Gradient Descent

Short Summary: Gradient Descent is a widely used optimization algorithm for efficiently training machine learning models. There are three primary categories of Gradient Descent: Batch Gradient Descent, Stochastic Gradient Descent, and Mini-Batch Gradient Descent. We discuss each.

Definitions:

- θ - Model parameters.
- $J(\theta)$ - Loss function.
- $\bigtriangledown_\theta J(\theta)$ - Gradient of the losss function w.r.t. the parameters θ.
- α - Learning rate.
- n - Mini-batch size.

- **Batch Gradient Descent**
  - The entire dataset is used for updating the model parameters.
  - Does not allow for online learning (i.e., on-the-fly updates with new samples).
  - Slow and problematic if the entire dataset does not fit into memory.
  - Converges to the global minimum for convex functions and to a local minimum for non-convex functions for "reasonable" learning rates.

$$\theta = \theta - \alpha \cdot \bigtriangledown_\theta J(\theta) \tag{1}$$

- **Stochastic Gradient Descent**
  - A single (x, y) feature-label pair is used for updating the model parameters.
  - Allows for online training and faster than Batch Gradient Descent (fits into memory).
  - Loss function fluctuates heavily with each iteration resulting in high variance updates.
  - Might fail to converge exactly due to fluctuations, but has a chance of jumping out of a local minimum and converging to a better one.

$$\theta = \theta - \alpha \cdot \bigtriangledown_\theta J(\theta; x^{(i)}; y^{(i)}) \tag{2}$$

- **Mini-Batch Gradient Descent**
  - A mini-batch (usually from $n = 8$ to $n = 512$) is used for updating the model parameters.
  - Allows for online training and is fast as it leverages highly optimized tensor operations.
  - Reduces the fluctuations of Stochastic Gradient Descent.
  - Does not guarantee good convergence out-of-the-box and utilizes advanced techniques.

$$\theta = \theta - \alpha \cdot \bigtriangledown_\theta J(\theta; x^{(i:i+n)}; y^{(i:i+n)}) \tag{3}$$