ESE 546, FALL 2020

MODULE 2 SUMMARY

SHEIL SARDA [SHEIL@SEAS]

Lecture outlines and key takeaways for Module 2

(1) Background on Optimization, Gradient Descent (Chapter 9)
   (a) Convexity
   (b) Introduction to Gradient Descent
      (i) Conditions for optimality
      (ii) Different types of convergence
   (c) Convergence rate for gradient descent
      (i) Some assumptions
      (ii) GD for convex functions
      (iii) Gradient descent for strongly convex functions
   (d) Limits on convergence rate of first-order methods

   Key takeaways:
   - Monotonicity of the gradient implies convexity.
   - The loss function l is always be a function of the entire dataset.
   - For gradient descent, if we pick the step-size $\eta \leq \frac{1}{L}$, the gradient descent always improves the value of the function with each iteration and also improves the distance of the weights to the global minimum at each iteration.
   - Strong convexity enables fewer iterations to converge. Compared to the $\mathcal{O}(1/\epsilon)$ iterations required for convex functions, strongly convex functions require only $\mathcal{O}(\log(1/\epsilon))$ iterations.
   - Nesterov's lower bound suggests the existence of gradient-based algorithms for convex functions which require $\mathcal{O}(1/\epsilon^2)$ iterations.

(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
14	Gradient Descent
15	Midterm Exam
Rec 9	Midterm Discussion
16	Momentum (heavy-ball, Nesterov), Adam, Early stopping
17	Stochastic Gradient Descent I
Rec 10	Tricks of Trade in training neural networks