Mini-Batch Stochastic Gradient Descent

1. Preparing dataset
   - proxy: quick test dataset for debugging
     - remove bugs
     - small, run fast
   - split into: training/validation/evaluation

2. Setting initial hyperparameters
   - Not network parameters
   - Used to train network
   - Not "learned"

3. Create the NN (model)

4. Train model
Gradient Descent (Batch Gradient Descent)

We average gradients across all training examples.

for each epoch
1. compute all $N$ gradients
2. average all $N$ gradients
3. update parameters using average gradients

+ very stable (loss nearly always goes down)
- very slow
Stochastic Gradient Descent

for each epoch
    shuffle the examples
    -> randomly/stochastically
    for each example
        1. compute gradients
        2. update parameters

How many times do we update parameters per epoch

for BGD: 1
for SGD: N (# of training examples)

+ much faster convergence
- susceptible outliers
  b) less general
Mini-Batch SGD

for each epoch
  randomly create batches
  for each batch
    1. compute gradients
    2. average gradients
    3. update parameters