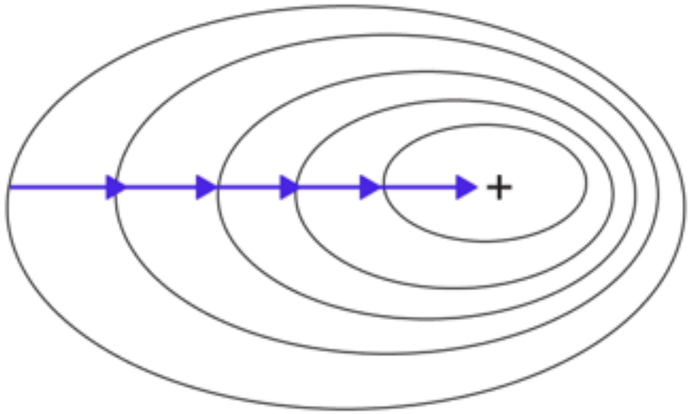
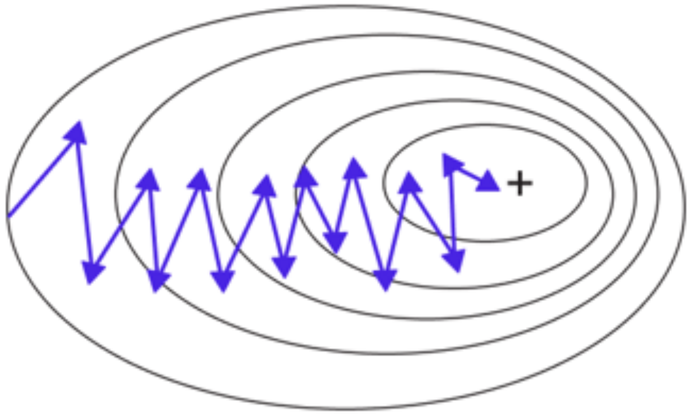
Versions of Gradient Descent

Batch Gradient Descent

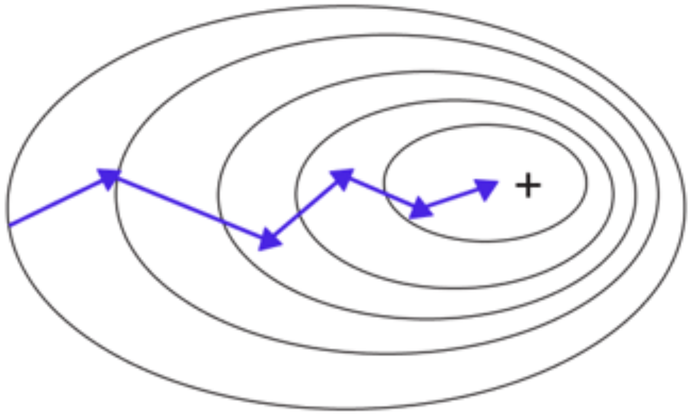


Stochastic Gradient Descent



works well for big data
with a lot redundancies

Mini-Batch Gradient Descent



mostly used for neural networks

Problems with Gradient Descent

- Choosing a learning rate is hard
- Setting learning rate schedules ahead of time is hard
- Using the same learning rate for all parameters
- Local minima and saddle points
- Variations exist to overcome some of these problems, e.g.
  - **Momentum:**
    - Momentum allows us to “build up speed” as we go downhill
    - It uses a moving average
    - Momentum often helps us converge faster
    - Momentum allows us to escape (some) local minima