Lecture #14: Gradient Descent for Strongly Convex Functions

Instructor: Aditya Bhaskara *Scribe: Alex Stewart*

CS 5966/6966: Theory of Machine Learning

March 3^{rd}, 2022

Abstract

This lecture recaps the basic theorem underpinning gradient descent and the effect of function smoothness on convergence. New material includes strong convexity and the Polyak-Lojasiewicz inequality which offer improved bounds and generalizations on gradient descent.

1 Basic Theorem Recap

Assume f is L Lipschitz, domain is all of R^d, $|w_0 - w^*| \leq B$. Without any other constraints of f, we have the following theorem.

Consider running T steps of gradient descent with a fixed learning rate η. Then we have:

$$\frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^*) \leq \frac{B^2}{2\eta T} + \frac{L^2\eta}{2} \tag{1}$$

When η is correctly tuned the RHS approximates $\frac{LB}{\sqrt{t}}$.
This theorem utilizes the basic inequality about convex functions that for any point on the function, the tangent of the point lies below the function. Mathematically this is equivalent to:

$$f(w^*) \geq f(w_t) + \langle w * -w_t, \Delta f(w_t) \rangle$$

In addition, this theorem uses the potential function:

$$\phi_t = |w_t - w^*|^2$$

2 Noisy Gradient Descent Recap

The intuition that for equation (1), f does not need to be the same function at every timestep allows us to generalize the theorem to the noisy case.

Let g be a "noisy gradient oracle" that returns a random variable $g(w)$ when given w, S.T. $E[g(w)] = \Delta f(w)$ with a variance bound $E[||g(w)||^2] \leq L^2$.

Given that g introduces unbiased noise with low variance, this concept allows gradient descent to generalize to stochastic sampling and gradient descent with privacy considerations.

3 Additional Structure: Smoothness Recap

Function f is M smooth if gradient of f is M-Lipschitz, mathematically this is:

$$||\Delta f(x) - \Delta f(y)|| \leq M||x-y|| \leftrightarrow ||\Delta^2 f(x)||_2 \leq M$$

$||\Delta^2 f(x)||_2$ is the magnitude of the largest eigenvalue.

This directly implies $\forall x, y$:

$$f(y) \leq f(x) + \langle \Delta f(x), y-x \rangle + M||y-x||^2$$

Intuitively, this states that the curvature of f is bounded by M, which also implies that every iteration of gradient descent yields a drop in the function value.

After T steps, $\sum_t |\Delta f(w_t)|^2$ is bounded by $4M(f(w_0) - f(w^*))$.

Key observations for gradient descent on smooth functions:

1. Convergence rate of 1/T.
2. Gradient descent on smooth non-convex functions converges to "approximately singular" points.

4 Matrix Basics

Let $A \in R^{d \times d}, z = (z_1, ..., z_d)$
The quadratic form in d variables:

$$z^T A z = \sum_{i,j} A_{ij} z_i z_j$$

Example: for the matrix $\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$, the quadratic form is $z_1^2 - z_1 z_2 + z_2^2$.

The max z with $||z|| = 1$ of the quadratic form is the largest eigenvector of A. Mathematically:

$$\max_{||z||=1} z^T A z = \max_{\lambda} A$$

5 Gradient Descent on Smooth Functions

Gradient descent update function:

$$w_{t+1} = w_t - \eta \Delta f(w_t)$$

Alternate definition of smoothness as it relates to the update function:

$$f(w_{t+1}) \leq f(w_t) + \langle \Delta f(w_t), w_{t+1} - w_t \rangle + M||w_{t+1} - w_t||^2$$

$w_{t+1} - w_t = -\eta \Delta f(w_t)$

Let $\eta = 1/2m$ and simplify:

$$f(w_t + 1) \leq f(w_t) - \frac{\eta}{2}||\Delta f(w_t)||^2$$

With the function being smooth, this shows convergence of $1/T$.

6 Can We Go Beyond 1/T Convergence?

Purely assuming smoothness we can get rate of $1/T^2$ (Nesterov 1983).

Formally, consider GD-like procedures, where $w_{t+1} = H(w_1, w_2, ..., w_t, \Delta f(w_1), \Delta f(w_2), ..., \Delta f(w_t))$. For all procedures of this kind, error after t iterations must be $\geq \frac{1}{t^2}$ in the worst case. This is also known as the oracle lower bound.

7 Strong Convexity

Function f is μ-strongly convex if we have a lower bound via a parabola. Mathematically:

$$f(y) \geq f(x) + \langle \Delta f(x), y - x \rangle + \mu ||y - x||^2.$$

If f is both μ-strongly convex and m-smooth, f is bounded by two parabolas. This equivalently means the hessian is bounded between two parabolas.

$$\forall n, \mu I \preceq \Delta^2 f(n) \preceq MI$$

Without strong convexity we had:

$$f(w^*) \geq f(w) + \langle \Delta f(w), w^* - w \rangle$$

With the addition of f being strongly convex we have an additional term on the RHS.

$$f(w^*) \geq f(w) + \langle \Delta f(w), w^* - w \rangle + \mu ||w^* - w||^2$$

Utilizing the potential function:

$||w^* - w||^2$ is the potential function ϕ_t

$$\phi_{t+1} = ||w_t - \eta \Delta f(w_t) - w^*||^2$$

$$= \phi_t - \eta \langle \Delta f(w_t), w_t - w^* \rangle + \eta^2 ||\Delta f(w_t)||^2$$

From the smoothness constraint we had:

$$f(w_{t+1}) \leq f(w_t) - \frac{\eta}{2} ||\Delta f(w_t)||^2, \eta < \frac{1}{2M}$$

$$||\Delta f(w_t)||^2 \leq \frac{2}{\eta}(f(w_t) - f(w_{t+1}))$$

Now, using μ-strong convexity:

$$\phi_{t+1} \leq \phi_t - \eta(f(w_t) - f(w^*)) - \eta\mu\phi_t + \frac{2}{\eta}(f(w_t) - f(w_{t+1}))$$

After T steps:

$$\phi_T \leq (1 - \frac{\mu}{8M})^T B^2 \leq e^{-\mu T/8M} B^2$$

Thus, if we want this to be $< \epsilon$, then we must pick $T \approx log(B^2/\epsilon) 8M/\mu$ or $T \approx (M/\mu) log(\frac{1}{\epsilon})$

M/μ is the condition number.

8 Gradient Descent Generalization

Polyak-Lojasiewicz inequality: suppose f satisfies:

$$|\Delta f(w)|^2 \geq c(f(w) - f(w^*))\forall w$$

This holds for strongly convex functions, but can also be satisfied for non-convex functions.
If this inequality holds for f then:

$$f(w_{t+1}) \leq f(w_t) - \frac{\eta}{2}||\Delta f(w_t)||^2$$

$$f(w_{t+1}) - f(w^*) \leq f(w_t) - f(w^*) - \frac{\eta}{2}||\Delta f(w_t)||^2$$

$$||\Delta f(w_t)||^2 \geq c(f(w_t) - f(w^*))$$

$$f(w_{t+1}) - f(w^*) \leq (1 - c\frac{\eta}{2})(f(w_t) - f(w^*))$$