Mathematics of Machine Learning - for Undergrads

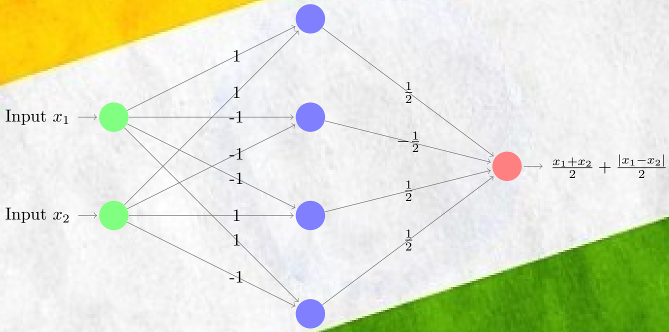
Anirbit

Computer Science, UManchester

Plan

1. Chapter-7 (a) : Introduction
2. Chapter-7 (b) : Some Key Mathematical Structures in M.L.
3. Chapter-8 : Formal Introduction to Gradient Descent
4. Chapter-9 : G.D. on Overparameterized Linear Regression
5. Chapter-10 : A Measure of Capacity : Rademacher Complexity
6. Chapter-11 : Computing the Rademacher Complexity

A depth 2, width 4 net computing the max of 2 numbers - via $\sigma : x \mapsto \max\{0, x\}$ “activation” at the (blue) gates.



Defining Neural Nets

[Practice with Formal Notation for Functions!]

The key component of a neural net is an “activation function” & a widely used one is the “Rectified Linear Unit (ReLU)”, $\forall n = 1, 2, \ldots$

$$\text{ReLU} : \mathbb{R}^n \ni \boldsymbol{x} \mapsto (\max\{0, x_1\}, \max\{0, x_2\}, \ldots, \max\{0, x_n\}) \in \mathbb{R}^n$$

Defining Neural Nets

[Practice with Formal Notation for Functions!]

The key component of a neural net is an "activation function" & a widely used one is the "Rectified Linear Unit (ReLU)", $\forall n = 1, 2, \ldots$

$$\mathrm{ReLU} : \mathbb{R}^n \ni \boldsymbol{x} \mapsto (\max\{0, x_1\}, \max\{0, x_2\}, \ldots, \max\{0, x_n\}) \in \mathbb{R}^n$$

Given a $q \times p$ matrix $\boldsymbol{W}$ and a $q-$dimensional vector $\boldsymbol{b}$, an "Affine Map" $\boldsymbol{A} : \mathbb{R}^p \to \mathbb{R}^q$ maps, $\mathbb{R}^p \ni \boldsymbol{x} \mapsto \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b} \in \mathbb{R}^q$

Defining Neural Nets

[Practice with Formal Notation for Functions!]

The key component of a neural net is an "activation function" & a widely used one is the "Rectified Linear Unit (ReLU)", $\forall n = 1, 2, \ldots$

$$\mathrm{ReLU} : \mathbb{R}^n \ni \boldsymbol{x} \mapsto (\max\{0, x_1\}, \max\{0, x_2\}, \ldots, \max\{0, x_n\}) \in \mathbb{R}^n$$

Given a $q \times p$ matrix $\boldsymbol{W}$ and a $q-$dimensional vector $\boldsymbol{b}$, an "Affine Map" $\boldsymbol{A} : \mathbb{R}^p \to \mathbb{R}^q$ maps, $\mathbb{R}^p \ni \boldsymbol{x} \mapsto \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b} \in \mathbb{R}^q$

What is a Neural Net?

Given affine maps, $\left\{ \boldsymbol{A}_i : \mathbb{R}^{w_{i-1}} \to \mathbb{R}^{w_i} \mid i = 1, \ldots, k+1 \right\}$, we can define a depth $k+1$ "Neural Net" as the following function,

$$\mathbb{R}^{w_0} \ni \boldsymbol{x} \mapsto \mathsf{N}(\boldsymbol{x}) = \boldsymbol{A}_{k+1} \circ \mathrm{ReLU} \circ \boldsymbol{A}_k \circ \cdots \circ \boldsymbol{A}_2 \circ \mathrm{ReLU} \circ \boldsymbol{A}_1(\boldsymbol{x}) \in \mathbb{R}^{w_{k+1}}$$

An Example of a Function Space -of Nets Mapping $\mathbb{R}^4 \to \mathbb{R}^3$ Using The Following Architecture

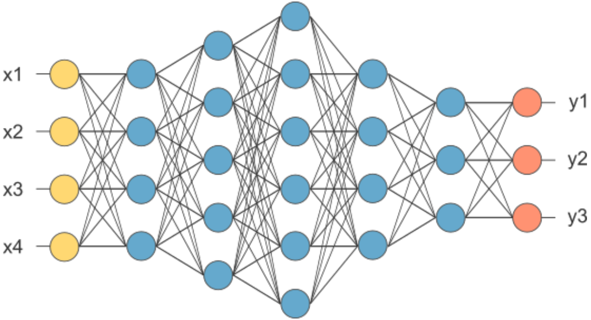


Figure 1: Unlike the max computing net, this is not a neural function because weights have not been assigned on the edges. **Such a diagram/architecture defines a certain set of neural functions.**

Some Specific Properties of Functions Are Crucial To ML

Three key properties of differentiable functions that we shall care a lot about are, whether they are,

1. "Lipschitz" or not,
2. "Convex" or not
3. "Smooth" or not.

Which of these properties are true for a given loss function makes a dramatic difference to how machine learning algorithms behave for them!

Plan

1. Chapter-7 (a) : Introduction
2. Chapter-7 (b) : Some Key Mathematical Structures in M.L.
3. Chapter-8 : Formal Introduction to Gradient Descent
4. Chapter-9 : G.D. on Overparameterized Linear Regression
5. Chapter-10 : A Measure of Capacity : Rademacher Complexity
6. Chapter-11 : Computing the Rademacher Complexity

[Concept 10] Introduction to the Euclidean Norm

Definition ($2-$**Norm)**

Given a vector $\boldsymbol{v} \in \mathbb{R}^p$ we define its $2-$Norm as,

$$\|\boldsymbol{v}\|_2 = \sqrt{v_1^2 + v_2^2 + \ldots + v_p^2}$$

i.e the Euclidean distance of the point $\boldsymbol{v}$ from the origin.

[Concept 10] Introduction to the Euclidean Norm

Definition ($2-$**Norm)**

Given a vector $\boldsymbol{v} \in \mathbb{R}^p$ we define its $2-$Norm as,

$$\|\boldsymbol{v}\|_2 = \sqrt{v_1^2 + v_2^2 + \ldots + v_p^2}$$

i.e the Euclidean distance of the point $\boldsymbol{v}$ from the origin.

Eg. Consider the vector $v = \begin{bmatrix} 1 \\ -2 \\ -1 \end{bmatrix}$**. Then it follows that**

$$\|\boldsymbol{v}\| = \sqrt{1^2 + (-2)^2 + (-1)^2} = \sqrt{6}$$

[Concept 9] Introduction to Lipschitzness

Definition (**Lipschitzness)**

A function $F : \mathbb{R}^p \to \mathbb{R}$ is said to be L–Lipschitz if $\forall\ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, we have,

$$|F(\boldsymbol{x}) - F(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$$

[Concept 9] Introduction to Lipschitzness

Definition (Lipschitzness)

A function $F : \mathbb{R}^p \to \mathbb{R}$ is said to be $L-$Lipschitz if $\forall\ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, we have,

$$|F(\boldsymbol{x}) - F(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$$

- Consider, $F(x) = x$.

  It's clear that this is $1-$Lipschitz.

[Concept 9] Introduction to Lipschitzness

Definition (**Lipschitzness)**

A function $F : \mathbb{R}^p \rightarrow \mathbb{R}$ is said to be $L-$Lipschitz if $\forall\ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, we have,

$$|F(\boldsymbol{x}) - F(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$$

- Consider, $F(x) = x$.

  It's clear that this is $1-$Lipschitz.

- But consider the function $F(x) = x^2$.

[Concept 9] Introduction to Lipschitzness

Definition (Lipschitzness)

A function $F : \mathbb{R}^p \to \mathbb{R}$ is said to be $L-$Lipschitz if $\forall\ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, we have,

$$|F(\boldsymbol{x}) - F(\boldsymbol{y})| \le L\|\boldsymbol{x} - \boldsymbol{y}\|$$

- Consider, $F(x) = x$.
  It's clear that this is $1-$Lipschitz.

- But consider the function $F(x) = x^2$.

  We can compare its values between say a point x and 0 and ask if it is true that there exists a $L > 0$ s.t for all x we will have, $|F(x) - F(0)| = x^2 \le L|x|$. But clearly for no $L > 0$ can the inequality $x^2 \le L|x|$ hold if $|x| > L$.

  Thus $F(x) = x^2$ is not a Lipschitz function.

[Concept 8] Introduction to Convexity

A convex function F can be informally understood as one whose graph between any two points x and y always lie below the straight line joining the points $(x, F(x))$ and $(y, F(y))$. *Note, that this notion of convexity does not need the function to be differentiable anywhere.*

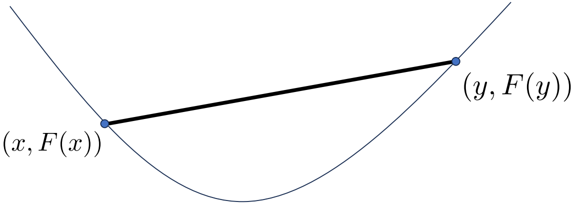


Figure 2: Convexity (Version 1)

[Concept 8] Introduction to Convexity

This view of convex function can be formally stated as,

Definition (**Convexity (Version 1)**)

A function $F : \mathbb{R}^p \to \mathbb{R}$ will be said to be a convex function if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ and $\forall \theta \in [0, 1]$ we have,

$$F(\theta \boldsymbol{x} + (1 - \theta)\boldsymbol{y}) \leq \theta F(\boldsymbol{x}) + (1 - \theta) F(\boldsymbol{y})$$

[Concept 8] Introduction to Convexity

This view of convex function can be formally stated as,

Definition (**Convexity (Version 1)**)

A function $F : \mathbb{R}^p \to \mathbb{R}$ will be said to be a convex function if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ and $\forall \theta \in [0, 1]$ we have,

$$F(\theta \boldsymbol{x} + (1 - \theta)\boldsymbol{y}) \leq \theta F(\boldsymbol{x}) + (1 - \theta) F(\boldsymbol{y})$$

This above view of convexity will be extremely important in the last lecture of this course.

[Concept 8] Introduction to Convexity

But if the function can be assumed to be differentiable (as is the case for all optimization examples that we shall consider) then the following notion of convexity also becomes useful.

Definition (**Convexity (Version 2)**)

An at least once differentiable function $F : \mathbb{R}^p \to \mathbb{R}$ is convex if, $\forall\ \boldsymbol{x}, \boldsymbol{y}$ we have,

$$F(\boldsymbol{x}) + \nabla F(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \leq F(\boldsymbol{y})$$

$$\text{where, } \nabla F(\boldsymbol{x}) \coloneqq \begin{bmatrix} \frac{\partial F}{\partial x_1} \\ \vdots \\ \frac{\partial F}{\partial x_p} \end{bmatrix} \Big|_{\boldsymbol{x}},$$

is the derivative of F that is assumed to be well-defined

[Concept 8] Introduction to Convexity

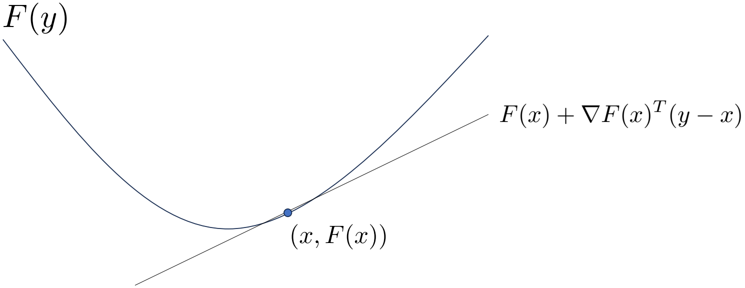


Figure 3: We can see that for differentiable functions, convexity can be thought of as the property that for any point x in the domain, the tangent to the function at that point is below the graph of the function.

Eg. $F(x) = x^2$, $F(x) = x^4$, $F(x) = e^{-x}$

[Concept 8] Introduction to Convexity

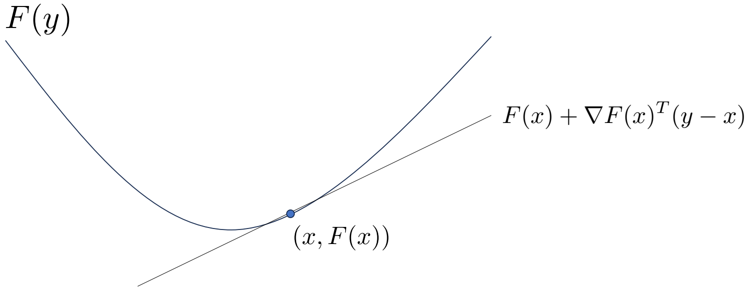


Figure 3: We can see that for differentiable functions, convexity can be thought of as the property that for any point x in the domain, the tangent to the function at that point is below the graph of the function.

Eg. $F(x) = x^2$, $F(x) = x^4$, $F(x) = e^{-x}$

Read the lecture notes for details on 2 special subsets of convex functions,
– functions which satisfy, "strict convexity" & "strong convexity"

[Concept 8] Introduction to Convexity

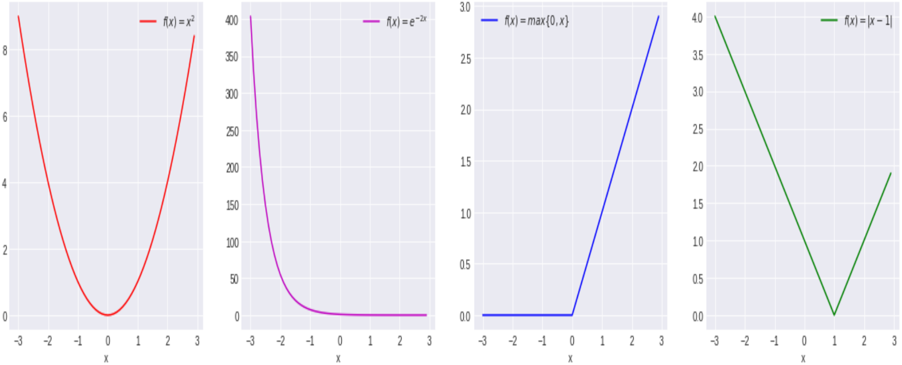


Figure 4: From left-to-right notice that we have examples of convex functions which (a) have a unique global minima (b) have no minima at all (c) are not differentiable everywhere and have an uncountable number of global minima and (d) are not differentiable everywhere but have a unique global minima.

[Concept 7] Introduction to Smoothness

Definition (**Lipschitz-Smoothness**)

For some $\beta > 0$, an at least once differentiable function $F : \mathbb{R}^p \to \mathbb{R}$ is said to be $\beta-$smooth or $\beta-$Lipschitz smooth or $\beta-$Gradient Lipschitz if its gradients are $\beta-$Lispschitz i.e $\forall\ \boldsymbol{x}, \boldsymbol{y}$, we have,

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|$$

[Concept 7] Introduction to Smoothness

Definition (**Lipschitz-Smoothness)**

For some $\beta > 0$, an at least once differentiable function $F : \mathbb{R}^p \to \mathbb{R}$ is said to be $\beta-$smooth or $\beta-$Lipschitz smooth or $\beta-$Gradient Lipschitz if its gradients are $\beta-$Lispschitz i.e $\forall\ \boldsymbol{x}, \boldsymbol{y}$, we have,

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|$$

Eg. $F(x) = x^2$ **is smooth with** $\beta = 2$**.**

[Concept 7] Introduction to Smoothness

Definition (**Lipschitz-Smoothness**)

For some $\beta > 0$, an at least once differentiable function $F : \mathbb{R}^p \to \mathbb{R}$ is said to be $\beta-$smooth or $\beta-$Lipschitz smooth or $\beta-$Gradient Lipschitz if its gradients are $\beta-$Lispschitz i.e $\forall\ \boldsymbol{x}, \boldsymbol{y}$, we have,

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|$$

Eg. $F(x) = x^2$ **is smooth with** $\beta = 2$**.** `Note that this terminology of ‘‘smoothness’’, differs from what is called a smooth function in calculus or analysis textbooks!`

[Concept 7] Introduction to Smoothness

Definition (**Lipschitz-Smoothness)**

For some $\beta > 0$, an at least once differentiable function $F : \mathbb{R}^p \to \mathbb{R}$ is said to be $\beta-$smooth or $\beta-$Lipschitz smooth or $\beta-$Gradient Lipschitz if its gradients are $\beta-$Lispschitz i.e $\forall\ \boldsymbol{x}, \boldsymbol{y}$, we have,

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|$$

Eg. $F(x) = x^2$ **is smooth with** $\beta = 2$**.** `Note that this terminology of ‘‘smoothness’’, differs from what is called a smooth function in calculus or analysis textbooks!`

In calculus, “smoothness” refers to a function being infinitely differentiable and we can see that **there are easy examples of infinitely differentiable functions which are not smooth by the above definition, like** $F(x) = x^3$ **(Prove!)**.

Non-Convex Neural Losses

Way too many real-world loss functions have local minima which are not global. Now we shall see an example of such a situation - and in this example we shall use a strongly convex regularizer to emphasize that its presence alone cannot ameliorate this complexity of having non-trivial local minima.

Non-Convex Neural Losses

Way too many real-world loss functions have local minima which are not global. Now we shall see an example of such a situation - and in this example we shall use a strongly convex regularizer to emphasize that its presence alone cannot ameliorate this complexity of having non-trivial local minima.

Consider training data of 4 data, $(x_1, y_1) = (0.5, -100), (x_2, y_2) = (-1, 300), (x_3, y_3) = (1, 1), (x_4, y_4) = (-0.5, -400)$. Recall that one of the simplest neural nets is one consisting of a single sigmoid gate of weight w which would be mapping,

$$\mathbb{R} \ni x \mapsto \frac{1}{1 + e^{-w \cdot x}} \in [0, \infty)$$

Non-Convex Neural Losses

A natural instance of a regularized squared loss function, which we call ℓ below, on the above gate for the above training data would be,

Non-Convex Neural Losses

A natural instance of a regularized squared loss function, which we call ℓ below, on the above gate for the above training data would be,

$$\mathbb{R} \ni w \mapsto F(w) \coloneqq \frac{1}{4}\sum_{i=1}^{4} \ell_{x_i,y_i}(w) + \lambda w^2 \tag{1}$$

$$\coloneqq \frac{1}{4}\sum_{i=1}^{4} \frac{1}{2} \cdot \left(y_i - \frac{1}{1 + e^{-wx_i}} \right)^2 + \lambda w^2 \in [0, \infty) \tag{2}$$

Notice how we think of losses as being univariate functions of the weights, while the gates are thought of as univariate functions of the input data.

Non-Convex Neural Losses

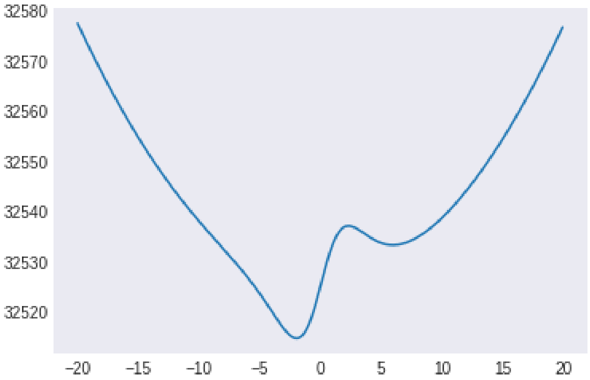


Figure 5: The plot of this F above, for $\lambda = 0.13$, as a function of w clearly shows that it is a non-Lipschitz function (owing to the regularizer) – and neither is it convex (since it has a local maxima) – and notice how it has two local minima, only one of which is a global minima.

[Concept 6] Introduction to Convergence

To match to usual conventions in mathematical literature on convergence, we shall rename the weight variable as x. Thus we can re-write the above example as,

$$\mathbb{R} \ni x \mapsto F(x) \coloneqq \frac{1}{4}\sum_{i=1}^{4} \ell_{x_i,y_i}(x) + 0.13 \cdot x^2 \tag{3}$$

$$\coloneqq \frac{1}{4}\sum_{i=1}^{4} \frac{1}{2} \cdot \left(y_i - \frac{1}{1+e^{-x \cdot x_i}} \right)^2 + 0.13 \cdot x^2 \in [0, \infty) \tag{4}$$

[Concept 6] Introduction to Convergence

To match to usual conventions in mathematical literature on convergence, we shall rename the weight variable as x. Thus we can re-write the above example as,

$$\begin{aligned} \mathbb{R} \ni x \mapsto F(x) &\coloneqq \frac{1}{4} \sum_{i=1}^{4} \ell_{x_i, y_i}(x) + 0.13 \cdot x^2 && (3) \\ &\coloneqq \frac{1}{4} \sum_{i=1}^{4} \frac{1}{2} \cdot \left(y_i - \frac{1}{1 + e^{-x \cdot x_i}} \right)^2 + 0.13 \cdot x^2 \in [0, \infty) && (4) \end{aligned}$$

On the above loss function, an implementation of the **"Gradient Descent"** algorithm at a constant "step-length" of $\eta > 0$ and starting from x_0 would be an iterative execution of the following method,

$$x_{t+1} = x_t - \eta \cdot \frac{\mathrm{d}F}{\mathrm{d}x}\Big|_{x_t}$$

[Concept 6] Introduction to Convergence

Let's see visually as to what the progress of G.D. looks like, when starting from $x_0 = 7$

[Concept 6] Introduction to Convergence

Let's see visually as to what the progress of G.D. looks like, when starting from $x_0 = 7$

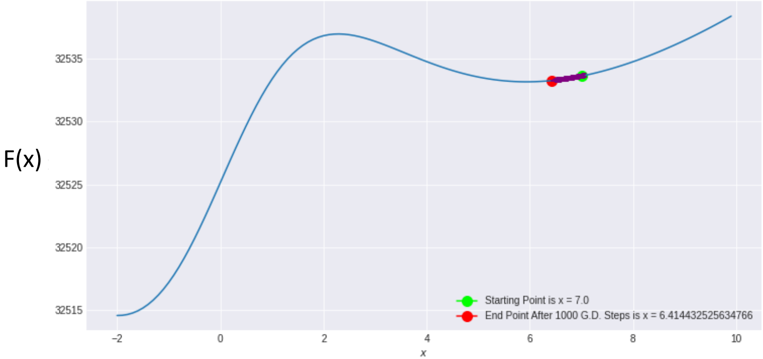


Figure 6: G.D. on F starting from $x_0 = 7$ and using a constant step-length of $\eta = 10^{-3}$ and running for 10^3 steps reaches $x \sim 6.414$

[Concept 6] Introduction to Convergence

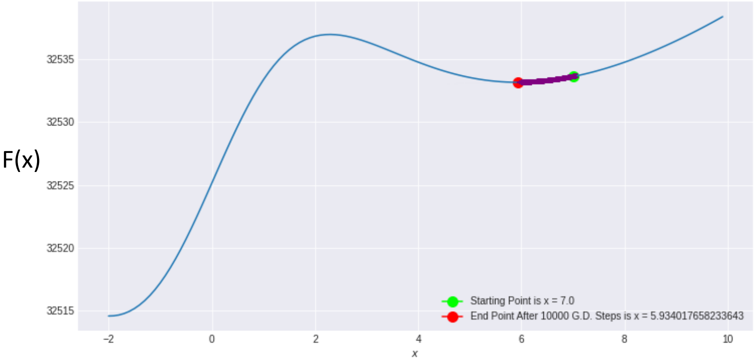


Figure 7: G.D. on F starting from $x_0 = 7$ and using a constant step-length of $\eta = 10^{-3}$ and running for 10^4 steps reaches $x \sim 5.934$

[Concept 6] Introduction to Convergence

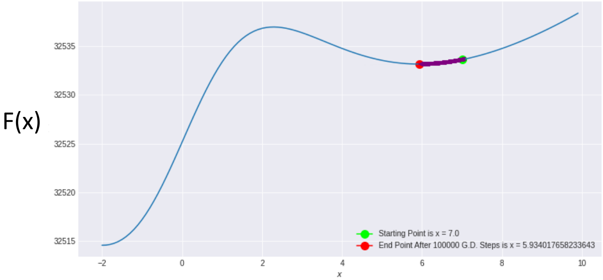


Figure 8: G.D. on F starting from $x_0 = 7$ and using a constant step-length of $\eta = 10^{-3}$ and running for 10^5 steps reaches $x \sim 5.934$

[Concept 6] Introduction to Convergence

It is clear from these figures that the algorithm makes significant progress between where it was after 10^3 steps and the next 9000 steps.

[Concept 6] Introduction to Convergence

It is clear from these figures that the algorithm makes significant progress between where it was after 10^3 steps and the next 9000 steps.

But for the next 90000 steps after $t = 10^4$, almost nothing seems to happen - the algorithm seems to have stalled. This is a tell-tale sign of "convergence" - that this sequence of numbers x_t seems to be getting arbitrarily close to a certain number as $t \to \infty$.

[Concept 6] Introduction to Convergence

Since here F was an actual example of a neural loss function – these experiments also point out that without a careful choice of starting point and step-lengths, G.D. can get "stuck" at a non-trivial local minima - and not reach the global minima of the function.

[Concept 6] Introduction to Convergence

Since here F was an actual example of a neural loss function – these experiments also point out that without a careful choice of starting point and step-lengths, G.D. can get "stuck" at a non-trivial local minima - and not reach the global minima of the function.

From an earlier figure we know that this F indeed has a unique global minima, which the G.D. settings in the above diagrams seem to fail to reach.

[Concept 6] Introduction to Convergence

Since here F was an actual example of a neural loss function – these experiments also point out that without a careful choice of starting point and step-lengths, G.D. can get “stuck” at a non-trivial local minima - and not reach the global minima of the function.

From an earlier figure we know that this F indeed has a unique global minima, which the G.D. settings in the above diagrams seem to fail to reach.

But can you find the right settings, when GD would reach the global minima?

[Concept 6] Introduction to Convergence

To understand how G.D. works, it's absolutely essential to formally define and understand the notion of convergence.

[Concept 6] Introduction to Convergence

To understand how G.D. works, it's absolutely essential to formally define and understand the notion of convergence.

Definition (**Convergence to a Limit Point)**

A sequence of points $\boldsymbol{x}_i, \forall i = 1, 2, \ldots$ in $\mathbb{R}^n$, will be said to converge to a point $\boldsymbol{x}_* \in \mathbb{R}^n$ if $\lim_{n\to\infty} \|\boldsymbol{x}_n - \boldsymbol{x}_*\|_2 = 0$. If this condition is satisfied then $\boldsymbol{x}_*$ shall be called the "limit point" or the "accumulation point" of the sequence.

[Concept 6] Introduction to Convergence

To understand how G.D. works, it's absolutely essential to formally define and understand the notion of convergence.

Definition (Convergence to a Limit Point)

A sequence of points $\boldsymbol{x}_i, \forall i = 1, 2, \ldots$ in $\mathbb{R}^n$, will be said to converge to a point $\boldsymbol{x}_* \in \mathbb{R}^n$ if $\lim_{n\to\infty} \|\boldsymbol{x}_n - \boldsymbol{x}_*\|_2 = 0$. If this condition is satisfied then $\boldsymbol{x}_*$ shall be called the "limit point" or the "accumulation point" of the sequence.

- The sequence $\frac{1}{2}, \frac{1}{2^2}, \frac{1}{2^3}, \ldots$ is a decreasing sequence, that converges to 0
- The sequence $\left(-1 - \frac{1}{2}\right), \left(-1 - \frac{1}{3}\right), \left(-1 - \frac{1}{4}\right), \ldots,$ is an increasing sequence that converges to the point -1

[Concept 4 & 5] Introduction to Supremums and Infimums

If one picks any $x \in [0,1)$ and wants to declare x as the maximum of this interval then one can always come up a very small number $\epsilon > 0$ s.t $x + \epsilon \in [0,1)$ and thus $x + \epsilon$ becomes a more credible candidate for being called the maximum - and starting from $x + \epsilon$ we can again repeat this argument and this can be continued ad infinitum!

[Concept 4 & 5] Introduction to Supremums and Infimums

If one picks any $x \in [0,1)$ and wants to declare x as the maximum of this interval then one can always come up a very small number $\epsilon > 0$ s.t $x + \epsilon \in [0,1)$ and thus $x + \epsilon$ becomes a more credible candidate for being called the maximum - and starting from $x + \epsilon$ we can again repeat this argument and this can be continued ad infinitum!

Thus we are led to realize that the language of "maximums" and "minimums" is simply not sufficient to quantify how large an infinite set is. **So what is a fix for this problem?**

[Concept 4 & 5] Introduction to Supremums and Infimums

If one picks any $x \in [0,1)$ and wants to declare x as the maximum of this interval then one can always come up a very small number $\epsilon > 0$ s.t $x + \epsilon \in [0,1)$ and thus $x + \epsilon$ becomes a more credible candidate for being called the maximum - and starting from $x + \epsilon$ we can again repeat this argument and this can be continued ad infinitum!

Thus we are led to realize that the language of "maximums" and "minimums" is simply not sufficient to quantify how large an infinite set is. **So what is a fix for this problem?**

Let's Think More Carefully!

[Concept 4 & 5] Introduction to Supremums and Infimums

Definition (**Upper Bound & Lower Bound)**

Let I be a subset of $\mathbb{R}$. Suppose there exists a number M s.t $x \leq M, \ \forall x \in I$. Then M is an upper bound on I. Suppose there exists a number m s.t $x \geq m, \ \forall x \in I$. Then m is a lower bound on I.

[Concept 4 & 5] Introduction to Supremums and Infimums

Definition (**Upper Bound & Lower Bound)**

Let I be a subset of $\mathbb{R}$. Suppose there exists a number M s.t $x \leq M,\ \forall x \in I$. Then M is an upper bound on I. Suppose there exists a number m s.t $x \geq m,\ \forall x \in I$. Then m is a lower bound on I.

For the case $I = [0,1)$ considered earlier realize that it has uncountably infinite number of upper and lower bounds.

(a) Every real number greater than 1 is an upper bound on I,
- and there are an uncountably infinite number of them.

[Concept 4 & 5] Introduction to Supremums and Infimums

Definition (**Upper Bound & Lower Bound)**

Let I be a subset of $\mathbb{R}$. Suppose there exists a number M s.t $x \leq M, \ \forall x \in I$. Then M is an upper bound on I. Suppose there exists a number m s.t $x \geq m, \ \forall x \in I$. Then m is a lower bound on I.

For the case $I = [0, 1)$ considered earlier realize that it has uncountably infinite number of upper and lower bounds.

(a) Every real number greater than 1 is an upper bound on I,
- and there are an uncountably infinite number of them.

(b) Every real number less than 0 is a lower bound on I,
- and there are an uncountably infinite number of them.

[Concept 4 & 5] Introduction to Supremums and Infimums

Definition (**Upper Bound & Lower Bound)**

Let I be a subset of $\mathbb{R}$. Suppose there exists a number M s.t $x \leq M,\ \forall x \in I$. Then M is an upper bound on I. Suppose there exists a number m s.t $x \geq m,\ \forall x \in I$. Then m is a lower bound on I.

For the case $I = [0, 1)$ considered earlier realize that it has uncountably infinite number of upper and lower bounds.

(a) Every real number greater than 1 is an upper bound on I,
- and there are an uncountably infinite number of them.

(b) Every real number less than 0 is a lower bound on I,
- and there are an uncountably infinite number of them.

Can we make these bounding numbers unique in some way while side-stepping the problem with maximums and minimums ?

[Concept 4 & 5] Introduction to Supremums and Infimums

Definition (**Supremum)**

Let I be a subset of $\mathbb{R}$. Suppose there exists an upper bound M on I s.t $M \leq M'$, $\forall$ upper bounds M' on I. Then M is the "least upper bound" on I and it's called the "supremum" of I and is denoted as $\sup I$. And when $\sup I \in I$ we call it the "maximum" of I, $\max I$.

[Concept 4 & 5] Introduction to Supremums and Infimums

Definition (**Supremum)**

Let I be a subset of $\mathbb{R}$. Suppose there exists an upper bound M on I s.t $M \leq M'$, $\forall$ upper bounds M' on I. Then M is the "least upper bound" on I and it's called the "supremum" of I and is denoted as $\sup I$. And when $\sup I \in I$ we call it the "maximum" of I, $\max I$.

Definition (**Infimum)**

Let I be a subset of $\mathbb{R}$. Suppose there exists a lower bound m on I s.t $m \leq m'$, $\forall$ lower bounds m' on I. Then m is the "greatest lower bound" on I and it's called the "infimum" of I and is denoted as $\inf I$. And when $\inf I \in I$ we call it the "minimum" of I, $\min I$.

For intuition, look at the previous example again to realize that (a) when $I = [0, 1)$, we have $\sup I = 1$ and $\inf I = 0 = \min I$.

[Concept 3] Introduction to Spectral Norms of Matrices

Definition

For any matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ we define its $2, 2$-norm as,

$$\|\boldsymbol{A}\|_{2,2} \coloneqq \sup_{\|\boldsymbol{x}\|_2 = 1} \|\boldsymbol{A}\boldsymbol{x}\|_2 = \sup_{\boldsymbol{x} \neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$$

When its clear from context, the above will also be called just the $2-$norm of $\boldsymbol{A}$ and denoted as $\|\boldsymbol{A}\|_2$ or spectral norm of $\boldsymbol{A}$.

[Concept 3] Introduction to Spectral Norms of Matrices

Let's See A Tricky Example Of It! :D

[Concept 3] Introduction to Spectral Norms of Matrices

Let's See A Tricky Example Of It! :D

Consider the following family of matrices,

$$\boldsymbol{A}(\theta) \coloneqq \begin{bmatrix} 0 & \theta \\ 0 & 0 \end{bmatrix}$$

One can solve the equation $\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$ to obtain the eigenvalues (λ) and eigenvectors $(\boldsymbol{v})$ for this matrix and realize that its only eigenvalue is 0.

[Concept 3] Introduction to Spectral Norms of Matrices

But, for an arbitrary vector on the unit circle in $\mathbb{R}^2$

- parameterized as say $\begin{bmatrix} \sin(\alpha) \\ \cos(\alpha) \end{bmatrix}$ we have,

$$\left\| \boldsymbol{A}(\theta) \begin{bmatrix} \sin(\alpha) \\ \cos(\alpha) \end{bmatrix} \right\| = \left\| \begin{bmatrix} \theta \cos(\alpha) \\ 0 \end{bmatrix} \right\| = |\theta| \cdot |\cos(\alpha)|$$

Thus we have,

$$\| \boldsymbol{A}(\theta) \| = \sup_{\alpha} |\theta| \cdot |\cos(\alpha)| = |\theta|$$

[Concept 3] Introduction to Spectral Norms of Matrices

But, for an arbitrary vector on the unit circle in $\mathbb{R}^2$

- parameterized as say $\begin{bmatrix} \sin(\alpha) \\ \cos(\alpha) \end{bmatrix}$ we have,

$$\left\| \boldsymbol{A}(\theta) \begin{bmatrix} \sin(\alpha) \\ \cos(\alpha) \end{bmatrix} \right\| = \left\| \begin{bmatrix} \theta \cos(\alpha) \\ 0 \end{bmatrix} \right\| = |\theta| \cdot |\cos(\alpha)|$$

Thus we have,

$$\| \boldsymbol{A}(\theta) \| = \sup_{\alpha} |\theta| \cdot |\cos(\alpha)| = |\theta|$$

So, by choosing θ arbitrarily large in magnitude we can make spectral norm of $\boldsymbol{A}(\theta)$ as large as we want, *while the largest eigenvalue magnitude (also called the "spectral radius" of a matrix) remains at* 0, $\forall \theta$.

[Concept 3] Introduction to Spectral Norms of Matrices

But, for an arbitrary vector on the unit circle in $\mathbb{R}^2$

- parameterized as say $\begin{bmatrix} \sin(\alpha) \\ \cos(\alpha) \end{bmatrix}$ we have,

$$\left\| \boldsymbol{A}(\theta) \begin{bmatrix} \sin(\alpha) \\ \cos(\alpha) \end{bmatrix} \right\| = \left\| \begin{bmatrix} \theta \cos(\alpha) \\ 0 \end{bmatrix} \right\| = |\theta| \cdot |\cos(\alpha)|$$

Thus we have,

$$\| \boldsymbol{A}(\theta) \| = \sup_{\alpha} |\theta| \cdot |\cos(\alpha)| = |\theta|$$

So, by choosing θ arbitrarily large in magnitude we can make spectral norm of $\boldsymbol{A}(\theta)$ as large as we want, *while the largest eigenvalue magnitude (also called the "spectral radius" of a matrix) remains at* 0*,* $\forall \theta$*.*

In general, one can show,

spectral radius $\leq$ spectral norm

– but the gap between them can be arbitrarily large!

[Concept 2] Introduction to Dimension of Sub-Spaces - of $\mathbb{R}^n$

Definition (**Span**)

Given a finite set of vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ each in the "vector space" $\mathbb{R}^n$, we denote,

$$\mathrm{Span}(\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}) \coloneqq \left\{ \sum_{i=1}^{k} w_i \boldsymbol{v}_i \mid w_i \in \mathbb{R}, \forall i = 1, \ldots, k \right\}$$

[Concept 2] Introduction to Dimension of Sub-Spaces - of $\mathbb{R}^n$

Definition (**Span**)

Given a finite set of vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ each in the "vector space" $\mathbb{R}^n$, we denote,

$$\text{Span}(\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}) \coloneqq \left\{ \sum_{i=1}^{k} w_i \boldsymbol{v}_i \mid w_i \in \mathbb{R}, \forall i = 1, \ldots, k \right\}$$

[Concept 2] Introduction to Dimension of Sub-Spaces - of $\mathbb{R}^n$

Definition (**Basis & Dimension)**

Given a finite set of mutually orthogonal vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$, each in $\mathbb{R}^n$, we denote,

$$p = \text{dimension}\left(\text{Span}(\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p\})\right)$$

$$\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p\} \text{ is an instance of a basis of } \text{Span}(\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p\})$$

[Concept 2] Introduction to Dimension of Sub-Spaces - of $\mathbb{R}^n$

Definition (**Basis & Dimension)**

Given a finite set of mutually orthogonal vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$, each in $\mathbb{R}^n$, we denote,

$$p = \text{dimension}\left(\text{Span}(\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p\})\right)$$

$$\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p\} \text{ is an instance of a basis of } \text{Span}(\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p\})$$

In other words – any set of vectors that can be written as a span of p mutually orthogonal vectors is said to be p–dimensional.

[Concept 1] Introduction to Rank of Matrices

Now, consider the following 3 matrices,

$$\boldsymbol{A} \coloneqq \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{B} \coloneqq \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}, \boldsymbol{C} \coloneqq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

[Concept 1] Introduction to Rank of Matrices

Now, consider the following 3 matrices,

$$\boldsymbol{A} \coloneqq \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{B} \coloneqq \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}, \boldsymbol{C} \coloneqq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

For any arbitrary vector $\boldsymbol{v} \coloneqq \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \in \mathbb{R}^2$,

we have its image under the matrices being,

$$\boldsymbol{A}[\boldsymbol{v}] = 2v_1\boldsymbol{e}_1 + v_2\boldsymbol{e}_2,\ \boldsymbol{B}[\boldsymbol{v}] = -v_1\boldsymbol{e}_1,\ \boldsymbol{C}[\boldsymbol{v}] = 0$$

[Concept 1] Introduction to Rank of Matrices

Now, consider the following 3 matrices,

$$\boldsymbol{A} \coloneqq \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{B} \coloneqq \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}, \boldsymbol{C} \coloneqq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

For any arbitrary vector $\boldsymbol{v} \coloneqq \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \in \mathbb{R}^2$,

we have its image under the matrices being,

$$\boldsymbol{A}[\boldsymbol{v}] = 2v_1\boldsymbol{e}_1 + v_2\boldsymbol{e}_2,\ \boldsymbol{B}[\boldsymbol{v}] = -v_1\boldsymbol{e}_1,\ \boldsymbol{C}[\boldsymbol{v}] = 0$$

Relating back to the earlier discussion we can observe,

$$\mathrm{dimension}\left(\{\boldsymbol{A}[\boldsymbol{v}] \mid \boldsymbol{v} \in \mathbb{R}^2\}\right) = 2,\ \mathrm{dimension}\left(\{\boldsymbol{B}[\boldsymbol{v}] \mid \boldsymbol{v} \in \mathbb{R}^2\}\right) = 1$$

$$\mathrm{dimension}\left(\{\boldsymbol{C}[\boldsymbol{v}] \mid \boldsymbol{v} \in \mathbb{R}^2\}\right) = 0$$

[Concept 1] Introduction to Rank of Matrices

The calculation above can be said to prove that that the **rank of the matrices A, B & C are $2, 1$ & 0 respectively**. *It is easy to see how to extend this discussion to arbitrary rectangular matrices.*

[Concept 1] Introduction to Rank of Matrices

The calculation above can be said to prove that that the **rank of the matrices $\boldsymbol{A}, \boldsymbol{B}$ & $\boldsymbol{C}$ are $2, 1$ & 0 respectively**. *It is easy to see how to extend this discussion to arbitrary rectangular matrices.*

Definition (**Image and Rank of a Matrix)**

Given any $m \times n$ matrix $\boldsymbol{T}$ i.e $\boldsymbol{T} : \mathbb{R}^n \to \mathbb{R}^m$, we define its "image" as,

$$\text{Image}(\boldsymbol{T}) \coloneqq \{\boldsymbol{T}[\boldsymbol{v}] \mid \boldsymbol{v} \in \mathbb{R}^n\}$$

. Then we can define,

$$\text{rank}(\boldsymbol{T}) \coloneqq \text{dimension}(\text{Image}(\boldsymbol{T}))$$

[Concept 1] Introduction to Rank of Matrices

The calculation above can be said to prove that that the **rank of the matrices A, B & C are $2, 1$ & 0 respectively**. *It is easy to see how to extend this discussion to arbitrary rectangular matrices.*

Definition (**Image and Rank of a Matrix**)

Given any $m \times n$ matrix $\boldsymbol{T}$ i.e $\boldsymbol{T} : \mathbb{R}^n \to \mathbb{R}^m$, we define its "image" as,

$$\text{Image}(\boldsymbol{T}) \coloneqq \{\boldsymbol{T}[\boldsymbol{v}] \mid \boldsymbol{v} \in \mathbb{R}^n\}$$

. Then we can define,

$$\text{rank}(\boldsymbol{T}) \coloneqq \text{dimension}(\text{Image}(\boldsymbol{T}))$$

- **If $\text{rank}(\boldsymbol{T}) = \min\{n, m\}$, then we call T to be of "full rank".**

Plan

1. Chapter-7 (a) : Introduction
2. Chapter-7 (b) : Some Key Mathematical Structures in M.L.
3. Chapter-8 : Formal Introduction to Gradient Descent
4. Chapter-9 : G.D. on Overparameterized Linear Regression
5. Chapter-10 : A Measure of Capacity : Rademacher Complexity
6. Chapter-11 : Computing the Rademacher Complexity

What is G.D?

In the last few years there has been a surge in literature on provable training of various kinds of neural nets in certain regimes of their widths or depths or for very specifically structured data

What is G.D?

In the last few years there has been a surge in literature on provable training of various kinds of neural nets in certain regimes of their widths or depths or for very specifically structured data

In the real world, for training the large neural net models, the algorithms that are used are far more complex – than just taking gradient steps.

What is G.D?

In the last few years there has been a surge in literature on provable training of various kinds of neural nets in certain regimes of their widths or depths or for very specifically structured data

In the real world, for training the large neural net models, the algorithms that are used are far more complex – than just taking gradient steps.

But to really appreciate the training technologies in use at the bleeding edge of modern data-science, we need to start from the most vanilla form of this algorithm i.e **"Gradient Descent" (G.D.)**

See, Sections 12.7 to 12.11 of "d2l.ai" for the various kinds of gradient based training methods that are most successful on modern massively over-parameterized neural nets.

What is G.D?

G.D. on a Differentiable Function $F : \mathbb{R}^p \to \mathbb{R}$

```
**Input:** An initial point $\boldsymbol{w}_1$ and a specification $T > 0$ of the number of steps for the G.D.
**Choose :** A step-size sequence $\eta_t > 0,\ \forall t = 1, 2, \ldots.$
**for** $t = 1, \ldots, T$ **do**
    $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \cdot \nabla F(\boldsymbol{w}_t)$
**end for**
**Output :** $\boldsymbol{w}_T$
```

What is G.D?

G.D. on a Differentiable Function $F : \mathbb{R}^p \to \mathbb{R}$

```
Input: An initial point $\boldsymbol{w}_1$ and a specification $T > 0$ of the number of steps for the G.D.
Choose : A step-size sequence $\eta_t > 0,\ \forall t = 1, 2, \ldots$.
for $t = 1, \ldots, T$ do
    $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \cdot \nabla F(\boldsymbol{w}_t)$
end for
Output : $\boldsymbol{w}_T$
```

The Key Question - Largely Unresolved!

"If F has global minima then when does G.D. find one of them,
– and which one and how fast?"

Introduction to Proof of Convergence of G.D.

Recall the idea of convergence of G.D. from the last class. In the following theorem we shall see an example of making that precise for a simple objective function $F(x) = x^2$
- that is strongly convex, differentiable and Lipschitz smooth.

Introduction to Proof of Convergence of G.D.

Recall the idea of convergence of G.D. from the last class. In the following theorem we shall see an example of making that precise for a simple objective function $F(x) = x^2$ - that is strongly convex, differentiable and Lipschitz smooth.

For this objective function, we can see that the G.D. steps are,

$$\begin{aligned} x_{t+1} =& x_t - \underbrace{\eta_t}_{\text{step–length}} F'(x_t) \\ =& x_t - \eta_t \cdot 2x_t \\ =& (1 - 2\eta_t)x_t \end{aligned}$$

Introduction to Proof of Convergence of G.D.

$$x_{t+1} = x_t - \underbrace{\eta_t}_{\text{step-length}} F'(x_t) = x_t - \eta_t \cdot 2x_t = (1 - 2\eta_t)x_t$$

Theorem (**Provable Rate of G.D. Convergence on x^2**)

Consider doing G.D. on the function $F(x) = x^2$, starting from any $x_0 \in \mathbb{R}$, $x_0 \neq 0$ and using any choice of constant step-length of $\frac{1-k}{2}$ for any $k \in (0,1)$. Then, (a) the iterates will asymptotically in time converge to the global minima at $x = 0$

&

(b) taking $\frac{\log\left(\frac{|x_0|}{\varepsilon}\right)}{\log \frac{1}{k}}$ number of steps is sufficient for the iterates to be within a distance of ε of the global minima, for all $\varepsilon > 0$.

Proof of Convergence of G.D. on x^2

Proof.

- Lets "unroll" $x_{t+1} = (1 - 2\eta_t)x_t$, to get,

$$\begin{aligned} x_{t+1} &= (1 - 2\eta_t)x_t = (1 - 2\eta_t)(1 - 2\eta_{t-1})x_{t-1} \\ &= (1 - 2\eta_t)(1 - 2\eta_{t-1})...(1 - 2\eta_{t-t})x_{t-t} \\ &= x_0 \prod_{i=0}^{t}(1 - 2\eta_i) \end{aligned}$$

Proof of Convergence of G.D. on x^2

Proof.

- Lets "unroll" $x_{t+1} = (1 - 2\eta_t)x_t$, to get,

$$\begin{aligned} x_{t+1} &= (1 - 2\eta_t)x_t = (1 - 2\eta_t)(1 - 2\eta_{t-1})x_{t-1} \\ &= (1 - 2\eta_t)(1 - 2\eta_{t-1})...(1 - 2\eta_{t-t})x_{t-t} \\ &= x_0 \prod_{i=0}^{t}(1 - 2\eta_i) \end{aligned}$$

- Lets observe a few things about this above equation!

Proof of Convergence of G.D. on x^2

Proof.

- Lets "unroll" $x_{t+1} = (1 - 2\eta_t)x_t$, to get,

$$\begin{aligned} x_{t+1} &= (1 - 2\eta_t)x_t = (1 - 2\eta_t)(1 - 2\eta_{t-1})x_{t-1} \\ &= (1 - 2\eta_t)(1 - 2\eta_{t-1})...(1 - 2\eta_{t-t})x_{t-t} \\ &= x_0 \prod_{i=0}^{t}(1 - 2\eta_i) \end{aligned}$$

- Lets observe a few things about this above equation!
  - If we had $x_0 = 0$ then it implies that for all times t, we would have $x_t = 0$. Thus the algorithm would not move at all. This is why it was important to assume in the theorem that $x_0 \neq 0$.

Proof of Convergence of G.D. on x^2

Proof.

- Lets "unroll" $x_{t+1} = (1 - 2\eta_t)x_t$, to get,

$$\begin{aligned} x_{t+1} &= (1 - 2\eta_t)x_t = (1 - 2\eta_t)(1 - 2\eta_{t-1})x_{t-1} \\ &= (1 - 2\eta_t)(1 - 2\eta_{t-1})...(1 - 2\eta_{t-t})x_{t-t} \\ &= x_0 \prod_{i=0}^{t}(1 - 2\eta_i) \end{aligned}$$

- Lets observe a few things about this above equation!
    - If we had $x_0 = 0$ then it implies that for all times t, we would have $x_t = 0$. Thus the algorithm would not move at all. This is why it was important to assume in the theorem that $x_0 \neq 0$. More generally if the algorithm were to start at a point x_0 s.t $f'(x_0) = 0$ i.e x_0 is a "critical point" of f, then the algorithm would not move at all. **Thus critical points are not to be used as starting points of G.D. algorithms.**

Proof of Convergence of G.D. on x^2

Proof.

- We had "unrolled" $x_{t+1} = (1 - 2\eta_t)x_t$, to get, $x_{t+1} = x_0 \prod_{i=0}^{t}(1 - 2\eta_i)$
- Lets observe a few things about this above equation!

Proof of Convergence of G.D. on x^2

Proof.

- We had "unrolled" $x_{t+1} = (1 - 2\eta_t)x_t$, to get, $x_{t+1} = x_0 \prod_{i=0}^{t} (1 - 2\eta_i)$
- Lets observe a few things about this above equation!
    - We want $x_t \to 0$ (the global minima of f) as $t \to \infty$,

Proof of Convergence of G.D. on x^2

Proof.

- We had "unrolled" $x_{t+1} = (1 - 2\eta_t)x_t$, to get, $x_{t+1} = x_0 \prod_{i=0}^{t} (1 - 2\eta_i)$
- Lets observe a few things about this above equation!
  - We want $x_t \to 0$ (the global minima of f) as $t \to \infty$,
    & from above it follows that a simple sufficient condition for that to happen is if for some constant $k \in (0,1)$, $(1 - 2\eta_i) = k, \ \forall i$
    – and this can be ensured by choosing a constant step-length as,

$$\eta_t = \frac{1}{2} \cdot (1 - k)$$

Proof of Convergence of G.D. on x^2

Proof.

- We had "unrolled" $x_{t+1} = (1 - 2\eta_t)x_t$, to get, $x_{t+1} = x_0 \prod_{i=0}^{t} (1 - 2\eta_i)$
- Lets observe a few things about this above equation!
  - We want $x_t \to 0$ (the global minima of f) as $t \to \infty$, & from above it follows that a simple sufficient condition for that to happen is if for some constant $k \in (0,1)$, $(1 - 2\eta_i) = k, \ \forall i$ – and this can be ensured by choosing a constant step-length as,

$$\eta_t = \frac{1}{2} \cdot (1 - k)$$

- Making the above choice we have, $(1 - 2\eta_i) = k, \ \forall i$, and we can rewrite what we previously had for x_{t+1} as, $x_{t+1} = x_0 \prod_{i=0}^{t} (k) = x_0 \cdot k^t$

□

Proof of Convergence of G.D. on x^2

Proof.

- Thus we have,

$$\lim_{t \to \infty} x_t = \lim_{t \to \infty} x_0 \cdot k^t = 0$$

Proof of Convergence of G.D. on x^2

Proof.

- Thus we have,

$$\lim_{t\to\infty} x_t = \lim_{t\to\infty} x_0 \cdot k^t = 0$$

- Thus, we have shown that for a range of choices of constant step-length, gradient descent converges on x^2 to its global minimum, regardless of x_0. **BUT, we can get more precise! Lets see!**

Proof of Convergence of G.D. on x^2

Proof.

- Thus we have,

$$\lim_{t\to\infty} x_t = \lim_{t\to\infty} x_0 \cdot k^t = 0$$

- Thus, we have shown that for a range of choices of constant step-length, gradient descent converges on x^2 to its global minimum, regardless of x_0. **BUT, we can get more precise! Lets see!**
- To obtain convergence time ("non-asymptotic") bounds, we check how long it takes to get within a $\varepsilon > 0$ interval of the global minima.

$$|x_{t+1}| \le \varepsilon \implies |x_0| \cdot k^t \le \varepsilon \implies t \ge \frac{\log\left(\frac{|x_0|}{\varepsilon}\right)}{\log \frac{1}{k}}$$

Proof of Convergence of G.D. on x^2

Proof.

- Thus we have,

$$\lim_{t\to\infty} x_t = \lim_{t\to\infty} x_0 \cdot k^t = 0$$

- Thus, we have shown that for a range of choices of constant step-length, gradient descent converges on x^2 to its global minimum, regardless of x_0. **BUT, we can get more precise! Lets see!**
- To obtain convergence time ("non-asymptotic") bounds, we check how long it takes to get within a $\varepsilon > 0$ interval of the global minima.

$$|x_{t+1}| \leq \varepsilon \implies |x_0| \cdot k^t \leq \varepsilon \implies t \geq \frac{\log\left(\frac{|x_0|}{\varepsilon}\right)}{\log \frac{1}{k}}$$

Thus we have arrived at the fact that we set out to prove.

This is Only the Beginning!

One of the key insights in the theory of G.D. is that one of the basic ways in which a function gets non-trivial for G.D. to work with is if the objective is **either non-convex or non-Lipschitz smooth**

– *as is almost always the case in the real world!*.

This is Only the Beginning!

One of the key insights in the theory of G.D. is that one of the basic ways in which a function gets non-trivial for G.D. to work with is if the objective is **either non-convex or non-Lipschitz smooth**
– *as is almost always the case in the real world!*.

Thus we need to start to go beyond the confines of smooth convex functions.

Demonstration of G.D. convergence, on a differentiable convex function which is – neither Lipschitz nor Lipschitz smooth.

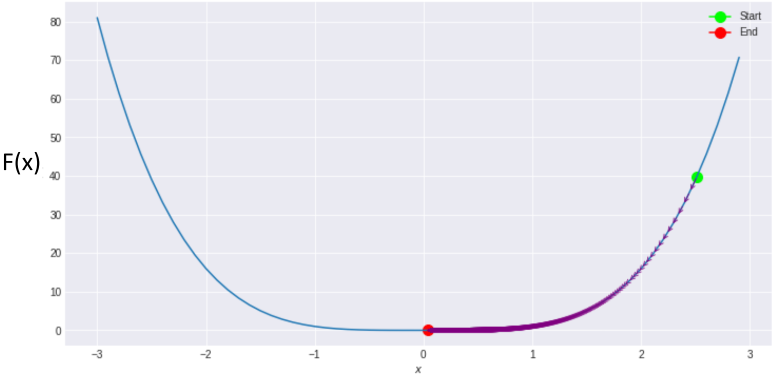


Figure 9: Progress to the unique global minima of G.D. on the function $F(x) = x^4$ for step-size $\eta = 10^{-3}$ and $T = 10^5$ and the algorithm starting near 2.5

Demonstration of G.D. convergence, on a differentiable function which is – neither convex, nor Lipschitz nor Lipschitz smooth.

This instance is more complex than the previous one - the target function here is not even convex! (Coming Up in Example Session!)

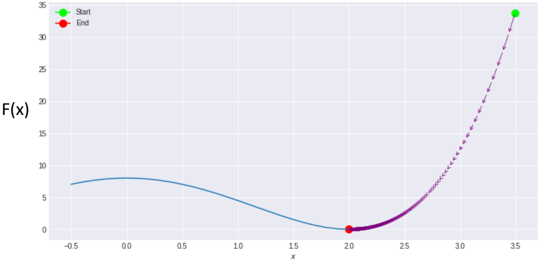


Figure 10: Progress to a global minima of G.D. on the function $F(x) = \frac{1}{2} \cdot (x^2 - 2^2)^2$ for $\eta = 10^{-3}$ and $T = 10^4$ and the algorithm starting near 3.5

Glimpses of the Road Beyond this Course

The situations as in this Mexican-Hat example above *are not at all rare* – in ways this is reflective of some of the key features of how actual deep-learning happens.

Glimpses of the Road Beyond this Course

The situations as in this Mexican-Hat example above *are not at all rare* – in ways this is reflective of some of the key features of how actual deep-learning happens.

Thus we saw two quite varied examples where this deceptively simple looking gradient descent algorithm seems to be finding the global minima and also while using constant step-sizes.

How does one prove convergence in such increasingly hard cases?

Glimpses of the Road Beyond this Course

The situations as in this Mexican-Hat example above *are not at all rare* – in ways this is reflective of some of the key features of how actual deep-learning happens.

Thus we saw two quite varied examples where this deceptively simple looking gradient descent algorithm seems to be finding the global minima and also while using constant step-sizes.

How does one prove convergence in such increasingly hard cases?

(Do a Ph.D.!)