Some notes on first-order ODEs

[These notes are under construction. Comments and criticism are welcome.]

Contents

Introduction

First-order ODEs come in two forms: *derivative form* and *differential form*. The two forms are closely related, but differ in subtle ways not addressed adequately in most textbooks (and often overlooked entirely)[1]. This often leads to an unclear or inadequate definition of "implicit solution" to an equation in derivative form, before differential-form equations (which are more easily relatable to "implicit solutions" than are derivative-form DEs) have even been introduced. I have not seen a single textbook whose definition of "implicit solution" I find wholly satisfactory. Exacerbating the problem is the usage of a relatively new term (or new, formal usage of an old, informal term) that has crept into textbooks in recent decades—"explicit solution" of a differential equation—that is at odds with the conventional meaning of "explicit", and is defined in these textbooks to mean *exactly* the same thing that mathematicians have always called simply a *solution* of a differential equation.

The purpose of these notes, originally, was simply to give a definition of "implicit solution" that is accurate, precise, complete, understandable by typical students in an introductory DE course, and sensible.[2] More topics, such as an attempt to give a usable meaning to the term "explicit solution" in which the word "explicit" is not superfluous and is less misleading than in current textbooks, were added as the writing went along. This has made for a rather lengthy, never-quite-finished set of notes, an ongoing project that I work on on only occasionally.

In order to make the presentation readable concurrently with a typical modern DE textbook, in these notes I define "implicit solutions of a DE in derivative form" before even introducing differential form. However, one cannot achieve a complete understanding of implicit solutions without investigating differential-form DEs in more depth than is typical for a first course in DEs. Therefore, after we cover differential-form DEs, we return to derivative-form equations to clean up the picture.

The "Notes for Instructors" section below is written for mathematicians (or, rather, *will be* written for mathematicians once I get around to writing it); it is intended to show why certain definitions commonly seen in textbooks are inadequate. Most students, in their first differential equations course, will not be in a position to appreciate these inadequacies. It is up to each instructor to decide whether, in a first

[1] Actually, it is only derivative-form DEs that can be written in the "standard form" $\frac{dy}{dx} = f(x, y)$ that are closely related to differential-form DEs. This is an important difference between the two types, but there are important differences even between standard-form derivative-form and differential-form DEs.

[2] (1) "Accurate" is a bit subjective in this case, since, to my knowledge, there exists no official definition of "implicit solution". In all textbooks I've seen from the era in which I was a student, the term "implicit solution" was not given a formal definition, and some books did not use the term at all. (2) What I mean by "sensible" is that the definition should not lead to anything being called an "implicit solution" that shouldn't be. The judgment of what "should" or "shouldn't" be called by a name that has no official definition is subjective too, of course, but these notes include my justification of why I think the most common definition of "implicit solution" I've seen in textbooks is not sensible.

course on ODEs, it is more important that a definition be short and (superficially) simple than that it be 100% accurate.

1 Notes for Instructors

[This section is not yet written. However, much of the content intended eventually for this section is in footnotes addressed to instructors in the "Notes for Students" section.]

2 Notes for Students

2.1 Review of "derivative form" and "solution"

In these notes, "differential equation", which we will frequently abbreviate as "DE", always means *ordinary* differential equation, of first order unless otherwise specified.

A DE in derivative form is a differential equation that (up to the names of the variables), using only the operations of addition and subtraction, can be put in the form

$$\mathsf{G}(x, y, \frac{dy}{dx}) = 0, \tag{2.1}$$

where G is a function of three variables. Such a DE has an *independent variable* (in this case x) and a *dependent variable* (in this case y). The notation "$\frac{dy}{dx}$" tells you which variable is which.

Definition 2.1 For a given G, a *solution of* (2.1) *on an open interval* I is a real-valued differentiable function ϕ on I such that when "$y = \phi(x)$" is substituted into (2.1), the resulting equation is a true statement for each $x \in I$ (equivalently, such that $\mathsf{G}(x, \phi(x), \phi'(x)) = 0$ for each $x \in I$.[3]

[3] See, for example, [1, p. 3]. Some current authors refer to what we have just defined as an *explicit solution* of (2.1) on I, terminology that did not exist when I was a student. (Note for instructors: Even worse, some authors would say not that ϕ is an explicit solution of (2.1), but that $\phi(x)$ is an explicit solution of (2.1). This perpetuates students' misunderstanding of what a *function* is, which can lead to problems when defining differential operators, or the Laplace Transform, as is usually done in an intro DE course.) This use of "explicit" has apparently been introduced to help students understand later, by way of contrast, what an *implicit solution* is. As commendable as this motivation may be, the terminology "explicit solution" suffers from several drawbacks: (1) It implies a meaning for the term *solution of an equation* that differs from pre-existing, completely standard meaning that is used throughout mathematics. (2) The terminology is misleading and potentially confusing. So-called "explicit solutions" can be functions for which it is effectively impossible to write down an explicit formula, which is usually what one means by "explicitly-defined function". We will deal with this terminological conundrum starting with Definition 2.2. I would prefer that textbook-authors and other instructors stop using the terminology "explicit solution", but since I

For a given G, we call a one-variable function ϕ a *solution of* (2.1) (no interval mentioned) if ϕ is a solution of (2.1) on *some* open interval I. A *solution curve* of (2.1) is the graph of a solution, i.e. the set $\{(x, \phi(x)) \mid x \in I\}$, where ϕ is a solution of (2.1) on the interval I. ■

(In these notes, the symbol ■ indicates the end of a definition, example, exercise, or theorem.)

Henceforth, whenever we say "solution of a differential equation on an interval I", we always mean an *open* interval I.[4]

Definition 2.2 (temporary) If ϕ is a solution of the DE (2.1) (perhaps with an interval specified, perhaps not), we will call the equation "$y = \phi(x)$" an *explicitsolution* (one word, for now), of the DE.[5] ■

Note that, according to Definition 2.1, an explicitsolution of a DE is *not* a solution of the DE. A solution of a DE is a *function*; an explicitsolution is an *equation*. A function and an equation are two different animals. An equation may be used to *define* a function, as in "$\phi(x) = e^x$". But "ϕ" is not the same thing as "the definition of ϕ", any more than an elephant is the same thing as the definition of an elephant.

Nonetheless, we allow ourselves to say, *technically incorrectly*, that "$y = x^2$ is a solution of $\frac{dy}{dx} = 2x$" (for example), because that wording is so much less awkward than "the function ϕ defined by $\phi(x) = x^2$ is a solution of $\frac{dy}{dx} = 2x$".[6] This is similar to allowing ourselves to say "$x = 5$ is a solution of $x^2 = 25$" in place of the

---

cannot make this happen, in these notes I give a definition that I believe is what the definition of "explicit solution" should have been all along, once the unfortunate decision to introduce this terminology was made.

[4] In order to avoid certain distracting technicalities, in these notes we stick to open intervals for the allowable domains of solutions to differential equations in derivative form. However, often it is important to study differential equations on non-open intervals as well. For example, in initial-value problems in which the independent variable is time t, we are generally interested only in what happens in the *future* of the initial time t_0, not in the past. In this case, the relevant intervals are of the form $[t_0, \infty)$, $[t_0, t_1)$, or $[t_0, t_1]$, where $t_1 > t_0$. Most of the statements made in these notes about differential equations on open intervals can be generalized to non-open intervals, but sometimes the statements have to be worded in a more complicated fashion. Your instructor can tell you which statements generalize, and what modifications need to be made.

[5] The one-word term "explicitsolution" is something invented just for these notes. It is used here to preserve logical clarity before replacing it with the two-word phrase "explicit solution" whose use in modern textbooks like [3] is consistent with neither the conventional meaning of the word "explicit", nor, in some cases, with the long-established and completely standard meaning of "solution of a differential equation" [1, p. 3].

[6] Only slightly more awkward than "$y = x^2$ is a solution of $\frac{dy}{dx} = 2x$" is the following type of phrasing that you may have seen instructors or textbook-authors use: "The function $\phi(x) = x^2$ is a solution of $\frac{dy}{dx} = 2x$." This phrasing is certainly much less awkward than, "The function ϕ defined

more precise "5 is a solution of $x^2 = 25$." Each of these examples (the differential-equation-solution and algebraic-equation-solution examples) is an example of "abuse of terminology", but this particular abuse is so standard, so convenient, so hard to avoid, and so unlikely to lead to any confusion that every mathematician regards it either as (i) a *permissible* abuse of terminology, or (ii) a second valid meaning of the phrase "solution of an equation." Because of this, we make the following definition:

Definition 2.3 An *explicit solution* (two words) of a DE is an explicitsolution of that DE. ■

This is not the same definition of "explicit solution" that appears in (for example) [3]. Rather, it is an attempt to reconcile any desire to use the phrase "explicit solution" with (i) the standard meaning of "solution of a differential equation, (ii) the expectation that an object called an "explicit solution" of a DE should, in particular, be a solution of that DE, and (iii) what appears to be the motive for introducing the phrase "explicit solution" into textbooks, namely "If there's some object we're going to call an *implicit* solution, we ought to call something by the name *explicit* solution."[7]

2.2 Implicit solution of a derivative-form DE

Key in understanding what "implicit solution of a differential equation" means is the understanding the concept of an implicitly defined *function* of one variable. You learned about implicitly defined functions as far back as Calculus 1, when you studied *implicit differentiation*, but we will review the concept here. In order to make sure the concept is clear, we go into more depth than you probably did in Calculus 1 (or even Calculus 3).

Suppose we are given an algebraic (i.e. non-differential) equation in variables x and y. We can always write such an equation in the form

by $\phi(x) = x^2$ is a solution of $\frac{dy}{dx} = 2x$." The reason we try not to use phrasing like "The function $\phi(x) = x^2 \ldots$" in these notes is that the function is ϕ, not $\phi(x)$. The object $\phi(x)$—a *number*—is the output of the function ϕ when the input is called x.

However, practically all math instructors at least occasionally use phrasing like "The function $\phi(x) = x^2$", and some use it all the time. The language needed to avoid such phrasing is often extremely convoluted (unless the student has been introduced to the notation "$x \mapsto x^2$"). So, while the author of these notes does not like it, this type of phrasing is generally regarded as "permissible abuse of terminology". Nonetheless it is important that the student understand the difference between a *function* and the *output of that function*. To help foster this understanding, we (mostly) avoid this particular abuse of terminology in these notes, even though we allow certain other abuses of terminology.

[7] As noted earlier, what would be far better than to use the definition of "explicit solution" in these notes would be for authors and instructors to abandon using the phrase "explicit solution" with anything other than its historical meaning: a solution for which we have an explicit *formula*. But until that happens, the definitions in these notes may be of some use.

$$F(x, y) = 0$$

for some two-variable function F. We may be interested in solving for y in terms of x. For example, if

$$x^2 + y^3 - 1 = 0 \tag{2.2}$$

then

$$y = (1 - x^2)^{1/3}. \tag{2.3}$$

In other words, if we define $F(x, y) = x^2 + y^3 - 1$ and $\phi(x) = (1 - x^2)^{1/3}$, then whenever the pair (x, y) satisfies $F(x, y) = 0$, it satisfies $y = \phi(x)$. Conversely, one may verify by direct substitution that if $y = (1 - x^2)^{1/3}$ then $F(x, y) = 0$. Thus

$$F(x, y) = 0 \quad \text{if and only if} \quad y = \phi(x). \tag{2.4}$$

Note that the "if" part of this implication is the "Conversely ..." statement above, and can be written equivalently as the equation

$$F(x, \phi(x)) = 0.$$

More generally than this example, any time (2.4) is true for a two-variable function F and one-variable function ϕ, we say that the equation $F(x, y) = 0$ *implicitly determines* (or *implicitly defines*) y as a function of x, and we call ϕ the function of x implicitly determined/defined by the equation $F(x, y) = 0$.

Now consider the equation

$$x^2 + y^2 - 1 = 0. \tag{2.5}$$

"Solving for y in terms of x" gives the relation

$$y = \pm\sqrt{1 - x^2}. \tag{2.6}$$

Looking just at (2.5), it is already clear that any numerical choice of x restricts the possible choices of y that will make the equation a true statement. Equation (2.6) tells us the only possible values for y that will work. It also tells us that for each x in the open interval $(-1, 1)$ there are at most two such values; for $x = 1$ and for $x = -1$ there is at most one such value; and for $|x| > 1$ there are no values of y that will work. Conversely, if we substitute $y = \pm\sqrt{1 - x^2}$ into (2.5), we see that all the values of y that we have labeled as "possible" actually do work. Thus, *for each pair* (x, y) *of real numbers*,

$$x^2+y^2-1 = 0 \text{ if and only if } |x| \leq 1 \text{ and either } y = \sqrt{1 - x^2} \text{ or } y = -\sqrt{1 - x^2}. \tag{2.7}$$

This is a *much* weaker statement than a statement of the form (2.4), because the sign in $\pm\sqrt{1-x^2}$ can be chosen independently for each x. On the domain $[-1,1]$, if we define

$$\phi_1(x) = \sqrt{1-x^2}, \tag{2.8}$$

$$\phi_2(x) = -\sqrt{1-x^2}, \tag{2.9}$$

$$\phi_3(x) = \begin{cases} \sqrt{1-x^2} & \text{if } x \text{ is a rational number,} \\ -\sqrt{1-x^2} & \text{if } x \text{ is an irrational number,} \end{cases} \tag{2.10}$$

then all three of these functions ϕ_i yield true statements, for each $x \in [-1,1]$, when $\phi_i(x)$ is substituted for y in (2.5). In fact, since the sign "$\pm$" can be assigned randomly for each $x \in [-1,1]$, there are *infinitely many* functions ϕ that work. What distinguishes ϕ_1 and ϕ_2 from all the others is that they are *continuous*. If we restrict their domains to the open interval $(-1,1)$, then they are even differentiable.

Now consider a more complicated equation, such as

$$e^x + x + 6y^5 - 15y^4 - 10y^3 + 30y^2 + 10xy^2 = 0. \tag{2.11}$$

Clearly, choosing a numerical value for x restricts the possible values for y that will make equation (2.11) a true statement. It turns out that, depending on the choice x, there can be anywhere from one to five values of y for which the pair (x,y) satisfies equation (2.11). As in the previous example, on any x-interval I for which there is more than one y-value that "works" for each x, there will be infinitely many functions ϕ for which $F(x,\phi(x)) = 0$, where $F(x,y)$ is the left-hand side of equation (2.11). However, there are not very many *continuous* ϕ's that work. In this example, whatever x-interval I we choose, there are at most five continuous functions ϕ defined on I for which $F(x,\phi(x)) = 0$. Writing out *explicit formulas* for them, analogous to the formulas for ϕ_1 and ϕ_2 in the previous example, is a hopeless task. But these continuous functions ϕ exist nonetheless. We can see this visually in Figure 1.

Definition 2.4 Let F be a function of two variables, ϕ a function of one variable, and I an interval. We say that the equation $F(x,y) = 0$ *implicitly determines* or *implicitly defines* the function ϕ, regarded as a function of x (or whatever name is used for the first variable of F), if $F(x,\phi(x)) = 0$ for all $x \in I$.[8]

[8] *Note to instructors:* I dislike this usage of the word *determines* (or *defines*)—which is the only one I've seen in Calculus 1-2-3 and Differential Equations textbooks that bother to give a definition at all—and would argue against using it if I knew a good substitute. The word "determines" is best be used only when there is a unique object being determined (as in (2.13) coming up soon); any other usage is a significant and unnecessary departure from standard English usage of this word. According to Definition 2.4, the equation $0 = 0$, viewed as an equation on $I \times \mathbf{R}$, implicitly determines (or even worse in this case, *defines*) *every* function $I \to \mathbf{R}$. For a less obvious example, the equation $(x^2+y)^{1/\ln(x^2+y)} - e = 0$, viewed as an equation on $\{(x,y) \in \mathbf{R}^2 \mid x^2 + y > 0\}$, determines every

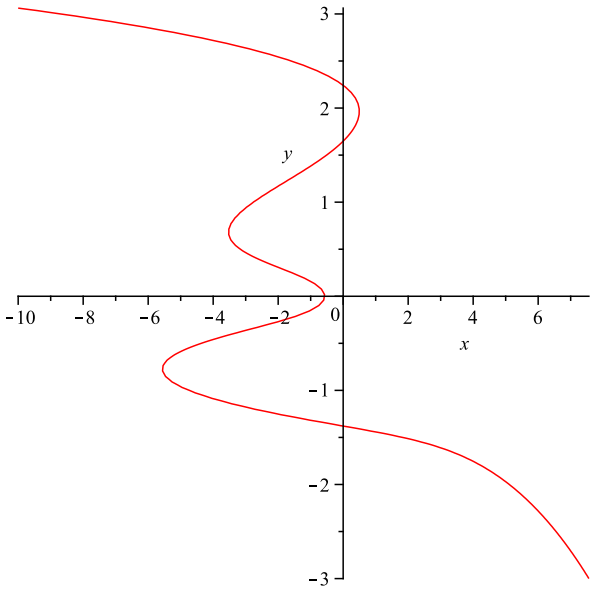

Figure 1: The graph of $e^x + x + 6y^5 - 15y^4 - 10y^3 + 30y^2 + 10xy^2 = 0$.

Without reference to a specific interval I, we say that the equation $F(x, y) = 0$ implicitly determines ϕ, regarded as a function of the first variable of F, if the equation $F(x, y) = 0$ implicitly determines ϕ (regarded as a function of x) on *some* open interval.

The same definitions apply if the "0" in $F(x, y) = 0$ is replaced by any other real number, or even by another function $H(x, y)$ (in the latter case, we replace "$F(x, \phi(x)) = 0$" with "$F(x, \phi(x)) = H(x, \phi(x))$". ■

---

function $\phi : \mathbf{R} \to \mathbf{R}$ for which $\phi(x) > -x^2$. And clearly we can cook up an arbitrarily complicated expression $F(x, y)$ such that "$F(x, y) = 0$" reduces to an identity, but does not *obviously* reduce to an identity, at least not in the eyes of a student. Altering Definition 2.4 so as to exclude any equation $F(x, y) = 0$ that restricts to an identity on some open subset of $\mathbf{R}^2$ would, of course, eliminate the examples just given, but would complicate Definition 2.4 without eliminating all of the problems intrinsic to using the word "determines" or "defines" as it is used in this definition. To see that the problem cannot be fixed (artificially, but with pedagogical simplicity) by making "implicitly-defined function" mean "function given by the conclusion of the Implicit Function Theorem", see Examples 2.14 and 2.15.

In the setting of Definition 2.4, there *is* a unique object determined: the *set* of all functions $\phi : I \to \mathbf{R}$ satisfying $F(x, \phi(x)) = 0$, not the elements of this set (unless it is a set with only one element, as in the conclusion of the Implicit Function Theorem). Even for the "0=0" example, this author feels much more comfortable saying that the equation 0=0 determines the set of all functions $I \to \mathbf{R}$, than saying that it determines the function $x \mapsto x^{53} + e^x - \sqrt{x^2 + 1} \tan^{-1} x$.

However, in these notes the author is compromising on this point. In this instance he feels that the benefit of rephrasing Definition 2.4 properly is outweighed by the risk that it would become incomprehensible to too many intro DE students.

Graphically, a function ϕ is implicitly determined by the equation $F(x, y) = 0$ if the graph of ϕ is part of the graph of $F(x, y) = 0$.[9] (For these purposes, "all of" is a special case of "part of".)

There are instances in which we want to know whether there is a one-variable function ϕ such that $F(\phi(y), y) = 0$. This comes up when we think of trying to solve the equation $F(x, y) = 0$ for x in terms of y, rather than for y in terms of x. To handle this case we can give a definition analogous to Definition 2.4, replacing the phrases "regarded as a function of x" and "first variable" with "regarded as a function of y" and "second variable", and replacing "$F(x, \phi(x)) = 0$ with "$F(\phi(y), y) = 0$". To simplify wording below, any time we say an equation $F(x, y) = 0$ implicitly determines (or defines) a function ϕ, we mean to regard ϕ as a function of x, unless we say otherwise.

Thus:

- Equation (2.2) implicitly determines the function ϕ given by the formula $\phi(x) = (1 - x^2)^{1/3}$.
- Equation (2.5) implicitly determines the functions ϕ_1, ϕ_2, ϕ_3 defined in (2.8)–(2.10), and infinitely many others on the interval $[-1, 1]$. The only *continuous* functions that (2.5) determines on $[-1, 1]$ are ϕ_1 and ϕ_2.
- Equation (2.11) implicitly determines infinitely many functions, but only a few continuous functions. In Figure 1, if we travel along the graph by starting at the upper left and moving along the curve, we encounter vertical tangents at points A, B, C, and D (labeled in the order that we encounter them). Let x_A, x_B, x_C, and x_D denote the x coordinates of these points. Then (2.11) implicitly determines a continuous function of x, say ϕ_1, with domain $(-\infty, x_A]$; another continuous function of x, say ϕ_2, with domain $[x_B, x_A]$; another, say ϕ_3, with domain $[x_B, x_C]$; another, say ϕ_4, with domain $[x_D, x_C]$; and another, say ϕ_5, with domain $[x_D, \infty]$. On the interval $[-3, -2]$, the equation $F(x, y) = 0$ determines five continuous functions (the restrictions of $\phi_1, \phi_2, \phi_3, \phi_4$, and ϕ_5 to this interval). On the interval $[-5, -4]$, $F(x, y) = 0$ determines three continuous functions (the restrictions of ϕ_1, ϕ_4, and ϕ_5 to this interval).

In some cases, an equation $F(x, y) = 0$ will implicitly determine one and only one function of x on some interval. That is a "best-case scenario". When we are in such a case, we can speak unambiguously of *the* function of x determined by this equation. Often we can achieve this result by "windowing" x and y; i.e., by agreeing to consider only pairs (x, y) where x lies in some specific interval I and y lies in some specific interval J. We denote the corresponding set in xy plane by $I \times J$:

[9] Recall that the *graph* of an equation in x and y is the solution-set of the equation: the set of points $(x, y) \in \mathbf{R}^2$ for which the equation is a true statement.

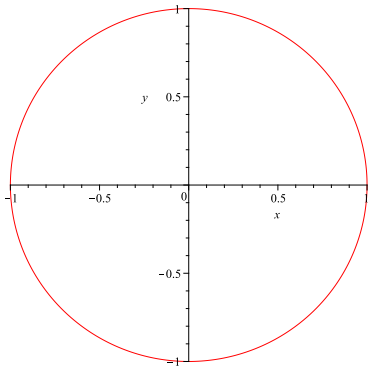

Figure 2: The graph of $x^2 + y^2 = 1$.

$$I \times J = \{(x, y) \mid x \in I \text{ and } y \in J\}.$$

In these notes we will call such a set a *rectangle*, even though we do not exclude the possibility that I and/or J extend(s) infinitely in one direction or both. Thus, for example, we consider the whole xy plane a rectangle; the set $[1, \infty) \times (-\infty, \infty)$ is a rectangle (consisting of all pairs (x, y) for which $x > 1$); the strip $(-\infty, \infty) \times (0, 1]$ is a rectangle (consisting of all pairs (x, y) with $0 < y \leq 1$). Of course, objects that Euclid would have called rectangles, such as $[1, 2] \times [3.1, 4.9]$, are also rectangles in our terminology. In these notes, we will be most interested in *open* rectangles, those we get by taking the intervals I and J to open.

When an equation $F(x, y) = 0$ implicitly determines more than function of x, "windowing" may allow us to single out one of them. For example, consider the graph of the circle $x^2 + y^2 = 1$ (Figure 2).

Let $P = (x_0, y_0)$ be any point on the circle *other than* $(1, 0)$ or $(-1, 0)$; thus $y_0 \neq 0$. For any such point, you can draw an open rectangle $R = I \times J$, containing (x_0, y_0), such that the portion of the circle lying in R is a portion of the graph of *exactly one* of the two functions ϕ_1, ϕ_2 in (2.8)–(2.9) ($\phi_1(x) = \sqrt{1 - x^2}$, $\phi_2(x) = -\sqrt{1 - x^2}$). For example, if $y_0 > 0$ you can take J to be any open subinterval of $(0, \infty)$ that contains y_0, and then take I to be any open interval whatsoever that contains x_0. Choose some points on the graph in Figure 2 and draw rectangles around them with the desired property.

Note that the closer your point (x_0, y_0) gets to $(1, 0)$ or $(-1, 0)$, the more limited your choices of I and J become, in the sense that one endpoint of I will have to be very close to x_0, and one endpoint of J will have to be very close to y_0. For example if $y_0 = -.01$ and $x_0 = \sqrt{.9999} \approx .99995$, then the right endpoint of I will have to lie between $\sqrt{.9999}$ and 1, while the right endpoint of J (which gives the location of the upper boundary of the rectangle) will have to lie between $-.01$ and $.01$. But as long as $(x_0, y_0) \neq (\pm 1, 0)$, *some* open rectangle will work.

If you take $(x_0, y_0) = (1, 0)$, then this windowing process fails in two ways to have the desired effect. First, for *no* open interval I containing 1 is there a function ϕ defined on all of I such that $x^2 + \phi(x)^2 = 1$ for all $x \in I$, because such an interval I will contain an x that is greater than 1 (so $x^2 + \phi(x)^2 > 1$ no matter what you choose for $\phi(x)$). Second, for any open rectangle $I \times J$ containing $(1, 0)$, for values of x very close to but less than 1, both the point $(x, \sqrt{1 - x^2})$ and $(x, -\sqrt{1 - x^2})$ will lie in $I \times J$. Thus $I \times J$ will include points of the graphs of both ϕ_1 and ϕ_2, no matter how small you take I and J.

Of course, similar statements are true for the point $(x_0, y_0) = (-1, 0)$.

The *Implicit Function Theorem* gives conditions under which the "windowing near a point (x_0, y_0)" idea works very nicely to guarantee that an equation such as "$F(x, y) = 0$" determines at least one function of x, and, if it determines more than one such function, to use (x_0, y_0) to single out one of them. Furthermore, the implicitly-defined functions given by this theorem are actually *differentiable* (in fact, continuously differentiable; i.e. the derivative of each implicitly-defined function is continuous).

Theorem 2.5 (Implicit Function Theorem) *Let F be a two-variable function whose first partial derivatives are continuous on an open rectangle $R = I \times J$. Suppose that $(x_0, y_0) \in R$ and that $\frac{\partial G}{\partial y}(x_0, y_0) \neq 0$, where $\frac{\partial G}{\partial y}$ denotes the partial derivative of F with respect to the second variable. Let $c_0 = F(x_0, y_0)$.*

Then there exists an open subinterval I_1 of I containing x_0, an open subinterval J_1 of J containing y_0, and a continuously differentiable function ϕ defined on I_1, such that

$$\begin{aligned} &\text{for all points } (x, y) \in I_1 \times J_1, \\ &\qquad F(x, y) = c_0 \text{ if and only if } y = \phi(x). \end{aligned} \tag{2.12}$$

■

In Theorem 2.5, since x_0 lies in I_1, we may look at what (2.12) tells us when $x = x_0$. What this statement reduces to when $x = x_0$ is the following:

$$\begin{aligned} &\text{for all } y \in J_1, \\ &\qquad F(x_0, y) = c_0 \text{ if and only if } y = \phi(x_0). \end{aligned}$$

But by the definition of c_0, we have $F(x_0, y_0) = c_0$. Therefore, since $y_0 \in J_1$, the "only if" part of the above statement tells us that $y_0 = \phi(x_0)$. Thus, the graph of the function ϕ that the Implicit Function Theorem gives us will always contain the point (x_0, y_0).

Let us pause to appreciate how strong the conclusion of this theorem is. Statement (2.12) says that for each $x \in I_1$, there is *one and only one* value $y \in J_1$ for

which $F(x, y) = c_0$, namely the value $\phi(x)$. Thus, (2.12) says that within $I_1 \times J_1$, the equation $F(x_0, y_0) = 0$ determines y *uniquely* as a function of x—not just uniquely among "nice" functions, like continuous functions or differentiable functions. Among *all* functions with domain I_1 and range contained in J_1, ϕ is the *only* function that satisfies $F(x, \phi(x)) = c_0$ identically in x. This function has the *additional nice feature* of being continuously differentiable (and hence continuous), but there is *no other function whatsoever* on I_1 that satisfies $F(x, \phi(x)) = c_0$ identically in x.

Compared statement (2.12) with statement (2.4). The only important difference is that to get the second line of (2.12), we had to make the windowing restriction in the first line. (The fact that we have "c_0" in (2.12) where we have "0" in (2.4) is an unimportant difference.)

The uniqueness (of the function ϕ) that is guaranteed by a statement of the form (2.12) allows us to use terminology that is less awkward than what we used in Definition 2.4. Specifically, whenever a statement of the form (2.12) holds true, we can dispense with the phrase "regarded as a function of the first variable of F" in that definition, or even introducing a letter for the function ϕ at all. We may simply say the following:

$$\begin{gathered}\text{Within the rectangle } I_1 \times J_1, \text{ the equation} \\ F(x, y) = c_0 \text{ determines } y \text{ as a function of } x.\end{gathered} \tag{2.13}$$

Optionally, we may put the word "implicitly" in front of "determines" above. Doing so emphasizes the fact that we are not saying we know how to produce a *formula* that tells us how to compute y from x (we may or may not be able to produce such a formula, depending on the function F); we are simply saying that for each $x \in I_1$, one and only one value of y is singled out. But an unambiguous assignment of a value y to each $x \in I_1$ is exactly what "function on I_1" means, by definition. No explicit formula is required in the definition of "function".

Similarly, if there exists a function ϕ defined on J_1 such that

$$\begin{gathered}\text{for all points } (x, y) \in I_1 \times J_1, \\ F(x, y) = c_0 \text{ if and only if } x = \phi(y)\end{gathered} \tag{2.14}$$

then we can say simply that within the rectangle $I_1 \times J_1$, the equation $F(x, y) = c_0$ determines x uniquely as a function of y. Thus, when condition (2.14) is met, we do not have to write a whole new definition analogous to Definition 2.4, with "regarded as a function of the first variable" replaced with "regarded as a function of the second variable", and with "$F(x, \phi(x)) = 0$" replaced with "$F(\phi(y), y) = 0$".

When either (2.12) or (2.14) holds for some rectangle $I_1 \times J_1$, we call ϕ an *implicitly-defined function.*[10]

[10]The informal terminology "implicit function" is a less precise but common phrase meaning "implicitly-defined function". The only good use of the term "implicit function" is in the title of the Implicit Function Theorem, where it provides a way to avoid the awkward title "Implicitly-Defined Function Theorem".

Exercise. Look back at Figure 1. For which points (x_0, y_0) on the graph is it *not* true that there is an open rectangle containing (x_0, y_0) on which the equation in caption determines y uniquely as a function of x? (Don't try to find the *values* of x_0 and y_0; just show with your pencil where these "bad" points are on the graph.) ■

Now, let us get back to differential equations:

Definition 2.6 (temporary) We call an equation $F(x, y) = 0$ an *implicitsolution* (one word, for now) of a differential equation

$$\mathsf{G}(x, y, \frac{dy}{dx}) = 0 \tag{2.15}$$

(for a given G) if

(i) the equation $F(x, y) = 0$ implicitly determines at least one function ϕ that is a solution of (2.15), and

(ii) *every* differentiable function ϕ determined by the equation $F(x, y) = 0$, with domain an open interval, is a solution of (2.15).[11] ■

Definition 2.7 If ϕ is a differentiable function determined implicitly by an implicitsolution $F(x, y) = 0$ of (2.15), then we call ϕ an *implicitly-defined* solution of (2.15). ■

Example 2.8 Consider the differential equation

$$x + y\frac{dy}{dx} = 0. \tag{2.16}$$

[11] *Note to instructors:* The definition of "implicit solution" does not, and should not, rely at all on implicit differentiation of the equation $F(x, y) = 0$. The function F need not even be continuous, let alone differentiable, for the concept of "implicit solution" to make sense (although dreaming up an artificial non-continuous or non-differentiable example to drive this point home to your students is more likely to confuse them.) An implicitly-defined solution of a DE is simply an implicitly-defined function that happens to be a solution of the DE. The *notion* of implicitly-defined function does not rely on calculus in any way.

Of course, it is tremendously important that the Implicit Function Theorem gives sufficient conditions under which we can confirm, via implicit differentiation, that we have an implicit solution of a DE is. When we launch too quickly into examples of implicit solutions, every one of which uses implicit differentiation, and never return to the conceptual definition, we obscure the fundamental issue of what an implicit solution actually *is*. Ask your students what an implicit solution of a DE is, and the *best* answer you're likely to get is, "It's an equation that, after I implicitly differentiate, I can rearrange back to the DE." Few students, if any, will mention any relation to the notion of implicitly-defined *function*, or to (true) solutions of the DE (what some authors call "explicit solutions"). And students are likely to mis-identify some equations as *not* being implicit solutions of a given DE, simply because implicit differentiation got them to a DE that was not algebraically equivalent to the given one. You may want to try Examples 2.15 and 2.16 on your students.

We claim that the equation

$$x^2 + y^2 - 1 = 0 \tag{2.17}$$

is an implicitsolution of (2.16). (Equivalently, so is the equation $x^2 + y^2 = 1$.) To verify this, we check that criteria (i) and (ii) of Definition 2.6 are satisfied:

- Criterion (i). Let $\phi_1(x) = \sqrt{1 - x^2}$ as in (2.8), but restricted to the open interval $(-1, 1)$. Note that $F(x, \phi_1(x)) = 1$ for all $x \in (-1, 1)$, so ϕ_1 is a function implicitly determined by the equation $F(x, y) = 1$ (the conditions of Definition 2.4) are met).

  We compute $\phi_1'(x) = \frac{-x}{\sqrt{1-x^2}}$. Thus if we substitute $y = \phi_1(x)$ into the left-hand side of (2.16), we have

$$\begin{aligned} & x + \sqrt{1 - x^2}\ \frac{-x}{\sqrt{1 - x^2}} \\ = \quad & 0 \quad \text{for all } x \in (-1, 1), \end{aligned}$$

  so ϕ_1 is a solution of (2.16). Thus criterion (i) is satisfied[12].

- Criterion (ii). Suppose ϕ is any differentiable function determined implicitly by (2.17) on some open interval I. Then we have

$$x^2 + \phi(x)^2 - 1 = 0$$

  identically in x on the interval I. Differentiating, we therefore have

$$2x + 2\phi(x)\phi'(x) = 0 \ \text{ for all } x \in I.$$

  Therefore ϕ is a solution of the equation

$$2x + 2y\frac{dy}{dx} = 0$$

  on I. Dividing by 2 we see that ϕ is a solution of (2.16) on I. Therefore criterion (ii) is satisfied.

[12]We could just as well have used the function ϕ_2 defined by $\phi_2(x) = -\sqrt{1 - x^2}$. But to show that criterion (i) is met it suffices to come up with *one* function ϕ that works, so we chose the ϕ that involves (slightly) less writing.

Hence (2.17) is an implicitsolution of (2.16), and the function ϕ_1 is an implicitly-defined solution of (2.16).

There are actually two implicitly-defined solutions in this example: ϕ_1 and $-\phi_1$ (the function that we called ϕ_2 in (2.9)). The first of these is the function implicitly defined by $x^2 + y^2 = 1$ on the rectangle $(-1, 1) \times (0, \infty)$; the second is the function implicitly defined by $x^2 + y^2 = 1$ on the rectangle $(-1, 1) \times (-\infty, 0)$. Both functions are solutions of (2.16). ■

Example 2.9 We claim that

$$(y - e^x)(x^2 + y^2 - 1) = 0 \tag{2.18}$$

is *not* an implicitsolution of (2.16). To verify this claim, it suffices to show that *at least one* of criteria (i) and (ii) in Definition 2.6 is not met. For this, we observe that if $y = e^x$, then (2.18) is satisfied. Thus, the function ϕ defined on any open interval I by $\phi(x) = e^x$ is a function determined implicitly by (2.18). However, if we substitute $y = e^x$ into (2.16), we get

$$x + e^{2x} = 0. \tag{2.19}$$

Is it possible to choose the interval I in such a way that (2.19) holds true for all $x \in I$? No, for if there were such an interval I, the left-hand side of (2.19) would be a differentiable function on I, so we could differentiate both sides of (2.19) and obtain

$$1 + 2e^{2x} = 0. \tag{2.20}$$

But there isn't even a single value of x for which this is true; $1 + 2e^x > 0$ for all x. Thus there is no open interval I on which ϕ is a solution of (2.16).

Thus ϕ is a differentiable function determined implicitly by (2.18) that is not a solution of (2.16). Therefore criterion (ii) in Definition 2.6 is not met, so equation (2.18) is not an implicitsolution of (2.16). (Of course, the same reasoning shows that the equation $y - e^x = 0$ is not an implicitsolution of (2.16).)

We mention that in this example, criterion (i) *is* met. The same function ϕ used in Example 2.8 is a solution of (2.16) that is defined implicitly by (2.18). ■

Example 2.10 The equation

$$x^2 + y^2 + 1 = 0 \tag{2.21}$$

is *not* an implicitsolution of (2.16), because it fails criterion (i) of Definition 2.6. There are no real numbers x, y at all for which (2.21) holds, let alone an open interval I on which (2.21) implicitly determines a function of x. Since (2.21) determines no functions ϕ whatsoever on any open interval I, criterion (ii) of Definition 2.6 is moot.

Similarly, the equation

$$x^2 + y^2 = 0 \tag{2.22}$$

is not an implicitsolution of (2.16). In this case there *is* a pair of real numbers (x, y) that satisfies (2.22), but there is no *open x-interval* I on which, for each $x \in I$, there is a real number y for which (2.22) is satisfied. ■

Now let us make an observation about implicitsolutions:

An implicitsolution of a DE is <u>not</u> a solution of that DE. (2.23)

The reason is simple. A solution of a DE is a *function* (of one variable). An implicitsolution of a DE is an *equation* (in two variables). These are two completely different animals.

However, in our earlier discussion of "explicit solutions", we said that if ϕ is a solution of a DE $\mathsf{G}(x, y, \frac{dy}{dx})$, we would permit ourselves to call the equation $y = \phi(x)$ an explicit solution of the DE, regarding this phrasing as "permissible abuse of terminology". Note that the equation "$y = \phi(x)$", which we are allowing ourselves to call a solution of a DE if ϕ is a solution of that DE, is equivalent to the equation "$y - \phi(x) = 0$", which is an equation of the form $F(x, y) = 0$. In the same spirit, we make the following definition:

Definition 2.11 We say that an equation $F(x, y) = 0$ is an *implicit solution* (two words) of a given differential equation if it is an implicitsolution (one word) of that differential equation, as defined in Definition 2.6. ■

Combining this definition with observation (2.23), we have a linguistic paradox:

An implicit solution of a DE is *not* a solution of that DE.

In other words, the meaning of "implicit solution" cannot be obtained by interpreting "implicit" as an adjective modifying "solution". One must regard the two-word phrase "implicit solution" as a single term, a compound noun whose meaning cannot be deduced from the meanings of the two words comprising it. That is why we initially used the made-up word "implicitsolution", which the student is not likely to find outside these notes. Of course, as we observed earlier (but did not display in a line like (2.2)), the term "explicit solution" has the same problem: an explicit solution of a DE, as defined by Definitions (2.2) and (2.3) is not literally a solution of the DE, according to the standard definition (2.1) of "solution of a differential equation".

Most textbooks that give a definition of the term "implicit solution" (some books essentially do not use the term at all; e.g. [1] uses it as a *topic heading*, but there

is no object that is actually called an "implicit solution"), give a definition that is similar to our definition of "implicitsolution"[13].

Of course, in English there are many compound nouns of the form "<adjective> <noun>" that do not mean "a special type of <noun>". A prairie dog is not a type of dog.

Note that the terminology "implicitly-defined solution" (Definition 2.7) does not suffer from any paradox. An implicitly-defined solution of a DE *is* a solution of that DE. It meets the criteria of Definition 2.1 perfectly.

Our approach to Example 2.8 above relied on our ability to produce an explicit formula for a "candidate solution" of the given DE. What if, in place of (2.17), we had been given an equation so complicated that we could not solve for y and produce a candidate-solution ϕ to plug into the DE? This is where the Implicit Function Theorem comes to the rescue.

Example 2.12 [14] Show that the equation

$$x + y + e^{xy} = 1 \tag{2.24}$$

is an implicit solution of

$$(1 + xe^{xy})\frac{dy}{dx} + 1 + ye^{xy} = 0. \tag{2.25}$$

To show this, we start with the observation that, writing $F(x, y) = x+y+e^{xy}$, we have $F(0, 0) = 1$. So, let us check whether the Implicit Function Theorem applies to the equation $F(x, y) = 1$ near the point $(0, 0)$ (i.e. taking $(x_0, y_0) = (0, 0)$ in Theorem 2.5). We compute

$$\begin{aligned} \frac{\partial G}{\partial x}(x, y) &= 1 + ye^{xy}, \\ \frac{\partial G}{\partial y}(x, y) &= 1 + xe^{xy}. \end{aligned}$$

Both of these functions are continuous on the whole xy plane, and $\frac{\partial G}{\partial y}(0, 0) = 1 \neq 0$. Thus, the hypotheses of Theorem 2.5 are satisfied (with $R = (-\infty, \infty) \times (\infty, \infty)$). Therefore the conclusion of the theorem holds. We do not actually need the whole conclusion; all we need is this part of it: there is an open interval I_1 containing 0, and a differentiable function ϕ defined on I_1, such that $F(x, \phi(x)) = 1$ for all $x \in I_1$.

[13] Except that criterion (ii) seems to have been either overlooked or deliberately omitted in all the textbooks I have seen. Example 2.9, as well as the discussion in footnote 8, show that omitting this criterion can lead to calling a function an "implicit solution of a (given) DE" when it is nonsensical to do so.

[14] This example is taken from Nagle, Saff, and Snider, *Fundamentals of Differential Equations and Boundary Value Problems*, 5th ed., Pearson Addison-Wesley, 2008.

Now we use the same method by which we checked criterion (ii) in Example 2.16: implicit differentiation (i.e. computing derivatives of an expression that contains an implicitly-defined function). Let us simplify the notation a little by writing $y(x) = \phi(x)$. Then

$$\begin{aligned} & x + y(x) + e^{xy(x)} = 1 \ \text{ for all } x \in I_1, \\ \Longrightarrow \qquad & 1 + \frac{dy(x)}{dx} + e^{xy(x)} \left(y(x) + x\frac{dy(x)}{dx} \right) = 0 \ \text{ for all } x \in I_1, \\ \Longrightarrow \qquad & (1 + xe^{xy(x)})\frac{dy(x)}{dx} + 1 + y(x)e^{xy(x)} = 0 \ \text{ for all } x \in I_1. \end{aligned}$$

Therefore ϕ is a solution of (2.25). Thus, criterion (i) in Definition 2.6 is satisfied. The exact same implicit-differentiation argument shows that if ψ is *any* differentiable function determined on an open interval by (2.24), then ψ is a solution of (2.25). Therefore criterion (ii) in Definition 2.6 is also satisfied. Hence (2.24) is an implicit solution of (2.25). ■

Looking back at Example 2.8, could we have shown that criterion (i) of Definition 2.6 is satisfied using the technique of Example 2.12, using the function $F(x, y) = x^2+y^2$? Absolutely! For (x_0, y_0) we could have taken any point of the circle $x^2+y^2 = 1$ other than $(\pm 1, 0)$. The partial derivatives are $\frac{\partial G}{\partial x}(x, y) = 2x$ and $\frac{\partial G}{\partial y}(x, y) = 2y$. As in Example 2.12, the partial derivatives of F are continuous on whole xy plane again[15], and since we are choosing a point (x_0, y_0) for which $y_0 \neq 0$, we have $\frac{\partial G}{\partial y}(x_0, y_0) \neq 0$. Thus, the Implicit Function Theorem applies, guaranteeing the existence of a differentiable, implicitly-defined function ϕ, with $\phi(x_0) = y_0$. We can then differentiate implicitly, as we did when we checked criterion (ii) in Example 2.8 (and as we did to check both criteria in Example 2.12), to show that ϕ is a solution of (2.16). If our point (x_0, y_0) has $y_0 > 0$, then the solution of (2.16) that we get is the function ϕ_1 defined by $\phi_1(x) = \sqrt{1 - x^2}$; if $y_0 < 0$ then the solution of (2.16) that we get is $-\phi_1$.

The student may wonder how we could have used the method of Example 2.12 had we not been clever (or lucky) enough to be able to find a point (x_0, y_0) that lay on the graph of our equation $F(x, y) =$ a given constant. The answer is that we could not have, unless we had some other argument showing that the graph contains at least one point, and, more restrictively, that it contains at least one point at which $\frac{\partial G}{\partial y}$ is not 0. For example, had we started with the equation

$$x + y + e^{xy} = 2 \tag{2.26}$$

[15]This does not always happen—Examples 2.8 and 2.12, and several other examples in these notes, just happen to have F's with this property.

instead of (2.24), we would have had a much harder time. We could show by implicit differentiation that every differentiable function determined by (2.26) is a solution of (2.25)—thus, that criterion (ii) of Definition 2.6 is satisfied—but that would not tell us that there is even a single function of x defined by (2.26), or even that the graph of (2.26) contains any points whatsoever. Conceivably, we could be in the same situation as in Example 2.10, in which all differentiable functions implicitly defined by (2.21)—all none of them—are solutions of our differential equation.

It so happens that we *can* show that the graph of (2.26) contains a point at which $\frac{\partial G}{\partial y}$ is not 0. However, doing that would require a digression that we do not want to take right now. Instead, let us consider a different type of problem that can be handled far more easily, even though the function $F(x, y)$ is much more complicated.

Example 2.13 Show that there is a number c_0 for which the equation

$$e^x + x + y^5 - y^4 + y^3 + y^2 + xy^2 = c_0 \tag{2.27}$$

is an implicit solution of the differential equation

$$e^x + 1 + y^2 + (5y^4 - 4y^3 + 3y^2 + 2y + 2xy)\frac{dy}{dx} = 0. \tag{2.28}$$

To approach this problem, we start with a variation on the second step of Examples 2.8 and 2.12: we assume that there is a number c_0 for which (2.27) implicitly determines a differentiable function ϕ, say on an interval I. On the interval I, we may then implicitly differentiate the equation (2.27)—i.e. differentiate with respect to x both sides of the equation we obtain by substituting "$y = \phi(x)$" into (2.27). To keep the notation as simple as possible, we will just write "y" instead of "$y(x)$" or "$\phi(x)$" when we differentiate. (This is usually what we do when we differentiate implicitly; we just haven't done it until now in these notes.) Then, using the chain rule and product rule, we find

$$e^x + 1 + 5y^4\frac{dy}{dx} - 4y^3\frac{dy}{dx} + 3y^2\frac{dy}{dx} + 2y\frac{dy}{dx} + y^2 + 2xy\frac{dy}{dx} = 0,$$

which is equivalent to equation (2.28).

Thus, all differentiable functions ϕ determined implicitly by an equation of the form (2.27) will be solutions of (2.28). Thus for any c_0 for which (2.27) implicitly determines a differentiable function, equation (2.27) will be an implicit solution of (2.28).

So, if we can show that there *is* such a c_0, we'll be done. For this, we look to the Implicit Function Theorem to help us out. Letting $F(x, y)$ denote the left-hand side of (2.27), we compute

$$\frac{\partial G}{\partial x}(x,y) = e^x + 1 + y^2, \tag{2.29}$$

$$\frac{\partial G}{\partial y}(x,y) = 5y^4 - 4y^3 + 3y^2 + 2y + 2xy. \tag{2.30}$$

Both partials are continuous on the whole xy plane, so whatever point we choose for (x_0, y_0), the Implicit Function Theorem's hypothesis that the partials be continuous on some open rectangle containing (x_0, y_0) will be satisfied. Let's look for a point (x_0, y_0) at which $\frac{\partial G}{\partial y}$ is not 0. From our computation above,

$$\frac{\partial G}{\partial y}(x,y) = y(5y^3 - 4y^2 + 3y + 2 + 2x). \tag{2.31}$$

So we definitely *don't* want to choose $y_0 = 0$. But if we choose y_0 to be anything other than 0, we can certainly find an x_0 for which the quantity inside parentheses isn't zero. Let's make things easy on ourselves and choose $y_0 = 1$. Then

$$\begin{aligned} 5y_0^3 - 4y_0^2 + 3y_0 + 2 + 2x_0 &= 6 + 2x_0 \\ &\neq 0 \text{ as long as } x_0 \neq -3. \end{aligned}$$

So if we take, for example, $(x_0, y_0) = (0, 1)$, then $\frac{\partial G}{\partial y}(x_0, y_0) \neq 0$. For this choice of (x_0, y_0), we have $F(x_0, y_0) = 3$. The Implicit Function Theorem then guarantees us that on some open x-interval containing 0, the equation $F(x, y) = 3$ implicitly determines a differentiable function of x. By the first part of our analysis (the part that involved implicit differentiation), this guarantees that the equation $F(x, y) = 3$ is an implicit solution of (2.28). So we have found a c_0 with the desired property. ■

As you probably noticed, in this example our expressions (2.29)–(2.30) for the partial derivatives of F appeared also in (2.28). This is no accident. As students who have taken Calculus 3 know, the multivariable chain rule implies that if we implicitly differentiate the equation $F(x, y) = c_0$ with respect to x, we obtain the equation

$$\frac{\partial G}{\partial x} + \frac{\partial G}{\partial y}\,\frac{dy}{dx} = 0. \tag{2.32}$$

With foresight, the author chose the DE (2.28) to be exactly the equation (2.32) for $F(x, y)$ equal to the left-hand side of (2.27). For *most* DEs, it will *not* be true that there is a value of c_0 for which (2.27) is an implicit solution.

It may seem to you that the author cheated, by choosing essentially the only DE for which the fact you were instructed to establish was actually a true fact. But you will see later that equations of the form (2.32) actually come up a lot.

You may also have noticed, in Example 2.13, that we could have come up with a whole lot of points (x_0, y_0) that "worked", in the sense that the hypotheses of the Implicit Function Theorem would have been met. All we needed was a point (x_0, y_0) for which $y(5y^3 - 4y^2 + 3y + 2 + 2x)|_{(x_0,y_0)} \neq 0$. But "almost every" choice (x_0, y_0) has this property; we just need $y_0 \neq 0$ and $x_0 \neq -\frac{1}{2}(5y_0^3 - 4y_0^2 + 3y_0 + 2)$. For each nonzero choice of y_0, there's only one "bad" choice of x_0; every other real number is a good choice of x_0. So the c_0's for which our method shows that (2.27) is an implicit solution of (2.28), are all the numbers $F(x_0, y_0)$ we can get by plugging in "good" choices of (x_0, y_0) (i.e. all choices with $y_0 \neq 0$ and $x_0 \neq -\frac{1}{2}(5y_0^3 - 4y_0^2 + 3y_0 + 2)$). We can expect this set of numbers to be a large subset of the range of F—perhaps the whole range of F. A challenging question for you to think about is this: are there any numbers c_0 for which (2.27) is *not* an implicit solution of (2.28)? Let's strip away the distracting complexity of the function F in (2.27) and pose the analogous question for a much simpler F, the one in Example 2.12:

Question: Are there any numbers c_0 for which the equation

$$x + y + e^{xy} = c_0$$

is *not* an implicit solution of (2.25)? (Note that (2.25) is the equation (2.32) for the function F defined by $F(x, y) = x + y + e^{xy}$.) ■

This question will not be answered in these notes; it is left as a challenge for the student. We point out that the answer to such a question will not be the same for all functions F that we could put on the left-hand side of "$F(x, y) = c_0$". For example, if we take $F(x, y) = x^2 + y^2$, then only for $c_0 > 0$ is the equation $F(x, y) = c_0$ an implicit solution of (2.16) (which is the equation (2.32) for this F, simplified by dividing by 2). But if we take $F(x, y) = x + y$, then for every real number c_0 the equation $F(x, y) = c_0$ is an implicit solution of the analogous differential equation, $1 + \frac{dy}{dx} = 0$, as you can see easily by explicitly solving the equation $x + y = c_0$ for y in terms of x.

The Implicit Function Theorem is one of the most important theorems in calculus, and it is crucial to the understanding of implicit solutions of differential equations. However, it does have its limitations: there are differential equations that have implicitly-defined solutions that are *not* functions given by the Implicit Function Theorem, as the next example shows.

Example 2.14 Consider the algebraic equation

$$x^2 - y^2 = 0 \tag{2.33}$$

and the differential equation

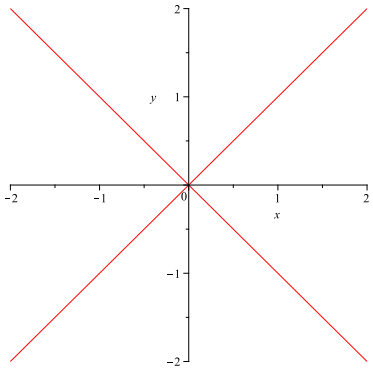

Figure 3: The graph of $x^2 - y^2 = 0$.

$$x - y\frac{dy}{dx} = 0. \tag{2.34}$$

Equation (2.33) is equivalent to $y = \pm x$. Thus on any interval I, equation (2.33) implicitly determines two differentiable functions ϕ of x, namely $\phi(x) = x$ and $\phi(x) = -x$. Both of these are solutions of (2.34). Therefore (2.33) is an implicit solution of (2.34), and the two functions ϕ above are implicitly-defined solutions of (2.34), on any interval.

The point $(x, y) = (0, 0)$ satisfies (2.33). But on no open rectangle containing the point $(0, 0)$ does (2.33) uniquely determine y as a function of x. Every such rectangle will contain both a portion of the graph of $y = x$ and a portion of the graph of $y = -x$ (see Figure 3; draw any rectangle enclosing the origin). Thus there are no intervals I_1 containing 0 (our x_0) and J_1 containing 0 (our y_0) for which (2.12) holds.

Does this contradict the Implicit Function Theorem? No—the theorem says only that there are I_1 and J_1 with the property (2.12) *if the hypotheses of the theorem are met*. But in the current example, the function F for which (2.33) is of the form $F(x, y) = c_0$ is given by $F(x, y) = x^2 - y^2$. Thus $\frac{\partial G}{\partial y}(x, y) = -2y$, and if we take $(x_0, y_0) = (0, 0)$ then $\frac{\partial G}{\partial y}(x_0, y_0) = 0$. One of the hypotheses of the theorem is not met, and therefore we can draw no conclusion from the theorem. The two functions ϕ above are perfectly good implicitly-defined solutions of (2.34); they just are not solutions that the Implicit Function Theorem finds. ■

For most two-variable functions F that we encounter in practice, the "bad points" (x_0, y_0) at which the Implicit Function Theorem does not apply are of two types: points at which the graph of $F(x, y) = F(x_0, y_0)$ has a vertical tangent (as is the case for the equations graphed in Figures 1 and 2), and points at which two or more smooth curves intersect (as in Figure 3; in this simplest of examples the intersecting curves are straight lines).

The equation $x^2 - y^2 = 0$ has another feature that none of our previous examples have illustrated. On any open x-interval containing the origin, the equation implicitly determines two *differentiable* functions of x, but four *continuous* functions of x: $\phi(x) = x, \phi(x) = -x$, $\phi(x) = |x|$, and $\phi(x) = -|x|$. In all of our previous examples, on any open interval the continuous implicitly-defined functions and the differentiable implicitly-defined functions were the same.

From the examples presented so far, and the examples in most textbooks, the student may get the false impression that "implicit solution" means "An equation that, after I implicitly differentiate, I can rearrange back to the DE." That is *not* the definition, however (Definition 2.6 does not mention implicit differentiation, or require the function F in the definition to be differentiable). Below are two examples that illustrate this point.

Example 2.15 Determine whether the equation

$$2|x| + |y| = 2 \tag{2.35}$$

is an implicit solution of

$$\left(\frac{dy}{dx}\right)^2 = 4|x| + 2|y|. \tag{2.36}$$

If we try to approach this just by implicit differentiation, we run into trouble because the function $F(x, y) = 2|x| + |y|$ is not differentiable anywhere that $x = 0$ or $y = 0$. However, if we run through all the sign-possibilities in (2.35) and solve for y in terms of x, we see that (2.35) the graph of (2.35), a "stretched diamond" with vertices at $(\pm 1, 0)$ and $(0, \pm 2)$, consists of the graphs of the following four equations:

$$\begin{aligned} y &= -2x + 2, \quad 0 \le x \le 1, \\ y &= 2x - 2, \quad 0 \le x \le 1, \\ y &= 2x + 2, \quad -1 \le x \le 0, \\ y &= -2x - 2, \quad -1 \le x \le 0. \end{aligned}$$

Therefore (2.35) determines the following four differentiable functions:

$$\begin{aligned} \phi(x) &= -2x + 2, \quad 0 < x < 1, \\ \phi(x) &= 2x - 2, \quad 0 < x < 1, \\ \phi(x) &= 2x + 2, \quad -1 < x < 0, \\ \phi(x) &= -2x - 2, \quad -1 < x < 0. \end{aligned}$$

Every differentiable function of x determined by (2.35), with domain an open interval, is one of these four functions (or the restriction of one of these to a smaller interval). For each of these functions we have $\phi'(x) \equiv 2$ or $\phi'(x) \equiv -2$, so for any of these functions if substitute $y = \phi(x)$ into (2.36), we find

$$\begin{aligned} \text{left-hand side of (2.35)} &\equiv 4, \\ \text{right-hand side of (2.35)} &= 2(2|x| + |y(x)|) \\ &\equiv 2 \times 2 \quad \text{(because of (2.35))} \\ &= 4. \end{aligned}$$

Therefore (2.36) is satisfied on the domain of ϕ, for all four choices of ϕ. Both criteria of (2.6) are satisfied, so (2.35) is an implicit solution of (2.36). ■

Example 2.16 Determine whether the equation

$$y^5 + y = x^5 + x \tag{2.37}$$

is an implicit solution of

$$\frac{dy}{dx} = \frac{5x^4 + 1}{5(x^5 + x - y)^{4/5} + 1} \ . \tag{2.38}$$

First, we observe that the graph of (2.37) has at least one point: the point $(0, 0)$.

Next, we rewrite (2.37) as $F(x, y) = 0$, where $F(x, y) = y^5 + y - x^5 - x$. Then $\frac{\partial G}{\partial y} = 5y^4 + 1$, which is continuous and positive on the whole xy plane. In particular, $\frac{\partial G}{\partial y}$ is continuous and nonzero at $(0, 0)$, so the Implicit Function Theorem guarantees us that (2.37) determines a differentiable function of x near the point $(0, 0)$ on the graph of $F(x, y) = 0$.

So (2.37) determines at least one differentiable function of x. If ϕ is any such function, then substituting $y = \phi(x)$ into (2.37) and differentiating implicitly, we find $(5y^4 + 1)\frac{dy}{dx} = 5x^4 + 1$, which implies

$$\frac{dy}{dx} = \frac{5x^4 + 1}{5y^4 + 1} \tag{2.39}$$

on the domain of ϕ (the denominator $5y^4 + 1$ is never zero). Hence ϕ is a solution of (2.39).

Now, (2.39) does not look like (2.38). The two DEs are not equivalent; there are points (x, y) at which the right-hand side of (2.38) is not equal to the right-hand side of (2.39). But that doesn't mean that (2.37) can't be an implicit solution of (2.38).

And, in fact, if we simply observe that on the graph of (2.37) we have $y^5 = x^5 + x - y$, implying $y^4 = (x^5 + x - y)^{4/5}$. Therefore for $y = \phi(x)$ we have

$$\frac{dy}{dx} = \frac{5x^4 + 1}{5y^4 + 1} = \frac{5x^4 + 1}{5(x^5 + x - y)^{4/5} + 1},$$

so ϕ is a solution of (2.38). Therefore (2.37) is an implicit solution of (2.38). ■

In the example above, it is irrelevant whether there are *some* solutions of (2.39) that are not solutions of (2.38). The question was not whether *every* solution of (2.39) was a solution of (2.38), but only whether a *specific* solution of (2.39), namely a function determined implicitly by (2.37), was a solution of (2.38).

Remark 2.17 (Families of implicit solutions) *Every* equation of the form $F(x, y) =$ constant that implicitly determines some differentiable function of x, and in which F is differentiable, is an implicit solution of the DE found by implicitly differentiating "$F(x, y) =$ constant", namely (2.32). But for any such F and constant C_0, the DE (2.32) is not the *only* DE of which "$F(x, y) = C_0$" is an implicit solution; there are always inequivalent DEs of which "$F(x, y) = C_0$" is an implicit solution.[16] However, you are unlikely to find examples like Example 2.15 or Example 2.16 in a DE textbook. In a typical DE course, implicit solutions tend to arise from solving two types of equations—separable derivative-form DEs and exact differential-form DEs. For any of these equations, there is always a *family* of solutions (not always an exhaustive family, in the separable-DE case) of the form

$$\{F(x, y) = C\}, \tag{2.40}$$

[16] *Note to instructors*: This point is not made in any textbook I have seen. This is one reason that I find the treatment of "implicit solution" in current textbooks to be misleading. Every example of implicit solution I see in textbooks that formalize the term, is an example of something much more restricted: an element of a *family* of implicit solutions $\{F(x, y) = c\}$. Part of the problem is that these books are defining something that they effectively never use, *single* implicit solutions rather than *families* of implicit solutions. This leaves the student with the impression that the meaning of "implicit solution" is something other than what his/her textbook-author has defined the term to mean. At least one older textbook, [4], entirely avoids this problem by introducing *families of curves* before any notion of "implicit solution" is used (the term "implicit solution" itself is not used in [4]). Indeed, there really is *no need ever to use the term "implicit solution"*. For example, an equation that meets the definition of "implicit solution" in these notes can be called "an *implicit formula* for a solution", or "a solution in *implicit form*". For another example, it is perfectly reasonable to say, "The general solution of $x + y\frac{dy}{dx} =$, *in implicit form*, is $\{x^2 + y^2 = c \mid c > 0\}$." (I do not agree that the term "general solution" needs to be avoided for all nonlinear equations, but if you don't like the use of "general solution" here, just substitute "the set of all solutions".) The reason I have given a definition for "implicit solution" in these notes is *not* that I think the term should be used; it is that *if* authors and instructors are going to continue using it in a formal manner, a definition that is needed that is accurate, precise, complete, understandable by students, and sensible.

where F is function that depends on the DE, and c is a constant ranging over some (often difficult to specify) interval I that may or may not be the whole real line. (I.e. for each c in this interval, the equation $F(x, y) = C$ is an implicit solution of the DE.) *Every* differentiable function implicitly determined by *any* member of the family (2.40) is a solution of the *same* DE, namely (2.32).

2.3 Maximal and general solutions of derivative-form DEs

Often we want to talk about the collection of all solutions of a given differential equation without pinning ourselves down to a specific interval I. For example, it may happen we can write down a family of solutions, distinguished from each other by the choice of some constant C, but for which the domain depends on the value of C and hence differs from solution to solution. This suggests making the following definition:

Definition 2.18 (temporary) For a given G, the *general solution* of the differential equation

$$\mathsf{G}(x, y, \frac{dy}{dx}) = 0 \tag{2.41}$$

is the collection of all solutions of (2.41), where "solution" is defined as in the second paragraph of Definition 2.1. Said another way, the general solution of (2.41) is the collection of pairs (I, ϕ), where I is an open interval and ϕ is a solution of (2.41) on I.

We warn the student that the terminology "general solution" (with or without the restriction "on an interval I") is not agreed upon by all mathematicians (except for linear equations in "standard linear form", which we have not yet discussed in these notes), for reasons discussed at the end of this subsection.

There is a problem with Definition 2.18 that we will discuss shortly. However, in their first exposure to the subject, many students will not have the mathematical sophistication needed to understand the problem or the way to fix it. Therefore **in a first course on differential equations, it is acceptable to use Definition 2.18 as the definition of "general solution", and students in this author's course will not be penalized for doing so.** Some students, however, may recognize (eventually, if not immediately) that there is a problem. The discussion below is for those students, and any others who might be interested in what the problem is. **Students who are not interested, or have trouble understanding the discussion, should skip to Example 2.23 and simply ignore the word "maximal" wherever it appears in these notes.**

To illustrate the problem, consider the separable equation $\frac{dy}{dx} = -y^2$. It is easy to show that for every solution ϕ other than the constant solution $\phi \equiv 0$, there is a constant C such that

$$\phi(x) = \frac{1}{x - C} \ . \tag{2.42}$$

on the domain of the solution. Remembering that the domain of a solution of a DE is required to be an *interval*, we look at equation (2.42) and say, "Okay, for each C this formula gives two solutions, one on $(-\infty, C)$ and (C, ∞)." But even this is not technically correct. These are not the only two intervals on which equation (2.42) defines solutions. If ϕ is a solution on (C, ∞), then it satisfies the DE at every point of this interval. Therefore it also satisfies the DE at every point of $(C, C + 1)$, at every point of $(C + 26.4, C + 93.7)$, and on any open subinterval of $(-\infty, C)$ or (C, ∞) whatsoever.

This example illustrates that the collection of pairs (I, ϕ) referred to in Definition 2.18 has a certain redundancy. There is terminology that allows us to speak more clearly about this redundancy:

Definition 2.19 Let ϕ be a function on an interval I and let I_1 be a subinterval of I. The *restriction of ϕ to I_1*, denoted $\phi|_{I_1}$, is defined by

$$\phi|_{I_1}(x) = \phi(x) \ \text{ for all } x \in I_1 \ .$$

(We leave $\phi|_{I_1}(x)$ undefined for x not in I_1.) We say that a function ψ is a restriction of ϕ if it is the restriction of ϕ to some subinterval.

If $\tilde{I}$ is an interval containing I, and $\tilde{\phi}$ is a function on $\tilde{I}$ whose restriction to I is ϕ, then we call $\tilde{\phi}$ an *extension* of ϕ.[17]

Equivalently: if $\tilde{I}$ is an interval of which I is a subinterval, and $\tilde{\phi}$ and ϕ are functions defined on $\tilde{I}$ and I respectively, then

$$\begin{aligned} \phi \text{ is a restriction of } \tilde{\phi} \quad &\Longleftrightarrow \quad \text{the graph of } \phi \text{ is part of the graph of } \tilde{\phi}, \\ &\Longleftrightarrow \quad \tilde{\phi} \text{ is an extension of } \phi. \end{aligned}$$

(The symbol " $\Longleftrightarrow$ " means "if and only if". When preceded by a comma, as in the transition from the first line above to the second, you should read the combination ", $\Longleftrightarrow$ " as "which is true if and only if".)

It may seem silly at first, and even outright confusing, to distinguish so carefully between a function and its restriction to a smaller domain, but there are many times in mathematics in which it is important to do this. For example, the sine function does not have an inverse, but the *restriction* of sine to the interval $[-\pi/2, \pi/2]$ does, and the inverse of this *restricted* function is the function we call $\sin^{-1}$ or arcsin.

[17] The same definition applies even when the domains of interest are not intervals; e.g. for a function ϕ with any domain whatsover, the restriction of ϕ to any subset of its domain is defined the same way. But for functions of one variable, the DE student should remain focused on domains that are intervals.

If a function ϕ is a solution of a given DE on some interval I then the restriction of ϕ to any subinterval I_1 is also a solution. But of course, if we know the function ϕ, then we know every speck of information about $\phi|_{I_1}$. Therein lies the redundancy of Definition 2.18: the definition names a much larger collection of functions than is needed to capture all the information there is to know about solutions of (2.41). We will see below that we can be more efficient.

While we can always restrict a solution ϕ of a given DE to a smaller interval and obtain a (technically different) solution, a more interesting and much less trivial problem is whether we can *extend* ϕ to a solution on a *larger* interval. The extension concept is always in the background whenever we talk about "the domain of a solution of an initial-value problem". When we say these words, it's always understood that we're looking for the *largest* interval on which the formula we're writing down is actually a solution of the given IVP. This is the differential-equations analog of what is often called the *implied domain* of a function represented by a formula, such as $f(x) = \frac{1}{x}$, in Calculus 1 or precalculus courses. The implied domain of this function f is $(-\infty, 0) \bigcup (0, \infty)$ (also frequently written as "$\{x \neq 0\}$"). However, if we are talking about "$y = \frac{1}{x}$" as a solution of the IVP

$$\frac{dy}{dx} = -x^{-2}, \quad y(3) = \frac{1}{3}, \tag{2.43}$$

then we would call "$y = \frac{1}{x}$" a solution of this IVP only on $(0, \infty)$, not on the whole domain of the formula " $\frac{1}{x}$ ".

With these ideas in mind, we call a solution ϕ of a given DE (or initial-value problem) on an interval I *maximal* or *inextendible* if ϕ cannot be extended to any open interval $\tilde{I}$ strictly containing I, while still remaining a solution of the DE.

Example 2.20 All the functions ϕ below are different functions, even though we are using the same letter for them.

- $\phi(x) = \frac{1}{x},\ 0 < x < 5$, is a solution of $\frac{dy}{dx} = -x^{-2}$, but not a maximal solution. It is also a solution of the IVP (2.43).
- $\phi(x) = \frac{1}{x},\ 2.9 < x < 16.204$, is another solution of $\frac{dy}{dx} = -x^{-2}$, and of the IVP (2.43), but not a maximal solution.
- $\phi(x) = \frac{1}{x},\ 3.1 < x < 16.204$, is another solution of $\frac{dy}{dx} = -x^{-2}$, but it is neither a maximal solution nor a solution of the IVP (2.43),
- $\phi(x) = \frac{1}{x},\ x \in (0, \infty)$ is *a* maximal solution of $\frac{dy}{dx} = -x^{-2}$, and is *the* maximal solution of the IVP (2.43).
- $\phi(x) = \frac{1}{x},\ x \in (-\infty, 0)$ is a *different* maximal solution of $\frac{dy}{dx} = -x^{-2}$. It is *not* a solution of the IVP (2.43).

- $\phi(x) = \frac{1}{x},\ x \in (-\infty, -\sqrt{2})$ is another non-maximal solution of $\frac{dy}{dx} = -x^{-2}$.
- $\phi(x) = \frac{1}{x} + 37,\ x \in (0, \infty)$ is yet another maximal solution of $\frac{dy}{dx} = -x^{-2}$. It is not a solution of the IVP (2.43).

Example 2.21 The maximal solutions of the differential equation $\frac{dy}{dx} = \sec^2 x$ are the functions ϕ defined by

$$\phi(x) = \tan x + C, \quad (n - \frac{1}{2})\pi < x < (n + \frac{1}{2})\pi, \quad n \text{ an integer}, \quad C \text{ a constant}$$

(one maximal solution for each pair of values (n, C) with n an integer and C real).

It can be shown that every non-maximal solution of a DE is the restriction of some maximal solution of that DE.[18] Thus the collection of maximal solutions "contains" all solutions in the sense that the graph of every solution is contained in the graph of some maximal solution. So, better than Definition 2.18 is this:

Definition 2.22 For a given G, the *general solution* of (2.41) is the collection of all maximal solutions of (2.41).

(This definition supersedes Definition 2.18.)

Example 2.20 demonstrates, we hope, the economy gained by including the word "maximal" in this definition. The student will probably agree that, even prior to writing down Definition 2.22, maximal solutions are what we really would have been thinking of had we been asked what all the solutions of "$\frac{dy}{dx} = -x^{-2}$" are—we just might not have realized consciously that that's what we were thinking of.

Example 2.23 The general solution of $\frac{dy}{dx} = x$ may be written in short-hand as

$$\left\{ y = \frac{1}{2}x^2 + C \right\}. \tag{2.44}$$

In this context equation (2.44) represents a one-parameter family of maximal solutions ϕ_C, each of which is defined on the whole real line. Here C is an arbitrary constant; every real number C gives one solution of the DE. (That's why the curly braces are written in (2.23); they tell us we're talking about a *set* of objects of the form within the braces.) We allow ourselves to write (2.44) as short-hand for "the collection of functions $\{\phi_C \mid C \in \mathbf{R}\}$, where $\phi_C(x) = \frac{1}{2}x^2 + C$".[19]

[18] Said another way, every solution can be extended to *at least one* maximal solution. Maximal extensions always exist, but they are not always unique.

[19] Students in my own classes are permitted to omit the curly braces in (2.44), but I am trying to maintain certain notational consistency across different sections of these notes.

Example 2.24

- The general solution of

$$\frac{dy}{dx} = -x^{-2}, \quad x > 0 \tag{2.45}$$

(meaning that we are interested in this differential equation only for $x > 0$) may be written as

$$\left\{ y = \frac{1}{x} + C \right\}, \ x > 0, \tag{2.46}$$

a one-parameter family of maximal solutions. Because the restriction $x > 0$ is stated explicitly in (2.45), it is permissible to leave out the "$x > 0$" when writing the general solution; we may simply write the general solution as

$$\left\{ y = \frac{1}{x} + C \right\} \tag{2.47}$$

- The general solution of

$$\frac{dy}{dx} = -x^{-2}, \tag{2.48}$$

with no interval specified, may also be written as (2.47)—i.e. it is *permissible* to write it this way, in the interests of saving time and space. However, because no interval was specified when the DE was written down, we must consider all possible intervals. Therefore, in this context, equation (2.47) does *not* represent a one-parameter family of maximal solutions; it represents *two* one-parameter families of maximal solutions[20]. Equation (2.47) is acceptable short-hand for

[20] Many calculus textbooks, and especially integral tables, foster a misunderstanding of the indefinite integral. *By definition*, for functions f that are continuous on an open interval or a union of disjoint open intervals, "$\int f(x)dx$" means "the collection of all antiderivatives of f". If the implied domain of f is an open interval, then this collection is the same as the general solution of $dy/dx = f(x)$. But we must be careful not to interpret formulas such as "$\int x^{-2}\ dx = -x^{-1} + C$" or "$\int \sec^2 x\ dx = \tan x + C$" as saying that every antiderivative of x^{-2} is of the form $x^{-1} + C$ *on the whole implied domain of the integrand* x^{-2}, or that every antiderivative of $\sec^2 x$ is of the form $\tan x + C$ *on the whole implied domain of the integrand* $\sec^2 x$.

The Fundamental Theorem of Calculus tells us that *on any open interval on which a function f is continuous*, any two antiderivatives of f differ by an additive constant. (Equivalently, if F is any *single* antiderivative of f on this interval, then *every* antiderivative of f on this interval is $F + C$ for some constant C.) It does *not* make any statement about antiderivatives on domains that are not connected, such as the implied domain of $f(x) = x^{-2}$ or the implied domain of $f(x) = \sec^2 x$.

$$\left.\begin{array}{l} \text{the union of the two families of functions} \\ \qquad \{\phi_C \mid \ C \in \mathbf{R}\}, \quad \{\psi_C \mid \ C \in \mathbf{R}\} \\ \text{where} \\ \qquad \phi_C(x) = \frac{1}{x} + C, \quad x > 0 \\ \text{and} \\ \qquad \psi_C(x) = \frac{1}{x} + C, \quad x < 0. \end{array}\right\} \tag{2.49}$$

(The *union* of the two families means the collection of functions that are in one family or the other.[21]) The solution $y = \frac{1}{x} + 6$ on $\{x < 0\}$ (the function ψ_6 in the notation of (2.49)) is no more closely related to the solution $y = \frac{1}{x} + 6$ on $\{x > 0\}$ (the function ϕ_6) than it is to the solution $y = \frac{1}{x} + 7$ on $\{x < 0\}$ (the function ψ_7) ; in fact it is *much less* closely related. (The function ψ_7 at least has the same domain as ψ_6, where as ϕ_6 does not.)

Alternative ways of writing the general solution of $\frac{dy}{dx} = -x^{-2}$ are

$$\text{“}\{y = \frac{1}{x} + C, x > 0\} \ \text{ and } \ \{y = \frac{1}{x} + C, x < 0\}\text{”} \tag{2.50}$$

and

$$\text{“}\{y = \frac{1}{x} + C_1, x > 0\} \ \text{ and } \ \{y = \frac{1}{x} + C_2, x < 0\}\text{”}. \tag{2.51}$$

In (2.50), it is understood that, *within each family*, C is an arbitrary constant, and that the two C's have nothing to do with each other. In (2.51), C_1 and C_2 again are arbitrary constants, and we have simply chosen different notation for them to emphasize that they have nothing to do with each other. But all three forms (2.47), (2.50), and (2.51) are acceptable ways of writing the general solution, as long as we understand what they mean, and are communicating with someone else who understands what they mean. These forms do not exhaust all permissible ways of writing the general solution; there are other notational variations on the same theme.

Example 2.25 The general solution of $\frac{dy}{dx} = \sec^2 x$ may be written as

[21] *Note to instructors:* Not wanting to over-burden students with new notation and terminology—of which there is already a fair bit in these notes—I have opted not to use the symbol $\bigcup$. You will notice later on, e.g. in (2.50), that in these notes I often write the union of two sets A, B as “A and B”. Of course, if I were describing the *elements* of the union, and had everything within just one pair of set-braces, I would have to use the conjunction “or”, not “and”, but I've deliberately avoided writing (2.50) and similar expressions this way. I felt that using the word “or” in these expressions would be confusing to students.

$$\{y = \tan x + C\}\,, \tag{2.52}$$

or as

$$\left\{y = \tan x + C, \quad (n - \frac{1}{2})\pi < x < (n + \frac{1}{2})\pi, \quad n \text{ an integer}\right\}, \tag{2.53}$$

or as

$$\left\{y = \tan x + C_n, \quad (n - \frac{1}{2})\pi < x < (n + \frac{1}{2})\pi, \quad n \text{ an integer}\right\}, \tag{2.54}$$

or in various other ways that impart the same information. As in the "$\frac{dy}{dx} = -x^{-2}$" example, it is understood that C and C_n above represent arbitrary constants (i.e. that they can assume all real values). But whichever of the forms (2.52)–(2.54) (or other variations on the same theme) that we choose for writing the general solution of $\frac{dy}{dx} = \sec^2 x$, we should not forget that each of these forms represents *an infinite collection of one-parameter families of maximal solutions*, one family for each interval of the form $(n - \frac{1}{2})\pi < x < (n + \frac{1}{2})\pi$ (where n is an arbitrary integer).

Example 2.26 The general solution of the separable equation

$$\frac{dy}{dx} = -y^2 \tag{2.55}$$

may be written as

$$\left\{y = \frac{1}{x - C}\right\} \text{ and } \{y = 0\}, \tag{2.56}$$

or in various other ways that impart the same information[22]. In the given context, the solution that is the constant function 0 may be written as "$y = 0$", as in (2.56) or as "$y \equiv 0$" (which, in this context, is read "y identically zero"). Since a solution of (2.55), expressed in terms of the variables in (2.55), is function of x, the only correct interpretation of "$y = 0$" in (2.56) is "y is the constant function whose value is zero for all x", *not* "y is a real number, specifically the number 0". An instructor may sometimes write a constant function using the identically-equal-to symbol "$\equiv$", especially in the early weeks of a DE course, to make sure that students are absolutely clear what is meant; at other times, when there is little possibility of confusion, (s)he may just use the ordinary "=" symbol.

Note that for each C, the equation "$y = \frac{1}{x-C}$" represents not one maximal solution, but two: one on the interval (C, ∞) and one on the interval $(-\infty, C)$.

[22]We do not discuss here how to *figure out* the general solution of this DE, since that is adequately covered outside these notes.

This example is very different from our previous ones. For the DE "$\frac{dy}{dx} = -x^{-2}$", every maximal solution had domain either $(-\infty, 0)$ or $(0, \infty)$, and on each of these intervals there were infinitely many maximal solutions. For the DE "$\frac{dy}{dx} = \sec^2 x$", there were infinitely many maximal solutions on every interval of the form $((n - \frac{1}{2})\pi, (n + \frac{1}{2})\pi)$. By contrast, for the differential equation (2.55):

1. The domain of every maximal solution is different from the domain of every other.

2. For every interval of the form (a, ∞) there is a maximal solution whose domain is that interval, namely $y = \frac{1}{x-a}$.

3. For every interval of the form $(-\infty, a)$ there is a maximal solution whose domain is that interval, namely $y = \frac{1}{x-a}$. (The *formula* is the same as for solution on (a, ∞) mentioned above, but we stress again that the fact that *as solutions of a differential equation*, "$y = \frac{1}{x-a},\ x > a$" and "$y = \frac{1}{x-a},\ x < a$" are *completely unrelated* to each other.)

4. There is one maximal solution whose domain includes the domain of every other, namely $y \equiv 0$.

The general solution of (2.55) also exhibits another interesting phenomenon. The way we have written the general solution in (2.56) isolates the maximal solution $y \equiv 0$ as not belonging to what appears to be a single nice family into which the other maximal solutions fall (there is no value of C for which the formula "$y = \frac{1}{x-C}$" produces the constant function 0). But for $C \neq 0$, writing $K = \frac{1}{C}$,

$$\frac{1}{x - C} = \frac{C^{-1}}{C^{-1}x - 1} = \frac{K}{Kx - 1} \ . \tag{2.57}$$

In the right-most formula in (2.57), we get a perfectly good function—the constant function 0—if we set $K = 0$. But this function is exactly what appeared to be the "exceptional" maximal solution in (2.56). Thus, we can rewrite the general solution (2.56) as

$$\left\{ y = \frac{K}{Kx - 1} \right\} \quad \text{and} \quad \left\{ y = \frac{1}{x} \right\} \ . \tag{2.58}$$

Here, K is an arbitrary constant, allowed to assume all real values, just as C was allowed to in (2.56); we could just as well use the letter C for it. Writing the general solution as in (2.58), the two solutions with formula $y = \frac{1}{x}$ (one for $x > 0$, one for $x < 0$) may be viewed as the exceptional ones, with all the others—including the constant function 0—falling into the "$\frac{K}{Kx-1}$" family. This illustrates that there be more than one way of expressing the collection of all maximal solutions as what looks like a "nice family" containing most of the maximal solutions, plus one or more

maximal solutions that don't fall into the family. This illustrates that *"falling into a family" can be in the eye of the beholder*, and not something intrinsic to a solution of a DE.

But this example also provides another instance of a theme to which we keep returning: how easy it is to mis-identify a family of *formulas* with a family of *solutions of a DE*. The maximal solutions described by $\{y = \frac{1}{x-C}\}$ in (2.56) do not form *one* one-parameter family; they form *two*.[23] Every value of C corresponds to two maximal solutions, one defined to the left of C and one defined to the right[24]. In (2.58), the "family" $\{y = \frac{K}{Kx-1}\}$ is even more deceptive: for each *nonzero* K, the formula $y = \frac{K}{Kx-1}$ yields two maximal solutions, one defined to the left of $1/K$ and one defined to the right, while for $K = 0$ the formula yields just one maximal solution.

In this example, one may reasonably decide that (2.56) is preferable to (2.58) as a way of writing down the general solution. The constant solution $y \equiv 0$ is distinguished from all the others not just by being constant, but by being the only solution defined on the whole real line. Furthermore, the collection of solutions described by $\{y = \frac{1}{x-C}\}$ is more "uniform" than is the collection described by $\{y = \frac{K}{Kx-1}\}$, in the sense that in the first collection, *every* value of the arbitrary constant corresponds to two maximal solutions, while in the second collection there is a value of the arbitrary constant, namely 0, for which the given formula defines only one maximal solution. However, in the next example, we will see two different ways of writing the general solution, neither of which can be preferred over the other by any such considerations.

Example 2.27 The general solution of the separable equation

$$\frac{dy}{dx} = y(1-y) \tag{2.59}$$

may be written as

$$\left\{ y = \frac{C}{e^{-x} + C} \right\} \quad \text{and} \quad \{y \equiv 1\}. \tag{2.60}$$

Using the same method as in the previous example, one sees that the same collection of functions also be written as

$$\left\{ y = \frac{1}{Ce^{-x} + 1} \right\} \quad \text{and} \quad \{y \equiv 0\}. \tag{2.61}$$

[23] This mistake—not necessarily with this particular DE—is made in many, if not all, current DE textbooks that use the phrase "one-parameter family of solutions" somewhere in their treatment of nonlinear first-order DEs.

[24] *Note to instructors:* Of course, the constant solution 0 may be viewed as the "$C = \infty$" case of "$y = \frac{1}{x-C}$", and you may even wish to tell your students that. However, this does *not* mean that the general solution is a one-parameter family parametrized by the one-point compactification of $\mathbf{R}$, i.e. the circle. Such a conclusion would be fine if we were talking the one-parameter family of *rational functions* defined by "$y = \frac{1}{x-C}$", but we are not; we are talking about solutions of an ODE, for which the *only* sensible domain is a connected one.

(Here, the analog of the previous example's K has been renamed to C.) In each case, in the family in curly braces, the formula giving $y(x)$ yields two maximal solutions for $C < 0$ and one maximal solution for $C \geq 0$. The $C = 0$ solution in (2.60) is the constant function 0, which is the "exceptional" solution in (2.61). The $C = 0$ solution in (2.61) is the constant function 1, which is the "exceptional" solution in (2.60). The situation is completely symmetric; neither of (2.60) and (2.61) can be preferred over the other.

The last example illustrates that for nonlinear DEs there may be no singled-out way to write the collection of all maximal solutions (or solutions on a specified interval) of a nonlinear equation as a one-parameter family, or as several one-parameter families, or as one or more one-parameter families of solutions plus some "exceptional" solutions. Because of this, many authors prefer to use the terminology "general solution" *only* for "nice" linear DEs (the meaning of "nice" is not important right now), and not to define the term at all for nonlinear DEs.[25]

2.4 General and implicit solutions on a region

For derivative-form DEs, so far we have defined "general solution on an interval I" and "general solution" (with no interval specified). There are two other types of general solution (of a derivative-form DE) that will be used later in these notes. To talk about these, we first must define what an *open set* is in the xy plane. A subset R of $\mathbf{R}^2$ is *open* if for every point (x_0, y_0) in R, there is some open rectangle that

[25] *Note to instructors:* I feel, however, that too much is lost this way. It is important for students to be able to know when they've found all (maximal) solutions, whether expressed explicitly or implicitly. I have not found a textbook that systematically addresses the question "Have we found all solutions (of a given nonlinear DE)?" at all, or even mentions the question explicitly. I fear that this omission reinforces the prevalent and unfortunate impression that the only thing one needs to do in DEs is push symbols around the page by whatever sets of rules one is told for the various types of equations, and that one does not need to question whether and/or why those rules yield all the solutions.

I feel that it is worthwhile to give the student a name for the collection of all solutions. Of course, "solution-set" would do this, but I fear that students at the level of an intro DE course may have heard this term in "solution-set of an algebraic equation or inequality"—and if so, have heard it *only* in this context—and are too likely to think of a "solution-set" as always being a subset of $\mathbf{R}$ or $\mathbf{R}^2$ or $\mathbf{R}^3$. Hence I have chosen the name "general solution", which is consistent with the use of this term for "nice" n^{th}-order linear DEs, i.e. those for which the solution-set is an n-dimensional affine space.

Of course, you (the instructor) may have a different convention that you prefer for use of the term "general solution". One convention I caution against, however, is to use "general solution" to refer to a *non-exhaustive* collection of solutions (or for a "typical" element of such a collection) for which (s)he has produced a nicely-parametrized family of formulas. As the simple examples 2.26 and 2.27 illustrate, the choice of which solutions should be considered part of a family, and which should be considered exceptional, can be in the eye of the beholder, and can be an artifact of the method used to find the solutions.

contains (x_0, y_0) and is contained in R. Another term we will use for "open subset of $\mathbf{R}^2$ " is *region*[26].

Do not expect to understand right away *why* the following refinement of "general solution" might be needed; this will begin to become clearer in the next section. If you find the definition hard to understand, just **skip over it for now**, and re-read it once it starts being used (fact (2.67)) and the subsequent examples in Section 2.5.

Definition 2.28 (General solution in a region) Let R be a region in the xy plane. The *general solution, in* R, of an equation $\mathsf{G}(x, y, \frac{dy}{dx}) = 0$, is the collection of all solutions whose graphs lie in R and that are maximal in R. Here, by "solution that is maximal in R" we mean a solution, defined on some open interval, that cannot be extended to a solution on a larger open interval without its graph leaving R. ■

In the next definition, you will see the imposing phrase "collection $\mathcal{E}$ of algebraic equations in x and y". An example of what this means is (2.40): for a given function F, each equation $F(x, y) = C$ is an algebraic equation (no derivatives appear), and as we vary C we get a collection of such equations. (We could also have called this collection simply a *set* of equations.)

You also may find criterion (ii) in this next definition hard to understand. If so, don't worry; it will be explained in the second paragraph after the definition.

Definition 2.29 (General solution on a region, in implicit form) For a given G, consider the derivative-form DE

$$\mathsf{G}(x, y, \frac{dy}{dx}) = 0. \tag{2.62}$$

Let R be a region in the xy plane. We call a collection $\mathcal{E}$ of algebraic equations in x and y *the general solution of* (2.62) *in* R*, in implicit form,* if

(i) each equation in the collection $\mathcal{E}$, restricted to R (i.e. with (x, y) required to lie in R) is an implicit solution of (2.62) (see Definitions 2.6 and 2.11), and

(ii) every solution-curve of (2.62) that lies in R, lies in the union of graphs of finitely many or countably many[27] equations in the collection $\mathcal{E}$.

---

[26] *Note to instructors*: I am taking some liberties here. The usual definition of "region" is *connected* non-empty open subset. I did not want to distract the student with a definition of *connected*, and felt that the student would understand from context that when "an open set in $\mathbf{R}^2$" is referred to in these notes, it is understood that the set is non-empty.

[27] The set $\mathbf{N}$ of natural numbers $\{1, 2, 3, \dots\}$ is an infinite set that is called *countable*, or *countably infinite*. More generally, the empty set and any set that can be indexed by a subset of $\mathbf{N}$ (for example, a collection of three curves $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$, or an infinite collection of curves $\{\mathcal{C}_n\}_{n=1}^{\infty}$) is called *countable*, and we say it has *countably many* elements. Every finite set is countable, so the phrase "finitely many or countably many" is redundant, but the author nonetheless wanted the student to see "finitely many" explicitly in Definition 2.29. Not every infinite set is countable; the set of all real numbers is an uncountable set.

Alternatively, we refer to such a collection $\mathcal{E}$ as *an implicit form of the general solution of* (2.62) *on* R.

If no region R is mentioned explicitly, it is understood that we are taking $R = \mathbf{R}^2$. We may rewrite the definition for this special case more simply:

We call a collection $\mathcal{E}$ of algebraic equations in x and y the *general solution of* (2.62) *in implicit form,* or *an implicit form of the general solution of* (2.62), if

(i)$'$ each equation in the collection $\mathcal{E}$ is an implicit solution of (2.62), and

(ii)$'$ every solution of (2.62) has its graph contained in the union of graphs of finitely many or countably many equations in the collection $\mathcal{E}$. ■

As mentioned earlier, one example of a collection of equations is a one-parameter family of equations $\{F(x, y) = C\}$, where F is a specific function and C is an arbitrary constant. But we do not limit ourselves to such a simple collection of equations in Definition 2.29; there are DEs whose general solutions cannot be given (at least not obviously), even in implicit form, by such a one-parameter family of equations.

Based on earlier definitions, such as Definition 2.7, what you might have expected to see in place of criterion (ii) in Definition 2.29 is the simpler, "Every solution-curve of (2.62) that lies in R, lies in the the graph of some equation in the collection $\mathcal{E}$." However, this would lead to an inadequate definition. The reason for "union of graphs" in (ii) is that sometimes we find ourselves with a collection $\mathcal{E}$ of equations for which some solution-curves of (2.62) are partially contained in the graph of one equation in the collection and partially contained in another (and perhaps partially contained in a third, etc.), without entirely being contained in the graph of any one of our equations. (We will see an example of this later, at the end of Example 2.63.) This can sometimes be fixed by throwing more equations into the collection $\mathcal{E}$. But adding more equations will almost always make the new collection of equations much harder to write down, and still may not handle cases in which the graph of a solution is not contained in any *finite* union of graphs of equations in the original collection $\mathcal{E}$, but only in a *countably infinite* union. (This will also be encountered in Example 2.63.)

A cautionary note: Do not be misled by the terminology "the general solution of (2.62) in R, *in implicit form.*" While there is only one general solution of (2.62) in R—the *collection* of all solutions whose graphs curves in R and that are maximal in R—there are infinitely many *implicit forms* of this general solution. This is the reason for the alternative terminology, "an implicit form of the general solution of (2.62) in R". Sometimes two different implicit forms of the same general solution in R may differ only in "trivial" ways; for example, if one implicit form of the general solution in R is a family of equations $F(x, y) = C$, then another is $2F(x, y) = C$, and another is $F(x, y)^3 = C$. But implicit forms of the same general solution can differ

in much less trivial ways. We saw this even for *ex*plicit ways of expressing general solutions in Examples 2.26 and 2.27.

2.5 Algebraic equivalence of derivative-form DEs

Some algebraic manipulations that, in general, are helpful when we are solving differential equations, have the potential to change the solution-set, either losing some solutions if the original or introducing spurious "solutions" that are not solutions of the original DE.[28] In this section of the notes, we discuss how to be aware of and deal with this problem.

Definition 2.30 We say that two derivative-form differential equations, with independent variable x and dependent variable y, are *algebraically equivalent on a region* R if one equation can be obtained from the other by the operations of (i) adding to both sides of the equation an expression that is defined for all $(x, y) \in R$ [29], and/or (ii) multiplying both sides of the equation by a function of x and y that is defined *and nonzero* at every point of R. When the region R is all of $\mathbf{R}^2$, we will often say simply that the two DEs are *algebraically equivalent*.

Note that subtraction of an expression A is the same as addition of $-A$, and division by a nonzero expression A is the same as multiplication by $\frac{1}{A}$, so subtraction and division are operations allowed in Definition 2.30, even though they are not mentioned explicitly.

Example 2.31 The differential equations

$$\frac{dy}{dx} = y(1-y) \tag{2.63}$$

and

$$\frac{1}{y(1-y)}\,\frac{dy}{dx} = 1 \tag{2.64}$$

are algebraically equivalent on the regions $\{(x, y) \mid y < 0\}$, $\{(x, y) \mid 0 < y < 1\}$, and $\{(x, y) \mid y > 1\}$. However, they are not algebraically equivalent on the whole xy plane. ■

[28] Unfortunately, this is rarely mentioned in textbooks outside the context of "losing constant solutions of separable DEs". In textbooks, it is common for some exercise-answers in the back of the book to be wrong because mistakes of the type discussed here were overlooked by the writer. Even some worked-out examples in some textbooks suffer from this problem.

[29] *Note to students*: The expression is allowed to involve $\frac{dy}{dx}$—i.e. it could be of the form $F(x, y, \frac{dy}{dx})$ for some three-variable function F—which is why we did not say "function of x and y" here. If the expression is $F(x, y, \frac{dy}{dx})$, our requirement that it be "defined for all $(x, y) \in \mathbf{R}$" is short-hand for: for each $(x, y) \in R$ there is *some* real number z such that $F(x, y, z)$ is defined.

Example 2.32 The differential equations

$$(x+y)\frac{dy}{dx} = 4x - 2y \tag{2.65}$$

and

$$\frac{dy}{dx} = \frac{4x-2y}{x+y} \tag{2.66}$$

are algebraically equivalent on the regions $\{(x,y) \mid y > -x\}$ and $\{(x,y) \mid y < -x\}$, but not on the whole xy plane. ■

Why this terminology? Mathematicians call two equations (of any type, not just differential equations) *equivalent* if they have the same set of solutions. For example, the equation $2x + 3 = 11$ is equivalent to the equation $3x = 12$. A general strategy for solving equations is to perform a sequence of operations, each of which takes us from one equation to an equivalent but simpler equation (or to an equivalent set of simpler equations, such as when we pass from "$(x-1)(x-2) = 0$" to "$x - 1 = 0$ or $x - 2 = 0$").

But often, when we manipulate equations in an attempt to find their solution-sets, we perform a manipulation that changes the solution-set.[30] This happens, for example, if we start with the equation $x^3 - 3x^2 = -2x$ and divide by x, obtaining $x^2 - 3x^2 = -2$. In this example, we lose the solution 0. (The solution set of the first equation is $\{0, 1, 2\}$, while the solution set of the second is just $\{1, 2\}$.) For another example, if start with the equation $\sqrt{x+4} = -3$, and square both sides, we obtain $x + 4 = 9$, and hence $x = 5$. But 5 is not a solution of the original equation; $\sqrt{5+4}$ is 3, not -3. Our manipulation has introduced a "spurious solution", a value of x that is a solution of the post-manipulation equation that we may mistakenly *think* is a solution of the original equation, when in fact it is not.

For this reason it is nice to have in our toolbox a large class of equation-manipulation techniques that are guaranteed to be "safe", i.e. not to change the set of solutions. For differential equations, the operations allowed in the definition of "algebraic equivalence" above are safe. The precise statement is:

$$\left.\begin{array}{l}\text{If two differential equations are algebraically equivalent on a}\\ \text{region } R\text{, then they have the same general solution on } R.\end{array}\right\} \tag{2.67}$$

(Here and throughout in these notes, "solution of a DE on (or in) a region R" means a solution whose graph is contained in R.) We may restate (2.67) more briefly as "Algebraically equivalent DEs have the same set of solutions," or "Algebraically equivalent

[30] Usually this is due to carelessness, but there are other times when we do not have much choice. In those cases, we try to keep track separately of any solutions we may have lost or spuriously gained in this step.

DEs are equivalent," sacrificing some precision by omitting reference to the region. But on regions that are not all of $\mathbf{R}^2$, the briefer wording must be interpreted more carefully to mean statement (2.67).

When we perform a sequence of algebraic operations in an attempt to solve a differential equation, especially a nonlinear one, we are rarely lucky enough to end up with a DE that is algebraically equivalent to the original one on the whole xy plane. But usually, we maintain algebraic equivalence on regions that fill out most of the xy plane, as in Examples 2.31 and 2.32 above.

To see why statement (2.67) is true, let us check that operation (ii) in Definition 2.30 does not change the set of solutions whose graphs lie in R. Let us suppose we start with a (first-order) derivative-form DE of the most general possible form:

$$\mathsf{G}_1(x, y, \frac{dy}{dx}) = \mathsf{G}_2(x, y, \frac{dy}{dx}). \tag{2.68}$$

(Of course, by subtracting $\mathsf{G}_2(x, y, \frac{dy}{dx})$ from both sides, we can put this in the simpler form $\mathsf{G}(x, y, \frac{dy}{dx}) = 0$, but since we often perform manipulations on equations without first putting them in the simple form (2.1), we will illustrate the solution-set-doesn't-change principle for DEs that have not been put in that form.) The equation obtained by multiplying both sides of (2.68) by a function h that is defined at every point of R and is nonzero on R is

$$h(x, y)\mathsf{G}_1(x, y, \frac{dy}{dx}) = h(x, y)\mathsf{G}_2(x, y, \frac{dy}{dx}). \tag{2.69}$$

Suppose that ϕ is a solution of (2.68). Then for all x in the domain of ϕ,

$$\mathsf{G}_1(x, \phi(x), \phi'(x)) = \mathsf{G}_2(x, \phi(x), \phi'(x)). \tag{2.70}$$

If the graph of ϕ lies in R, then for all x in the domain of ϕ, the point $(x, \phi(x))$ lies in R, so the number $h(x, \phi(x))$ is defined, and equality is maintained if we multiply both sides of (2.70) by this number. Therefore

$$h(x, \phi(x))\mathsf{G}_1(x, \phi(x), \phi'(x)) = h(x, \phi(x))\mathsf{G}_2(x, \phi(x), \phi'(x)) \tag{2.71}$$

for all x in the domain of ϕ. Hence ϕ is a solution of (2.69). Thus every solution of (2.68) whose graph lies in R is also a solution of (2.69) whose graph lies in R.

Conversely, suppose that ϕ is a solution of (2.69) whose graph lies in R. Then (2.71) is satisfied for all x in the domain of ϕ. By hypothesis, $h(x, y) \neq 0$ for every point $(x, y) \in R$, so for each x in the domain of ϕ, $\frac{1}{h(x,\phi(x))}$ is some number, and equality is maintained if we multiply both sides of (2.71) by this number. Therefore (2.70) is satisfied for all x in the domain of ϕ, so ϕ is a solution of (2.68). Thus every solution of (2.69) whose graph lies in R is also a solution of (2.68) whose graph lies in R.

This completes the argument that multiplying by h has not changed the set of solutions whose graphs lie in R. The argument that operation (i) in Definition 2.30 does not change this set of solutions is similar, and is left to the student.

We mention that it is possible for two differential equations to be equivalent without being *algebraically* equivalent. Performing operations other than those in Definition 2.30 does not *always* change the set of solutions.[31] But because they *might* change the set of solutions, any time we perform one of these "unsafe" operations we must check, by some other method, that we properly account for any lost solutions or spurious solutions.

Students should already be familiar with this fact from their experience with separable equations. For example, in passing from equation (2.63) to (2.64), we potentially lose any solution whose graph intersects the horizontal line $\{y = 0\}$ or the horizontal line $\{y = 1\}$. Are there any such solutions? Yes: the two constant solutions $y \equiv 0$ and $y \equiv 1$, whose graphs happen to be exactly these two horizontal lines.

When we are dealing with separable equations $\frac{dy}{dx} = g(x)p(y)$, and there is any number r for which $p(r) = 0$, when we separate variables we don't just *potentially* lose solutions, we *always* lose solutions (unless we make an error later in the process). For every number r for which $p(r) = 0$, the constant function $y = r$ is a solution that separation of variables, carried out with no errors, *cannot* find[32]. But fortunately, it finds all the others (in implicit form).

We can see why in the context of Example 2.31. The right-hand side of (2.63) is a function of y whose partial derivative with respect to y is continuous everywhere. Therefore for *every* initial-condition point (x_0, y_0) in the xy plane, the fundamental Existence and Uniqueness Theorem for initial-value problems applies, and so through each such point there is the graph of one and only one maximal solution. If there were a non-constant solution of (2.63) whose graph intersected the graph of the constant solution $y \equiv 1$ (the line $\{y = 1\}$), say at the point $(x_0, 1)$, we would have a contradiction to uniqueness of the solution of the IVP with differential equation (2.63) and with initial condition $y(x_0) = 1$. Similarly, no non-constant solution of (2.63) can have a graph that intersects the graph of the constant solution $y \equiv 0$ (the line $\{y = 0\}$). Therefore the graph of every non-constant solution lies entirely in one of the three regions mentioned in Example 2.31. Since equations (2.63) and (2.64) are algebraically equivalent on each of these three regions, the general solution of (2.64) is precisely the set of all solutions of (2.63) other than the two constant solutions that we have already accounted for.

[31] For example, it can be shown that the DEs (2.65) and (2.66) have the same general solution, of which an implicit form is the family of equations $(y + 4x)^3(y - x)^2 = C$.

[32] Unfortunately it is quite common, even in textbooks, to make a pair of canceling errors in solving separable equations, leading to the false impression that the separation-of-variables procedure of variables may lose only *some* of the constant solutions, when in fact it *always* loses *all* of them if no mistakes are made.

Thus, if we manage to solve (2.64)—which we leave the student to do—and then add to its general solution the two constant functions $y \equiv 0$ and $y \equiv 1$, we obtain all solutions of (2.63).

Let us now look at the algebraic-equivalence concept for some linear DEs.

Example 2.33 The equations

$$\frac{dy}{dx} + 3y = \sin x \tag{2.72}$$

and

$$e^{3x}\frac{dy}{dx} + 3e^{3x}y = e^{3x}\sin x \tag{2.73}$$

are algebraically equivalent on the whole xy plane. The second equation can be obtained from the first by multiplying by e^{3x}, which is nowhere zero. Similarly, the first equation can be obtained from the second by multiplying by e^{-3x}, which is nowhere zero. ■

The student familiar with integrating-factors will recognize that the e^{3x} in the example above is an integrating factor for the first equation. To solve linear DEs by the integrating-factor method, the only functions we ever need to multiply by are functions of x alone. Of course, every such function can be viewed as a function of x and y that simply happens not to depend on y. More explicitly, given a function one-variable function μ, we can define a two-variable function $\tilde{\mu}$ by $\tilde{\mu}(x, y) = \mu(x)$. If $\mu(x)$ is nonzero for every x in an interval I, then $\tilde{\mu}(x, y)$ is nonzero at every (x, y) in the region $I \times R$ (an vertical strip, infinite in the $\pm y$-directions). So we will add a bit to Definition 2.30 to have language better suited to linear equations:

Definition 2.34 We say that two linear differential equations, with independent variable x and dependent variable y, are *algebraically equivalent on an interval I* if they are algebraically equivalent on the region $I \times \mathbf{R}$. This happens if and only if one equation can be obtained from the other by the operations of (i) adding to both sides of the equation either a function of x that is defined at every point of the interval I, or y times such function of x, or $\frac{dy}{dx}$ times such a function of x; and/or (ii) multiplying both sides of the equation by a function of x that is defined and nonzero at every point of the interval I.

Example 2.35 The equations

$$x\frac{dy}{dx} - 2y = 0 \tag{2.74}$$

and

$$x^3\frac{dy}{dx} - 2x^2y = 0 \tag{2.75}$$

are algebraically equivalent on the interval $(0, \infty)$, and also on the interval $(-\infty, 0)$, but not on $(-\infty, \infty)$ or on any other interval that includes 0. (Thus, in accordance with Definition 2.30, we do not simply call them "algebraically equivalent"; we specify *an interval on which* they are algebraically equivalent.) The second can be obtained from the first by multiplying by x^2, which satisfies the "nowhere zero" criterion on any interval not containing 0, but violates it on any interval that includes 0.

The first equation can be obtained from the second by multiplying by x^{-2}, which is not zero *anywhere*, but does not yield a function of x on any interval that contains 0. ■

Example 2.36 The equations

$$x\frac{dy}{dx} - 2y = 0 \tag{2.76}$$

(the same equation as (2.74) and

$$x^{-2}\frac{dy}{dx} - 2x^{-3}y = 0 \tag{2.77}$$

are algebraically equivalent on the interval $(0, \infty)$, and also on the interval $(-\infty, 0)$, but not on $(-\infty, \infty)$ or on any other interval that includes 0. In fact, the second equation does not even make sense on any interval that includes 0. The second equation can be obtained from the first by multiplying by x^{-3}, which is not zero *anywhere*, but is not defined at $x = 0$, hence does yield a function that we can multiply by on any interval that includes 0.

The first equation can be obtained from the second by multiplying by x^3, which is defined for all x, but violates the "nowhere zero" condition on any interval that contains 0. ■

In the context of linear DEs, fact (2.67) reduces to the following simpler statement:

$$\left.\begin{array}{l}\text{Two linear DEs that are algebraically equivalent}\\ \text{on an interval } I \text{ have exactly the same solutions on } I.\end{array}\right\} \tag{2.78}$$

Two linear DEs that are not algebraically equivalent on an interval I may or may not have the same set of solutions on I. When we manipulate a linear DE in such a way that we "turn it into" an algebraically inequivalent DE, we run the risk that we will not find the true set of solutions. The next example illustrates this trap.

Example 2.37 Find the general solution of

$$x\frac{dy}{dx} - 2y = 0 \tag{2.79}$$

(the same equation as (2.76) and (2.74)).

Since this is a linear equation, our first step is to "put it in standard linear form" by dividing through by x. This yields the equation

$$\frac{dy}{dx} - \frac{2}{x}\, y = 0. \tag{2.80}$$

However, (2.79) and (2.80) are not algebraically equivalent on the whole real line, but only on $(-\infty, 0)$ and $(0, \infty)$. Equation (2.80) does not even make sense at $x = 0$, while (2.79) makes perfectly good sense there.[33]

As the student may verify, equation (2.80) has an integrating factor $\mu(x) = x^{-2}$. Putting our brains on auto-pilot, we multiply through by x^{-2}, and write

$$\begin{aligned} & (x^{-2}y)' &&= 0, \\ \Longrightarrow \quad & \int (x^{-2}y)'dx &&= \int 0\ dx, \\ \Longrightarrow \quad & x^{-2}y &&= C, \\ \Longrightarrow \quad & y &&= Cx^2. \end{aligned} \tag{2.81}$$

(Even worse than putting our brains on auto-pilot is to ignore warnings to learn the *integrating-factor method* rather than to memorize a formula it leads to for the general solution of a first-order linear DE in "most" circumstances. That formula has its limitations and will also lead, incorrectly, to (2.81).)

Neither in the original DE (2.79) nor in (2.81) do we see any of the clues we are used to seeing, such as a "$\frac{1}{x}$", that warn us that there may be a problem with (2.81) at $x = 0$. (There were clues in the intermediate steps, in which negative powers of x appeared, but we ignored them.) The functions given by (2.81) form a 1-parameter family of functions defined on the whole real line, and it is easy to check that each member of this family is a solution of (2.79). We have been taught that the general solution of a first-order linear DE is a 1-parameter family of solutions—*under certain hypotheses*. (We have ignored the fact that those hypotheses were not met, however.) Having found what we expected to find, we write "$y = Cx^2$" as our final, but wrong, answer.

[33] Standard terminology related to this problem is *singular point*. Roughly speaking, a first-order linear DE does not "behave well" on an interval I if, when the DE is put in standard linear form $\frac{dy}{dx} + p(x)y = g(x)$, there is a point $x_0 \in I$ for which $\lim_{x\to x_0+} |p(x)| = \infty$ or $\lim_{x\to x_0-} |p(x)| = \infty$. Such points x_0 are called *singular points* of the linear DE. The point $x = 0$ is a singular point of both (2.79) and (2.80).

Let us go back to square-one and correct our work. The transition from equation (2.79) to (2.80) involves dividing by x, and therefore is not valid on any interval that contains 0. These two equations are algebraically equivalent on $(0, \infty)$ and on $(-\infty, 0)$, and therefore have the same solutions on these intervals. But the general solution of (2.79) might include solutions on intervals that contain 0, while the general solution of (2.80) cannot.

We can still use the basic procedure that led us to (2.81); we just have to be more careful with it. Auto-pilot will not work.

Because (2.80) makes no sense at $x = 0$, we must solve it separately on $(-\infty, 0)$ and $(0, \infty)$. We can do the work for both of these intervals simultaneously, as long as we keep track of the fact that that's what we're doing.

So suppose ϕ is a differentiable function on *either* on $I = (0, \infty)$ or on $I = (-\infty, 0)$, and let $y = \phi(x)$. On I, x^{-2} is an integrating factor. Multiplying both sides of our equation on I by x^{-2}, we find that ϕ is a solution of (2.80) if and only if $(x^{-2}y)' = 0$. Because I is an interval, $(x^{-2}y)' = 0$ if and only if $x^{-2}y$ is constant. Therefore:

- ϕ is a solution of (2.80) on $(0, \infty)$ if and only if there is a constant C for which $x^{-2}\phi(x) \equiv C$; equivalently, for which ϕ is given by

$$\phi(x) = Cx^2. \tag{2.82}$$

- Exactly the same conclusion holds on the interval $(-\infty, 0)$.

Thus the general solution of (2.80) on $(0, \infty)$ is

$$y = Cx^2, \quad x > 0, \tag{2.83}$$

while the general solution of (2.80) on $(-\infty, 0)$ is

$$y = Cx^2, \quad x < 0. \tag{2.84}$$

Now return to the equation we originally were asked to solve, (2.79), and suppose that ϕ is a solution of this equation on $(-\infty, \infty)$. (The argument we are about to give would work on any interval containing 0.) Let ϕ_1 be the restriction of ϕ to the domain-interval $(0, \infty)$, and let ϕ_2 be the restriction of ϕ to the domain-interval $(-\infty, 0)$. Since (2.79) and (2.80) are algebraically equivalent on $(0, \infty)$, ϕ_1 must be one of the solutions given by (2.83). Thus there is some constant C_1 for which $\phi_1(x) = C_1x^2$. Similarly, ϕ_2 must be one of the solutions given by (2.84), so $\phi_2(x) = C_2x^2$.

Therefore $\phi(x) = C_1x^2$ for $x > 0$, and $\phi(x) = C_2x^2$ for $x < 0$. But we assumed that ϕ was a solution on $(-\infty, \infty)$, so it also has a value at 0. We can deduce this value by using the fact that every solution of an ODE is continuous on its domain (since, by definition, solutions are differentiable functions, and differentiable functions are continuous). Therefore $\phi(0) = \lim_{x\to 0} \phi(x)$. Whether we approach 0 from the left

(using $\phi(x) = C_2x^2$) or the right (using $\phi(x) = C_1x^2$), we get the same limit, namely 0. Hence $\phi(0) = 0$.[34] Since 0 also happens to be the value of C_1x^2 at $x = 0$ (as well as the value of C_2x^2 at $x = 0$), we can write down a formula for ϕ in several equivalent ways, one of which is

$$\phi(x) = \begin{cases} C_1x^2 & \text{if } x \geq 0, \\ C_2x^2 & \text{if } x < 0, \end{cases} \tag{2.85}$$

(We could have chosen to absorb the "$x = 0$" case into the second line instead of the first, or to use both "≥ 0" in the top line and "≤ 0" in the bottom line, since that would not lead to any inconsistency. Or we could have chosen to write a three-line formula, with one line for $x > 0$, one line for $x = 0$, and one line for $x < 0$. All of these ways are equally valid; we just chose one of them.)

Conversely, as the student may check, every function of the form (2.85) is a solution of (2.79). Therefore the general solution solution-set of (2.79) on $(-\infty, \infty)$ is the *two-parameter* family of functions given by (2.85), with C_1 and C_2 arbitrary constants[35]. This collection of solutions contains all the solutions on every other interval, in the sense that the general solution on any interval I is obtained by restricting the functions (2.85) to the interval I. (For the student who read and understood the material on maximal solutions: the two-parameter family (2.85) is the general solution of (2.79) as defined in Definition 2.22.) ■

We do not want the previous example to give the student the wrong impression. For the vast majority, if not 100%, of n^{th}-order linear DEs you are likely to encounter in your first course on DEs, you will be shown how to solve them (or asked to solve them) only on intervals for which the general solution solution-set is an n-parameter family of functions. You are unlikely to see a two-parameter family of functions as

[34] Another way to find the value of $\phi(0)$ in this example is as follows. Since ϕ is differentiable on its domain, the whole real line, $\phi'(0)$ is *some* real number. Whatever this value is, when we plug $x = 0$ and $y = \phi(x)$ into (2.79), the term "$x\frac{dy}{dx}$" becomes $0 \times \phi'(0)$, which is 0. Hence $\phi(0) = y(0) = 0$.

While this second method works for (2.79), it does not work for (2.75)—which the student will later be asked to solve—but the first method we presented does.

[35] Some authors, with a different definition of "general solution", would say that the first-order linear equation (2.79) *does not have* a general solution on $(-\infty, \infty)$, because the set of all solutions on $(-\infty, \infty)$ is a two-parameter family rather than a one-parameter family. To the author of these notes, this seems an odd convention to apply to a solution-set with a completely systematic description.

Note to instructors: The solution-set of *any* homogeneous linear DE on *any* interval is a vector space. We already show this to our students, in different language (0 is a solution, and any linear combination of solutions is a solution). It does not make sense to me to say that the DE *does not have* a general solution if the dimension of this vector space happens not to be the same as the order of the DE. It makes far more sense to me to define the general solution on an interval to be the set of all solutions on that interval (especially for a linear DE), and simply teach, as we already do—usually without the vector-space terminology—that for a standard-form linear n^{th}-order homogeneous DE on an interval on which all of the coefficients are continuous, the general solution is a vector space *of dimension n*.

the general solution of a DE unless the equation is second-order. Example 2.37 is the exception, not the rule. But we wanted the student to see another example of the perils of what can happen when algebraic equivalence is not maintained during the manipulation of equations.

As mentioned earlier, algebraically inequivalent linear DEs do not *always* have different solution-sets. The student should test his/her understanding of the example above by showing that equations (2.74) and (2.75) have the same set of solutions.

2.6 First-order equations in differential form

Definition 2.38 A *differential* in the variables (x, y) is an expression of the form

$$M(x, y)dx + N(x, y)dy \tag{2.86}$$

where M and N are functions defined on some region in $\mathbf{R}^2$. We often abbreviate (2.86) as just

$$Mdx + Ndy, \tag{2.87}$$

leaving it understood that M and N are functions of x and y. When a region R is specified, we call $Mdx + Ndy$ a *differential on* R.

The functions M, N in (2.86) and (2.87) are called the *coefficients* of dx and dy in these expressions. ■

The following definition provides an important source of examples of differentials.

Definition 2.39 (a) If F is a continuously differentiable function on a region R (i.e. if both first partial derivatives of F are continuous on R), and the variables we use for $\mathbf{R}^2$ are x and y, then the *differential of* F *on* R is the differential dF defined by

$$dF = \frac{\partial G}{\partial x}\,dx + \frac{\partial G}{\partial y}\,dy. \tag{2.88}$$

(b) A differential $Mdx + Ndy$ on a region R is called *exact on* R if there is some continuously differentiable function F on R for which $Mdx + Ndy = dF$ on R. ■

Note that the "continuously differentiable" requirement in part (b) implies that the coefficient functions M, N in any exact differential are continuous.

Note that we have not yet ascribed *meaning* to "dx" or "dy"; effectively, they are just place-holders for the functions M and N in (2.86) and (2.87). Similarly, so far the expression "$Mdx + Ndy$" is just *notation*; its information-content is just the

pair of functions M, N (plus the knowledge of which function is the coefficient of dx and which is the coefficient of dy).

You (the student) may have come across the noun "differential" in your previous calculus courses. The sense in which we use this noun in these notes is more sophisticated than the notion used in Calculus 1-2-3. (For interested students, Section 3.1 discusses what a differential actually *is*, in the sense used in these notes.) There is a relation between the two notions, but it is beyond the scope of these notes to state exactly what that relation is.

If $Mdx + Ndy$ is a differential on a region R, and (x_0, y_0) is a point in R, we call the expression $M(x_0, y_0)dx + N(x_0, y_0)dy$ the *value* of the differential $Mdx + Ndy$ at (x_0, y_0). However, this "value" is not a real number; so far it is only a piece of notation of the form "(real number times dx) + (real number times dy)", and we still have attached no meaning to "dx" and "dy". The value of a differential at a point is actually a certain type of *vector*, but not the type you learned about in Calculus 3. (The type of vector that it *is* will not be described in these notes; the necessary concepts require a great deal of mathematical sophistication to appreciate, and are usually not introduced at the undergraduate level.[36])

We next define rules for algebraic operations involving differentials. These definitions are necessary, rather than being "obvious facts", because so far differentials are just pieces of notation to which we have attached no meaning. **However, in an introductory course on DEs, it is generally permissible for students to treat the rules in Definition 2.40 as "obvious facts".** If you have trouble understanding why Definition 2.40 is necessary, don't worry about it; just make sure that the way you manipulate differentials agrees with these rules.

Definition 2.40 Let R be an open set in $\mathbf{R}^2$, let x, y be the usual coordinate-functions on $\mathbf{R}^2$, and let M, N, M_1, M_2, N_1, N_2, and f be functions defined on R. (Thus $Mdx + Ndy$, $M_1dx + N_1dy$, and $M_2dx + N_2dy$ are differentials on R.) Then we make the following definitions for differentials in (x, y).

1. Equality of differentials: $M_1dx + N_1dy = M_2dx + N_2dy$ on R if and only if $M_1(x, y) = M_2(x, y)$ and $N_1(x, y) = N_2(x, y)$ for all $(x, y) \in R$.

2. Abbreviation by omitting terms with coefficient zero:

$$\begin{aligned} Mdx &= Mdx + 0dy, \\ Ndy &= 0dx + Ndy. \end{aligned}$$

[36] However, for students who have taken enough linear algebra to know what the *dual of a vector space* is, the value of a differential at a point can be treated as an element of the dual space of $\mathbf{R}^2$. *Note to instructors*: More precisely, a differential at a point is a *covector* or *cotangent vector*, an element of the cotangent space of $\mathbf{R}^2$ at that point.

3. Abbreviation by omitting the coefficient 1 (the constant function whose constant value is the real number 1):

$$\begin{aligned} dx &= 1dx, \\ dy &= 1dy. \end{aligned}$$

4. Insensitivity to which term is written first:

$$Ndy + Mdx = Mdx + Ndy.$$

5. Addition of differentials:

$$(M_1dx + N_1dy) + (M_2dx + N_2dy) = (M_1 + M_2)dx + (N_1 + N_2)dy.$$

6. Subtraction of differentials:

$$(M_1dx + N_1dy) - (M_2dx + N_2dy) = (M_1 - M_2)dx + (N_1 - N_2)dy.$$

7. Multiplication of a differential by a function of (x, y):

$$f(Mdx + Ndy) = fMdx + fNdy.$$

   (Here, the left-hand side is read "f times $Mdx+Ndy$", not "f of $Mdx+Ndy$". The latter would make no sense, since f is a function of two real variables, not a function of a differential.)

8. The *zero differential* on R is the differential $0dx+0dy$, which we often abbreviate just as "0". (We tell from context whether the symbol "0" is being used to denote the *real number* zero, the *constant function* whose value at every point is the real number zero, or the zero differential. In the equation "$0dx + 0dy = 0$", context tells us that each zero on the left-hand side of the equation is to be interpreted as *the constant function with constant value* 0, while the zero on the right-hand side is to be interpreted as the zero differential[37]. ■

[37] As a general rule, it's a bad idea to use the same symbol to represent different objects, and it's *usually* a particularly awful idea to let the same symbol have two different meanings in the same equation. We allow certain—very few—exceptions to this rule, in order to avoid cumbersome notation, such as having three different symbols such "$0_{\mathbf{R}}$", "0_{fcn}," and "0_{diff}," fot the zero number, zero function, and zero differential respectively.

Note that our definition of subtraction is the same as what we would get by combining the operations "addition" and "multiplication by the constant function -1":

$$(M_1dx + N_1dy) - (M_2dx + N_2dy) = (M_1dx + N_1dy) + (-1)(M_2dx + N_2dy).$$

Note also that *we do not define the product or quotient of two differentials.* In particular we don't (yet) attempt to relate the differentials dx and dy to a derivative $\frac{dy}{dx}$. (When we do relate them later, $\frac{dy}{dx}$ still will not be the quotient of two differentials.)

Finally, we are ready to bring differential equations back into the picture!

Definition 2.41 A *differential equation in differential form, with variables* (x, y), is an equation of the form

$$\text{one differential in } (x, y) = \text{another differential in } (x, y). \tag{2.89}$$

We write such an equation only when where there is some region R on which both differentials are defined. When the region R is specified, we use phrasing like "a DE on R in differential form" or "a DE in differential form on R."

Example 2.42 Whenever we separate variables in a separable, derivative-form DE, we go through a step in which we write down a differential-form DE, such as

$$ydy = e^x dx. \tag{2.90}$$

■

A **very important difference** between a DE in derivative form and a DE in differential form is that **a DE in differential form has no "independent variable" or "dependent variable".** The two variables are on an equal footing. We do have a "first variable" and "second variable" (for which we are using the letters x and y, respectively, in these notes), but *only* because we need to put names to our first and second variables in order to specify the functions M and N (e.g. to write a formula such as "$M(x, y) = x^2y^3$"). *Do not* make the mistake of thinking that whenever you see "x" and "y" in a DE, x is automatically the independent variable and y the dependent variable. Also, even when it's been decided that the letters x and y will be used, there is no law that says x has to be the first variable and y the second. In these notes we *choose* the conventional order so that the student will feel on more familiar ground. But notice that if we were to choose different names for our variables, and for the sake of being ornery write something like

$$\aleph \, d\aleph = e^{\mathfrak{a}} d\mathfrak{a},$$

you would not have a clue as to which variable to call the first—*nor would it matter which choice you made.*

Here is the differential-form analog of Definition 2.30:

Definition 2.43 We say that two DEs in differential form, with variables (x, y), are *algebraically equivalent on a region* R if one can be obtained from the other by the operations of (i) addition of differentials and/or (ii) multiplication by a function of (x, y) that is defined at every point of R and is nowhere zero on R. ■

So, for example, each of the differential-form ODEs

$$2x^2 y dx = \tan(x + y) dy,$$

$$2x^2 y dx - \tan(x + y) dy = 0,$$

and

$$e^x (2x^2 y dx - \tan(x + y) dy) = 0,$$

is algebraically equivalent to the other two on $\mathbf{R}^2$ (and on any region in $\mathbf{R}^2$). On the open set $\{(x, y) \mid x \neq 0\}$ these equations are also algebraically equivalent to

$$x(2x^2 y dx - \tan(x + y) dy) = 0, \tag{2.91}$$

but are *not* algebraically equivalent to (2.91) on the whole plane $\mathbf{R}^2$, since the plane contains points at which $x = 0$.

Note that by subtracting the differential on the right-hand side of (2.89) from both sides of the equation, we obtain an algebraically equivalent equation of the form

$$M dx + N dy = 0.$$

Later, after we have defined "solution of a DE in differential form", we will see that algebraically equivalent equations have the same solutions. Therefore we lose no generality, in our discussion of solutions of DEs in differential form, if we restrict attention to equations of the form (2.93). (However, there is one instance in which it is convenient to consider differential-form DEs that have a nonzero term on each side: the case of separated variables, of which (2.90) is an example.)

In our discussion of derivative-form DEs, we defined, and frequently used, the notion of *solution curve*. Soon we will define *solution curve* for differential-form DEs. This notion is even more important for differential-form DEs than it is for derivative-form DEs. But defore defining *solution curve* of a differential-form DE, we need to discuss the basics of curves in general.

2.6.1 Curves, parametrized curves, and smooth curves

In Calculus 2 and 3 you learned about *parametrized curves* (not necessarily by that name, however). We review the concept and some familiar terminology, and introduce what may be some unfamiliar terminology.

Definition 2.44 A *parametrized curve* or *curve-parametrization* in $\mathbf{R}^2$ is an ordered pair of continuous real-valued functions (f, g) defined on an interval[38]. The set

$$\{(f(t), g(t)) \mid t \in I\} \tag{2.92}$$

(where I is an interval) is called the *range*, *trace*, or *image* of the parametrized curve.

A *curve* in $\mathbf{R}^2$ is a point-set $\mathcal{C} \subset \mathbf{R}^2$ that is the image of some parametrized curve[39].

Given a curve $\mathcal{C}$, if (f, g) is a parametrized curve with image $\mathcal{C}$, then we say that (f, g) is a *parametrization of* $\mathcal{C}$ or that (f, g) *parametrizes* $\mathcal{C}$. ■

In other words, a curve $\mathcal{C}$ is a point-set that is "traced out" by the parametric equations

$$\begin{aligned} x &= f(t), \\ y &= g(t), \end{aligned}$$

as t ranges over a parameter-interval; hence the terminology "trace". Unfortunately, the word "trace" has several different meanings in mathematics, each of them completely unrelated to the others. The next course in students encounter this word it is likely to mean something totally different, so it will not be our preferred term in these notes. The word *range* is often used by teachers because the student is familiar with it from precalculus and Calculus 1. The concept is the same here: the range of (f, g), thought of as a single $\mathbf{R}^2$-valued function γ (defined by $\gamma(t) = (f(t), g(t))$) rather than as a pair of real-valued functions. A synonym for *range* is *image*, which is the term we will use in these notes. For vector-valued functions (and other functions more exotic than real-valued functions), mathematicians generally prefer "image" to "range" because it is more geometrically suggestive.

Note that we are now using the letter I for a *parameter-interval* ("t-interval"), not an x-interval.

Most of the time it is simpler to write "$(x(t), y(t))$" than to introduce the extra letters f, g and write "$(f(t), g(t))$" for the point in the xy plane defined by "$x =$

[38] In these notes, intervals are required to have more than one point; we never mean "degenerate intervals" $[a, a]$.

[39] The "$\mathcal{C}$" used in these notes for a curve is in a different font from the C that we use for a constant.

$f(t), y = g(t)$". We will often use the simpler notation $(x(t), y(t))$ when there is no danger of misinterpretation. Thus we we also sometimes write "$\gamma(t) = (x(t), y(t))$". When we do not want to introduce a name (e.g. γ) for such an $\mathbf{R}^2$-valued function, we will say "the parametrized curve (or curve-parametrization) $t \mapsto (x(t), y(t))$." (Read the symbol "$\mapsto$" as "goes to".)

Note that in Definition 2.44, we do not require the interval I to be open. This is so that we can present certain examples below simply, without bringing in too many concepts at once that may be new to the student. Eventually, we will want to consider only parametrized curves that have an open domain-interval, but we will not impose that requirement just yet.

Example 2.45 Let $x(t) = 2\cos t, y(t) = 2\sin t, t \in [0, 2\pi]$. Then for all t we have $x(t)^2+y(t)^2 = 4$, so the trace of this parametrized curve lies along the circle $x^2+y^2 = 4$. It is not hard to see that every point on the circle is in the image of this parametrized curve, so the *curve* traced out by the parametrized curve $t \mapsto (x(t), y(t)), t \in [0, 2\pi]$, is the whole circle $x^2 + y^2 = 4$. Had we used the same formulas for $x(t)$ and $y(t)$, but restricted t to the interval $[0, \pi]$, the range would still have lain along the circle $x^2 + y^2 = 4$, but would have been only a semicircle. Had we used the same formulas, but used a slightly larger, open interval, say $(-0.1, 2\pi + 0.1)$, then we would have obtained the whole circle again, with some small arcs traced-out twice. ■

Every curve has infinitely many parametrizations. For example, "$x(t) = 2\cos 7t$, $y(t) = 2\sin 7t, t \in [0, 2\pi/7]$" traces out the same curve as in first part of the example above. So does "$x(t) = 2\cos t^3, y(t) = 2\sin t^3, t \in [-\pi^{1/3}, \pi^{1/3}]$".

Definition 2.46 A curve-parametrization $(x(t), y(t)), t \in I$ is called

- *differentiable* if the derivatives $x'(t)$, $y'(t)$ exist[40] for all $t \in I$;
- *continuously differentiable* if it is differentiable and $x'(t)$, $y'(t)$ are continuous in t; and
- *non-stop* if it is differentiable and $x'(t)$ and $y'(t)$ are never simultaneously zero (i.e. there is no t_0 for which $x'(t_0) = 0 = y'(t_0)$).

■

[40] When I contains an endpoint (i.e. I is of the form $[a, b)$, $[a, b]$, or $(a, b]$, the first two of which contain their left endpoints and the last two of which contain their right endpoints), then *derivative* at an endpoint that I contains is interpreted as the appropriate *one-sided* derivative. Thus, if I contains a left endpoint a, then what we mean by "$x'(a)$", or "$\frac{dx}{dt}$ at a", is $\lim_{t\to a^+} \frac{x(t)-x(a)}{t-a}$. Similarly if I contains a right endpoint b, then what we mean by "$x'(b)$", or "$\frac{dx}{dt}$ at b", is $\lim_{t\to b^-} \frac{x(t)-x(b)}{t-b}$.

Definition 2.47 A curve $\mathcal{C}$ in $\mathbf{R}^2$ is *smooth* if for every point (x_0, y_0) on the curve, there is a number $\epsilon_0 > 0$ such that for all positive $\epsilon < \epsilon_0$, the portion of $\mathcal{C}$ lying inside the open square of side-length ϵ centered at (x_0, y_0) admits a continuously differentiable, nonstop parametrization, with domain an open interval. ■

"Admits", as used in Definition 2.47, is essentially another word for "has". We use the word "admits" because "has" might mislead the student into thinking that the curve has already been dropped on his/her plate with a continuously differentiable, nonstop parametrization; "*admits* a continuously differentiable, nonstop parametrization" does not lend itself to this misinterpretation.

The open-interval requirement at the end of Definition 2.47 implies that if a curve contains an endpoint, then the curve does not meet our definition of "smooth curve". This is necessary in order to make various other definitions and theorems reasonably short; curves with endpoints are messier to handle.

The student should re-read the end of Example 2.45 to convince him/herself that a circle meets our definition of "smooth curve".

Observe that Definition 2.47 uses a "windowing" idea similar to the one that we used to talk about implicitly-defined functions in Section 2.2. We will later give an equivalent definition of "smooth curve" that is even more reminiscent of that earlier discussion.

Every curve admits parametrizations that are not continuously differentiable and/or are not non-stop. Every *smooth* curve admits continuously differentiable parametrizations that do not meet the "non-stop" criterion, as well as those that *do* meet this criterion. But curves with corners, such as the graph of $y = |x|$, admit *no* continuously differentiable, nonstop parametrizations. We can parametrize the graph of $y = |x|$ continuously differentiably—for example, by $t \mapsto (t^3, |t|^3)$, with parameter-interval $(-\infty, \infty)$—but observe that for this parametrization, $x'(0) = 0 = y'(0)$, so the parametrization is not non-stop. The corner forces us to stop in order to instantaneously change direction.

The graph of $y = |x|$ is one example of a non-smooth curve. Other examples of non-smooth curves are:

- The letter X. You can draw this without your pencil leaving the paper, so it satisfies the definition of "curve". (When you draw a curve $\mathcal{C}$, you are parametrizing $\mathcal{C}$ using time as the parameter. The condition "without your pencil leaving the page" corresponds to the domain of the parametrization being an interval. Nothing in the definition of "parametrized curve" prohibits you from stopping, reversing direction, and retracing parts of the curve that you've already drawn). But you you need to violate the "non-stop" criterion in order to draw the X.

- A figure-8. The whole curve does admit a continuously differentiable, non-stop parametrization, but the point (x_0, y_0) at which the curve crosses itself causes

the definition of "smooth" not to be met. For small ϵ, the portion of the curve that lies in the square of side ϵ centered at (x_0, y_0) is essentially an X, and has the same problem that the X did.

Warning about terminology. Many calculus textbooks refer to a continously differentiable, non-stop parametrization as a *smooth* parametrization. This usage of "smooth" is unfortunate. It conflicts with the modern meaning of "smooth function" in advanced mathematics[41]. A preferable one-word term is "regular", and the only reason we are not using it in these notes is that the meaning of "regular" is not self-evident, and we did not want to present the student with extra terminology to remember. "Regular" is flexible term that mathematicians use with a contextually varying meaning, which usually is "having the most common features" or "having no nasty or inconvenient features" (where the context determines what features are important). The meaning of *non-stop* is self-evident (regarding $\gamma'(t) = (x'(t), y'(t))$ as the velocity vector $\mathbf{v}(t)$ at time t associated with the parametrization, "non-stop" is the condition that the velocity $\mathbf{v}(t)$ is not the zero vector for any t), but the author of these notes has never seen it in any textbook[42].

We make one more definition before moving on to the next section.

Definition 2.48 A smooth curve $\mathcal{C}$ lying in a region R in $\mathbf{R}^2$ is *inextendible in* R if either

1. $\mathcal{C}$ is a closed curve (i.e. $\mathcal{C}$ has a continuously differentiable, non-stop parametrization γ, with domain a closed interval $[a, b]$, for which $\gamma(a) = \gamma(b)$), or

2. $\mathcal{C}$ is an "open curve without endpoints" (i.e. $\mathcal{C}$ has a continuously differentiable, non-stop parametrization with domain an open interval), and there is no non-stop, continuously differentiable, parametrized curve γ whose image lies in R

[41] *Note to instructors:* in differential topology and differential geometry, "smooth parametrization" simply means "C^k map" (from an open interval to $\mathbf{R}^2$, in the setting of these notes) for some pre-specified k, usually 1 or ∞. There is no requirement that the parametrization be non-stop to be called smooth. Even *constant* maps, whose images are a single point, are considered smooth parametrized curves—and it is indispensable to the definition of "tangent space" to include these when one talks about the collection of all smooth parametrized curves passing through a given point.

[42] *Note to instructors*: in differential topology and geometry, what we are calling here a (continuously differentiable) non-stop parametrization is called an *immersion*, so one would never see "non-stop" in a research paper. Introductory courses and textbooks would be the only places to use this term. When teaching about curves in Calculus 3, I use "non-stop" as a separate condition, rather than part of the definition of"smooth parametrization", because (i) it is pedagogically useful, (ii) it is more self-explanatory than the calculus-textbook definition of "smooth parametrization", which has the awkward feature that (with this bad definition) all smooth curves admit non-smooth parametrizations, (iii) the calculus-textbook definition of "smooth parametrization" conflicts with the definition used by mathematicians who specialize in studying smooth topological or geometric objects, and (iv) the term "non-stop" presents no such conflict.

and contains $\mathcal{C}$ as a proper subset.[43] ■

A smooth curve that "runs off to infinity in both directions", like either branch of the hyperbola $xy = 1$, is inextendible in any region that contains it. For a smooth curve that is *not* closed, and does not "run off to infinity in both directions", "inextendible" essentially means that we cannot add points at either end of the curve without leaving the region R. For example, the portion of the graph of $y = x$ that lies in the region R between the lines $y = 1$ and $y = -1$ is inextendible in R. The portion of the same graph that lies in the open first quadrant R_1 is inextendible in R_1.

2.6.2 Solution curves for DEs in differential form

Now we get to the heart of the difference between DEs in derivative form and those in differential form: unlike a DE in derivative form, a DE in differential form is not an equation that is looking for a *function*. It is an equation that is looking for a *curve*.

Definition 2.49 A *solution curve* [44] of a differential equation

$$M(x, y)dx + N(x, y)dy = 0 \tag{2.93}$$

on a region R is a smooth curve $\mathcal{C}$, contained in R, for which some continuously differentiable, non-stop parametrization $\gamma(t) = (x(t), y(t))$ of $\mathcal{C}$ satisfies

$$M(x(t), y(t))\frac{dx}{dt} + N(x(t), y(t))\frac{dy}{dt} = 0 \tag{2.94}$$

for all t in the domain-interval I of the parametrization. In this context, we will call γ a *parametric solution* of (2.93).[45]

When no region R is specified, it is understood that the region of interest is the interior of the common implied domain of M and N. Here, "common implied domain" means the set of points at which both M and N are defined, and "interior" means that we don't count points that are on the boundary of the common domain[46]. ■

[43] The condition that $\mathcal{C}$ is an "open curve without endpoints" turns out to be redundant in this part of the definition, but is included here as a visual aid.

[44] It would be more logical to use the term *solution* for what we are calling *solution curve*. However, this would conflict with the meaning of "solution of a DE in differential form" that students are likely to see used in a textbook. Even if not stated explicitly, the meaning in a textbook is likely to be close to Definition 2.55 later in these notes.

[45] The terminology "solution curve" and "parametric solution" for a DE in differential form were invented for these notes; they are not standard.

[46] *Note to instructor:* The author has avoided giving a careful definition of "boundary" here, and therefore of "interior", to avoid distracting the student.

Note that we have not yet defined "*solution* of a DE in differential form"; we have defined only *solution curves* and *parametric solutions*. The definition of *solution* for such DEs is deferred to Section 2.6.4.

As we noted previously, in a differential-form DE (2.93) there is neither an independent nor a dependent variable; x and y are treated symmetrically. This symmetry is preserved in (2.94), but in a surprising way: in (2.94), *both* x and y are dependent variables! The independent variable is t—a variable that is not even visible in (2.93).

Definition 2.49 implies more about solution curves and parametric solutions than is obvious just from reading the definition.

To start with, equation (2.94) has a geometric interpretation. Let $(x(t), y(t))$ be a continuously differentiable, non-stop parametrization of some solution curve $\mathcal{C}$ of $M\,dx + N\,dy = 0$. Let $\mathbf{v}(t) = \gamma'(t) = x'(t)\mathbf{i} + y'(t)\mathbf{j}$, where $\mathbf{i}$ and $\mathbf{j}$ are the standard basis vectors in the xy plane. Then $\mathbf{v}(t)$, the velocity-vector function associated with the parametrization, is tangent to the smooth curve $\mathcal{C}$ at the point $(x(t), y(t))$. We can rewrite equation (2.94) using the dot-product you learned in Calculus 3:

$$(M(x(t), y(t))\mathbf{i} + N(x(t), y(t))\mathbf{j}) \cdot \mathbf{v}(t) = 0. \tag{2.95}$$

This says that, for each t, the vector $\mathbf{v}(t)$ is perpendicular to the vector $M(x(t), y(t))\mathbf{i} + N(x(t), y(t))\mathbf{j}$. Thus for each point (x_0, y_0) on $\mathcal{C}$, the velocity vector at that point (i.e. $\mathbf{v}(t_0)$, where $(x(t_0), y(t_0)) = (x_0, y_0)$) is perpendicular to $M(x_0, y_0)\mathbf{i} + N(x_0, y_0)\mathbf{j}$.

Suppose we have another non-stop parametrization of the same curve $\mathcal{C}$. To speak clearly of both parametrizations, we must temporarily abandon the notation "$(x(t), y(t))$" in favor of $(f_1(t), g_1(t))$ $(t \in I_1)$ and $(f_2(t), g_2(t))$ $(t \in I_2)$. At a given point (x_0, y_0), the velocity vectors $\mathbf{v}_1, \mathbf{v}_2$ coming from the two parametrizations will be parallel, both being nonzero vectors tangent to $\mathcal{C}$ at that point. (I.e. if t_1, t_2 are such that $(f_1(t_1), g_1(t_1)) = (x_0, y_0) = (f_2(t_2), g_2(t_2))$, then $\mathbf{v}_2(t_2) = c\mathbf{v}_1(t_1)$ for some nonzero scalar c.) But then

$$\begin{aligned}
(M(x_0, y_0)\mathbf{i} + N(x_0, y_0)\mathbf{j}) \cdot \mathbf{v}_2(t_2) &= (M(x_0, y_0)\mathbf{i} + N(x_0, y_0)\mathbf{j}) \cdot c\mathbf{v}_1(t_1) \\
&= c\,(M(x_0, y_0)\mathbf{i} + N(x_0, y_0)\mathbf{j}) \cdot \mathbf{v}_1(t_1) \\
&= c\ 0 \\
&= 0.
\end{aligned}$$

Since this holds for all points (x_0, y_0) on $\mathcal{C}$, it follows that the parametrization $x = f_2(t), y = g_2(t)$ also satisfies (2.94).[47] Thus if one continuously differentiable, non-stop parametrization of $\mathcal{C}$ satisfies (2.94), so does every other. Therefore, even though Definition 2.49 requires only that there be *some* continuously differentiable, non-stop parametrization of $\mathcal{C}$ satisfying (2.94), once we know that even *one* continuously

[47] This can also be shown using the Inverse Function Theorem that you may have learned in Calculus 1, plus the Chain Rule.

differentiable, non-stop parametrization of $\mathcal{C}$ has this property, they all do. Said another way:

$$\left.\begin{array}{l}\textit{Every} \text{ continuously differentiable, non-stop parametrization of a} \\ \text{solution curve of a differential equation } Mdx + Ndy = 0 \text{ is a} \\ \text{parametric solution of this equation.}\end{array}\right\} \tag{2.96}$$

This gets back to the statement we made just prior to Definition 2.49: that a DE in differential form is looking for a curve. We did not say "*parametrized* curve". A curve is a geometric object, a certain type of point-set in the plane. The concept of *parametrized curve* is needed to define which point-sets are curves and which aren't. It's also needed to define many other features or properties of a curve, such as whether a curve is a solution curve of a (given) DE in differential form. Any property that is defined via parametrizations (such as being a solution curve of a DE in differential form) can potentially hold true for one parametrization but not for another. A property defined in terms of parametrizations is intrinsic to a (smooth) *curve* $\mathcal{C}$ — the point-set traced out by any parametrization— if and only if the property holds true for *all* continuously differentiable, non-stop parametrizations of $\mathcal{C}$. These are the properties that are truly *geometric*. What statement (2.96) is saying is that the property "I am a solution curve of this differential-form DE" is an intrinsic, geometric property.

Although the concepts of "solution of a DE in derivative form" and "solution curve of a DE in differential form" are fundamentally different—the former is a function (of one variable); the latter is a *geometric object*, a smooth curve—they are still related to each other. We will see precisely what the relation is in a later section of these notes. For now, we mention just that a solution curve of any derivative-form DE in derivative form is a solution curve for a related differential-form DE. (We will see make this precise in Section 2.8.) The converse is not true, because not every smooth curve in $\mathbf{R}^2$ is the graph of a function of one variable (consider a circle).

Many smooth curves in $\mathbf{R}^2$ that are not graphs of one-variable functions can still be expressed entirely or "mostly" as a union of graphs of equations of the form "y = differentiable function of x." But for many smooth curves, including those that arise as solution curves of differential equations in differential form, this is often neither necessary nor desirable[48]. This is another fundamental difference between derivative-form DEs and differential-form DEs.

[48]We emphasize that this "neither necessary nor desirable" applies *only* to DEs that *from the start* are written in differential form, such as in orthogonal-trajectories problems. When differential-form equations are used as a tool to solve derivative-form equations, say with dependent variable y and independent variable x, then it usually *is* desirable to write solutions in the explicit form "y = differentiable function of x"—and your instructor may em require you to do this whenever it is algebraically possible.

Example 2.50 Consider the equation

$$xdx + ydy = 0. \tag{2.97}$$

Suppose we are interested in a solution curve of this DE that passes through the point $(0, 5)$. As the student may check, the parametrized curve

$$\begin{aligned} x(t) &= 5\cos t, \\ y(t) &= 5\sin t, \end{aligned}$$

$t \in [0, 2\pi]$, is a parametric solution. The solution curve it parametrizes is the circle with equation $x^2 + y^2 = 25$. The circle is not the graph of a function of x, but it is a beautiful smooth curve, and as far as the DE (2.97) is concerned, there is no reason to exclude any point of it.

But we run into trouble if we try to express this curve using graphs of differentiable functions of x alone. The circle can be expressed "mostly" as the union of the graphs of $y = \sqrt{25 - x^2}, -5 < x < 5$, and $y = -\sqrt{25 - x^2}, -5 < x < 5$. (The endpoints of the x-interval $[-5, 5]$ must be excluded since $\frac{d}{dx}\sqrt{25 - x^2}$ does not exist at $x = \pm 5$.) But we cannot get the whole circle. ■

2.6.3 Existence/uniqueness theorem for DEs in differential form

Recall that an initial-value problem, with dependent variable y and independent variable x, consists of a derivative-form differential equation together with an initial condition of the form $y(x_0) = y_0$. The differential-form analog of an initial-value problem is a differential-form DE together with a point (x_0, y_0) of the xy plane. The analog of "solution of an initial value problem" is a solution curve of a differential-form DE that passes through the given point (x_0, y_0). In such a context we may (loosely) refer to the point (x_0, y_0) as an "initial condition" or "initial-condition point", and to the combination "differential-form DE, together with point (x_0, y_0)" as an "initial-value problem in differential form". But because there is neither an independent variable nor a dependent variable in a differential-form DE, this terminology is not as well-motivated as it is for derivative-form DEs, where the terminology stems from thinking of the independent variable as *time*.

Just as for derivative-form IVPs, there is an Existence and Uniqueness Theorem for differential-form IVPs, which we will state shortly. To understand what's behind a restriction that will appear in the statement of this theorem, let us look again at equation (2.95). Suppose (x_0, y_0) lies on a smooth solution curve $\mathcal{C}$ of $Mdx + Ndy = 0$. If $M(x_0, y_0)$ and $N(x_0, y_0)$ are not both zero, then $\mathbf{w} = M(x_0, y_0)\mathbf{i} + N(x_0, y_0)\mathbf{j}$ is a nonzero vector, and (2.95) tells us that the velocity vector at (x_0, y_0) of any continuously differentiable, non-stop parametrization of $\mathcal{C}$ must be perpendicular to

$\mathbf{w}$. *Hence* $\mathbf{w}$ *completely determines the slope of the line tangent to* $\mathcal{C}$ *at* (x_0, y_0). This places a very strong restriction on possible solution curves through (x_0, y_0): there is one and only one possible value for the slope of their tangent lines.

But if $M(x_0, y_0)$ and $N(x_0, y_0)$ *are* both zero, then $M(x_0, y_0)\mathbf{i} + N(x_0, y_0)\mathbf{j}$ is the zero vector, and *every* vector is perpendicular to it. Said another way, if $(x(t), y(t))$ is a parametrization of *any* smooth curve passing through (x_0, y_0), say when $t = t_0$, then (2.95) is satisfied at $t = t_0$, and so is (2.94). There is no restriction at all on the slope!

Therefore at such a point (x_0, y_0), in general we cannot expect solutions of the differential equation $Mdx+Ndy = 0$ to be as "predictable" as they are when $M(x_0, y_0)$ and $N(x_0, y_0)$ are not both zero. In this sense, the points (x_0, y_0) at which $M(x_0, y_0)$ and $N(x_0, y_0)$ are both zero are "bad", so we give them a special name:

Definition 2.51 A point (x_0, y_0) is a *singular point* of the differential $Mdx + Ndy$ if $M(x_0, y_0) = 0 = N(x_0, y_0)$.[49] ■

Recall that a derivative-form DE, with independent variable x and dependent variable y, is said to be in *standard form* if the DE is of the form

$$\frac{dy}{dx} = f(x, y). \tag{2.98}$$

If the graph of a solution of (2.98) passes through (x_0, y_0), then the slope of the graph must be $f(x_0, y_0)$. This is true even if the IVP

$$\frac{dy}{dx} = f(x, y), \quad y(x_0) = y_0 \tag{2.99}$$

has more than one solution (which can happen if the hypotheses of the Existence and Uniqueness Theorem for derivative-form IVPs are not met, e.g. if $\frac{\partial f}{\partial y}$ is not continuous at (x_0, y_0)). So in some sense, a singular point (x_0, y_0) of a differential $Mdx + Ndy$ is a worse problem for the differential-form IVP "$Mdx + Ndy = 0$ with initial condition (x_0, y_0)" than we ever see for the derivative-form IVP (2.99). This is another important difference between derivative-form DEs and differential-form DEs.

It is difficult to define "maximal solution curve" *satisfactorily* for an equation $Mdx+Ndy = 0$ on a region in which $Mdx+Ndy$ has a singular point. But in regions free of singular points, there are no difficulties. We make the following definition:

Definition 2.52 Let R be a region in which the differential $Mdx + Ndy$ has no singular points. A solution curve $\mathcal{C}$ of the equation $Mdx + Ndy = 0$ is *maximal in* R if $\mathcal{C}$ is inextendible in R (see Definition 2.48).

[49] "Singular point" here does not mean the same thing as in footnote 33.

While it may appear that this definition could be made without the "no singular points" assumption, it would not be a *satisfactory* definition, for technical reasons that will not be discussed here.

We can now state the differential-form analog of the Existence and Uniqueness Theorem for derivative-form initial-value problems:

Theorem 2.53 *Suppose M and N are continuously differentiable functions on an open region R in $\mathbf{R}^2$, and that $Mdx + Ndy$ has no singular points in R. Then for every point $(x_0, y_0) \in R$, there exists a unique solution curve of $Mdx + Ndy = 0$ that that passes through (x_0, y_0) and is maximal in R.*

Like the analogous theorem for derivative-form initial-value problems, this theorem gives *sufficient* conditions under which a desirable conclusion can be drawn, not *necessary* conditions. There are differential-form equations $Mdx+Ndy = 0$ that have a unique intextendible solution curve through a point (x_0, y_0), even though (x_0, y_0) is a singular point of the differential. But there are also differentials $Mdx+Ndy$ for which M and N are continuously differentiable in the whole xy plane, for which $Mdx+Ndy$ has a singular point (x_0, y_0), and for which the equation $Mdx + Ndy = 0$, on some region R containing (x_0, y_0), has no solution curve through (x_0, y_0), or has several inextendible solution curves through (x_0, y_0), or has infinitely many inextendible solution curves through (x_0, y_0).

Under another name, singular points of *exact* differentials are familiar to students who've taken Calculus 3:

Example 2.54 Suppose $Mdx + Ndy$ is exact on a region R, and let F be a function on R for which $Mdx + Ndy = dF$. Then $M = \frac{\partial G}{\partial x}$ and $N = \frac{\partial G}{\partial y}$. Hence, for a given point $(x_0, y_0) \in R$,

$$\begin{aligned} & (x_0, y_0) \text{ is a singular point of } dF \\ \iff \quad & M(x_0, y_0) = 0 = N(x_0, y_0), \\ \iff \quad & \frac{\partial G}{\partial x}(x_0, y_0) = 0 = \frac{\partial G}{\partial y}(x_0, y_0), \\ \iff \quad & (x_0, y_0) \text{ is a critical point of } F. \end{aligned}$$

Thus, the singular points of dF are exactly the critical points of F.

2.6.4 Implicit solutions of DEs in differential form

The fact that derivative-form and differential-form DEs are intrinsically very different animals is generally not mentioned in DE textbooks. Consequently, textbooks' definitions of "solution of a differential-form DE" tends to look very similar to their definitions of "solution of a derivative-form DE". Usually this is accomplished by saying, early on, "We're going to use the word 'solution' to refer to both 'explicit' and

implicit solutions (of derivative-form DEs)," and then effectively take the definition of "solution of a DE in differential form" to be "implicit solution of a related derivative-form DE".[50] Since this is what students are most likely to see in a textbook, we make here a similar definition of "(implicit) solution of a DE in differential form" that is consistent with textbooks' treatment of this concept, but relate it carefully to the concepts we have developed earlier.

Definition 2.55 An equation

$$F(x, y) = 0 \quad (\text{or } F(x, y) = \text{ any real number } C_0) \tag{2.100}$$

is an *implicit solution* of a differential-form equation

$$M(x, y)dx + N(x, y)dy = 0 \tag{2.101}$$

on a region R if

(i) the portion of the graph of (2.100) that lies in R contains a smooth curve, and

(ii) every smooth curve in R contained in the graph of (2.100) is a solution curve of (2.101).[51]

If $R = \mathbf{R}^2$ then we usually omit mention of the region, and say just that (2.100) is an *implicit solution* of (2.101). ■

Remark 2.56 Note that we have not defined the term "solution of a DE in differential form". The most sensible definition of "solution of a DE in differential form" is what we have defined to be a *solution curve* of such a DE. We have used the two-word phrase *solution curve* only for pedagogical reasons. But temporarily, let us call a solution *curve* of a differential-form DE simply a *solution* of that DE; this will help with the discussion of our next point: The fundamental differences between derivative-form DEs and differential-form DEs make it awkward to come up with good terminology for what equation (2.100) is in relation to (2.101). Because a curve is a point-set in the plane, an equation of the form $F(x, y) = 0$ is actually a very *explicit* description of a curve $\mathcal{C}$ (when this equation does define a curve): a point (x, y) is on $\mathcal{C}$ if and only if $F(x, y) = 0$. "Implicitly-defined function" (of one variable, for DEs) is a perfectly sensible concept and term; "implicitly-defined *curve*" is not. The only thing that is really "implicit" about "*implicit solution* of a differential-form equation" as defined

[50] *Note to instructors:* The reason that we have a good relation at all between differential-form ODEs and (certain) derivative-form ODEs is, literally, because $1 + 1 = 2$. A curve in $\mathbf{R}^2$ has both dimension 1 and codimension 1. Graphs of equations $y = \phi(x)$ have *dimension* 1. Graphs of equations $F(x, y) = 0$ have *codimension* 1 (generically).

[51] *Note to instructor*: Observe that again, we do not assume that F is differentiable, or even continuous. Of course any F we are likely to find by any standard method *will* be differentiable, but for the purposes of *concept* and *definition*, that is beside the point.

above, is that the equation (2.100) *itself* is not a solution of (2.101)—the *solutions* of (2.101) related to (2.100) are smooth curves contained in the *graph* of equation (2.100).

Example 2.57 The equation

$$xy = 1$$

is a solution of

$$ydx + xdy = 0. \tag{2.102}$$

The graph, a hyperbola, consists of two solution curves, one lying in the first quadrant of the xy plane, the other lying in the third quadrant. One of the solution curves admits the continuously differentiable, non-stop parametrization $x(t) = t,\ y(t) = \frac{1}{t},\ t \in (0, \infty)$, while the other admits the continuously differentiable, non-stop parametrization $x(t) = t,\ y(t) = \frac{1}{t},\ t \in (-\infty, 0)$.

More generally, for every real number C, the equation

$$xy = C$$

is a solution of the same DE (2.102). For most C, the graph is a hyperbola, but the case $C = 0$ is exceptional. The graph of

$$xy = 0 \tag{2.103}$$

is a pair of crossed lines, the x- and y-axes. Note that this graph is not a smooth curve, nor is it the *disjoint* union of two smooth curves the way a hyperbola is ("disjoint" meaning that the two curves have no points in common). We can verify that (2.103) is indeed a solution of (2.102) by observing that the parametrized curves given by $x(t) = t, y(t) = 0,\ t \in \mathbf{R}$ (a continuously differentiable, non-stop parametrization of the x-axis) and $x(t) = 0, y(t) = t,\ t \in \mathbf{R}$ (a continuously differentiable, non-stop parametrization of the y-axis) both satisfy

$$y(t)\frac{dx}{dt} + x(t)\frac{dy}{dt} \equiv 0.$$

So we can express the graph of $xy = 0$ as the union of two solution curves of (2.102)—the graph of $y = 0$ and the graph of $x = 0$—but, unlike for the graph of $xy = C$, with $C \neq 0$ we cannot do it without having the two solution curves intersect. The source of this difference is that only for $C = 0$ does the graph of $xy = C$ contain $(0, 0)$, a singular point of $ydx + xdy$. ■

Remark 2.58 You may wonder to what extent criterion (i) in Definition 2.55 is necessary. An example of a graph that we would not want to call a solution curve of any DE is the graph of $x^2+y^2 = 0$: the graph is a single point, and includes no smooth curves at all. Obviously, we would also want to exclude graphs that consist of just two points, just ten points, etc. Criterion (i) does this, but does it do anything else? Could we get away with just excluding graphs that consist of a bunch of disconnected points?

Pushing this question a little further: suppose that we have an equation $F(x, y) = 0$ whose graph in the open set R is a *curve*, or a union of curves. Is it possible for this graph not to have *any* smooth portion, not even a teeny-tiny one?

You've seen many curves that were not *entirely* smooth, like the graph of $y = |x|$, but the curves you're accustomed to seeing are *mostly* smooth—there may be one or several points at which they're not smooth, but those points are joined by smooth sub-curves. These curves are the *piecewise smooth* curves that you may have seen in Calculus 3.

If you try to draw a curve (or, more generally, the graph of an equation $F(x, y) = 0$) that contains *no* smooth portions, you will not succeed. But the key word here is *draw*. There are, indeed, curves that contain no smooth portions at all. An example you may have seen is the infinitely jagged "snowflake curve", which is defined as a limit of certain piecewise-smooth curves, each of which is obtained from the preceding one by making it more jagged in a certain way. The best representation you can draw is an approximation of the limiting curve, obtained by stopping the iterative process at some stage. You may have heard of *fractals*, of which the snowflake curve is one example, but there are examples even more badly-behaved than fractals.

An equation $F(x, y) = 0$ can have a graph as bad as what we have just described, even if F is continuously differentiable. The graph does not care whether you can draw it. It is what it is. That's why we need a criterion like (i) in Definition 2.55. ■

2.6.5 Exact equations

The next example is very general. It is key to understanding the differential equations that are called *exact*.

Example 2.59 Suppose $M dx + N dy$ is an exact differential on a region R (see Definition 2.39), and let F be a differentiable function on R for which $M dx + N dy = dF$. Then (2.93) becomes

$$\frac{\partial G}{\partial x}\, dx + \frac{\partial G}{\partial y}\, dy = 0. \tag{2.104}$$

Suppose that $\mathcal{C}$ is a solution curve of (2.104), and that $t \mapsto (x(t), y(t))$, $t \in I$, is a continuously differentiable parametrization of $\mathcal{C}$. Then (2.94) says

$$\frac{\partial G}{\partial x}(x(t), y(t))\frac{dx}{dt} + \frac{\partial G}{\partial y}(x(t), y(t))\frac{dy}{dt} = 0. \tag{2.105}$$

By the Chain Rule, the left-hand side of (2.105) is just $\frac{d}{dt}F(x(t), y(t))$. Thus, (2.94) simplifies, in this case, to

$$\frac{d}{dt}F(x(t), y(t)) = 0 \quad \text{for all } t \in I. \tag{2.106}$$

Since I is an interval, this implies that $F(x(t), y(t))$ is constant in t. Thus, for every parametric solution $(x(t), y(t))$ of the equation $dF = 0$ on R, there is a (specific, non-arbitrary) constant c_0 such that

$$F(x(t), y(t)) = c_0 \tag{2.107}$$

for all $t \in I$. This implies that *every solution curve of* (2.104) *in* R *is contained in the graph of* (2.107) *for some value of the constant* c_0.

Now, fix a number c_0, and consider the equation

$$F(x, y) = c_0. \tag{2.108}$$

Is this equation a solution of (2.104) in R, according to Definition 2.55? The answer is yes, provided that criterion (i) of the definition is met. If this criterion is met, let $\mathcal{C}$ be a smooth curve in R that is contained in the graph of (2.108). Let γ be such a continuously differentiable parametrization of $\mathcal{C}$, and write $\gamma(t) = (x(t), y(t))$, $t \in I$. Since every point of $\mathcal{C}$ lies on the graph of (2.108), equation (2.107) is satisfied for all $t \in I$. Differentiating both sides of (2.107) with respect to t, we find that equation (2.106) is satisfied. But, by the Chain Rule, the left-hand side of (2.106) is exactly the left-hand side of (2.105), so equation (2.105) is satisfied. Therefore $\mathcal{C}$ is a solution curve of the differential equation (2.104). Hence criterion (ii) of Definition 2.55 is met, so (2.108) is a solution of the DE (2.104) in R. ■

Example 2.60 Suppose we are given a differential-form equation

$$M dx + N dy = 0 \tag{2.109}$$

that is exact on a region R, and we have found a function F such that $M dx + N dy = dF$ on R. Then Example 2.59 shows that the set of all solutions of (2.104) on R, in implicit form, is the collection of equations

$$\{F(x, y) = C\}, \tag{2.110}$$

where C is a "semi-arbitrary" constant: the allowed values of C are those for which the graph of (2.110) contains a smooth curve in R. ■

In view of this example, we make the following definition, stating ahead of time that we are choosing the letter C to stand for an (unspecified) constant:

Definition 2.61 In the setting of Example 2.60, we call the set

$$\{F(x,y) = C \mid C \in \mathbf{R}\}, \text{ also denoted simply } \{F(x,y) = \text{constant}\} \text{ or } \{F(x,y) = C\} \tag{2.111}$$

the *general solution, in implicit form,* of (2.109). In the notation (2.111), it is understood that C is a constant and that the set of C's for which the graph of "$F(x,y) = C$" contains a smooth curve (and therefore for which "$F(x,y) = C$" is a solution of (2.109)), the set of "allowed" C's is some subset of the range of F that we are not specifying explicitly. If we are able to specify this set explicitly, then we may substitute the corresponding restrictions on C (e.g. "$C > 0$") for "$C \in \mathbf{R}$" in (2.111). ■

For any C, the graph of $F(x,y) = C$ is called a *level set* of F. A level set can *contain* a smooth curve without *being* a smooth curve. One familiar example is the graph of $xy = 0$, which consists of two crossed lines. But in this example, every point of the level-set lies on at least one smooth curve (either the x-axis or the y-axis) contained in the level-set. The next example shows that this is not always the case.

Example 2.62 (Level-set with a corner) Let $F(x,y) = y^3 - |x|^3$. This function has continuous second partial derivatives on the whole plane $\mathbf{R}^2$ (for example $\frac{\partial G}{\partial x}(x,y) = \begin{cases} -3x^2, & x \geq 0 \\ 3x^2, & x \leq 0 \end{cases}$, so $\frac{\partial^2 G}{\partial x^2}(x,y) = \begin{cases} -6x, & x \geq 0 \\ 6x, & x \leq 0 \end{cases}$). It has one critical point, the origin. The level-set containing this critical point is the graph of

$$y^3 - |x|^3 = 0, \tag{2.112}$$

which is simply the graph of $y = |x|$. The portion of this graph in the open first quadrant ($y = x,\ x > 0$) is a smooth curve contained in this level-set, and so is the portion of this graph in the open second quadrant. But the origin is a point of this level-set that is not contained in any smooth curve in the level-set.

Equation (2.112) is a solution of

$$y^2 dy + \left\{ \begin{array}{ll} -3x^2, & x \geq 0 \\ 3x^2, & x \leq 0 \end{array} \right\} dx = 0; \tag{2.113}$$

it meets both criteria in Definition 2.55. However, the graph of (2.112) contains a point, $(0,0)$, that is not on any solution *curve* of (2.113) (see Definitions 2.49 and 2.47). Thus, in general, the graph of a solution "$F(x,y) = C$" of $dF = 0$ can include points that do not lie on any solution *curve* of $dF = 0$. ■

Note that the "corner" of the level set $F(x, y) = 0$ in Example 2.62 was a critical point of F (hence a singular point of the differential dF). In the absence of singular points, we can be much more concrete about the implicit-form general solution of an exact equation:

If a differential $Mdx + Ndy$ is exact on a region R and has no singular points in R, then the set of C's allowed in (2.110) is simply the range of F on the region C, and every point in R is contained in a unique solution curve that is maximal in R. (2.114)

To see why this is true, the interested student may read Example 3.1 in the optional-reading Section 3.2.

2.7 Algebraic equivalence of DEs in differential form

Algebraic equivalence (see Definition 2.43) has the same importance for DEs in differential form that it has for DEs in derivative form. Suppose that two equations $M_1dx + N_1dy = 0$ and $M_2dx + N_2dy = 0$ are algebraically equivalent on a region R . Then there is a function f on R, nonzero at every point of R, such that $M_2 = fM_1$ and $N_2 = fN_1$. If $\mathcal{C}$ is a solution curve of $M_1dx + N_1dy = 0$ and $t \mapsto (x(t), y(t))$, $t \in I$, is a continuously differentiable, non-stop parametrization of $\mathcal{C}$, then

$$\begin{aligned} & M_2(x(t), y(t))\frac{dx}{dt} + N_2(x(t), y(t))\frac{dy}{dt} \\ = \;& f(x(t), y(t)) \left(M_1(x(t), y(t))\frac{dx}{dt} + N_1(x(t), y(t))\frac{dy}{dt} \right) \\ = \;& f(x(t), y(t)) \times 0 \\ = \;& 0. \end{aligned}$$

Thus $\mathcal{C}$ is a solution curve of $M_2dx + N_2dy = 0$, and $t \mapsto (x(t), y(t))$ is a parametric solution of this DE. Hence every solution curve of $M_1dx + N_1dy = 0$ is a solution curve of $M_2dx + N_2dy = 0$, and the same goes for parametric solutions.

Similarly, since f is nowhere zero on R, we have $M_1 = \frac{1}{f}M_2$ and $N_1 = \frac{1}{f}N_2$. The same argument as above, with the subscripts "1" and "2" interchanged and with f replaced by $\frac{1}{f}$, shows that every solution curve or parametric solution of $M_2dx + N_2dy = 0$ is a solution curve or parametric solution of $M_1dx + N_1dy = 0$. Adding Definition 2.55 to this analysis, we have the following:

If two differential-form DEs are algebraically equivalent on a region R, then in R they have exactly the same solution curves, exactly the same parametric solutions, and exactly the same solutions. (2.115)

Observe that if $M_2 = fM_1$ and $N_2 = fN_1$, but f is allowed to be zero somewhere on R, then every solution curve (or parametric solution) of $M_1dx + N_1dy = 0$ is a solution curve (or parametric solution) of $M_2dx + N_2dy = 0$, but the reverse may not be true. (A similar statement holds for equations in derivative form.) Thus, just as for derivative form, when we algebraically manipulate differential-form DEs, *if we multiply or divide by functions that can be zero somewhere, we can gain or lose solutions*, and therefore wind up with a set of solutions that is *not* the set of all solutions of the DE we started with.

The next example (in which the DE is *not* exact), is included to illustrate an interesting phenomenon. The student should be able to follow the author's steps, but is not expected to understand how the author knew to take these steps.

Example 2.63 Consider the DE

$$2xy\ dx + (y^2 - x^2)dy = 0. \tag{2.116}$$

This DE is not exact on any region in the xy plane. However, the functions $M(x, y) = 2xy$ and $N(x, y) = y^2 - x^2$ are continuously differentiable on the whole plane, and the only point at which they are both zero is $(0, 0)$. So, as with (2.102), we have a differential with one singular point, which happens to be the origin[52]. Letting $R = \{\mathbf{R}^2 \text{ minus the origin}\}$, Theorem 2.53 guarantees us that through each point $(x_0, y_0) \neq (0, 0)$, there exists a unique solution curve of (2.116).

Observe that the positive x-axis is a solution-curve: if we set $x(t) = t, y(t) = 0, t \in (0, \infty)$, then the image of this parametrized curve is the positive x-axis, and for all $t \in (0, \infty)$ we have

$$2x(t)y(t)\ \frac{dx}{dt} + (y(t)^2 - x(t)^2)\frac{dy}{dt}\ =\ 2t \times 0 \times 1 + (-t^2) \times 0\ =\ 0.$$

Similarly, the negative x-axis is a solution-curve. The uniqueness statement in Theorem 2.53 guarantees us that the positive and negative x-axes are the *only* solution curves containing a point on either of these open half-axes. Therefore no other solution curve in R contains a point (x, y) for which $y = 0$; every other solution curve in R lies either entirely in the region $R_+ = \{(x, y) \mid y > 0\}$ (the half-plane above the x-axis), or entirely in the region $R_- = \{(x, y) \mid y < 0\}$ (the half-plane below the x-axis).

On R_+, and also on R_-, equation (2.116) is algebraically equivalent to

$$\frac{1}{y^2}\left(2xy\ dx + (y^2 - x^2)dy\right) = 0. \tag{2.117}$$

But as the student may verify,

[52]In general, singular points can occur anywhere in the xy plane. The origin is used in most examples in these notes just to simplify the algebra, so that the student may focus more easily on the concepts.

$$\begin{aligned}\frac{1}{y^2}\left(2xy\ dx + (y^2 - x^2)dy\right) &= 2\frac{x}{y}\ dx + (1 - \frac{x^2}{y^2})dy \\ &= d\left(\frac{x^2}{y} + y\right) \\ &= d\left(\frac{x^2 + y^2}{y}\right).\end{aligned}$$

So on R_+, and also on R_-, the left-hand side of (2.117) is exact; it is dF, where $F(x, y) = \frac{x^2+y^2}{y}$. Hence the general solution of (2.117), in either of these regions, is

$$\left\{\frac{x^2 + y^2}{y} = C\right\}. \tag{2.118}$$

where, from fact (2.114), the set of allowed values of C is the range of F on each region. Since the sign of $\frac{x^2+y^2}{y}$ is the same as the sign of y, this means that on R_+, only positive C's will be allowed, and on R_-, only negative C's will be allowed. To see that these are the only restrictions on C, just observe that from the definition of of F, we have $F(0, C) = C$.

Now for some algebraic rearrangement. Let us write $C = 2b$ in (2.118). Then b is a semi-arbitrary constant with $b > 0$ for solution curves in R_+, and $b < 0$ for solution curves in R_-. On each of these two regions,

$$\begin{aligned}&\frac{x^2 + y^2}{y} = 2b \\ \iff\quad & x^2 + y^2 = 2by, \\ \iff\quad & x^2 + y^2 - 2by = 0, \\ \iff\quad & x^2 + y^2 - 2by + b^2 = b^2, \\ \iff\quad & x^2 + (y - b)^2 = b^2.\end{aligned} \tag{2.119}$$

The graph of (2.119) in $\mathbf{R}^2$ is a circle of radius $|b|$ centered at $(0, b)$ on the y-axis; the graph in R is the circle with the origin deleted. Thus, these circles-with-origin-deleted are the maximal solution curves of (2.117) on R_+ and on R_-. But since (2.117) is algebraically equivalent to (2.116) on these regions, the same curves are all the solution curves of (2.116) in these regions.

We have now found all the solution curves of (2.116) in R that do not intersect the x-axis, as well as all those that do intersect it. So we have all the solution curves in $R = \{\mathbf{R}^2$ minus the origin$\}$. If we now re-include the origin, we see that the origin lies on every one of the circles (2.119), as well as on the x-axis. With the origin re-included, it is easy to see that the full x-axis is a solution curve of (2.116). We leave the student to check that each full circle (2.119), with the origin included, is also a solution curve of (2.116).

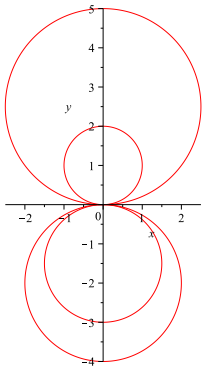

Figure 4: Some solution curves of $2xy\,dx + (y^2 - x^2)dy = 0$. (The graphing utility used to render this diagram does not do a good job near the origin; there should be no gap in any of the circles.)

Thus, *among* the solution curves of (2.116) are all circles centered on the y axis, plus one "exceptional" curve, the x-axis. We can write the set of all implicit solutions corresponding to this set of solution curves as

$$\{x^2 + (y-b)^2 = b^2 \mid b \neq 0\} \quad \text{and} \quad \{y = 0\}. \tag{2.120}$$

An alternative way of expressing this set of solutions is as follows. In (2.118), C can be any nonzero constant, so we may write C as $\frac{1}{K}$, where the allowed values of K are also anything other than zero. We can then rewrite (2.118) as $y = K(x^2 + y^2)$. The solution curves that lie in R_+ have $K > 0$; those that lie in R_+ have $K < 0$. These give all the solutions in the "b-family", just expressed in different-looking but algebraically equivalent way. But magically, if we now allow $K = 0$, we get the lonely $y = 0$ solution as well. So we can also write the set (2.120) in a unified way as

$$\left\{y = C(x^2 + y^2) \mid C \in \mathbf{R}\right\}. \tag{2.121}$$

(We have renamed K back to C just to emphasize that the letter chosen an arbitrary or "semi-arbitrary" constant does not matter, as long as it is clear that this is what the letter represents.)

From the foregoing analysis, it may appear that the set of all solution *curves* of (2.116) on $\mathbf{R}^2$ consists of all circles centered on the y axis, plus one "exceptional" curve, the x-axis. Similarly, it may appear that the set of all *implicit solutions* of (2.116) is (2.120), or equivalently (2.121).

But both of these conclusions are wrong! To see why, in Figure 4 start at a point P other than the origin. This point lies on a unique circle in the figure. Move along this circle in either direction till you reach the origin. When you reach the origin continue moving, but go out along a different circle, either on the same side of the y-axis as the first circle or on the opposite side, whatever you feel like. Stop

at a point Q before you reach the origin again. Erase the endpoints P and Q (see the second paragraph after Definition 2.47), and you have a perfectly good, smooth, solution curve that is not contained in any circle or in the x-axis.

You can let the x-axis into this game as well. For example, start on the positive x-axis, move left till you reach the origin, and then move out along one of the circles.

Thus there are solution curves of (2.117) that are not contained in any of the "circles plus one straight line" family given by (2.120) or (2.121). ■

In Example 2.63, all the solution curves in $\mathbf{R}^2$ intersected at the origin (a singular point of $Mdx+Ndy$), but all had the same slope there (zero). Next we give an example of a very simple equation of the form $Mdx+Ndy=0$ in which all the solution curves in $\mathbf{R}^2$ intersect at a singular point of $Mdx+Ndy$, but with all different slopes—in fact, with every possible slope.

Example 2.64 Consider the DE

$$xdy - ydx = 0. \tag{2.122}$$

The student may check that every straight line through the origin—whether horizontal, vertical, or oblique—is a solution curve.

The only singular point of $xdy-ydx$ is the origin. Therefore in $R = \{\mathbf{R}^2$ minus the origin$\}$, there is a unique solution curve through every point. If we take the straight lines through the origin, and delete the origin, we get the collection of open rays emanating from the origin. Every point of R lies on one and only one such ray. Therefore these are all the solution curves of (2.122) in $\{\mathbf{R}^2$ minus the origin$\}$. It follows that there are no solution curves of (2.122) in $\mathbf{R}^2$ other than what we get by re-including the origin. Thus the set of solution curves in $\mathbf{R}^2$ is the family of all straight lines through the origin. ■

2.8 Relation between differential form and derivative form

Definition 2.65 Let M, N be functions on a region R in $\mathbf{R}^2$. Consider the equations

$$M(x,y)dx + N(x,y)dy = 0, \tag{2.123}$$

$$M(x,y) + N(x,y)\frac{dy}{dx} = 0, \tag{2.124}$$

$$M(x,y)\frac{dx}{dy} + N(x,y) = 0. \tag{2.125}$$

We call equations (2.124) and (2.125) the *derivative-form DEs associated with the differential-form DE* (2.123). Similarly, we call equation (2.123) the *differential-form*

DE associated with the derivative-form DE (2.124), and also the *differential-form DE associated with the derivative-form DE* (2.125).

More generally, if a derivative-form equation is algebraically equivalent to (2.124) or (2.125) on a region R, we call the equation *a derivative form of* (2.123) on R. Similarly, if a differential-form equation is algebraically equivalent to (2.123) on a region R, we call the equation a differential form of (2.124) and (2.125) on R.[53]

It is easy to remember how to associate a differential-form DE to a derivative-form DE, and vice-versa: **Pretend** that $\frac{dy}{dx}$ and $\frac{dx}{dy}$ are actual fractions with the numerators and denominators that the notation suggests, and formally "divide" equation (2.123) by dx or dy to obtain the associated equation (2.124) or (2.125), or formally "multiply" equation (2.124) or (2.125) by dx or dy to obtain the associated equation (2.123). *This is an* **extremely** *useful memory-device, and the student should not hesitate to use it,* **but mathematically it is garbage**.[54] The Leibniz notation "$\frac{dy}{dx}$" for derivatives has many extraordinarily useful features, but the student must remember that it is *only notation*, in which neither dy nor dx is a real number, and which *does not* represent a fraction with numerator dy and denominator dx.

In this section of the notes we will see how and why equations (2.123)–(2.125) *actually* are related to each other.

To start, suppose that $\mathcal{C}$ is smooth curve, and γ a continuously differentiable, non-stop parametrization of $\mathcal{C}$, with domain-interval I. Write $\gamma(t) = (f(t), g(t))$ (for what we are about to do, writing "$\gamma(t) = (x(t), y(t))$" would lead to confusion). Let's call a subinterval I_1 of I "x-monotone" if $f'(t)$ is nowhere 0 on I_1, and "y-monotone" if $g'(t)$ is nowhere 0 on I_1.[55] (These are not mutually exclusive: if both $f'(t)$ and $g'(t)$ are nowhere zero on I_1, then I_1 is both x-monotone and y-monotone. For example, if we parametrize a circle by $\gamma(t) = (\cos t, \sin t)$, then the interval $(0, \pi/2)$, in which γ traces out the quarter-circle in the open first quadrant, is both x-monotone and y-monotone. The interval $(0, \pi)$, in which γ traces out the half-circle lying above the x-axis, is x-monotone but not y-monotone.)

Since γ is a non-stop parametrization, for every $t_0 \in I$ at least one of the two

[53]The last paragraph of this definition is more restrictive than any analogous statement in textbooks from which I've taught in the past, all of which omit the (important!) requirement of algebraic equivalence. Except in the context of separable equations, current textbooks tend to omit any mention whatsoever of the *logical* relation between a given DE, and the DE obtained from the given one by multiplying it through by a function. Current textbooks allow (and, by setting an example, implicitly encourage) multiplication/division by functions that are zero somewhere. But this can lead to losing one or more solutions of the original DE, or gaining one or more spurious "solutions"—functions (or curves) that are not solutions (or solution curves) of the original DE.

[54]Unfortunately, most DE textbooks do not mention that this way of viewing the relations among (2.123), (2.124), and (2.125) is mathematical nonsense, and simply encourage the formal multiplication/division without giving any explanation whatsoever of why the derivative-form and differential-form equations are related to each other.

[55]This is *very temporary* terminology, invented *only* for this part of these notes.

numbers $f'(t_0), g'(t_0)$ is nonzero. If $f'(t_0) \neq 0$, then since f' is assumed to be continuous, there is some open interval containing t_0 on which $f'(t)$ is nonzero and has the same sign as $f'(t_0)$. A similar statement holds if $g'(t_0) \neq 0$. Thus, every $t \in I$ lies in a subinterval I_1 that is either x-monotone or y-monotone.

Let I_1 be an x-monotone interval. Then $f'(t)$ not zero for any $t \in I_1$. The Inverse Function Theorem that you may have learned in Calculus 1 assures us that there is an inverse function f^{-1}, with domain an interval I_2 and with range I_1, and that f^{-1} is continuously differentiable[56]. Let $\mathcal{C}_1$ be the smooth curve parametrized by $(f(t), g(t))$ using just the x-monotone open interval I_1 rather than the whole original interval I. On this domain, "$x = f(t)$" is equivalent to "$t = f^{-1}(x)$". So, temporarily writing $t_{\text{new}} = x$, for $(x, y) = (f(t), g(t)) \in \mathcal{C}_1$ we have

$$\begin{aligned} x &= t_{\text{new}}, \\ y = g(t) = g(f^{-1}(x)) &= g(f^{-1}(t_{\text{new}})) \\ &= \phi(t_{\text{new}}) \end{aligned}$$

where $t_{\text{new}} \in I_2$ and $\phi = g \circ f^{-1}$. Since g and f^{-1} are continuously differentiable, so is h. Furthermore, $dx/dt_{\text{new}} \equiv 1 \neq 0$. Therefore the equations above give us a new continuously differentiable, non-stop parametrization γ_{new} of $\mathcal{C}_1$:

$$\gamma_{\text{new}}(t_{\text{new}}) = (t_{\text{new}}, \phi(t_{\text{new}})). \tag{2.126}$$

The variable in (2.126) is a "dummy variable"; we can give it any name we like. Since the x-component of $\gamma_{\text{new}}(t_{\text{new}})$ is simply the parameter t_{new} itself, we will simply use the letter x for the parameter; thus

$$\gamma_{\text{new}}(x) = (x, \phi(x)). \tag{2.127}$$

Thus, this parametrization uses x itself as the parameter, treats x as an independent variable, and treats y as a dependent variable related to x by $y = \phi(x)$.

Now suppose that our original curve $\mathcal{C}$ is a solution curve of a given differential-form DE

$$M(x, y)dx + N(x, y)dy = 0. \tag{2.128}$$

[56]This important theorem *used* to be stated, though usually not proved, in Calculus 1. Unfortunately, it seems to have disappeared from many Calculus 1 syllabi. The theorem says that if f is a differentiable function on an interval J, and $f'(t)$ is not 0 for any $f \in J$, then (i) the range of f is an interval K, (ii) an inverse function f^{-1} exists, with domain K and range J, and (iii) f^{-1} is differentiable, with its derivative given by $(f^{-1})'(x) = 1/f'(f^{-1}(x))$. (If we write $x = f(t)$ and $t = f^{-1}(x)$, then the formidable-looking formula for the derivative of f^{-1} may be written in the more easily remembered, if somewhat less precise, form $\frac{dt}{dx} = \frac{1}{dx/dt}$.) If the derivative of h is continuous, so is the derivative of h^{-1}.

Then $\mathcal{C}_1$, a subset of $\mathcal{C}$, is also a solution curve, so *every* continuously differentiable, non-stop parametrization $(x(t), y(t))$ of $\mathcal{C}_1$ satisfies

$$M(x(t), y(t))\frac{dx}{dt} + N(x(t), y(t))\frac{dy}{dt} = 0 \tag{2.129}$$

In particular this is true for the parametrization (2.127), in which the parameter t is x itself, and in which have $y(t) = \phi(t) = \phi(x) = y(x)$. Therefore, for all $x \in I_2$,

$$\begin{aligned} 0 &= M(x, \phi(x))\frac{dx}{dx} + N(x, \phi(x))\ \phi'(x) \\ &= M(x, \phi(x)) + N(x, \phi(x))\ \phi'(x). \end{aligned} \tag{2.130}$$

The right-hand side of (2.130) is exactly what we get if we substitute "$y = \phi(x)$" into $M(x, y) + N(x, y)\frac{dy}{dx}$. Hence ϕ is a solution of

$$M(x, y) + N(x, y)\frac{dy}{dx} = 0. \tag{2.131}$$

Therefore the portion $\mathcal{C}_1$ of $\mathcal{C}$ is the graph of a solution (namely ϕ) of the derivative-form differential equation (2.131). The argument above also gives us the following an important fact to which we will want to refer later:

If a solution curve of the differential-form equation $Mdx + Ndy = 0$ can be parametrized by $\gamma(x) = (x, \phi(x))$, where ϕ is a differentiable function, then ϕ is a solution of the associated derivative-form equation $M + N\frac{dy}{dx} = 0$. (2.132)

Similarly, if $\mathcal{C}_2$ is a portion of $\mathcal{C}$ obtained by restricting the original parametrization γ to a y-monotone interval I_2, then $\mathcal{C}_2$ is the graph of a differentiable function $x(y)$—more precisely, the graph of the equation $x = \phi(y)$ for some differentiable function ϕ—that is a solution of the derivative-form differential equation

$$M(x, y)\frac{dx}{dy} + N(x, y) = 0. \tag{2.133}$$

Therefore:

Every solution curve of the differential-form equation (2.123) **is a union of solution curves of the derivative-form equations** (2.124) **and** (2.125)**.** (2.134)

Note that the graphs mentioned in (2.134) will overlap, in general, since the x-monotone intervals and y-monotone intervals of a continuously differentiable, non-stop parametrization γ will usually overlap. (The only way there will not be an overlap

is if $f'(t) \equiv 0$ or $g'(t) \equiv 0$, in which case $\mathcal{C}$ is a vertical or horizontal straight line, respectively, and there are, respectively, no x-monotone or y-monotone subintervals.)

Now compare (2.131) with the general first-order derivative-form DE with independent variable x and dependent variable y,

$$\mathsf{G}(x, y, \frac{dy}{dx}) = 0. \tag{2.135}$$

Equation (2.131) is a special case of (2.135), in which the dependence of G on its third variable is very simple. If we use a third letter z for the third variable of G, then (2.131) corresponds to taking $\mathsf{G}(x, y, z) = M(x, y) + N(x, y)z$, a function that can depend in any conceivable way on x and y, but is linear separately in z. In general, (2.135) could be a much more complicated equation, such as

$$\left(\frac{dy}{dx}\right)^3 + (x + y)\sin(\frac{dy}{dx}) + xe^y = 0. \tag{2.136}$$

Solving equations such as the one above is *much* harder than is solving equations of the simpler form (2.131). For certain functions G that are more complicated than (2.131), but much less complicated than (2.136), methods of solution are known[57]. But the general theory and techniques for working with equation (2.135) for general G's are much less highly developed than they are for equations in the standard form (2.138) or in the form (2.131).

One of the features of (2.131) that makes it so special is that on any region on which $N(x, y) \neq 0$, (2.131) is algebraically equivalent to

$$\frac{dy}{dx} = -\frac{M(x, y)}{N(x, y)}, \tag{2.137}$$

which is of form

$$\frac{dy}{dx} = f(x, y). \tag{2.138}$$

Recall that equation (2.138) is exactly the "standard form" equation that appears in the fundamental Existence and Uniqueness Theorem for initial-value problems. This theorem is absolutely crucial in enabling us to determine whether our techniques of finding solutions actually give us *all* solutions.

If you re-read these notes, you will see that all the *general* facts about DEs in derivative form—such as the definition of "solution" and "implicit solution", and the fact that algebraically equivalent DEs have the same set of solutions—were stated

[57] One such type equation is a *Clairaut equation* $y = x\frac{dy}{dx} + g(\frac{dy}{dx})$, which is equivalent to (2.135) with $\mathsf{G}(x, y, z) = xz + g(z) - y$. Students using the textbook Nagle, Saff, and Snider, *Fundamentals of Differential Equations*, 8th ed., Pearson Addison-Wesley, 2012 can learn about these equations by doing Group Project 2F.

for the general first-order DE (2.135). These facts apply just as well to nasty DEs like (2.136) as they do to (relatively) nice ones like (2.138). However, in all of our *examples*, we used equations that were algebraically equivalent to (2.124) (hence also to (2.138)) on some region. The reason is that although the concept of "the set of all solutions" makes perfectly good sense for the general equation (2.135), the author wanted to use examples in which he could show the student easily that the set of all solutions had actually been found.

Nowadays, students in an introductory DE course rarely see any first-order derivative-form equations that are not algebraically equivalent, on some region, to a DE in the standard form (2.138). Because of this, it is easy to overlook a significant fact: **the *only* derivative-form DEs that are related to differential-form DEs are those that are algebraically equivalent to (2.138) on some region**. The two types of equations, in full generality, are not merely two sides of the same coin.

However, for derivative-form DEs that can be "put into standard form"—which are exactly those that are algebraically equivalent to a DE of the form (2.124) on some region—there is a very close relation between the two types of DEs. We are able to relate many, and sometimes all, solutions of a DE of one type to solutions of the associated DEs of the other type. Statement (2.134) gives one such relation.

To have a name for equations that are explicitly of the form (2.124) or (2.125), let us say that a derivative-form equation, with independent variable x and dependent variable y, is in "almost-standard form"[58] if it is in the form (2.124), or can be put in that form just by subtracting the right-hand side from the left-hand side. If you re-inspect the argument leading to the conclusion (2.134), you will see that it also shows that every solution curve of (2.124) or (2.125) is a solution curve of (2.123). Thus:

$$\left.\begin{array}{l}\textbf{Every solution curve of a derivative-form}\\ \textbf{equation in almost-standard form is a solution}\\ \textbf{curve of the associated differential-form equation.}\end{array}\right\} \tag{2.139}$$

Combining (2.134) and (2.139), we conclude the following:

$$\left.\begin{array}{l}\textbf{A smooth curve } \mathcal{C} \textbf{ is a solution curve of an equation}\\ \textbf{in differential form if and only if } \mathcal{C} \textbf{ is a union of}\\ \textbf{solution curves of the associated derivative-form}\\ \textbf{equations.}\end{array}\right\} \tag{2.140}$$

We emphasize that in deriving these relations, the transition from the differential-form DE (2.128) to the derivative-form DEs (2.131) and (2.133) was NOT obtained

[58] This is another bit of terminology invented only for these notes, just to have a name to distinguish (2.124) from (2.137) on regions in which $N(x, y)$ may be zero somewhere.

by the nonsensical process of "dividing by dx" or "dividing by dy", even though the notation makes it look that way. The transition was achieved by understanding that what we are looking for when we solve (2.123) are curves whose parametrizations satisfy (2.129), and that for particular choices of the parameter on the intervals that we called "x-monotone" or "y-monotone", (2.129) reduces to (2.124) or (2.125).

Similarly, transitions from derivative form to differential form are NOT achieved by the nonsensical process of "multiplying by dx" or "multiplying by dy". The benefit of the Leibniz notation " $\frac{dy}{dx}$ " for derivatives is that it can be used to help remember many true statements by *pretending, momentarily*, that you can multiply or divide by a differential just as if it were a real number[59]. In particular, we can use this principle help us easily *remember* that the differential-form equation (2.123) is related to (but not the same as!) the derivative-form equations (2.124) and (2.125). But this notational trick doesn't tell us everything, such as the *precise relationship* among these equations, which is statement (2.139) (of which statement (2.134) is the "only if" half).

2.9 Using differential-form equations to help solve derivative-form equations

The standard procedure taught in DE courses for using differential-form equations to help solve derivative-form equations is essentially the following:

Step 1. Write down a differential-form equation associated with the derivative-form DE.

Step 2. If this differential-form DE is exact, go to Step 3. Otherwise, attempt by algebraic manipulation to "turn the equation into" an exact DE or a separated DE, the latter meaning one of the form $h(y)dy = g(x)dx$. If you succeed, go on to Step 3. (If you do not succeed, then differential-form equations will not help you solve the original derivative-form equation.)

Step 3. If the new DE is exact, solve it by the "exact equations" method. If the new DE is separated, solve it by integrating both sides.

Step 4. Write down your final answer in the form "$\{F(x, y) = C\}$" (or, for separable equations, "$\{H(y) = G(x) + C\}$"), and hope that this is the general solution, in implicit form, of the original derivative-form DE.

Step 5. If the equations in your final answer can be solved explicitly for y in terms of x, then (usually) you should do so. Otherwise, stop after Step 4.

[59] Simultaneously, the *drawback* of the Leibniz notation is that it promotes some incorrect or lazy thought-patterns. It encourages the manipulation of symbols without the understanding of what the symbols means. It may lead the student to think something is "obviously true" when it isn't obvious, and often when it isn't true.

No doubt you noticed the phrase "and hope that this is the general solution, in implicit form, of the original derivative-form DE." All we did above is write down a sequence of steps, pushing symbols around a page. Our outline of this general procedure did not involve asking whether every solution of the equation we started yielded a solution-curve of the diffferential-form equation written in Step 1, or vice-versa; whether DE written in Step 2 had the same set of solution curves as the DE written in Step 1. So, why should we expect our final answer we've given to be the general solution (in implicit form) of the original derivative-form DE we were asked to solve?

Before discussing how to turn the "autopilot" procedure outlined above into a more reliable one, let us look an example that illustrates one of the problems with the procedure as outlined.

Example 2.66 Solve the differential equation

$$(10xy^9 + 2xy)\frac{dy}{dx} = -(3x^2 + 1 + y^{10} + y^2). \tag{2.141}$$

(As always, the instruction "solve the DE" means "find *all* the [maximal] solutions", i.e. the general solution.)

This DE is neither separable or linear. The standard method of attack is to look at the associated differential-form DE, of the form "differential=0", and hope that it is exact. In this case, the associated differential-form DE is[60]

$$(3x^2 + 1 + y^{10} + y^2)dx + (10xy^9 + 2xy)dy = 0. \tag{2.142}$$

The coefficients $M(x, y)$ of dx and $N(x, y)$ of dy are continuously differentiable on the whole xy plane, and we see that our differential $Mdx + Ndy$ passes the exactness test "$M_y = N_x$", so we know that there is some F, continuously differentiably on all of $\mathbf{R}^2$, for which the left-hand side of (2.142) is dF. Using our usual method, we find that an F with this property is

$$F(x, y) = x^3 + x + xy^{10} + xy^2. \tag{2.143}$$

From Example 2.60, we know that the general solution of (2.142) is

$$\{x^3 + x + xy^{10} + xy^2 = C\}, \tag{2.144}$$

where C is (at worst) a semi-arbitrary constant. Fact (2.114) shows that the set of allowed values of C is simply the range of F, provided that $Mdx + Ndy$ has no

[60]More precisely, in this sentence and the last, we should have said "one of the two" associated differential-form DEs. One of these is obtained by first subtracting the right-hand side of (2.141) from the left-hand side; the other is obtained by first subtracting the leftt-hand side of (2.141) from the right-hand side. Each of these equations is just the other with both sides multiplied by -1.

singular points. Looking at $M(x, y)$, we observe that x^2, y^{10}, and y^2 are all ≥ 0, so $M(x, y) \geq 1$. Therefore $M(x, y)$ is nowhere zero, so $Mdx + Ndy$ has no singular points. So fact (2.114) applies, and the set of allowed values of C is simply the range of F. We can easily see that this range is the entire real line $(-\infty, \infty)$. (Just set $y = 0$ in (2.143) and observe that $\lim_{x\to\infty} F(x, 0) = \infty$ and $\lim_{x\to-\infty} F(x, 0) = -\infty$.)

Therefore the general solution of (2.142), in implicit form, is the family of equations (2.144), with C a *completely* arbitrary constant; all real values are allowed.

But the equation we wanted to solve was (2.141), not (2.142), so we ask: is this same family of equations the set of all solutions of (2.141), in implicit form? The answer is no.

To see why, note that for (2.144) to be the set of all solutions of (2.141), in implicit form, two criteria must be satisfied: (i) every solution of (2.141) must be implicitly defined by one of the equations in the collection (2.144), and (ii) the collection of equations (2.144) cannot contain any "spurious implicit solutions"—equations that are not implicit solutions of (2.141). Fact (2.139) assures us that criterion (i) is met, so we need worry only about (ii).

Let's look at the $C = 0$ case of (2.144):

$$x^3 + x + xy^{10} + xy^2 = 0. \tag{2.145}$$

(Don't worry about "why this choice of C?" The author contrived this example so that $C = 0$ would be useful to look at; he is using information that the student doesn't have.) Observe that this equation can be rewritten as

$$x(x^2 + 1 + y^{10} + y^2) = 0. \tag{2.146}$$

The quantity inside parentheses is strictly positive, so (2.146) is equivalent to just $x = 0$. The graph of (2.146) is simply the y-axis, a perfectly nice smooth curve, and a perfectly good solution curve of (2.142), but it does not contain the graph of any function of x on any open interval. Therefore it does not contain the graph of any solution of (2.141). Therefore (2.146) is *not* an implicit solution of (2.141).

So our set of all solutions of (2.142) is *not* simply an implicit form of the general solution (2.141); it has at least one equation, namely (2.145), that doesn't belong in the latter. This demonstrates the main point of this example:

$$\left.\begin{array}{l}\textbf{The general solution of an almost-standard-form derivative-form DE is not always the same as the general solution of the associated differential-form DE.}\end{array}\right\} \tag{2.147}$$

To complete the current example, we would need to answer this question: Are there any values of C other than 0 for which $x^3 + x + xy^{10} + xy^2 = C$ is not an implicit solution of (2.141)? The answer is no. (This can be shown using the Implicit Function Theorem, but in the interests of brevity, and since demonstrating (2.147)

was the main point of the current example, we will omit the argument.) Thus the general solution of (2.141), in implicit form, is

$$\left\{x^3 + x + xy^{10} + xy^2 = C, \quad C \neq 0\right\}. \tag{2.148}$$

■

What Example 2.66 shows is that if you try to solve a differential equation by mindlessly pushing differentials around the page as if they were numbers, the answer you wind up with may not be the set of solutions to the equation you were trying to solve. In fact, when you realize how dissimilar differentials and numbers are, it should initially strike you as miraculous that you can even get *close* to the correct set of solutions by such manipulations. One of the chief purposes of these notes is to explain this miracle, but another is to get the student to appreciate that there is something to explain. Writing a derivative using fraction-notation doesn't make it a true fraction, any more than calling a hippopotamus a lollipop makes it a lollipop. Treating "$\frac{dy}{dx}$" as if it were a fraction is an *abuse of notation*, and conclusions we reach from treating it like a fraction need to be justified some other way.

Despite this warning, **statement (2.147) should not discourage the student from using an associated differential-form DE to help solve a derivative-form DE.** In fact, to become good at solving first-order DEs, it is essential that you develop facility in passing back and forth between the two types of equations. The "autopilot" procedure is not worthless; it's simply not perfect. The behavior seen in Example 2.66 is rather exceptional. **For "most" continuously differentiable functions M and N** ("most" in a sense that cannot be made precise at the level of these notes), **if a collection $\mathcal{E}$ of equations is an implicit form of the general solution of a DE $M(x,y)dx+N(x,y)dy=0$, the same collection $\mathcal{E}$ will also be an implicit form of the general solution of the associated derivative-form DE $M(x,y)+N(x,y)\frac{dy}{dx}=0$.** In "most" of the exceptions to this rule, we need only delete one or a few of the equations from $\mathcal{E}$ to obtain the general solution of the derivative-form DE (in implicit form).

The simplest of these exceptions are equations that are explicitly of the form "x = some specific constant", or are equivalent to an equation of this form, such as equation (2.146). It is obvious that equations *written this way* are not implicit solutions of a derivative-form DE with x as independent variable, but when a whole *family* of equations is given, such as $x^3 + x + xy^{10} + xy^2 = C$ (equation (2.144)), it may take some work and cleverness to determine whether there are members of this family that are equivalent to "x = specific constant".

The next example involves simpler equations than Example 2.66, but a more complicated "spurious solution".

Example 2.67 Consider the differential-form DE

$$(y^2+1)\cos x\ dx + 2y\sin x\ dy = 0 \tag{2.149}$$

and the associated derivative-form DE

$$(y^2+1)\cos x + 2y\sin x\ \frac{dy}{dx} = 0. \tag{2.150}$$

Equation (2.149) is exact. Its general solution, in implicit form, is

$$(y^2+1)\sin x = C \tag{2.151}$$

where C is an arbitrary constant. For $C \neq 0$, every point (x, y) in the graph of (2.151) has $\sin x \neq 0$, hence $y^2 + 1 = \frac{C}{\sin x}$. As the student may check, the latter equation is an implicit solution of (2.150); the general solution of (2.150), in implicit form, is

$$(y^2+1)\sin x = C, \quad C \neq 0. \tag{2.152}$$

However, for $C = 0$, equation (2.151) is equivalent to $\sin x = 0$, whose graph in $\mathbf{R}^2$ is the infinite collection of vertical lines of the form $x = n\pi$, where n is an integer. None of these vertical lines is the graph (or contains the graph) of a solution of (2.150).

So in this example, we again need to throw away only one equation from the given implicit form (2.151) of the general solution of the differential-form DE in order to get an implicit form of the general solution of the associated derivative-form DE, but the graph of the discarded equation consists of infinitely many inextendible solution *curves* of the differential-form DE. ■

If a collection $\mathcal{E}$ of equations is an implicit form of the general solution of a DE $M(x,y)dx+N(x,y)dy = 0$, then fact (2.139) guarantees that the collection $\mathcal{E}$ *contains* all the solutions of the associated derivative-form DE $M(x,y) + N(x,y)\frac{dy}{dx} = 0$, in implicit form. If we are trying to obtain the general solution of $M(x,y)+N(x,y)\frac{dy}{dx} = 0$ from having solved $M(x,y)dx + N(x,y)dy = 0$, we need only worry whether $\mathcal{E}$ contains any equations that are *not* implicit solutions of the derivative-form equation with x as independent variable.

In general, an algebraic equation (say $F(x,y) = 0$) is an implicit solution of the differential-form DE $M(x,y)dx + N(x,y)dy = 0$ but not the associated derivative-form DE $M(x,y) + N(x,y)\frac{dy}{dx} = 0$ if and only if the graph $\mathcal{G}$ of $F(x,y) = 0$ has both of the following properties:

- $\mathcal{G}$ contains at least one vertical line segment, and
- the *only* smooth curves that $\mathcal{G}$ contains are vertical lines or line segments.

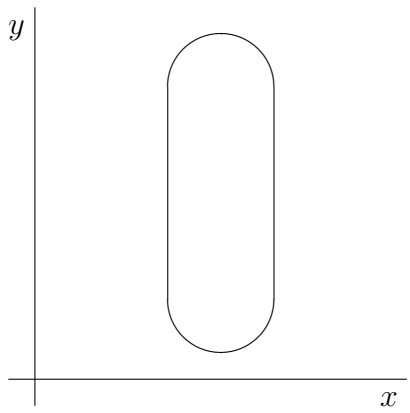

Figure 5:

If we have a collection $\mathcal{E}$ of equations that is an implicit form of the general solution of the differential-form DE, and we remove from this collection all equations whose graphs have the two properties above, then the remaining collection of equations is an implicit form of the general solution of the associated derivative-form DE. "Most of the time", there will be *no* such equations in our original collection $\mathcal{E}$, in which case the same collection $\mathcal{E}$ serves as an implicit form of both the solution of the differential-form DE and the associated derivative-form DE.

It should be noted that even when an algebraic equation, say $F(x, y) = 0$, is an implicit solution of both $M(x, y)dx + N(x, y)dy = 0$ and $M(x, y) + N(x, y)\frac{dy}{dx} = 0$, its graph may contain smooth curves that have vertical segments, and therefore are not solution curves of the derivative-form DE. For example, there is an infinitely differentiable function F (whose formula we will not write down) whose graph is the oval in Figure 5. The entire oval is a solution curve of $\frac{\partial F}{\partial x}dx + \frac{\partial F}{\partial y}dy = 0$, but the vertical line segments in the oval are not contained in graphs of any solutions of $\frac{\partial F}{\partial x} + \frac{\partial F}{\partial y}\ \frac{dy}{dx} = 0$. The equation $F(x, y) = 0$ is still an implicit solution of the derivative-form DE because (i) the graph of $F(x, y) = 0$ contains curves that are graphs of differentiable functions of x (the semicircles at the top and bottom of the oval, with the endpoints of the semicircles deleted), and (ii) all such curves are solutions of the derivative-form DE.

The previous examples in this section focused on problems caused by passing mindlessly between derivative-form and differential-form DEs (Step 1 of the autopilot procedure outlined earlier). The other source of problems in the autopilot procedure is that when carrying out the procedure, we often perform some algebraic manipulations. Sometimes we do these manipulations on the derivative-form DE, prior to writing down an associated differential-form DE; sometimes we do the manipulations on the differential-form DE; and sometimes we do both. The allowed algebraic manipulations of the derivative-form DE are addition/subtraction of a function and multipli-

cation/division by a function; the allowed algebraic manipulations of the differential-form DE are addition/subtraction of a differential and multiplication/division by a function (however, once our differential-form DE is in the form $Mdx + Ndy = 0$, adding/subtracting differentials will take it out of this form). *Any time we perform such a manipulation, we must check whether the new DE is algebraically equivalent to the old one on the entire region of interest.* If algebraic equivalence is not maintained, then there is the potential of either losing solutions or introducing spurious ones.

Now let's try to nail down how to modify the autopilot procedure into one that neither loses solutions nor introduces spurious ones. Suppose we want to solve a standard-form DE

$$\frac{dy}{dx} = f(x, y) \tag{2.153}$$

or, more generally, an "almost-standard form" DE

$$f_1(x, y)\frac{dy}{dx} = f_2(x, y) \tag{2.154}$$

If (2.153) or (2.154) is separable or linear, we can use standard techniques for such equations in order to find the general solution. (For separable equations, the only modification needed for the autopilot procedure is to add to "$\{H(y) = G(x)+C\}$" any constant solutions that the original DE had.) If our starting DE is not separable or linear, we can look at the associated differential-form DE, which for the two equations above would be

$$-f(x, y)dx + dy = 0 \tag{2.155}$$

and

$$-f_2(x, y)dx + f_1(x, y)dy = 0. \tag{2.156}$$

If we are extremely lucky, then (2.155) or (2.156) will be exact.

In the case of (2.155), this virtually never happens: we would need $\frac{\partial f}{\partial y} \equiv 0$. If we are working on a rectangular region R, this condition is equivalent to saying that f is a function of x alone; i.e. $f(x, y) = g(x)$ for some one-variable function g. But then (2.153) was already of the form $\frac{dy}{dx} = g(x)$, solvable just by integrating g; there is no need even to look at equation (2.155).

More commonly, however, our equation $\frac{dy}{dx} = f(x, y)$ or $f_1(x, y)\frac{dy}{dx} = f_2(x, y)$ may be *algebraically equivalent* to a DE whose associated differential-form DE is exact, perhaps just on some region R. (In a best-case scenario, algebraic equivalence and exactness will hold on the whole plane $\mathbf{R}^2$. Usually, however, we will have to restrict attention to a region R that is not all of $\mathbf{R}^2$ to maintain algebraic equivalence. We may have to shrink the region further to achieve exactness.) For the sake of concreteness, let us focus on the case in which our starting equation is the of the form $\frac{dy}{dx} = f(x, y)$;

the principles for working with the more general $f_1(x,y)\frac{dy}{dx} = f_2(x,y)$ are essentially identical.

The derivative-form equation $\frac{dy}{dx} = f(x,y)$ is algebraically equivalent on R to an exact DE on R if and only if the differential-form equation $-f(x,y)dx + dy$ is algebraically equivalent on R to an exact DE on R. To make use of this fact, we relate the equation $\frac{dy}{dx} = f(x,y)$ to a differential-form DE by a two-step process—one step of which is algebraic manipulation of the DE (this may involve several sub-steps, in each of which we keep track of the algebraic-equivalence issue), and the other of which is the passage from a derivative-form DE to the associated differential-form DE—hoping to arrive at an exact DE. The order in which we do these steps and sub-steps does not matter. For example, if we start with the equation $\frac{dy}{dx} = \frac{2y^3 \sin x \cos x}{3y^2 \cos^2 x + 1}$, we could go through the procedure

$$\frac{dy}{dx} = \frac{2y^3 \sin x \cos x}{3y^2 \cos^2 x + 1}$$

$$\downarrow \text{ multiply by } 3y^2 \cos^2 x + 1 \text{ (this yields an algebraically equivalent DE on } \mathbf{R}^2 \text{ since } 3y^2 \cos^2 x + 1 \text{ is nowhere zero)}$$

$$(3y^2 \cos^2 x + 1)\frac{dy}{dx} = 2y^3 \sin x \cos x$$

$$\downarrow \text{ subtract } 2y^3 \sin x \cos x \text{ (yielding an algebraically equivalent DE)}$$

$$-2y^3 \sin x \cos x + (3y^2 \cos^2 x + 1)\frac{dy}{dx} = 0$$

$$\downarrow \text{ write the associated differential-form DE}$$

$$-2y^3 \sin x \cos x \; dx + (3y^2 \cos^2 x + 1)dy = 0,$$

or through the procedure

$$\frac{dy}{dx} = \frac{2y^3 \sin x \cos x}{3y^2 \cos^2 x+1}$$

$$\downarrow \qquad \begin{array}{l}\text{subtract } \frac{2y^3 \sin x \cos x}{3y^2 \cos^2 x+1} \\ \text{(yields an algebraically equivalent equation)}\end{array}$$

$$-\frac{2y^3 \sin x \cos x}{3y^2 \cos^2 x+1} + \frac{dy}{dx} = 0,$$

$$\downarrow \qquad \text{write the associated differential-form DE}$$

$$-\frac{2y^3 \sin x \cos x}{3y^2 \cos^2 x+1} dx + dy = 0,$$

$$\downarrow \qquad \text{multiply by } 3y^2 \cos^2 x + 1$$

$$-2y^3 \sin x \cos x \; dx + (3y^2 \cos^2 x + 1) dy = 0.$$

Whichever procedure we use, we end up with the same differential-form DE. As the student may check, this last DE is exact on $\mathbf{R}^2$, so we may find its general solution by our standard exact-equation method. Depending on how we choose to integrate $\sin x \cos x$, there are several different implicit forms in which we could choose to write the general solution, one of which is

$$\left\{ y + y^3 \cos^2 x = C \right\} \tag{2.157}$$

(Note: "obvious" manipulations such as clearing fractions, plus writing down the associated differential-form DE, will not always lead to an exact DE. The author contrived the current example so that the technique above *would* lead to an exact equation, in order to illustrate further the relation between derivative and differential form. Your textbook probably has similarly contrived examples and homework exercises, in order to give you practice with the techniques you are learning.) But what relation do the solutions of the equation $-2y^3 \sin x \cos x + (3y^2 \cos^2 x + 1)\frac{dy}{dx} = 0$ bear to the solutions of our original derivative-form DE?

Fact (2.139) guarantees us that any implicit form of the general solution of $-2y^3 \sin x \cos x \; dx + (3y^2 \cos^2 x + 1) dy = 0$—in particular, the family of equations (2.157)—*contains* a general solution, in implicit form, of the derivative-form equation $-2y^3 \sin x \cos x \; + (3y^2 \cos^2 x + 1)\frac{dy}{dx} = 0$. This derivative-form equation is algebraically equivalent to the DE we started with, $\frac{dy}{dx} = \frac{2y^3 \sin x \cos x}{3y^2 \cos^2 x+1}$, hence has the same solutions. Therefore (2.157) contains a general solution, in implicit form, of our original derivative-form DE. The only question is whether the family (2.157) contains "spurious solutions"—equations that are implicit solutions of $-2y^3 \sin x \cos x \; dx + (3y^2 \cos^2 x + 1) dy = 0$, but not of $\frac{dy}{dx} = \frac{2y^3 \sin x \cos x}{3y^2 \cos^2 x+1}$ (equivalently, not of $-\frac{2y^3 \sin x \cos x}{3y^2 \cos^2 x+1} + \frac{dy}{dx} = 0$). We have seen that the graph $\mathcal{G}$ of a spurious solution must contain a vertical line segment, i.e. a set of the form $\{(x_0, y) \mid y \in J\}$ where x_0 is a constant

and J is some interval over which y may vary. But it is easily seen that none of the equations (2.157) has such a graph[61]. Therefore (2.157) is the general solution of the derivative-form equation that we started with, $\frac{dy}{dx} = \frac{2y^3 \sin x \cos x}{3y^2 \cos^2 x + 1}$.

So, we may use differential-form DEs to help us find solutions of derivative-form DEs that are in almost-standard form, or are algebraically equivalent to a DE in almost-standard form, as follows:

1. Perform any algebraic manipulations that may be necessary to put the DE into "almost-standard" form $f_1(x,y)\frac{dy}{dx} = f_2(x,y)$ or $-f_2(x,y) + f_1(x,y)\frac{dy}{dx} = 0$. Each time we perform an algebraic manipulation, keep track of the region(s) on which the manipulation gives us an algebraically equivalent DE.

2. Write down the differential-form DE associated with our last derivative-form DE. If this DE does not pass the test for exactness, look for additional algebraic manipulations that may yield an exact DE (we may not find any). Again, keep track of the region(s) on which any algebraic manipulations we use give us an algebraically equivalent DE.

3. Assuming we have now produced an exact DE on some region(s) $R_1, R_2, \ldots$, find the general solution of that DE on each R_i, in implicit form. This will be a collection $\mathcal{E}_i$ of equations of the form $F_i(x,y) = C$ on R_i, where C is a "semi-arbitrary" constant as discussed earlier in these notes. Amalgamate all the collections $\mathcal{E}_i$—hopefully there will only be one or two—into one large collection $\mathcal{E}$ (which may take several lines to write down if there is more than one region R_i).

4. Discard from $\mathcal{E}$ any spurious solutions—those equations whose graphs contain a vertical line segment, and contain no smooth curves except vertical lines or line segments. The collection $\mathcal{E}'$ of equations that remain is the general solution of the original derivative-form DE, in implicit form, on the union of the regions R_i.

5. If any of the algebraic manipulations used above did not preserve algebraic equivalence on the region (or union of regions) R on which we were interested in the original differential equation, check whether these manipulations may have resulted in the loss of solutions or the inclusion of spurious solutions. Adjust $\mathcal{E}'$ accordingly.

[61] One argument is as follows. Suppose that the graph of $y + y^3 \cos^2 x = c_0$ contained a vertical line segment $\{(x_0, y) \mid y \in J\}$. Then for all $y \in J$ we would have $y + y^3 \cos^2 x_0 = c_0$. Differentiating with respect to y, we would have $1 + 3y^2 \cos^2 x_0 = 0$ for all $y \in J$. But this is impossible, since $1 + 3y^2 \cos^2 x_0 \geq 1$.

The last step in the procedure above is not one for which we will try to state general rules; instead, we will illustrate with an example the sort of work that must be done.

Example 2.68 Solve the differential equation

$$\frac{dy}{dx} = -\frac{2x+2y}{2x+3y^2} \ . \tag{2.158}$$

First we observe that since the right-hand side of (2.158) is not defined when $2x + 3y^2 = 0$, the only regions in which "solution of (2.158)" has any meaning are $R_1 = \{(x, y) \mid 2x + 3y^2 > 0\}$ and $R_2 = \{(x, y) \mid 2x + 3y^2 < 0\}$. On each of these regions, (2.158) is algebraically equivalent to

$$(2x + 2y) + (2x + 3y^2)\frac{dy}{dx} = 0, \tag{2.159}$$

whose associated differential-form equation is

$$(2x + 2y)dx + (2x + 3y^2)dy = 0. \tag{2.160}$$

Equation (2.160) is exact on the whole plane $\mathbf{R}^2$; its left-hand side is dF, where $F(x, y) = x^2 + 2xy + y^3$. Thus the general solution of (2.160) is $x^2 + 2xy + y^3 = C$. We will see shortly that in this example C can be arbitrary, but we do not need that fact yet.

Every solution of (2.158) is guaranteed to be a solution of (2.159), so in passing from (2.158) to (2.159) we have not lost any solutions; the only question is whether we have introduced spurious solutions. We must also check whether we introduced spurious solutions when passing from (2.159) to (2.160). The latter possibility is easy to rule out: it is easy to see that (2.160) has no solutions of the form $x =$ constant. (If $x = c$ were a solution, then we could use y as a parameter for a parametric solution, yielding $(2c+2y) \times 0 + (2c+3y^2)\frac{dy}{dy} = 0 = 2c+3y^2$, impossible since the parameter y must range over an interval.) Thus every solution curve of (2.160) is a solution curve of (2.159)

To see whether the graph of $x^2 + 2xy + y^3 = C$, for a given C, is an implicit solution of (2.158) on R_1 (or R_2) we must check whether its graph contains a smooth curve in this region. First let us consider the allowed values of C. The only critical point of F is the origin, so fact (2.114) assures us that the general solution of (2.160) on $\{\mathbf{R}^2$ minus the origin$\}$ is $x^2 + 2xy + y^3 = C$, where C can be any value in the range of F on this domain. By holding x fixed (say $x = 1$) and letting y vary over $\mathbf{R}$, we see that the range of F on this domain is the set of all real numbers. Therefore the general solution of (2.160) in $\{\mathbf{R}^2$ minus the origin$\}$ is $x^2 + 2xy + y^3 = C$, where C is arbitrary.

Now we must check whether multiplying by $2x + 3y^2$ in passing from (2.158) to (2.159) introduced any spurious solutions: equations $x^2 + 2xy + y^3 = C$ that are not

implicit solutions of (2.158). For this, we must check whether for some C, the graph of $x^2+2xy+y^3 = C$ fails to contain a smooth curve lying in R_1 or R_2. But (for any C), the points of the the graph of $x^2+2xy+y^3 = C$ not lying in R_1 or R_2 lie on the graph of $2x+3y^2 = 0$. But the graph of $x^2+2xy+y^3 = C$ intersects the graph of $2x+3y^2 = 0$ only at those points (x, y) for which $x = -\frac{3}{2}y^2$ and $\left(-\frac{3}{2}y^2\right)^2 + 2\left(-\frac{3}{2}y^2\right) + y^3 = C$, the latter equation simplifying to $\frac{9}{4}y^4 - 2y^3 = C$. No matter what the value of C is, there are at most four numbers y for which $\frac{9}{4}y^4 - 2y^3 = C$, so the graph of $2x + 3y^2 = 0$ intersects the graph of $x^2 + 2xy + y^3 = C$ in at most four points. But the portion of the graph of $x^2 + 2xy + y^3 = C$ that lies in $\{\mathbf{R}^2$ minus the origin$\}$— the whole graph unless $C = 0$—is a smooth curve $\mathcal{C}$. Deleting from $\mathcal{C}$ the at-most-four points of $\mathcal{C}$ for which $2x + 3y^2 = 0$, what remains is one or more curves each of which lies entirely in R_1 or R_2, and hence is a solution-curve of (2.158). Therefore there are no values of C that we need to exclude, and no spurious solutions. The general solution of (2.158) is $\{x^2 + 2xy + y^3 = C \mid C \in \mathbf{R}, 2x + 3y^2 \neq 0\}$. (Writing the "$2x + 3y^2 \neq 0$" explicitly is optional, since that constraint is imposed from the moment we write down (2.158).) ■

In the preceding example, the passage from (2.158) to (2.160) did not introduce any spurious solutions. In the earlier examples, whenever spurious solutions were introduced, they were of the form $x =$ constant. So it is natural to ask whether, starting with an "almost-standard" derivative-form DE $f_1(x, y)\frac{dy}{dx} = f_2(x, y)$ or $-f_2(x, y) + f_1(x, y)\frac{dy}{dx} = 0$, algebraic manipulations can ever introduce spurious solutions that are *not* of the form $x =$ constant. The answer is yes. Failure to preserve algebraic equivalence can lead to spurious solutions not of the form "one variable = constant" whether we are working with derivative-form or differential-form DEs. The next example could have been presented before we ever talked about differential form, but we have placed it in this section of the notes as a reminder.

Example 2.69 (A spurious solution not of the form $x =$ constant) Let

$$f(x, y) = \begin{cases} \frac{e^y - e^x}{y - x} & \text{if } y \neq x, \\ e^x & \text{if } y = x. \end{cases}$$

It can be shown that this function is continuously differentiable on the whole xy plane. (The student should be able to show at least that f is continuous everywhere, including at points of the line $\{y = x\}$.) Therefore, for every initial condition $y(x_0) = y_0$, the corresponding initial-value problem for the DE $\frac{dy}{dx} = f(x, y)$ has a unique solution. In particular, this is true when $y_0 = x_0$. So for every $x_0 \in \mathbf{R}$, the initial-value problem

$$\frac{dy}{dx} = f(x, y), \quad y(0) = 0 \tag{2.161}$$

has a unique maximal solution.

If we substitute the definition of $f(x,y)$ into (2.161), the DE becomes

$$\frac{dy}{dx} = \begin{cases} \frac{e^y - e^x}{y-x} & \text{if } y \neq x, \\ e^x & \text{if } y = x. \end{cases} \tag{2.162}$$

This equation is neither linear nor separable, so in an attempt to solve we might write down the associated differential-form equation, which is

$$-\left\{ \begin{array}{ll} \frac{e^y - e^x}{y-x} & \text{if } y \neq x \\ e^x & \text{if } y = x \end{array} \right\} dx + dy = 0. \tag{2.163}$$

It is natural to try to rewrite (2.163) more simply by multiplying through by $y - x$. Observing that $(y-x)f(x,y) = e^y - e^x$ for all $(x,y) \in \mathbf{R}^2$ (even for those points with $y = x$), if we multiply both sides of (2.163) by $y - x$ we obtain

$$-(e^y - e^x)dx + (y-x)dy = 0, \tag{2.164}$$

which certainly *looks* much simpler than (2.163). This DE is not exact, and the student will not succeed in solving it—i.e. finding *all* solutions—by any method taught in an introductory DE course. However, *one* solution is obvious: $y = x$. This solution also satisfies the initial condition $y(0) = 0$. Does this mean that $y = x$ is the solution of the IVP (2.161)?

The answer is a resounding "No!". If we define $\phi(x) = x$, and substitute $y = \phi(x)$ into "$\frac{dy}{dx} = f(x,y)$", then the left-hand side is identically 1, while the right-hand side is e^x. There is no x-interval on which $e^x \equiv 1$. The function ϕ is not a solution of $\frac{dy}{dx} = f(x,y)$.

It is easy to see what went wrong if, instead of writing (2.163) with the two-line formula for f, we write it as

$$-f(x,y)dx + dy = 0 \tag{2.165}$$

and if, when we multiply through by $y - x$, we write the result as

$$-(y-x)f(x,y)dx + (y-x)dy = 0 \tag{2.166}$$

rather than in the "simpler" form (2.164). It is obvious that $y = x$ is a solution of (2.166), whether or not it is a solution of (2.165). Less obvious, but true, is what we checked above: that $y = x$ is *definitely not* a solution of (2.161), hence not a solution of (2.165).

In this example, the general solution of (2.166) consists of the general solution of (2.165) *plus* the straight line $\{y = x\}$. The equation (2.165) has no solutions of the form $x =$ constant, so any implicit form of the general solution of (2.165) is also an implicit form of the general solution of $\frac{dy}{dx} = f(x,y)$. Thus, in passing from $\frac{dy}{dx} = f(x,y)$ to the differential-form equation (2.164), we gained a spurious solution $y = x$ that is not a solution of the DE we started with.

In this instance, it was not the transition from derivative form to differential form that introduced the spurious solution; it was multiplication by the function $y - x$, which is zero at lots of points. The equations (2.163) and (2.164) are algebraically equivalent on the region $R_1 = \{(x, y) \mid y > x\}$, and also on the region $R_2 = \{(x, y) \mid y < x\}$. On each of these regions, the two equations have the same general solution. But they are not algebraically equivalent on the whole xy plane, and their general solutions on the whole xy plane are different. ■

2.10 Using derivative-form equations to help solve differential-form equations

Not yet written.

3 Optional Reading

3.1 The meaning of a differential

Now we are ready to ascribe meaning to a differential.[62] However, don't worry if you don't understand the meaning given below. Understanding it is not essential to the use of differentials in differential equations. In fact, in this section of the notes, there are no differential *equations*—just differentials.

A differential $Mdx + Ndy$ is a machine with an input and an output. What it takes as input is a (differentiably) parametrized curve γ. What it then outputs is a *function*, defined on the same interval I as γ. If we write $\gamma(t) = (x(t), y(t))$, then the output is the function whose value at $t \in I$ is $M(x(t), y(t))\frac{dx}{dt} + N(x(t), y(t))\frac{dy}{dt}$.

We use the language "$Mdx + Ndy$ *acts on* γ" to refer to the fact that the differential takes γ as an input and then "processes" it to produce some output. Notation we will use for the output function is $(Mdx + Ndy)[\gamma]$. This is the same function that we expressed in terms of t in the previous paragraph:

[62] Differentials can be understood at different levels of loftiness. The level chosen for these notes is a higher than in Calculus 1-2-3 and introductory DE textbooks, but it is not the highest level.

$$\overbrace{\underbrace{(M\,dx + N\,dy)[\gamma]}_{\substack{\text{value of the function}\\ (M\,dx+N\,dy)[\gamma]\\ \text{at } t}}\ (t)}^{\substack{\text{the function obtained}\\ \text{when the differential}\\ \text{acts on } \gamma}} = M(x(t), y(t))\frac{dx}{dt} + N(x(t), y(t))\frac{dy}{dt}\ . \tag{3.1}$$

The notation on the left-hand side of (3.1) may look intimidating and unwieldy, but it (or something like it) is a necessary evil for this section of the notes. It will not be used much outside this section.

Let us make contact between the meaning of differential given above, and what the student may have seen about differentials before. The easiest link is to differentials that arise as *notation* in the context of line integrals in Calculus 3. (Students who haven't completed Calculus 3 should skip down to the paragraph that includes equation (3.5), read that paragraph, and skip the rest of this section.) Recall that one notation for the line integral of a vector field $M(x, y)\mathbf{i} + N(x, y)\mathbf{j}$ over a smooth, oriented curve $\mathcal{C}$ in the xy plane is

$$\int_C M(x, y)dx + N(x, y)dy. \tag{3.2}$$

To see that the integrand in (3.2) is the same gadget we described above, let's review the rules you learned for computing such an integral:

1. Choose a continuously differentiable, nonstop parametrization γ of $\mathcal{C}$. Write this as $\gamma(t) = (x(t), y(t))$, $t \in [a, b]$.[63] Depending on your teacher and textbook, you may or may not have been introduced to using a single letter, such as γ or $\mathbf{r}$, for the parametrization. But almost certainly, one ingredient of the notation you used was "$(x(t), y(t))$".

2. In (3.2), make the following substitutions: $x = x(t), y = y(t), dx = \frac{dx}{dt}dt, dy = \frac{dy}{dt}dt$, and $\int_C = \int_a^b$. The integral obtained from these substitutions is

$$\int_a^b \left\{ M(x(t), y(t))\frac{dx}{dt} + N(x(t), y(t))\frac{dy}{dt} \right\} dt. \tag{3.3}$$

[63]The parametrization should also consistent with the given orientation of $\mathcal{C}$, and to be one-to-one, except that "$\gamma(a) = \gamma(b)$" is allowed in order to handle closed curves. These technicalities is unimportant here; the author is trying only to jog the student's memory, not to review line integrals thoroughly.

3. Compute the integral (3.3). The definition of (3.2) is the value of (3.3):

$$\int_{\mathcal{C}} M(x,y)dx + N(x,y)dy = \int_a^b \left\{ M(x(t),y(t))\frac{dx}{dt} + N(x(t),y(t))\frac{dy}{dt} \right\} dt. \tag{3.4}$$

(You also learn in Calculus 3 that this definition is self-consistent: no matter what continuously differentiable, non-stop parametrization of $\mathcal{C}$ you choose[64], you get the same answer.)

A casual glance at (3.4) suggests that we have used the following misleading equality:

$$\text{“}M(x,y)dx + N(x,y)dy = \left\{ M(x(t),y(t))\frac{dx}{dt} + N(x(t),y(t))\frac{dy}{dt} \right\} dt.\text{”} \tag{3.5}$$

But that is not quite right. The left-hand side and right-hand side are not the same object. Only *after we are given a parametrized curve* γ can we produce, from the object on the left-hand side, the function of t in braces on the right-hand side.

In addition, in constructing the integral on the right-hand side of (3.4), we did not confine our substitutions to the *integrand* of the integral on the left-hand side. We made the substitution “$\int_{\mathcal{C}} \to \int_a^b$ ” as well. Attempting to equate *pieces* of the notation on the left-hand side with *pieces* of the notation on the right-hand side helps lead to a wrong impression of what is equal to what. Instead of making this fallacious attempt, understand that (3.4) is simply a definition of the whole left-hand side. The data on the left-hand side are reflected in the computational prescription on the right-hand side as follows:

1. The right-hand side involves functions $x(t), y(t)$ on a t-interval $[a,b]$. These two functions and the interval $[a,b]$ give us a parametrized curve γ, defined by $\gamma(t) = (x(t), y(t))$. The curve $\mathcal{C}$ on the left-hand side tells us which γ's are allowed: only those having image $\mathcal{C}$.

2. Once we choose such a γ, what is the integrand on the right-hand side? It is exactly the quantity $(Mdx+Ndy)[\gamma](t)$ in (3.1). The effect of the “$M(x,y)dx+N(x,y)dy$” on the left-hand side has been to produce the function $(Mdx + Ndy)[\gamma]$ when fed the parametrized curve γ.

Thus, the differential that appears as the integrand on the left-hand side is exactly the machine we described at the start of this section.

[64]Subject to the other conditions in the previous footnote.

There is one other topic in Calculus 3 that makes reference to differentials (if the instructor chooses to discuss them at that time): the tangent-plane approximation of a function of two variables. The differentials you learned about in that context are not quite the same gadgets as the machines we have defined. They are related, but different. To demonstrate the precise relation, there are two things we would need to do: (1) restrict attention to exact differentials, and (2) discuss what kind of gadget the *value of a differential at a point*—an expression of the form $M(x_0, y_0)dx + N(x_0, y_0)dy$—is. This would require a digression that we omit, in the interests of both brevity and comprehensibility.

3.2 Exact equations: further exploration

Example 3.1 In the setting of Example 2.60, assume that $Mdx+Ndy$ has no singular points (equivalently, F has no critical points) in R. We claim that in this case, the general solution of (2.104) on R, in implicit form, is (2.110), but where the allowed values of C are those for which the graph of (2.110) contains even a single *point* of R. Equivalently, *the set of allowed values of C is the range of F on the domain R.*

To see that this is the case, it suffices to show that if, for a given C, the graph of (2.110) contains a point (x_0, y_0) of R, then the graph contains a smooth curve in R. So, with C held fixed, assume there is such a point (x_0, y_0). Remember that, by definition of "exact", the functions $\frac{\partial G}{\partial x}, \frac{\partial G}{\partial y}$ are continuous on R. Since we are assuming that F has no critical points in R, the point (x_0, y_0) is not a critical point of F, so at least one of the partial derivatives $\frac{\partial G}{\partial x}(x_0, y_0), \frac{\partial G}{\partial y}(x_0, y_0)$ is not zero. Then:

- If $\frac{\partial G}{\partial y}(x_0, y_0) \neq 0$, then, since we are assuming that $\frac{\partial G}{\partial x}$ and $\frac{\partial G}{\partial y}$ are continuous on R, we can apply the Implicit Function Theorem (Theorem 2.5) to deduce that is an open rectangle $I_1 \times J_1$ containing (x_0, y_0), and a continuously differentiable function ϕ with domain I_1 such that the portion of the graph of (2.108) contained in $I_1 \times J_1$ is the graph of $y = \phi(x)$, i.e. the set of points $\{(x, \phi(x)) \mid x \in I_1\}$. This same set is the image of the parametrized curve given by

$$\left\{ \begin{array}{l} x(t) = t \\ y(t) = \phi(t) \end{array} \right\}, \ t \in I_1.$$

  This parametrized curve γ is continuously differentiable, and it is non-stop since $\frac{dx}{dt} = 1$ for all $t \in I_1$. Hence the image of γ is a smooth curve contained in the graph of (2.110). Since $(x_0, y_0) \in R$, and R is an open set, a small enough segment of this curve, passing through (x_0, y_0), will be contained in R.

- If $\frac{\partial G}{\partial x}(x_0, y_0) \neq 0$, then (reversing the roles of x and y in the Theorem—e.g. by defining $\tilde{G}(x, y) = F(y, x)$), the Implicit Function Theorem tells us that there is an open rectangle $I_1 \times J_1$ containing (x_0, y_0), and a continuously differentiable function ϕ with domain J_1 such that the portion of the graph of

(2.108) contained in $I_1 \times J_1$ is the graph of $x = \phi(y)$, i.e. the set of points $\{(\phi(y), y) \mid y \in J_1\}$. This graph is exactly the image of the parametrized curve γ given by

$$\left\{ \begin{array}{l} x(t) = \phi(t) \\ y(t) = t \end{array} \right\}, \ t \in J_1.$$

As in the previous case, γ is continuously differentiable and non-stop. Hence the image of γ is again a smooth curve contained in the graph of (2.110), and again a small enough segment of it, passing through (x_0, y_0), will be contained in R. ■

Example 3.2 Consider again the DE

$$xdx + ydy = 0. \tag{3.6}$$

Defining $F(x, y) = \frac{1}{2}(x^2 + y^2)$ (on the whole plane $\mathbf{R}^2$), the left-hand side of (3.6) is the exact differential dF. The function F has only one critical point, $(0, 0)$, and the functions $M(x, y) = x$ and $N(x, y) = y$ are continuous on the whole xy plane. So if we let $R = \{\mathbf{R}^2$ minus the origin$\}$, F has no critical points in R, and Example 3.1 applies. The range of F on R is the set of positive real numbers, which for the sake of Definition 2.61, we view as $\{C \in \mathbf{R} \mid C > 0\}$. Therefore the general solution of $xdx + ydy = 0$ in R is $\left\{\frac{1}{2}(x^2 + y^2) = C \mid \ C > 0\right\}$, which, by renaming the constant, we can write more simply as

$$\left\{x^2 + y^2 = C \mid C > 0\right\}. \tag{3.7}$$

The graph of each solution is a circle. The collection of these circles is what we call the general solution of (3.6) in R (according to Definition 2.61), and the general solution in R "fills out" the region R (every point of R lies on the graph of $x^2 + y^2 = C$ for some $C > 0$).

If we look at (3.6) on the whole xy plane rather than just R, then Example 3.1 no longer applies (because of the critical point at the origin), but Example 2.60 still applies. From the analysis above, every point of the xy plane other than the origin lies on a solution curve with equation $x^2 + y^2 = C$ with $C > 0$. For $C = 0$, the equation "$F(x, y) = C$" becomes $x^2 + y^2 = 0$. The graph of this equation is the single point $(0, 0)$, and contains no smooth curves. For $C < 0$, the graph of $x^2 + y^2 = C$ is empty. Hence the general solution of (3.6) in implicit form, with no restriction on the region, is the same as the general solution on R in implicit form, namely (3.7).
■

Example 3.3 Consider again the DE from Example 2.57,

$$ydx + xdy = 0. \tag{3.8}$$

The left-hand side is the exact differential dF (on the whole plane $\mathbf{R}^2$), where $F(x, y) = xy$. The function F has only one critical point, $(0, 0)$, and the functions $M(x, y) = y$ and $N(x, y) = x$ are continuous on the whole xy plane. So, as in the previous example if we let $R = \{\mathbf{R}^2$ minus the origin$\}$, there are no critical points in R, and Example 3.1 applies. This time, for every $C \in \mathbf{R}$ there is a point in R for which $xy = C$. Therefore the general solution of $ydx + xdy = 0$ in R is

$$xy = C, \tag{3.9}$$

where C is a "true" arbitrary constant—every real value of C is allowed.

Note that for $C \neq 0$, the graph of $xy = C$ consists of two solution curves (the two halves of a hyperbola) in R. For $C = 0$, there are four solution curves in R: the positive x-axis, the negative x-axis, the positive y-axis, and the negative y-axis. The set of solution-curves in R again fills out R.

If we look at (3.8) on the whole xy plane rather than just R, then from the preceding, the only point we do not yet know to be on a solution curve is the origin. But, as we saw in Example 2.57, the origin *is* on a solution curve; in fact it is on two of them: the x-axis and the y-axis. So the general solution (without the words "in implicit form", and with no restriction on the region) is the set of the half-hyperbolas noted above, plus the x-axis and the y-axis. The general solution of (3.8), with no restriction on the region, is again (3.9). But in contrast to Example 3.2, this time the general solution fills out the whole plane $\mathbf{R}^2$. ■

Students who've taken Calculus 3 have studied equations that are explicitly of the form "$F(x, y) = C$" before. For a given constant C and function F, the graph of $F(x, y) = C$ is called a **level-set** of F. (Your calculus textbook may have used the term "level curve" for a level-set of a function of two variables, because most of the time—though not always—a non-empty level-set of a function of two variables is a smooth curve or a union of smooth curves.[65]) A level-set may have more than one

[65] *Note to students.* This is true provided that the second partial derivatives of the function exist and are continuous on the domain of F. The definition of "most of the time" is beyond the scope of these notes. However, one instance of "most of the time" is the case in which there are only finitely many C's for which the graph of $F(x, y) = C$ is a non-empty set that is not a union of one or more smooth curves. For example, for the equation $x^2 + y^2 = C$, only for $C = 0$ is the graph both non-empty and not a smooth curve.

Note to instructors: The "most of the time" statement is a combination of the Regular Value Theorem and Sard's Theorem for the case of a C^2 real-valued function F on a two-dimensional domain. The Regular Value Theorem asserts that if C is not a critical value of F (i.e. if $F^{-1}(C)$ contains no critical points), then $F^{-1}(C)$ is a submanifold of the domain, which for the dimensions

connected component, such as the graph of $xy = 1$: there is no way to move along the portion of this hyperbola in the first quadrant, and reach the portion of the hyperbola in the third quadrant. Our definition of "smooth curve" prevents any level-set with more than one connected component from being called a smooth curve. However, it is often the case that a level-set is the union of several connected components, each of which is a smooth curve. From Examples 2.60 and 3.1 we can deduce the following:

$$\left.\begin{array}{r}\textbf{If } F \textbf{ has continuous second partial derivatives in the region } \\ R\textbf{, then the set of solution curves of } dF = 0 \textbf{ on } R \textbf{ is the} \\ \textbf{set of smooth curves in } R \textbf{ that are contained in level-sets of } F\textbf{.}\end{array}\right\} \tag{3.10}$$

Statement (3.10) is not an "if and only if". For example, the function $F(x, y) = xy$ has a critical point at the origin, but the general solution of $dF = 0$ is still the set of smooth curves in $\mathbf{R}^2$ that are contained in level-sets of F. (One of these smooth curves is the x-axis, one is the y-axis, and the others are half-hyperbolas.) For an example of a level-set that contains smooth curves, but is not a union of smooth curves (i.e. has a point that's not contained in any of the smooth curves in the level-set), see Example 2.62 elsewhere in these notes.

4 Appendix

4.1 The Fundamental Theorem of ODEs

The "Fundamental Theorem of ODEs" ("FTODE") is the theorem asserting that, under certain rather general conditions, an initial-value problem has a unique solution. The first-order case is the theorem below. In this theorem, "$\frac{\partial f}{\partial y}$" denotes the partial derivative of f with respect to its second variable.

Theorem 4.1 (FTODE) *Let f be a function of two variables, and consider the initial-value problem*

$$\frac{dy}{dx} = f(x, y), \quad y(x_0) = y_0\,. \tag{4.1}$$

Assume that f and $\frac{\partial f}{\partial y}$ are continuous on some rectangle $\{(x, y) : a < x < b, c < x < d\}$ containing the point (x_0, y_0). Then there exists a number $\delta > 0$ such that for every interval I contained in $(x_0 - \delta, x_0 + \delta)$, and containing x_0, the initial-value problem (4.1) *has a unique solution on I.*

Most textbooks (including [1], [3], and [4]), state a version of this theorem that is far too weak to be useful, effectively replacing the last sentence with, "Then there

involved here means "empty or a union of smooth curves". Sard's Theorem asserts that the set of critical *values* (not critical *points*!) of F has measure zero.

exists a number $\delta > 0$ such that the initial-value problem (4.1) has a unique solution on $(x_0 - \delta, x_0 + \delta)$." This weaker statement allows for the possibility that (4.1) has a unique solution on $(x_0 - \delta, x_0 + \delta)$, but has more than one solution on a smaller interval, e.g. $(x_0 - \frac{\delta}{2}, x_0 + \frac{\delta}{2})$, a phenomenon ruled out by the more-carefully stated theorem above.[66] Textbooks that state the weaker theorem sometimes implicitly *use* Theorem 4.1, without observing it is not implied by the weaker version.

References

[1] W.E. Boyce and R.C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, 4th edition, John Wiley & Sons, 1986.

[2] L.H. Loomis and S. Sternberg, *Advanced Calculus*, Addison-Wesley, 1968.

[3] R.K. Nagle, E.B. Saff, and A.D. Snider, *Fundamentals of Differential Equations*, 8th edition, Addison-Wesley, 2012.

[4] E.D. Rainville and P.E. Bedient, *A Short Course in Differential Equations*, 5th edition, Macmillan Publishing Co., 1974.

[66] *Note to instructors*: From Theorem 4.1, but not from the weaker version, the following can be deduced: If I_1, I_2 are intervals contained in (a, b) and containing x_0, and ϕ_1 and ϕ_2 are solutions of (4.1) on I_1 and I_2, respectively, then ϕ_1 and ϕ_2 agree on $I_1 \bigcap I_2$. This fact is of critical importance to the notion of "maximal domain of a unique solution". The only textbook I've looked at recently that states this useful a form of the FTODE is [2].