are eliminated).
* L1 regularization is the preferred choice when having a high number of features as it provides sparse solutions. Even, we obtain the computational advantage because features with zero coefficients can be avoided.
* L2 regularization can deal with the multicollinearity (independent variables are highly correlated) problems through constricting the coefficient and by keeping all the variables.
* L2 regression can be used to estimate the significance of predictors and based on that it can penalize the insignificant predictors.

[Link](https://www.analyticssteps.com/blogs/l2-and-l1-regularization-machine-learning)