Princess Tara Zamani
Adv Opt for Machine Learning
Homework 2
Gradient Descent Implementation

The objective of this assignment was to apply gradient descent to solve the least-mean-square problem defined as:

$$min_{x \in R^n} f(x) = \frac{1}{2}\|Ax - b\|_2^2, \text{ where } x \in R^n, A \in R^{m \times n}, b \in R^m.$$

The gradient for the definition above is given as $\nabla f(x) = A^T(Ax - b)$.

The underlying signal (z), and the initialization of x_0 were generated randomly using the numpy.random.randn function. The initial noise variance was chosen to be 0.1.

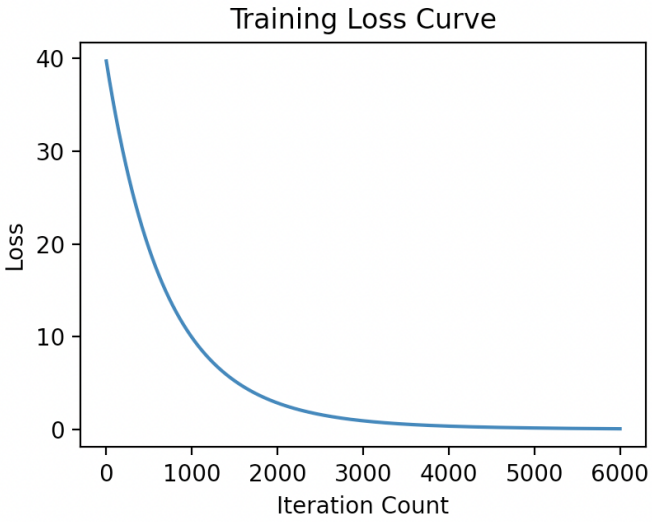
The gradient algorithm implementation is iterative and takes the following form:

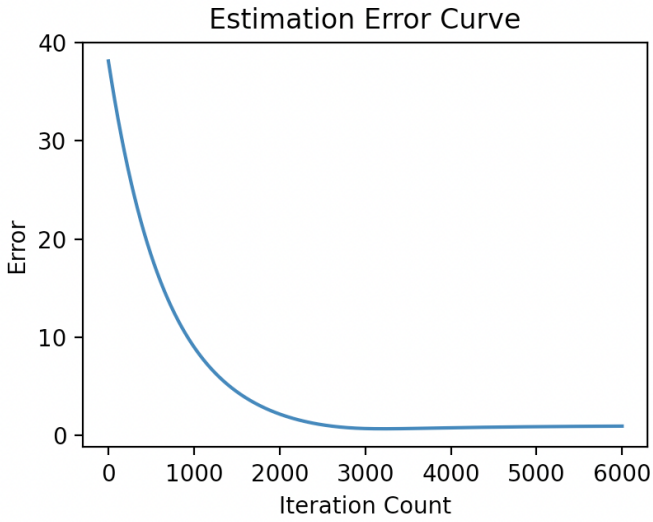
$$x_{k+1} = x_k - \eta \nabla f(x_k), \text{ where } \eta \text{ is the learning rate.}$$

The results of the gradient descent implementation are shown below.

First Run Results

The following training loss curve and estimation error curve showcase the functionality of the gradient descent algorithm implementation. The results were obtained with the following settings: learning rate = 1e-3, noise variance = 0.1, iterations k = 6000.



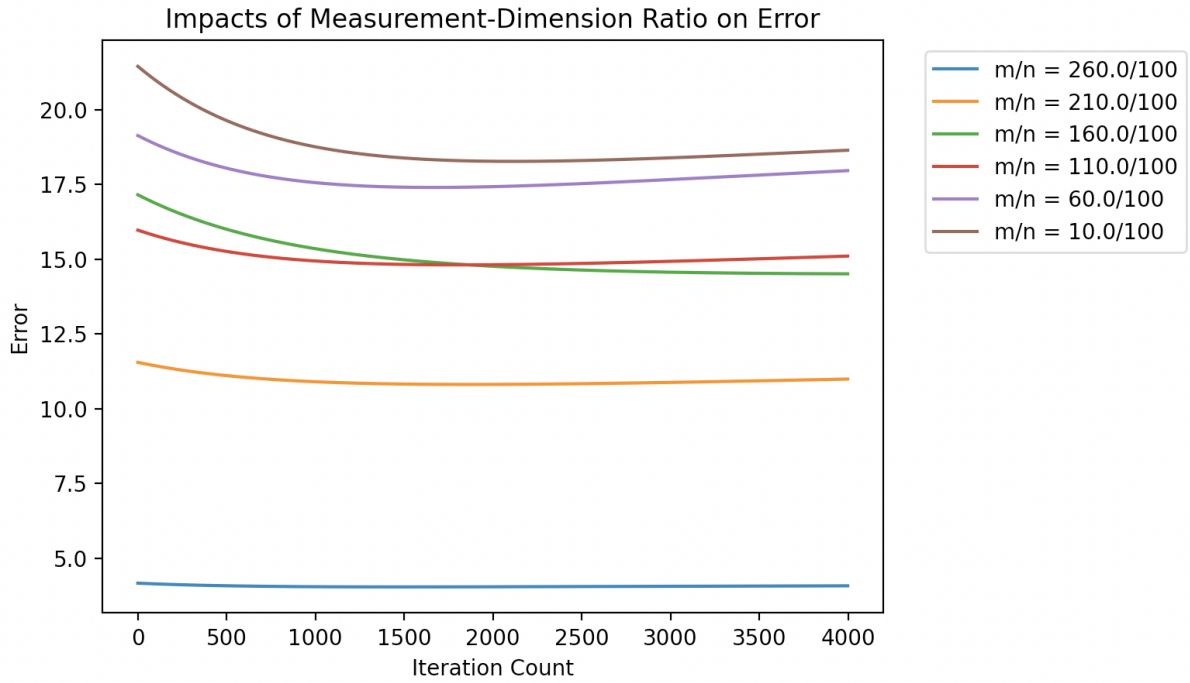


Both the curves behave as expected. The loss and error should both decrease and training continues. The similarity in values is predicted to come from the fact that normal gaussian distributions were used to generate the sub-components of the algorithm, such as A or the δ offset in the generation of b.

Hyper-Parameter Exploration

Measurement - Dimension Ratio

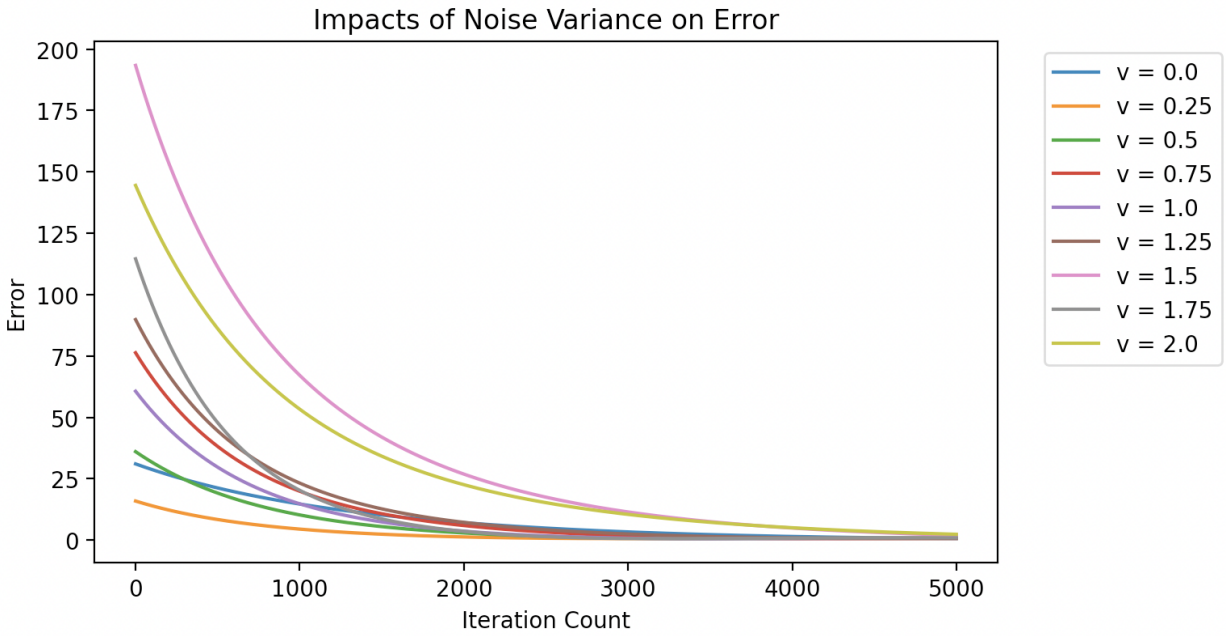
It is expected that as the measurement value approaches the dimension value, the estimation error will decrease because there are more samples to train on. The plot below confirms this relationship. The results were obtained with the following settings: learning rate = 1e-3, variance = 0.1, iterations k = 4000. The n dimension was held constant at 100. The m measurements ranged from 0.1*n up to 3*n with a step size of 0.5*n.



Note: Although the bottom curve for the highest ratio (260/100) does seem to be flat in this plot due to its relation to the other curves, it is actually still curved and decreasing when zoomed in.

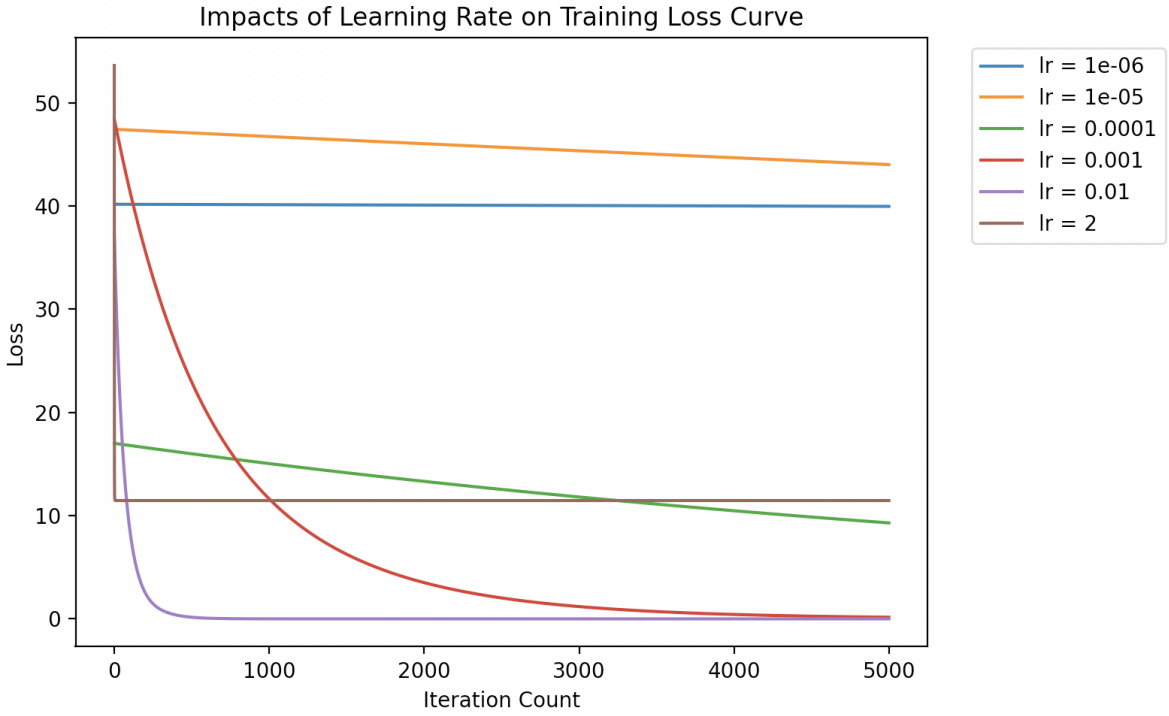
Noise Variance

It is expected that as the noise variance increases, the error will increase as well due to the introduction of more noise. The plot below confirms this expectation and the impact that noise variance has on the error. The results were obtained with the following settings: n = 100, m = 10, learning rate = 1e-3, iterations k = 5000. The variance ranged from 0 to 2 with a step size of 0.25.



Learning Rate

The learning rate often has a strong impact on the behavior of gradient descent training. The plot below shows the impact of learning rate on the training loss curve. The results were obtained with the following settings: n = 100, m = 10, variance = 0.1, iterations k = 5000. The learning rate ranged from 1e-6 to 2.



The plot indicates that the best learning rate is 0.001 or 0.01. The high learning rate of 2 is seen to drop quickly and flatten such that no more progress is made. The very small learning rates, such as 1e-6 or 1e-5, clearly get stuck at the stop where they make very little progress over all the iterations. These behaviors are as expected for variations in learning rates.