

MAT237 Multivariable Calculus with Proofs



Due Friday November 19, 2021 by 13:00 ET

Instructions

This problem set is based on **Module C: Derivatives** (C6 to C11). Please read the **Problem Set FAQ** for details on submission policies, collaboration rules, and general instructions.

- **Problem Set 3 sessions are held on Tuesday November 16, 2021 in tutorial.** You will work with peers and get help from TAs. Before attending, seriously attempt these problems and prepare initial drafts.
- **Submissions are only accepted by Gradescope.** Do not send anything by email. Late submissions are not accepted under any circumstance. Remember you can resubmit anytime before the deadline.
- **Submit your polished solutions using only this template PDF.** You will submit a single PDF with your full written solutions. If your solution is not written using this template PDF (scanned print or digital) then you will receive zero. Do not submit rough work. Organize your work neatly in the space provided.
- **Show your work and justify your steps** on every question, unless otherwise indicated. Put your final answer in the box provided, if necessary.

We recommend you write draft solutions on separate pages and afterwards write your polished solutions here. You must fill out and sign the academic integrity statement below; otherwise, you will receive zero.

Academic integrity statement

Full Name: **Jace Alloway** _____

Student number: **1006940802** _____


Full Name: _____

Student number: _____

I confirm that:

- I have read and followed the policies described in the **Problem Set FAQ**.
- I have read and understand the rules for collaboration on problem sets described in the Academic Integrity subsection of the syllabus. I have not violated these rules while writing this problem set.
- I understand the consequences of violating the University's academic integrity policies as outlined in the **Code of Behaviour on Academic Matters**. I have not violated them while writing this assessment.

By signing this document, I agree that the statements above are true.

Signatures: 1)  _____

Problems

1. Caleb is a data scientist building a machine learning algorithm for a neural network. At each step of the algorithm, he needs to choose n variables which minimize a C^1 error function $E : \mathbb{R}^n \rightarrow [0, \infty)$. Unfortunately, since n is extremely large and E is quite complicated, it is hopeless to solve this optimization problem algebraically. Caleb tries to numerically solve this problem using the method of *gradient descent*.

(Optional: Watch [this video on gradient descent](#). No specific formulas are needed.)

Caleb has an initial guess $x_0 \in \mathbb{R}^n$ for the minimum and, based on a numerical methods textbook that he reads, he fixes a constant $h > 0$ and computes the sequence

$$\forall k \in \mathbb{N}, \quad x_{k+1} = x_k - h \nabla E(x_k).$$

When the algorithm updates from x_k to x_{k+1} , Caleb claims that the error reduces by approximately $h \|\nabla E(x_k)\|^2$. Justify Caleb's claim.

Fix $k \in \mathbb{N}$ and let $x_k \in \mathbb{R}^n$. Caleb is wanting to find the linear approximation of E at x_k towards x_{k+1} in the direction of the negative of the gradient of E , the direction of steepest descent.

This direction is given by the vector $v = -\nabla E(x_k)$.

Since E is C^1 on \mathbb{R}^n , E is differentiable on \mathbb{R}^n and its first derivative is continuous at every point in \mathbb{R}^n .

By **Definition 3.5.1**, the differential of E at x_k dE_{x_k} exists. By **Lemma 3.4.3** and **Theorem 3.5.8**, the differential is equivalent to

$$dE_{x_k}(v) = \nabla E(x_k) \cdot v$$

For a small fixed $h > 0$, the linear approximation of E in the direction $-\nabla E(x_k)$ is then given by

$$E(x_{k+1}) \approx E(x_k) + h dE_{x_k}(v),$$

which then becomes

$$E(x_{k+1}) - E(x_k) \approx h dE_{x_k}(-\nabla E(x_k)) = -h \|\nabla E(x_k)\|^2.$$

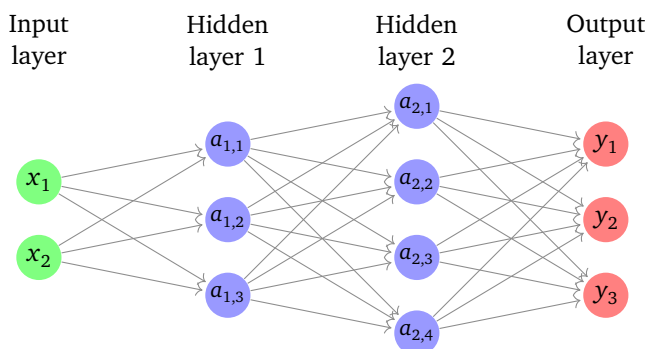
This implies that for every k , the error of the linear approximation is given by $\Delta E = E(x_{k+1}) - E(x_k)$, meaning that the value of ΔE decreases (or reduces) by the value $h \|\nabla E(x_k)\|^2$ for each k , hence the minus sign.

This justifies Caleb's claim.

2. Neural networks are used in machine learning to solve complicated problems. The basic idea is twofold: first, design a mechanism to predict the correct output given some inputs; second, update this mechanism using training data to improve the quality of predictions. A critical aspect of the second step is *backpropagation*. Backpropagation is an application of the chain rule.

(Optional: Watch [this video summarizing neural networks](#). Ignore the specific formulas.)

Here is an example of a neural network. Each node is called a neuron.



Each layer has links to the next layer, i.e. a neuron is a function of each neuron in the previous layer. Here is a brief description of a general neural network with L hidden layers, n inputs, and m outputs.

- Input layer has n inputs $x = (x_1, x_2, x_3, \dots, x_n) \in \mathbb{R}^n$.
- There are $L \in \mathbb{N}^+$ hidden layers. The input layer is layer 0 and the output layer is the layer $L + 1$.
- Fix $\ell \in \{1, \dots, L\}$. Hidden layer ℓ has $N(\ell)$ neurons $a_\ell = (a_{\ell,1}, a_{\ell,2}, \dots, a_{\ell,N(\ell)})$. Each component is a real-valued C^1 function of the previous layer. That is, for every $j \in \{1, 2, \dots, N(\ell)\}$, the quantity $a_{\ell,j}$ is a real-valued C^1 function of $a_{\ell-1}$. If $\ell = 1$ then $a_{1,j}$ is a real-valued C^1 function of the input layer x .
- Output layer has m predicted outputs $y = (y_1, y_2, \dots, y_m)$. Each component is a real-valued C^1 function of hidden layer L . That is, for $i \in \{1, 2, \dots, m\}$, the quantity y_i is a real-valued C^1 function of a_L .

Ignore the diagram's neural network and assume the parameters $L, N(1), \dots, N(L), m$, and n are arbitrary.

No justification is necessary for any part of this question.

- (2a) You are writing an algorithm that will update some neurons at each step. You need to efficiently calculate how much y_i will change depending on x_j assuming you only modify some neurons in hidden layers.

Fix $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. Use Leibniz notation to express $\frac{\partial y_i}{\partial x_j}$ in terms of the intermediate

partial derivatives of consecutive layers such as $\frac{\partial a_{2,4}}{\partial a_{1,1}}$. You will need L iterated sums $\sum \sum \dots \sum$.

$$\frac{\partial y_i}{\partial x_j} = \sum_{n_L=1}^{N(L)} \sum_{n_{L-1}=1}^{N(L-1)} \dots \sum_{n_3=1}^{N(3)} \sum_{n_2=1}^{N(2)} \sum_{n_1=1}^{N(1)} \frac{\partial y_i}{\partial a_{L,n_L}} \frac{\partial a_{L,n_L}}{\partial a_{L-1,n_{L-1}}} \dots \frac{\partial a_{2,n_2}}{\partial a_{1,n_1}} \frac{\partial a_{1,n_1}}{\partial x_j}$$

- (2b) For each $\ell \in \{0, 1, \dots, L\}$, use Leibniz notation to express the Jacobian matrix A_ℓ associated to layer ℓ and $\ell + 1$ in terms of partial derivatives of real-valued functions. Also, specify the dimensions of A_ℓ . Note that you will need to define A_0 and A_L separately.

$$DA_\ell = \begin{bmatrix} \frac{\partial a_{(\ell+1),1}}{\partial a_{\ell,1}} & \frac{\partial a_{(\ell+1),1}}{\partial a_{\ell,2}} & \cdots & \frac{\partial a_{(\ell+1),1}}{\partial a_{\ell,N(\ell)}} \\ \frac{\partial a_{(\ell+1),2}}{\partial a_{\ell,1}} & \frac{\partial a_{(\ell+1),2}}{\partial a_{\ell,2}} & \cdots & \frac{\partial a_{(\ell+1),2}}{\partial a_{\ell,N(\ell)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial a_{(\ell+1),N(\ell+1)}}{\partial a_{\ell,1}} & \frac{\partial a_{(\ell+1),N(\ell+1)}}{\partial a_{\ell,2}} & \cdots & \frac{\partial a_{(\ell+1),N(\ell+1)}}{\partial a_{\ell,N(\ell)}} \end{bmatrix}$$

where ℓ is the layer, $N(\ell)$ is the number of neurons in layer ℓ . The rate of change of each component of layer $\ell + 1$ depends on each component in layer ℓ . At the ends of the network,

$$DA_0 = \begin{bmatrix} \frac{\partial a_{0,1}}{\partial x_1} & \frac{\partial a_{0,1}}{\partial x_2} & \cdots & \frac{\partial a_{0,1}}{\partial x_n} \\ \frac{\partial a_{0,2}}{\partial x_1} & \frac{\partial a_{0,2}}{\partial x_2} & \cdots & \frac{\partial a_{0,2}}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial a_{0,N(0)}}{\partial x_1} & \frac{\partial a_{0,N(0)}}{\partial x_2} & \cdots & \frac{\partial a_{0,N(0)}}{\partial x_n} \end{bmatrix}, \quad DA_L = \begin{bmatrix} \frac{\partial y_1}{\partial a_{L,1}} & \frac{\partial y_1}{\partial a_{L,2}} & \cdots & \frac{\partial y_1}{\partial a_{L,N(L)}} \\ \frac{\partial y_2}{\partial a_{L,1}} & \frac{\partial y_2}{\partial a_{L,2}} & \cdots & \frac{\partial y_2}{\partial a_{L,N(L)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial a_{L,1}} & \frac{\partial y_m}{\partial a_{L,2}} & \cdots & \frac{\partial y_m}{\partial a_{L,N(L)}} \end{bmatrix},$$

where of course the components of layer $\ell = 0$ depend on the input layer x components and the output layer y components depend on the $\ell = L$ layer components.

The dimensions of each DA_ℓ for $\ell = 1, \dots, L$ are $N(\ell + 1) \times N(\ell)$, while DA_0 are $N(0) \times j$ and DA_L are $i \times N(L)$.

- (2c) A piece of training data is a fixed input $\hat{x} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n) \in \mathbb{R}^n$ with a fixed correct output $\hat{y} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m) \in \mathbb{R}^m$. Once you've specified the functions for a_1, \dots, a_L and y , your neural network predicts an output of $y(\hat{x})$ instead of the fixed correct \hat{y} . The cost function $C : \mathbb{R}^n \rightarrow [0, \infty)$ for this piece of training data is

$$C(x) = \|\hat{y} - y(x)\|^2.$$

The value $C(\hat{x})$ measures the quality of your prediction for this piece of training data. Backpropagation quantifies how the cost function changes when you modify neurons and links in your neural network. This step in machine learning algorithms helps choose how to best modify the neural network. Express the gradient $\nabla C(\hat{x})$ in terms of your matrices from (2b) and one additional matrix.

$$\nabla C(\hat{x}) = 2 \cdot (DA_L DA_{L-1} \dots DA_\ell \dots DA_1 DA_0)^T \begin{bmatrix} y(\hat{x}) - \hat{y} \\ y(\hat{x}) - \hat{y} \\ \vdots \\ y(\hat{x}) - \hat{y} \end{bmatrix}.$$

3. Let $A \subseteq \mathbb{R}^n$ be a compact set. Fix $a \in A^\circ$. Let f be a real-valued function that is continuous on $A \setminus \{a\}$ and is differentiable on its interior. Assume f has a unique critical point p in $A \setminus \{a\}$. Prove that if $\lim_{x \rightarrow a} f(x) = -\infty$ and $f(x) < f(p)$ for all $x \in \partial A$ with $x \neq a$, then f attains a global maximum at p . *Hint: Apply PS2 Q7.*

Proof.

- Since $\forall x \in \partial A$ with $x \neq a$, $f(x) < f(p)$, it follows that $p \notin \partial(A \setminus \{a\})$ if p is a unique critical point. By **Lemma 2.2.21**, it must be that $p \in (A \setminus \{a\})^\circ$.
- By **PS2 Q7**, since $\lim_{x \rightarrow a} f(x) = -\infty$, f attains a maximum on the set $A \setminus \{a\}$. Since $f(p)$ is unique, it then follows that $f(p)$ is the only critical point in $(A \setminus \{a\})^\circ$ but also the maximum of f on $(A \setminus \{a\})^\circ$.
- If $\forall x \in \partial(A \setminus \{a\})$, $f(x) < f(p)$, it directly follows that if there are any critical points in $\partial(A \setminus \{a\})$, they will be all less than $f(p)$.
- Therefore $f(p)$ is the global maximum of f on $A \setminus \{a\}$.

■

4. Mega-gaming company SO-KNEE is preparing to simultaneously launch two of their next-generation consoles: the PlayStation 3 and the Nintendo Ritch. They want to sell each console but their prices will affect each other's demand. If they price the PlayStation at p_1 dollars per unit and the Ritch at p_2 dollars per unit, then they will sell q_1 thousand PlayStation units and q_2 thousand Ritch units. A consulting company models how much they will sell depending on the prices and they suggest that

$$q_1 = 2500 - 2p_1 - p_2, \quad q_2 = 2000 - p_1 - 3p_2.$$

What is their projected maximum revenue and how should they price their consoles to achieve it? Round to the nearest million dollars. Summarize your final answer in a full sentence. You may use **WolframAlpha** to do basic algebraic calculations; indicate when you have done so. As always, justify your answer.

Do not write any part of your solution to Question 4 on this page. Write it on the next page.

Write your entire solution to **Question 4** on this page.

The revenue function $R : [0, \infty)^2 \rightarrow \mathbb{R}$ is given by

$$R(p_1, p_2) = 1000q_1p_1 + 1000q_2p_2$$

because p defines the console cost per unit ($\frac{\text{cost}}{\text{unit}}$) and q defines the number of consoles sold in thousands, implying that q is a scaling factor for 1000 (ie, $1000 * q$).

Since q_1 and q_2 are given by

$$q_1 = 2500 - 2p_1 - p_2, \quad q_2 = 2000 - p_1 - 3p_2,$$

we can substitute them in R :

$$\begin{aligned} R(p_1, p_2) &= 1000(2500 - 2p_1 - p_2)p_1 + 1000(2000 - p_1 - 3p_2)p_2 \\ &= 2'500'000p_1 - 2'000p_1^2 - 2000p_1p_2 - 3000p_2^2 + 2'000'000p_2 \end{aligned}$$

This is the function we want to maximize. To begin, we need to see if points of extrema of R exist on $M = [0, \infty)^2$. M is closed and R is continuous on M and real-valued, so by **Lemma 2.9.12** we have that

$$\begin{aligned} \lim_{\|p_1, p_2\| \rightarrow \infty} R(p_1, p_2) &= \lim_{\|p_1, p_2\| \rightarrow \infty} [2'500'000p_1 - 2'000p_1^2 - 2000p_1p_2 - 3000p_2^2 + 2'000'000p_2] \\ &= -\infty \end{aligned}$$

and therefore R attains a maximum somewhere on M .

Next, we check the interior of the domain $(0, \infty)^2$ by finding when $\nabla R(p_1, p_2) = 0$:

$$\nabla R(p_1, p_2) = (2'500'000 - 4000p_1 - 2000p_2, 2'000'000 - 6000p_2 - 2000p_1).$$

This yields the matrix

$$\begin{bmatrix} 4000 & 2000 & 2'500'000 \\ 2000 & 6000 & 2'000'000 \end{bmatrix} \xrightarrow{r_1, r_2 / 2000} \begin{bmatrix} 2 & 1 & 1'250 \\ 1 & 3 & 1'000 \end{bmatrix}.$$

After row reducing, we have the matrix $\begin{bmatrix} 1 & 0 & 550 \\ 0 & 1 & 150 \end{bmatrix}$, implying that R has a critical point at

$(p_1, p_2) = (550, 150)$. Evaluating, $R(550, 150) = 837'500'000$.

To check the boundary of the domain (the x and y axes), define the parametric functions $g : [0, \infty) \rightarrow \mathbb{R}^2$ and $h : [0, \infty) \rightarrow \mathbb{R}^2$ by $g(t) = (t, 0)$ and $h(s) = (0, s)$, respectively. Then we want to find the critical points of $R(g(t))$ and $R(h(s))$:

$$\begin{aligned} R(g(t)) &= R(t, 0) = 2'500'000t - 2000t^2 \\ \implies R'(g(t)) &= 2'500'000 - 4000t \\ \implies R'(g(t)) = 0 &\iff t = 625, \quad R(g(625)) = 781'250'000 \\ R(h(s)) &= R(0, s) = 2'000'000s - 3000s^2 \\ \implies R'(h(s)) &= 2'000'000 - 6000s \\ \implies R'(h(s)) = 0 &\iff s = 333.33 \dots, \quad R(h(333.33 \dots)) = 333'333'333.33 \dots \end{aligned}$$

In comparison, our set of critical points is then

Critical Points (p_1, p_2)	Value of R	Rounded R
(550, 150)	837'500'000	838M
(625, 0)	781'250'000	781M
(0, 333.33)	333'333'333.33	333M

Therefore, if SO-KNEE wants to maximize their revenue, it would be best for them to sell the Paystation at \$550 per unit and the Ritch as \$150 per unit. Their maximum revenue is then approximately \$838M.

5. With a pinch of linear algebra, the tangent space of a graph at a point can be written in a more computationally convenient form than Theorem 3.10.11. Here you will study this equivalent formulation.

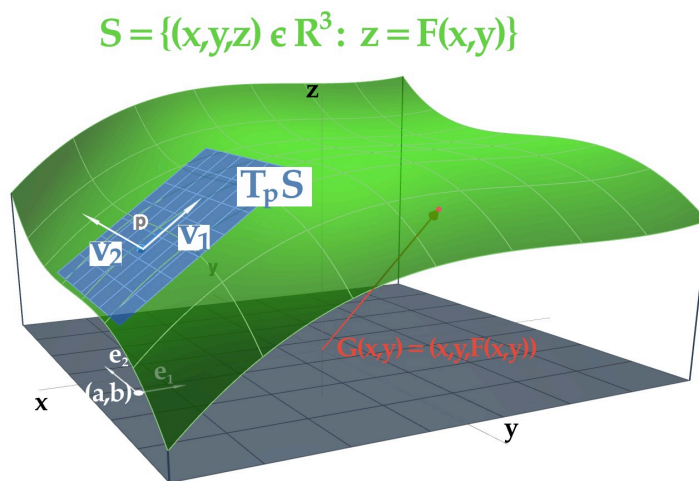
(5a) Begin by visualizing the special case of a graph in \mathbb{R}^3 .

Lemma A. Let the set $S = \{(x, y, z) \in \mathbb{R}^3 : z = F(x, y)\} \subseteq \mathbb{R}^3$ be the graph of a C^1 function $F : \mathbb{R}^2 \rightarrow \mathbb{R}$. Fix $(a, b) \in \mathbb{R}^2$. The tangent space of S at $p = (a, b, F(a, b))$ is given by

$$T_p S = \text{span}\{(1, 0, \partial_1 F(a, b)), (0, 1, \partial_2 F(a, b))\}.$$

Illustrate Lemma A with a detailed labelled picture of a transformation from \mathbb{R}^2 to \mathbb{R}^3 . Use the function $G : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ given by $G(x, y) = (x, y, F(x, y))$. Some quantities in your picture should be labelled using both G and F . You may label some objects with words, but do not write phrases or full sentences.

Hint: Draw specific tangent vectors on the tangent plane.



- $p = (a, b, F(a, b)) = G(a, b)$
- $v_1 = (e_1, \partial_1 F(a, b)) = (1, 0, \partial_x F(a, b))$
- $v_2 = (e_2, \partial_2 F(a, b)) = (0, 1, \partial_y F(a, b))$
- $T_p S = \text{span}\{v_1, v_2\}$

(5b) This lemma can be extended to higher dimensions.

Lemma B. Let $S \subseteq \mathbb{R}^n$ be the graph of a C^1 function $F : \mathbb{R}^k \rightarrow \mathbb{R}^{n-k}$. Fix $a \in \mathbb{R}^k$. The tangent space of S at $p = (a, F(a))$ is given by $T_p S = \text{span}\{e_j, \partial_j F(a) : 1 \leq j \leq k\}$.

Do not prove this lemma. Let $S \subseteq \mathbb{R}^5$ be the graph of $F(x, y, z) = (xe^z + y, ye^{-z} + x)$ and $a = (2, 1, 0)$. Use Lemma B to express the tangent space $T_p S$ at $p = (a, F(a))$ as a span of 3 vectors.

$S \subseteq \mathbb{R}^5$ and $F : \mathbb{R}^3 \rightarrow \mathbb{R}^2$, which implies that $n = 5$ and $k = 3$. If $k = 3$, then we will need three unit vectors e_j for $1 \leq j \leq 3$. They are

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Because the domain of F is \mathbb{R}^3 , we will also need to take the partial derivatives of the components of F first with respect to x , then y and z . We have

$$\begin{aligned} \frac{\partial F(x, y, z)}{\partial x} &= (e^z, 1) \\ \frac{\partial F(x, y, z)}{\partial y} &= (1, e^z) \\ \frac{\partial F(x, y, z)}{\partial z} &= (xe^z, ye^z). \end{aligned}$$

Evaluating at $a = (2, 1, 0)$,

$$\begin{aligned} \frac{\partial F(a)}{\partial x} &= (1, 1) \\ \frac{\partial F(a)}{\partial y} &= (1, 1) \\ \frac{\partial F(a)}{\partial z} &= (2, 1). \end{aligned}$$

By **Lemma B**, we can express $T_p S$ at $p = (a, F(a))$ as

$$T_p S = \text{span} \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 2 \\ 1 \end{pmatrix} \right\}.$$

Lastly, it is important to note that this is a basis for $T_p S$ because each of the vectors are linearly independent, neither of the three are scalar multiples of each other.

6. Using open balls in the definition of regular surfaces is topologically natural, but it is tricky to choose the correct radius and verify that it is small enough. This makes such proofs rather tedious and cumbersome. There is a better equivalent definition, which you may assume without proof.

Lemma C. Let $S \subseteq \mathbb{R}^n$ be a set and let p be a point in S . The set S is a k -dimensional regular surface at p if and only if there exists an open set $N \subseteq \mathbb{R}^n$ containing p such that $S \cap N$ is a graph of a C^1 function $F : U \rightarrow \mathbb{R}^{n-k}$ with U an open subset of \mathbb{R}^k .

Use Lemma C to formally prove that the set $S = \{(x, y) \in \mathbb{R}^2 : x^{2/3} + y^{2/3} = 2\}$ is a 1-dimensional regular surface at the point $p = (1, 1)$. Remember to carefully define any functions with their domain and codomain. (Choose your open set N in Lemma C to make the proof as easy as possible. Do not choose a ball.)

I want to prove that there exists an open set $N \subseteq \mathbb{R}^2$ containing the point $(1, 1)$ such that the intersection $S \cap N$ is a graph of a C^1 function $F : U \rightarrow \mathbb{R}$ with $U \subseteq \mathbb{R}$, then S is a 1-dimensional regular surface at $p = (1, 1)$.

Proof.

- Choose the open set $N = (0, \infty)^2 \subseteq \mathbb{R}^2$. The point $p = (1, 1) \in N$ because for each component, $1 > 0$.
- Then the intersection between N and S becomes $N \cap S = \{(x, y) \in N : x^{2/3} + y^{2/3} = 2\}$.
- Similarly, $p = (1, 1) \in N \cap S$ because $1^{2/3} + 1^{2/3} = 2$ and $p \in N$.
- Define the open subset $U = (0, \infty) \subseteq \mathbb{R}$ and the function $F : U \rightarrow \mathbb{R}$ given by $F(x) = (2 - x^{2/3})^{3/2}$.
- F is a C^1 function because F is differentiable on U for $x > 0$. Similarly, F is continuous on U by **Corollary 2.7.18** because the domain of \sqrt{x} is $(0, \infty) = U$ and the domain of $x^{2/3}$ is \mathbb{R} .

The derivative of is $F'(x) = -\frac{\sqrt{2 - x^{2/3}}}{x^{1/3}}$. F' exists and is continuous on $U = (0, \infty)$ because the numerator and denominators are both continuous on $(0, \infty)$.

Therefore F is C^1 on U .

- The graph of F is the set

$$\begin{aligned} H &= \{(x, y) \in (0, \infty) : y = F(x)\} \\ &= \{(x, y) \in (0, \infty) : y = (2 - x^{2/3})^{3/2}\} \\ &= \{(x, y) \in (0, \infty) : y^{2/3} = 2 - x^{2/3}\} \\ &= \{(x, y) \in (0, \infty) : x^{2/3} + y^{2/3} = 2\} \\ &= N \cap S, \end{aligned}$$

the intersection $N \cap S$ containing p .

- Therefore by **Lemma C**, S is a 1-dimensional surface at $p = (1, 1)$.

■