Some notes on first-order ODEs

version date: 10/10/2024

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

Contents

1 Introduction

First-order ODEs seen in an introductory course come in two forms: *derivative form* and *differential form*. Some textbooks *use* differential-form DEs without ever defining them at all; the student is given the impression that differential-form DEs and derivative-form DEs are simply different ways of writing the same thing. *They are not.*[1]

The two forms are closely related, but differ in subtle ways not addressed adequately in most textbooks (and often overlooked entirely)[2]. In particular, the essential nature of what constitutes a *solution* is different for the two forms. Even just for derivative-form equations, the definition and concept of what a *solution of a differential equation* is—arguably the most fundamental concept in the study of ODEs—has, in my opinion, become increasingly muddled in recent editions of introductory DE textbooks.[3] The confusion may have started with a well-intentioned effort to define the term "implicit solution"—a term that is truly unnecessary, but seems now to be so widely used that there ought at least to be a good definition of it. Unfortunately, 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 the original version of these notes 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.[4] As the writing went along, I became aware of more deficiencies in the textbook (a department-wide adoption)

[1] *Note to instructors*: In fact, a "differential-form DE" is not a true differential equation at all; it has no distinct independent or dependent variable. What a "differential-form DE" *is*, is the simplest example of what differential geometers call a *differential system*.

[2] 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 one important difference between the two types, but there are important differences even between standard-form derivative-form DEs and differential-form DEs.

[3] These notes contain numerous opinions of mine, but henceforth, qualifiers like "in my opinion" are mostly left implicit to avoid tortured writing.

[4] (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.

I was teaching from at that time, which led me to add more topics and examples. Then, newer editions of the textbook came along that had even more deficiencies and inaccuracies than the edition used in 2010, leading me to rewrite whole sections and to add others. This has made for a rather lengthy, never-quite-finished set of notes, an ongoing project that I work 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 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 I cover differential-form DEs, I return to derivative-form equations to clean up the picture. In a more efficient presentation (which I hope eventually to achieve in some future version of these notes), I would introduce differential-form DEs before talking about implicit solutions of derivative-form DEs.

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

2 Notes for Instructors

[This section is under construction. However, much of the content intended for this section is in footnotes addressed to instructors in the "Notes for Students" section.]

3 Notes for Students

Throughout these notes, unless otherwise specified,"function" always means "real-valued function defined on a domain that lies in $\mathbf{R}$, or in $\mathbf{R}^n$ for some n." A *function of n (real) variables* is a function whose domain lies in $\mathbf{R}^n$,

3.1 Functions: domains, restrictions, and extensions

There is a difference between *domain of a formula* (or *expression*) and *domain of a function*. A *function* f is given by specifying (i) a domain, and (ii) an assignment of a real number $f(p)$ to each element p of the domain. (I've written p for "point", rather than a letter like x or t that's commonly used for functions of *one* variable, since

the functions under discussion right now may or may not be single-variable functions; I haven't specified the number of variables. If we were talking only about, say, two-variable functions, the domains would lie in $\mathbf{R}^2$, and instead of 'p' I could write an ordered pair of real numbers, e.g. (x, y).) The way in which $f(p)$ is assigned to p is often, *but not always*, given by an explicit formula. When we write, say, the formula $\frac{1}{1-x^2}$, where x is a real variable, the domain of the formula is the set of all real numbers x for which the formula yields another real number, in this case all x except 1 and -1. This is the set often called the *implied domain* of the formula in calculus and precalculus courses.

However, in some situations we want to restrict attention to a smaller domain. For example, if $-1 < x < 1$, then

$$1 + x^2 + x^4 + x^6 + \ldots \;=\; \sum_{n=0}^{\infty} x^{2n} \;=\; \frac{1}{1-x^2}\;, \tag{3.1}$$

but if $|x| \geq 1$ then the series on the left-hand side of equation (3.1) diverges.[5] Thus, if we define a function f only on the domain $(-1, 1)$ (the open interval with endpoints ± 1, "centered" at 0) by $f(x) = \frac{1}{1-x^2}$, then f has a convergent power series expansion centered at 0. If we define a function g whose domain is $\{x \in \mathbf{R} : x \neq \pm 1\}$ by $g(x) = \frac{1}{1-x^2}$, then f is a *restriction* of g; specifically, f is the restriction of g to the interval $(-1, 1)$. But only the function f, not the function g, can be represented on its domain by a power series centered at 0.

The above example illustrates only *one* reason that we might want to restrict a function, defined on some domain, to a smaller domain. Another reason has to do with *inverse functions*. The sine function, for example does not have an inverse, but the *restriction* of sine to the interval $[-\pi/2, \pi/2]$ does; the inverse of this *restricted* function is the function we call $\sin^{-1}$ or arcsine. There are other reasons that we won't go into at this time.

"Opposite" (informally) to the notion of restriction is *extension*. Some times we are given a function f on some domain D, and wish to extend f to a function $\tilde{f}$ on a larger domain $\tilde{D}$ that contains D, without changing any function-values on D (i.e., we want $\tilde{f}$ to have the property that $\tilde{f}(x) = f(x)$ for all $x \in D$). Such a function $\tilde{f}$ is called an *extension* of f. Thus, given two arbitrary functions f and $\tilde{f}$, the function $\tilde{f}$ is an extension of f if and only if f is a restriction of $\tilde{f}$. As an example, in the next-to-last paragraph above, g is an extension of f, and f is a restriction of g.

Note that, in general, a function defined on one domain D will have many (in fact, infinitely many) extensions to any larger domain $\tilde{D}$. For example, for the function f

[5] As you may recall from Calculus 2, there is a special convention for Sigma-notation for power series: the expression "x^0" is interpreted as meaning 1 for all x, including for $x = 0$. This is *not* a definition of 0^0; it is *only* a *convention for Sigma-notation for power series*, without which we would have to write "$\sum_{n=0}^{\infty} x^{2n}$" as "$1 + \sum_{n=1}^{\infty} x^{2n}$."

defined on $[0, \infty)$ by the formula $f(x) = x$, each of the following is an extension of f to the whole real line:

- The function $\tilde{f}_1$ defined by $\tilde{f}_1(x) = x$ for all $x \in \mathbf{R}$.
- The function $\tilde{f}_2$ defined by $\tilde{f}_1(x) = |x|$ for all $x \in \mathbf{R}$.
- The function $\tilde{f}_3$ defined by

$$\tilde{f}_3(x) = \begin{cases} x & \text{if } x \geq 0, \\ x^2 & \text{if } x < 0. \end{cases}$$

- The function $\tilde{f}_4$ defined by

$$\tilde{f}_4(x) = \begin{cases} x & \text{if } x \geq 0 \\ 1 & \text{if } x < 0 \text{ and } x \text{ is rational,} \\ 23 & \text{if } x < 0 \text{ and } x \text{ is irrational.} \end{cases}$$

Usually when we extend a function that has some nice property (e.g. continuity), we want the extended function also to have that nice property, not to be some "random" extension like $\tilde{f}_4$ above. Later in these notes, what will matter to is extending functions that are solutions of a differential equation on some interval (see Definition 3.1, coming up soon), to solutions of the same differential equation on a larger interval.[6]

But one thing that the examples above already show is that a real-valued function f should really be thought of as a *pair* (D, f), where D is the domain. (Mathematicians use the efficient notation "$f : D \to \mathbf{R}$" to emphasize this.) If we change D, we get a *different function* (by definition; see the handout "Sets and Functions"), even if the computation rule for producing $f(x)$ from x is the same.

Rather than give students extra, unfamiliar notation to deal with, I will not use notation of the form "$f : D \to \mathbf{R}$" in these notes (after this sentence!). Instead, I will use wording of the form "a function f, with domain a set D," or "a function f defined on a set D." In the first of these wordings, the student must remember that D need not be the whole domain of a formula used to express f. *A convention for these notes* is that when we use the wording "a function f defined on a set D", we mean that we are treating D as the domain of f, even though the wording literally allows the domain of f to be a larger set that contains D.

[6] An *interval* is a non-empty subset I of $\mathbf{R}$ with the "betweenness property": given any two distinct elements c, d of I, every real number between c and d lies in I. See Section 5.1 for terminology concerning intervals.

3.2 First-order DEs in derivative form

3.2.1 Definition 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.

An *algebraic* equation[7] in variables x and y is an equation of the form

$$F_1(x, y) = F_2(x, y), \tag{3.2}$$

where F_1 and F_2 are functions defined on some domains in $\mathbf{R}^2$. A special case is an equation of the form $F(x, y) = 0$; a more general special case is $F(x, y) = C$, where C is some real number (any constant).[8] Note that (3.2) makes sense only on the *common domain* of F_1 and F_2 (the set of pairs (x, y) for which both $F_1(x, y)$ and $F_2(x, y)$ are defined. On this common domain, equation (3.2) is equivalent to $F_3(x, y) = 0$, where $F_3(x, y) = F_1(x, y) - F_2(x, y)$. Thus, every algebraic equation in x and y can be put in the form $F(x, y) = 0$ (i.e. is equivalent to an equation in this form).

An algebraic equation in two variables is sometimes referred to as a *relation* between the two variables.

For any pair of real numbers (x, y) for which both sides of an algebraic equation in variables x and y are defined, the equation makes a *statement* that is either true or false. When the statement is true, we say that the pair (x, y) *satisfies* the algebraic equation, and call the pair (x, y) a *solution* of that equation. For example, the pairs $(1, 0)$ and $(0, 2)$ satisfy the equation $x^2 + \frac{y^2}{4} = 1$, and (synonymously) are solutions of this equation.[9] Of course, this equation has infinitely many solutions; the set of all solutions is an ellipse in the xy plane.

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

$$\mathsf{G}(x, y, \frac{dy}{dx}) = 0, \tag{3.3}$$

where G is a function of three variables. Such a DE, written in the notation in (3.3), has an *independent variable* (in this case x) and a *dependent variable* (in this case

[7] *Note to instructors*: In these notes, we use the term "algebraic equation" just to distinguish a non-differential equation from a differential equation. My "algebraic equation" has nothing to do with *algebraic functions*, a term that I have tried to make sure not to use.

[8] In these notes, the letter C (possibly with subscripts), in plain-italic font, denotes a constant unless otherwise specified.

[9] A convention for these notes: when the variables in an algebraic equation are denoted by the specific letters x and y, then unless otherwise specified, we regard x as the first element in an ordered pair (x, y), and regard y as the second element.

y). The notation "$\frac{dy}{dx}$" tells you which variable is which. The *independent* variable is the domain-variable for a function for which that DE is "looking." The *dependent* variable is a letter chosen for the *output* of such a function.

Definition 3.1 (*solution* of a derivative-form DE) Given a function G as above:

(a) A 1-variable function ϕ defined on an open set[10] D in $\mathbf{R}$ is said to *satisfy* equation (3.3) if (i) ϕ is differentiable on D and (ii) when "$y = \phi(x)$" is substituted into equation (3.3), the resulting equation is a true statement for each $x \in D$. (Criterion (ii) can be stated equivalently, without naming a depending variable, as: $\mathsf{G}(x, \phi(x), \phi'(x)) = 0$ for each $x \in D$).

Similarly, a 1-variable function ϕ defined on a (positive-length) interval[11] I is said to *satisfy* equation (3.3) if conditions (i) and (ii) above are satisfied with "D" replaced by "I".

(b) A *solution of* (3.3) is a function ϕ, with domain an interval (of positive length), that satisfies (3.3). If the domain-interval of the solution ϕ is I, we say that ϕ is a *solution of* (3.3) *on* I.[12]

(c) We call a one-variable function ϕ a *solution of* (3.3) (no interval mentioned) if ϕ is a solution of (3.3) on *some* open interval I.

(d) A *solution curve* of (3.3) is the graph of a solution, i.e. the set

$$\{(x, \phi(x)) : x \in I\},$$

where ϕ is a solution of (3.3) on the interval I. ■

[10] "Open set" (in $\mathbf{R}$) is a generalization of "open interval". A set D in $\mathbf{R}$ is called *open* if for every x_0 in D, there is an open interval centered at x_0 that is entirely contained in D. It can be shown that every nonempty open set in $\mathbf{R}$ is either a single open interval, or a union of non-intersecting open intervals.

[11] See section 5.1. For the case of a non-open positive-length interval I, see Section 5.1.1 for the interpretation of $\frac{dy}{dx}$ at any endpoint that I contains.)

[12] See, for example, [1, p. 3]. Some current textbooks refer to a *solution* of a DE as an *explicit solution* of that DE, 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 (3.3), but that $\phi(x)$ is an explicit solution of (3.3). 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 the 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. Such an "explicit solution" is, then, *not* an "explicitly-defined function" under any customary meaning of "explicitly defined". (3) The terminology leads to the absurd-sounding, "The functions implicitly defined by $F(x, y) = 0$ are explicit solutions (of the appropriate DE)."

(In these notes, the symbol ■ indicates the end of a definition, example, exercise, proof of a theorem, or just the *statement* of a theorem if a proof is not given. We often omit this symbol if it is clear that the definition, example, etc., has ended, e.g. if the next line of text is the start of a new labeled definition, example, etc.)

Remark 3.2 In the setting of Definition 3.1, if ϕ is a solution of (3.3), we allow ourselves the convenience of calling the equation "$y = \phi(x)$" a solution of (3.3), even though this is not in agreement with the precise definition of "solution" above. (An *equation* and a *function* 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.) For example, we allow ourselves to say, *technically incorrectly*, that "$y = x^2$ is a solution of $\frac{dy}{dx} = 2x$", because that wording is so much less awkward than "the function ϕ defined by $\phi(x) = x^2$ is a solution of $\frac{dy}{dx} = 2x$".[13] This is similar to allowing ourselves to say "$x = 5$ is a solution of $x^2 = 25$" in place of the more precise "5 is a solution of $x^2 = 25$." The wordings "$y = x^2$ is a solution ..." and"$x = 5$ is a solution ..." are a particular type of something called "abuse of terminology", in which we (often unconsciously) use terminology in a way that gets the point across but is technically incorrect. The "$x = 5$ is a solution of $x^2 = 25$" type of abuse of terminology 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 a equation."

Remark 3.3 (Constant solutions) A derivative-form DE may have one or more *constant solutions*, or none. A constant solution is simply a constant function that is a solution of the DE. For example consider the differential equation

$$\frac{dy}{dx} = (y-7)\sin\big(\,(y+3)x\,\big)\,, \tag{3.4}$$

and define functions ϕ_1, ϕ_2, ϕ_3 by $\phi_1(x) = 7$, $\phi_2(x) = -3$, $\phi_3(x) = 0$ (for all $x \in \mathbf{R}$, in each case). All three of these are constant functions, so their derivatives are identically

[13] Slightly more awkward than "$y = x^2$ is a solution of $\frac{dy}{dx} = 2x$", but suffering from a similar inaccuracy, is the following type of phrasing that you may have seen: "The function $\phi(x) = x^2$ is a solution of $\frac{dy}{dx} = 2x$." This is certainly much less awkward than, "The function ϕ defined by $\phi(x) = x^2$ is a solution of $\frac{dy}{dx} = 2x$." The reason I (mostly) avoid phrasing like "The function $\phi(x) = x^2$ ..." 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 $f(x) = x^2$", 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 phrasing like "the function $\phi(x) = x^2$" 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.*

zero. If we plug "$y = \phi_i(x)$" into equation (3.4) (for $i = 1, 2$, or 3), the left-hand side is 0 for all x. If we plug $y = \phi_1(x)$ into (3.4), the right-hand side is also 0 for all x. The same is true for ϕ_2. Thus, ϕ_1 and ϕ_2 are constant solutions of (3.4) on $(-\infty, \infty)$ and, indeed, on any interval. But if we plug $y = \phi_3(x)$ into (3.4), we obtain the equation $0 = -7\sin(3x)$. In any interval there are values of x for which $\sin(3x) \neq 0$, hence for which "$0 = -7\sin(3x)$" is a false statement.[14] Hence ϕ_3 is *not* a solution of equation (3.4) on *any* interval.

The constant solution ϕ_1 of (3.4) may be expressed any of the following ways:

(i) $y = 7$.

(ii) $y \equiv 7$.

(iii) $y(x) = 7$.

(iv) $y(x) \equiv 7$.

Interpretation of notation (i) depends very strongly on context. In the present context, (i) does not mean "y is the number 7." *When it's understood that what we're writing down is a solution of a DE in which y is the dependent variable, "$y = 7$" represents a* constant function *whose value at every point in the domain is* 7. The graph of this "$y = 7$" is a *horizontal line* in the xy plane (assuming x is the independent variable, as it is in equation (3.4)), not a point on the real line.

The symbol "$\equiv$" in (ii) is read "is identically equal to". This notation is sometimes used as a reminder that the object on the left-hand side is a *function*, an object whose value could potentially depend on an (unwritten) independent variable. The equation "$y \equiv 7$" means *exactly* the same as what "$y = 7$" means *in the current context*. Similarly, in Definition 3.1, instead of writing "$\mathsf{G}(x, \phi(x), \phi'(x)) = 0$ for each $x \in D$", we could have written "$\mathsf{G}(x, \phi(x), \phi'(x)) \equiv 0$ on D."

The notation (iii) is simply another way of reminding ourselves (or informing a reader) that we are using the letter y to represent the output of a function whose input we're representing by the letter x. Notation (iv) (in which $\equiv$ is again read "is identically equal to") is simply an extra-forcible reminder that we're talking about a constant function for which we've chosen the letter x for the indepedent variable and the letter y as the dependent variable.

In these notes, the notation-form we use most often for constant solutions of DEs with dependent variable y is (i), since this is most consistent with our notation for *any* solution of a derivative-form DE. **It is critical that the student understand that such an equation, in that context, is describing a *constant function*, not the value of a single *number* y.**

[14] The fact that $7\sin(-3x) = 0$ for *some* values of x is irrelevant. Solutions of a derivative-form DE are *functions* of the independent variable, not *values* of the independent variable.

Note that an equation of the form "$x =$ constant" (whose graph in the xy plane would be a *vertical* line) can *never* be a solution of a derivative-form DE in which x is the independent variable. An independent variable has to be able to *vary*.

When a DE has any constant solutions, these solutions are almost always *very important*. In real-life DEs in which the independent variable is *time*, and the dependent variable is some important measurable quantity whose behavior is being modeled by the DE—e.g. the temperature in a room, or the concentration of some chemical species—a constant solution represents *equilibrium*. For this reason, constant solutions are often called *equilibrium solutions*.[15]

Despite their importance, constant solutions of DEs can *almost never* be found by manipulating the DE (unless the DE is linear). They are found by substituting "$y = c$" into the DE (where c represents a real number—of course any other letter could be used—and "$y = c$" has the meaning above) and seeing which values of c, if any, make the resulting equation a true statement for all x (or whatever letter is being used for the independent variable). For example, if we substitute $y = c$ into equation (3.4), the equation we obtain is $0 \equiv (c-7)\sin((c+3)x)$, where we have used the "$\equiv$" symbol as a reminder that for $y = c$ to be a solution of the DE, this last equation has to hold for *all* x (or for all x in some specified interval). The student should be able to show that the only values of c that work are $c = 7$ and $c = -3$. Thus, the functions ϕ_1 and ϕ_2 defined earlier are the *only* constant solutions of (3.4). ■

3.2.2 Maximal and general solutions of derivative-form DEs

Definition 3.4 Let I be a (positive-length) interval. For a given differential equation, the *general solution on* I is the collection of all solutions (of that DE) on I.

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. You'll see an example shortly in the paragraph containing equation (3.6). This suggests making the following definition:

Definition 3.5 (temporary) For a given three-variable function G, the *general solution* of the differential equation

$$\mathsf{G}(x, y, \frac{dy}{dx}) = 0 \tag{3.5}$$

[15]This terminology is most common for *autonomous* DEs: equations of the form $\frac{dy}{dx} = p(y)$, a particular type of *separable* equation.

is the collection of all solutions of (3.5), where "solution of (3.5)" is defined as in Definition 3.1(c). Said another way, the general solution of (3.5) is the collection of pairs (I, ϕ), where I is an open interval and ϕ is a solution of (3.5) 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 Section 3.2.4.

There is a "redundancy" problem with Definition 3.5 that we will discuss shortly. However, in a first course on differential equations, many students will not have the mathematical sophistication needed to appreciate the problem or the way we will fix it. Therefore **in a non-honors first course on differential equations, it is acceptable to use Definition 3.5 as the definition of "general solution", and students in my non-honors classes will not be penalized for doing so.** Some students, however, may recognize (eventually, if not immediately) that while Definition 3.5 has no *logical* problem, it undesirably "overcounts" solutions. The discussion below is for those students, and any others who might be interested in learning what the problem is. **Non-honors students who are not interested, or have trouble understanding the discussion, should skip to Example 3.11 and simply ignore the word "maximal" wherever it appears in these notes.** But honors students should *not* skip ahead; they should continue on with the next paragraph.

To illustrate the problem, consider the rather simple DE $\frac{dy}{dx} = -y^2$. 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

$$\phi(x) = \frac{1}{x - C} \tag{3.6}$$

on the domain of the solution. Remembering that the domain of a solution of a derivative-form DE is required to be an *interval*, we look at equation (3.6) and say, "Okay, for each C this formula gives two solutions, one on $(-\infty, C)$ and (C, ∞)." But these are not actually *all* the solutions, because $(-\infty, C)$ and (C, ∞) are not the *only* two intervals on which equation (3.6) defines solutions; they are simply the *largest* (i.e. most inclusive) such intervals. 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[16] of $(-\infty, C)$ or (C, ∞) whatsoever.

This example illustrates that the collection of pairs (I, ϕ) referred to in Definition 3.5 has a certain redundancy. The concepts of *restriction* and *extension* introduced

[16] A *subinterval* of an interval I is subset of I that is an interval.

in Section 3.1 allows us to speak precisely about this redundancy. The following definition simply restates these concepts in the context of greatest importance to us (domains that are intervals), and gives some notation for restrictions.

Definition 3.6 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

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

(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 (again restating something from Section 3.1) : if I is a subinterval of an interval $\tilde{I}$, and ϕ and $\tilde{\phi}$ are functions defined on I and $\tilde{I}$ respectively, then

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

(The symbol " $\iff$ " 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 ", $\iff$ " as "which is true if and only if".)

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 3.5: the definition names a much larger collection of functions than is needed to capture all the information there is to know about solutions of (3.5). We will see shortly 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

[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.

actually a solution of the given IVP. This is the differential-equations analog of the "implied domain" of a function expressed 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

$$\frac{dy}{dx} = -x^{-2}, \quad y(3) = \frac{1}{3}, \tag{3.7}$$

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 make the following definition.

Definition 3.7 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 a solution $\tilde{\phi}$ of the DE on any open interval $\tilde{I}$ strictly containing I. The graph of a maximal solution is called a *maximal* (or *inextendible*) *solution curve*.

Example 3.8 All the functions ϕ_i below are different functions (because we have specified different domains for them).

- $\phi_1(x) = \frac{1}{x},\ 0 < x < 5$, is a solution of $\frac{dy}{dx} = -x^{-2}$, but not a maximal solution. It is also a solution of the IVP (3.7).
- $\phi_2(x) = \frac{1}{x},\ 2.9 < x < 16.204$, is another solution of $\frac{dy}{dx} = -x^{-2}$, and of the IVP (3.7), but not a maximal solution.
- $\phi_3(x) = \frac{1}{x},\ 3.1 < x < 16.204$, is another solution of $\frac{dy}{dx} = -x^{-2}$, but it is neither a maximal solution nor a solution of the IVP (3.7),
- $\phi_4(x) = \frac{1}{x},\ x \in (0, \infty)$, is *a* maximal solution of $\frac{dy}{dx} = -x^{-2}$, and is *the* maximal solution of the IVP (3.7).
- $\phi_5(x) = \frac{1}{x},\ x \in (-\infty, 0)$, is a *different* maximal solution of $\frac{dy}{dx} = -x^{-2}$. It is *not* a solution of the IVP (3.7).
- $\phi_6(x) = \frac{1}{x},\ x \in (-\infty, -\sqrt{2})$ is another non-maximal solution of $\frac{dy}{dx} = -x^{-2}$.
- $\phi_7(x) = \frac{1}{x} + 37,\ x \in (0, \infty)$ is yet another maximal solution of $\frac{dy}{dx} = -x^{-2}$. It is not a solution of the IVP (3.7).

Example 3.9 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][19] 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, more useful than the (temporary) Definition 3.5 is this:

Definition 3.10 For a given G, the *general solution* of the DE (3.5) is the collection of all maximal solutions of (3.5). ■

(Definition 3.10 supersedes Definition 3.5.)

Example 3.8 demonstrates the economy gained by including the word "maximal" in this definition. The student will probably agree that, even prior to writing down Definition 3.10, 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 3.11 The general solution of $\frac{dy}{dx} = x$ may be written in short-hand as

$$\left\{ y = \frac{1}{2}x^2 + C \right\}. \tag{3.8}$$

In this context equation (3.8) 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 (3.11); they tell us we're talking about a *set* of objects of the form within the braces.) We allow ourselves to write (3.8) as short-hand for "the collection of functions $\{\phi_C : C \in \mathbf{R}\}$, where $\phi_C(x) = \frac{1}{2}x^2 + C$".[20] A convention in these notes is that "$\left\{y = \frac{1}{2}x^2 + C\right\}$" means the same thing as $\left\{y = \frac{1}{2}x^2 + C : C \in \mathbf{R}\right\}$.

[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] *Note to instructors*: Existence of a maximal extension is not so obvious, absent any conditions that ensure local uniqueness of solutions. However, this existence follows easily from Zorn's Lemma. The set of S of extensions of a solution ϕ is partially ordered by the extension/restriction relation. Every chain has an upper bound (in fact, a maximal element). Hence, by Zorn's Lemma, S has at least one maximal element.

[20] Students in my own classes are permitted to omit the curly braces in (3.8), but I am trying to maintain certain notational consistency across different sections of these notes.

Example 3.12

- The general solution of
$$\frac{dy}{dx} = -x^{-2}, \quad x > 0 \tag{3.9}$$
(meaning that we are interested in this differential equation only for $x > 0$) may be written as
$$\left\{ y = \frac{1}{x} + C \right\}, \ x > 0, \tag{3.10}$$
a one-parameter family of maximal solutions. Because the restriction $x > 0$ is stated explicitly in (3.9), it is permissible to omit the "$x > 0$" when writing the general solution; we may simply write the general solution as
$$\left\{ y = \frac{1}{x} + C \right\} \tag{3.11}$$

- The general solution of
$$\frac{dy}{dx} = -x^{-2}, \tag{3.12}$$
with no interval specified, may also be written as (3.11)—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 (3.12) was written down, we must consider all possible intervals. Therefore, in this context, equation (3.11) does *not* represent a one-parameter family of maximal solutions; it represents *two* one-parameter families of maximal solutions[21]. Equation (3.11) is acceptable short-hand for

[21]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". (See, for example [5, p. 240].) 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 he Fundamental Theorem of Calculus, reviewed in Section 5.3) assures us that *on any positive-length 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 such an 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—i.e. are not single intervals—such as the implied domain of $f(x) = x^{-2}$ or the implied domain of $f(x) = \sec^2 x$.

$$\left.\begin{array}{l} \text{the union of the two families of functions} \\ \qquad \{\phi_C \mid \ C \in \mathbf{R}\}, \quad \{\psi_C \mid \ C \in \mathbf{R}\} \\ \text{where} \\ \qquad \phi_C(x) = \frac{1}{x} + C, \quad x > 0 \\ \text{and} \\ \qquad \psi_C(x) = \frac{1}{x} + C, \quad x < 0. \end{array}\right\} \tag{3.13}$$

(The *union* of the two families means the collection of functions that are in one family or the other.[22]) The solution $y = \frac{1}{x} + 6$ on $\{x < 0\}$ (the function ψ_6 in the notation of (3.13)) 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

$$\text{“}\{y = \frac{1}{x} + C, \ x > 0\} \ \text{ and } \ \{y = \frac{1}{x} + C, \ x < 0\}\text{”} \tag{3.14}$$

and

$$\text{“}\{y = \frac{1}{x} + C_1, \ x > 0\} \ \text{ and } \ \{y = \frac{1}{x} + C_2, x < 0\}\text{”}.^{23} \tag{3.15}$$

In (3.14), 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 (3.15), 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 (3.11), (3.14), and (3.15) 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 variations on the same theme.

[22] *Note to instructors:* In these notes I have opted not to use the symbol $\bigcup$ for union of sets of functions, out of concern that this might confuse some students. You will notice later on, e.g. in (3.14), 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 “or”, not “and”, but I've deliberately avoided writing (3.14) and similar expressions this way. I felt that using the word “or” in these expressions would be confusing to students.

[23] In both (3.14) and (3.15), where we combine two or more sets of solutions using the word “and”, the union-symbol $\bigcup$ would be more precise.

Example 3.13 The general solution of $\frac{dy}{dx} = \sec^2 x$ may be written as

$$\{y = \tan x + C\}, \tag{3.16}$$

or as

$$\left\{y = \tan x + C, \quad (n - \frac{1}{2})\pi < x < (n + \frac{1}{2})\pi, \quad n \text{ an integer}\right\}, \tag{3.17}$$

or as

$$\left\{y = \tan x + C_n, \quad (n - \frac{1}{2})\pi < x < (n + \frac{1}{2})\pi, \quad n \text{ an integer}\right\}, \tag{3.18}$$

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 (3.16)–(3.18) (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 3.14 The general solution of the separable equation

$$\frac{dy}{dx} = -y^2 \tag{3.19}$$

may be written as

$$\left\{y = \frac{1}{x - C}\right\} \text{ and } \{y = 0\}, \tag{3.20}$$

or in various other ways that impart the same information. (This fact will be justified in Section 3.2.10; just assume it for now.) In the given context, the solution that is the constant function 0 may be written as "$y = 0$", as in (3.20) or as "$y \equiv 0$" (which, in this context, is read "y identically zero"). Since a solution of (3.19), expressed in terms of the variables in (3.19), is function of x, the only correct interpretation of "$y = 0$" in (3.20) is "y is the constant function whose value is zero for all x", *not* "y is a real number, specifically the number 0". As mentioned in Remark 3.3, it is also okay to use the notation "$y \equiv 0$" for this constant solution.

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 $(-\infty, C)$.

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 (3.19):

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 $y \equiv 0$.

Example 3.15 The general solution of the separable equation

$$\frac{dy}{dx} = y(1 - y) \tag{3.21}$$

may be written as

$$\left\{ y = \frac{C}{e^{-x} + C} : C \neq 0 \right\} \quad \text{and} \quad \{y \equiv 0\} \quad \text{and} \quad \{y \equiv 1\}\,. \tag{3.22}$$

(As with the previous example, this fact will be justified in Section 3.2.10; just assume it for now.)

Line (3.22) is not the *only* good way to write down the collection of all maximal solutions of the given DE, which can also be said of line (3.20) in Example 3.19. This is an important phenomenon discussed later in Section 3.2.4, under "The myopic eye of the beholder."

3.2.3 "Standard Forms" and solutions in a region

In Section 3.2.1, the equation-form

$$\mathsf{G}(x, y, \frac{dy}{dx}) = 0 \tag{3.23}$$

was used simply as a way to talk about all first-order DEs at once, and to define "solution". For DEs that are *so* general, with no algebraic restrictions on the function G, there isn't much that we can say about the set of solutions. Fortunately, a very important class (arguably, *the* most important class) of first-order DEs can be put in the form

$$\frac{dy}{dx} = f(x, y), \tag{3.24}$$

where f is a function defined on some region in $\mathbf{R}^2$ (often all of $\mathbf{R}^2$). Every equation of the form (3.24) is equivalent to an equation of the form (3.23); simply take $\mathsf{G}(x, y, z) = z - f(x, y)$. However, the converse is false; for example,

$$\left(\frac{dy}{dx}\right)^5 + \left(\frac{dy}{dx}\right)^2 + \sin\left(x + y\frac{dy}{dx}\right) - x^2 + y^3 - 17 = 0$$

cannot be put in the form (3.24).

Equation (3.24) is often referred to as "standard form" for a general first-order ODE. However, for a *linear* first-order DE, "standard form" means something else, namely the form

$$\frac{dy}{dx} + P(x)y = Q(x), \tag{3.25}$$

(where P and Q are defined on some interval) To avoid confusion, in these notes we will refer to (3.24) as *standard general form*, and to (3.25) as *standard linear form*.

Thanks to the "integrating factor" approach to linear DEs (not presented in these notes), we already know "all there is to know" about linear equations (3.25), at least for functions P and Q that are continuous on an interval I: we know that for every initial conditon $y(x_0) = y_0$ with $x_0 \in I$, the corresponding IVP has a unique solution that is maximal in I; that the domain of this solution is the entire interval I; and that we have an explicit *formula* for the solution in terms of integrals of P and Q. (We may or may not be able to "do" the integrals explicitly—i.e. to find antiderivatives that are elementary functions—but the formula *in terms of those integrals* is explicit.) Thus, general results about equations of the form "$\frac{dy}{dx} = f(x, y)$" are of interest to us mainly when they are *non*-linear, and for practical purposes we may *think* of this form as "standard *non*-linear form." The problem with the latter terminology is that linear DEs can be put in this form as well: given functions P and Q on an interval I, if we define $f(x, y) = Q(x) - P(x)y$, then the standard-linear-form equation (3.25) is algebraically equivalent to $\frac{dy}{dx} + P(x)y = Q(x)$ on the region $I \times \mathbf{R}$ (see Section 5.2 for the notation "$I \times \mathbf{R}$"), and see Section 3.2.8 for the meaning of "algebraically equivalent on a region"). To avoid the contradictory-sounding "$\frac{dy}{dx} = Q(x) - P(x)y = 0$ is a linear equation in standard non-linear form," and to avoid using terminology as awkward as "standard not-necessarily-linear form" as a remedy, we are opting to use the term "standard *general* form" for $\frac{dy}{dx} = f(x, y)$.

Unsurprisingly, the behavior of solutions of DEs of the form $\frac{dy}{dx} = f(x, y)$ depends on some properties of the function f, e.g. continuity or differentiability. When $f(x, y) = Q(x) - P(x)y$ for some given one-variable functions P, Q defined on an interval I (corresponding to the linear case, as discussed above), the behavior of P and Q on the interval $I \subseteq \mathbf{R}$ completely determine all relevant properties of f on the region $I \times \mathbf{R} \subseteq \mathbf{R}^2$, a "vertical strip" (see Section 5.2 for the notation "$I \times \mathbf{R}$"). This is not true for a more-general function f, which may therefore not have "nice" properties on an entire vertical strip $I \times \mathbf{R}$, but may have them on some rectangle or more complicated region $R \subseteq \mathbf{R}^2$. Since a solution ϕ of equation (3.24) satisfies $\phi'(x) = f(x, \phi(x))$, properties of f can inform us about the behavior of ϕ only at points $(x, \phi(x))$ that lie in R. In other words, if that graph of ϕ *leaves* R, we can expect our "nice" properties of f to inform us only about portion of the graph that lies in R.[24] If the function ϕ is defined on an interval I, and its graph lies partially in R, we may need to restrict ϕ to a smaller interval to obtain a solution whose graph lies in R.

Thus, for a given DE, although the set of solutions on a given interval I is still important, we need some terminology that takes into account the considerations above. This terminology applies whether or not our DE is in standard form, so we define it for any derivative-form DE.

Definition 3.16 (Solution, and solution curve, in a region) Let R be a region in the xy plane. Given a differential equation $\mathsf{G}(x, y, \frac{dy}{dx}) = 0$, we say that a solution ϕ defined on an interval I is a *solution (of the given DE) in* R, if the graph of ϕ (the graph of the equation $y = \phi(x)$, with x required to lie in I) is contained in R. We call the graph of a solution in R a *solution curve in* R.

Definition 3.17 (Maximal solution, and maximal solution curve, in a region) Let R be a region in the xy plane. Given a differential equation $\mathsf{G}(x, y, \frac{dy}{dx}) = 0$, we say that a solution ϕ is *maximal in* R if ϕ is a solution in R that cannot be extended to a solution $\tilde{\phi}$ whose graph is still contained in R. When feasible, to avoid awkward or lengthy wording (such as "maximal-in-R solution" or "solution that is maximal in R"), we also use the term *maximal solution in* R for a solution that is maximal in R.[25] For the same reason, we use the term *maximal solution curve in* R for the graph of a maximal solution in R.

Definition 3.18 (General solution in a region) Let R be a region in the xy plane. The *general solution, in* R, of a differential equation $\mathsf{G}(x, y, \frac{dy}{dx}) = 0$, is the collection of all maximal solutions in R. ■

[24] Recall that the *graph* of a one-variable function ϕ defined on a set I is the set $\{(x, \phi(x)) : x \in I\} \subseteq \mathbf{R}^2$.

[25] We're stating this convention explicitly because otherwise "maximal solution in R" could be interpreted to mean "maximal solution, with no restriction on the region, that happens to lie in R.

Thus the general solution of a derivative-form DE, as defined in Definition 3.10, is the same as the general solution of that DE in the region $\mathbf{R}^2$.

The DEs about which we can draw the most systematic conclusions are those that satisfy the conditions of the Fundamental Theorem of Ordinary Differential Equations (Theorem 5.8, henceforth "the FTODE") on some region R that may or may not be all of $\mathbf{R}^2$. (For example, the DE $\frac{dy}{dx} = y^{1/3}$ satisfies these conditions on the region $\{(x, y) \in \mathbf{R}^2 : y > 0\}$ and on the region $\{(x, y) \in \mathbf{R}^2 : y < 0\}$, but not on all of $\mathbf{R}^2$, or on any region that includes a point on the x-axis.)

The FTODE has several consequences that we will be making use of. The corollary below, which appears also in the Appendix as Corollary 5.11, states three of these that are very closely related. (Essentially they are the same result stated three ways.)

Corollary 3.19 *Let f be a function of two variables and suppose that R is an open region in $\mathbf{R}^2$ on which f and $\partial f/\partial y$ are continuous. Then:*

(a) For every (x_0, y_0) in R, the initial-value problem

$$\frac{dy}{dx} = f(x, y), \quad y(x_0) = y_0 \tag{3.26}$$

(5.13) has a unique solution that is maximal in R. This solution $\phi_{\max}$ has the property that every solution of (5.13) in R is a restriction of $\phi_{\max}$.

(b) Every point (x_0, y_0) in R lies on a unique maximal solution curve in R. [26]

(c) No two distinct maximal solution curves in R intersect.

(In parts (b) and (c), "solution curve" means "solution curve of the DE in (3.26)".) ■

How can we ever be sure we have found *all* solutions (maximal in a given region R, perhaps all of $\mathbf{R}^2$) of a derivative-form DE? The key principle is always the following:

> The set of all solutions of a differential equation is the same as the set of solutions of all initial-value problem for that DE. (3.27)

Statement (3.27) is true because every solution of a differential equation is a solution of *some* initial-value problem for the same DE: for any point (x_0, y_0) on the graph of

[26] *Note to instructors*: In differential-geometric terminology, the maximal solution curves *foliate* R.

a solution ϕ of a DE, the function ϕ is a solution of the initial-value problem for that DE with initial condition $y(x_0) = y_0$.

For linear DEs we apply this principle in a rather special way that involves the integrating-factor method. For nonlinear DEs *for which the hypotheses of the Fundamental Theorem are met*, we apply this theorem to be sure we've found all solutions of all IVP's for the given DE. This will be illustrated later for separable derivative-form DEs.

For *some* DEs for which the hypotheses on f and $\partial f/\partial y$ in Theorem 5.8 fail are not satisfied on the entire region R of interest, but are satisfied on open regions R that comprise "most" of the region of definition, some additional analysis enables us to use our knowledge of the solutions on these smaller regions to deduce what all the solutions are in R. This will also be illustrated later; see Example 3.43 and Example 3.48 (The DE in Example 3.43 is linear, so you don't see *regions* of interest mentioned explicitly there. For linear DEs, the open regions of interest are always of the form $I \times \mathbf{R}$, where I is an open interval, so we only need to identify the relevant *intervals*.)

3.2.4 One-parameter families of solutions

It is easiest to get a sense of what the term "one-parameter family of solutions" means by seeing it used in examples, before attempting to give a definition:

1. The collection of equations $\{y = x^2/2 + C : C \in \mathbf{R}\}$ represents a one-parameter family of solutions of $\frac{dy}{dx} = x$. A more precise way of writing this family is

$$\{\phi_C : C \in \mathbf{R}\},$$
$$\text{where } \phi_C(x) = x^2/2 + C \text{ and the domain of each } \phi_C \text{ is } (-\infty, \infty).$$

2. The collection of equations $\{y = \frac{1}{x-C} : C \in J, \text{ and } x > C\}$, where J is any positive-length interval (possibly all of $\mathbf{R}$, possibly smaller), represents a one-parameter family of solutions of $\frac{dy}{dx} = -y^2$, the DE in Example 3.14. A more precise way of writing this family is

$$\{\phi_C : C \in \mathbf{R}\},$$
$$\text{where } \phi_C(x) = \tfrac{1}{x-C} \text{ and the domain of } \phi_C \text{ is } (C, \infty).$$

In each case, the *parameter* is C; it distinguishes one solution from another. (The parameter is a "variable constant": within the given fomula for any of the solutions, C is a constant, but by varying C we get different solutions.)

Note that, for a given C, the ψ defined by $\psi(x) = \frac{1}{x-C}$ (defined for all $x \neq C$) satisfies $\frac{dy}{dx} = -y^2$ both on the interval (C, ∞) and on the interval $(-\infty, C)$, and therefore represents *two* maximal solutions of $\frac{dy}{dx} = -y^2$, not *one*.

Generally, the term *one-parameter family of solutions* of a given differential equation $\mathsf{G}(x, y, \frac{dy}{dx}) = 0$ is used for a collection of solutions $\{\phi_C : C \in J\}$ of that DE, where J is some positive-length interval. Usually some additional restrictions are understood, if not required explicitly, to ensure that the solution ϕ_C varies "nicely" as the parameter c changes. Typically, these restrictions amount to the requirement that (i) for each $c \in J$, the function ϕ_c is a maximal solution of the given DE and that (ii) there is a continuous, two-variable function Φ defined on the set

$$\{(C, x) : C \in J \text{ and } x \in I_C\} \subseteq \mathbf{R}^2$$

for which

$$\phi_C(x) = \Phi(C, x), \quad x \in I_C = \text{domain-interval of } \phi_C.$$

In general, the parameter-interval J (the "C-interval") need not have any fixed relation to the intervals I_C (the "x-intervals") which themselves may or may not vary as C varies. (In the second collection-of-equations example above, the corresponding functions and domains are $\Phi(C, x) = \frac{1}{x-C}$, restricted to the domain $\{(C, x) \in \mathbf{R}^2 : x > c\}$; and, for each $C \in \mathbf{R}$, the function $\phi_C(x) = \frac{1}{x-C}$, restricted to $I_C = (C, \infty)$.)

The myopic eye of the beholder

The general solution of $\frac{dy}{dx} = -y^2$ exhibits (non-obviously) another phenomenon that needs closer examination. The way we have written the general solution in (3.20) isolates the maximal solution $y \equiv 0$ as not belonging to what appears to be a single nice family (of *equations*, not *solutions*), namely $\left\{y = \frac{1}{x-C}\right\}$, into which all 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 we could also write the general solution (3.20) as

$$\left\{y = \frac{1}{x-C} : C \neq 0\right\} \text{ and } \left\{y = \frac{1}{x}\right\} \text{ and } \{y = 0\}, \tag{3.28}$$

(since "$\frac{1}{x}$" if what "$\frac{1}{x-C}$" reduces to if $C = 0$). But for $C \neq 0$, writing $K = \frac{1}{C}$,

$$\frac{1}{x-C} = \frac{C^{-1}}{C^{-1}x - 1} = \frac{K}{Kx - 1}\ . \tag{3.29}$$

In the right-most formula in (3.29), 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 (3.20). Thus, we can rewrite the general solution (3.20) as

$$\left\{y = \frac{K}{Kx - 1}\right\} \text{ and } \left\{y = \frac{1}{x}\right\}\ . \tag{3.30}$$

Here, K is an arbitrary constant, allowed to assume all real values, just as C was allowed to in (3.20); we could just as well use the letter C for it. Writing the

general solution as in (3.30), 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" (not required to be a single *one-parameter* 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.

This example also illustrates another 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 (3.20) do not form *one* one-parameter family of solutions; they form *two*.[27] Every value of C corresponds to two maximal solutions, one defined to the left of C and one defined to the right[28]. In (3.30), 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 (3.20) is preferable to (3.30) as a way of writing down the general solution (although both are correct). 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 $\{y = \frac{1}{x-C}\}$ is more "uniform" than is the collection $\{y = \frac{K}{Kx-1}\}$, 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 of the given DE, neither of which can be preferred over the other by any such considerations.

Example 3.20 In line (3.22), we wrote the general solution of the DE (3.21) as

$$\left\{ y = \frac{C}{e^{-x} + C} : C \neq 0 \right\} \quad \text{and} \quad \{y \equiv 0\} \quad \text{and} \quad \{y \equiv 1\}\,. \tag{3.31}$$

[27] 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.

[28] *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 decide to tell your students that. (That's how I viewed this picture until I had taught differential equations for 15 years or so.) 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 (another misconception I held for many years). Such a conclusion would be fine if we were talking about the one-parameter family of *rational functions* defined by "$y = \frac{1}{x-C}$", but we are not; we are talking about *solutions of a derivative-form ODE*, for which the *only* sensible domain is a connected one.

Following the same steps used above to rewrite (3.28) a different way, the general solution (3.31) can be rewritten as

$$\left\{y = \frac{1}{Ce^{-x}+1} : C \neq 0\right\} \quad \text{and} \quad \{y \equiv 0\} \quad \text{and} \quad \{y \equiv 1\}\,. \tag{3.32}$$

In line (3.31), we can absorb the constant solution $y \equiv 0$ into the first family by removing the "$C \neq 0$" restriction within that family's curly braces; if we set $C = 0$, then "$y = \frac{C}{e^{-x}+C}$" becomes "$y = 0$" (which we have written as "$y \equiv 0$" in line (3.22) as an (optional) reminder that this equation represents the constant *function* with equation $y(x) = 0$, not the *number* 0). Similarly, in line (3.32), we can absorb the constant solution $y = 1$ into the first family by removing the "$C \neq 0$" restriction within that family's curly braces; if we set *that* C equal to 0, "$y = \frac{1}{Ce^{-x}+1}$" becomes "$y = 1$." Hence, a third and fourth equivalent way of writing the general solution (3.31) are

$$\left\{y = \frac{C}{e^{-x}+C}\right\} \quad \text{and} \quad \{y \equiv 1\} \tag{3.33}$$

and

$$\left\{y = \frac{1}{Ce^{-x}+1}\right\} \quad \text{and} \quad \{y \equiv 0\}\,. \tag{3.34}$$

(In accordance with our previously stated convention, C is intended to be completely arbitrary in the sets in curly braces above, since we have placed no restrictions on it.) In each of the lines (3.33) and (3.34), in the family in curly braces the formula giving $y(x)$ yields two maximal solutions when $C < 0$ and one maximal solution when $C \geq 0$. The $C = 0$ solution in (3.33) is the constant function 0, which is the "exceptional" solution in (3.34). The $C = 0$ solution in (3.34) is the constant function 1, which is the "exceptional" solution in (3.33). The situation is completely symmetric; neither of (3.33) and (3.34) 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) 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 linear DEs—and then, 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.[29]

[29] *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

3.2.5 Implicitly defined functions

To understand what "implicit solution of a differential equation" (defined in Section 3.2.6) means, it is essential to understand what "implicitly defined *function*" (of one variable) means. You were introduced to the concept of *implicitly defined functions* as far back as Calculus 1, when you studied *implicit differentiation*, but we will review the concept here. 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 equation in variables x and y, say $F_1(x, y) = F_2(x, y)$. We can always write such an equation in the form $F(x, y) = 0$ for some two-variable function F. However, for the purposes of these notes, it will be helpful to consider equations written in the less restrictive form

$$F(x, y) = c_0 \ , \tag{3.35}$$

where c_0 is a constant that may or may not be 0. We are often interested in solving a two-variable equation such as (3.35) for one variable in terms of the other, e.g. solving for y in terms of x. For example, if x and y are real numbers for which

$$x^2 + y^3 = 1, \tag{3.36}$$

then

---

implicitly. For example, for autonomous DEs, equilibrium solutions are extremely important, and are *never* found by separating variables unless a mistake is made. 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 students at the level of an intro DE course may have heard this term before in "solution-set of an algebraic equation [or inequality]"—and if so, have heard it *only* in this context—and might be 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". Other terminology I have considered for the set of maximal solutions is "full solution" and "complete solution", and I may adopt one of those (or something else) in future versions of these notes. One convention I strongly advise against, however, is to use "general solution" to refer to a *non-exhaustive* collection of solutions (or for a *generic*—i.e. "typical"—element of such a collection) for which (s)he has produced a nicely-parametrized family of formulas. As the simple examples 3.14 and 3.15 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.

$$y = (1 - x^2)^{1/3}. \tag{3.37}$$

In other words, if we define $F(x, y) = x^2 + y^3$, let $c_0 = 1$, and define $\phi(x) = (1 - x^2)^{1/3}$, then whenever the pair (x, y) satisfies $F(x, y) = c_0$, it satisfies $y = \phi(x)$. Conversely, one may verify by direct substitution that if $y = (1 - x^2)^{1/3}$ then $F(x, y) = c_0$. Thus, for *any* real numbers x and y,

$$F(x, y) = c_0 \quad \text{if and only if} \quad y = \phi(x). \tag{3.38}$$

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

$$F(x, \phi(x)) = c_0. \tag{3.39}$$

Geometrically, what (3.39) says is that *the graph of $y = \phi(x)$ is part of the graph of $F(x, y) = c_0$*.

More generally than the example above, any time (3.38) is true for a given two-variable function F, real number c_0, and one-variable function ϕ, we say that the equation $F(x, y) = c_0$ *(implicitly) defines*, or *(implicitly) determines*, y as a function of x.[30] (Using the word "implicitly" is optional, but can be a helpful reminder that even if we have an explicit formula for $F(x, y)$, we may not be able to find one for $\phi(x)$.) We call ϕ an *implicitly defined function*.

Now consider the equation

$$x^2 + y^2 = 1. \tag{3.40}$$

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

$$y = \pm\sqrt{1 - x^2}. \tag{3.41}$$

Looking just at (3.40), it is already clear that any numerical choice of x restricts the possible choices of y that will make this equation a true statement. Equation (3.41) tells us the only *possible* values for y that might 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 (3.40), we see that all the

[30] Since the letters used for a function's independent variable(s) are not part of the function, the wording "$F(x, y) = 0$ (implicitly) determines the relation $y = \phi(x)$" would be more precise. Another alternative the avoids the "function of x" wording is "$F(\cdot, \cdot) = 0$ determines the second variable as a function of the first." (The dots in $F(\cdot, \cdot)$ represent the unnamed independent variables of F But we allow " ... determines [or defines] y as a function of x" since, in addition to being the least clumsy wording, it reflects the fact that what we're thinking of is solving the equation $F(x, y) = 0$ for y in terms of x.

values of y that we have labeled as "possible" actually do work. Thus, *for each pair* (x, y) *of real numbers*,

$$x^2 + y^2 = 1 \text{ if and only if (i) } |x| \leq 1 \text{ and (ii) either } y = \sqrt{1 - x^2} \text{ or } y = -\sqrt{1 - x^2}. \tag{3.42}$$

This is a ***much*** weaker statement than a statement of the form (3.38), because the sign in "$\pm\sqrt{1 - x^2}$" can be chosen *independently for each* x. On the domain $[-1, 1]$, if we define

$$\phi_1(x) = \sqrt{1 - x^2}, \tag{3.43}$$

$$\phi_2(x) = -\sqrt{1 - x^2}, \tag{3.44}$$

$$\phi_3(x) = \begin{cases} \sqrt{1 - x^2} & \text{if } x \text{ is a rational number,} \\ -\sqrt{1 - x^2} & \text{if } x \text{ is an irrational number,} \end{cases} \tag{3.45}$$

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 (3.40). 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

$$e^x + x + 6y^5 - 15y^4 - 10y^3 + 30y^2 + 10xy^2 = 0. \tag{3.46}$$

Clearly, choosing a numerical value for x restricts the possible values for y that will make equation (3.46) a true statement. It turns out that, depending on the choice of x, there can be anywhere from one to five values of y for which the pair (x, y) satisfies equation (3.46). 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 (3.46). 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.

In some cases, an equation $F(x, y) = c_0$ implicitly determines one and only one function of x on the whole real line (equation (3.36) is one such example). That is a "best-case scenario". In the next-best scenario, $F(x, y) = c_0$ implicitly determines one and only one function of x on at least one interval I, allowing us to speak unambiguously of *the* function of x, on I, determined by this equation. But even when we are not in one of these nice situations, we may still be able to achieve a similar outcome

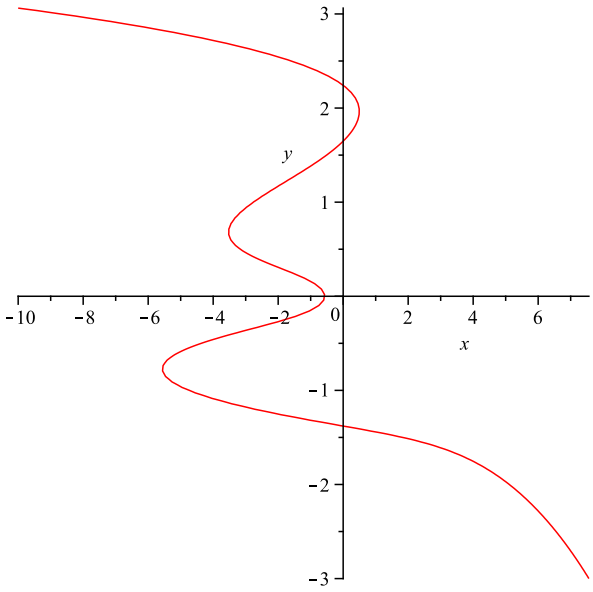

Figure 1: The graph of $e^x + x + 6y^5 - 15y^4 - 10y^3 + 30y^2 + 10xy^2 = 0$.

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. The corresponding set in the xy plane is the rectangle

$$I \times J = \{(x, y) : x \in I \text{ and } y \in J\}. \tag{3.47}$$

(See Definition 5.2 for this notation and terminology.)

When two or more functions ϕ on the same interval satisfy $F(x, \phi(x)) \equiv c_0$ (for a given two-variable function F and number c_0), "windowing" near a point (x_0, y_0) that satisfies $F(x_0, y_0) = c_0$. 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 (3.43)–(3.44) ($\phi_1(x) = \sqrt{1 - x^2}$, $\phi_2(x) = -\sqrt{1 - x^2}$). 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 subinterval of $[-1, 1]$ that contains x_0. To make sure you understand this, 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

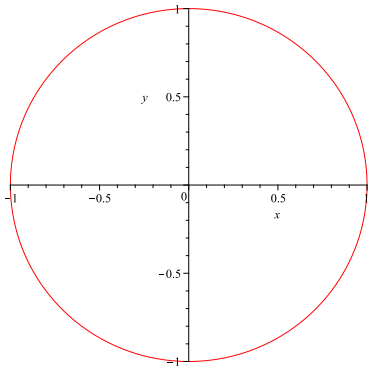

Figure 2: The graph of $x^2 + y^2 = 1$.

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 "windowing" idea underlies the following definition.

Definition 3.21 (implicitly defined function) Let F be a function of two variables and let $c_0 \in \mathbf{R}$.

(a) Let I and J be open intervals.[31]

(i) If for each number $x \in I$, there exists one and only one number $y \in J$ for which $F(x, y) = c_0$, then we say that the equation $F(x, y) = c_0$ *determines* (or *implicitly defines*) y *as a function of* x *in* $I \times J$. When this condition holds, and for each $x \in I$ we let $\phi(x)$ denote the unique $y \in J$ for which $F(x, y) = c_0$, we call ϕ *the function of* x *determined by* (or *implicitly defined*

[31]Note for instructors: Openness of I and J is not essential for part (a) of Definition 3.21, but matters for part (b) if we don't want to have to mention the word "open" each time we use the term "implicitly defined function".

by) *the equation* $F(x,y) = c_0$ *in the rectangle* $I \times J$, and we say that, in the rectangle $I \times J$, the equation $F(x,y) = c_0$ *determines the relation* $y = \phi(x)$.

(ii) Similarly terminology applies with the roles of x and y reversed: If for each number $y \in J$, there exists one and only one number $x \in I$ for which $F(x,y) = c_0$, then we say that the equation $F(x,y) = c_0$ *determines* (or *implicitly defines*) x *as a function of* y *in* $I \times J$. When this condition holds, and for each $y \in J$ we let $\phi(y)$ denote the unique $x \in I$ for which $F(x,y) = c_0$, we call ϕ *the function of* y *determined by* (or *implicitly defined by*) *the equation* $F(x,y) = c_0$ *in the rectangle* $I \times J$, and we say that, in the rectangle $I \times J$, the equation $F(x,y) = c_0$ *determines the relation* $x = \phi(y)$.

(b) We say that a one-variable function ϕ is an *implicitly defined function determined by the equation* $F(x,y) = c_0$ if there is some open rectangle $I \times J$ for which, in that rectangle, the equation $F(x,y) = c_0$ determines either the relation $y = \phi(x)$ or the relation $x = \phi(y)$.

■

Observe that, with notation as in Definition 3.21, "$F(x,y) = c_0$ implicitly determines y as a function of x in $I \times J$" is equivalent to the following:

> There exists one and only one (real-valued) function ϕ defined on I such that (i) $\phi(x) \in J$ for each $x \in I$ and (ii) $F(x, \phi(x)) = c_0$ for each $x \in I$.

Note also that the condition above is equivalent to the following modified version of statement (3.38):

$$\begin{array}{c} \text{For every pair } (x,y) \in I \times J, \\ F(x,y) = c_0 \quad \text{if and only if} \quad y = \phi(x). \end{array} \tag{3.48}$$

The only difference between statement (3.48) and statement (3.38) is that to get the second line of (3.48), we had to make the windowing restriction in the first line. This is usually the best we can do; only occasionally do we have situations in which we can take the "window" to be the whole xy plane and still get a unique implicitly-defined function.

These ideas motivate the following definition:

Definition 3.22 (implicitly defined function) Let F be a two-variable function defined on an open set R and let $c_0 \in \mathbf{R}$. Suppose that I_1 and J_1 are open intervals for which the the rectangle $R_1 = I_1 \times J_1$ is contained in R, and that either

(i) ϕ is a function with domain I_1 and with range contained in J_1, having the property that

$$\begin{array}{c} \text{for every point } (x,y) \in I_1 \times J_1, \\ F(x,y) = c_0 \text{ if and only if } y = \phi(x), \end{array} \tag{3.49}$$

or

(ii) ϕ is a function with domain J_1 and with range contained in I_1, having the property that

$$\begin{array}{c} \text{for every point } (x,y) \in I_1 \times J_1, \\ F(x,y) = c_0 \text{ if and only if } x = \phi(y). \end{array} \tag{3.50}$$

Then we call ϕ an *implicitly defined function* determined by the equation $F(x,y) = c_0$.[32] In case (i), we say that the equation $F(x,y) = c_0$ defines y as a function of x in R_1; in case (ii), we say that this equation defines x as a function of y in R_1. ■

To simplify wording, **henceforth, unless we say otherwise, whenever we speak of an implicitly defined function ϕ determined by an equation $F(x,y) = c_0$, we mean to "regard ϕ as a function of x"—i.e. that the relevant relation determined by $F(x,y) = c_0$ is of the form $y = \phi(x)$ (case (i) of Definition 3.21(a) for some open intervals I and J).**

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 the caption determines y 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.) ■

The *Implicit Function Theorem*, stated and discussed in Section 5.5, gives conditions under which an equation of the form $F(x,y) = c_0$, where $c_0 = F(x_0, y_0)$, determines an implicitly defined function in any small enough rectangle containing

[32] The informal terminology "implicit function" is a less precise (and rather lazy) but, nowadays, unfortunately common phrase meaning "implicitly *defined* function". The only *good* use of the term "implicit function" is as the first two words in the title of the Implicit Function Theorem, where it spares us from having to call this theorem the "Implicitly-Defined-Function Theorem".

(x_0, y_0). Furthermore, the implicitly defined functions ϕ given by this theorem are actually *differentiable* (in fact, continuously differentiable; i.e. the derivative of each implicitly-defined function is continuous).

Given an equation $F(x, y) = c_0$, the single condition (3.39) on a function ϕ (the condition "$F(x, \phi(x)) = c_0$ on some interval") is **much weaker** than the if-and-only-if statement in the second line of (3.48); i.e., weaker than "ϕ is an implicitly defined function determined by the equation $F(x, y) = c_0$". Therefore it should not be said that a function ϕ known only to satisfy equation (3.39) is *defined by* the equation $F(x, y) = c_0$, or even (somewhat less objectionably) that such a ϕ is *determined by* the equation $F(x, y) = c_0$. However, equation (3.39) is still an important condition all by itself; it's all that's needed for implicit differentiation to be valid. For this reason, we give this property its own name:

Definition 3.23 Let F be a function of two variables and let $c_0 \in \mathbf{R}$. If a differentiable function ϕ of one variable satisfies $F(x, \phi(x)) \equiv c_0$ on some open interval I, we will say that the equation $F(x, y) = c_0$ *semi-determines* the function ϕ. (Note that this condition is equivalent to: *the graph of $y = \phi(x)$, over the interval I, is part of the graph of $F(x, y) = c_0$.*) ■

Warning: "Semi-determines" is **not** standard terminology; it's something I *made up* for this version of these notes.[33] Some current DE textbooks, either explicitly or implicitly (no pun intended) take the condition $F(x, \phi(x)) \equiv c_0$ as their definition of "$F(x, y) = c_0$ *defines*, or *determines*, the function ϕ." [34]

Note: **Definitions 3.21 and 3.23 apply with "$F(x, y) = c_0$" replaced by the more general equation-form $F_1(x, y) = F_2(x, y)$, in which case equation "$F(x, \phi(x)) = c_0$" is replaced by $F_1(x, \phi(x)) = F_2(x, \phi(x))$ in Definition 3.23 and in any references to equations** (3.38) **and**(3.39)**.**

[33] I did this in order to avoid terminological distinctions that I felt were too subtle in earlier versions of the notes, in which I was trying to use terminology that was closer to poor terminology used in most DE textbooks—an attempt that was putting me in a straightjacket. I will change this terminology in future versions of these notes, if I find a better alternative to handling the important distinctions between the concepts being defined in Definition 3.21, and 3.23.

[34] For example, [3] does this in (what it labels as) its definition of "implicit solution of a DE" (which we have not yet defined in these notes). In that book, the only clue as to the meaning the authors ascribe to "$F(x, y) = 0$ defines one or more [functions of x]" is in an exercise that states a very weak theorem that the book misidentifies as the Implicit Function Theorem.

3.2.6 Implicit solutions, and implicitly *defined* solutions, of derivative-form DEs

Now, let's get back to differential equations. Shortly, we will define several terms involving the word "solution": *implicitly defined solution*, *strongly implicitly defined solution, implicit solution*, and *strong implicit solution.* Current DE textbooks use the term "implicit solution", without giving a definition that is clear, accurate, complete, or sensible.[35]

A term that every DE textbook *should* define (but doesn't), and distinguish from "implicit solution", is "implicitly *defined* solution". (Of these two objects, only an implicitly *defined* solution is truly a solution of a DE.) However, **"*strong implicit solution*" and "*strongly implicitly defined solution*" are terminology that I made up purely for these notes**[36] (and just to be used with honors students), in order to have some "solution" terminology that corresponded to the (correct) definition of *implicitly defined function*.

If the only DEs we wished to consider were those that are algebraically equivalent, on the whole xy plane, to a standard-form DE $\frac{dy}{dx} = f(x, y)$, where f satisfies the conditions of the FTODE on the whole xy plane, then we could define "implicit defined solution of a DE" to be what the terminology suggests it should mean: an implicitly defined *function* that is a solution of the DE. However, even in an introductory DE class we consider many DEs that are algebraically equivalent to a standard-form DE only on an open region that is not all of $\mathbf{R}^2$. For even some rather simple DEs of this type, the relation expressed in Definition 3.23 *is* important, even though it is much weaker than "implicitly defined function" (Definition 3.21); for dome DEs, implicitly defined functions are not a wide enough class to let us express all solutions efficiently. (See Example 3.30, later in these notes, for example.)

An adequate definition of "solution of a DE implicitly determined by $F(x, y) = 0$" should allow us to say, for example, that the family of equations $\{x^2 - y^2 = \text{constant}\}$ implicitly determines *all* solutions of $x - y\frac{dy}{dx} = 0$. [Students in my Fall 2024 class: At the time this page of notes is part of your reading assignment, you have not yet seen why this ought to be true.] However, among the solutions of this DE are $y = \phi_1(x) = x$ and $y = \phi_2(x) = -x$, both on the interval $\mathbf{R}$. Definition 3.24, below, allows us to say that these functions are *implicitly defined solutions* of (3.70) determined by the

[35] The term "implicit solution" was not a formally defined term in DE textbooks when I was a student—at least in any textbook that I saw at the time, or textbook of similar vintage that I've tracked down since then. The (unnecessary and very flawed) definitions introduced into textbooks since then seem not to have been being subjected to critical examination by textbook-reviewers or by most instructors.

[36] I could make an argument that these are the objects that should be called "implicit solution" and "implicitly defined solution", respectively, and that the objects that I'm currently using those terms for should have the word "weak" or "weakly" in their title. But since the DE course I teach uses a popular modern textbook, I wanted my definition of "implicit solution" to be what I think current textbook-authors meant their (careless and/or ambiguous) definitions to achieve.

equation $x^2 - y^2 = 0$. But for neither ϕ_1 nor ϕ_2 is there an open rectangle $I \times J$ containing $(0, 0)$ (a point on both solution curves) such that condition (3.49) holds.

Definition 3.24 (implicitly defined solution) Let G be a three-variable function and let F_1 and F_2 be two-variable functions. If ϕ is a differentiable (one-variable) function that is semi-determined by the equation

$$F_1(x, y) = F_2(x, y) \tag{3.51}$$

(see Definition 3.23, and ϕ is also a solution of the differential equation

$$\mathsf{G}(x, y, \frac{dy}{dx}) = 0, \tag{3.52}$$

(3.52), we say that ϕ is an *implicitly defined solution* of equation (3.52), determined (implicitly) by equation (3.51).

Furthermore, if the solution ϕ is not just semi-determined by equation (3.51), but is actually implicitly *defined* by (3.51)—i.e. if ϕ is truly an implicitly defined function (see Definition 3.21)—then we will call ϕ a *strongly implicitly defined solution* of equation (3.52), determined (implicitly) by equation (3.51). ■

Definition 3.25 (implicit solution) Let G, F_1, F_2 be as in Definition 3.24. We call equation (3.51) an *implicit solution* of the differential equation (3.52) if

(i) equation (3.51) semi-determines (see Definition 3.23) at least one solution of equation DE (3.52), and

(ii) *every* differentiable function ϕ that is semi-determined by equation (3.51) is a solution of (3.52).[37]

[37] *Note to instructors:* (a) The given definition of "implicitly defined solution" (Definition 3.24) *does not, AND SHOULD NOT, rely at all on implicit differentiation of the equation* $F_1(x, y) = F_2(x, y)$, and neither should the definition of "implicit solution" (Definition 3.25). The function F need not even be continuous, let alone differentiable, for the concept of "solution implicitly defined by $F_1(x, y) = F_2(x, y)$" 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 than enlighten them.) An implicitly defined solution of a DE is simply an implicitly semi-defined function that happens to be a solution of the DE. The *notion* of implicitly defined, or semi-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. But 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." (Unfortunately, even some instructors may think this answer is correct.)

If equation (3.51) is an implicit solution of (3.52), and every differentiable function ϕ that is semi-determined by (3.51) is actually implicitly defined by (3.51), then we will call (3.51) a *strong* implicit solution of (3.52). ■

The notion of *maximality* (or *inextendibility*) of solutions of DEs applies also in the context of Definiton 3.25: for every function satsifying criterion (ii), there is some maximal open interval I to which ϕ can be extended to a differentiable function $\tilde{\phi}$ satisfying $F_1(x, \tilde{\phi}(x)) = F_2(x, \tilde{\phi}(x))$ on I. To check whether criterion (ii) is satisified, it suffices to check that every function that is *maximal* (= *inextendible*) among differentiable functions semi-determined by equation (3.51), is implicitly *defined* by (3.51).

Example 3.26 Consider the differential equation

$$x + y\frac{dy}{dx} = 0. \tag{3.53}$$

We claim that the equation

$$x^2 + y^2 = 1 \tag{3.54}$$

is a strong implicit solution of (3.53). To verify this, first we check that criteria (i) and (ii) of Definition 3.25 are satisfied:

- Criterion (i). Let $F(x, y) = x^2 + y^2$ and, as in (3.43)–(3.44), and let S be the the graph of $F(x, y) = 1$ (the *unit circle*). Let $\phi_1(x) = \sqrt{1 - x^2}$ and $\phi_2(x) = -\sqrt{1 - x^2}$, but restricted to the open interval $(-1, 1)$. Then ϕ_1 is the function implicitly defined by $F(x, y) = 1$ in the rectangle $(-1, 1) \times (0, \infty)$

  An easy computation yields $\phi_1'(x) = \frac{-x}{\sqrt{1-x^2}}$. Substituting $y = \phi_1(x)$ into the left-hand side of (3.53), we then find that

---

Few students, if any, will mention any relation to the notion of implicitly defined (or semi-defined) *function*, or to any relation to (true) solutions of the DE ("explicit solutions", in the needless and very misleading terminology introduced into recent editions of textbooks such as [3]). In fact, if you ask your students what an implicitly defined *function* is, most may very well reply as if you'd asked "What's an implicit solution of a DE?" even though you haven't mentioned "DE" or "solution". 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. I suggest trying Example 3.32 on your students, and perhaps also Example 3.31.

(b) The textbooks I've seen that attempt to define "implicit solution" take criterion (i) alone as the definition (and assume that students are already clear on what it means for an equation $F_1(x, y) = F_2(x, y)$ to determine a function of x, an assumption I think is perilous since the modern Calculus 1-2-3 textbooks I've seen do not cover this with any clarity). But criterion (i) alone leads to a nonsensical definition, as illustrated shortly in Example 3.27. Many older textbooks avoid this problem by *not attempting to formally define "implicit solution"*; see footnote 42.

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

so ϕ_1 is a solution of (3.53). Thus ϕ_1 is an implicitly defined solution of (3.53). Hence criterion (i) in Definition 3.25 is satisfied.

- Criterion (ii). Suppose ϕ is any differentiable function semi-determined by (3.54) on some open interval I. Then we have

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

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

$$2x + 2\phi(x)\phi'(x) = 0 \ \text{ for all } x \in I. \tag{3.55}$$

Therefore ϕ is a solution of the equation

$$2x + 2y\frac{dy}{dx} = 0 \tag{3.56}$$

on I. Dividing by 2 we see that ϕ is a solution of (3.53) on I. Therefore criterion (ii) is satisfied, (in addition to criterion (i)), and the equation $x^2 + y^2 = 1$ is an implicit solution of (3.53).

Finally, with ϕ_1, ϕ_2 as in the first bullet point, work similar to what we did with ϕ_1 shows that the function ϕ_2 is implicitly defined by $F(x,y) = 1$ in the rectangle $(-1,1)\times(-\infty,0)$). The graphs of ϕ_1 and ϕ_2 contain every point of the circle S except $(1,0)$ and $(-1,0)$. Neither of the latter two points is contained in the graph of any solution of equation (3.53), since if $(1,0)$ or $(-1,0)$ were in the graph of a solution ϕ, then equation (3.55) would imply that $2\cdot(\pm 1) + 2\cdot 0 = 0$. Thus the implicitly defined functions ϕ_1 and ϕ_2 are maximal among all differentiable functions of x that are *semi*-defined by $x^2 + y^2 = 1$, so this equation is a *strong* implicit solution of the DE (3.53). ■

Example 3.27 We claim that

$$(y - e^x)(x^2 + y^2 - 1) = 0 \tag{3.57}$$

satisfies criterion (i) in Definition 3.25 but not criterion (ii), and hence is *not* an implicit solution of (3.53).

To see this, first note that, from Example 3.26, the function ϕ_1 defined by $\phi_1(x) = \sqrt{1-x^2}$ is a solution of (3.53) on the interval $(-1, 1)$. On this interval, if we subsitute $y = \sqrt{1-x^2}$ into (3.57), the factor "x^2+y^2-1" is identically 0, so $(y-e^x)(x^2+y^2-1)$ is also identically 0. Thus ϕ_1 is a differentiable function that is implicitly semi-determined by (3.57), and is also a solution of (3.53). Hence criterion (i) in Definition 3.25 is satisfied: equation (3.57) semi-determines at least one solution of (3.53).

However, if we substitute $y = e^x$ into (3.57), we also get a true statement (for all real x). Thus, the function ϕ defined on any open interval I by $\phi(x) = e^x$ is semi-determined by (3.57). However, if we substitute $y = e^x$ into (3.53), we get

$$x + e^{2x} = 0. \tag{3.58}$$

Is it possible to choose the interval I in such a way that (3.58) holds true for all $x \in I$? No, since, for any interval I, if we were to define a function ϕ on I by $\phi(x) =$ the left-hand side of (3.58), then ϕ would be differentiable, and we could differentiate both sides of (3.58), obtaining

$$1 + 2e^{2x} = 0. \tag{3.59}$$

But there isn't even a single value of x for which equation (3.59) is true; $1 + 2e^x > 0$ for all x. Thus there is no open interval I on which ϕ is a solution of the DE (3.53).

Thus ϕ is a differentiable function that is semi-determined by (3.57) but is not a solution of (3.53). Therefore criterion (ii) in Definition 3.25 is not met, so equation (3.57) is not an implicit solution of the DE (3.53). ■

Example 3.28 The equation

$$x^2 + y^2 + 1 = 0 \tag{3.60}$$

is *not* an implicit solution of (3.53) (even though implicitly differentiating (3.60) with respect to x yields equation (3.53)), because it fails criterion (i) of Definition 3.25. There are no real numbers x, y at all for which (3.60) holds, let alone an open interval I on which (3.60) semi-determines a function of x. Since (3.60) determines no functions ϕ whatsoever on any open interval I, criterion (ii) of Definition 3.25 is moot.

Similarly, the equation

$$x^2 + y^2 = 0 \tag{3.61}$$

is not an implicit solution of (3.53). In this case there *is* a pair of real numbers (x, y) that satisfies (3.61), but there is no *open x-interval* I on which, for each $x \in I$, there is a real number y for which (3.61) is satisfied. ■

Now let us make a paradoxical, but true, observation about implicit solutions:

An implicit solution of a DE is not a solution of that DE. (3.62)

The reason is simple. A solution of a DE is a *function* (of one variable). An implicit solution of a DE is an *equation* (in two variables). These are two completely different animals.[38]

Note that the terminology "implicitly defined solution" suffers from no such problem. An implicitly *defined* solution is a function of one variable. Similarly, "solution *in implicit form*" has no such problem. Both are careful not to conflate *function of one variable* with *more-general relation between two variables*. The reason for the terminological paradox (3.62) is that *in recent decades, textbook authors decided that the term "implicit solution" ought to have a formal meaning, with a definition, and the terminology caught on (perhaps because of its brevity).*

We have actually seen a special instance of the paradox (3.62) once before, in a situation in which it did not appear paradoxical. This was in Section 3.2.1, when we said that if ϕ is a solution of a differential equation $\mathsf{G}(x, y, \frac{dy}{dx}) = 0$, we would permit ourselves to call the equation $y = \phi(x)$ a solution of the DE, regarding this phrasing as "permissible abuse of terminology". Note that "$y = \phi(x)$" is an equation of the form $F_1(x, y) = F_2(x, y)$, just with very special functions F_1 (defined by $F_1(x, y) = y$, with no dependence on x) and F_2 (which has no dependence on y). When a formula for ϕ is given, "$y = \phi(x)$" is effectively a *definitional* equation for a function ϕ (which has no named input or output variable), couched as as a *restrictive* equation in *two variables with specific names*. So, ironically, what your textbook may call an "explicit solution" of a DE, is actually an *implicit* solution! It's just a *very special type* of implicit solution. But allowing this *minor* "abuse of terminology" doesn't mean it was a good idea to open the floodgates with "implicit solution", completely blurring the distinction between "function of one variable" and "equation in two variables". For many students, the term "implicit solution" *does* lead to a misunderstanding of what a solution of an ODE really *is*.

[38] *Note to instructors*: This is why I cannot tolerate textbooks' increasingly sloppy usage of the term "implicit solution". One of our jobs as teachers of a DE course is to make sure that students understand that a solution of a derivative-form ODE is a *function of one variable*. Lumping "implicit solutions" together with *true* solutions (and, even worse, muddying the water further by introducing the horrible term "explicit solution") may make a DE instructor's life easier, may make it easier to cover more topics in a semester, and may make it easier for students to get answers the instructor will count as correct, but I don't believe that these outcomes justify keeping students in a fog about what ODEs and their solutions *are*.

Our approach to Example 3.26 relied on our ability to produce an explicit formula for a "candidate solution" of the given DE. What if, in place of (3.54), 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 can come to the rescue.

Example 3.29 [39] Show that the equation

$$x + y + e^{xy} = 1 \tag{3.63}$$

is an implicit solution of

$$(1 + xe^{xy})\frac{dy}{dx} + 1 + ye^{xy} = 0. \tag{3.64}$$

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 5.13). We compute

$$\frac{\partial F}{\partial x}(x, y) = 1 + ye^{xy}, \tag{3.65}$$

$$\frac{\partial F}{\partial y}(x, y) = 1 + xe^{xy}. \tag{3.66}$$

Both of these functions are continuous on the whole xy plane, and $\frac{\partial F}{\partial y}(0, 0) = 1 \neq 0$. Thus, the hypotheses of Theorem 5.13 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 $x \in I_1$.

Now we use the same method by which we checked criterion (ii) in Example 3.53: 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

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

$$x + y(x) + e^{xy(x)} = 1 \quad \text{for all } x \in I_1,$$

$$\implies \qquad 1 + \frac{dy(x)}{dx} + e^{xy(x)} \left(y(x) + x\frac{dy(x)}{dx} \right) = 0 \quad \text{for all } x \in I_1,$$

$$\implies \qquad (1 + xe^{xy(x)})\frac{dy(x)}{dx} + 1 + y(x)e^{xy(x)} = 0 \quad \text{for all } x \in I_1.$$

Therefore ϕ is a solution of (3.64). Thus, criterion (i) in Definition 3.25 is satisfied. The exact same implicit-differentiation argument shows that if ψ is *any* differentiable function semi-determined on an open interval by (3.63), then ψ is a solution of (3.64). Therefore criterion (ii) in Definition 3.25 is also satisfied. Hence (3.63) is a (strong) implicit solution of (3.64). ■

Looking back at Example 3.26, could we have shown that criterion (i) of Definition 3.25 is satisfied using the technique of Example 3.29, 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 $\frac{\partial F}{\partial x}(x, y) = 2x$ and $\frac{\partial F}{\partial y}(x, y) = 2y$. As in Example 3.29, the partial derivatives of F are continuous on whole xy plane again[40], and since we are choosing a point (x_0, y_0) for which $y_0 \neq 0$, we have $\frac{\partial F}{\partial y}(x_0, y_0) \neq 0$. 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 3.26 (and as we did to check both criteria in Example 3.29), to show that ϕ is a solution of (3.53). If our point (x_0, y_0) has $y_0 > 0$, then the solution of (3.53) 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 (3.53) that we get is $-\phi_1$.

The student may wonder how we could have used the method of Example 3.29 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 $\frac{\partial F}{\partial y}$ is not 0. For example, had we started with the equation

$$x + y + e^{xy} = 2 \tag{3.67}$$

[40]This does not always happen—Examples 3.26 and 3.29, and several other examples in these notes, just happen to have F's with this property.

instead of (3.63), we would have had a much harder time. We could show by implicit differentiation that every differentiable function determined by (3.67) is a solution of (3.64)—thus, that criterion (ii) of Definition 3.25 is satisfied—but that would not tell us that there is even a single function of x defined by (3.67), or even that the graph of (3.67) contains any points whatsoever. Conceivably, we could be in the same situation as in Example 3.28, in which all differentiable functions implicitly defined by (3.60)—all none of them—are solutions of our differential equation.

As you probably noticed, in Example 3.29 our expressions (3.65)–(3.66) for the partial derivatives of F appeared also in (3.64). 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

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial y}\,\frac{dy}{dx} = 0. \tag{3.68}$$

With foresight, the author chose the DE to be exactly the equation (3.68) for $F(x, y)$ equal to the left-hand side of (3.63).

It may seem to you that I cheated, by choosing essentially the only DE for which the fact you were instructed to establish was actually true. But you will see later that equations of the form (3.68) actually come up a lot.

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 3.30 Consider the algebraic equation

$$x^2 - y^2 = 0 \tag{3.69}$$

and the differential equation

$$x - y\frac{dy}{dx} = 0. \tag{3.70}$$

For each *fixed real number* x, (3.69) is equivalent to the assertion that the real *number* y is either x or $-x$, a statement that we may write as "$y = \pm x$." However, if we consider (3.69) as a relation involving an independent *variable* x, and regard the unknown object as a *function* of x represented by the letter y (a dependent variable), then (3.69) is equivalent to $y = s(x)x$, where s can be *any* function satisfying $s(x) = \pm 1$, not necessarily the same sign for each x. (See the discussion in the paragraph that starts with the line above equation (3.40) and concludes a few lines

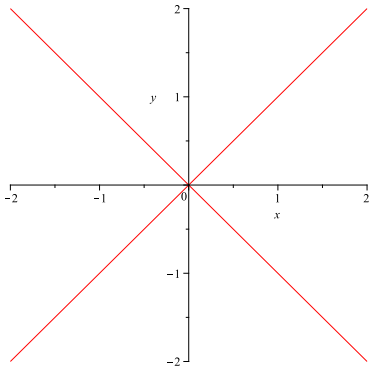

Figure 3: The graph of $x^2 - y^2 = 0$.

below equation (3.45).) If we decide that the only unknowns y we are interested in are *differentiable* functions of x on some interval I—as is automatically the case in equation (3.70)—then (3.69) is equivalent to $y = \pm x$, where the sign is the same for all $x \in I$. Thus on any interval I, equation (3.69) semi-determines exactly two differentiable functions ϕ of x on any open interval including 0, namely $\phi(x) = x$ and $\phi(x) = -x$. Both of these are solutions of (3.70). Therefore (3.69) is an implicit solution of (3.70), and the two functions ϕ above are implicitly-defined solutions of (3.70), on any interval.

The point $(x, y) = (0, 0)$ satisfies (3.69). But on no open rectangle containing the point $(0, 0)$ does (3.69) 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 (5.15) holds.

Does this contradict the Implicit Function Theorem? No—the theorem says only that *if the hypotheses of the theorem are met*, then there are intervals I_1 and J_1 with the property (5.15)). But in the current example, the function F and number c_0 for which (3.69) is the equation $F(x, y) = c_0$ are $F(x, y) = x^2 - y^2$ and $c_0 = 0$. Thus $\frac{\partial F}{\partial y}(x, y) = -2y$, and if we take $(x_0, y_0) = (0, 0)$ (a point satisfying $x^2 - y^2 = c_0$) then $\frac{\partial F}{\partial y}(x_0, y_0) = 0$. 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 (3.70); they just are not *strongly* implicitly-defined solutions, the only ones that the Implicit Function Theorem informs us about. ■

For most two-variable functions F that we encounter in practice, the "bad" points (x_0, y_0) in the domain of F—the points at which the Implicit Function Theorem does not apply to the equation $F(x, y) = F(x_0, y_0)$—are of two types: points at which the graph of $F(x, y) = F(x_0, y_0)$ has a vertical tangent (all of the "bad" points in Figures

1 and 2 are of this type), and points at which two or more smooth curves intersect (the only "bad" point in Figure 3 is of this type).

The nature of the "bad" point in Figure 3 leads to a phenomenon that was not present in our earlier examples. On any open x-interval containing 0, the equation $x^2 - y^2 = 0$ 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 our previous examples, the *continuous* implicitly-defined functions and the *differentiable* implicitly-defined functions were the same (on any open interval).

From the examples presented so far, and the treatment in most textbooks, the student[41] may get the false impression that "implicit solution" means "An equation that, after I implicitly differentiate, I can rearrange back to the given DE." That is *not* the definition, however (Definition 3.25 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 3.31 Determine whether the equation

$$2|x| + |y| = 2 \tag{3.71}$$

is an implicit solution of

$$\left(\frac{dy}{dx}\right)^2 = 4|x| + 2|y|. \tag{3.72}$$

If we try to approach this problem just by implicit differentiation, we run into trouble because the function $F(x, y) = 2|x| + |y|$ is not differentiable at any point at which $x = 0$ or $y = 0$. However, if we run through all the sign-possibilities in equation (3.71) and solve for y in terms of x, we see that the graph of (3.71), a "stretched diamond" with vertices at $(\pm 1, 0)$ and $(0, \pm 2)$, consists of the graphs of the following four equations:

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

Therefore equation (3.71) determines the following four differentiable functions:

[41] Or even the instructor!

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

Every differentiable function of x determined by equation (3.71), with domain an open interval, is one of these four functions (or the restriction of one of these functions 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 equation (3.72), we find

$$\begin{aligned}
\text{left-hand side of (3.71)} &\equiv 4,\\
\text{right-hand side of (3.71)} &= 2(2|x| + |y(x)|)\\
&\equiv 2\times 2 \quad \text{(because of equation (3.71))}\\
&= 4.
\end{aligned}$$

Therefore for all four choices of ϕ, equation (3.72) is satisfied on the domain of ϕ. Both criteria in the definition of "implicit solution" (Definition 3.25) are satisfied, so equation (3.71) is an implicit solution of the DE (3.72). ■

Example 3.32 Determine whether the equation

$$y^5 + y = x^3 + x \tag{3.73}$$

is an implicit solution of

$$\frac{dy}{dx} = \frac{3x^2+1}{5(x^3+x-y)^{4/5}+1}\ . \tag{3.74}$$

First, we observe that the graph of (3.73) has at least one point: the point $(0,0)$.

Next, we rewrite (3.73) as $F(x,y) = 0$, where $F(x,y) = y^5 + y - x^3 - x$. Then $\frac{\partial F}{\partial x} = -(3x^2+1)$ and $\frac{\partial F}{\partial y} = 5y^4 + 1$, both of which are continuous on the whole xy plane; furthermore, $\frac{\partial F}{\partial y}$ is nowhere zero (it's positive at every point, in particular at $(0,0)$). Hence the Implicit Function Theorem guarantees us that (3.73) determines a differentiable function of x near the point $(0,0)$ (a point on the graph of $F(x,y) = 0$).

So (3.73) determines at least one differentiable function of x. If ϕ is any such function, then substituting $y = \phi(x)$ into (3.73) and differentiating implicitly, we find $(5y^4+1)\frac{dy}{dx} = 3x^2 + 1$, which implies

$$\frac{dy}{dx} = \frac{3x^2+1}{5y^4+1} \tag{3.75}$$

on the domain of ϕ (the denominator $5y^4+1$ is never zero). Hence ϕ is a solution of (3.75).

Now, (3.75) does not look like (3.74). The two DEs are not equivalent; there are points (x, y) at which the right-hand side of (3.74) is not equal to the right-hand side of (3.75). But that doesn't mean that (3.73) can't be an implicit solution of (3.74). And, in fact, on the graph of (3.73) we have $y^5 = x^3 + x - y$, implying $y^4 = (x^3 + x - y)^{4/5}$. Therefore for $y = \phi(x)$ we have

$$\frac{dy}{dx} = \frac{3x^2+1}{5y^4+1} = \frac{3x^2+1}{5(x^3+x-y)^{4/5}+1} \ ,$$

so ϕ is a solution of (3.74). Therefore (3.73) is an implicit solution of (3.74). ■

In the example above, it is irrelevant whether there are *some* solutions of (3.75) that are not solutions of (3.74). The question was not whether *every* solution of (3.75) was a solution of (3.74), but only whether a *specific* solution of (3.75), namely a function determined implicitly by (3.73), was a solution of (3.74).

Remark 3.33 (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 (3.68). But for any such F and constant C_0, the DE (3.68) 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.[42] However, you are unlikely to find examples like Example 3.31 or Example 3.32 in a

[42] *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've seen in textbooks that formalize the term, is an example of something much more restricted: an element of a *family* of implicit solutions $\{F(x, y) = C\}$. These books are unnecessarily defining something that they effectively never use, *single* implicit solutions outside the context of some easily-expressed *family* of equations. 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 $x + y\frac{dy}{dx} = 0$, *in implicit form*, is $\{x^2 + y^2 = c : c > 0\}$." (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

DE textbook. In a typical DE course, implicit solutions tend to arise from solving DEs that are either separable or exact (types of equations we will cover in class, but may not yet have covered at the time you're reading this). For any of these DEs, there is always a *family* of implicit solutions (which does not always yield all of the DE's solutions, in the separable case) of the form

$$\{F(x,y) = C\}\,, \tag{3.76}$$

where F is function that depends on the DE, and C is a constant ranging over some subset I of the real line.[43] (I.e. for each C in I, the equation $F(x,y) = C$ is an implicit solution of the DE.) *Every* differentiable function that is semi-determined by *any* member of the family (3.76) is a solution of the *same* DE, namely (3.68).

For simplicity, when a DE has a family of solutions of the form (3.76), that family is often called a "*one-parameter* family of implicit solutions" (the parameter being C), even when the set I of allowed values of C is not the whole real line. More generally, a collection of equations of the form $\{\tilde{F}(x,y,C) = 0\}$, where $\tilde{F}$ is a three-variable function, may be called a "one-parameter family of equations in x and y". (Note that the family (3.76) can be recast in this form, with $\tilde{F}(x,y,C) = F(x,y) - C$.) For a given differential equation $\mathsf{G}(x, y, \frac{dy}{dx}) = 0$, and a specific function $\tilde{F}$ for which "$\tilde{F}(x,y,C) = 0$" is an implicit solution of the DE for all C ranging over some subset I of the real line that contains an open interval, the set $\{\tilde{F}(x,y,C) = 0 : C \in I\}$ is also often called one-parameter family of implicit solutions. For example[44] the family of equations $\{x^2 + Cy^2 = 1 : C \neq 0\}$ is a one-parameter family of implicit solutions of $\frac{dy}{dx} = \frac{xy}{1-x^2}$.

3.2.7 General solutions in implicit form (for a derivative-form DE)

Sometimes we can write down an explicit expression for every solution of a derivative-form DE.[45] In this case, we usually write the general solution as a collection of equations expressing these formulas, as in Examples 3.11–3.15. (In those examples, we did

reason I've 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 is needed that is accurate, precise, complete, understandable by students, and sensible.

[43] The subset I is often difficult to specify. However, in typical examples I is an interval, and is sometimes the whole real line.

[44] This is exercise 1.2/16 in [3].

[45] *Note to instructors*: An equation of the form "$\phi(x) =$ explicit formula in terms of x," or an equation of the form "(dependent variable) = (explicit formula for a solution, in terms of the independent variable)," can *reasonably* be called an "explicit solution". *Nothing else merits this terminology.* Using it with any other meaning *inflicts harm.* Of course, there is some subjectivity as to what formulas are "explicit"—e.g. most mathematicians would regard "$\int_0^x e^{t^2}\,dt$" as explicitly defining a function of x, but most students in an introductory DE course would not. However, (i) that discrepancy only *exacerbates* the ambiguity in what "explicit solution" might mean, and (ii) there are plenty of differentiable functions for which virtually *no* mathematican would say there is

not fully *justify* that we'd found all the solutions; at that time we simply wanted to illustrate the concept of "general solution" in a few examples with which students might already be familiar. For the linear DEs in Examples 3.11–3.13, the "Fundamental Theorem of *Linear* ODEs" [not included in these notes at this time] guarantees that, in each of these examples, the collection of equations we wrote down *does* give *all* maximal solutions of the given DE. For the *nonlinear* examples 3.14 and 3.15, this conclusion follows from results proven in Section 3.2.10.)

There are other times in which we can't find explicit formulas for all (or perhaps any!) solutions of a DE, but are able to to find a collection of implicit solutions that (at least semi-) determine every solution. In this case it seems reasonable to say that we have found the general solution *in implicit form*. For example, for a given DE, we may be able to show that there is a two-variable function F such that every solution of the DE is (at least semi-) determined by the equation $F(x, y) = C$ for some constant C, and conversely for which every such equation (possibly with some restrictions on C) is an implicit solution. In this case we would like to be able to say that the collection of equations $\{F(x, y) = C\}$ "is" the general solution, at least in implicit form.

Thus, whether we can find all solutions explicitly, or can find them only in implicit form, when we want to *write down* a general solution of a DE we almost always do so (or attempt to) by writing down a *collection of equations in the independent and dependent variables* (for which we will continue to use the letters x and y, respectively).[46] These equations in this collection are *algebraic equations*, not differential equations; they are of the form $F_1(x, y) = F_2(x, y)$ for some two-variable functions F_1 and F_2. (This form includes the case in which F_1 or F_2 is zero or some other constant function.) Sometimes one of these functions may depend only on x, and the other may depend only on y. Sometimes we may have $F_1(y) = y$, and F_2 an explicitly expressed function of x alone, in which case $F_1(x, y) = F_2(x, y)$ expresses y explicitly as a function of x. In all cases, whenever we find it convenient we can write "$F_1(x, y) = F_2(x, y)$" as $F(x, y) = 0$, where $F = F_1 - F_2$.

[**Note to MAP2302 students**: You may find criterion (ii) in the following definition very difficult to understand, because of the term "locally contained". **Don't worry**; that terminology is above the level appropriate for an intro DE class. **Don't spend too much time trying to understand it; move on.** For *separable* DEs satisfying all the hypotheses I used in class—the only type for which we currently need a definition of "general solution in implicit form"—the word "locally" can be deleted from the definition below.]

---

an explicit formula. Calling a function an "explicit solution" in the absence of any explicit formula for that function, invents a new meaning for "explicit" that is precisely the *opposite* of its dictionary definition.

[46] A *collection* of equations is no different from a *set* of equations; we are using a different word simply to emphasize that we are not talking about a set in the xy plane.

Definition 3.34 (General solution on a region, in implicit form) [47] For a given three-variable function G, consider the derivative-form DE

$$\mathsf{G}(x, y, \frac{dy}{dx}) = 0. \tag{3.77}$$

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* (3.77) *in* R*, in (an) implicit form,* if the following two conditions are satisfied:

(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 the DE (3.77) (see Definition 3.25).

(ii) Every solution curve $\mathcal{C}$ of (3.77) in R is locally contained in the graph of one and only one of the equations E in the collection $\mathcal{E}$. Here, "a curve $\mathcal{C}$ is locally contained in the graph of an equation E" means that for every point (x_0, y_0) of $\mathcal{C}$, there is an open rectangle R' containing (x_0, y_0) for which the portion of $\mathcal{C}$ in R' lies in the graph of E.

Alternatively, we refer to such a collection $\mathcal{E}$ as *an implicit form of the general solution of* (3.77) *in* R. Note that, if such a collection $\mathcal{E}$ exists, it will not be the *only* such collection (for example, any equation in $\mathcal{E}$ could be replaced by an equivalent equation); hence the terminology "*an* implicit form", not "*the* implicit form".

If no region R is mentioned explicitly, it is understood that we are taking R to be the largest region in $\mathbf{R}^2$ on which the DE (3.77) makes sense: the set $\{(x, y) \in \mathbf{R}^2 : \text{the expression } \mathsf{G}(x, y, z) \text{ is defined for } \textit{some} \text{ real number } z\}$.[48] ■

As mentioned earlier, one example of a collection of algebraic 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 3.34. There are DEs for which we can write down the general solution, in implicit form, perfectly well, but for which it may be difficult or undesirable (if even possible) to express the general solution by a one-parameter family of equations.

Note that we are *not* asserting that we will always be able to *find* a general solution of a DE (with or without the "in a region R"). However, there are several

[47] The terminology in this definition was invented purely for these notes; it is not standard.

[48] For example, for the equation $\frac{dy}{dx} = x^2 + y^2 - 1$ this set is all of $\mathbf{R}^2$; for the equation $\frac{dy}{dx} = \frac{1}{x^2+y^2-1}$ this set is $\mathbf{R}^2$ with the circle $x^2 + y^2 = 1$ deleted. For a completely arbitrary three-variable function G, this set might not be open, in which case it would not fit our definition of *region*. However, under extremely mild conditions on G, satisfied by all DEs we consider in these notes, this set *will* be open.

types of DEs—which tend to be the ones studied in an introductory course on the subject—for which we *can* write down a general solution in implicit form. There are some closely related types for which we cannot quite do this without quite a lot of extra bookkeeping, but for which we can still write down a collection of equations that may *not* be the general solution (in implicit form), but for which the full general solution, in implicit form, can still be constructed in a systematic way.

The "locally contained" in criterion (ii) of Definition 3.34 may come as an (unpleasant and confusing) surprise; you might have hoped for, or expected, just the word "contained". That hope represents *almost* the best-case scenario (see Remark 3.35), and it does occur for many DEs, such as for separable DEs satisfying some not-very-stringent requirements (see Theorem 3.44 later in these notes). But for some DEs, our methods of finding solutions lead us to a collection of equations $\mathcal{E}$ for which some solution-curves 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 this happen in Example 3.47.) 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 $\mathcal{E}$; infinitely many may be needed.

A cautionary note: Do not be misled by the terminology "the general solution of (3.77) in R, *in implicit form*." While there is only one general solution of (3.77) in R—the *collection* of all solutions whose graphs lie 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 ..." and "the general solution ... in an implicit form". 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 $\{F(x,y) - C = 0\}$, 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 3.14 and 3.15.

Remark 3.35 Suppose that $\mathcal{E}$ is an implicit form of the general solution of a given DE. Condition (ii) in Definition 3.34 does not imply that the graphs of equations in $\mathcal{E}$ don't intersect each other; the definition does not even prevent two graphs from *overlapping* along some segment. A solution curve (maximal or otherwise) could *intersect* the graph of more than one equation in $\mathcal{E}$ without lying entirely in more than one graph. However, *if* it happens that the graphs of no two equations in $\mathcal{E}$ intersect, then condition (ii) implies that *every* solution curve in R—not just *maximal* solution curves—lies in the graph of one and only one equation in $\mathcal{E}$. This is a "best

of all worlds" situation for general solutions in implicit form. ■

3.2.8 Algebraic equivalence and general solutions of derivative-form DEs

Some algebraic manipulations that help us solve DEs have the potential to change the solution-set, either losing some solutions of the original DE or introducing spurious "solutions" that are not solutions of the original DE.[49] In this section of the notes, we discuss how to be aware of whether a given algebraic manipulation on a given DE has the potential to cause such a problem, and to deal with this problem if it is actually present.

Definition 3.36 We say that two derivative-form differential equations, with independent variable x and dependent variable y, are *algebraically equivalent n 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 $(x, y) \in R$ [50], 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 3.36, even though they are not mentioned explicitly.

Example 3.37 The differential equations

$$\frac{dy}{dx} = y(1-y) \tag{3.78}$$

and

$$\frac{1}{y(1-y)}\,\frac{dy}{dx} = 1 \tag{3.79}$$

[49] 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. Even some worked-out examples in some textbooks suffer from this problem.

[50] *Note to students*: The expression is allowed to involve $\frac{dy}{dx}$—i.e. it could be of the form $\mathsf{G}(x, y, \frac{dy}{dx})$ for some three-variable function G—which is why we did not say "function of x and y" here. If the expression is $\mathsf{G}(x, y, \frac{dy}{dx})$, 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 $\mathsf{G}(x, y, z)$ is defined.

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. ■

Example 3.38 The differential equations

$$(x+y)\frac{dy}{dx} = 4x - 2y \tag{3.80}$$

and

$$\frac{dy}{dx} = \frac{4x-2y}{x+y} \tag{3.81}$$

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 their solution-sets are the same. 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.[51] 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:

[51] 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.

$$\left.\begin{array}{l}\text{If two differential equations are algebraically equivalent on a}\\ \text{region } R\text{, then they have the same general solution in } R.\end{array}\right\} \tag{3.82}$$

We may restate (3.82) more briefly as "Algebraically equivalent DEs have the same set of solutions," or "Algebraically equivalent 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 as meaning statement (3.82).

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 3.37 and 3.38 above.

To see why statement (3.82) is true, let us check that operation (ii) in Definition 3.36 does not change the set of solutions in R. Let us suppose we start with a (first-order) derivative-form DE of the most general possible form:

$$\mathsf{G}_1(x, y, \frac{dy}{dx}) = \mathsf{G}_2(x, y, \frac{dy}{dx}). \tag{3.83}$$

The equation obtained by multiplying both sides of (3.83) by a function h that is defined at every point of R and is nonzero on R is

$$h(x, y)\mathsf{G}_1(x, y, \frac{dy}{dx}) = h(x, y)\mathsf{G}_2(x, y, \frac{dy}{dx}). \tag{3.84}$$

Suppose that ϕ is a solution of (3.83). Then for all x in the domain of ϕ,

$$\mathsf{G}_1(x, \phi(x), \phi'(x)) = \mathsf{G}_2(x, \phi(x), \phi'(x)). \tag{3.85}$$

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 (3.85) by this number. Therefore

$$h(x, \phi(x))\mathsf{G}_1(x, \phi(x), \phi'(x)) = h(x, \phi(x))\mathsf{G}_2(x, \phi(x), \phi'(x)) \tag{3.86}$$

for all x in the domain of ϕ. Hence ϕ is a solution of (3.84). Thus every solution of (3.83) whose graph lies in R is also a solution of (3.84) whose graph lies in R.

Conversely, suppose that ϕ is a solution of (3.84) whose graph lies in R. Then (3.86) 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 (3.86) by this number. Therefore (3.85) is satisfied for all x in the domain of ϕ, so ϕ is a solution of (3.83). Thus every

solution of (3.84) whose graph lies in R is also a solution of (3.83) whose graph lies in R.

This completes the argument that multiplying by h has not changed the set of solutions in R. The argument that operation (i) in Definition 3.36 does not change this set of solutions is similar, and is left to the student. Note that subtracting $\mathsf{G}_2(x, y, \frac{dy}{dx})$ from both sides of equation (3.83) is a special case of operation (i), and is exactly what we do when we "put (3.83) in the simpler form $\mathsf{G}(x, y, \frac{dy}{dx}) = 0$." Thus, all the way back in Section 3.2.1, we were tacitly using the notion of algebraic equivalence, and the fact that operation (i) does not change the set of solutions in any given regon.

It is possible for two differential equations to be equivalent (i.e. to have the same set of solutions) without being *algebraically* equivalent. For example, performing operations other than those in Definition 3.36 does not *always* change the set of solutions. But because they *might* change the set of solutions, any time we perform one of these "unsafe" operations we must use other methods to check whether we've lost any solutions or have added any spurious solutions.

3.2.9 Algebraic equivalence and general solutions of linear DEs

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

Example 3.39 The equations

$$\frac{dy}{dx} + 3y = \sin x \tag{3.87}$$

and

$$e^{3x}\frac{dy}{dx} + 3e^{3x}y = e^{3x}\sin x \tag{3.88}$$

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 $I \times \mathbf{R}$ (an vertical strip, infinite in the $\pm y$-directions). So we will add a bit to Definition 3.36 to have language better suited to linear equations:

Definition 3.40 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 I, or y times such a 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 3.41 The equations

$$x\frac{dy}{dx} - 2y = 0 \tag{3.89}$$

and

$$x^3\frac{dy}{dx} - 2x^2y = 0 \tag{3.90}$$

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 3.36, 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 3.42 The equations

$$x\frac{dy}{dx} - 2y = 0 \tag{3.91}$$

(the same equation as (3.89) and

$$x^{-2}\frac{dy}{dx} - 2x^{-3}y = 0 \tag{3.92}$$

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 not yield a function that we can multiply by on any interval that contains 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 (3.82) reduces to the following simpler statement:

$$\begin{array}{l}\text{Two linear DEs that are algebraically equivalent}\\ \text{on an interval } I \text{ have exactly the same solutions on } I.\end{array} \tag{3.93}$$

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 3.43 Find the general solution of

$$x\frac{dy}{dx} - 2y = 0 \tag{3.94}$$

(the same equation as (3.91) and (3.89)).

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

$$\frac{dy}{dx} - \frac{2}{x}\, y = 0. \tag{3.95}$$

However, (3.94) and (3.95) are not algebraically equivalent on the whole real line, but only on $(-\infty, 0)$ and $(0, \infty)$. Equation (3.95) does not even make sense at $x = 0$, while (3.94) makes perfectly good sense there.[52]

[52] 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

As the student may verify, equation (3.95) has an integrating factor $\mu(x) = x^{-2}$. Putting our brains on auto-pilot, we multiply through by x^{-2}, and write

$$\begin{aligned} (x^{-2}y)' &= 0, \\ \Longrightarrow \quad \int (x^{-2}y)'dx &= \int 0\ dx, \\ \Longrightarrow \quad x^{-2}y &= C, \\ \Longrightarrow \quad y &= Cx^2. \end{aligned} \tag{3.96}$$

(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 (3.96).)

Neither in the original DE (3.94) nor in (3.96) do we see any of the red flags we are used to seeing, such as a "$\frac{1}{x}$", that warn us that there may be a problem with (3.96) at $x = 0$. (There were red flags in the intermediate steps, in which negative powers of x appeared, but we ignored them.) The functions given by (3.96) 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 (3.94). 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.

Let us go back to square-one and correct our work. The transition from equation (3.94) to (3.95) 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 (3.94) might include solutions on intervals that contain 0, while the general solution of (3.95) cannot.

We can still use the basic procedure that led us to (3.96); we just have to be more careful with it. Auto-pilot will not work.

Because (3.95) 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

$\frac{dy}{dx} + p(x)y = g(x)$, 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 (3.94) and (3.95).

sides of our equation on I by x^{-2}, we find that ϕ is a solution of (3.95) 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 (3.95) 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

$$\phi(x) = Cx^2. \tag{3.97}$$

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

Thus the general solution of (3.95) on $(0, \infty)$ is

$$y = Cx^2, \quad x > 0, \tag{3.98}$$

while the general solution of (3.95) on $(-\infty, 0)$ is

$$y = Cx^2, \quad x < 0. \tag{3.99}$$

Now return to the equation we originally were asked to solve, (3.94), 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 (3.94) and (3.95) are algebraically equivalent on $(0, \infty)$, ϕ_1 must be one of the solutions given by (3.98). 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 (3.99), 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 $\phi(0) = 0$.[53] 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

$$\phi(x) = \begin{cases} C_1x^2 & \text{if } x \geq 0, \\ C_2x^2 & \text{if } x < 0, \end{cases} \tag{3.100}$$

[53] 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 (3.94), 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 (3.94), it does not work for (3.90)—which the student will later be asked to solve—but the first method we presented does.

(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 (3.100) is a solution of (3.94) on $(-\infty, \infty)$. Therefore the general solution of (3.94) on $(-\infty, \infty)$ is the *two-parameter* family of functions given by (3.100), with C_1 and C_2 arbitrary constants[54]. 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 (3.100) to the interval I. (For the student who read and understood the material on maximal solutions: the two-parameter family (3.100) is the general solution of (3.94) as defined in Definition 3.10.) ■

You should not draw the wrong impression from Example 3.43. 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 is an n-parameter family of functions. You are unlikely to see a two-parameter family of functions as the general solution of a DE unless the equation is second-order. Example 3.43 is the exception, not the rule. But it does provide a simple example of the perils of what can happen when algebraic equivalence is not maintained during the manipulation of DEs.

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 (3.89) and (3.90) have the same set of solutions.

3.2.10 General solutions of separable DEs

Consider any separable DE

[54]Some authors, with a different definition of "general solution", would say that the first-order linear equation (3.94) *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. I find this an odd convention to apply to a solution-set with a completely systematic and very explicit 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*.

$$\frac{dy}{dx} = g(x)p(y), \tag{3.101}$$

for which

$$\left.\begin{array}{l} g \text{ is continuous on some open interval } I, \text{ and} \\ g \text{ is not identically zero on } I, \text{ and} \\ p \text{ and } p' \text{ are continuous on some open set } D \text{ in } \mathbf{R}. \end{array}\right\} \tag{3.102}$$

There is a redundancy in the third line of (3.102): if p' even *exists* on D, then automatically p is continuous on D. However, we will find it convenient below to have the continuity of p stated explicitly.

We are interested in making the strongest always-true statements we can about solutions of the DE (3.101) under hypotheses of the form (3.102). For this reason, if g is given by an explicit formula, we generally take I to be a "maximal open interval of continuity", i.e. an open interval on which the formula defines a continuous function, but for which the formula does not yield a continuous function on any larger open interval containing I. Similarly, if p is given by an explicit formula, we generally take D to be a "maximal open domain of continuity" of p'. Theorem 3.44 below is true whether or not we choose I or D this way; the conclusion is simply stronger if we choose I and D this way than if we don't. Very commonly, we can take I to be the whole real line.

Writing $f(x, y) = g(x)p(y)$, we have $\frac{\partial f}{\partial y}(x, y) = g(x)p'(y)$. Thus both f and $\frac{\partial f}{\partial y}$ are continuous on the region $R = I \times D$, so for any point (x_0, y_0) in R, the Fundamental Theorem and Corollary 5.11 part (a) apply to the initial-value problem for (3.101) with initial condition $y(x_0) = y_0$, and Corollary 5.11 parts (b) and (c) apply to the DE (3.101) on R.

Suppose that r is a number in D for which $p(r) = 0$. Consider the constant function ϕ defined by $\phi(x) = r$. Then $\phi'(x) = 0$ (because ϕ is constant) and $p(\phi(x))g(x) = p(r)g(x) = 0 \cdot g(x) = 0$ for all $x \in I$. Hence the constant function ϕ, with domain I, is a solution of equation (3.101) in R, and is *maximal* in R—the domain is already as large as it can be without the graph leaving R. The horizontal line $y = r$ is a maximal solution curve in R, which (by Corollary 5.11(c)) no other maximal solution curve in R can intersect. Therefore if $y_0 \neq r$, and ϕ a solution of the IVP for (3.101) with initial condition $y(x_0) = y_0$, then for *every x in the domain of ϕ, we have $\phi(x) \neq r$.*

Note that if r is a number for which $p(r) \neq 0$, the constant function ϕ defined by $\phi(x) = r$ has derivative $\phi'(x) = 0$ (identically), but $p(\phi(x))g(x) = p(r)g(x)$ is not identically zero (since g is not identically 0), so ϕ is *not* a solution of (3.101) on I.

Combining the preceding facts:

- **For each r in D, the equation $y = r$ is a (constant) solution of** (3.101)

on I if $p(r) = 0$, and is not a solution of (3.101) **on I if $p(r) \neq 0$** (cf. Remark 3.3).

- **If ϕ is a *non*-constant solution of** (3.101) **in R, then the graph of ϕ does not intersect the graph of any of the constant solutions on I.** If there are any numbers r for which $p(r) = 0$, then the graph of any non-constant solution is "trapped" in an open region bounded above and/or below by horizontal lines that are graphs of constant solutions.

Notation for Theorem 3.44 below:

- Z denotes the set $\{r \in D : p(r) = 0\}$ (the set of *zeroes* of p; if p is a polynomial these numbers are also called *roots* of p).

- Let D_1 be the set of elements of D that are *not* in Z. (Note that, depending on p, the set Z can be *empty*—p may have no zeros—in which case D_1 is all of D. If p is *identically* zero, then Z is all of D.) The set D_1 is open, because p is continuous: if $p(y_0) \neq 0$, then $p(y) \neq 0$ for all numbers y sufficiently close to y_0. For every number y in D_1, let $h(y) = \frac{1}{p(y)}$. Then the function h is continuous on D_1, and g is continuous on I, so the Fundamental Theorem of Calculus ensures us that h and g have antiderivatives on these domains.[55]

  Let H be any fixed antiderivative of h on D_1, and let G be any fixed antiderivative of g on I.

Theorem 3.44 *Assume the hypotheses* (3.102) *are satsified. Then the general solution of* (3.101) *in the region $R = I \times D$, in implicit form, is the collection of equations*

$$\mathcal{E} = \mathcal{E}_1 \bigcup \mathcal{E}_2, \tag{3.103}$$

where

$$\mathcal{E}_1 = \{H(y) = G(x) + C : C \in \mathbf{R}\} \quad \textit{and} \quad \mathcal{E}_2 = \{y = r : r \in Z\}. \tag{3.104}$$

(Note that the set Z may be empty, in which case the collection $\mathcal{E}_2$ is empty.) *The collection $\mathcal{E}_1$ is precisely the set of all* non-constant *solutions, in implicit form, of*

[55] The Fundamental Theorem of Calculus establishes, among other things, that every continuous function on an open *interval* has an antiderivative. Since an open set in $\mathbf{R}$ is (at worst) a union of nonintersecting open intervals, this implies that every continuous function on an open *set*, such as D_1, has an antiderivative. However, if D_1 is not an *interval*, then the difference between two antiderivatives of h need not be a constant.

(3.101), *while* $\mathcal{E}_2$ *is precisely the set of all* constant *solutions, in explicit (and therefore also in implicit) form. Every solution curve in* R*, whether maximal or not, lies in the graph of one and only one equation in the collection* $\mathcal{E}$*, and its graph does not intersect the graph of any other equation in* $\mathcal{E}$*.*

The symbol "$\bigcup$" in (3.103) denotes *union*: the collection $\mathcal{E}$ consists of equations that lie either in the collection $\mathcal{E}_1$ or the collection $\mathcal{E}_2$.

In Examples 3.14 and 3.15, we asserted that we had written down the general solutions of the DEs in those examples. Those assertions can now be justified using Theorem 3.44, in conjunction with some algebra that we omit from these notes. (In both of these examples we may take I and D to be the whole real line $\mathbf{R}$. In Example 3.14 we may take $p(y) = -y^2$, take $D = (-\infty, 0) \bigcup (0, \infty)$, take $g(x) = 1$, take $H(y) = \frac{1}{y}$, take $G(x) = x$, use simple algebra to solve "$H(y) = G(x) + C$" explicitly for y in terms of x, and rewrite the collection $\mathcal{E}_1$ in (3.104) as $\left\{y = \frac{1}{x-C}\right\}$. In Example 3.15 we may take $p(y) = y(1-y)$, take $D = (-\infty, 0) \bigcup (0, 1) \bigcup (1, \infty)$, take $g(x) = 1$, take $H(y) = \ln\left|\frac{y}{1-y}\right|$, take $G(x) = x$, use somewhat more-involved algebra to solve "$H(y) = G(x) + C$" explicitly for y in terms of x, and rewrite the collection $\mathcal{E}_1 \bigcup \{y \equiv 0\}$ as $\left\{y = \frac{C}{e^{-x}+C}\right\}$ [with C being an arbitrary real constant, but not having the same numerical value for a given solution as in "$H(y) = G(x) + C$"].)

Proof of Theorem 3.44. In the discussion preceding the theorem, we established that $\mathcal{E}_2$ is the set of constant maximal solutions of (3.101) on I, and that if ϕ is any non-constant solution in R, then the graph of ϕ cannot intersect the graph of any of these constant solutions. In fact, the graph of an equation in $\mathcal{E}_1$ cannot intersect the graph of an equation in $\mathcal{E}_2$ at all, since, by the definition of H, no element of the set Z is in the domain of H.

Let ϕ be a non-constant solution of equation (3.101) in R, with domain I_1 (some sub-interval of I). Then, by the preceding, *for all* $x \in I_1$ *we have* $\phi(x) \in D_1$*, and therefore* $p(\phi(x)) \neq 0$. Therefore, throughout the interval I_1 we have

$$\frac{1}{p(\phi(x))}\,\phi'(x) = g(x). \tag{3.105}$$

But by definition of the functions h and H, we have that $H' = h = \frac{1}{p}$ on D_1, so $\frac{1}{p(\phi(x))} = h(\phi(x)) = H'(\phi(x))$. But then the left-hand side of equation (3.105) is $H'(\phi(x))\phi'(x)$, which, by the Chain Rule, is precisely $\frac{d}{dx}H(\phi(x))$. Hence, on the interval I_1 we have

$$\frac{d}{dx}(H(\phi(x)) - G(x)) = g(x) - G'(x) = 0, \tag{3.106}$$

and therefore $H(\phi(x)) - G(x)$ is constant. Thus, for some $C \in \mathbf{R}$,

$$H(\phi(x)) = G(x) + C, \tag{3.107}$$

so the relation $y = \phi(x)$ satisfies the equation

$$H(y) = G(x) + C. \tag{3.108}$$

(Alternatively, instead of using (3.106), we could have reached (3.107) as follows: Let $x_0 \in I_1$. Then, for any $x \in I_1$, the Fundamental Theorem of Calculus tells us that

$$H(\phi(x)) - H(\phi(x_0)) = \int_{x_0}^{x} \frac{d}{dt} H(\phi(t))\, dt = \int_{x_0}^{x} g(t)\, dt = G(x) - G(x_0),$$

so $H(\phi(x)) = G(x) + C$, where $C = H(\phi(x_0)) - G(x_0)$.)

This establishes that

$$\left.\begin{array}{l}\text{the graph of every non-constant solution}\\ \text{of (3.101) in } R \text{ lies in the graph of one}\\ \text{of the equations in the collection } \mathcal{E}_1.\end{array}\right\} \tag{3.109}$$

Next, we claim that

$$\left.\begin{array}{l}\text{for each } C \in \mathbf{R} \text{ for which the graph of equation (3.108) in } R\\ \text{contains at least one point, there is an implicitly defined}\\ \text{function of } x \text{ determined by this equation (see Definition 3.21).}\end{array}\right\} \tag{3.110}$$

To see this, note that the graph of equation (3.108) lies in the set $I \times D_1$, since for a point (x, y) to lie on this graph we must have x in the domain of G (which is I) and must have y in the domain of H (which is D_1). On the domain $I \times D_1$, define $F(x, y) = H(y) - G(x)$, so that "$H(y) = G(x) + C$" is equivalent to "$F(x, y) = C$". We compute $\frac{\partial F}{\partial x}(x, y) = -G'(x) = -g(x)$ and $\frac{\partial F}{\partial y}(x, y) = H'(y) = h(y)$, both of which are continuous on $I \times D_1$. Moreover, $h(y_0) = \frac{1}{p(y_0)} \neq 0$. Hence the hypotheses of the Implicit Function Theorem are satisfied for the equation $F(x, y) = C$ and the point (x_0, y_0), so there is some open rectangle $I_1 \times J_1$ containing (x_0, y_0) on which the equation $F(x, y) = C$ determines y uniquely as a function of x. (Said another way: any function of x that is implicitly *semi*-defined by the equation $F(x, y) = C$ [Definition 3.23] is, truly, implicitly *defined* by the same equation.)

Now suppose that ϕ is a differentiable function of x, with domain an open interval I_1, that is semi-determined implicitly by one of the equations in $\mathcal{E}_1$. Then, for some constant C, equation (3.107) is satisfied on I_1. In particular, for all $x \in I_1$, the number $\phi(x)$ lies in the domain of H—the set D_1, on which H is differentiable, with derivative $H' = h = \frac{1}{p}$. Hence, differentiating both sides of (3.107) with respect to x, we obtain $H'(\phi(x))\phi'(x) = G'(x)$, implying that equation (3.105) holds on I_1, hence that $\phi'(x) = p(\phi(x))g(x)$ on I_1. Thus ϕ is a solution of (3.101). This establishes that

$$\left.\begin{array}{l}\text{every differentiable function of } x \text{ that is semi-determined}\\ \text{by equation (3.108) is a solution of the DE (3.101).}\end{array}\right\} \tag{3.111}$$

Together, facts (3.110) and (3.111) imply that for each $C \in \mathbf{R}$ for which the graph of equation (3.108) in R has any points, equation (3.108) is an implicit solution of (3.101) (see Definitions 3.25 and 3.25). Combining this with fact (3.109) and our observation that the collection $\mathcal{E}_2$ is the set of *constant* maximal solutions, we conclude that $\mathcal{E}$ is the general solution of (3.101) on R, in implicit form (Definition 3.34).

As noted earlier, the graphs of equations in $\mathcal{E}_1$ don't intersect the graphs of equations in $\mathcal{E}_2$. It is clear that the graphs of two equations in $\mathcal{E}_1$ can't intersect each other, and that that the graphs of two equations in $\mathcal{E}_2$ can't intersect each other. (If $C_1 \neq C_2$ and the graphs of $H(y) = G(x) + C_1$ and $H(y) = G(x) + C_2$ intersected at a point (x_0, y_0), we would have $C_1 = H(y_0) - G(x_0) = C_2$, contradicting $C_1 \neq C_2$.)

Thus, by Remark 3.35, every solution curve of (3.101) in R, whether maximal or not, lies in the graph of a unique equation in $\mathcal{E}$. This completes the proof of Theorem 3.44. ■

Remark 3.45 In the proof above, instead of using (3.106), we can use definite integration to achieve next step in the argument:

> Let $x_0 \in I_1$. Then, for any $x \in I_1$, the Fundamental Theorem of Calculus tells us that
>
> $$H(\phi(x)) - H(\phi(x_0)) = \int_{x_0}^{x} \frac{d}{dt} H(\phi(t))\, dt = \int_{x_0}^{x} g(t)\, dt = G(x) - G(x_0),$$
>
> so $H(\phi(x)) = G(x) + C$, where $C = H(\phi(x_0)) - G(x_0)$.

With all the data as in the above theorem, observe that if the function p is zero anywhere—i.e. if the set Z is not empty—then the DE (3.101) is *not* algebraically equivalent, on R, to the DE

$$\frac{1}{p(y)} \frac{dy}{dx} = g(x) \tag{3.112}$$

that arises in the process of separating variables in equation (3.101). However, these two DEs *are* equivalent on the region $I \times D_1$, and therefore have the same general solution *on this region.* Note that this region can be described simply as the region we obtain by removing from R every horizontal line that corresponds to a constant solution.

The proof of our theorem shows that the collection of equations $\{H(y) = G(x) + C\}$ is the general solution, in implicit form, of each of the DEs (3.101) and (3.112) in $I \times D_1$. Thus, assuming the conditions (3.102) are met, separation of variables always finds every non-constant solution (in implicit form), and yields no "spurious solutions" (equations that are not even implicit solutions), but *always* fails to find *any* constant solutions, (which are in one-to-one correspondence

with the zero-set Z of the function p). ■

If you re-examine the proof of Theorem 3.44 more closely, looking to see where all the assumptions in (3.102) were used, you will see that the continuity of the function p was used explicitly, but that its derivative p' does not appear anywhere. Differentiability of p (and continuity of the derivative) entered only indirectly, namely through the (first sentence of) the second bullet-point stated shortly before Theorem 3.44, a fact that we used the hypotheses (3.102) to establish. If you trace back the argument for this fact, you will see that it relied on p' being continuous *at each point* $r \in Z$; the continuity of p' elsewhere was never used. Thus we can relax the continuity assumption on p' in (3.102) somewhat without altering the conclusion of Theorem 3.44.

But suppose we weaken the conditions (3.102) more significantly by omitting all reference to p', thus requiring p to be continuous but not requiring it to be differentiable. Then the second bullet-point shortly before Theorem 3.44 no longer is valid, but the first bullet-point is, and the only parts of the proof of Theorem 3.44 that become invalid are those that relied on what that second bullet-point stated. Thus, the argument we gave to prove Theorem 3.44 actually proves the following more general theorem:

Theorem 3.46 *For a given separable DE* (3.101)*, assume that the first two hypotheses in* (3.102) *are met, and assume that the function p is continuous on some open set D. Again let $Z = \{r \in D : p(r) = 0\}$, and $D_1 = \{y \in D : p(y) \neq 0\}$. With all other notation as in Theorem 3.44, the following are true:*

1. *$\mathcal{E}_1$ is the general solution of the DE* (3.101) *in $I \times D_1$, in implicit form.*
2. *Every solution curve in the the region $I \times D_1$ is contained in the graph of a unique equation in $\mathcal{E}_1$, and does not intersect the graph of any other equation in $\mathcal{E}_1$.*
3. *$\mathcal{E}_2$ is the collection of all maximal* constant *solutions of* (3.101) *in $I \times D$.*
4. *Every solution curve of* (3.101) *in $I \times D$ is contained in a union of graphs of equations in the collection $\mathcal{E} = \mathcal{E}_1 \bigcup \mathcal{E}_2$.*

■

(The reason for "*union* of graphs of equations" in conclusion 4 is that a solution curve may lie partly in the graph of one equation in the collection, and partly in the graph of at least one other. Example 3.47 below, illustrates this phenomenon.)

What Theorem 3.46 does *not* assert, unlike Theorem 3.44, is that $\mathcal{E}_1$ is the set of *all* non-constant solutions (in implicit form) in the whole region $R = I \times D$, or that each solution curve in R is wholly contained in the graph of *one* of the equations in $\mathcal{E}$. This is the price of having weakened the hypotheses. This price can be very high, as we are about to see.

Example 3.47 Consider the DE

$$\frac{dy}{dx} = 6x(y-2)^{2/3}. \tag{3.113}$$

We wish to find the general solution. (Recall that this is the same thing as the general solution in $\mathbf{R}^2$.) Writing $g(x) = 6x$, $p(y) = (y-2)^{2/3}$, the functions g and p are continuous on the whole real line. However, p' is defined only on the set $D_1 = \{y \in \mathbf{R} : y \neq 0\} = (-\infty, 2) \bigcup (2, \infty)$. In particular, p' is not a continuous function on $\mathbf{R}$.

Observe that in $\mathbf{R} \times D_1$, equation (3.113) is algebraically equivalent to the DE

$$(y-2)^{-2/3}\frac{dy}{dx} = 6x \tag{3.114}$$

that we might write in the separation-of-variables process, but the two DEs are *not* algebraically equivalent on $\mathbf{R}^2$. Doing the relevant integrals, and solving explicitly for y in terms of x (since we can do that easily in this example), we find from Theorem 3.46 that the general solution of (3.114) on $\mathbf{R} \times D_1$, in implicit form, is

$$\mathcal{E}_1 = \{y = 2 + (x^2 + C)^3 : C \in \mathbf{R}\}.$$

The set $\mathcal{E}_2$ consists only of the one constant solution, $y = 2$.

But there are solutions whose graphs do not lie in the graph of any of the equations in $\mathcal{E}_1$ or $\mathcal{E}_2$. For example, all of the functions $\phi_1, \ldots, \phi_7$ defined below are solutions of the DE (3.114), but only for ϕ_1 does the solution-curve lie in the graph of an equation in $\mathcal{E}_1 \bigcup \mathcal{E}_2$; each of the other solution-curves lies only in a union of two or more such graphs (see Figures 4 and 5).

$$
\begin{aligned}
\phi_1(x) &= 2 + (x^2-1)^3 \\
\phi_2(x) &= \begin{cases} 2, & x \le 1, \\ 2 + (x^2-1)^3, & x \ge 1. \end{cases} \\
\phi_3(x) &= \begin{cases} 2 + (x^2-1)^3, & x \le 1, \\ 2 & x \ge 1. \end{cases} \\
\phi_4(x) &= \begin{cases} 2 + (x^2-1)^3, & x \le -1, \\ 2, & x \ge -1. \end{cases} \\
\phi_5(x) &= \begin{cases} 2 & x \le -1, \\ 2 + (x^2-1)^3, & x \ge -1. \end{cases} \\
\phi_6(x) &= \begin{cases} 2, & x \le -1, \\ 2 + (x^2-1)^3, & -1 \le x \le 1, \\ 2, & x \ge 1. \end{cases} \\
\phi_7(x) &= \begin{cases} 2 + (x^2 - (1.2)^2), & x \le -1.2, \\ 2, & -1.2 \le x \le -0.9\,, \\ 2 + (x^2 - (0.9)^2)^3, & -0.9 \le x \le 0.9, \\ 2, & 0.9 \le x \le 1.4\,, \\ 2 + (x^2 - (1.4)^2)^3, & x \ge 2. \end{cases}
\end{aligned}
$$

The solution $y = 2$ of the DE (3.113) is an example of a *singular solution*: for every point (x_0, y_0) on the corresponding solution curve, and every open interval I (no matter how small) containing x_0, the initial-value problem for this DE with initial conditions $y(x_0) = y_0$ has more than one solution on I. (In this example, $y_0 = 2$ at every point on the singular solution curve; I am defining what "singular solution" means in general.)

In all examples discussed previously, there were no singular solutions. *For separable DEs, the conditions* (3.102) *guarantee that there are no singular solutions.*

The presence of a singular solution gives rise to another phenomenon we have not seen before. The DE (3.113) has (non-maximal) solutions that can be extended to *infinitely many* maximal solutions (because solution-curves can *bifurcate* if they intersect the line $y = 2$). In all our previous examples, every non-maximal solution could be extended to a *unique* maximal solution. The singular solution in our current example fails, spectacularly, to have this property. **Every point on the graph of this singular solution curve is a disaster waiting to happen.**

Bifurcation is terrible behavior for solutions of a DE, the very opposite of the hoped-for predictability for solutions of initial-value problems, so it is worth knowing when we can rule out this behavior. For DEs in the form "$\frac{dy}{dx} = f(x, y)$", bifurcation of solutions is ruled out on any region in which the

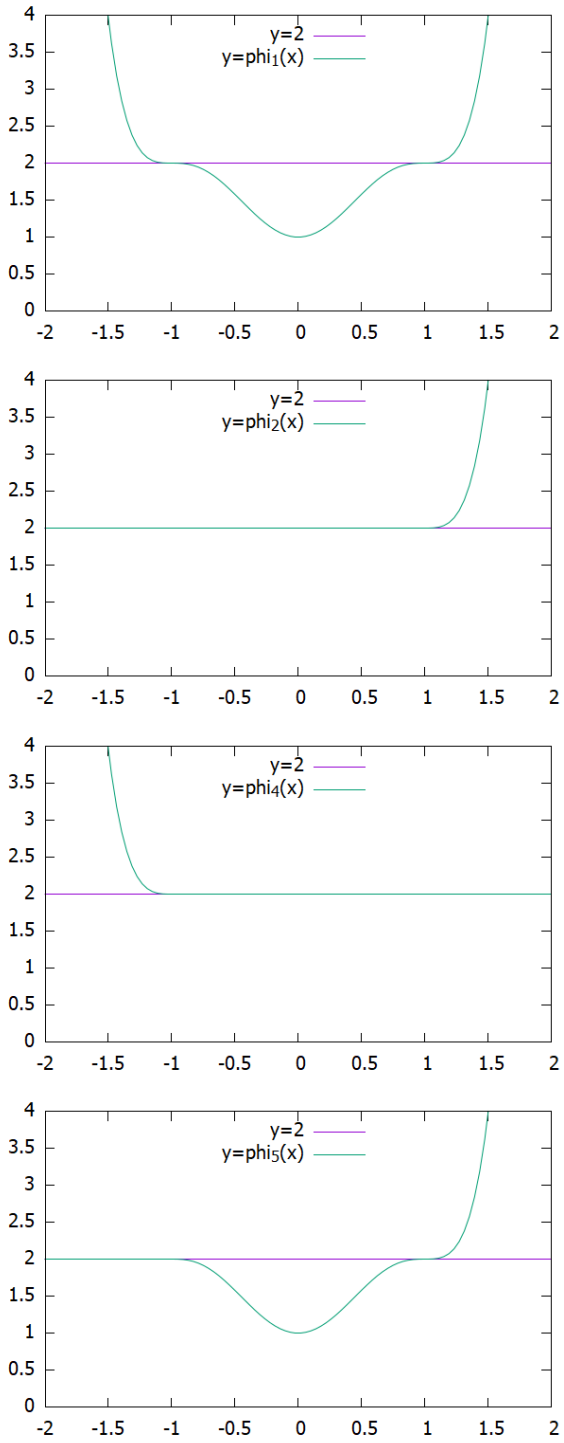

Figure 4: For Example 3.47: The solutions ϕ_1, ϕ_2, ϕ_4, and ϕ_5, plotted simultaneously with the constant solution 2. The coordinate axes (not shown) are the usual x and y axes. The graph of ϕ_1 *intersects* the graph of the constant solution, but does not *overlap* it. The graphs of ϕ_2, ϕ_4, and ϕ_5 *do* overlap with the graph of the constant solution. The graph of ϕ_3 (not shown) is the mirror image of the graph of ϕ_5.

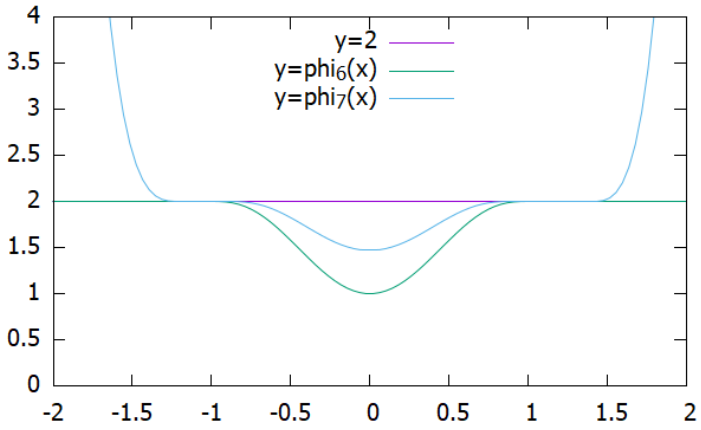

Figure 5: For Example 3.47: The solutions ϕ_6 and ϕ_7, plotted simultaneously. The two graphs overlap for $-1.2 \leq x \leq -1$ and for $1 \leq x \leq 1.4$. The solution curve $y = 2$ (not shown) overlaps the graph of ϕ_6 for $x \leq -1$ and for $x \geq 1$, and overlaps the graph of ϕ_7 for $-1.2 \leq x \leq -0.9$ and for $0.9 \leq x \leq 1.4$.

hypotheses on f in the Fundamental Theorem's hypotheses are met. This is one reason that the Fundamental Theorem is so important.[56] ■

In Example 3.47, although the collection $\mathcal{E}$ is not an implicit (or explicit) form of the general solution, it can be used to *construct* one. We simply have to write down all the (additional) solutions that are piecewise-expressed functions that, between "break-points", satisfy either $y = 2$ or one of the equations in $\mathcal{E}_1$. As the student may check, the only possibilities for the number of break-points are one (as exemplified by ϕ_2, ϕ_3, ϕ_4, and ϕ_5), two (as exemplified by ϕ_6), three (for an example, take our formula for ϕ_7, and either replace the top two lines by the single line "2, [for] $x \leq -0.9$" or replace the bottom two lines by the single line "2, [for] $x \geq 0.9$"), or four (as exemplified by ϕ_7). The bookkeeping is laborious, but it can be done. The

[56] *Note to instructors using [3]*: I find it very unfortunate that this textbook never mentions this type of bifurcation as such, since it is the most visual and basic bifurcation phenomenon in the entire study of ODEs. Golden opportunities to do this are neglected in Example 9 and Exercise 29 of Section 1.2, and in the related Project G of Chapter 2.

The only time the book *does* mention bifurcation, it is of a more subtle type. In Project B of Chapter 1 of [3], the book considers a *parametrized family* of DEs of the type $dy/dt = p(y) - s$, where s is a "perturbation parameter" and where p has a zero of order 2 at $y = 0$. As s passes through 0, there is a bifurcation in the *set of equilibria*: for s slightly positive, there are no equilibrium values near 0; for $s = 0$ there is a unique (and semistable) equilibrium value near 0, namely 0 itself; and for s slightly negative there are two equilibrium values (one stable, one unstable), near 0. Unfortunately, the diagrams in this project do not depict, at all, the bifurcation discussed in the project, and could lead students to misinterpret what "bifurcation" means, at least initially.

constructibility of a general solution (in implicit or explicit form) from a smaller collection of (explicit or implicit) solutions is a phenomenon that occurs frequently for DEs on regions in which the Fundamental Theorem does not apply directly.

Separable DEs do not always come to us in the standard form (3.101):

Example 3.48 Consider the differential equation

$$x\frac{dy}{dx} = \sin y. \tag{3.115}$$

This DE makes sense on all of $\mathbf{R}^2$, so there is no reason we should not try to solve it there. But it is not written in the form to which Theorems 3.44 and 3.46 apply. However, on the regions $R_1 = (0,\infty) \times \mathbf{R} = \{(x,y) \in \mathbf{R} : x > 0\}$ and $R_2 = (-\infty, 0) \times \mathbf{R} = \{(x,y) \in \mathbf{R} : x < 0\}$ equation (3.115) and the DE

$$\frac{dy}{dx} = \frac{\sin y}{x}\ . \tag{3.116}$$

are algebraically equivalent on the regions $R_1 = (0,\infty) \times \mathbf{R} = \{(x,y) \in \mathbf{R} : x > 0\}$ and $R_2 = (-\infty, 0) \times \mathbf{R} = \{(x,y) \in \mathbf{R} : x < 0\}$, hence have the same general solution on each of these regions. But Theorem 3.44 *does* apply to the DE (3.116), on R_1 and R_2, regions that together comprise almost the whole xy plane (everything but the y-axis). Hence we can solve (3.115) in R_1 and R_2, and then see if we can infer from our answer whether there are solutions (3.115) that are not confined to R_1 or R_2 (and if so, what these solutions are).

First consider (3.116) on R_1. Separating variables, doing the relevant integrals, and simplifying (partly by using the trig identity $\csc\theta - \cot\theta = \tan(\theta/2)$), we find that the set of non-constant solutions of equation (3.116) on R_1, in implicit form, is

$$\mathcal{E}_1 = \{\tan\frac{y}{2} = Cx : C \neq 0\},$$

and the set of constant solutions of equation (3.116) on R_1 is

$$\mathcal{E}_2 = \{y = n\pi : n \text{ is any integer}\}.$$

By Theorem 3.44, every solution of (3.116) satisfies exactly *one* of the equations in $\mathcal{E}_1 \bigcup \mathcal{E}_2$, and every solution-curve corresponding to $\mathcal{E}_1$ is trapped between the graphs of two consecutive constant solutions. In such a sub-region of R_1, we can solve "$\tan\frac{y}{2} = Cx$" for y in terms of x and find that

$$y = 2\tan^{-1}(Cx) + 2m\pi$$

for some integer m;[57] the corresponding solution-curve in R_1 lies between the graphs of $y = 2m\pi$ and $(2m+1)\pi$ if $C > 0$, and between the graphs of $y = 2m\pi$ and $(2m-1)\pi$ if $C < 0$. (Here "$\tan^{-1}$" denotes the inverse-tangent function, also known as "arctan"; it does *not* denote the reciprocal of the tangent function, i.e. the cotangent function. Recall that the range of $\tan^{-1}$ is the interval $(-\pi/2, \pi/2)$.) Thus, the set of non-constant solutions of equation (3.116) on R_1, in explicit form, is

$$\mathcal{E}_1' = \{y = 2\tan^{-1}(Cx) + 2m\pi : m \text{ is an integer and } C \neq 0\}.$$

Since the DEs (3.115) and (3.116) are equivalent on R_1, they have the same general solution in this region. Hence the general solution of the DE (3.115) on R_1 is $\mathcal{E} = \mathcal{E}_1' \bigcup \mathcal{E}_2$. As with many such expressions of general solutions of DEs, we can look to see whether any restrictions on any constants that distinguish one equation in $\mathcal{E}$ from another are necessary to ensure that every equation represents a *solution* (or implicit solution), or whether these restrictions are simply artifacts of the method we used to find *some* way to express the general solution. If we can remove these restrictions, we may be able to write the general solution more simply. In the current example, observe that if we set $C = 0$ in "$y = 2\tan^{-1}(Cx) + 2m\pi$", we get the constant solution $y = 2m\pi$ (recall that $\tan^{-1}(0) = 0$). Thus we can recover the constant solutions currently labeled by *even* integers n in $\mathcal{E}_2$ this way, but not those labeled by odd n. The resulting, somewhat simpler, way of expressing the general solution of equation (3.115) on R_1 is

$$\begin{aligned} \mathcal{E} \;=\; & \{y = 2\tan^{-1}(Cx) + 2m\pi : m \text{ is an integer and } C \in \mathbf{R}\} \\ & \quad \text{and} \\ & \{y = (2m+1)\pi : m \text{ is an integer}\}. \end{aligned} \tag{3.117}$$

Similar analysis on R_2 reveals that the general solution of (3.115) on R_2 can be written as exactly the same set of equations $\mathcal{E}$. That does not mean that the *solutions* in R_1 are the same as the *solutions* in $\mathbf{R}^2$; the domain of each maximal solution in R_1 is the interval $(0, \infty)$, while the domain of each maximal solution in R_2 is $(-\infty, 0)$.

The solutions found above are maximal solutions of (3.115) in R_1 and in R_2, but what we are looking for are solutions of (3.115) that are maximal *in the whole plane* $\mathbf{R}^2$.

To find all of these, first observe that for each $C \in \mathbf{R}$ and integer m, the function by $\phi(x) = 2\tan^{-1}(Cx) + 2m\pi$ is differentiable on the whole real line and we already know that it satisfies equation (3.115) on $(-\infty, 0)$ and on $(0, \infty)$. At $x = 0$ we have $x\phi'(x) = 0 \times \phi'(0) = 0$, and $\sin(\phi(x)) = \sin(\phi(0)) = \sin(2m\pi) = 0$. Hence ϕ is a solution of (3.115) on the whole real line (and is therefore a maximal solution).

[57] Remember that "$\tan u = v$" does not imply that $u = \tan^{-1} v$! Because the tangent function has period π, "$\tan u = v$" implies only that $u = \tan^{-1} v + m\pi$ for some integer m.

Similarly, for any integer n, the constant function $\phi(x) = n\pi$ is a maximal solution on the whole real line.

Are there any other maximal solutions whose domains include 0? To answer this, suppose that we have such a solution ϕ on an open interval I containing 0. Let I_+ and I_- be the portions of I to the right and left of 0, respectively, and let ϕ_+ and ϕ_- be the restrictions of ϕ to these intervals. Then ϕ_+ must be one of our maximal solutions in $(0,\infty)\times\mathbf{R}$, so $I_+ = (0,\infty)$ and for some C_1 and m_1 the function ϕ_+ is given by either

$$\begin{gathered}\phi_+(x) = 2\tan^{-1}(C_1 x) + 2m_1\pi \\ \text{or} \\ \phi_+(x) = (2m_1+1)\pi.\end{gathered} \tag{3.118}$$

Similarly, $I_- = (-\infty, 0)$, and for some C_2 and m_2 the function ϕ_- is given by formulas to those for ϕ_+. Since ϕ is a solution of a DE, ϕ is continuous. Therefore $\phi(0) = \lim_{x\to 0}\phi(x) = \lim_{x\to 0+}\phi_+(x)$, which has the value $2m_1\pi$ or $(2m_1+1)\pi$ accordingly as ϕ_+ is given by the top or bottom line of (3.117). Similarly, we also have $\phi(0)$ equal to either $2m_2\pi$ or $(2m_2+1)\pi$. Hence $m_1 = m_2$ (and we may call both of these simply m), and either both ϕ_+ and ϕ_- are of the form on the top line of (3.117), or both are of the form on the bottom line. In the latter case, we have $\phi(x) = (2m+1)\pi$ for all x, a constant function on $(-\infty,\infty)$. In the former case, ϕ_+ and ϕ_- extend to differentiable functions on the whole real line, and we have with $\phi'_+(0) = 2C_1$. Thus $\phi'(0) = \lim_{x\to 0}\frac{\phi(x)-\phi(0)}{x-0} = \lim_{x\to 0+}\frac{\phi_+(x)-\phi_+(0)}{x} = \phi'_+(0) = 2C_1$. Similarly $\phi'(0) = 2C_2$. Hence $C_1 = C_2$. Letting C denote both of these numbers, we then have $\phi(x) = 2\tan^{-1}(Cx) + m\pi$ on $(-\infty,\infty)$.

Thus there are no maximal solutions other than the ones we found earlier, the ones given by the equations in (3.117). Therefore the general solution of equation (3.115) is (3.117), with x now running over $(-\infty,\infty)$.

Note that every solution $y(x)$ of the DE (3.115) on an interval containing 0 has $\sin(y(0)) = 0\times y'(0) = 0$, implying that $y(0) = n\pi$ for some integer n. Thus for every y_0 that is not a multiple of π, the initial-value problem $x\frac{dy}{dx} = \sin y,\ \ y(0) = y_0$, has *no* solution on *any* interval containing 0. At the same time, for every *even* integer n, the IVP $x\frac{dy}{dx} = \sin y,\ \ y(0) = n\pi$ has infinitely many solutions, while for every *odd* integer n this IVP has a *unique* solution, the constant function $y = n\pi$. These facts are illustrated in Figure 6, where many solutions are plotted. The solution curves of $x\frac{dy}{dx} = \sin y$ "fill out" the region $\{(x,y)\in\mathbf{R}^2 : x\neq 0\}$, but not the whole plane $\mathbf{R}^2$. Every point (x_0,y_0) with $x_0\neq 0$ lies on exactly one maximal solution curve, as does every point $(0,n\pi)$ with n an *odd* integer. Every point $(0,n\pi)$ with n an *even* integer lies on infinitely many maximal solutions curves, and every point $(0,y_0)$ with y_0 not an integer multiple of π lies on *no* solution curve. ■

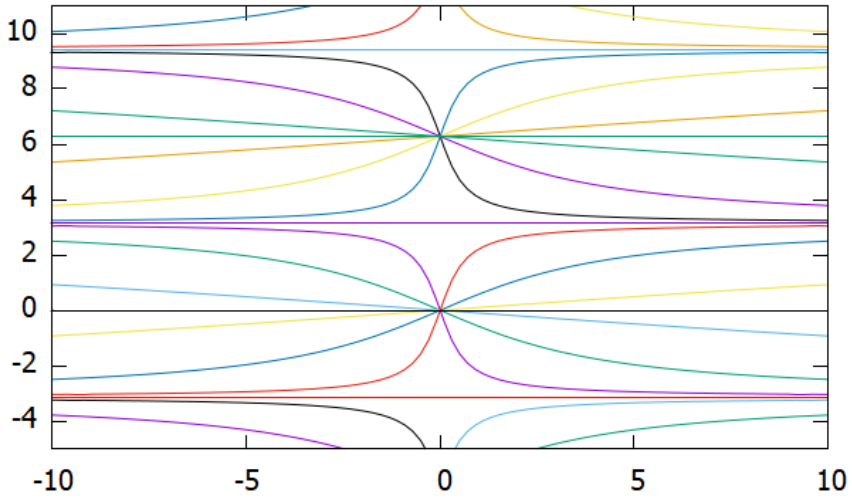

Figure 6: For Example 3.48: Several solutions of $x\frac{dy}{dx} = \sin y$, plotted in the rectangle $-10 \leq x \leq 10,\ -5 \leq y \leq 11$. (The displayed solution curves are maximal in this region.) The horizontal lines in the figure are the constant solutions $y = -\pi,\ y = 0, y = \pi,\ y = 2\pi$, and $y = 3\pi$.

3.3 First-order equations in differential form

3.3.1 Differentials and differential-form DEs

Definition 3.49 A *differential* in the variables (x, y) is an expression of the form

$$M(x, y)dx + N(x, y)dy \tag{3.119}$$

where M and N are functions defined on some region in $\mathbf{R}^2$. We often abbreviate (3.119) as just

$$Mdx + Ndy, \tag{3.120}$$

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 (3.119) and (3.120) are called the *coefficients* of dx and dy in these expressions. ■

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

Definition 3.50 (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

$$dF = \frac{\partial F}{\partial x}\,dx + \frac{\partial F}{\partial y}\,dy. \tag{3.121}$$

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

Remember that for a function F of two variables, "continuously differentiable (on R)" implies that the function F itself is continuous (on R). Thus 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, so far they are just place-holders for the functions M and N in (3.119) and (3.120). Similarly, so far the expression "$M dx + N dy$" 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 4.1 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 $M dx + N dy$ 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 $M dx + N dy$ 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.[58])

We next define rules for algebraic operations involving differentials. These def-

[58] 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.

initions 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 3.51 as "obvious facts".**[59] If you have trouble understanding why Definition 3.51 is necessary, don't worry about it; just make sure that the way you manipulate differentials agrees with these rules.

Definition 3.51 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 $M dx + N dy$, $M_1 dx + N_1 dy$, and $M_2 dx + N_2 dy$ are differentials on R.) Then we make the following definitions for differentials in (x, y).

1. Equality of differentials: $M_1 dx + N_1 dy = M_2 dx + N_2 dy$ 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:

$$\begin{aligned} M dx &= M dx + 0 dy, \\ N dy &= 0 dx + N dy. \end{aligned}$$

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

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

4. Insensitivity to which term is written first:

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

5. Addition of differentials:

$$(M_1 dx + N_1 dy) + (M_2 dx + N_2 dy) = (M_1 + M_2) dx + (N_1 + N_2) dy.$$

6. Subtraction of differentials:

$$(M_1 dx + N_1 dy) - (M_2 dx + N_2 dy) = (M_1 - M_2) dx + (N_1 - N_2) dy.$$

[59] This tends to be what DE textbooks do: the algebraic rules in Definition 3.51 are *used* without ever stating them.

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[60]. ■

Note 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 $\frac{dy}{dx}$.

Note also 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":

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

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

Definition 3.52 A *differential equation in differential form* (or *differential-form DE*), *with variables* (x, y), is a (non-definitional) equation of the form

$$\text{one differential in } (x, y) = \text{another differential in } (x, y). \tag{3.122}$$

We (should) 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

[60] 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.

"a DE on R in differential form" or "a DE in differential form on R." ■

Above, "non-definitional" means that the equation is not simply a definition of expressions on one side or the other. The equation "$dF = \frac{\partial F}{\partial x}\,dx + \frac{\partial F}{\partial y}$" is an example of a *definitional* equation.

Example 3.53 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

$$y\,dy = e^x\,dx. \tag{3.123}$$

■

Note that when we write equation (3.123), or any other differential-form DE, we are not asserting that the left-hand side and right-hand side are equal differentials. Like other equations, *a differential-form DE makes a statement that will be true when certain things of the appropriate type are plugged in, and false when other things of that type are plugged in.* We will reveal in Section 3.3.3 what are the "things of appropriate type" to plug in; we must lay some groundwork first.

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.[61] 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.*

[61] Some textbook authors implicitly encourage students to this mistake, either through answers in the back of the book (or official solutions manuals) or through example. This is very unfortunate.

Here is the differential-form analog of Definition 3.36:

Definition 3.54 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 DEs

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

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

and

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

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

$$x(2x^2y\,dx - \tan(x+y)\,dy) = 0, \tag{3.124}$$

but are *not* algebraically equivalent to (3.124) 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 (3.122) from both sides of the equation, we obtain an algebraically equivalent equation of the form

$$M\,dx + N\,dy = 0. \tag{3.125}$$

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 (3.125). (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 (3.123) 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. Some of these basics will look familiar to you from Calculus 2 or 3, but not all of them.

3.3.2 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 3.55 A *parametrized curve* or *curve-parametrization* in $\mathbf{R}^2$ is an ordered pair of continuous real-valued functions (f, g) defined on a positive-length interval[62]. The set

$$\{(f(t), g(t)) : t \in I\} \tag{3.126}$$

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

A *curve* in $\mathbf{R}^2$ is a set $\mathcal{C} \subseteq \mathbf{R}^2$ that is the image of some parametrized curve and has more than one point. [63] [64] (A concrete example is coming up shortly, Example 3.56.)

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 which 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: thinking of (f, g) as a single $\mathbf{R}^2$-valued function γ (defined by $\gamma(t) = (f(t), g(t))$), the range of γ is the set of points (x_0, y_0) in $\mathbf{R}^2$ for which $\gamma(t_0) = (x_0, y_0)$. (In case the

[62] Recall that a *positive-length* interval is any interval that contains more than a single point; i.e. any interval other than one of the form $[a, a]$.

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

[64] If the functions f, g in (3.126) are *constant* functions, then $\mathcal{C}$ will contain just one point. In advanced mathematics, in the definition of "curve" we usually omit the requirement that $\mathcal{C}$ have more than one point. Although it is counterintuitive to think of a *point* as an example of a *curve*, for some purposes (in advanced mathematics) it is essential to allow this.

symbol "γ" is new to you: it's the Greek letter gamma, in lower case.) 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 extra letters f, g and write "$(f(t), g(t))$" for the point in the xy plane defined by "$x = f(t)$, $y = g(t)$". We will often use the simpler notation $(x(t), y(t))$ when there is no danger of misinterpretation. Thus we 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 write "the parametrized curve (or curve-parametrization) $t \mapsto (x(t), y(t))$." (Read the symbol "$\mapsto$" as "goes to". The little vertical bar at start of the arow is *essential* for this arrow to have this meaning. The "$\mapsto$" arrow is a very special arrow.)

Note that in Definition 3.55, 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.

Example 3.56 Let $x(t) = 2\cos t,\ y(t) = 2\sin t,\ \ 0 \le t \le 2\pi$. Then for all t we have $x(t)^2 + y(t)^2 = 4$, so the image of this parametrized curve lies on 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 parametrif *curve* traced out by the parametrized curve $t \mapsto (x(t), y(t)),\quad t \in [0, 2\pi]$, 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 3.57 A curve-parametrization $(x(t), y(t)), t \in I$ is called

- *differentiable* if the derivatives $x'(t)$, $y'(t)$ exist[65] for all $t \in I$;

[65] When I contains an endpoint (i.e. I is closed or half-closed), *derivative* at a contained endpoint is interpreted as the appropriate *one-sided* derivative. See Section 5.1.1. 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 $\lim_{t \to b^-} \frac{x(t) - x(b)}{t - b}$.

- *continuously differentiable* if it is differentiable and $x'(t)$, $y'(t)$ are continuous in t;
- *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 $x'(t_0) = 0 = y'(t_0)$);[66] and
- *regular* if it is continuously differentiable and non-stop.

■

Definition 3.58 A curve $\mathcal{C}$ in $\mathbf{R}^2$ is *smooth* if for every point (x_0, y_0) on the curve, there is an open rectangle R containing (x_0, y_0) such that the portion of $\mathcal{C}$ lying inside R admits a regular parametrization, with domain an open interval.[67] ■

"Admits", as used in Definition 3.58, means that the indicated portion of $\mathcal{C}$ is the image of *some* parametrization with the indicated properties.

The open-interval requirement at the end of Definition 3.58 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 3.56 to convince him/herself that a circle meets our definition of "smooth curve".

Observe that Definition 3.58 uses a "windowing" idea similar to the one that we used to talk about implicitly-defined functions in Section 3.2.5. 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

[66] "Differentiable" and "continuously differentiable" are completely standard terminology; "non-stop" is not. See the "Warning about terminology" coming up soon.

[67] *Note to instructors*: Shortly, I will be defining *solution curves* for differential-form DEs, and will require such curves to be smooth. "Smooth curve", as we have defined it here, is synonymous with "connected, 1-dimensional, smooth (C^1) submanifold of $\mathbf{R}^2$". Every such curve is the image of an embedding of either $\mathbf{R}$ or S^1 in the plane. For some purposes, it might be better to allow "smooth immersed curves", i.e. images of *immersions*, not just embeddings. These are *all* the curves admitting regular parametrizations. A smooth immersed curve can (among other things) cross itself, limit to one of its interior points, or "wrap around" infinitely close to itself (like a line of irrational slope in the standard torus). But for a course at this level, I thought it best to stick to a definition of "smooth curve" that eliminates at least the first two of these possibilities. Eliminating these while not eliminating the third would just lead to a distraction, and could cause confusion.

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 paper" 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 regular parametrization, but the point (x_0, y_0) at which the curve crosses itself causes the definition of "smooth" not to be met. For any open rectangle R containing (x_0, y_0), however small, the portion of the curve inside R is essentially an X, and has the same problem that the whole X did.

Warning about curve terminology. Many calculus textbooks refer to a regular parametrization as a *smooth* parametrization. This usage of "smooth" is unfortunate (and has been since the 1950s or earlier, though it has historical roots); it conflicts with the modern meaning of "smooth function" in advanced mathematics.[68]

We make one more definition before moving on to the next section. **The paragraph after the definition explains the terminology less formally; skip ahead to this paragraph if you have trouble understanding the formal definition.**

[68] *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 prespecified 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 are indispensable to the definition of "tangent space". "Regular" is a flexible term that mathematicians use with a contextually varying meaning, which usually is "having the most common features" or "having no important nasty or inconvenient features" (where the context determines what features are important).

Definition 3.59 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 regular 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 regular parametrization with domain an open interval), and there is no regular parametrized curve whose image lies in R and contains $\mathcal{C}$ as a proper subset.[69] ■

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", *inextendible* essentially means that we cannot add points at either end of the curve without leaving the region R. For example, if R is the region that lies strictly between the horizontal lines $y = 1$ and $y = -1$, the portion of the graph of $y = x$ that lies in R is inextendible in R. The portion of the same graph that lies in the open first quadrant R_1 is inextendible in R_1.

3.3.3 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 3.60 A *solution curve* [70] of a differential equation

$$M(x, y)\,dx + N(x, y)\,dy = 0 \tag{3.127}$$

on a region R is a smooth curve $\mathcal{C}$, contained in R, admitting a regular parametrization $t \mapsto \gamma(t) = (x(t), y(t))$ that satisfies

$$M(x(t), y(t))\frac{dx}{dt} + N(x(t), y(t))\frac{dy}{dt} = 0 \tag{3.128}$$

[69] 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.

[70] 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 in a textbook. That meaning, even if not stated explicitly, is likely to be close to Definition 3.66 later in these notes.

for all t in the domain-interval I of the parametrization. In this context, we will call γ a *parametric solution* of (3.127) (in R).[71]

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[72]. ■

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 3.3.5.

As we noted previously, in a differential-form DE (3.127) there is neither an independent nor a dependent variable; x and y are treated symmetrically. This symmetry is preserved in (3.128), but in a surprising way: in (3.128), *both* x and y are dependent variables! The independent variable is t—a variable that is not even visible in (3.127). (Of course, in place of t we could have used any other letter not appearing elsewhere in Definition 3.60.)

Definition 3.60 implies more about solution curves and parametric solutions than is obvious just from reading the definition.

To start with, equation (3.128) has a geometric interpretation.[73] Let $t \mapsto (x(t), y(t))$ be a regular parametrization of some solution curve $\mathcal{C}$ of $M\,dx + N\,dy = 0$. Let $\mathbf{v}(t) = x'(t)\mathbf{i} + y'(t)\mathbf{j}$, where $\mathbf{i}$ and $\mathbf{j}$ are the standard basis vectors in the xy plane. Then $\mathbf{v}(t)$, the velocity-vector of the parametrization at "time" t, is tangent to the smooth curve $\mathcal{C}$ at the point $(x(t), y(t))$. We can rewrite equation (3.128) using the dot-product you learned in Calculus 3:

$$[M(x(t), y(t))\,\mathbf{i} + N(x(t), y(t))\,\mathbf{j}\,] \cdot \mathbf{v}(t) = 0. \tag{3.129}$$

This says that, for each t, the vector $\mathbf{v}(t)$ is perpendicular to the vector $M(x(t), y(t))\mathbf{i} + N(x(t), y(t))\mathbf{j}$.

Suppose we have a second non-stop parametrization of the same curve $\mathcal{C}$. To speak clearly of both parametrizations, we temporarily abandon the notation "$(x(t), y(t))$" in favor of $\gamma_1(t)$ (with t-domain I_1) for the first parametrization, and

[71] The terminology "parametric solution" for a DE in differential form was **invented for these notes**; it is **not** standard.

[72] *Note to instructor:* I have avoided giving a careful definition of "boundary" here, and therefore of "interior", to avoid distracting the student.

[73] If you have not yet taken Calculus 3, either (i) look up "standard basis vectors" and "dot product" before proceeding, or (ii) skip to statement (3.130) below, taking it on faith that statement (3.130) was established in the part you skipped over.

$\gamma_2(t)$ (with t-domain I_2) for the second. At a given point $(x_0, y_0) = \gamma_1(t_1) = \gamma_2(t_2)$ of $\mathcal{C}$, the velocity vectors $\mathbf{v}_1(t_1), \mathbf{v}_2(t_2)$ coming from the two parametrizations are parallel, both being nonzero vectors tangent to $\mathcal{C}$ at that point. Thus $\mathbf{v}_2(t_2) = c\mathbf{v}_1(t_1)$ for some nonzero scalar c. But then

$$\begin{aligned}
(M(x_0, y_0)\mathbf{i} + N(x_0, y_0)\mathbf{j}) \cdot \mathbf{v}_2(t_2) &= (M(x_0, y_0)\mathbf{i} + N(x_0, y_0)\mathbf{j}) \cdot c\mathbf{v}_1(t_1) \\
&= c\,(M(x_0, y_0)\mathbf{i} + N(x_0, y_0)\mathbf{j}) \cdot \mathbf{v}_1(t_1) \\
&= c \times 0 \quad (\text{“times”, not cross-product}) \\
&= 0.
\end{aligned}$$

Since this holds for all points (x_0, y_0) on $\mathcal{C}$, it follows that the parametrization $t \mapsto (x(t), y(t)) = \gamma_2(t)$ also satisfies (3.128).[74] Thus if one regular parametrization of $\mathcal{C}$ satisfies (3.128), so does every other.

Therefore, even though Definition 3.60 requires only that *some* regular parametrization of $\mathcal{C}$ satisfy (3.128), once we know that even *one* regular parametrization of $\mathcal{C}$ satisfies (3.128), we know that they *all* do. Said another way:

$$\left.\begin{array}{l} \textit{Every} \text{ regular parametrization of a solution curve} \\ \text{of a differential equation } M\,dx + N\,dy = 0 \\ \text{is a parametric solution of this equation.} \end{array}\right\} \tag{3.130}$$

This gets back to what we said just before to Definition 3.60: that a DE in differential form is looking for a *curve*. We carefully 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. But it is not the *same thing* as “curve” in the geometric sense.

The blue paragraph below is optional reading.

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* regular parametrizations of $\mathcal{C}$. These are the properties that are truly *geometric*. What statement (3.130) is saying is that the property “I am a solution curve of this differential-form DE” is an intrinsic, geometric property.

[74] This can also be shown without appealing to geometry, using the Chain Rule plus the Inverse Function Theorem that you may have learned in Calculus 1. (This theorem is stated in footnote 86.)

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*—they are still related to each other. We will see the precise relation in Section 3.4. For now, we mention just that a solution curve of any derivative-form DE is a solution curve for a related differential-form DE. 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 (possibly overlapping) 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, expressing the curves this way is often neither necessary nor desirable[75]. This is another fundamental difference between derivative-form DEs and differential-form DEs.

Example 3.61 Consider the equation

$$x\,dx + y\,dy = 0. \tag{3.131}$$

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 that passes through this point. The solution curve that 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 (3.131) 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. ■

[75]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 just as a tool to solve derivative-form equations, say with independent variable x and dependent variable y, then it usually *is* desirable to write solutions in the explicit form "$y =$ differentiable function of x"—and your instructor may *require* you to do this whenever it is algebraically possible.

3.3.4 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); the analog of "*solution* of an initial value problem" is a solution curve that passes through this point. 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. For derivative-form DEs, the terminology stems from there being a definite independent variable that, for many DEs in the sciences, is *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 (3.129). Suppose (x_0, y_0) lies on a smooth solution curve $\mathcal{C}$ of $M\,dx + N\,dy = 0$. 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 (3.129) 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 (3.129) is satisfied at $t = t_0$, and so is (3.128). 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 $M\,dx + N\,dy = 0$ 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 3.62 A point (x_0, y_0) is a *singular point* of the differential $M\,dx + N\,dy$ if $M(x_0, y_0) = 0 = N(x_0, y_0)$.[76] ■

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

[76] "Singular point" here does not mean the same thing as in footnote 52.

$$\frac{dy}{dx} = f(x, y). \tag{3.132}$$

If the graph of a solution of (3.132) passes through (x_0, y_0), then the slope of the graph at this point must be $f(x_0, y_0)$. This is true even if the IVP

$$\frac{dy}{dx} = f(x, y), \quad y(x_0) = y_0 \tag{3.133}$$

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 $M\,dx+N\,dy$ is a worse problem for the differential-form IVP "$M\,dx + N\,dy = 0$ with initial condition (x_0, y_0)" than we ever see for the derivative-form IVP (3.133).[77] 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 $M\,dx + N\,dy = 0$ on a region in which $M\,dx + N\,dy$ has a singular point. But in regions free of singular points, there are no difficulties. We make the following definition:

Definition 3.63 Let R be a region in which the differential $M\,dx + N\,dy$ has no singular points. Suppose $\mathcal{C}$ is a curve lying in R and is a solution curve of the equation $M\,dx + N\,dy = 0$. We say that $\mathcal{C}$ is *maximal in* R if $\mathcal{C}$ is inextendible in R (see Definition 3.59).

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 we will not discuss here (but one of which is a phenomenon exhibited later in Example 3.76).[78]

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

Theorem 3.64 *Suppose M and N are continuously differentiable functions on an open region R in $\mathbf{R}^2$, and that $M\,dx + N\,dy$ has no singular points in R. Then for*

[77] However, for derivative-form DEs that are *not* in the standard form "$\frac{dy}{dx} = f(x, y)$", there can be points (x_0, y_0) for which the the initial condition $y(x_0) = y_0$ also imposes no restriction on $\frac{dy}{dx}$ at that point, and for which the corresponding IVP has infinitely many solutions, each of whose graphs has a different slope at (x_0, y_0). One example is the IVP $x\frac{dy}{dx} = \sin y, \quad y(0) = 0$. See Example 3.48 and Figure 6.

[78] *Note to instructors*: The main problem is that some solution curves would not lie in any maximal solution curve. While this is not truly a problem with *defining* maximal solution curve, it does make the notion less useful in regions with singular points.

every point $(x_0, y_0) \in R$, there exists a unique solution curve of $M\,dx + N\,dy = 0$ 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 $M\,dx + N\,dy = 0$ that have a unique inextendible solution curve through a singular point of the differential. But there are also differentials $M\,dx + N\,dy$ for which (i) M and N are continuously differentiable in the whole xy plane, (ii) $M\,dx+N\,dy$ has a singular point (x_0, y_0), and (iii) the equation $M\,dx + N\,dy = 0$ 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 3.65 Suppose $M\,dx+N\,dy$ is exact on a region R, and let F be a function on R for which $M\,dx + N\,dy = dF$. Then $M = \frac{\partial F}{\partial x}$ and $N = \frac{\partial F}{\partial y}$. Hence, for a given point $(x_0, y_0) \in R$,

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

Thus, *the singular points of dF are precisely the critical points of F.*

3.3.5 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" tend 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' solutions [i.e. *true* solutions] implicit solutions (of derivative-form DEs)," and then effectively taking the definition of "solution of a DE in differential form" to be "implicit solution of a related derivative-form DE".[79] Here we wish to maintain the conceptual

[79] *Note to instructors:* The only reason there is a good relation at all between differential-form ODEs and (certain) derivative-form ODEs is, literally, the fact that $1 + 1 = 2$. 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). The solutions of $x\,dx + y\,dy + z\,dz = 0$ in $\mathbf{R}^3$ are *spheres*.

difference between solutions of derivative-form and differential-form DEs. With this in mind, we introduce different terminology for an equation that might be an implicit solution of a derivative-form DE, but that we wish to regard as some sort of solution of a related differential-form DE.

Definition 3.66 We will say that an equation

$$F(x, y) = 0 \quad (\text{or } F(x, y) = \text{any fixed real number}) \tag{3.134}$$

is an *algebraic solution*, or *non-parametric solution*, of a differential-form DE

$$M(x, y)\, dx + N(x, y)\, dy = 0 \tag{3.135}$$

on a region R if

(i) the graph of (3.134) contains a smooth curve in R , and

(ii) every smooth curve in R contained in the graph of (3.134) is a solution curve of (3.135).[80]

If $R = \mathbf{R}^2$ then we usually omit mention of the region, and say just that (3.134) is an *algebraic solution*, or *non-parametric solution*, of (3.135). ■

Note that we have not yet defined the term "*solution* of a DE in differential form". The reason for the delay is discussed in Remark 3.67 below, after which we give the missing definition.

Remark 3.67 Since a DE in differential form is looking for a curve, 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 (just for the remainder of this Remark), 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 (3.134) is in relation to (3.135). An equation of the form $F(x, y) = 0$ is a very *explicit* description of a set $\mathcal{C}$: a point (x_0, y_0) is in $\mathcal{C}$ if and only if $F(x_0, y_0) = 0$. In "nice" situations (which we will be more specific about later) $\mathcal{C}$ will be a curve, or at least a finite or countably infinite collection of curves.

[80] *Note to instructors*: 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.

Because a curve is a point-set in the plane, an equation of the form $F(x, y) = 0$ is 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 $F(x, y) = 0$. In this context, "implicitly defined *function*" is a perfectly sensible concept and term; "implicitly defined *curve*" is not.

However, terminology *is* needed to distinguish a *parametric description* of a curve (as the range of some given function $t \mapsto (x(t), y(t))$ on some interval) from a non-parametric description (as the solution-set of some given algebraic equation in x and y), and when writing equations for curves, many people use the word "implicit" simply to mean "non-parametric". This is a practice I would like to discourage. The only thing that is really "implicit" about "*implicit solution* of a differential-form equation" as defined above, is that the equation (3.134) *itself* is not a solution of equation (3.135)—rather, the solutions of equation (3.135) related to (3.134) are smooth curves contained in the *graph* of equation (3.134). (The solutions of equation (3.135) won't *all* be related to (3.134), even if $dF = M\,dx + N\,dy$, because we chose a specific value for the constant on the right-hand side in (3.134); Definition (3.66) started with the words "*An* equation".) *For these reasons, Definition 3.66 does not contain the word "implicit".* The modifiers "algebraic" and "non-parametric" provide a distinction between *equation(s)* for a curve and the curve itself—the latter being the right type of animal to call a solution to a differential-form DE—without calling an explicit *equation* something it is not, namely "implicit". ■

Early in these notes, after defining *solution* of a derivative-form DE, we stated in Remark 3.2 that, in the interests of brevity, we would allow ourselves one common abuse of terminology: we would allow ourselves to say, e.g., that the equation "$y = x^2$ is a solution of $\frac{dy}{dx} = 2x$" it *can't* be, literally (because an *equation*—in this case $y = x^2$—isn't a *function*). In a similar spirit, we are going to allow ourselves the following abuse of terminology for solutions of differential-form equations.

Definition 3.68 In the setting of Definition 3.66, we also call an algebraic solution of equation (3.135) (on R) simply a *solution* of equation (3.135) (on R).

In other words, we are allowing ourselves to drop the modifier "algebraic" (or "non-parametric").

Example 3.69 The equation

$$xy = 1$$

is a solution of

$$y\,dx + x\,dy = 0. \tag{3.136}$$

The graph, a hyperbola, consists of two inextendible solution curves, one lying in the first quadrant and the other lying in the third. One of the these solution curves

admits the regular parametrization $x(t) = t,\ y(t) = \frac{1}{t}$ on the t-interval $(0, \infty)$, while the other admits the regular parametrization $x(t) = t,\ y(t) = \frac{1}{t}$ on the t-interval $(-\infty, 0)$.

More generally, for every real number C, the equation

$$xy = C$$

is a solution of the same differential-form equation (3.136). For most C, the graph is a hyperbola, but the case $C = 0$ is exceptional. The graph of

$$xy = 0 \tag{3.137}$$

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 (where "disjoint" means that the two curves have no points in common). We can verify that (3.137) is indeed a solution of (3.136) by observing that the parametrized curves given by $x(t) = t, y(t) = 0,\ t \in \mathbf{R}$ (a regular parametrization of the x-axis) and $x(t) = 0, y(t) = t,\ t \in \mathbf{R}$ (a regular 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 (3.136)—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 this 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 a singular point of $y\,dx + x\,dy$. (See Definition 3.62. The only singular point in the present example is $(0, 0)$.) ■

Remark 3.70 (Horizontal and vertical solution curves) A *derivative*-form DE can potentially have some solution curves that are *horizontal* lines (assuming that we plot the independent variable horizontally, and the dependent variable vertically), but can never have solution curves that are *vertical lines*. A vertical line isn't even a *candidate* for a solution curve of a derivative-form DE; it's not the graph of *any* function of the independent variable. The horizontal-line solution curves are exactly the graphs of *constant* solutions (see Remark 3.3). By contrast, a differential-form DE

$$M\,dx + N\,dy = 0 \tag{3.138}$$

can have solution curves that are vertical lines (or horizontal lines; a given differential-form DE may have both, neither, or one but not the other). The DE (3.136) has both a horizontal solution-curve (the x-axis) and a vertical solution-curve (the y-axis).

Suppose that the functions M, N in equation (3.138) are defined on a region R that may or may not be the whole y plane. To simplify the wording in the rest of this remark, we allow the terms "horizontal line" and "vertical line" to mean not just whole lines, but segments of these lines that are contained in R.

For any parametrization $t \mapsto (x(t), y(t))$ of a vertical line, the function $x(t)$ is constant. Hence for any regular parametrization of a vertical line $x = C$ in R, we have $\frac{dx}{dt} \equiv 0$, implying that $\frac{dy}{dt}$ is nowhere 0, and reducing equation (3.128) simply to $N(C, y(t)) \equiv 0$. Thus if x_0 is a number such that $N(x_0, y) \equiv 0$ (for all y such that (x_0, y) lies in R), the any vertical line in R with equation $x = x_0$ is a solution curve of equation (3.138). Conversely, these are the *only* vertical lines that are solution curves of (3.138) in R (and there may be none). Similarly, the horizontal lines in R that are solution curves are exactly the lines $y = y_0$ for which $M(x, y_0) \equiv 0$.

Because $\frac{dx}{dt} = 0$ for any parametrization of a vertical line, we say that the differential dx *evaluates to zero* on vertical lines. More loosely, we may say that dx "is" zero on any vertical line, consistent with the Calculus-1 notion that dx represents "infinitesimal change in x", and the fact that x is not changing at all on a vertical line. Similarly, we say that dy evaluates to zero on horizontal curves, and allow ourselves to say more loosely that dy "is" zero on horizontal lines. ■

Remark 3.71 You may wonder to what extent criterion (i) in Definition 3.66 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 isolated 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 a sequence of 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 curves that are 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 3.66. ■

Defining "general solution" for equations in differential form is trickier than it is for derivative form. One reason is that, logically, what we are calling a *solution curve* of a DE in differential form is what we really should be calling just a *solution* (see Remark 3.67). *Logically*, we could define the general solution of $Mdx + Ndy = 0$ in R to be the set of all solution curves in R. But as a practical matter, to "write down" a curve we must write down an equation or equations (possibly parametric, possibly not) to describe that curve. So, having allowed ourselves to call *algebraic* solutions of a differential-form DE simply *solutions* of that DE, we will content ourselves with a definition of "general solution" that is similar to Definition 3.34:

Definition 3.72 (General solution of a differential-form DE in a region) For a given differential-form DE

$$Mdx + Ndy = 0 \tag{3.139}$$

and a region R in the xy plane, we say that a collection $\mathcal{E}$ of algebraic equations in x and y is *an algebraic form of the general solution of* (3.139) *on* R if

(i) each equation in the collection $\mathcal{E}$ is a solution of (3.139) on R (see Definitions 3.68 and 3.66), and

(ii) every solution curve of (3.139) in R is contained in the graph of an equation in the collection $\mathcal{E}$.

For simplicity's sake, when a collection $\mathcal{E}$ of equations meets the conditions above, we will allow ourselves to drop the words "an algebraic form of", and simply call $\mathcal{E}$ *the general solution of equation* (3.139) *on* R—with the understanding that such a collection $\mathcal{E}$ is never unique.

When no region R is mentioned explicitly, it is assumed that R is the common implied domain of M and N. ■

3.3.6 Exact equations

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

Example 3.73 Suppose $M\,dx + N\,dy$ is an exact differential on a region R (see Definition 3.50), and let F be a differentiable function on R for which $M\,dx + N\,dy = dF$. Then (3.127) becomes

$$\frac{\partial F}{\partial x}\,dx + \frac{\partial F}{\partial y}\,dy = 0. \tag{3.140}$$

Suppose that $\mathcal{C}$ is a solution curve of (3.140), and that $t \mapsto (x(t), y(t))$, $t \in I$, is a continuously differentiable parametrization of $\mathcal{C}$. Then (3.128) says

$$\frac{\partial F}{\partial x}(x(t), y(t))\frac{dx}{dt} + \frac{\partial F}{\partial y}(x(t), y(t))\frac{dy}{dt} = 0. \tag{3.141}$$

By the multivariable Chain Rule (learned in Calculus 3), the left-hand side of (3.141) is just $\frac{d}{dt}F(x(t), y(t))$. Thus equation (3.128) simplifies, in this case, to

$$\frac{d}{dt}F(x(t), y(t)) = 0 \quad \text{for all } t \in I. \tag{3.142}$$

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

$$F(x(t), y(t)) = c_0 \tag{3.143}$$

for all $t \in I$. This implies that *every solution curve of* (3.140) *in* R *is contained in the graph of* (3.143) *for some value of the constant* c_0.

Now, fix a number c_0, and consider the equation

$$F(x, y) = c_0. \tag{3.144}$$

Is this equation a solution of (3.140) in R, according to Definition 3.68? The answer is yes, provided that the graph of (3.134) contains a smooth curve in R (criterion (i) of Definition 3.66). If this criterion is met, let $\mathcal{C}$ be a smooth curve in R that is contained in the graph of (3.144). Let γ be a regular 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 (3.144), equation (3.143) is satisfied for all $t \in I$. Differentiating both sides of (3.143) with respect to t, we find that equation (3.142) is satisfied. But, by the Chain Rule, the left-hand side of (3.142) is exactly the left-hand side of (3.141), so equation (3.141) is satisfied.

Therefore $\mathcal{C}$ is a solution curve of the differential equation (3.140). Hence criterion (ii) of Definition 3.66 is met, so (3.144) is a solution of the DE (3.140) in R. ■

Example 3.74 (General solution of an exact equation) Suppose we are given a differential-form equation

$$M\,dx + N\,dy = 0 \tag{3.145}$$

that is exact on a region R, and F is a function for which $M\,dx + N\,dy = dF$ on R. Then Example 3.73 shows that **one algebraic form of the general solution of** (3.145) **on** R **is the collection of equations**

$$\{F(x,y) = C\}\,, \tag{3.146}$$

where C **is a "semi-arbitrary" constant: the allowed values of** C **are those for which the graph of** $F(x,y) = C$ **contains a smooth curve in** R**.** To simplify the notation and terminology, we allow ourselves not to state this restriction on C explicitly in equation (3.146), and to refer to "$\{F(x,y) = C\}$" as the general solution of (3.145) on R. However, in cases in which we are able to identify the set of allowed values of C concretely (e.g. "$C > 0$"), we may incorporate the restrictions on C into equation (3.146).

Note that for $C_1 \neq C_2$, the graphs of $F(x,y) = C_1$ and $F(x,y) = C_2$ never intersect. Hence for the exact DE (3.145) and function F above, we can say something stronger than that (3.146) is an algebraic form of the general solution of the DE: every solution curve is contained in the graph of *one and only one* equation in the collection $\{F(x,y) = C\}$. ■

Blue portion below is optional reading.

For any real number C and (two-variable) function F, the graph of $F(x,y) = C$ is called a *level-set* of F.[81] 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 that 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 3.75 (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 $\frac{\partial F}{\partial x}(x,y) = \begin{cases} -3x^2, & x \geq 0 \\ 3x^2, & x \leq 0 \end{cases}$, so $\frac{\partial^2 F}{\partial x^2}(x,y) = \begin{cases} -6x, & x \geq 0 \\ 6x, & x \leq 0 \end{cases}$). It has one critical point, the origin. The level-set containing this critical point is the graph of

[81] The same terminology is used for functions of any number of variables.

$$y^3 - |x|^3 = 0, \tag{3.147}$$

which is simply the graph of $y = |x|$. The portion of this graph in the open first quadrant, namely $\{(x, x) : x > 0\}$ 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 (3.147) is a solution of

$$y^2\,dy + \left\{ \begin{array}{ll} -3x^2, & x \geq 0 \\ 3x^2, & x \leq 0 \end{array} \right\} dx = 0; \tag{3.148}$$

it meets both criteria in Definition 3.66. But as seen above, the graph of (3.147) contains a point, $(0, 0)$, that is not on any solution *curve* of (3.148) (see Definitions 3.60 and 3.58). 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 3.75 was a critical point of F (hence a singular point of the differential dF; see Example 3.65). In the absence of singular points, we can be much more concrete about the general solution of an exact equation:

If a differential $M\,dx + N\,dy$ is equal to dF on a region R, and has no singular points in R, then the set of C's allowed in (3.146) is simply the range of F on the region R, and every point in R is contained in a unique solution curve that is maximal in R. (3.149)

To see why this is true, the interested student may read Example 4.1 in the optional-reading Section 4.2.

3.3.7 Algebraic equivalence of DEs in differential form

Algebraic equivalence (see Definition 3.54) has the same importance for DEs in differential form that it has for DEs in derivative form. Suppose that two equations $M_1\,dx + N_1\,dy = 0$ and $M_2\,dx + N_2\,dy = 0$ 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 $M_1\,dx + N_1\,dy = 0$ and $t \mapsto (x(t), y(t))$, $t \in I$, is a regular 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 $M_2\,dx + N_2\,dy = 0$, and $t \mapsto (x(t), y(t))$ is a parametric solution of this DE. Hence every solution curve of $M_1\,dx + N_1\,dy = 0$ is a solution curve of $M_2\,dx + N_2\,dy = 0$, 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 $M_2\,dx + N_2\,dy = 0$ is a solution curve or parametric solution of $M_1\,dx + N_1\,dy = 0$. Adding Definition 3.66 to this analysis, we have the following:

$$\left.\begin{array}{l} \text{If two differential-form DEs are algebraically equivalent on a region} \\ R\text{, then in } R \text{ they have exactly the same solution curves, exactly} \\ \text{the same parametric solutions, and exactly the same solutions.} \end{array}\right\} \quad (3.150)$$

Combining this fact with Example 3.74, we have the following:

$$\left.\begin{array}{l} \textbf{If the equation } Mdx + Ndy = 0 \textbf{ is algebraically} \\ \textbf{equivalent to an exact equation } dF = 0 \textbf{ on} \\ \textbf{a region } R\textbf{, then } \{F(x,y) = C\} \textbf{ is the} \\ \textbf{general solution of } Mdx + Ndy = 0 \textbf{ in } R\textbf{.} \\ \textbf{The same understanding concerning the allowed} \\ \textbf{values of the constant } C \textbf{ in "}\{F(x,y) = C\}\textbf{"} \\ \textbf{applies as in Example 3.74.} \end{array}\right\} \quad (3.151)$$

Observe that if $M_2 = fM_1$ and $N_2 = fN_1$, but f is zero somewhere in R, then every solution curve (or parametric solution) of $M_1\,dx + N_1\,dy = 0$ is a solution curve (or parametric solution) of $M_2\,dx + N_2\,dy = 0$, 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 are 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 related to singular points of differentials (and to the reason that, in Definition 3.63 we required $M\,dx + N\,dy$ to have no singular points in R). [Blue portion below is optional reading.] 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 3.76 Consider the DE

$$2xy\,dx + (y^2 - x^2)dy = 0. \tag{3.152}$$

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 (3.136), we have a differential with one singular point, which happens to be the origin[82]. Letting $R = \{\mathbf{R}^2 \text{ minus the origin}\}$, Theorem 3.64 guarantees us that through each point $(x_0, y_0) \neq (0,0)$, there exists a unique solution curve of (3.152) that is maximal in R.

Observe that the positive x-axis is a solution-curve: if we set $x(t) = t$, $y(t) = 0$, on the t-interval $(0,\infty)$, 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 3.64 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 (3.152) is algebraically equivalent to

$$\frac{1}{y^2}\left(2xy\,dx + (y^2 - x^2)dy\right) = 0. \tag{3.153}$$

But as the student may verify, on both R_+ and R_- we have that

[82] 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 (3.153) is exact; it is dF, where $F(x, y) = \frac{x^2+y^2}{y}$. Hence one form of the general solution of (3.153), in either of these regions, is

$$\left\{\frac{x^2 + y^2}{y} = C\right\}, \tag{3.154}$$

where, from fact (3.149), 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 (3.154). 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,

$$\begin{aligned}&\frac{x^2 + y^2}{y} = 2b \\ \iff\quad & x^2 + y^2 = 2by, \\ \iff\quad & x^2 + y^2 - 2by = 0, \\ \iff\quad & x^2 + y^2 - 2by + b^2 = b^2, \\ \iff\quad & x^2 + (y - b)^2 = b^2.\end{aligned} \tag{3.155}$$

The graph of (3.155) 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 the DE (3.153) on R_+ and on R_-. But since equations (3.153) and (3.152) are algebraically equivalent on these regions, the same curves are all the maximal solution curves of the DE (3.152) in these regions.

We have now found all the solution curves of (3.152) 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

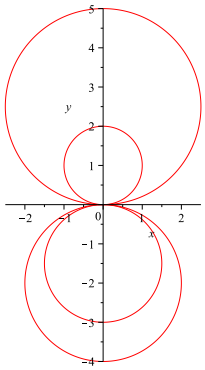

Figure 7: Some solution curves of $2xy\,dx + (y^2 - x^2)dy = 0$. (The graphing utility used to render this diagram does not do a good job near the origin; there should be no gap in any of the circles.)

lies on every one of the circles described by (3.155), 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 (3.152). We leave the student to check that each full circle (3.155), with the origin included, is also a solution curve of (3.152).

Thus, *among* the solution curves of (3.152) are all circles centered on the y axis, and an "exceptional" curve, the x-axis. A collection of algebraic solutions of (3.152) corresponding exactly to this set of solution curves is

$$\{x^2 + (y-b)^2 = b^2 : b \neq 0\} \quad \text{and} \quad \{y = 0\}. \tag{3.156}$$

From the foregoing analysis, it may appear that the set of all solution curves of (3.152) 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 *algebraic solutions* of (3.152) is (3.156). But both of these conclusions are this are wrong!

To see why, in Figure 7 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 x-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 3.58), and you have a perfectly good, smooth, solution curve that is not contained in any circle or in the x-axis.

In addition to the circular arcs (from P to the origin, and from the origin to Q) used above, you can let the x-axis into this game. 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 (3.153) that are not contained in any of the

"circles plus one straight line" family given by (3.156). It is possible to write down an algebraic equation for each of these other solution curves, but these will be new equations that aren't in the collection (3.156).

However, every regular parametrization γ of each of the "non-obvious" solution-curves just described has the property that there are two equations in the collection (3.156), and some t_0 in the domain of γ, such that $(x(t), y(t)) = \gamma(t)$ satisfies one of these equations for $t \leq t_0$ and satisfies the other for $t \geq t_0$. With a bit more work it can be shown that these are the only solution curves of the DE (3.152) that do not lie in the graph of a *single* equation in the collection (3.156). Thus (3.156) is (one algebraic form of) the general solution of the DE (3.152). ■

In the example above, an alternative way of expressing the collection (3.156) is as follows. In (3.154), 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 (3.154) 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" above, 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 collection (3.156) in a unified way as

$$\left\{ y = C(x^2 + y^2) \mid C \in \mathbf{R} \right\}. \tag{3.157}$$

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

Both (3.156) and (3.157) are algebraic forms of the general solution of (3.152). This serves as a reminder that, in general, an impression like "this solution (or equation) falls into a one-parameter family, while this other does not," can be *purely in the eye of the beholder*, depending heavily on the *form* in which you choose to express the set of all solutions, not on anything intrinsic to the set of all solutions itself.

We can also use Example 3.76 to exhibit one of the reasons it is difficult to give a satisfactory, useful, general definition of "maximal solution curve" of $M dx + N dy = 0$ in a region that includes singular points of $M dx + N dy$. For the sake of concreteness, using Figure 7 for reference, start at the point $P = (0, 1)$ and move counterclockwise along the "upper circle" $x^2 + (y - 1)^2 = 1$. When you reach the origin, continue by moving along the mirror-image "lower circle" $x^2 + (y + 1)^2 = 1$, clockwise, until you reach the point $Q = (0, -1)$. Deleting the endpoints in order to meet our definition of "smooth curve", you now have an open S-shaped curve smooth from P to Q. This curve is extendible to a larger solution curve: imagine dragging the starting-point P

clockwise along the upper circle, and dragging Q clockwise along the lower circle. We can drag P to any point in the open first quadrant lying on the upper circle, and can drag Q to any point in the open third quadrant lying on the lower circle. No matter how far we drag P or Q (subject to the quadrant restrictions), the curve we get is a solution curve of (3.152) that is extendible to a larger solution curve; we can always drag the endpoints farther, getting them closer and closer to the origin. Were we to allow P or Q to *reach* the origin, we would violate our definition of "smooth curve" (e.g. were we to let them both reach the origin, we'd have a figure-8). So there is no largest smooth solution curve that contains our S-shaped solution curve.

Thus there are solution curves of (3.153) that are not contained in any of the "circles plus one straight line" family given by (3.156). It is possible to write down an algebraic equation for each of these other solution curves, but these will be new equations that aren't in the collection (3.156).

However, every regular parametrization γ of each of the "non-obvious" solution-curves just described has the property that there are two equations in the collection (3.156), and some t_0 in the domain of γ, such that $(x(t), y(t)) = \gamma(t)$ satisfies one of these equations for $t \leq t_0$ and satisfies the other for $t \geq t_0$. With a bit more work it can be shown that these are the only solution curves of the DE (3.152) that do not lie in the graph of a *single* equation in the collection (3.156). Thus (3.156) is (one algebraic form of) the general solution of the DE (3.152).

In Example 3.76, all the solution curves in $\mathbf{R}^2$ intersected at the origin (a singular point of $M\,dx + N\,dy$) if these curves were extended far enough, but all had the same slope (zero) at the origin. Next we give an example of a very simple equation of the form $M\,dx + N\,dy = 0$ in which all the solution curves in $\mathbf{R}^2$ intersect at a singular point of $M\,dx + N\,dy$, but with all different slopes—in fact, with every possible slope.

Example 3.77 Consider the DE

$$x\,dy - y\,dx = 0. \tag{3.158}$$

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 $x\,dy - y\,dx$ is the origin. Therefore in $R = \{\mathbf{R}^2$ minus the origin$\}$, there is a unique maximal 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 inextendible solution curves of (3.158) in $\{\mathbf{R}^2$ minus the origin$\}$. Therefore every solution curve $\mathcal{C}$ in $\mathbf{R}^2$ that is *not* contained in one of these rays must pass through the origin. If we delete the origin from $\mathcal{C}$, what remains are two solution curves in R, so each of these must be a subset of a ray. For $\mathcal{C}$ to be

smooth, the two rays must be "opposite" to each other, so $\mathcal{C}$ is contained in a straight line though the origin. The full straight lines are inextendible. Thus the family of all straight lines through the origin is the set of inextendible solution curves of (3.158) in $\mathbf{R}^2$. Every point in $\mathbf{R}^2$ other than the origin lies on a unique one of these lines, and the origin lies on *all* of them. ■

3.4 Relation between differential form and derivative form

Definition 3.78 Let M, N be functions on a region R in $\mathbf{R}^2$. Consider the equations

$$M(x,y)\,dx + N(x,y)\,dy = 0, \tag{3.159}$$

$$M(x,y) + N(x,y)\frac{dy}{dx} = 0, \tag{3.160}$$

$$M(x,y)\frac{dx}{dy} + N(x,y) = 0. \tag{3.161}$$

We call equations (3.160) and (3.161) the *derivative-form DEs associated with the differential-form DE* (3.159). Similarly, we call equation (3.159) the *differential-form DE associated with the derivative-form DE* (3.160), and also the *differential-form DE associated with the derivative-form DE* (3.161).

More generally, if a derivative-form equation is algebraically equivalent to (3.160) or (3.161) on a region R, we call the equation *a derivative form of* (3.159) on R. Similarly, if a differential-form equation is algebraically equivalent to (3.159) on a region R, we call the equation a *differential form of* (3.160) *and* (3.161) on R.[83] ■

Remark 3.79 (Constant solutions of differential-form DEs) Note that equation (3.159) can conveivably have solutions of the form "$x =$ constant", and/or solutions of the form "$y =$ constant". As discussed in Remark 3.70, these correspond to solution curves of (3.159) that are vertical and horizontal lines, respectively. However, equation (3.160) has *no* vertical solution curves, and equation (3.161) has no horizontal solution curves.

[83]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.

If equation (3.159) has any solutions of the form $x = \text{constant}$, *we are guaranteed to lose these solutions if we replace equation* (3.159) *by the associated derivative-form equation* (3.160). The notation in (3.159)–(3.160) provides a convenient mnemonic device for remembering this. If we think of dx as "being" 0 on any vertical curve (see the end of Remark 3.70), and pretend that (3.160) is obtained from (3.159) by the (nonsensical) operation of "dividing by dx", then we can think of the loss of solutions $x = \text{constant}$ in passing from (3.159) to (3.160) as begin a result of dividing by zero. Similar comments apply to the relationship among equation (3.159), equation (3.161), and solutions of (3.159) of the form $y = \text{constant}$. ■

As observed in Remark 3.79, 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 (3.159) by dx or dy to obtain the associated equation (3.160) or (3.161), or formally "multiply" equation (3.160) or (3.161) by dx or dy to obtain the associated equation (3.159). *This is an* **extremely** *useful memory-device, and the student should not hesitate to use it,* **but mathematically it is garbage**.[84] 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 true fraction with numerator dy and denominator dx.

[Magenta portion below is optional reading]

We will see next just how and why equations (3.159)–(3.161) *actually* are related to each other.

To start, suppose that $\mathcal{C}$ is smooth curve, and γ a regular 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.[85] (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.)

[84] Unfortunately, most DE textbooks do not mention that this way of viewing the relations among (3.159), (3.160), and (3.161) 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.

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

Since γ is a non-stop parametrization, for every $t_0 \in I$ at least one of the two 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 an open subinterval I_1 that is either x-monotone or y-monotone.

Let I_1 be an open 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[86] assures us that $f|_{I_1}$ has an inverse function—which we will denote simply f^{-1}, rather than the more accurate $(f|_{I_1})^{-1}$—with domain an open interval I_2 and with range I_1, and that f^{-1} is continuously differentiable. 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 $\mathcal{C}_1$:

$$\gamma_{\text{new}}(t_{\text{new}}) = (t_{\text{new}}, \phi(t_{\text{new}})). \tag{3.162}$$

The variable in (3.162) 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

$$\gamma_{\text{new}}(x) = (x, \phi(x)). \tag{3.163}$$

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

[86] 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 h is a differentiable function on an open interval J, and $h'(t)$ is not 0 for any $h \in J$, then (i) the range of h is an open interval K, (ii) an inverse function h^{-1} exists, with domain K and range J, and (iii) h^{-1} is differentiable, with its derivative given by $(h^{-1})'(x) = 1/h'(h^{-1}(x))$. (If we write $x = h(t)$ and $t = h^{-1}(x)$, then the formidable-looking formula for the derivative of h^{-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}.

$$M(x,y)\,dx + N(x,y)\,dy = 0. \tag{3.164}$$

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

$$M(x(t),y(t))\frac{dx}{dt} + N(x(t),y(t))\frac{dy}{dt} = 0. \tag{3.165}$$

In particular this is true for the parametrization (3.163), in which the parameter t is x itself, and in which have $y(t) = \phi(t) = \phi(x) = y(x)$. Therefore, for all $x \in I_2$,

$$\begin{aligned} 0 &= M(x,\phi(x))\frac{dx}{dx} + N(x,\phi(x))\ \phi'(x) \\ &= M(x,\phi(x)) + N(x,\phi(x))\ \phi'(x). \end{aligned} \tag{3.166}$$

The right-hand side of (3.166) 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

$$M(x,y) + N(x,y)\frac{dy}{dx} = 0. \tag{3.167}$$

Therefore the portion $\mathcal{C}_1$ of $\mathcal{C}$ is the graph of a solution (namely ϕ) of the derivative-form differential equation (3.167). 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 $M\,dx + N\,dy = 0$ 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 $M + N\frac{dy}{dx} = 0$. (3.168)

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

$$M(x,y)\frac{dx}{dy} + N(x,y) = 0. \tag{3.169}$$

Therefore:

Every solution curve of the differential-form equation (3.159) **is a union of solution curves of the derivative-form equations** (3.160) **and** (3.161)**.** (3.170)

Example 3.61 provides an illustration of fact (3.170). We observed in that example that the circle $x^2 + y^2 = 25$ is a solution curve of $x\,dx + y\,dy = 0$ but cannot be expressed as a union of graphs of functions of x alone. Both $y = \sqrt{25 - x^2}$, and $y = -\sqrt{25 - x^2}$ are solutions of the associated derivative-form DE $x + y\frac{dy}{dx} = 0$ on the *open* interval $-5 < x < 5$, but the union of the corresponding graphs (solution curves of $x + y\frac{dy}{dx} = 0$) is not the whole dircle. *No* solution-curve of $x + y\frac{dy}{dx} = 0$ can include the point $(5, 0)$ or $(-5, 0)$, because the circle's tangent line at these points is vertical. However, the circle *is* the union of the graphs of $y = \sqrt{25 - x^2}$, and $y = -\sqrt{25 - x^2}$ (both for $-5 < x < 5$) and the graphs of $x = \sqrt{25 - y^2}$, and $x = -\sqrt{25 - y^2}$ (both for $-5 < y < 5$). The first two of these graphs are solution curves of the associated derivative-form equation $x + y\frac{dy}{dx} = 0$, while the other two are solution curves of the associated derivative-form equation $x\frac{dx}{dy} + y = 0$.

In general, just as in the circle example above, the solution-curves mentioned in (3.170) will overlap, since the x-monotone intervals and y-monotone intervals of a regular parametrization γ will usually overlap. (The only instances in which there will not be overlap are those in which the solution curve of the differential-form DE is a horizontal or vertical line.)

[Magenta portion below is optional reading.]

Now compare (3.167) with the general first-order derivative-form DE with independent variable x and dependent variable y,

$$\mathsf{G}(x, y, \frac{dy}{dx}) = 0. \tag{3.171}$$

Equation (3.167) is a special case of (3.171), 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 (3.167) 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, (3.171) could be a much more complicated equation, such as

$$\left(\frac{dy}{dx}\right)^3 + (x + y)\sin(\frac{dy}{dx}) + xe^y = 0. \tag{3.172}$$

Solving equations such as the one above is *much* harder than is solving equations of the simpler form (3.167). For certain functions G that are more complicated than

(3.167), but much less complicated than (3.172), methods of solution are known[87]. But the general theory and techniques for working with equation (3.171) for general G's are much less highly developed than they are for equations in the standard form (3.174) or in the form (3.167).

One of the features of equation (3.167) that makes it so special is that on any region on which $N(x, y) \neq 0$, (3.167) is algebraically equivalent to

$$\frac{dy}{dx} = -\frac{M(x, y)}{N(x, y)}, \tag{3.173}$$

which is of form

$$\frac{dy}{dx} = f(x, y). \tag{3.174}$$

Recall that equation (3.174) 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 (when applicable) of finding solutions of nonlinear DEs actually give us *all* solutions.

[Magenta portion below is optional reading.]

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 for the general first-order DE (3.3). These facts apply just as well to nasty DEs like (3.172) as they do to (relatively) nice ones like (3.174). However, in all of our *examples*, we used equations that were algebraically equivalent to (3.160) (hence also to (3.174)) on some region. The reason is that although the concept of "the set of all solutions" makes perfectly good sense for the general equation (3.171), I kept to examples in which I 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 (3.174). 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 (3.174) on some region**.

[87] One such type equation is a *Clairaut equation* $y = x\frac{dy}{dx} + g(\frac{dy}{dx})$, which is equivalent to (3.171) 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.

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 (3.160) 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 (3.170) gives one such relation.

To have a name for equations that are explicitly of the form (3.160) or (3.161), let us say that a derivative-form equation, with independent variable x and dependent variable y, is in "almost-standard form"[88] if it is in the form (3.160), 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 (3.170), you will see that it also shows that every solution curve of (3.160) or (3.161) is a solution curve of (3.159). Thus:

$$\left.\begin{array}{l}\textbf{Every solution curve of a derivative-form DE}\\ \textbf{in almost-standard form is a solution curve}\\ \textbf{of the associated differential-form equation.}\end{array}\right\} \tag{3.175}$$

Combining (3.170) and (3.175), we conclude the following:

$$\left.\begin{array}{l}\textbf{A smooth curve } \mathcal{C} \textbf{ is a solution curve of a DE}\\ \textbf{in differential form if and only if } \mathcal{C} \textbf{ is a union of}\\ \textbf{solution curves of the associated derivative-form}\\ \textbf{equations.}\end{array}\right\} \tag{3.176}$$

We emphasize that in deriving these relations, the transition from the differential-form DE (3.159) to the derivative-form DEs (3.167) and (3.169) was NOT obtained 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 $M\,dx + N\,dy = 0$ are curves whose parametrizations satisfy (3.165), and that for particular choices of the parameter on the intervals that we called "x-monotone" or "y-monotone", (3.165) reduces to (3.160) or (3.161).

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

[88] This is another bit of terminology invented only for these notes, just to have a name to distinguish (3.160) from (3.173) on regions in which $N(x, y)$ may be zero somewhere.

many true statements by *pretending, momentarily*, that you can multiply or divide by a differential just as if it were a real number[89]. In particular, we can use this principle help us easily *remember* that the differential-form equation (3.159) is related to (but not the same as!) the derivative-form equations (3.160) and (3.161). But this notational trick doesn't tell us *everything*, such as the *precise relationship* among these equations, which is statement (3.175) (of which statement (3.170) is the "only if" half).

3.5 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 nonlinear derivative-form equations is essentially the following. Below, assume that you are given a derivative-form equation with independent variable x and dependent variable y, and that this DE can be "put in standard form".

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 the result of Step 3 in the form "$\{F(x,y) = C\}$" if you used the equation by the "exact equations method", or in the form "$\{H(y) = G(x)+C\}$" if you separated variables. Then hope that what you've just written is set of all solutions, in implicit form, of the original derivative-form DE—or at least that you've written down enough solutions that your instructor will mark your answer as correct.

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

No doubt you noticed the phrase, "[H]ope that what you've just written is set of all solutions, in implicit form, of the original derivative-form DE." All we did above

[89]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.

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 differential-form equation written in Step 1, or vice-versa, or whether the 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 at an example that illustrates one of the problems with the procedure as outlined.

Example 3.80 Solve the differential equation

$$(10xy^9 + 2xy)\frac{dy}{dx} = -(3x^2 + 1 + y^{10} + y^2). \tag{3.177}$$

(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[90]

$$(3x^2 + 1 + y^{10} + y^2)dx + (10xy^9 + 2xy)dy = 0. \tag{3.178}$$

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 $M\,dx + N\,dy$ passes the exactness test "$M_y = N_x$", so we know that there is some F, continuously differentiable on all of $\mathbf{R}^2$, for which the left-hand side of (3.178) is dF. Using our usual method, we find that an F with this property is

$$F(x, y) = x^3 + x + xy^{10} + xy^2. \tag{3.179}$$

From Example 3.74, we know that the general solution of (3.178) is

$$\{x^3 + x + xy^{10} + xy^2 = C\}, \tag{3.180}$$

where C is (at worst) a semi-arbitrary constant. Fact (3.149) shows that the set of allowed values of C is simply the range of F, provided that $M\,dx+N\,dy$ has no singular

[90]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 (3.177) from the left-hand side; the other is obtained by first subtracting the lefft-hand side of (3.177) from the right-hand side. Each of these equations is just the other with both sides multiplied by -1.

points. Looking at $M(x, y)$, we observe that x^2, y^{10}, and y^2 are all non-negative, so $M(x, y) \geq 1$. In particular, $M(x, y)$ is nowhere zero, so $M\,dx + N\,dy$ has no singular points. So fact (3.149) 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 (3.179) 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 (3.178) is the family of equations (3.180), with C a *completely* arbitrary constant; all real values are allowed.

But the equation we wanted to solve was the derivative-form equation (3.177), not the differential-form equation (3.178), so we ask: is this same family (3.180) the general solution of (3.177), in implicit form? The answer is no.

To see why, consider the equation in $\mathcal{E}$ corrsponding to $C = 0$:

$$x^3 + x + xy^{10} + xy^2 = 0. \tag{3.181}$$

(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

$$x(x^2 + 1 + y^{10} + y^2) = 0. \tag{3.182}$$

The quantity inside parentheses is strictly positive, so (3.182) is equivalent to just $x = 0$. The graph of (3.182) is simply the y-axis, a perfectly nice smooth curve, and a perfectly good solution curve of (3.178), 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 the derivative-form DE 3.177, so equation (3.182) is not an implicit solution of (3.177). Therefore $\mathcal{E}$ does not meet Definition 3.34's first criterion for "general solution, in implicit form, of a derivative-form DE".

This demonstrates the main point of this example:

A collection of equations can be an implicit form of the general solution of an almost-standard-form derivative-form DE, yet not an algebraic form of the general solution of the associated differential-form DE, and vice-versa. (3.183)

("Almost-standard form" was defined a few lines before statement 3.175.)

However, while the two general solutions in (3.183) need not be *identical*, our conclusions in the previous section show that they are closely related. A consequence of fact (3.175) is the following:

If a collection $\mathcal{E}$ of algebraic equations is the general solution of a DE $M(x,y)\,dx + N(x,y)\,dy = 0$, and the graphs of no two equations in $\mathcal{E}$ overlap (i.e., if the graphs intersect at all, there is no *curve* contained in the intersection), then $\mathcal{E}$ *contains* an implicit form of the general solution of $M(x,y) + N(x,y)\frac{dy}{dx} = 0$.

Thus, if we are trying to obtain the general solution of $M(x,y) + N(x,y)\frac{dy}{dx} = 0$ from having found $\mathcal{E}$ as the general solution of $M(x,y)\,dx + N(x,y)\,dy = 0$, we need only worry whether $\mathcal{E}$ has any equations that are *not* implicit solutions of the derivative-form equation, or has any equations whose graphs overlap each other. (3.184)

The reason for the no-overlap restriction in fact (3.184) is that in Definition 3.34, for collection of algebraic equations to be an implicit form of the general solution of a derivative-form DE , we required that every maximal solution-curve of the DE lie in the graph of a *unique* equation in the collection; in the analogous Definition 3.72 for differential-form DEs, no uniqueness was required. If an algebraic form $\mathcal{E}$ of the general solution of $M\,dx + N\,dy = 0$ has two equations whose graphs overlap in a curve $\mathcal{C}$, and $\mathcal{C}$ is not a vertical line segment, then $\mathcal{C}$ is a solution curve of $M+N\frac{dy}{dx} = 0$ lying in the graphs of two equations in $\mathcal{E}$. Potentially, $\mathcal{C}$ is a *maximal* solution curve, in which case $\mathcal{E}$ would not meet our definition of "general solution, in implicit form" for a derivative-form DE.

But when the differential $M dx + N dy$ is *exact*, we can simplify fact (3.184) a great deal. As noted in Example 3.74, for an equation-family of the form $\{F(x,y) = C\}$, no two graphs intersect, let alone overlap. Thus:

If the differential $M\,dx + N\,dy$ is exact, and $\mathcal{E}$ is any algebraic form of the general solution of $M(x,y)\,dx + N(x,y)\,dy = 0$, then $\mathcal{E}$ *contains* an implicit form of the general solution of $M(x,y) + N(x,y)\frac{dy}{dx} = 0$. We can obtain the general solution of $M(x,y) + N(x,y)\frac{dy}{dx} = 0$ by removing from $\mathcal{E}$ any "spurious solutions" of the derivative-form DE, i.e. equations that in $\mathcal{E}$ that are not implicit solutions of the derivative-form equation. (3.185)

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 (3.177)? The answer is no. (This can be shown using the Implicit Function Theorem, but in the interests of brevity, and since demonstrating fact (3.183) was the main point of the current example, we will omit the argument.) Thus the general solution of (3.177), in implicit form, is

$$\left\{x^3 + x + xy^{10} + xy^2 = C, \quad C \neq 0\right\}. \tag{3.186}$$

■

In the example above, facts 3.183, 3.184, and 3.185 were stated without reference to a region R, for simplicity's sake; the relevant region in the example was the whole xy plane. However, they remain true with the words "on a region R" inserted in the appropriate places.

What Example 3.80 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 of 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 chief purpose 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 (3.183) 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. You can "shoot first and ask questions later", as long as you don't forget the "ask questions later" part. The "autopilot" procedure is not worthless; it's simply not perfect. The behavior seen in Example 3.80 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 the general solution of a DE $M(x,y)\,dx + N(x,y)\,dy = 0$, then $\mathcal{E}$ will also be an the general solution of the associated derivative-form DE $M(x,y) + N(x,y)\frac{dy}{dx} = 0$, in implicit form.** 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, as was the case with equation (3.182). It is obvious that equations written in the form "$x =$ constant" are not implicit solutions of a derivative-form DE whose independent variable is x, but when a whole *family* of equations is given, such as $x^3 + x + xy^{10} + xy^2 = C$ (equation (3.180)), 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 differential equations than Example 3.80, but a more complicated "spurious solution".

Example 3.81 Suppose we wish to find the general solution of

$$(y^2 + 1)\cos x\, dx + 2y \sin x\, dy = 0. \tag{3.187}$$

One of the associated derivative-form DEs is

$$(y^2 + 1)\cos x + 2y \sin x\, \frac{dy}{dx} = 0. \tag{3.188}$$

Equation (3.187) is exact. Its general solution is

$$\left\{(y^2 + 1)\sin x = C\right\} \tag{3.189}$$

where C is an arbitrary constant. For $C \neq 0$, every point (x, y) in the graph of (3.189) has $\sin x \neq 0$, hence $y^2 + 1 = \frac{C}{\sin x} = C \csc x$. As the student may check, the latter equation is an implicit solution of (3.188); the general solution of (3.188), in implicit form, can be written as

$$\left\{y^2 + 1 = C \csc x, \quad C \neq 0\right\} \tag{3.190}$$

or as

$$\left\{(y^2 + 1)\sin x = C, \quad C \neq 0\right\}. \tag{3.191}$$

However, for $C = 0$, equation (3.189) 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 (3.188).

So in this example, we again need to throw away only one *equation* from the algebraic form (3.189) 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. ■

In general, an algebraic equation (say $F(x, y) = 0$) is a solution of the differential-form equation $M(x, y)\,dx + N(x, y)\,dy = 0$, yet not an implicit solution of 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.

If we have an equation-collection $\mathcal{E}$ that is (an algebraic form of) the general solution of the differential-form DE, and we remove from $\mathcal{E}$ all equations whose graphs have the two properties above, then the remaining collection of equations is the general solution, in implicit form, 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 both (an algebraic form of) the general solution of the differential-form DE, and an implicit form of the general solution of the associated derivative-form DE.

It should be noted that even when an algebraic equation, say $F(x, y) = 0$, is a solution of both $M(x, y)dx + N(x, y)\,dy = 0$ and $M(x, y) + N(x, y)\frac{dy}{dx} = 0$ (implicitly, in the latter case), 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 two-variable function F (whose formula we will not write down) for which the graph of $F(x, y) = 0$ is the oval in Figure 8. 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 multiplication/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 $M\,dx + N\,dy = 0$, adding/subtracting differentials will take it out of this form). *Any time we perform*

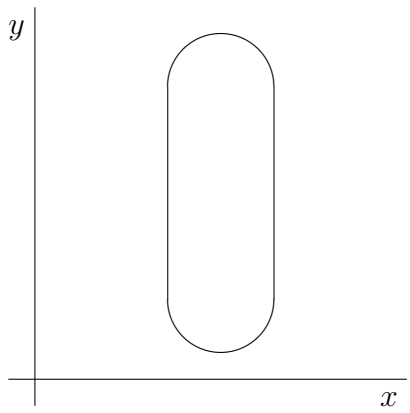

Figure 8:

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

$$\frac{dy}{dx} = f(x, y) \tag{3.192}$$

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

$$f_1(x, y)\frac{dy}{dx} = f_2(x, y) \tag{3.193}$$

If (3.192) or (3.193) 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

$$-f(x, y)\,dx + dy = 0 \tag{3.194}$$

and

$$-f_2(x, y)\,dx + f_1(x, y)\,dy = 0. \tag{3.195}$$

If we are extremely lucky, then (3.194) or (3.195) will be exact.

In the case of (3.194), 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 (3.192) was already of the form $\frac{dy}{dx} = g(x)$, solvable just by integrating g; there is no need even to look at equation (3.194).

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 one whose associated differential-form DE is exact (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

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

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

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

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

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

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

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

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 forms in which we could choose to write the general solution, one of which is

$$\left\{ y + y^3 \cos^2 x = C \right\} \tag{3.196}$$

(Note: "obvious" manipulations such as clearing fractions, plus writing down the associated differential-form DE, will not always lead to an exact DE. I 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 (3.185) guarantees us that the family of equations (3.196) *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 (3.196) contains a general solution, in implicit form, of our original derivative-form DE. The only question is whether the family (3.196) contains "spurious solutions"—equations that are 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 (3.196) has such a graph[91]. Therefore (3.196) 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 (begin aware that 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. 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) 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.

[91] 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.

[Magenta portion below is optional reading.]

Example 3.82 Solve the differential equation

$$\frac{dy}{dx} = -\frac{2x+2y}{2x+3y^2} \ . \tag{3.197}$$

First we observe that since the right-hand side of (3.197) is not defined when $2x + 3y^2 = 0$, the only regions in which "solution of (3.197)" 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, (3.197) is algebraically equivalent to

$$(2x + 2y) + (2x + 3y^2)\frac{dy}{dx} = 0, \tag{3.198}$$

whose associated differential-form equation is

$$(2x + 2y)dx + (2x + 3y^2)dy = 0. \tag{3.199}$$

Equation (3.199) 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 (3.199) 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 (3.197) is guaranteed to be a solution of (3.198), so in passing from (3.197) to (3.198) 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 (3.198) to (3.199). The latter possibility is easy to rule out: it is easy to see that (3.199) 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 (3.199) is a solution curve of (3.198)

To see whether the graph of $x^2 + 2xy + y^3 = C$, for a given C, is an implicit solution of (3.197) 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 (3.149) assures us that the general solution of (3.199) in $\{\mathbf{R}^2$ 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 (3.199) 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 (3.197) to (3.198) introduced any spurious solutions: equations $x^2+2xy+y^3=C$ that are not implicit solutions of (3.197). 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 $\frac{9}{4}y^4-2y^3=C$ (a polynomial of degree four has at most four distinct roots), 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 (3.197). Therefore there are no values of C that we need to exclude, and no spurious solutions. The general solution of (3.197) 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 the original DE (3.197).) ■

In all examples we've looked at so far, in which we used an associated differential-form DE to help us solve a derivative-form DE, the only spurious solutions this process ever introduced 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, using a derivative-form DE, could have been presented before we ever talked about differential-form DEs, but we have placed it in this section of the notes as a reminder.

[Magenta portion below is optional reading.]

Example 3.83 (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 $y_0 = x_0$. Hence for every $x_0 \in \mathbf{R}$, the initial-value problem

$$\frac{dy}{dx} = f(x, y), \quad y(0) = 0, \tag{3.200}$$

has a unique maximal solution.

If we substitute the definition of $f(x, y)$ into (3.200), the DE becomes

$$\frac{dy}{dx} = \begin{cases} \frac{e^y - e^x}{y - x} & \text{if } y \neq x, \\ e^x & \text{if } y = x. \end{cases} \tag{3.201}$$

This equation is neither linear nor separable, so in an attempt to solve we might write down the associated differential-form equation, which is

$$-\left\{ \begin{array}{ll} \frac{e^y - e^x}{y - x} & \text{if } y \neq x \\ e^x & \text{if } y = x \end{array} \right\} dx + dy = 0. \tag{3.202}$$

It is natural to try to rewrite (3.202) 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 (3.202) by $y - x$ we obtain

$$-(e^y - e^x)dx + (y - x)dy = 0, \tag{3.203}$$

which certainly *looks* much simpler than (3.202). 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 (3.200)?

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 $e^x \equiv 1$ (other than the single-point interval $[0, 0]$. degenerate. Thus 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 (3.202) with the explicit two-line formula for f, we write it simply as

$$-f(x, y)dx + dy = 0, \tag{3.204}$$

and if, when we multiply through by $y - x$, we write the result as

$$-(y-x)f(x,y)dx + (y-x)dy = 0 \tag{3.205}$$

rather than in the "simpler" form (3.203). It is obvious that $y = x$ is a solution of (3.205), whether or not it is a solution of (3.204). Less obvious, but true, is what we checked above: that $y = x$ is *definitely not* a solution of (3.200), hence not a solution of (3.204).

In this example, the general solution of (3.205) consists of the general solution of (3.204) *plus* the straight line $\{y = x\}$. The equation (3.204) has no solutions of the form $x =$ constant, so any algebraic form of the general solution of (3.204) 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 (3.203), 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 (3.202) and (3.203) 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. ■

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

When trying to solve a differential-form DE, passing to an associated derivative-form DE is generally not useful unless (at least) one of the associated derivative-form DEs is linear. (If an associated derivative-form DE is separable, you will wind up converting back to a differential-form DE in the separation-of-variables process.) When one of the associated derivative-form DEs *is* linear, that fact can be exploited, as the following two examples illustrate.

Example 3.84 Consider the equation

$$(2y + 3x)dy + 5dx = 0. \tag{3.206}$$

This equation is not exact. The simplest associated derivative-form DEs are

$$(2y + 3x)\frac{dy}{dx} + 5 = 0 \tag{3.207}$$

and

$$2y + 3x + 5\frac{dx}{dy} = 0. \tag{3.208}$$

Equation (3.207) is nonlinear, but (3.208) is linear. We can solve the latter DE by our usual method for linear DEs (remembering that in equation (3.208), y is the independent variable and x is the dependent variable, not the other way around), obtaining the general solution

$$\left\{x = -\frac{2}{3}y + \frac{10}{9} + Ce^{-\frac{3}{5}y}\right\}. \tag{3.209}$$

From the discussion in Section 3.4, the graph of each equation in the collection (3.209) is a solution-curve of the differential-form DE (3.206), and the only *potential* solution curves of (3.206) not represented in (3.209) are graphs of equations of the form $y = c$, where c is a constant. (Since y is now the independent variable, *horizontal* lines in the xy planes are the only potential solution-curves lost in passing from (3.206) to (3.208).) But since $5 \neq 0$, there are no values of c for which "$y = c$" is a solution of (3.206) (see Remarks 3.70 and 3.79). Thus the collection (3.209) is (an algebraic form of) the general solution of equation (3.206).

In the next example, the linearity of an associated derivative-form DE enables us to find *most* of the solutions fairly quickly. But finding *all* of them, and knowing that we've found them all, is much trickier; several subtleties discussed in Sections 3.2.9 and 3.4 are involved.

Example 3.85 We will solve the equation

$$(2y - xe^x)dx + x\,dy = 0. \tag{3.210}$$

Equation (3.210) is not exact on any region in the xy plane, but the associated derivative-form DE

$$2y - xe^x + x\frac{dy}{dx} = 0 \tag{3.211}$$

is linear, so we can obtain *most* (if not all) of the solutions of equation (3.210) by solving equation (3.211).

To solve (3.211) by our usual integrating-factor method, we divide-through by x and rewrite the new equation as

$$\frac{dy}{dx} + \frac{2}{x}y = e^x. \tag{3.212}$$

However, because of the division by x, the linear equations (3.211) and (3.212) are not algebraically equivalent on the whole real line; they are algebraically equivalent

only on the x-intervals $(-\infty, 0)$ and $(0, \infty)$. Thus, we know in advance that we may (potentially) miss any solution of equation (3.211) whose domain includes $x = 0$.

Solving equation (3.212) by the integrating-factor method yields the collection of equations

$$\{y = e^x(1 - 2x^{-1} + 2x^{-2}) + Cx^{-2}\}, \tag{3.213}$$

representing one 1-parameter family of solutions of (3.212) on $(-\infty, 0)$ and another on $(0, \infty)$.

To see if we lost any solutions of (3.211) when passing to (3.212)—i.e. any solutions that are defined on an open x-interval containing 0—suppose that $y = \phi(x)$ is a solution on some such interval (a, b) (with $a < 0$ and $b > 0$). Then on the interval $(a, 0)$ our function ϕ is (the restriction of) one of the solutions represented in (3.212), and on $(0, b)$ our ϕ is (the restriction of) another. Thus, for some constants C_1, C_2 we have

$$\phi(x) = \begin{cases} e^x(1 - 2x^{-1} + 2x^{-2}) + C_1x^{-2}) & \text{if } a < x < 0, \\ e^x(1 - 2x^{-1} + 2x^{-2}) + C_2x^{-2}) & \text{if } 0 < x < b. \end{cases}$$

Since ϕ is a solution of a differential equation, ϕ is continuous, so the two one-sided limits of ϕ at $x = 0$ must exist and be equal. With some work (which the student should be able to do, though not effortlessly), it can be shown that $\lim_{x\to 0+} \phi(x)$ exists if and only if $C_1 = -2$, and $\lim_{x\to 0-} \phi(x)$ exists if and only if $C_2 = -2$, in which case both limits are 0. It can also be shown that the corresponding function on $(-\infty, \infty)$,

$$\phi_{\text{special}}(x) = \begin{cases} e^x(1 - 2x^{-1} + 2x^{-2}) - 2x^{-2}) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases} \tag{3.214}$$

is differentiable at $x = 0$. Thus, substituting "$y = \phi_{\text{special}}(x)$" into equation (3.211) yields a true statement at $x = 0$, as well as everywhere else. (The value of $\phi'_{\text{special}}(0)$ happens to be $\frac{1}{3}$, but this value does not affect whether equation (3.211) is satisfied at $x = 0$, since in this equation $\frac{dy}{dx}$ is multiplied by 0 at $x = 0$.) Thus, the set of inextendible solution curves of (3.211) consists of (i) the graphs of all the equations in (3.213) on the x-interval $(-\infty, 0)$, (ii) the graphs of all the equations in (3.213) on the x-interval $(0, \infty)$, and (iii) the graph of $y = \phi_{\text{special}}(x)$.

We have now found all the maximal solutions of (3.211), but must determine whether any solution-curves of (3.210) were lost when we passed from (3.210) to (3.211). From the discussion in Section 3.4, the only potential solution-curves of (3.210) that are not solution-curves of (3.211) are vertical lines, graphs of equations of the form $x = c$ (where c is a constant). Plugging into (3.210), there is one and only one value of c, namely 0, for which $x = c$ is a solution of (3.210). (Thus, we *did* lose a solution-curve in passing from the differential-form DE to an associated derivative-form DE, but we lost only one. Again see Remarks 3.70 and 3.79).

Thus (an algebraic form of) the general solution of equation (3.210) is

$$\{y = e^x(1 - 2x^{-1} + 2x^{-2}) + Cx^{-2}\} \bigcup \{y = \phi_{\text{special}}(x)\} \bigcup \{x = 0\}.^{92}$$

■

[Blue portion below is optional reading.]

Remark 3.86 The differential on the left-hand side of equation (3.210) has one (and only one) singular point: the origin, $(0, 0)$. We have cautioned that general solutions of DEs in differential form can be difficult to write down in regions that include a singular point of the differential, because there may be more than one solution-curve passing through a given singular point; cf. Example 3.76. In Example 3.85, we exhibited two solution-curves that pass through $(0, 0)$—the graph of $x = 0$ and the graph of $y = \phi_{\text{special}}(x)$. Every solution-curve coincides with one of these *except possibly at the origin*. However, a careful analysis of (3.210) should cover the possibility that there might be some bifurcation of solution-curves at the origin; e.g. could we approach the origin along the solution curve $x = 0$ and go out from the origin along the solution curve $y = \phi_{\text{special}}(x)$? The answer is no, because such a curve would not be smooth: the graph of $y = \phi_{\text{special}}(x)$ has finite slope at the origin, while the graph of $x = 0$ has infinite slope. Thus, the analysis in Example 3.85 did indeed find all the solution curves of (3.210).

3.7 Summary of some results about differential-form DEs

In this section, R is a region in the xy plane, M and N are functions on R, and we consider the differential-form DE

$$M\,dx + N\,dy = 0 \tag{3.215}$$

on R. Almost all definitions in this summary are abridged, with a referral given to the complete definition.

3.7.1 Definitions

1. *Curve* and *curve-parametrization*, defined.

[92] On a timed exam in a class at this level, the author would give full credit for the final answer "$y = e^x(1 - 2x^{-1} + 2x^{-2}) + Cx^{-2}$ and $x = 0$"; he would not expect students to find the solution $y = \phi_{\text{special}}(x)$, or even to realize that there might be *some* value(s) of C for which "$e^x(1 - 2x^{-1} + 2x^{-2}) + Cx^{-2}$" has a finite limit as $x \to 0$.

A curve $\mathcal{C}$ in $\mathbf{R}^2$ is the set of points "traced out" by the parametric equations $x = f(t),\ y = g(t)$, as t (the *parameter*) varies over some (positive-length) interval I. The $\mathbf{R}^2$-valued function $t \mapsto (f(t), g(t)),\ \ t \in I$, is called a *parametrization of* $\mathcal{C}$ (or simply a *curve-parametrization* when the curve $\mathcal{C}$ is not mentioned explicitly). We often write $(f(t), g(t))$ simply as $(x(t), y(t))$. [See Definition 3.55 for more details.]

2. *Regular parametrization* of a curve, defined.

A curve-parametrization $t \mapsto (x(t), y(t)),\ \ t \in I$, is called *regular* if it is *continuously differentiable* (meaning: the derivatives $x'(t), y'(t)$ exist at every $t \in I$ and are continuous in t) and *non-stop* (meaning: the *velocity vector* $(x'(t), y'(t)$ is not the zero vector $(0, 0)$ for any $t \in I$). [See Definition 3.57 for more details.]

3. *Smooth curve*, defined.

A curve $\mathcal{C}$ in $\mathbf{R}^2$ is *smooth* if for every point (x_0, y_0) on $\mathcal{C}$, there is an open rectangle containing (x_0, y_0) such that the portion of $\mathcal{C}$ lying inside that rectangle admits a regular parametrization $t \mapsto (x(t), y(t))$ on some open t-interval. [See Definition 3.58 for more details.]

4. *Solution curve,* defined.

A *solution curve* of the differential-form DE (3.215) on R is a smooth curve $\mathcal{C}$, contained in R, admitting a regular parametrization $t \mapsto (x(t), y(t))$ that satisfies

$$M(x(t), y(t))\frac{dx}{dt} + N(x(t), y(t))\frac{dy}{dt} = 0 \tag{3.216}$$

at every t in the domain-interval of the parametrization. [See Definition 3.60 for more details.]

6. *Singular point of a differential*, defined.

A point (x_0, y_0) is a *singular point* of the differential $M\,dx + N\,dy$ if $M(x_0, y_0) = 0 = N(x_0, y_0)$. [(Definition 3.62)]

6. *Inextendible in* R (condition on *any* smooth curve), defined.

A smooth curve $\mathcal{C}$ lying in a region R in $\mathbf{R}^2$ is *inextendible in* R if either (i) $\mathcal{C}$ is a closed curve , or (ii) $\mathcal{C}$ is an "open curve without endpoints" and there is no smooth curve in R that contains $\mathcal{C}$ as a proper subset.[See Definition 3.59 and the paragraph below it for details.]

7. *Maximal in* R (condition on a *solution* curve), defined.

Assume that $M\,dx + N\,dy$ has no singular points (otherwise this definition does

not apply). Suppose $\mathcal{C}$ is a curve in R that is a solution curve of the equation $M\,dx + N\,dy = 0$. We say that $\mathcal{C}$ is *maximal in* R (as a solution curve of this DE) if $\mathcal{C}$ is inextendible in R. [See Definition 3.63.]

8. *"Almost standard form"* (terminology just for these notes) for derivative-form DEs, defined.

In these notes, we say that a derivative-form equation, with independent variable x and dependent variable y, is in "almost-standard form" if it is in the form $M+N\frac{dy}{dx} = 0$, or can be put in that form just by subtracting the right-hand side from the left-hand side.

3.7.2 Results (facts *shown* to be true)

1. (*General solution of an exact equation*) (Derived fact, not definition.) Suppose that F is a function for which $M\,dx + N\,dy = dF$ on R. Then one algebraic form of the general solution of (3.215) on R is the collection of equations

$$\{F(x, y) = C\}\,, \tag{3.217}$$

where C is a "semi-arbitrary" constant: the allowed values of C are those for which the graph of $F(x, y) = C$ contains a smooth curve in R. (See Example 3.74 for more details.)

2. *General solution of equation that's algebraically equivalent exact equation.*

If the equation $Mdx + Ndy = 0$ is algebraically equivalent to an exact equation $dF = 0$ on a region R, then $\{F(x, y) = C\}$ is the general solution of $Mdx + Ndy = 0$ in R. The same understanding concerning the allowed values of the constant C in "$\{F(x, y) = C\}$" applies as in ' General solution of an exact equation" above. [Fact (3.151).]

3. If a solution curve of $M\,dx + N\,dy = 0$ 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 $M + N\frac{dy}{dx} = 0$.[Fact (3.168).] The analogous statement with x and y reversed also holds. —small [Paragraph after fact (3.168)].

4. Every solution curve of a derivative-form DE in almost-standard form $M+N\frac{dy}{dx} = 0$ or $M\frac{dx}{dy} + N = 0$ is a solution curve of the associated differential-form equation $Md + Ndy = 0$. [Fact 3.175.] Every solution curve of this differential-form equation is a *union* of solution curves of the two associated derivative-form equations. [Fact 3.170.]

It follows that a smooth curve $\mathcal{C}$ is a solution curve of a differential-form DE *if and only if* $\mathcal{C}$ is a union of solution curves of the two associated derivative-form DEs. [Fact 3.176.]

3.8 "Tricks"

[This is a reminder to myself to write this section, some day.]

4 Optional Reading

4.1 The meaning of a differential

For the interested student, in this section we ascribe meaning to a differential.[93] Understanding this meaning 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 $M\,dx + N\,dy$ 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 "$M\,dx + N\,dy$ *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 $(M\,dx + N\,dy)[\gamma]$. This is the same function that we expressed in terms of t in the previous paragraph:

$$\overbrace{\underbrace{(M dx + N\,dy)[\gamma]\ (t)}_{\substack{\text{value of the function}\\ (M\,dx + N\,dy)[\gamma]\\ \text{at } t}}}^{\substack{\text{the function obtained}\\ \text{when the differential}\\ \text{acts on } \gamma}} = M(x(t), y(t))\frac{dx}{dt} + N(x(t), y(t))\frac{dy}{dt}\ . \tag{4.1}$$

The notation on the left-hand side of (4.1) may look intimidating and unwieldy, but it (or something like it) is a necessary evil for this section of the notes.

[93] 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.

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 (4.5), 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

$$\int_{\mathcal{C}} M(x,y)\,dx + N(x,y)\,dy. \tag{4.2}$$

To see that the integrand in (4.2) is the same gadget we described above, let's review the rules you learned for computing such an integral:

1. Choose a regular parametrization γ of $\mathcal{C}$. Write this as $\gamma(t) = (x(t), y(t))$, $t \in [a,b]$.[94] 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 (4.2), make the following substitutions: $x = x(t), y = y(t), dx = \frac{dx}{dt}dt, dy = \frac{dy}{dt}dt$, and $\int_{\mathcal{C}} = \int_a^b$. The ordinary Calc-1 integral obtained from these substitutions is
$$\int_a^b \left\{ M(x(t),y(t))\frac{dx}{dt} + N(x(t),y(t))\frac{dy}{dt} \right\} dt. \tag{4.3}$$

3. Compute the integral (4.3). The definition of (4.2) is the value of (4.3):

$$\int_{\mathcal{C}} M(x,y)dx + N(x,y)dy = \int_a^b \left\{ M(x(t),y(t))\frac{dx}{dt} + N(x(t),y(t))\frac{dy}{dt} \right\} dt. \tag{4.4}$$

(You also learn in Calculus 3 that this definition is self-consistent: no matter what regular parametrization of $\mathcal{C}$ you choose[95], you get the same answer.)

A casual glance at (4.4) suggests that we have used the following misleading equality:

$$\text{“}M(x,y)dx + N(x,y)dy = \left\{ M(x(t),y(t))\frac{dx}{dt} + N(x(t),y(t))\frac{dy}{dt} \right\} dt.\text{”} \tag{4.5}$$

[94]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.

[95]Subject to the other conditions in the previous footnote.

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 (4.4), 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 (4.4) 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 $(M\,dx + N\,dy)[\gamma](t)$ in (4.1). The effect of the "$M(x, y)dx + N(x, y)dy$" on the left-hand side has been to produce the function $(M\,dx + N\,dy)[\gamma]$ 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.

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.

4.2 Exact equations: further exploration

Example 4.1 In the setting of Example 3.74, assume that $M\,dx + N\,dy$ has no singular points (equivalently, F has no critical points) in R. We claim that in this case, (one form of) the general solution of $M\,dx + N\,dy = 0$ on R is $\{F(x, y) = C\}$, but where the allowed values of C are those for which the graph of $F(x, y) = C$ 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 (3.146) 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 $\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}$ 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 $\frac{\partial F}{\partial x}(x_0, y_0), \frac{\partial F}{\partial y}(x_0, y_0)$ is not zero. Then:

- If $\frac{\partial F}{\partial y}(x_0, y_0) \neq 0$, then, since we are assuming that $\frac{\partial F}{\partial x}$ and $\frac{\partial F}{\partial y}$ are continuous on R, we can apply the Implicit Function Theorem (Theorem 5.13) 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 (3.144) 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 (3.146). 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 $\frac{\partial F}{\partial x}(x_0, y_0) \neq 0$, then (reversing the roles of x and y in the Theorem—e.g. by defining $\tilde{F}(x, y) = F(y, x)$), 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 (3.144) 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 (3.146), and again a small enough segment of it, passing through (x_0, y_0), will be contained in R. ■

Example 4.2 Consider again the DE

$$x\,dx + y\,dy = 0. \tag{4.6}$$

Defining $F(x,y) = \frac{1}{2}(x^2+y^2)$ (on the whole plane $\mathbf{R}^2$), the left-hand side of (4.6) 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 4.1 applies. The range of F on R is the set of positive real numbers, which for the sake of Definition 3.74, we view as $\{C \in \mathbf{R} \mid C > 0\}$. Therefore the general solution of $x\,dx + y\,dy = 0$ 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

$$\left\{x^2 + y^2 = C \mid C > 0\right\}. \tag{4.7}$$

The graph of each solution is a circle. The collection of these circles is what we call the general solution of (4.6) in R (according to Definition 3.74), 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 (4.6) on the whole xy plane rather than just R, then Example 4.1 no longer applies (because of the critical point at the origin), but Example 3.74 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 (4.6), with no restriction on the region, is the same as the general solution on R, namely (4.7). ■

Example 4.3 Consider again the DE from Example 3.69,

$$y\,dx + x\,dy = 0. \tag{4.8}$$

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 4.1 applies. This time, for every $C \in \mathbf{R}$ there is a point in R for which $xy = C$. Therefore the general solution of $y\,dx + x\,dy = 0$ in R is

$$xy = C, \tag{4.9}$$

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 (4.8) 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 3.69, the origin *is* on a solution curve; in fact it is on two of them: the x-axis and the y-axis. So the set of solution curves (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 (4.8), with no restriction on the region, is again (4.9). But in contrast to Example 4.2, 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.[96]) A level-set may have more than one *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 3.74 and 4.1 we can deduce the following:

$$\left.\begin{array}{l}\textbf{If } F \textbf{ has continuous second partial derivatives in the region}\\ R\textbf{, then the set of solution curves of } dF = 0 \textbf{ in } R \textbf{ is the set}\\ \textbf{of smooth curves in } R \textbf{ that are contained in level-sets of } F.\end{array}\right\} \quad (4.10)$$

Statement (4.10) 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

[96] *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 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.

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 3.75 elsewhere in these notes.

4.3 One-parameter families of equations

Suppose that G is a three-variable function with the property that

$$\begin{array}{l}\text{for each } z \in \mathbf{R}\text{, there is } \textit{some} \text{ point } (x, y) \in \mathbf{R}^2 \\ \text{for which } (x, y, z) \text{ is in the domain of } G.\end{array} \tag{4.11}$$

Then for each $C \in \mathbf{R}$, the equation $G(x, y, C) = 0$ is an algebraic equation in the variables x and y.[97] The collection of equations

$$\{G(x, y, C) = 0 \mid C \in \mathbf{R}\} \tag{4.12}$$

is an example of a *one-parameter family of (algebraic) equations* in variables x and y.[98] (The word "algebraic" is understood, even if omitted.) The third variable of C, the *parameter*, is a constant in each equation $G(x, y, C) = 0$, but as we vary C we get different equations in x and y. With these notes' convention that when the letter C appears in an equation, it is playing the role of a parameter in the sense above, we may also write (4.12) simply as "$\{G(x, y, C) = 0\}$".

More generally, a one-parameter family of (algebraic) equations in variables x and y is a collection of the form

$$\{G_1(x, y, C) = G_2(x, y, C) \mid C \in \mathbf{R}\} \tag{4.13}$$

where G_1 and G_2 are three-variable functions having the property stated above for G. (We may also write (4.13) simply as $\{G_1(x, y, C) = G_2(x, y, C)\}$.) Of course, "$G_1(x, y, C) = G_2(x, y, C)$" is equivalent to "$G_1(x, y, C) - G_2(x, y, C) = 0$", an equation of the form in (4.12), so all statements we might want to make a collection of the form (4.13) can be deduced from statements about collections of the form (4.12).

For a given function G having the domain-property above, there may be values of C for which the equation $G(x, y, C) = 0$ has no solutions. For example, observe that if $C > 0$, the equation $x^2 + y^2 + C = 0$ is not satisfied by any point $(x, y) \in \mathbf{R}^2$. Thus if G is the function defined by $G(x, y, z) = x^2 + y^2 + z$, then even though domain of G is all of $\mathbf{R}^3$, for $C > 0$ the equation $G(x, y, C) = 0$ has no solutions. When talking

[97] For an example of a function G that does not have property, (4.11) consider $G(x, y, z) = x + y + \ln z$. For $C < 0$, $\ln(C)$ is not even defined, so the equation $G(x, y, C) = 0$ makes no sense.

[98] This G should not be confused with the G in equation (3.3), which is being used to describe a *single, differential* equation, not a collection of algebraic equations.

about one-parameter families of *equations*, we do not exclude values of C for which the equation $G(x, y, C) = 0$ makes sense but simply has no solutions.

Key here is the word "equations" in "one-parameter family of *equations*". Where we can get into trouble is when we use similar-sounding terminology, "one-parameter family of *solutions* of a DE", or "one-parameter family of *implicit solutions* of a DE". When solving a DE we frequently write down a one-parameter family of equations that is intended to represent a collection of solutions, or implicit solutions, of the DE—perhaps even the whole general solution. However, it is very easy to get carried away by the ease of writing down such a family of equations[99], especially when solving derivative-form DEs, fall into the trap of forgetting what "a solution of a DE" *means*", and muddle some conceptually important distinctions. Before giving examples, let us make one more definition to reinforce the meaning of "a solution of derivative-form DE".

Definition 4.4 For a given derivative-form DE

$$\mathcal{H}(x, y, \frac{dy}{dx}) = 0 \tag{4.14}$$

a *one-parameter family of solutions* is a collection of pairs $\{(I_C, \phi_C)\}$ where, for each $C \in \mathbf{R}$, the set I_C is an interval and the function ϕ_C is a solution of (4.14) on I_C. (Here, keep in mind the distinction between the domain of a *formula*—the "implied domain" in the terminology of precalculus and Calculus 1—and the domain of a *function*. The interval I_C may not be the whole domain of a formula that is given for ϕ_C.) An *explicit form* of a one-parameter family of solutions of (4.14) is a collection of restricted equations

$$\{y = \phi_C(x) \mid x \in I_C\} \tag{4.15}$$

where $\{(I_C, \phi_C)\}$ is a one-parameter family of solutions of (4.14). (The "$x \in I_C$" may be written in other formats, such as "$a < x < b$", "$x > a$", etc.) **If $\phi_C(x)$ is presented by a formula whose domain, for every C, is an interval,** then we may omit the "$x \in I_C$" in (4.15) and simply write

$$\{y = \phi_C(x)\} \tag{4.16}$$

in place of (4.15).

In the spirit of our convention for individual solutions, we allow ourselves to call (4.15) (and, when applicable, (4.16)) simply a one-parameter family of solutions (omitting the words "an explicit form of").

A one-parameter family of equations in x and y (e.g. a family of the form $\{G(x, y, C) = 0\}$) is a *one-parameter family of implicit solutions* of the DE (4.14) if

[99] "Ease", when the DE falls into one of the categories for which systematic methods of solution are taught in an introductory course in DEs.

each equation in the family is an implicit solution of the DE.[100] ■

Example 4.5 Consider the differential equation

$$\frac{dy}{dx} = y^2 \tag{4.17}$$

and the one-parameter family of equations

$$\mathcal{E} = \left\{ y = -\frac{1}{x - C} \right\} \tag{4.18}$$

(equivalently, $\left\{ y + \frac{1}{x-C} = 0 \right\}$.) The equation $y = -\frac{1}{x-C}$ represents *two* solutions of (4.17), one on the interval $(-\infty, C)$ and one on the interval (C, ∞). Thus, $\mathcal{E}$ is *not* a one-parameter family of solutions of the DE (4.17).

However, for each C the equation $y = -\frac{1}{x-C}$ *does* meet our definition of "*implicit* solution" of (4.17), and the collection (4.18) is a one-parameter family of *implicit* solutions of this DE.[101] ■

[100] *Note to instructors*: It may strike you that my definition of one-parameter family is too restrictive, even for "one-parameter family of implicit solutions", especially if you are used to allowing "$C = \infty$" in order to include in a family some solution(s) that would otherwise be excluded. For professional mathematicians, a reasonable definition of "one-parameter family" of some type of object is a parametrized set of those objects, where the parameter space is a *connected* 1-dimensional topological space (with a definition of "dimension" appropriate to that type of topological space) and whose topology is related canonically to the set of objects being parametrized. The interval $[-\infty, \infty]$ and the circle $[-\infty, \infty]/(-\infty \sim \infty)$ satisfy at least the first part this definition. However, whether or not it satisfies the second, it does not truly parametrize a set of *equations*, in two real variables, that we might write down as solutions of an ODE. As a tool to inspire curiosity, it can be valuable to show students at this level that, for example, the solution $y = 0$ of $dy/dx = y^2$ can be viewed as a $C \to \infty$ limit, in a sense that need not be made precise, of the solutions $y = -1/(x - C)$, or that the equilibrium solutions of the logistic equation can be obtained similarly as a limit of non-constant solutions. However, while it's good to show them this thought-provoking phenomenon once or twice, I think it is a mistake to encourage the use of infinite parameter-values in any generality, simply to allow some expression for the set of all solutions to appear to capture an otherwise outlying solution. I don't want to encourage students in an intro ODE course to use the extended reals; they already have too much of a propensity to treat infinity as a real number. Furthermore, in the setting of one-parameter families of solutions, it is likely to cause them to not to realize that, say, in the $dy/dx = y^2$ example above, $\{y = -1/(x - C)\}$ is *not* a one-parameter family of solutions; it is the union of *two* one-parameter families of solutions. Thus, in my definitions of one-parameter families in these notes, I am allowing only real parameters.

[101] *Note to instructors*: It may seem odd that I am using this terminology here, when I argued in Remark 3.67 against using the word "implicit" to describe an explicit equation. But there are two important differences here: (1) There *is* something implicit when we refer to the equation $y = -\frac{1}{x-C}$ as an implicit solution of a DE, namely that this equation represents *two* true solutions, one for $x > C$ and another for $x < C$. (2) In Remark 3.67 we were talking, simultaneously, about *all* explicit algebraic equations in x and y. In such a discussion there is no reasonable way to exclude

Example 4.6 Let F be a function of two variables defined on some region R. Then the collection of equations

$$\mathcal{E} = \{F(x, y) = C\}$$

is a one-parameter family of equations. If F is continuously differentiable on R, then this one-parameter family *contains* an algebraic form of the general solution of the exact differential equation $dF = 0$, namely the sub-collection $\mathcal{E}_1$ in which the values of C are restricted to those for which the graph of $F(x, y) = C$ contains a smooth curve in R (see Example 3.74). It is not *terrible* to say that $\{F(x, y) = C\}$ is a one-parameter family of solutions of $dF = 0$, but in doing so we must keep in mind that we are using the term "one-parameter family" more loosely than in "one-parameter family of *equations*": there may be some values of C for which the equation $F(x, y) = C$ is *not* a solution of $dF = 0$.

Of course, for many functions F, the equation $F(x, y) = C$ *is* a solution of $dF = 0$ for every value of C. In this case, the one-parameter family of *equations* $\{F(x, y) = C\}$ *is* (one algebraic form of) the general solution of $dF = 0$, and we may say that this collection is a one-parameter family of *solutions* of $dF = 0$ without any alteration in the meaning of "one-parameter family". ■

Example 4.7 Consider a separable DE

$$\frac{dy}{dx} = g(x)p(y) \tag{4.19}$$

for which the functions g and p satisfy the conditions (3.102). For simplicity, let us assume that both the interval I and set D in (3.102) are the whole real line, and assume that there is at least one $r \in \mathbf{R}$ for which $p(r) = 0$. We have seen that one implicit form of the general solution of (4.19) is the collection of equations $\mathcal{E} = \mathcal{E}_1 \bigcup \mathcal{E}_2$ defined in (3.103)–(3.104). The collection $\mathcal{E}_1$ is a one-parameter family of equations; the collection $\mathcal{E}$ is not.

But, in the notation of Theorem 3.44, consider the collection of equations

$$\mathcal{E}' = \{p(y)(H(y) - G(x) - C) = 0\}.$$

This *is* a one-parameter family of equations. Furthermore, the graph of $p(y) = 0$ is simply the union of the graphs of all the equations in $\mathcal{E}_2$, none of which intersects the graph of any of the equations $H(y) - G(x) - C = 0$ (as shown in Theorem 3.44, since $H(y) - G(x) - C = 0$ is equivalent to $H(y) = G(x) + C$). Thus, for each C, the

equations that happen to express y explicitly in terms of x. If we are going to define the terminology "implicit solution", there is no reasonable way to exclude "explicit solutions" from meeting the definition, or from excluding explicit equations like $y = 1/(x - C)$ that represent more than one solution.

graph of $p(y) = 0$ does not intersect the graph of $H(y) - G(x) - C = 0$. Therefore every smooth curve lying in the graph of $p(y)(H(y) - G(x) - C) = 0$ either lies entirely in the graph of $p(y) = 0$, or lies entirely in the graph of $H(y) = G(x) + C$, and every such curve is the graph of a solution of the DE (4.19). Furthermore, each equation $H(y) - G(x) - C = 0$ determines at least one solution of (4.19), hence so does each equation $p(y)(H(y) - G(x) - C) = 0$. Thus, for each C, the equation $p(y)(H(y) - G(x) - C) = 0$ is an implicit solution of this DE.

Hence $\mathcal{E}'$ is a one-parameter family of implicit solutions of (4.19) that determines all solutions of this DE, and determines no differentiable function that isn't a solution of this DE.

However, the graph of each constant solution of (4.19) (of which there is at least one, since we assumed that p was zero *somewhere*) is contained in the graph of *every* equation in $\mathcal{E}'$, not only one equation in $\mathcal{E}'$. Thus $\mathcal{E}'$ is not what we are calling *an implicit form of the general solution of* (4.19), as defined by Definition 3.34. ■

Example 4.7 illustrates that while it may be *possible* to express the general solution of a DE as a one-parameter family of implicit solutions, it may not be *desirable* to do so. More generally than Example 4.7, if there is *any* one-parameter family of implicit solutions $\{P(x, y, C) = 0\}$ of some derivative-form DE, we can brutally force any other solution to lie in a new one-parameter family: if ϕ is *any* solution of (4.7), the collection of equations $\{(y - \phi(x))P(x, y, C) = 0\}$ is a new one-parameter family of implicit solutions that determines all the solutions determined by the original family and also determines the solution ϕ. Similarly, if there is an implicit solution $F(x, y) = 0$ not in the original family, the collection $\{F(x, y)P(x, y, C) = 0\}$ is a new one-parameter family of implicit solutions that determines all the solutions determined by the original family as well as those determined by $F(x, y) = 0$.

5 Appendix

5.1 Intervals in R

An *interval* is a non-empty subset I of $\mathbf{R}$ with the "betweenness property": given any two distinct elements c, d of I, every real number between c and d lies in I.

Every interval is of exactly one of the following forms:

$$[a, a], \qquad \text{where } a \in \mathbf{R}; \tag{5.1}$$

$$(a, b),\ (a, b],\ [a, b),\ \text{or } [a, b], \qquad \text{where } a, b \in \mathbf{R} \text{ and } a < b; \tag{5.2}$$

$$(-\infty, c),\ (-\infty, c],\ (c, \infty),\ \text{or } [c, \infty), \qquad \text{where } c \in \mathbf{R}; \tag{5.3}$$

$$\text{or} \quad (-\infty, \infty). \tag{5.4}$$

Intervals may have two endpoints, one endpoint, or no endpoints. The intervals of the form (5.2) have two endpoints; the intervals of the form (5.1) and (5.3) have one endpoint, and the interval (5.4) (the whole real line) has no endpoints.

An interval is called *open* if it does not contain any of its endpoints, and *closed* if it contains all its endpoints. Thus, intervals of the form $(a, b), (-\infty, c), (c, \infty)$, and $(-\infty, \infty)$ are open, while intervals of the form $[a, a], (-\infty, c], [c, \infty)$, and $(-\infty, \infty)$ are closed. (Hence the interval $(-\infty, \infty)$ is both open and closed.) Intervals of the form $(a, b]$ and $[a, b)$ are neither open nor closed; these are sometimes called "half-open", "half-closed", or both.

The intervals of the forms (5.1) and (5.2) are called *bounded*; the others are called *unbounded*. Among the unbounded intervals, the ones on line (5.3) are called *semi-bounded*. Intervals of the form (5.1) are called *singletons* (or *singleton sets*, or *singleton intervals*) and are said to have *zero length*; all other intervals are said to have *positive length*. In particular, all open intervals have positive length.

Remark 5.1 (A common mistake) If a function has a property that holds only at a single point x_0, it is not technically correct to say, as at least one DE textbooks does, that "the function does not have this property on an interval," since $[x_0, x_0]$ is an interval. It *is* correct to say, in this case, that "the function does not have this property on a *positive-length interval*," or "the function does not have this property on an *open* interval." However, authors and instructors can avoid this issue if they say, early enough, something to the effect of "In this book (or class), whenever we use the word *interval*, we mean *positive-length* interval." ■

5.1.1 DEs on non-open positive-length intervals

The concept of *differentiability of a function ϕ at a point x_0*, as defined in most Calculus 1 courses, requires ϕ to be defined on some open interval containing x_0. Thus, if the domain of ϕ is an interval containing x_0 as an endpoint, this definition of differentiability does not allow us to say either that "ϕ is differentiable at x_0" or "ϕ is *not* differentiable at x_0," let alone to define a number "$\phi'(x_0)$."

A generalized definition is made to fill this gap. If ϕ is a positive-length interval containing a point x_0, then:

- If the domain of ϕ includes an interval $[x_0, x_0 + \delta)$ for some $\delta > 0$, we define the *right-hand derivative of ϕ at x_0* to be the one-sided limit

$$\lim_{x \to x_0+} \frac{\phi(x) - \phi(x_0)}{x - x_0}, \tag{5.5}$$

  provided this limit exists. One notation used for this limit is $\phi'(x_0+)$.

- If the domain of ϕ includes an interval $(x_0 - \delta, x_0]$ for some $\delta > 0$, we define the *right-hand derivative of ϕ at x_0* to be the one-sided limit

$$\lim_{x \to x_0-} \frac{\phi(x) - \phi(x_0)}{x - x_0}, \tag{5.6}$$

  provided this limit exists. One notation used for this limit is $\phi'(x_0-)$.

The limits $\phi'(x_0+)$ and $\phi'(x_0-)$ are called the *one-sided* derivatives of ϕ at x_0.

If x_0 is an *interior* point of the domain of f (i.e. if the domain contains an interval $(x_0 - \delta, x_0 + \delta)$ for some $\delta > 0$), it is not hard to see that ϕ is differentiable at x_0 if and only if both one-sided derivatives of ϕ at x_0 exist and are equal.

If the domain of ϕ is a positive-length non-open interval I that contains x_0 as an *endpoint*, then we *define* ϕ to be differentiable at x_0 if the corresponding one-sided derivative exists. In such a case, when substituting "$y = \phi(x)$" into a differential equation $\mathsf{G}(x, y, \frac{dy}{dx}) = 0$, we intepret $\frac{dy}{dx}(x_0)$ as the corresponding one-sided derivative $\phi'(x_0\pm)$.

5.2 Open rectangles and open sets in $\mathbf{R}^2$

Definition 5.2 An *open rectangle* is a subset of $\mathbf{R}^2$ of the form

$$I \times J := \{(x, y) \in \mathbf{R}^2 : x \in I \text{ and } y \in J\}, \tag{5.7}$$

where I and J are open intervals (*Note*: the notation ":=" means that we are *defining* the notation on the left to mean the object to the right of the equals-sign. Thus, the sentence above defines both the term "open rectangle" and the notation "$I \times J$", the latter of which is read "I cross J.") ■

The notation "$I \times J$" is defined by equation (5.7) for *any two sets* I and J, not just for intervals (open or otherwise). For *closed, bounded* intervals $I = [a, b]$ and $J = [c, d]$, where $a < b$ and $c < d$, the set $I \times J$ is a Cartesian-coordinate representation of what we would have called a rectangle in high school geometry, with sides parallel to the coordinate axes:

$$[a, b] \times [c, d] = \{(x, y) \in \mathbf{R}^2 : a \le x \le b \text{ and } c \le y \le d\}. \tag{5.8}$$

Analogously to the definition of "closed interval", and in the spirit of Definition 5.2, we call $[a, b] \times [c, d]$ a *closed rectangle*; it contains all the points on its boundary (the four sides of the rectangle). For the open rectangle $(a, b) \times (c, d)$, all the "$\le$" signs in equation (5.8) are replaced by strict "$<$" signs. Thus, this open rectangle is the set we get by *removing* all the boundary points of the closed rectangle.

Note that if either of the open intervals I, J is *unbounded* (see Section 5.1), then the set $I \times J$ does not *look* rectangular. For example, if $I = (a, b)$ and $J = (-\infty, \infty) = \mathbf{R}$, then $I \times J$ is an infinite vertical strip, the region strictly between the vertical lines $x = a$ and $x = b$.

"Open set in $\mathbf{R}^2$" generalizes the notion of open *rectangles*:

Definition 5.3 A set $R \in \mathbf{R}^2$ is an *open set* if for every point $(x_0, y_0) \in R$, there is *some* open rectangle (no matter how small) that contains (x_0, y_0) and is contained entirely in R.[102]

In these notes, we often call an open set in $\mathbf{R}^2$ an open *region* in $\mathbf{R}^2$. [103].

Example 5.4 The sets $R_1 = \{(x, y) \in \mathbf{R} : y > x\}$ and $R_2 = \{(x, y) \in \mathbf{R} : y < x\}$ are open regions in $\mathbf{R}^2$. To see this, observe that R_1 is the set of points in $\mathbf{R}^2$ lying *above* the line $y = x$, while B is the set of points lying *below* this line L. Given $(x_0, y_0) \in A$, let $\delta = (y_0 - x_0)/2$; note that $\delta > 0$. Then the open rectangle $S_{(x_0,y_0)} = (x_0-\delta, x_0+\delta)\times(y_0-\delta, y_0+\delta)$ lies in A. (You should be able to convince yourself of this easily with a picture. The point of L closest point to (x_0, y_0) is $(\frac{x_0+y_0}{2}\ \frac{x_0+y_0}{2})$, which is exactly the lower right corner of the square $S_{(x_0,y_0)}$, since $x_0 + \delta = \frac{x_0+y_0}{2} = y_0 - \delta$. Thus $S_{(x_0,y_0)}$ is an open rectangle that contains (x_0, y_0) and is contained in A. Since this holds for *any* $(x_0, y_0) \in R_1$, the set R_1 is open. The argument that R_3 is open is similar. ■

In Example 5.4, we showed that A is open by explicitly identifying, for every $(x_0, y_0) \in A$, a rectangle $S_{(x_0,y_0)}$ that contains (x_0, y_0) and is contained in A. The following powerful generalization of Example 5.4 assures us of the openness of all regions R of a certain, very common, type, without ever having to explicitly identify a rectangle $S_{(x_0,y_0)}$ that "works" for a given point $(x_0, y_0) \in R$.

Theorem 5.5 *Let f be a function of two variables, and let $c \in \mathbf{R}$. Assume that f is continuous on the whole plane $\mathbf{R}^2$. Then $\{(x, y) \in \mathbf{R}^2 : f(x, y) > c\}$ and $\{(x, y) \in \mathbf{R}^2 : f(x, y) < c\}$ are open sets in $\mathbf{R}^2$.* ■

[102] You may have seen a different definition of "open set in $\mathbf{R}^2$" in which "open rectangle that contains (x_0, y_0)" is replaced by "open *disk* centered at (x_0, y_0)". The latter, more standard, definition of "open set in $\mathbf{R}^2$" is equivalent to Definition 5.3.

[103] *Note to instructors*: I am taking some liberties here. Although "region" has no universal definition in mathematics, most definitions require at least that a region be *connected* and *non-empty*. 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, the set is assumed to be non-empty.

(We do not prove Theorem 5.5 in these notes. If you take a course in topology or advanced calculus course, you should see it proven there.)

To see that the openness of R_1 and R_2 in Example 5.4 can be obtained directly from Theorem 5.5, take $f(x, y) = y - x$ and $c = 0$.

5.3 Review of the Fundamental Theorem of Calculus

"The" Fundamental Theorem of Calculus (FTC) really consists of two related theorems, either of which can be used to derive the other. Roughly speaking, one involves the derivative of an integral, while the other involves the derivative of an integral. In many textbooks, these two parts are presented as two separate theorems, and are called the first and second *form*, or *version*, of the FTC, but there is no consistency among authors as to which part is called "first" and which is called "second". In the statement of the FTC below, I could have put the parts in either order.

Theorem 5.6 (Fundamental Theorem of Calculus) *Let I be a positive-length interval, and let f be a continuous function on I.*

(a) Let x_0 be a point in I. Define a function F on I by

$$F(x) = \int_{x_0}^{x} f(t)\,dt \quad (\text{for each } x \text{ in } I). \tag{5.9}$$

Then F is differentiable on I, and $F' = f$ (i.e. $F'(x) = f(x)$ for each x in I). In other words, F is an anti*derivative of f on I.*

(b) Let G be any antiderivative of f on I. Then for any points a, b in I,

$$\int_{a}^{b} f(x)\,dx = G(b) - G(a). \tag{5.10}$$

■

To appreciate what the FTC is saying, keep in mind that *definite* integrals are defined as *limits of Riemann sums*; their definition does not involve derivatives at all. The fact that definite integrals of continuous functions *exist* (i.e. that the relevant limits exist) and have various properties you learned in Calculus 1, is something proven in *advanced* calculus. Part (b) of Theorem 5.6 tells us that antiderivatives can be used to *compute* definite integrals (when we're fortunate enough to know an explicit formula for some antiderivative of the function we're integrating).

Note that part (a) of Theorem 5.6 can be stated without ever introducing a letter for the function F, by simply writing the conclusion as

$$\frac{d}{dx}\left(\int_{x_0}^{x} f(t)\,dt\right) = f(x) \quad \text{(for each } x \text{ in } I\text{)}. \tag{5.11}$$

(Thus part (a) is the "derivative of an integral" form of the FTC.)

Similarly, part (b) of Theorem 5.6 can be stated without ever introducing a letter for the function f, as follows:

Let G be a continuously differentiable function on I (i.e. a differentiable function on I whose derivative is continuous). Then for any points a, b in I,

$$\int_a^b G'(x)\,dx = G(b) - G(a). \tag{5.12}$$

(Thus part (b) is the "integral of a derivative" form of the FTC.)

Note also that part (a) of 5.6 has the following important corollary:

Corollary 5.7 *Every continuous function on a positive-length interval I has an antiderivative on I.*

5.4 The "Fundamental Theorem of Ordinary Differential Equations"

The "Fundamental Theorem of ODEs" ("FTODE"), also known by names such as the Existence and Uniqueness Theorem (for IVPs, or for ODEs), is the theorem asserting that, under certain rather general conditions, an initial-value problem has a unique solution (with "uniqueness" appropriately defined). The first-order case is the theorem below.[104]

Theorem 5.8 (FTODE) *Let f be a function of two variables, let (x_0, y_0) be a point in $\mathbf{R}^2$, and consider the initial-value problem*

$$\frac{dy}{dx} = f(x, y), \quad y(x_0) = y_0\,. \tag{5.13}$$

[104] *Note to instructors*: There are several stronger versions of this theorem; see [2, Section 6.1]. In one version, the hypothesis that $\partial f/\partial y$ is continuous is relaxed to "f is locally uniformly Lipschitz in its second variable" without altering the conclusion. In a different strengthening, *more* differentiability of f is assumed, and one shows that the solution of (5.8) depends differentiably on parameters. (Parameters include the initial-condition point (x_0, y_0) and any extra parameters on which function f may explicitly depend.) However, no theorem with a *weaker conclusion* than Theorem 5.8's has any use whatsoever in understanding uniqueness of solutions of IVP's—see Remark 5.12. Nothing so useless should ever be presented to DE students as an important theorem, and presenting a theorem that is of no consequence (and giving exercises on it, to boot!) is a waste of class time.

Suppose that f and $\frac{\partial f}{\partial y}$ are continuous on a given open set R (see Section 5.2) *containing the point (x_0, y_0). Then there exists a number $\delta > 0$ such that for every (not necessarily open) subinterval I_1 of $(x_0 - \delta, x_0 + \delta)$ containing x_0, the initial-value problem* (5.13) *has a unique solution-in-R on I_1.*

(We do not prove Theorem 5.8 in these notes.)

Remark 5.9 The conclusion of Theorem 5.8 can be restated qualitatively as: *On* **every** *sufficiently small interval containing x_0, the IVP* (5.13) *has a unique solution.* This is extremely important; see Remark 5.12. ■

Even without any hypotheses on f, if ϕ is solution of the IVP (5.13) on an interval I containing x_0, then the *restriction* of ϕ to any subinterval I_1 containing x_0 is still a solution of (5.13). But if the hypotheses of Theorem 5.8 are met, and δ is as in the theorem, and I_1 is any subinterval of $I = (x_0 - \delta, x_0 + \delta)$ containing x_0, then the theorem tells us that (5.13) has *exactly one* solution on I_1. This immediately yields the following important corollary of Theorem 5.8:

Corollary 5.10 *Let f, R, x_0, y_0, and δ be as in Theorem 5.8. If I_1 is* **any** *subinterval of $(x_0-\delta, x_0+\delta)$ containing x_0, and ϕ is the unique solution of the initial-value problem* (5.13) *on $(x_0 - \delta, x_0 + \delta)$, then $\phi|_{I_1}$* (the restriction of ϕ to I_1) *is the unique solution of* (5.13) *on I_1.* ■

Another important corollary of Theorem 5.8 is the following (which was stated earlier as Corollary 3.19). The three parts are very closely related; essentially they are the same result stated three ways.

Corollary 5.11 *Let f and R be as in Theorem 5.8. Then:*

(a) For every (x_0, y_0) in R, the initial-value problem (5.13) *has a unique solution that is maximal in R. This maximal-in-R solution $\phi_{\max}$ has the property that every solution of* (5.13) *in R is a restriction of $\phi_{\max}$.*

(b) Every point (x_0, y_0) in R lies on a unique maximal solution curve in R (i.e. the graph of a unique solution that is maximal in R).[105]

(c) No two distinct maximal solution curves in R can intersect.

[105] *Note to instructors*: In differential-geometric terminology, the maximal solution curves *foliate* R.

(In parts (b) and (c), "solution curve" means "solution curve of the differential equation $\frac{dy}{dx} = f(x, y)$.") ■

(We do not include a proof of Corollary 5.11 in these notes at this time, but may add one later.)

The essence of Corollary 5.11 is that, under the given hypotheses, solutions of $\frac{dy}{dx} = f(x, y)$ in R cannot "bifurcate": If ϕ is a non-maximal solution of the initial-value problem (5.13) on an open interval I_1, then there is only one way to extend ϕ to a solution on slightly larger open interval. (There can't be a different solution that "peels off".) More precisely: if ϕ is a solution of (5.13) on an open interval I_1, and can be extended to a solution $\tilde{\phi}$ on a larger open interval I (with I small enough for the graph of $\tilde{\phi}$ to remain in the region R), then then $\tilde{\phi}$ is the *only* solution of (5.13) on I. Another way of stating this uniqueness is that if ϕ_1 and ϕ_2 are two solutions of the IVP (5.13) in R, with ϕ_1 having domain-interval I_1 and with ϕ_2 having domain-interval I_2, then ϕ_1 and ϕ_2 are identically equal on the intersection of I_1 and I_2.[106] (Note that the intersection of any two intervals containing x_0 is another interval containing x_0.)

Remark 5.12 Most DE textbooks (including [1], [3], and [4]) state a version of Theorem 5.8 that is *essentially* **useless** *as far as uniqueness is concerned*. In this version, R is taken to be an open rectangle $I \times J$, and the last sentence of Theorem 5.8 is replaced with

$$\begin{array}{l}\text{"Then there exists a number } \delta > 0 \text{ such that the initial-value} \\ \text{problem (5.13) has a unique solution on } (x_0 - \delta, x_0 + \delta)\text{"}\end{array} \tag{5.14}$$

(or an equivalent sentence).

There are two differences, one major and one minor, between Theorem 5.8 and this weaker theorem. The minor difference is the use of "open rectangle" in the weaker theorem vs. "open set" in Theorem 5.8. Presumably, the reason that most textbooks state only an "open rectangle" version is to avoid burdening students with the definition of "open set". I do not believe that this definition imposes much of a burden, and the benefits of being able to use "open set" are significant.

The major difference between the weaker theorem and Theorem 5.8 is statement (5.14). Statement (5.14) says only that there exists **one** open interval, centered at x_0,

[106] *Note to instructors*: This fact is of critical importance to showing that, under the hypotheses of Theorem 5.8 there is a "maximal domain of uniqueness" for a solution of an IVP, which is essentially what Corollary 5.11 states three different ways. But this critical fact *cannot* be deduced from the versions of Theorem 5.8 that assert only the weak conclusion (5.14). The only textbook I've looked at recently that states an existence/uniqueness theorem as useful as Theorem 5.8 is [2].

possibly very large, on which the IVP (5.13) has a unique solution. *This has no useful uniqueness-implications whatsoever*; it is barely any stronger than simply the *existence* of a solution.[107] But the latter conclusion can be proven with only the assumption that f itself is continuous; $\partial f/\partial y$ need not even *exist*.[108] **The whole point of Theorem 5.8, stated qualitatively, is Remark 5.9.** Under the hypotheses of Theorem 5.8, and with δ as in the theorem, we can never *gain* solutions of (5.13), and thereby break uniqueness, by *shrinking* the interval to any neighborhood of x_0 *smaller* than $(x_0 - \delta, x_0 + \delta)$. However, we *can* potentially gain solutions, and break uniqueness, by *extending* this interval to a *larger* one).

The weak conclusion (5.14) does not rule out the possibility that the IVP (5.13) has a unique solution on $(x_0 - \delta, x_0 + \delta)$, but has more than one solution on a smaller interval, e.g. $(x_0 - \frac{\delta}{2}, x_0 + \frac{\delta}{2})$ or $(x_0 - \frac{\delta}{2}, x_0 + \delta)$ or $[x_0, x_0 + \delta)$. (Even replacing (5.14) by "Then there exist *arbitrarily small* numbers $\delta > 0$ such that the IVP (5.13) has a unique solution on $(x_0 - \delta, x_0 + \delta)$" would not solve this problem.) This phenomenon is ruled out by the more-carefully stated Theorem 5.8.

Textbooks that state the weaker theorem (the one with (5.14) as its conclusion) tend to *use* Theorem 5.8 without ever stating it, as if it were implied by the weaker version (**which it is not!**).

5.5 The Implicit Function Theorem

Theorem 5.13 (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 $\frac{\partial F}{\partial y}(x_0, y_0) \neq 0$, where $\frac{\partial F}{\partial y}$ 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 from I_1 to J_1 (i.e. a function defined on I_1 and whose range is contained in J_1), such that

$$\begin{aligned} &\text{for every point } (x, y) \in I_1 \times J_1, \\ &\qquad F(x, y) = c_0 \text{ if and only if } y = \phi(x). \end{aligned} \tag{5.15}$$

■

[107] *Note to instructors*: Personally, I do not see **any reason whatsoever** to teach this version of the theorem; it's a waste of time. It is simultaneously useless and uninteresting. Even worse, it has *none* of the consequences that some textbooks, e.g. [3], say that it has. Thus, if you're teaching from one of those textbooks, and you don't go out of your way to correct these misstatements, you're (implicitly or explicitly) teaching your students something that's **false**. I would therefore exhort any instructor either to teach Theorem 5.8 (just the statement, not the proof) or *not to state an existence/uniqueness theorem at all*.

[108] *Note to instructors*: Nor do we need to assume that f is locally uniformly Lipschitz in its second variable.

Note: In Theorem 5.13, if we replace "Let $c_0 = F(x_0, y_0)$" by "Assume that $F(x_0, y_0) = 0$," and replace the c_0 in statement (5.15) by 0, the theorem we obtain is equivalent to Theorem 5.13. In fact, the Implicit Function Theorem is usually stated that way (with 0 rather than c_0). We have stated it with c_0 only to make the theorem more convenient to use, as stated, in our discussion of *exact* DEs.

In the setting of Theorem 5.13, we may think of the point (x_0, y_0) as a defining a function ϕ whose domain is the single number x_0, and for which we set $\phi(x_0)$ equal to y_0. (This the graph of $y = \phi(x)$ is the single point (x_0, y_0).) Essentially, the questions that the Implicit Function Theorem is aimed at addressing are:

1. Can we extend the graph-is-a-single-point function ϕ above to a function defined on at least a *small* open interval I_1 containing x_0, such that as x varies over the interval I_1, the graph of $y = \phi(x)$ (a) is still contained in the graph of $F(x, y) = c_0$ (i.e., such that $F(x, \phi(x)) = c_0$ for every $x \in I_1$)?

2. If so, is such an extension unique (at least if we take I' small enough)?

   Note that the question "Is there a unique such extension?" is simply another way of asking the following: Does the equation "$F(x, y) = c_0$" *define* (or *determine*) y as a function of x, at least in a small enough "window" $R = I_1 \times J_1$ around (x_0, y_0)? To understand precisely what the latter question is asking, recall that to define a (particular) function on a domain I_1, all we need do is to specify, for each $x \in I_1$, a *criterion* (or *rule*)—not necessarily a *formula*—for selecting one and only number y (which we may then reasonably denote "$y(x)$").

3. When there *is* a unique extension, does it have "nice properties" such as continuity, differentiability, continuous differentiability (i.e. having a continuous derivative), etc.? (With the continuity condition, we're asking whether the point (x_0, y_0) can be "extended" to at least a little *curve* that's (i) the graph of a function of x and (ii) is still contained in the graph of $F(x, y) = c_0$. With just the continuous differentiability condition, we're asking whether this can be done with a *smooth* curve.)

4. There are definitely functions F, with points (x_0, y_0) in the domain of F for which the answers is *no* to at least one of the questions above. (See Figure 1 or Figure 2, for example.) What are some conditions on F and (x_0, y_0) that are sufficient to guarantee that all the answers are yes?

Theorem 5.13 says, that, under its hypotheses: if we take I_1 and J_1 small enough, then for any $x \in I_1$, the criterion "y lies in J_1 and satisfies $F(x, y) = c_0$," does indeed single out one and only one number y. In other words, the small-enough-window condition, combined with the condition "$F(x, y) = c_0$," singles out exactly one $y \in J_1$

for each $x \in I_1$. "One y for each x in some domain" is *exactly* what a function is (with these names for output and input variables). The rule for selecting y from x is *unambiguous*. In this sense, we have have "solved for y in terms of x" (in the rectangle $I_1 \times J_1$), even though we have not given any algebraic formula telling us *explicitly* how to *calculate* y from x. (That's why we say such a function $y(x)$ is *implicitly* defined by the equation $F(x, y) = c_0$.)

As will be discussed below, Theorem 5.13 is a **very strong theorem**. Its full strength, which includes a *uniqueness* implication for the function ϕ, depends on very careful wording of the conclusion. Of course, as with any theorem, there are other ways this conclusion can be worded without changing what it's telling us. However, almost any attempt to *simplify* or *shorten* the wording leads to a much weaker theorem. Unfortunately, this is exactly what has happened in many textbooks below the level of Advanced Calculus (especially differential equations textbooks). In the other direction, there is a *stronger* version of Theorem 5.13 in which all the partial derivatives of F up through order n are assumed continuous on a rectangle R (where n can be any positive integer), and which conclude that ϕ is n-times continuously differentiable. It is valid to call this *strengthening* of Theorem 5.13 the Implicit Function Theorem. But **no theorem whose conclusion is weaker than that of Theorem 5.13 is the Implicit Function Theorem.**

To see how the theorem addresses these questions, let's examine some implications of the conclusion of Theorem 5.13 that you don't see stated explicitly in the theorem. First, in Theorem 5.13, since x_0 lies in I_1, we may look at what statement (5.15) 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 $y_0 = \phi(x_0)$.[109] Thus, the graph of the function ϕ in statement (5.15) always contain the point (x_0, y_0), no matter how large or small the intervals I_1 and J_1 are.

Further examining the conclusion of Theorem 5.13, statement (5.15) 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, (5.15) 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, e.g. 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

[109] The theorem called the "Implicit Function Theorem" in at least one DE textbook does not imply even this much.

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.[110]

The following corollary of the Implicit Function Theorem allows us to dispense with possibly having to shrink the interval J to a subinterval J_1, at the expense of weakening the uniqueness property to uniqueness among *continuous functions satsifying* $\phi(x_0) = y_0$.

Remark 5.14 Under the hypotheses of Theorem 5.13, the conclusion tells us that for any sufficiently small open subinterval I_1 of I containing x_0, there exists a unique *continuous* function $\phi : I_1 \to \mathbf{R}$ satisfying $F(x, \phi(x)) = c_0$ for all $x \in I_1$. If the interval I_1 is small enough, this function ϕ is continuously differentiable.

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*, 9th edition, Addison-Wesley, 2018.

[4] E.D. Rainville and P.E. Bedient, *A Short Course in Differential Equations*, 5th edition, Macmillan Publishing Co., 1974.

[5] G.B. Thomas, Jr., *Elements of Calculus and Analytic Geometry*, Addison-Wesley, 1959.

[6] Zill and Wright, *Differential Equations with Boundary Value Problems*, 8th edition, Brooks/Cole, 2013.

[110] *Note to instructors*: Many differential equations textbooks state such a weak version of the Implicit Function Theorem that this crucial point is missed, or is at best stated ambiguously. For example, many books state the conclusion in the form "Then there exists a unique differentiable function such that ..." or "Then there exists a unique continuously differentiable function such that ...", and assert only the "if" part of the "if and only if" in (5.15), i.e. that $F(x, \phi(x)) = 0$. Such statements are so weak as to be nearly useless.