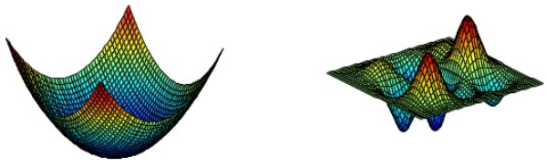
Gradient Descent Implementation Issues

- If we are asked to use gradient descent on data that follows the pattern of non-convex optimization, then optimization becomes difficult. This is because there exist several locally optimal points and this makes the process of finding out whether a solution exists, or the solution is global.
- If there exists a saddle point (point which has zero as the gradient value but not categorized as an optimal point), then we enter an area of low gradient values and get stuck here, especially when you arrive from a ridge. Then, you will have to decrease your algorithm step size, and this results in premature convergence.



Convex vs Non-convex Optimization

- The learning rate must be carefully selected since if it is too large, you will skip the correct local minimum, and if it is too small, the gradient descent will never converge.
- Vanishing gradient is a problem that occurs when applying gradient descent to ANNs (artificial neural networks). Each of the neural network's weights receives an update with respect to the weights in each iteration of training. In some instances, the gradient is vanishingly small, and this results in prevention of the weights from modifying their values, or even entirely halting the neural network from further training.
- Exploding gradient is a problem that occurs when large errors in gradients accumulate and result in large updates in the weights of the neural network during training. This makes our model unstable and prevents further learning from training data.
- When you are applying gradient descent to neural networks. It is necessary to keep track of resource utilization by networks, since if the memory space provided is not large enough for your model the model will fail.
- Computing the gradient descent for large data sets is very time consuming since it takes O(NK) time where 'N' is the size of the data set and 'K' is the update iterations.
- Gradient descent is not invariant to linear transformations. For instance, for some function f(x), if the gradient is almost orthogonal to the direction that leads to the minimum, several iterations will be needed to reach the minimum. If we set f'(x) = f(M.x) for some linear transformation 'M', then the gradient descent may minimize f quickly but not f'.