Exercise 1

Task 1:

Batch Gradient Descent:

computes the gradient of the loss function using the entire dataset in each update step. It provides a solution with minimal noise and results in stable convergence, but it is computationally very expensive.

Stochastic Gradient Descent:

uses only a single data point per iteration, which makes an iteration faster and less expensive, but it increases the noise. This noise may help to explore new areas but also slow convergence down.

Mini-batch SGD:

is a middle way that computes the gradient on a small subset which is not as expensive as BGD but still introduces some noise.

Task 2:

a) A fixed learning rate might be too high which leads to divergence or too low which leads to slow progress. To navigate a landscape of a neural network you could use a large learning rate for fast exploration and a low one to fine tune.

b) A learning rate schedule is a technique where we adjust the learning rate over time so that it gives us to highest benefit for training.

Exponential Decay:

$$\eta_t = \eta_0 \cdot e^{-\lambda t}$$

- ηt is the learning rate at epoch t.
- η0 is the initial learning rate at the start of training.
- t is the current epoch during training.
- λ is the exponential decay constant.

As t increases the learning rate decreases exponentially. This allows large steps initially and reduces the size of the steps overtime to avoid overshooting and a stable pace to finetune.