Lecture 2 - Gradient Descent

Gradient descent is an optimization algorithm used to minimize the cost function in machine learning. It iteratively adjusts model parameters in the direction of steepest descent.

The update rule: theta = theta - alpha * gradient(J), where alpha is the learning rate.

Batch gradient descent: Uses the entire dataset for each update. Stable but slow for large datasets.

Stochastic gradient descent (SGD): Uses one sample at a time. Faster but noisier updates.

Mini-batch gradient descent: Uses small batches. Balances speed and stability.

Practical Tips

Learning rate scheduling: Start with a larger rate, then decay. Helps escape local minima early and converge precisely later.

Momentum: Accelerates in consistent directions, dampens oscillations. v = beta*v + (1-beta)*gradient(J), theta = theta - alpha*v

Adam optimizer: Combines momentum and RMSprop. Often the default choice for neural networks.