ESE 546, FALL 2020

MODULE 4 SUMMARY

SHEIL SARDA [SHEIL@SEAS]

Lecture outlines and key takeaways for Module 4

(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)

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
Rec 12	Generalization Bounds