# Machine Learning

Pawel Wocjan

University of Central Florida

Fall 2020

# Stochastic Gradient Descent

- In gradient descent, a **batch** is the total number of examples you use to calculate the gradient in a single iteration.
- So far, we've assumed that the batch has been the entire data set.
- But often data sets contain huge numbers of examples with huge numbers of features.
- Consequently, a batch can be enormous. A very large batch may cause even a single iteration to take a very long time to compute.
- A large data set with randomly sampled examples probably contains redundant data. In fact, redundancy becomes more likely as the batch size grows.
- Some redundancy can be useful to smooth out noisy gradients, but enormous batches tend not to carry much more predictive value than large batches.

# Stochastic Gradient Descent

- What if we could get the right gradient on average for much less computation?
- By choosing examples at random from our data set, we could estimate (albeit, noisily) a big average from a much smaller one.
- **Stochastic gradient descent (SGD)** takes this idea to the extreme–it uses only a single example (a batch size of 1) per iteration.
- Given enough iterations, SGD works but is very noisy. The term "stochastic" indicates that the one example comprising each batch is chosen at random.

# Reducing Loss

- **Mini-batch stochastic gradient descent (mini-batch SGD)** is a compromise between full-batch iteration and SGD. A mini-batch is typically between 10 and $1,000$ examples, chosen at random.
- Mini-batch SGD reduces the amount of noise in SGD but is still more efficient than full-batch.

# Key Terms

- batch
- batch size
- mini-batch
- stochastic gradient descent (SGD)