Name: *Caleb McWhorter— Solutions*
MAT 397— Fall 2020
Applied Problems:
Gradients & Direc. Deriv.

"Premature optimization is the root of all evil."
–Donald Knuth

Gradient Descent & Machine Learning

We have been promised intelligent AI for decades. So what is the holdup? For years, computer programs tried to program intelligence line-by-line—how ironic! This means for a given computer task, one would have to program every possibility that the computer could face. For example, if you were programming an AI which could respond to human asking the computer how it was, you would have to program all the possible ways to form that question: "How are you?", "How are you doing?", "How are you feeling?", "What's up?", etc. While much progress was possible with this restrictive approach to AI, clearly, a better approach was needed. It wasn't until the 1980s, and really the 1990s–2000s with more technology available, that the probable 'correct' approach was found—machine learning. Rather than tell the computer what to do in every situation, you would 'teach' the computer how to learn. Then you train the computer by applying this learning algorithm to datasets, and 'intelligence' emerges. Machine learning is a specific type of artificial intelligence, and the most powerful of machine learning techniques is called *deep learning*, which relies on *neural nets*.

Essentially, (supervised) machine learning works as follows: suppose you wanted to predict which people might default on their home loans. A professional, such as a banker, could look at the data for an individual and make an assessment of how likely a person is to default. How do we teach a computer to do the same? We gather a lot of relevant data on home loans: age, income level, credit rating, geographic location, number of dependents, etc, and whether the person defaulted on their home loan or not. We then feed this data to a computer, asking it for each person in the dataset whether given their associated data, whether it thinks they will default or not. Based on whether it was right or wrong, the algorithm alters how it makes predictions. We repeat the process many times, and in 'most' cases, the predictions improve. This is how computers are now better than humans at tasks like identifying images, and the games Chess, Go, Jeopardy, etc.

Notice deep learning hinges on the ability for the computer to systematically 'tweak' its guessing system at each stage. This is ultimately an optimization problem, where the computer seeks to find a 'best possible' weighting system of the variables that minimizes its chances of being incorrect. [This is a vast oversimplification, but carries the essence of what happens.] Of course, how does the computer know how to change the weightings at each stage? A common method is *gradient descent*.

Imagine the probability of being incorrect, L, for a computer prediction depends on two weights, say x, y. Then $L(x, y)$ is a surface in three dimensions. Given a set of weights, (x, y), we want to know how to change our weights, x, y, to minimize $L(x, y)$. Imagine standing on the surface $L(x, y)$ at the point (x, y). The gradient, ∇L, will give a direction we can travel in to increase $L(x, y)$ the quickest. Therefore, $-\nabla L(x, y)$ will give a direction we can move in to decrease $L(x, y)$ the fastest—and this is our goal! So we could take a 'step' in the direction of $-\nabla L(x, y)$. How far should we step? That's a difficult problem in these computer learning methods, but say we use step size dt. Then $-\nabla L(x, y)\ dt$ will tell us how much to change x and y; that is, the x and y components of $-\nabla L(x, y)\ dt$ tells us how much to change x and y, respectively. We then repeat the process again and again. Hopefully, this 'walks' us to a minimum value for $L(x, y)$. This process is called gradient descent. If you wanted to maximize $L(x, y)$, you would use this process with $\nabla L(x, y)$—gradient ascent.

Problem:

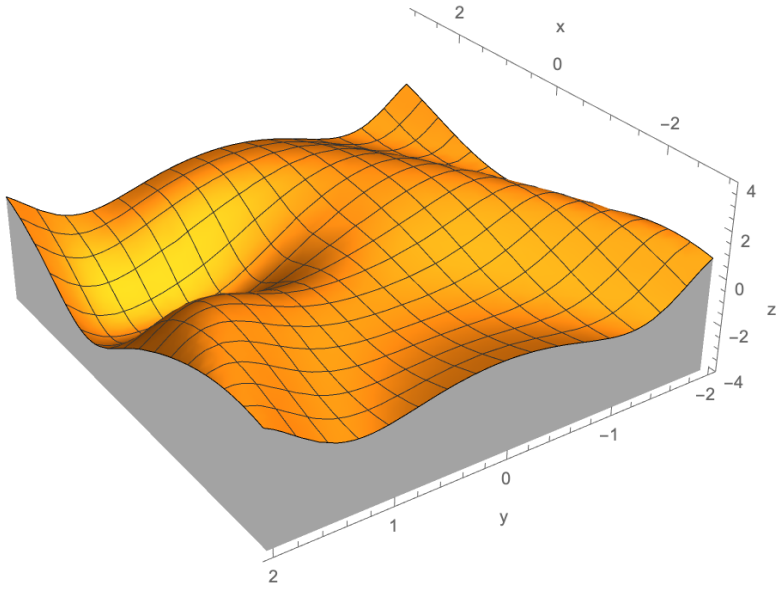
(a) Suppose you want to minimize $F(x, y)$. You decide to use a gradient descent method to find an approximate minimum value for $F(x, y)$. You start with a point $\mathbf{x}_0 = (x_0, y_0)$, and will use step size Δt. Explain why $\mathbf{x}_{n+1} = \mathbf{x}_n - \Delta t \nabla F(\mathbf{x}_n)$ is your gradient descent method.

(b) Is it true that $F(\mathbf{x}_{n+1}) \leq F(\mathbf{x}_n)$? Explain.

(c) There are many problems that can arise in gradient descent. Take the function $f(x, y) = x^2 + y^2$. Try to approximate a minimum value for $f(x, y)$ by using two steps of the gradient descent method using $\Delta t = 1$ and initial point $(x_0, y_0) = (2, 2)$. What goes wrong? What can you change about Δt to fix this?

(d) What can be problematic with your suggestion in (c)? [Hint: Take your suggestion in (c) to an extreme.]

(e) Consider again the problem that arose in (c). A possible solution might be to adjust your choice of $\mathbf{x}_0 = (x_0, y_0)$. Consider applying the gradient descent method to some starting point on the surface given in the figure below. By carefully considering what happens when choosing different initial points (x_0, y_0), explain another issue that can arise when using the gradient descent method.



(f) Let $f(x, y) = y^4 - 2xy^2 + x^3 - x + 3$. This function has minima at $(1, 1)$ and $(1, -1)$. Moreover, $f(1, 1) = f(1, -1) = 2$. Using $\Delta t = 0.1$ and start point $(x_0, y_0) = (1.5, 2.2)$, perform five steps of the gradient descent method. How close to the minima point and minimum value does this procedure produce?

(g) For the function given in (f), suppose you use the gradient descent method starting at $(1, 0)$. Because the point $(1, 0)$ is 'midway' between the minima $(1, 1)$ and $(1, -1)$, will the gradient descent method simply 'stay' at the point $(1, 0)$—unable to 'decide' between $(1, 1)$ and $(1, -1)$? Explain.

(h) Is $\nabla F(\mathbf{x_n}) = \mathbf{0}$ possible in the gradient descent method, either at the initial or some subsequent point? What would happen to the method, assuming it could happen?

Solution.

(a) The point $\mathbf{x}_0 = (x_0, y_0)$ is the starting point. Then $\nabla F(\mathbf{x}_0)$ gives the direction of maximum increase for F at $\mathbf{x}_0$. Therefore, -$\nabla F(\mathbf{x}_0)$ gives the direction of maximum decrease for F at $\mathbf{x}_0$. Then $\Delta t \nabla F(\mathbf{x}_0)$ scales $\nabla F(\mathbf{x}_0)$. Then $\mathbf{x}_1 = \mathbf{x}_0 - \Delta t \nabla F(\mathbf{x}_0)$ takes the point $\mathbf{x}_0$ and adjusts it by $-\Delta t \nabla F(\mathbf{x}_0)$ to create a new guess, $\mathbf{x}_1$. Then generally, $\mathbf{x}_{n+1} = \mathbf{x}_n - \Delta t \nabla F(\mathbf{x}_n)$ is the new adjusted guess based on the previous guess $\mathbf{x}_n$.

(b) Just because the gradient descent method is trying to find a minimum does not necessarily mean that we decrease at each step (or ever). For example, imagine standing at the edge of a slowly steeped hole. The vector $-\lambda F(x, y)$ will point down towards the center of the hole. But with a large enough Δt, you will over step and be on the other side the hole, perhaps even higher up than before. We see an example of this in (c).

(c) First, note that $\nabla F(x, y) = \langle 2x, 2y \rangle$. We have chosen $\Delta t = 1.$ and $\mathbf{x}_0 = (x_0, y_0) = (2, 2)$. Then we have Notice because the step size is so large, we overstep the minimum and go to the 'other

n	(x_n, y_n)	$\nabla F(x_n, y_n)$	$\Delta t \nabla F(x_n, y_n)$	$(x_n, y_n) - \Delta t \nabla F(x_n, y_n)$
0	$(2, 2)$	$\langle 4, 4 \rangle$	$\langle 4, 4 \rangle$	$(-2, -2)$
1	$(-2, -2)$	$\langle -4, -4 \rangle$	$\langle -4, -4 \rangle$	$(2, 2)$
2	$(2, 2)$	$\langle 4, 4 \rangle$	$\langle 4, 4 \rangle$	$(-2, -2)$
3	$(-2, -2)$	$\langle -4, -4 \rangle$	$\langle -4, -4 \rangle$	$(2, 2)$

side' of it, and so on and so forth. One could change the step size to be smaller to avoid this type of issue.

(d) If you make Δt 'very' small, then at each stage, you aren't really moving anywhere, i.e. $\Delta \mathbf{x}_n$ is very small. Therefore, it will take many steps for the gradient descent method to converge (if it converges at all). This increases the computational expense of the gradient descent method.

(e) Choosing one point and 'heading downhill' via the gradient, you may end up at one minimum that is not an optimum minimum. For example, choose a starting point $(-2, -1)$ seems to take you towards the point on the surface given by $(-4, 1)$ or $(-4, -1)$. However, choosing starting point $(1, 0)$ would seem to take you towards the 'pit' on the surface at $(0, 1)$, which seems to be a better minimum.

(f) We end at the point $(1.0175, -1.01367)$. We know the actual minimum is at $(1, -1)$. So we

n	(x_n, y_n)	$\nabla F(x_n, y_n)$	$\Delta t \nabla F(x_n, y_n)$	$(x_n, y_n) - \Delta t \nabla F(x_n, y_n)$	$F(x_{n+1}, y_{n+1})$
0	(1.5,2.2)	$\langle -3.93, 29.392 \rangle$	$\langle -0.393, 2.9392 \rangle$	$(1.893, -0.7392)$	3.98158
1	$(1.893, -0.7392)$	$\langle 8.65751, 3.98158 \rangle$	$\langle 0.865751, 0.398158 \rangle$	$(1.02725, -1.13736)$	2.07244
2	$(1.02725, -1.13736)$	$\langle -0.421446, -1.21167 \rangle$	$\langle -0.0421446, -0.121167 \rangle$	$(1.06939, -1.01619)$	2.01132
3	$(1.06939, -1.01619)$	$\langle 0.365518, 0.149377 \rangle$	$\langle 0.0365518, 0.0149377 \rangle$	$(1.03284, -1.03113)$	2.00312
4	$(1.03284, -1.03113)$	$\langle 0.073832, -0.125322 \rangle$	$\langle 0.0073832, -0.0125322 \rangle$	$(1.02546, -1.0186)$	2.00146
5	$(1.02546, -1.0186)$	$\langle 0.0796168, -0.0492199 \rangle$	$\langle 0.00796168, -0.00492199 \rangle$	$(1.0175, -1.01367)$	2.00072

are a distance of 0.0222063 away from the minimum (x, y) location. Our predicted minimum value is 2.00072, whereas the actual minimum value is 2. Therefore, our predicted minimum value has magnitude of error 0.00072.

(g) No. First, while $(1, 0)$ is not even the midpoint of $(1, 1)$ and $(1, -1)$. Second, we are 'moving' on the surface and the points in the plane have little to do with the topology of the surface. Third, what would it mean to speak of the 'midpoint' on a surface? But fourth and most importantly, $\nabla F \neq \mathbf{0}$ unless we are at either a maximum, minimum, or saddle (because we will deal with continuously differentiable functions). So if we do not start at a maximum, minimum, or saddle point, then $\nabla F \neq \mathbf{0}$, and F 'already has a direction in mind.' And in our iterative process, if we 'land' at a point where $\nabla F = \mathbf{0}$, then we must be at a saddle or minimum by our algorithm construction.

(h) We have already answered this in (g). We know $\nabla F \neq \mathbf{0}$ unless we are at either a maximum, minimum, or saddle (because we will deal with continuously differentiable functions). So if $\nabla F = \mathbf{0}$ at the start, we are either done or need choose a new initial point. And if we find $\nabla F = \mathbf{0}$ in our algorithm, we have reached either a minimum or saddle point. If we do get $\nabla F = \mathbf{0}$ in the algorithm, you want to make sure the algorithm terminates because all subsequent points will be the same (because you are no longer moving).