E0 306: Deep Learning: Theory and Practice 5^{th} March 2019

Scribe Notes for Lecture 4

Introduction to Neural Networks

Lecturer: Anand Louis *Scribe*: C Murti

1 Introduction

Machine learning typically involves solving large scale optimization problems; for example, training a neural networks involves solving an optimization problem with millions of variables. The most commonly used algorithm for solving such problems is **Stochastic Gradient Descent** (SGD), first introduced in [3]. SGD is an extension of **gradient descent** (also known as **steepest descent**), a fundamental technique for optimization.

First, consider the following optimization problem:

$$\begin{aligned} &\inf \ f(x) \\ &\text{subject to } \ x \in K \end{aligned} \tag{P}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a *cost function* and $K \subseteq \mathbb{R}^n$ is a convex set, and typically, f is assumed to convex. Problem (P) is referred to as a *constrianed problem*, and if $K \equiv \mathbb{R}^n$, we say that (P) is an *unconstrained problem*. When f is differentiable, we can obtain gradient information about f; often, this is referred to as first order information [1, 2]. Informally, gradient descent is an algorithm which, at each iteration, minimizes the first order approximation of the function centered at the previous iterate. We describe this algorithm formally in the sequel.

In this lecture, we introduce gradient descent, an optimization algorithm that is a crucial tool for large scale machine learning. After providing some foundational results on smoothness and convexity, we prove that under mild assumptions, gradient descent converges to a minimum in $O(1/\varepsilon^2)$ steps.

Why do we use Gradient Descent?

Gradient descent (and it's variants) are far and away the most widely used algorithms for optimization in high dimensional spaces. This is because, while other algorithms such as conjugate gradient methods and Newton's method [1] have better convergence rates, they require efficient computations of large matrices; this is particularly true for Newton's method, which requires the computation of the Hessian at each iteration, which in turn costs $\Theta(n^2)$ operations to solve. When training Neural Networks in particular, the dimension of the problem can be in the order of millions. Gradients, on the other hand, can be computed very efficiently thanks to backpropagation.

2 Notation

In this section, we introduce our notation.

Let $f \in C^k(X, Y)$ denote that $f : X \to Y$ be k times differentiable. $\nabla f(x)$ denotes the gradient of f at x, and $\nabla^2 f(x)$ denotes the Hessian of f at x. If $f : \mathbb{R} \to \mathbb{R}$, $f'(x)$ denotes the derivative of f, and $f''(x)$ denotes the second derivative of f. For any $A, B \in \mathbb{S}^n$, where $\mathbb{S}^n$ is the space of symmetric matrices, $A \succeq B$ means $A - B$ is positive semidefinite. $\langle x, y \rangle$ denotes the inner product between x and y; in $\mathbb{R}^n$, $\langle x, y \rangle = x^T y$.

3 Some preliminaries on convexity and smoothness

3.1 Convex Sets and Functions

First, we define convex sets and functions, and prove some properties of each. After we establish our initial results, we will derive and prove the convergence of gradient descent under the assumption of convexity.

Definition 1. *A set $K \subseteq \mathbb{R}^n$ is said to be convex if, for each $x,\ y \in K$*

$$\lambda x + (1 - \lambda) y \in K \tag{1}$$

for each $\lambda \in [0, 1]$ (the condition holds trivially for the end points).

Examples of convex sets include polyhedra, ℓ_p balls, half spaces, and $\mathbb{R}^n$.

Definition 2. *A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be convex if, for each $x,\ y \in \mathbb{R}^n$, we have*

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y) \tag{2}$$

for each $\lambda \in [0, 1]$ (the condition holds trivially for the end points). Furthermore, if

$$f(\lambda x + (1 - \lambda) y) < \lambda f(x) + (1 - \lambda) f(y)$$

for each $\lambda \in (0, 1)$, we say that f is strictly convex.

Examples of convex functions are all affine functions, positive semidefinite quadratics, cross entropy loss and functions of the form $f(x) = \|x\|_p$; for more examples, refer to [1].

3.2 First order characterizations of convex functions

We now derive a fundamental property of differentiable convex functions.

Lemma 1. *Suppose $f \in C^1(K, \mathbb{R})$, where K is convex. Then, the following statements are equivalent.*

1. *f is convex.*
2. *For every $x,\ y \in K$, we have*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle. \tag{3}$$

Proof. Here, we prove that this holds for the 1-dimensional case ($K \subseteq \mathbb{R}$, and then generalize this result to the higher dimensional setting.

First, we prove $1 \Rightarrow 2$. From (2), we have

$$\begin{aligned} f((1-\lambda)x + \lambda y) &\leq \lambda f(y) + (1-\lambda) f(x) \\ &= f(x) + \lambda \left(f(y) - f(x) \right). \end{aligned} \tag{4}$$

We rearrange terms, giving us

$$\begin{aligned} f((1-\lambda)x + \lambda y) - f(x) &\leq \lambda \left(f(y) - f(x) \right) \\ \frac{f((1-\lambda)x + \lambda y) - f(x)}{\lambda} &\leq f(y) - f(x) \\ \frac{f((1-\lambda)x + \lambda y) - f(x)}{\lambda} + f(x) &\leq f(y) \\ \frac{f((1-\lambda)x + \lambda y) - f(x)}{\lambda(y-x)} (y-x) + f(x) &\leq f(y) \end{aligned} \tag{5}$$

We then take the limit as λ approaches 0 from the right of (5).

$$\begin{aligned} \lim_{\lambda \to 0^+} \frac{f((1-\lambda)x + \lambda y) - f(x)}{\lambda(y-x)} (y-x) + f(x) &\leq f(y) \\ \Rightarrow f'(x)(y-x) + f(x) &\leq f(y). \end{aligned} \tag{6}$$

Hence, we have shown that $1 \Rightarrow 2$.

To show that $2 \to 1$, fix $x, y \in \mathbb{R}$ and $\lambda \in (0,1)$. Suppose $z = (1-\lambda)x + \lambda y$. Then, we have

$$\begin{aligned} f(x) &\geq f(z) + f'(z)(x-z) \\ f(y) &\geq f(z) + f'(z)(y-z) \\ \Rightarrow (1-\lambda) f(x) &\geq (1-\lambda) \left(f(z) + f'(z)(x-z) \right) \\ \lambda f(y) &\geq \lambda \left(f(z) + f'(z)(y-z) \right). \end{aligned}$$

We then add the latter two expressions to get

$$\begin{aligned} (1-\lambda) f(x) + \lambda f(y) &\geq f(z) + f'(z) \left(\lambda(y-z) + (1-\lambda)(x-z) \right) \\ &\geq f(z) + f'(z) \left(\lambda(1-\lambda)(y-x) + (1-\lambda)\lambda(x-y) \right) \\ (1-\lambda) f(x) + \lambda f(y) &\geq f(z). \end{aligned}$$

Hence, we have shown that $2 \Rightarrow 1$ for the 1 dimensional case.

To prove the case for $f \in C^1(\mathbb{R}^n, \mathbb{R})$, define

$$g(t) = f(x + t(y-x)). \tag{7}$$

We can think of g as the restriction of f to the line segment connecting x and y. Then, we have

$$g'(t) = \frac{d}{dt} g(t) = \langle \nabla f(x + t(y-x)), y - x \rangle$$

. Since f is convex, it follows that g is convex. Thus, it follows that

$$g(1) \geq g(0) + g'(0)(1-0) \Rightarrow f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

Thus, we prove the Lemma. □

We now prove the monotonicity of the gradient (the gradient is a monotonic mapping). This result was not covered in class; see [1] for reference.

Lemma 2. *Suppose we have a convex function $f \in C^1(K, \mathbb{R})$, where K is convex. Then, for each $x, y \in K$, we have*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0. \tag{8}$$

Proof. The proof follows directly from the previous Lemma: substitute the pairs (x, y) and (y, x) into (3), and add. □

3.3 Second order characterizations of Convexity

We now state a second order characterization of convexity.

Lemma 3. *Suppose $f \in C^2(K, \mathbb{R})$, where $K \subseteq \mathbb{R}^n$ is convex, is convex. Then, for each $x \in K$, we have $\nabla^2 f(x) \succeq 0$.*

Refer to [1] for the proof. Note that it is essential that K is a convex set. An example of a function that has a positive semidefinite Hessian, but is not convex is $f(x) = x^{-2}$, as it's domain is nonconvex (0 is excluded from it).

Remark: The condition in Lemma 2 is necessary and sufficient; to show this, simply define a scalar function on the projection of f onto the line segment connecting x, y. If this function is nondecreasing, then (8) holds, and by the second order characterization of convexity, we can show that that if the gradient is a monotone mapping, the function is convex.

3.4 L-smoothness

In this section, we introduce a notion of smoothness of functions. Fundamnetally, we define some properties of functions whose gradients are Lipschitz-smooth. We define this property formally below.

Definition 3. *Suppose $f \in C^1(K, \mathbb{R})$, where $K \subseteq \mathbb{R}^n$ and $k \geq 1$. We say that f is L smooth on K if, for each $x,\ y, \in K$, we have*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \tag{9}$$

L smooth functions have a number of useful properties, which we state below.

Lemma 4. *Suppose $f \in C^1(K, \mathbb{R})$, where $k \geq 1$ and $K \subseteq \mathbb{R}^n$, is convex and L-smooth. Then, for each $x,\ y \in K$, we have*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + L\|y - x\|^2. \tag{10}$$

Proof. From the convexity of f, we have

$$\begin{aligned}
&f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle \\
\Rightarrow &f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle \\
&f(y) - f(x) \leq \langle \nabla f(y) - \nabla f(x) + \nabla f(x), y - x \rangle \\
&f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \langle \nabla f(y) - \nabla f(x), y - x \rangle \\
&f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \|\nabla f(y) - \nabla f(x)\| \|y - x\| && \text{(Cauchy-Schwartz Inequality)} \\
&f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + L\|y - x\|^2 && (L - \text{smoothness}).
\end{aligned}$$

□

We now state a property of twice-differentiable L-smooth functions.

Lemma 5. *Suppose $f \in C^2(\mathbb{R}^n, \mathbb{R})$ is L-smooth. Then,*

$$-LI \preceq \nabla^2 f(x) \preceq LI \tag{11}$$

for all $x \in \mathbb{R}^n$.

The proof of this Lemma is given in assignment 2.

4 A First proof for Gradient Descent under Mild Assumptions

In this section, we prove that gradient descent, when applied to a convex function, converges to an ϵ ball containing a minimum in $O(1/\epsilon^2)$ iterations. We formally state the theorem below.

Theorem 1. *Suppose $f \in C^1(\mathbb{R}^n, \mathbb{R})$ is convex with a minimum at x^*, and that $\|\nabla f(x)\| \leq G \; \forall \; x \in S := \{x : \|x - x^*\| \leq D, \}$, where $D, G \in \mathbb{R}_{>0}$. For any $\epsilon \in (0, 1)$, if we generate a sequence $\{x_t\}_{t=0}^T$, where each $x_t \in S$, using*

$$x_{t+1} = x_t - \eta \nabla f(x_t), \tag{12}$$

we have

$$f\left(\frac{1}{T}\sum_{t=0}^{T} x_t\right) - f(x^*) \leq \epsilon, \tag{13}$$

if we choose

$$T = \frac{DG}{\epsilon^2} \;\; \textit{and} \;\; \eta = \frac{D}{G\sqrt{T}}. \tag{14}$$

Proof. Recall (3). Choose $x = x_t$ and $y = x^*$, giving us

$$\begin{aligned}
f(x^*) &\geq f(x_t) + \langle \nabla f(x_t), x^* - x_t \rangle \\
\Rightarrow \langle \nabla f(x_t), x_t - x^* \rangle &\geq f(x_t) - f(x^*).
\end{aligned}$$

Recall from (12) that

$$\frac{1}{\eta}(x_t - x_{t+1}) = \nabla f(x_t).$$

Substituting this, we get

$$f(x_t) - f(x^*) \leq \langle \frac{1}{\eta}(x_t - x_{t+1}), x_t - x^* \rangle.$$

We then apply the property $\langle a, b \rangle = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a-b\|^2\right)$, giving us

$$(x_t) - f(x^*) \leq \frac{1}{2\eta}\left(\|x_t - x_{t+1}\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right). \tag{15}$$

We know that $\|x_{t+1} - x^*\|^2 = \eta^2 \|\nabla f(x_t)\|^2 \leq \eta^2 G^2$. Substituting this back, we get

$$f(x_t) - f(x^*) \leq \frac{1}{2\eta}\left(\eta^2 G^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right).$$

Then. taking a telescoping sum (from $t = 0$ to T), we get

$$\begin{aligned}
\sum_{t=0}^{T} f(x_t) - f(x^*) &\leq \sum_{t=0}^{T} \frac{1}{2\eta}\left(\eta^2 G^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right) \\
&= \frac{T\eta G}{2} + \frac{\|x_0 - x^*\|^2}{2\eta} - \frac{\|x_{T+1} - x^*\|^2}{2\eta} \\
&\leq \frac{T\eta G^2}{2} + \frac{D^2}{2\eta}.
\end{aligned}$$

In the last inequality, we ignore $-\|x_{T+1} - x^*\|^2$ since it is negative, and we want an upper bound. From here, we divide both sides by T (taking the average over T iterations), giving us

$$\frac{1}{T}\sum_{t=0}^{T} f(x_t) - f(x^*) \leq \frac{\eta G^2}{2} + \frac{D^2}{2\eta T}.$$

Since f is convex, it follows that

$$\frac{1}{T}\sum_{t=0}^{T} f(x_t) \geq f\left(\frac{1}{T}\sum_{t=0}^{T} x_t\right).$$

Thus, we get

$$f\left(\frac{1}{T}\sum_{t=0}^{T} x_t\right) - f(x^*) \leq \frac{\eta G^2}{2} + \frac{D^2}{2\eta T}. \tag{16}$$

From here, we we find the value of η that minimizes the previous expression. Differentiating w.r.t η, we get

$$\eta = \frac{D}{G\sqrt{T}}. \tag{17}$$

Substituting (17) into (16) gives us

$$f\left(\frac{1}{T}\sum_{t=0}^{T} x_t\right) - f(x^*) \leq \frac{DG}{\sqrt{T}}.$$

If we choose

$$\frac{DG}{\sqrt{T}} \leq \epsilon$$

we get

$$T = \frac{DG}{\epsilon^2}. \tag{18}$$

Thus, the theorem is proved. □

Remark If we want a very accurate approximation of the minimum (ϵ is small), this requires a large number of steps with a very small stepsize (η is $O(1\sqrt{T})$). However, in practice, many functions are at least locally L-smooth. In such cases, we have

$$f(x_T) - f(x^*) \leq O(1/T).$$

We state this formally in the following theorem.

Theorem 2. *Suppose $f \in C^1(\mathbb{R}^n, \to \mathbb{R})$ is convex and L-smooth. For any $\epsilon \in (0, 1)$, the sequence $\{x_t\}_{t=1}^T$, where*

$$\|x_T - x^*\| \leq D \text{ for all } t \in \{1, ..., T\},$$

generated by (12) satisfies

$$f(x_T) - f(x^*) \leq \epsilon \tag{19}$$

with

$$\eta = \frac{2}{L} \text{ and } T = O\left(LD^2/\epsilon\right). \tag{20}$$

The proof is covered in the next lecture.

Bonus: A Tighter Bound for L-smooth functions (without assuming convexity!)

In this section, we provide a tighter version of the previous inequality, without relying on any assumptions of convexity. This result can also be found in [1, 2]. Note that we did not go over this result in class.

First, we prove the following technical Lemma.

Lemma 6. *Suppose $f \in C^k(K, \mathbb{R})$, where $k \geq 1$ and $K \subseteq \mathbb{R}^n$ is convex, is L-smooth. Then, the function*

$$g(x) = \frac{L}{2}\|x\|^2 - f(x) \tag{21}$$

is convex.

Proof. We need to show that $g(\cdot)$ satisfies (3). From the L-smoothness of $f(\cdot)$, we get

$$\begin{aligned}
&\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \\
&\|\nabla f(x) - \nabla f(y)\|\|x - y\| \leq L\|x - y\|^2 \\
&\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L\langle x - y, x - y\rangle \quad \text{(by Cauchy-Schwartz)} \\
\Rightarrow &\langle Lx - \nabla f(x) - (Ly - \nabla f(y)), x - y\rangle \geq 0.
\end{aligned}$$

Observe that $\nabla[\frac{L}{2}\|x\|^2 - f(x)] = Lx - \nabla f(x)$. Thus, from (8), we have $\frac{L}{2}\|x\|^2 - f(x)$ is convex. □

Lemma 7. *Suppose $f \in C^k(K, \mathbb{R})$, where $k \geq 1$ and $K \subseteq \mathbb{R}^n$ is convex, is L-smooth. Then, for each x, $y \in K$, we have*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2. \tag{22}$$

Proof. The proof follows by defining $g(x) = \frac{L}{2}\|x\|^2 - f(x)$, and substituting $g(\cdot)$ into (3). □

These results are useful in proving deriving the gradient descent rule

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k),$$

and proving that

$$\frac{1}{K}\sum_{k=1}^{K} \|\nabla f(x_k)\| \leq O(1\sqrt{K})$$

for any $f \in C^1(\mathbb{R}^n, \mathbb{R})$.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[3] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.