ESE 546, FALL 2020

MODULE 3 SUMMARY

SHEIL SARDA [SHEIL@SEAS]

Lecture outlines and key takeaways for Module 3

(1) Stochastic Gradient Descent (Chapter 11)
  (a) SGD for least-squares regression
  (b) Convergence of SGD
    (i) Strongly convex functions
    (ii) What is the appropriate notion of convergence?
    (iii) Descent Lemma for SGD
    (iv) Typical assumptions in the analysis of SGD
      (A) Stochastic gradients are unbiased
      (B) Second moment of gradient norm does not grow too quickly
    (v) Descent Lemma for SGD with additional assumptions
    (vi) Convergence rate of SGD for strongly convex functions
    (vii) Optimality gap for SGD (Theorem)
    (viii) Heuristic for training neural networks
    (ix) Convergence rate of SGD for decaying step-size (Theorem)
    (x) Convergence rate for mini-batch SGD
    (xi) When should one use SGD in place of Gradient Descent?
  (c) Accelerating SGD using momentum
    (i) Polyak-Ruppert averaging
    (ii) Momentum methods do not accelerate SGD
    (iii) Why do we use Nesterov's method to train neural networks?
  (d) Understanding SGD as a Markov Chain
    (i) Gradient flow
    (ii) Markov chains
    (iii) Invariant distribution of a Markov chain
    (iv) Time spent at a particular state by the Markov chain
    (v) A Markov chain model of SGD
      (A) Transition probability of SGD
      (B) Variance of SGD weight updates
      (C) SGD is like GD with Gaussian noise
    (vi) The Gibbs distribution

(vii) Convergence of a Markov chain to its invariant distribution
(viii) KL Divergence monotonically decreases (Lemma)

Key takeaways:

- A

(2) Accelerated Gradient Descent (Chapter 10)
  (a) Polyak's Heavy Ball Method
    (i) Polyak's method can fail to converge
  (b) Nesterov's method
    (i) Yet another way to write Nesterov's updates
    (ii) How to pick the momentum parameter?

Key takeaways:

- We can think of the gradient applied to the weight at time t as a force that acts on a particle to update its position between time steps. This particle has no inertia, so the force applied directly affects its position.
- If we give the particle a point mass and some inertia, instead of the force directly affecting the position, we can apply Newton's second law of motion $F = ma$.
- The caveat with relying on inertia to make progress is overshooting behavior around the global minimum since inertia is often very different from the gradient. This results in oscillating behavior.
- Nesterov's method removes the oscillation problem of Polyak by incorporating dampening or friction like in the case of a simple harmonic oscillator.

(3) Stochastic Gradient Descent (Chapter 11)
  (a) SGD for least-squares regression
  (b) Convergence of SGD
    (i) Typical assumptions in the analysis of SGD
    (ii) Convergence rate of SGD for strongly-convex functions
    (iii) When should one use SGD in place of GD?
  (c) Accelerating SGD using momentum
    (i) Momentum methods do not accelerate SGD
  (d) Understanding SGD as Markov Chain

Key takeaways:

- It is difficult to do gradient descent if the number of samples n is large because the gradient is a summation of a large number of terms.
- Epochs is a construct introduced in the deep learning libraries for book-keeping purposes, allowing apples-to-apples comparisons between different algorithms.
- After $w^t \in (w_{min}, w_{max})$ (the zone of confusion), there is no real convergence of the weights.
- If the learning rate is large, SGD makes quick progress outside of the zone of confusion but bounces around a lot inside the zone of confusion.

- If the learning rate is too small, SGD is slow outside of the zone of confusion but does not bounce around too much inside the zone.

Table of lecture and recitation topics:

Lecture	Topic
18	Stochastic Gradient Descent I
19	Stochastic Gradient Descent II
Rec 11	Vignette: Object Detection
20	Markov Chains