

Some notes on first-order ODEs

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

Contents

1	Introduction	2
2	Notes for Instructors	3
3	Notes for Students	3
3.1	First-order DEs in derivative form	4
3.1.1	Review of “derivative form” and “solution”	4
3.1.2	Implicitly defined functions	9
3.1.3	Implicit solutions of derivative-form DEs	17
3.1.4	Maximal and general solutions of derivative-form DEs	31
3.1.5	General and implicit solutions on a region in \mathbf{R}^2	41
3.1.6	Algebraic equivalence of derivative-form DEs	46
3.1.7	General solutions of separable DEs	50
3.1.8	Algebraic equivalence and general solutions of linear DEs	61
3.2	First-order equations in differential form	67
3.2.1	Curves, parametrized curves, and smooth curves	72
3.2.2	Solution curves for DEs in differential form	77
3.2.3	Existence/uniqueness theorem for DEs in differential form	80
3.2.4	Solutions of DEs in differential form	82
3.2.5	Exact equations	88
3.2.6	Algebraic equivalence of DEs in differential form	90
3.3	Relation between differential form and derivative form	96
3.4	Using differential-form equations to help solve derivative-form equations	102
3.5	Using derivative-form equations to help solve differential-form equations	116
4	Optional Reading	119
4.1	The meaning of a differential	119
4.2	Exact equations: further exploration	122
4.3	One-parameter families of equations	125
5	Appendix	129
5.1	The Fundamental Theorem of ODEs	129

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 they are simply different ways of writing the same thing. *They are not*. The two forms are closely related, but differ in subtle ways not addressed adequately in most textbooks (and often overlooked entirely)¹. 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.² 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 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 these notes, originally, was simply to give a definition of “implicit solution” that is accurate, precise, complete, understandable by typical students in an introductory DE course, and sensible.³ More topics, such as an attempt to give a usable meaning to the term “explicit solution” in which the word “explicit” is not superfluous and is less misleading than in current textbooks, were added as the writing went along. This has made for a rather lengthy, never-quite-finished set of notes, an ongoing project that I work on on only occasionally.

In order to make the presentation readable concurrently with a typical modern DE textbook, in these notes I define “implicit solutions of a DE in derivative form”

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

²These notes contain numerous opinions of mine, but henceforth, qualifiers like “in my opinion” are mostly left implicit to avoid tortured writing.

³(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.

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 we cover differential-form DEs, we return to derivative-form equations to clean up the picture.

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

Domain of a function. 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”, since the domains I’m talking about may or may not be subsets of \mathbf{R} . If we were talking about, say, two-variable functions only, 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 $\frac{1}{x^2-1}$, 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 that is often called the *implied domain* of the formula in calculus and precalculus courses. However, in some instances we want to restrict attention to a smaller domain.

We can define a function f on, say, the interval $(1.1, 2.345)$ by setting $f(x) = \frac{1}{x^2-1}$. Using the same formula, this function can be *extended* to a function g on a larger domain, such as the interval $(1, \infty)$ or $\{x \in \mathbf{R} \mid x \neq \pm 1\}$. But it is often important to distinguish between two functions related this way, or, more generally, functions defined on different domains but computed using the same formula. 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 “ $f : D \rightarrow \mathbf{R}$ ” to emphasize this.

Rather than give students extra, unfamiliar notation to deal with, I will not use notation of the form “ $f : D \rightarrow \mathbf{R}$ ” 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 .

3.1 First-order DEs in derivative form

3.1.1 Review of “derivative form” and “solution”

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

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

$$F_1(x, y) = F_2(x, y), \quad (3.1)$$

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).⁴ Note that (3.1) 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 of the special case $F(x, y) = 0$ (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 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.⁵ Of course, this equation has infinitely many solutions; the set of all solutions is an ellipse in the xy plane.

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

$$\mathbf{G}\left(x, y, \frac{dy}{dx}\right) = 0, \quad (3.2)$$

⁴In these notes, the letter C (possibly with subscripts), in plain-italic font, denotes a constant unless otherwise specified.

⁵A 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.

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

Definition 3.1 For a given function G as above, a 1-variable function ϕ defined on an open set D in \mathbf{R} is said to *satisfy* equation (3.2) if ϕ is differentiable on D and when “ $y = \phi(x)$ ” is substituted into (3.2), the resulting equation is a true statement for each $x \in D$ (equivalently, such that $G(x, \phi(x), \phi'(x)) = 0$ for each $x \in D$).⁶

A *solution of* (3.2) is a function ϕ , with domain an open interval, that satisfies (3.2). If the domain-interval of the solution ϕ is I , we say that ϕ is a *solution of* (3.2) *on* I .⁷

For a given G , we call a one-variable function ϕ a *solution of* (3.2) (no interval mentioned) if ϕ is a solution of (3.2) on *some* open interval I . A *solution curve* of (3.2) is the graph of a solution, i.e. the set $\{(x, \phi(x)) \mid x \in I\}$, where ϕ is a solution of (3.2) on the interval I . ■

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

Henceforth, whenever we say “solution of a differential equation on an interval I ”, we always mean an *open interval* I .⁸

⁶“Open set” 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. In Definition 3.1, we require D to be an open set in order for $\phi'(x)$ to be defined (using the Calculus 1 definition) for each $x \in D$. However, this condition can be relaxed when D is an interval; see footnote 8.

⁷ 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 (3.2), but that $\phi(x)$ is an explicit solution of (3.2). 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, which is usually what one means by “explicitly-defined function”. We will deal with this terminological conundrum starting with Definition 3.2. I would prefer that textbook-authors and other instructors abandon the terminology “explicit solution”, but since I cannot make this happen, in these notes I give a definition that I believe is what the definition of “explicit solution” should have been all along, once the unfortunate decision to introduce this terminology was made. (3) The terminology leads to the absurd-sounding, “The functions implicitly defined by $F(x, y) = 0$ are explicit solutions (of the appropriate DE).”

⁸ In order to avoid certain distracting technicalities, in these notes we stick to *open intervals*

Definition 3.2 (temporary) If ϕ is a solution of the DE (3.2) (perhaps with an interval specified, perhaps not), we will call the equation “ $y = \phi(x)$ ” an *explicit solution* (one word, for now), of the DE.⁹ ■

Note that, according to Definition 3.1, an explicit solution of a DE is *not* a solution of the DE. A solution of a DE is a *function*; an explicit solution is an *equation*. A function and an equation are two different animals. An equation may be used to *define* a function, as in “ $\phi(x) = e^x$ ”. But “ ϕ ” is not the same thing as “the definition of ϕ ”, any more than an elephant is the same thing as the definition of an elephant.

Nonetheless, we allow ourselves to say, *technically incorrectly*, that “ $y = x^2$ is a solution of $\frac{dy}{dx} = 2x$ ” (for example), because that wording is so much less awkward than “the function ϕ defined by $\phi(x) = x^2$ is a solution of $\frac{dy}{dx} = 2x$ ”.¹⁰ 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. This particular 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.”

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

⁹The one-word term “explicit solution” was invented just for these notes. It is used here to preserve logical clarity before replacing it with the two-word phrase “explicit solution” whose use in modern textbooks like [3] is consistent with neither the conventional meaning of the word “explicit”, nor, in some cases, with the long-established and completely standard meaning of “solution of a differential equation” [1, p. 3].

¹⁰Only slightly more awkward than “ $y = x^2$ is a solution of $\frac{dy}{dx} = 2x$ ” is the following type of phrasing that you may have seen: “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 we (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 $\phi(x) = x^2$ ”, and some use it all the time. The language needed to avoid such phrasing is often extremely convoluted (unless the student has been introduced to the notation “ $x \mapsto x^2$ ”), so 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*.

Because of this, we make the following definition:

Definition 3.3 An *explicit solution* (two words) of a DE is an explicit solution of that DE. ■

This is not the same definition of “explicit solution” that appears in (for example) [3]. Rather, it is an attempt to reconcile any desire to use the phrase “explicit solution” with (i) the standard meaning of “solution of a differential equation, (ii) the expectation that an object called an “explicit solution” of a DE should, in particular, be a solution of that DE, and (iii) what appears to be the motive for introducing the phrase “explicit solution” into textbooks, namely “If there’s some object we’re going to call an *implicit* solution, we ought to call something by the name *explicit* solution.”¹¹

For later reference, we record the convention that motivated Definition 3.3:

**We allow ourselves to call an explicit solution
of a DE simply a *solution* of that DE.** (3.3)

Remark 3.4 (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

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

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 each case). All three of these are constant functions, so their derivatives are identically 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). 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 $\sin(3x) \neq 0$ (in fact, this is the case for *most* x in any interval), hence for which “ $0 = -7 \sin(3x)$ ” is a false statement.¹² Hence ϕ_3 is *not* a solution of equation (3.4).

The constant solution ϕ_1 above may be expressed any of the following ways:

- (i) $y = 7$.

¹¹As noted earlier, far better than using the definition of “explicit solution” in these notes would be for authors and instructors to abandon using the phrase “explicit solution” with anything other than its historical meaning: a solution that we are expressing by an explicit *formula*. But until that happens, the definitions in these notes may be of some use.

¹²The 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.

(ii) $y \equiv 7$.

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

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

The notation (i) is the one most consistent with Definitions 3.2 and 3.3 of “explicit solution”, but its interpretation depends very strongly on context. In the present context, (i) does not mean “ y is the number 7.” *When it is 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 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 “ $G(x, \phi(x), \phi'(x)) = 0$ for each $x \in D$ ”, we could have written “ $G(x, \phi(x), \phi'(x)) \equiv 0$ 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) is simply an extra-forcible reminder that we’re talking about a constant function with the indicated variable-names.

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 .

Note that an equation of the form “ $x = \text{constant}$ ” (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*.¹³

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

¹³This terminology is most common for *autonomous* DEs: equations of the form $\frac{dy}{dx} = p(y)$, a particular type of *separable* equation.

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). ■

3.1.2 Implicitly defined functions

Key to understanding what “implicit solution of a differential equation” (defined in the next section) means is to understand what “implicitly defined *function*” (of one variable) means. You learned about 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 . We can always write such an equation in the form

$$F(x, y) = 0 \tag{3.5}$$

for some two-variable function F . We may be interested in solving (3.5) 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

$$x^2 + y^3 - 1 = 0, \tag{3.6}$$

then

$$y = (1 - x^2)^{1/3}. \tag{3.7}$$

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

$$F(x, y) = 0 \quad \text{if and only if} \quad y = \phi(x). \tag{3.8}$$

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

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

More generally than this example, any time (3.8) is true for a two-variable function F and one-variable function ϕ , we say that the equation $F(x, y) = 0$ *implicitly determines* (or *implicitly defines*) y as a function of x , and we call ϕ the function of x implicitly determined/defined by the equation $F(x, y) = 0$.

Now consider the equation

$$x^2 + y^2 - 1 = 0. \quad (3.9)$$

“Solving for y in terms of x ” gives the relation

$$y = \pm\sqrt{1 - x^2}. \quad (3.10)$$

Looking just at (3.9), it is already clear that any numerical choice of x restricts the possible choices of y that will make the equation a true statement. Equation (3.10) tells us the only possible values for y that will work. It also tells us that for each x in the open interval $(-1, 1)$ there are at most two such values; for $x = 1$ and for $x = -1$ there is at most one such value; and for $|x| > 1$ there are no values of y that will work. Conversely, if we substitute $y = \pm\sqrt{1 - x^2}$ into (3.9), we see that all the values of y that we have labeled as “possible” actually do work. Thus, *for each pair (x, y) of real numbers,*

$$x^2 + y^2 - 1 = 0 \text{ if and only if } |x| \leq 1 \text{ and either } y = \sqrt{1 - x^2} \text{ or } y = -\sqrt{1 - x^2}. \quad (3.11)$$

This is a *much* weaker statement than a statement of the form (3.8), because the sign in $\pm\sqrt{1 - x^2}$ can be chosen independently for each x . On the domain $[-1, 1]$, if we define

$$\phi_1(x) = \sqrt{1 - x^2}, \quad (3.12)$$

$$\phi_2(x) = -\sqrt{1 - x^2}, \quad (3.13)$$

$$\phi_3(x) = \begin{cases} \sqrt{1 - x^2} & \text{if } x \text{ is a rational number,} \\ -\sqrt{1 - x^2} & \text{if } x \text{ is an irrational number,} \end{cases} \quad (3.14)$$

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 (3.9). 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

$$e^x + x + 6y^5 - 15y^4 - 10y^3 + 30y^2 + 10xy^2 = 0. \quad (3.15)$$

Clearly, choosing a numerical value for x restricts the possible values for y that will make equation (3.15) a true statement. It turns out that, depending on the choice x ,

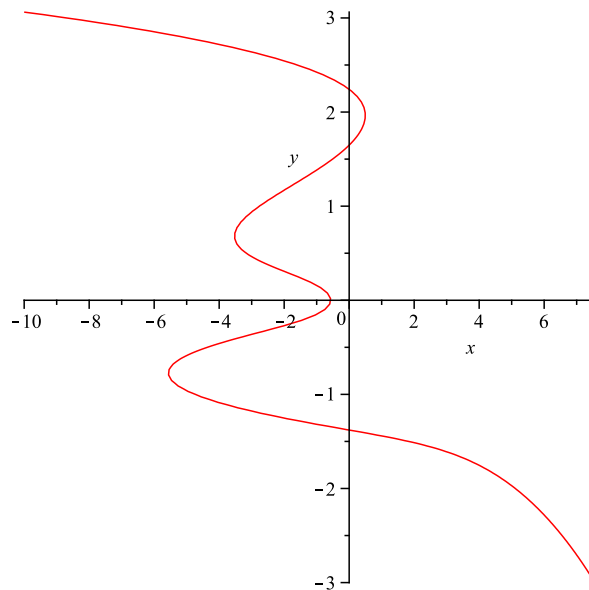


Figure 1: The graph of $e^x + x + 6y^5 - 15y^4 - 10y^3 + 30y^2 + 10xy^2 = 0$.

there can be anywhere from one to five values of y for which the pair (x, y) satisfies equation (3.15). 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 (3.15). However, there are not very many *continuous* ϕ ’s that work. In this example, whatever x -interval I we choose, there are at most five continuous functions ϕ defined on I for which $F(x, \phi(x)) = 0$. Writing out *explicit formulas* for them, analogous to the formulas for ϕ_1 and ϕ_2 in the previous example, is a hopeless task. But these continuous functions ϕ exist nonetheless. We can see this visually in Figure 1.

Definition 3.5 Let F be a function of two variables, ϕ a function of one variable, and I an interval. We say that the relation (or equation) $y = \phi(x)$ *satisfies* the relation (or equation) $F(x, y) = 0$ if $F(x, \phi(x)) = 0$ for all $x \in I$. We say that the equation $F(x, y) = 0$ *implicitly determines* (or, simply, *determines*) the function ϕ , regarded as a function of the first variable of F , if $F(x, \phi(x)) = 0$ for all $x \in I$.¹⁴

¹⁴*Note to instructors:* I dislike this usage of the word *determines*. I decided to use it here because the even more objectionable *defines* appears in at least one current textbook’s definition of “implicit solution of a DE”, without the authors having said what this “defines” means. It can be deduced from this textbook’s wording of a later exercise (in which a theorem much weaker than the Implicit Function Theorem is called the Implicit Function Theorem) that what the authors mean by “a function of x implicitly defined by $F(x, y) = 0$ ” is exactly what Definition 3.5 calls “a function of x implicitly *determined* by $F(x, y) = 0$ ”. Unfortunately, this usage of “a function of x implicitly defined by $F(x, y) = 0$ ” is at odds with what “implicitly defined” means in the standard definition

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

of “implicitly defined function” (Definition 3.7). As bad as it is to have “implicitly determined” not mean the same thing as “implicitly defined”, I felt that this was preferable to having “function implicitly defined by an equation” not mean something different from “implicitly defined function”, or to give students the wrong definition of what it means for an equation $F(x, y) = 0$ to define y locally as a function of x .

Both “determined” and “defined” and are best used only when, for any potentially non-unique object being “determined”, there are conditions that guarantee uniqueness (as in (3.18)). Any other usage is a significant and unnecessary departure from standard English. According to Definition 3.5 the equation $0 = 0$, viewed as an equation on $I \times \mathbf{R}$, implicitly determines *every* function $I \rightarrow \mathbf{R}$. For a less obvious example, on the set $\{(x, y) \in \mathbf{R}^2 \mid x^2 + y > 0\}$ the equation $(x^2 + y)^{1/\ln(x^2+y)} - e = 0$ determines every function $\phi : \mathbf{R} \rightarrow \mathbf{R}$ for which $\phi(x) > -x^2$. And clearly we can cook up an arbitrarily complicated expression $F(x, y)$ such that “ $F(x, y) = 0$ ” *non-obviously* reduces to an identity. Altering Definition 3.5 to exclude any equation that restricts to an identity on some nonempty open subset of \mathbf{R}^2 would, of course, eliminate the examples just given, but would complicate the definition without eliminating all of its intrinsic problems.

There really is no need to have *both* of the last two sentences in Definition 3.5. Anything that can be said using one of these definitions can be said using the other; they express the same relation between F and ϕ . I’ve included the second sentence since using “satisfies” expresses this relation in a problem-free way. I’ve included the last sentence only to make contact with a textbook. My preference would be to simply erase the last sentence of Definition 3.5.

However, I’ve found no substitute for the last sentence of Definition 3.5 that is adequate if one insists on defining “implicit solution” using wording of the sort “an equation/relation that (implicitly) defines/determines one or more solutions of the DE”. Note that the relation expressed in Definition 3.5 *is* important; implicitly defined functions (Definition 3.7) are not a rich enough class to let us express all solutions of even some rather simple DEs. See Example 3.16. 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 (3.40). (We can get around the “implicitly determines” *wording*, at the expense of brevity, using the “satisfies” terminology, but we can’t get around needing a way to say that all solutions of the DE can be extracted from the equations $\{x^2 - y^2 = \text{constant}\}$.) 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 3.5 allows us to say that these solutions are implicitly determined by the 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 (3.18) holds.

Thus, taking “function of x implicitly determined by $F(x, y) = 0$ ” to mean “implicitly defined function” (as defined in Definition 3.7) would lead to an inadequate definition of “*solution of a DE* implicitly determined by $F(x, y) = 0$ ”. But far worse would be to compound that by changing the standard meaning of “implicitly defined function” to “function given by applying the Implicit Function Theorem”. No matter how much this would simplify the lives of DE instructors, it would be an artificial, essentially tautological definition that would beg the question, “Why do we call the Implicit Function Theorem the Implicit Function Theorem?”

In the setting of Definition 3.5, there *is* an object that is uniquely determined by the function F and interval I : the *set* of all functions $\phi : I \rightarrow \mathbf{R}$ satisfying $F(x, \phi(x)) = 0$, rather than individual elements of this set. Even for the “ $0=0$ ” example, I feel much more comfortable saying that the equation $0=0$ determines the set of all functions $I \rightarrow \mathbf{R}$, than saying that it determines the function $x \mapsto x^{53} + e^x - \sqrt{x^2 + 1} \tan^{-1} x$. But that is not something to mention to intro DE students.

on *some* open interval.

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

Graphically, a function ϕ is implicitly determined by the equation $F(x, y) = 0$ if the graph of ϕ is part of the graph of $F(x, y) = 0$.¹⁵ (For these purposes, “all of” is a special case of “part of”.)

There are instances in which we want to know whether there is a one-variable function ϕ such that $F(\phi(y), y) = 0$. This comes up when we think of trying to solve the equation $F(x, y) = 0$ for x in terms of y , rather than for y in terms of x . To handle this case we can give a definition analogous to Definition 3.5, replacing the phrases “regarded as a function of x ” and “first variable” with “regarded as a function of y ” and “second variable”, and replacing “ $F(x, \phi(x)) = 0$ ” with “ $F(\phi(y), y) = 0$ ”. To simplify wording below, **any time we say that an equation $F(x, y) = 0$ implicitly determines a function ϕ , we mean to regard ϕ as a function of the first variable of F —i.e. we mean that $F(x, \phi(x)) = 0$ identically on the interval in question** (unless we say otherwise). Thus:

- Equation (3.6) determines the function ϕ on $(-\infty, \infty)$ given by the formula $\phi(x) = (1 - x^2)^{1/3}$.
- Equation (3.9) determines the functions ϕ_1, ϕ_2, ϕ_3 defined in (3.12)–(3.14), and infinitely many others on the interval $[-1, 1]$. The only *continuous* functions that (3.9) determines on $[-1, 1]$ are ϕ_1 and ϕ_2 .
- Equation (3.15) determines infinitely many functions, but only a few *continuous* functions. In Figure 1, if we travel along the graph by starting at the upper left and moving along the curve, we encounter vertical tangents at four points. Label these points $A, B, C,$ and D (in the order that we encounter them), and let $x_A, x_B, x_C,$ and x_D denote their x -coordinates. Then (3.15) implicitly determines a continuous function of x , say ϕ_1 , with domain $(-\infty, x_A]$; another continuous function of x , say ϕ_2 , with domain $[x_B, x_A]$; another, say ϕ_3 , with domain $[x_B, x_C]$; another, say ϕ_4 , with domain $[x_D, x_C]$; and another, say ϕ_5 , with domain $[x_D, \infty)$. On the interval $[-3, -2]$, the equation $F(x, y) = 0$ determines five continuous functions (the restrictions of $\phi_1, \phi_2, \phi_3, \phi_4,$ and ϕ_5 to this interval). On the interval $[-5, -4]$, $F(x, y) = 0$ determines three continuous functions (the restrictions of $\phi_1, \phi_4,$ and ϕ_5 to this interval).

¹⁵Recall that the *graph* of an equation in x and y is the solution-set of the equation: the set of points $(x, y) \in \mathbf{R}^2$ for which the equation is a true statement.

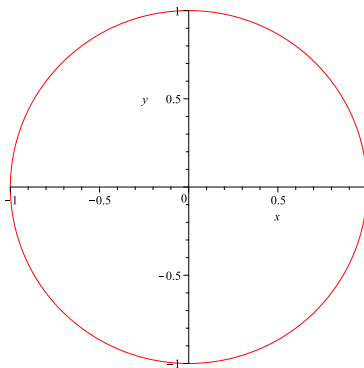


Figure 2: The graph of $x^2 + y^2 = 1$.

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

$$I \times J = \{(x, y) \mid x \in I \text{ and } y \in J\}. \quad (3.16)$$

In these notes we will call such a set a *rectangle*, even though we do not exclude the possibility that I and/or J extend(s) infinitely in one direction or both.¹⁶ Thus, for example, we consider all of the following to be rectangles: the whole xy plane; the set $[1, \infty) \times (-\infty, \infty)$ (the set of all pairs (x, y) for which $x > 1$); the strip $(-\infty, \infty) \times (0, 1]$ (the set of all pairs (x, y) with $0 < y \leq 1$). Of course, objects that Euclid would have called rectangles, such as $[1, 2] \times [3.1, 4.9]$, are also rectangles in our terminology. In these notes, we will be most interested in *open* rectangles, those we get by taking the intervals I and J to open.

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

Let $P = (x_0, y_0)$ be any point on the circle *other than* $(1, 0)$ or $(-1, 0)$; thus $y_0 \neq 0$. For any such point, you can draw an open rectangle $R = I \times J$, containing (x_0, y_0) , such that the portion of the circle lying in R is a portion of the graph of *exactly one* of the two functions ϕ_1, ϕ_2 in (3.12)–(3.13) ($\phi_1(x) = \sqrt{1 - x^2}$, $\phi_2(x) = -\sqrt{1 - x^2}$). For example, if $y_0 > 0$ you can take J to be any open subinterval of $(0, \infty)$ that contains y_0 , and then take I to be any open interval whatsoever that contains x_0 . Choose some

¹⁶The notation “ $I \times J$ ” is defined exactly as in (3.16) whether or not I and J are intervals, but $I \times J$ is not called a rectangle unless I and J are intervals.

points on the graph in Figure 2 and draw rectangles around them with the desired property.

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

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

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

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

Theorem 3.6 (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 defined on I_1 , such that

$$\begin{aligned} &\text{for all points } (x, y) \in I_1 \times J_1, \\ &F(x, y) = c_0 \text{ if and only if } y = \phi(x). \end{aligned} \tag{3.17}$$

■

In Theorem 3.6, since x_0 lies in I_1 , we may look at what (3.17) tells us when $x = x_0$. What this statement reduces to when $x = x_0$ is the following:

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

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

Let us pause to appreciate how strong the conclusion of this theorem is. Statement (3.17) 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, (3.17) says that within $I_1 \times J_1$, the equation $F(x, y) = c_0$ determines y *uniquely* as a function of x —not just uniquely among “nice” functions, like continuous functions or differentiable functions. Among *all* functions with domain I_1 and range contained in J_1 , ϕ is the *only* function that satisfies $F(x, \phi(x)) = c_0$ identically in x . This function has the *additional nice feature* of being continuously differentiable (and hence continuous), but there is *no other function whatsoever* on I_1 that satisfies $F(x, \phi(x)) = c_0$ identically in x .

Compared statement (3.17) with statement (3.8). The only important difference is that to get the second line of (3.17), we had to make the windowing restriction in the first line. (The fact that we have “ c_0 ” in (3.17) where we have “0” in (3.8) is an unimportant difference.)

The uniqueness (of the function ϕ) that is guaranteed by a statement of the form (3.17) allows us to use terminology that is less awkward than what we used in Definition 3.5. Specifically, whenever a statement of the form (3.17) holds true (whether or not the other hypotheses of the Implicit Function Theorem are met), we can dispense with the phrase “regarded as a function of the first variable of F ” in that definition, or even introducing a letter for the function ϕ at all. We may simply say the following:

$$\begin{aligned} &\text{Within the rectangle } I_1 \times J_1, \text{ the equation} \\ &F(x, y) = c_0 \text{ determines } y \text{ as a function of } x. \end{aligned} \tag{3.18}$$

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

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

$$\begin{aligned} &\text{for all points } (x, y) \in I_1 \times J_1, \\ &F(x, y) = c_0 \text{ if and only if } x = \phi(y) \end{aligned} \tag{3.19}$$

then we can say simply that within the rectangle $I_1 \times J_1$, the equation $F(x, y) = c_0$ determines x uniquely as a function of y . Thus, when condition (3.19) is met, we do not have to write a whole new definition analogous to Definition 3.5, with “regarded as a function of the first variable” replaced with “regarded as a function of the second variable”, and with “ $F(x, \phi(x)) = c_0$ ” replaced with “ $F(\phi(y), y) = c_0$ ”.

Definition 3.7 Let F be a two-variable function defined on an open rectangle $I \times J$. When either (3.17) or (3.19) holds for some open sub-rectangle $I_1 \times J_1$, we call ϕ an *implicitly defined function* determined by the equation $F(x, y) = c_0$.¹⁷

Note that there may be functions implicitly determined by an equation $F(x, y) = c_0$ (see Definition 3.5) that do *not* meet this definition of “implicitly defined function”.

Reminder: “ $F(x, y) = c_0$ ” is equivalent to “ $F_{\text{new}}(x, y) = 0$,” where $F_{\text{new}}(x, y) = F(x, y) - c_0$. The c_0 in (3.17) and (3.19) is simply a remnant of the way phrased the Implicit Function Theorem; we are not talking about equations that are any more general than those of the form “function of $(x, y) = 0$.”

To simplify wording, henceforth whenever we speak of an implicitly defined function determined by an equation $F(x, y) = c_0$, we mean to “regard ϕ as a function of x ”—i.e. to mean that (3.17) is satisfied (unless we say otherwise).

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 uniquely as a function of x ? (Don’t try to find the *values* of x_0 and y_0 ; just show with your pencil where these “bad” points are on the graph.) ■

3.1.3 Implicit solutions of derivative-form DEs

Now, let us get back to differential equations:

Definition 3.8 (temporary) We call an equation $F(x, y) = 0$ an *implicit solution* (one word, for now) of a differential equation

$$\mathbf{G}\left(x, y, \frac{dy}{dx}\right) = 0 \tag{3.20}$$

(for a given three-variable function \mathbf{G}) if

(i) there is at least one solution ϕ of (3.20) that is an implicitly defined function determined by the equation $F(x, y) = 0$, and

¹⁷The informal terminology “implicit function” is a less precise but common phrase meaning “implicitly defined function”. The only good use of the term “implicit function” is in the title of the Implicit Function Theorem, where it provides a way to avoid the awkward title “Implicitly-Defined-Function Theorem”.

(ii) every differentiable function ϕ determined by the equation $F(x, y) = 0$, with domain an open interval, is a solution of (3.20).¹⁸ ■

Definition 3.9 If ϕ is a differentiable, implicitly defined function determined by an implicit solution $F(x, y) = 0$ of (3.20), then we call ϕ an *implicitly defined solution* of (3.20). ■

Example 3.10 Consider the differential equation

$$x + y \frac{dy}{dx} = 0. \quad (3.21)$$

We claim that the equation

$$x^2 + y^2 - 1 = 0 \quad (3.22)$$

is an implicit solution of (3.21). (Equivalently, so is the equation $x^2 + y^2 = 1$.) To verify this, we check that criteria (i) and (ii) of Definition 3.8 are satisfied:

¹⁸*Note to instructors:* (a) The definition of “implicitly defined solution” does not rely at all on implicit differentiation of the equation $F(x, y) = 0$, and neither should the definition of “implicit solution”. The function F need not even be continuous, let alone differentiable, for the concept of “solution implicitly defined by $F(x, y) = 0$ ” 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 defined function that happens to be a solution of the DE. The *notion* of implicitly defined function does not rely on calculus in any way.

Of course, it is tremendously important that the Implicit Function Theorem gives sufficient conditions under which we can confirm, via implicit differentiation, that we have an implicit solution of a DE. 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.” Few students, if any, will mention any relation to the notion of implicitly defined *function*, or to (true) solutions of the DE (“explicit solutions”, in the terminology whose existence I bemoan). 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 3.18 and on your students, and perhaps also Example 3.17.

(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 $F(x, y) = 0$ 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 3.11. Many older textbooks avoid this problem by *not attempting to formally define “implicit solution”*; see footnote 24.

- Criterion (i). Let $F(x, y) = x^2 + y^2$ and let $\phi_1(x) = \sqrt{1 - x^2}$ as in (3.12), but restricted to the open interval $(-1, 1)$. Note that $F(x, \phi_1(x)) = 1$ for all $x \in (-1, 1)$, so ϕ_1 is a function implicitly determined by the equation $F(x, y) = 1$ (the conditions of Definition 3.5) are met).

We compute $\phi_1'(x) = \frac{-x}{\sqrt{1-x^2}}$. Thus if we substitute $y = \phi_1(x)$ into the left-hand side of (3.21), we have

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

so ϕ_1 is a solution of (3.21). Thus criterion (i) is satisfied¹⁹.

- Criterion (ii). Suppose ϕ is any differentiable function determined implicitly by (3.22) on some open interval I . Then we have

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

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

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

Therefore ϕ is a solution of the equation

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

on I . Dividing by 2 we see that ϕ is a solution of (3.21) on I . Therefore criterion (ii) is satisfied.

Hence (3.22) is an implicit solution of (3.21), and the function ϕ_1 is an implicitly defined solution of (3.21).

There are actually two implicitly defined solutions in this example: ϕ_1 and $-\phi_1$ (the function that we called ϕ_2 in (3.13)). The first of these is the function implicitly defined by $x^2 + y^2 = 1$ on the rectangle $(-1, 1) \times (0, \infty)$; the second is the function implicitly defined by $x^2 + y^2 = 1$ on the rectangle $(-1, 1) \times (-\infty, 0)$. Both functions are solutions of (3.21). ■

¹⁹We could just as well have used the function ϕ_2 defined by $\phi_2(x) = -\sqrt{1 - x^2}$. But to show that criterion (i) is met it suffices to come up with *one* function ϕ that works, so we chose the ϕ that involves (slightly) less writing.

Example 3.11 We claim that

$$(y - e^x)(x^2 + y^2 - 1) = 0 \tag{3.23}$$

satisfies criterion (i) in Definition 3.8 to be an implicit solution of (3.21), yet is *not* an implicit solution of this DE.

To see this, first note that, from Example 3.10, the function ϕ_1 defined by $\phi_1(x) = \sqrt{1 - x^2}$ is a solution of (3.21) on the interval $(-1, 1)$. On this interval, if we substitute $y = \sqrt{1 - x^2}$ into (3.23), 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 function implicitly determined by (3.23), and is a solution of (3.21). Hence criterion (i) in Definition 3.8 is satisfied: equation (3.23) implicitly determines at least one solution of (3.21).

However, if we substitute $y = e^x$ into (3.23), 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 a function determined implicitly by (3.23). However, if we substitute $y = e^x$ into (3.21), we get

$$x + e^{2x} = 0. \tag{3.24}$$

Is it possible to choose the interval I in such a way that (3.24) holds true for all $x \in I$? No, for if there were such an interval I , the left-hand side of (3.24) would be a differentiable function on I , so we could differentiate both sides of (3.24) and obtain

$$1 + 2e^{2x} = 0. \tag{3.25}$$

But there isn't even a single value of x for which this is true; $1 + 2e^{2x} > 0$ for all x . Thus there is no open interval I on which ϕ is a solution of the DE (3.21).

Thus ϕ is a differentiable function determined implicitly by (3.23) that is not a solution of (3.21). Therefore criterion (ii) in Definition 3.8 is not met, so equation (3.23) is not an implicit solution of the DE (3.21). ■

Example 3.12 The equation

$$x^2 + y^2 + 1 = 0 \tag{3.26}$$

is *not* an implicit solution of (3.21), because it fails criterion (i) of Definition 3.8. There are no real numbers x, y at all for which (3.26) holds, let alone an open interval I on which (3.26) implicitly determines a function of x . Since (3.26) determines no functions ϕ whatsoever on any open interval I , criterion (ii) of Definition 3.8 is moot.

Similarly, the equation

$$x^2 + y^2 = 0 \tag{3.27}$$

is not an implicit solution of (3.21). In this case there *is* a pair of real numbers (x, y) that satisfies (3.27), but there is no *open* x -interval I on which, for each $x \in I$, there

is a real number y for which (3.27) is satisfied. ■

Now let us make an observation about implicit solutions:

An implicit solution of a DE is not a solution of that DE. (3.28)

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.

However, in Section 3.1.1, we said that if ϕ is a solution of a DE $G(x, y, \frac{dy}{dx})$, 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 the equation “ $y = \phi(x)$ ”, which we are allowing ourselves to call a solution of a DE if ϕ is a solution of that DE, is equivalent to the equation “ $y - \phi(x) = 0$ ”, which is an equation of the form $F(x, y) = 0$. So in the same spirit as in Definition 3.3, we make the following definition:

Definition 3.13 We say that an equation $F(x, y) = 0$ (or, more generally, $F_1(x, y) = F_2(x, y)$) is an *implicit solution* (two words) of a given differential equation if it is an implicit solution (one word) of that differential equation, as defined in Definition 3.8. ■

Combining Definition 3.13 with observation (3.28), we have a linguistic paradox:

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

In other words, the meaning of “implicit solution” cannot be obtained by interpreting “implicit” as an adjective modifying “solution”. One must regard the two-word phrase “implicit solution” as a single term, a compound noun whose meaning cannot be deduced from the meanings of the two words comprising it. That is why we initially used the made-up word “implicitsolution”, which the student is not likely to find outside these notes. Of course, as we observed earlier (but did not display in a line like (3.1.3)), the term “explicit solution” has the same problem: an explicit solution of a DE, as defined by Definitions (3.2) and (3.3) is not literally a solution of the DE, according to the standard definition (3.1) of “solution of a differential equation”.²⁰

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

²⁰Again, the problem is that the term “explicit solutions”, as it is used in in current textbooks (and defined in these notes *only* because it is used in current textbooks), is intrinsically flawed.

is no object that is actually called an “implicit solution”), give a definition that is similar to our definition of “implicitsolution”²¹.

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

Note that the terminology “implicitly defined solution” (Definition 3.9) does not suffer from any paradox. An implicitly defined solution of a DE *is* a solution of that DE. It meets the criteria of Definition 3.1 perfectly.

Also observe the following:

Every *explicit* solution of a given DE (as defined in Definitions 3.3 and 3.2) is also an *implicit* solution. (3.29)

Our approach to Example 3.10 relied on our ability to produce an explicit formula for a “candidate solution” of the given DE. What if, in place of (3.22), we had been given an equation so complicated that we could not solve for y and produce a candidate-solution ϕ to plug into the DE? This is where the Implicit Function Theorem comes to the rescue.

Example 3.14 ²² Show that the equation

$$x + y + e^{xy} = 1 \tag{3.30}$$

is an implicit solution of

$$(1 + xe^{xy})\frac{dy}{dx} + 1 + ye^{xy} = 0. \tag{3.31}$$

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 3.6). We compute

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

²¹Except that criterion (ii) seems to have been either overlooked or deliberately omitted in all the textbooks I have seen. Example 3.11, as well as the discussion in footnote 14, show that omitting this criterion can lead to calling a function an “implicit solution of a (given) DE” when it is nonsensical to do so.

²²This example is taken from Nagle, Saff, and Snider, *Fundamentals of Differential Equations and Boundary Value Problems*, 5th ed., Pearson Addison-Wesley, 2008.

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 3.6 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 3.21: implicit differentiation (i.e. computing derivatives of an expression that contains an implicitly defined function). Let us simplify the notation a little by writing $y(x) = \phi(x)$. Then

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

Therefore ϕ is a solution of (3.31). Thus, criterion (i) in Definition 3.8 is satisfied. The exact same implicit-differentiation argument shows that if ψ is *any* differentiable function determined on an open interval by (3.30), then ψ is a solution of (3.31). Therefore criterion (ii) in Definition 3.8 is also satisfied. Hence (3.30) is an implicit solution of (3.31). ■

Looking back at Example 3.10, could we have shown that criterion (i) of Definition 3.8 is satisfied using the technique of Example 3.14, 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 3.14, the partial derivatives of F are continuous on whole xy plane again²³, 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 3.10 (and as we did to check both criteria in Example 3.14), to show that ϕ is a solution of (3.21). If our point (x_0, y_0) has $y_0 > 0$, then the solution of (3.21) 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 (3.21) that we get is $-\phi_1$.

The student may wonder how we could have used the method of Example 3.14 had we not been clever (or lucky) enough to be able to find a point (x_0, y_0) that lay on

²³This does not always happen—Examples 3.10 and 3.14, and several other examples in these notes, just happen to have F 's with this property.

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

$$x + y + e^{xy} = 2 \quad (3.32)$$

instead of (3.30), we would have had a much harder time. We could show by implicit differentiation that every differentiable function determined by (3.32) is a solution of (3.31)—thus, that criterion (ii) of Definition 3.8 is satisfied—but that would not tell us that there is even a single function of x defined by (3.32), or even that the graph of (3.32) contains any points whatsoever. Conceivably, we could be in the same situation as in Example 3.12, in which all differentiable functions implicitly defined by (3.26)—all none of them—are solutions of our differential equation.

Blue text in these notes is *optional* reading, except in honors classes.

It so happens that we *can* show that the graph of (3.32) contains a point at which $\frac{\partial G}{\partial y}$ is not 0. However, doing that would require a digression that we do not want to take right now. Instead, let us consider a different type of problem that can be handled far more easily, even though the function $F(x, y)$ is much more complicated.

Example 3.15 Show that there is a number c_0 for which the equation

$$e^x + x + y^5 - y^4 + y^3 + y^2 + xy^2 = c_0 \quad (3.33)$$

is an implicit solution of the differential equation

$$e^x + 1 + y^2 + (5y^4 - 4y^3 + 3y^2 + 2y + 2xy) \frac{dy}{dx} = 0. \quad (3.34)$$

To approach this problem, we start with a variation on the second step of Examples 3.10 and 3.14: we assume that there is a number c_0 for which (3.33) implicitly determines a differentiable function ϕ , say on an interval I . On the interval I , we may then implicitly differentiate the equation (3.33)—i.e. differentiate with respect to x both sides of the equation we obtain by substituting “ $y = \phi(x)$ ” into (3.33). To keep the notation as simple as possible, we will just write “ y ” instead of “ $y(x)$ ” or “ $\phi(x)$ ” when we differentiate. (This is usually what we do when we differentiate implicitly; we just haven’t done it until now in these notes.) Then, using the chain rule and product rule, we find

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

which is equivalent to equation (3.34).

Thus, all differentiable functions ϕ determined implicitly by an equation of the form (3.33) will be solutions of (3.34). Thus for any c_0 for which (3.33) implicitly

determines a differentiable function, equation (3.33) will be an implicit solution of (3.34).

So, if we can show that there *is* such a c_0 , we'll be done. For this, we look to the Implicit Function Theorem to help us out. Letting $F(x, y)$ denote the left-hand side of (3.33), we compute

$$\frac{\partial F}{\partial x}(x, y) = e^x + 1 + y^2, \quad (3.35)$$

$$\frac{\partial F}{\partial y}(x, y) = 5y^4 - 4y^3 + 3y^2 + 2y + 2xy. \quad (3.36)$$

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

$$\frac{\partial F}{\partial y}(x, y) = y(5y^3 - 4y^2 + 3y + 2 + 2x). \quad (3.37)$$

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

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

So if we take, for example, $(x_0, y_0) = (0, 1)$, then $\frac{\partial F}{\partial y}(x_0, y_0) \neq 0$. For this choice of (x_0, y_0) , we have $F(x_0, y_0) = 3$. The Implicit Function Theorem then guarantees us that on some open x -interval containing 0, the equation $F(x, y) = 3$ implicitly determines a differentiable function of x . By the first part of our analysis (the part that involved implicit differentiation), this guarantees that the equation $F(x, y) = 3$ is an implicit solution of (3.34). So we have found a c_0 with the desired property. ■

As you probably noticed, in this example our expressions (3.35)–(3.36) for the partial derivatives of F appeared also in (3.34). 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

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial y} \frac{dy}{dx} = 0. \quad (3.38)$$

With foresight, the author chose the DE (3.34) to be exactly the equation (3.38) for $F(x, y)$ equal to the left-hand side of (3.33). For *most* DEs, it will *not* be true that there is a value of c_0 for which (3.33) is an implicit solution.

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 (3.38) actually come up a lot.

You may also have noticed, in Example 3.15, that we could have come up with a whole lot of points (x_0, y_0) that “worked”, in the sense that the hypotheses of the Implicit Function Theorem would have been met. All we needed was a point (x_0, y_0) for which $y(5y^3 - 4y^2 + 3y + 2 + 2x)|_{(x_0, y_0)} \neq 0$. But “almost every” choice (x_0, y_0) has this property; we just need $y_0 \neq 0$ and $x_0 \neq -\frac{1}{2}(5y_0^3 - 4y_0^2 + 3y_0 + 2)$. For each nonzero choice of y_0 , there’s only one “bad” choice of x_0 ; every other real number is a good choice of x_0 . So the c_0 ’s for which our method shows that (3.33) is an implicit solution of (3.34), are all the numbers $F(x_0, y_0)$ we can get by plugging in “good” choices of (x_0, y_0) (i.e. all choices with $y_0 \neq 0$ and $x_0 \neq -\frac{1}{2}(5y_0^3 - 4y_0^2 + 3y_0 + 2)$). We can expect this set of numbers to be a large subset of the range of F —perhaps the whole range of F . A challenging question for you to think about is this: are there any numbers c_0 for which (3.33) is *not* an implicit solution of (3.34)? Let’s strip away the distracting complexity of the function F in (3.33) and pose the analogous question for a much simpler F , the one in Example 3.14:

Question: Are there any numbers c_0 for which the equation

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

is *not* an implicit solution of (3.31)? (Note that the differential equation (3.31) is simply equation (3.38) for the function F defined by $F(x, y) = x + y + e^{xy}$.) ■

This question will not be answered in these notes; it is left as a challenge for the student. We point out that the answer to such a question will not be the same for all functions F that we could put on the left-hand side of “ $F(x, y) = c_0$ ”. For example, if we take $F(x, y) = x^2 + y^2$, then only for $c_0 > 0$ is the equation $F(x, y) = c_0$ an implicit solution of (3.21) (which is the equation (3.38) for this F , simplified by dividing by 2). But if we take $F(x, y) = x + y$, then for every real number c_0 the equation $F(x, y) = c_0$ is an implicit solution of the analogous differential equation, $1 + \frac{dy}{dx} = 0$, as you can see easily by explicitly solving the equation $x + y = c_0$ for y in terms of x .

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

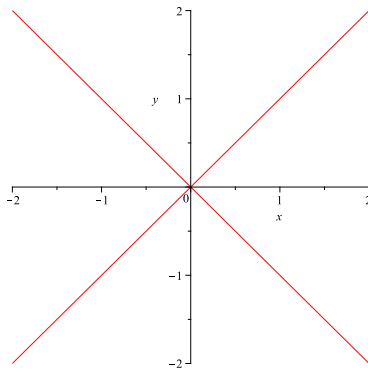


Figure 3: The graph of $x^2 - y^2 = 0$.

Example 3.16 Consider the algebraic equation

$$x^2 - y^2 = 0 \tag{3.39}$$

and the differential equation

$$x - y \frac{dy}{dx} = 0. \tag{3.40}$$

Equation (3.39) is equivalent to $y = \pm x$. Thus on any interval I , equation (3.39) implicitly determines two differentiable functions ϕ of x , namely $\phi(x) = x$ and $\phi(x) = -x$. Both of these are solutions of (3.40). Therefore (3.39) is an implicit solution of (3.40), and the two functions ϕ above are implicitly-defined solutions of (3.40), on any interval.

The point $(x, y) = (0, 0)$ satisfies (3.39). But on no open rectangle containing the point $(0, 0)$ does (3.39) 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 (3.17) holds.

Does this contradict the Implicit Function Theorem? No—the theorem says only that there are I_1 and J_1 with the property (3.17) *if the hypotheses of the theorem are met*. But in the current example, the function F for which (3.39) is of the form $F(x, y) = c_0$ is given by $F(x, y) = x^2 - y^2$. Thus $\frac{\partial F}{\partial y}(x, y) = -2y$, and if we take $(x_0, y_0) = (0, 0)$ 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 (3.40); they just are not solutions that the Implicit Function Theorem finds. ■

For most two-variable functions F that we encounter in practice, the “bad” points

(x_0, y_0) in the domain of F —the points at which the Implicit Function Theorem does not apply—are of two types: points at which the graph of $F(x, y) = F(x_0, y_0)$ has a vertical tangent (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 of our previous examples, on any open interval the continuous implicitly-defined functions and the differentiable implicitly-defined functions were the same.

From the examples presented so far, and the treatment in most textbooks, the student may get the false impression that “implicit solution” means “An equation that, after I implicitly differentiate, I can rearrange back to the given DE.” That is *not* the definition, however (Definition 3.8 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 3.17 Determine whether the equation

$$2|x| + |y| = 2 \tag{3.41}$$

is an implicit solution of

$$\left(\frac{dy}{dx}\right)^2 = 4|x| + 2|y|. \tag{3.42}$$

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 (3.41) and solve for y in terms of x , we see that (3.41) the graph of (3.41), a “stretched diamond” with vertices at $(\pm 1, 0)$ and $(0, \pm 2)$, consists of the graphs of the following four equations:

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

Therefore (3.41) determines the following four differentiable functions:

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

Every differentiable function of x determined by (3.41), 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 (3.42), we find

$$\begin{aligned}
\text{left-hand side of (3.41)} &\equiv 4, \\
\text{right-hand side of (3.41)} &= 2(2|x| + |y(x)|) \\
&\equiv 2 \times 2 \quad (\text{because of (3.41)}) \\
&= 4.
\end{aligned}$$

Therefore (3.42) is satisfied on the domain of ϕ , for all four choices of ϕ . Both criteria of (3.8) are satisfied, so (3.41) is an implicit solution of (3.42). ■

Example 3.18 Determine whether the equation

$$y^5 + y = x^3 + x \tag{3.43}$$

is an implicit solution of

$$\frac{dy}{dx} = \frac{3x^2 + 1}{5(x^3 + x - y)^{4/5} + 1}. \tag{3.44}$$

First, we observe that the graph of (3.43) has at least one point: the point $(0, 0)$.

Next, we rewrite (3.43) as $F(x, y) = 0$, where $F(x, y) = y^5 + y - x^3 - x$. Then $\frac{\partial F}{\partial y} = 5y^4 + 1$, which is continuous and positive on the whole xy plane. In particular, $\frac{\partial F}{\partial y}$ is continuous and nonzero at $(0, 0)$, so the Implicit Function Theorem guarantees us that (3.43) determines a differentiable function of x near the point $(0, 0)$ on the graph of $F(x, y) = 0$.

So (3.43) determines at least one differentiable function of x . If ϕ is any such function, then substituting $y = \phi(x)$ into (3.43) and differentiating implicitly, we find $(5y^4 + 1)\frac{dy}{dx} = 3x^2 + 1$, which implies

$$\frac{dy}{dx} = \frac{3x^2 + 1}{5y^4 + 1} \tag{3.45}$$

on the domain of ϕ (the denominator $5y^4 + 1$ is never zero). Hence ϕ is a solution of (3.45).

Now, (3.45) does not look like (3.44). The two DEs are not equivalent; there are points (x, y) at which the right-hand side of (3.44) is not equal to the right-hand side of (3.45). But that doesn't mean that (3.43) can't be an implicit solution of (3.44). And, in fact, on the graph of (3.43) 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 (3.44). Therefore (3.43) is an implicit solution of (3.44). ■

In the example above, it is irrelevant whether there are *some* solutions of (3.45) that are not solutions of (3.44). The question was not whether *every* solution of (3.45) was a solution of (3.44), but only whether a *specific* solution of (3.45), namely a function determined implicitly by (3.43), was a solution of (3.44).

Remark 3.19 (Families of implicit solutions) *Every* equation of the form “ $F(x, y) = \text{constant}$ ” 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 “ $F(x, y) = \text{constant}$ ”, namely (3.38). But for any such F and constant C_0 , the DE (3.38) 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.²⁴ However, you are unlikely to find examples like Example 3.17 or Example 3.18 in a DE textbook. In a typical DE course, implicit solutions tend to arise from solving

²⁴*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 that are not elements 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 $\{x^2 + y^2 = c \mid c > 0\}$.” (I do not agree that the term “general solution” needs to be avoided for all nonlinear equations, but if you don’t like the use of “general solution” here, just substitute “the set of all solutions”.) The reason I’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.

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 solutions (not always an exhaustive family, in the separable case) of the form

$$\{F(x, y) = C\}, \quad (3.46)$$

where F is function that depends on the DE, and C is a constant ranging over some subset I of the real line.²⁵ (I.e. for each C in I , the equation $F(x, y) = C$ is an implicit solution of the DE.) *Every* differentiable function implicitly determined by *any* member of the family (3.46) is a solution of the *same* DE, namely (3.38).

3.1.4 Maximal and general solutions of derivative-form DEs

Definition 3.20 Let I be an open 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. This suggests making the following definition:

Definition 3.21 (temporary) For a given three-variable function G , the *general solution* of the differential equation

$$G(x, y, \frac{dy}{dx}) = 0 \quad (3.47)$$

is the collection of all solutions of (3.47), where “solution” is defined as in the second paragraph of Definition 3.1. Said another way, the general solution of (3.47) is the collection of pairs (I, ϕ) , where I is an open interval and ϕ is a solution of (3.47) on I .

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

There is a “redundancy” problem with Definition 3.21 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**

²⁵The subset I is often difficult to specify. However, in typical examples I is an interval, and is sometimes the whole real line.

acceptable to use Definition 3.21 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 3.21 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 3.27 and simply ignore the word “maximal” wherever it appears in these notes.**

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

$$\phi(x) = \frac{1}{x - C} . \tag{3.48}$$

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 (3.48) 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 (3.48) 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 of $(-\infty, C)$ or (C, ∞) whatsoever.

This example illustrates that the collection of pairs (I, ϕ) referred to in Definition 3.21 has a certain redundancy. There is terminology that allows us to speak precisely about this redundancy:

Definition 3.22 Let ϕ be a function on an interval I and let I_1 be a subinterval of I . The *restriction of ϕ to I_1* , denoted $\phi|_{I_1}$, is defined by

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

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

If \tilde{I} is an interval containing I , and $\tilde{\phi}$ is a function on \tilde{I} whose restriction to I is ϕ , then we call $\tilde{\phi}$ an *extension* of ϕ .²⁶

Equivalently: 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

²⁶The same definition applies even when the domains of interest are not intervals; e.g. for a function ϕ with any domain whatsoever, 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.

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

(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”.)

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

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 3.21: the definition names a much larger collection of functions than is needed to capture all the information there is to know about solutions of (3.47). We will see below that we can be more efficient.

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

$$\frac{dy}{dx} = -x^{-2}, \quad y(3) = \frac{1}{3}, \tag{3.49}$$

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 3.23 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 3.24 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 (3.49).
- $\phi_2(x) = \frac{1}{x}$, $2.9 < x < 16.204$, is another solution of $\frac{dy}{dx} = -x^{-2}$, and of the IVP (3.49), 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 (3.49),
- $\phi_4(x) = \frac{1}{x}$, $x \in (0, \infty)$, is a maximal solution of $\frac{dy}{dx} = -x^{-2}$, and is *the* maximal solution of the IVP (3.49).
- $\phi_5(x) = \frac{1}{x}$, $x \in (-\infty, 0)$, is a *different* maximal solution of $\frac{dy}{dx} = -x^{-2}$. It is *not* a solution of the IVP (3.49).
- $\phi_6(x) = \frac{1}{x}$, $x \in (-\infty, -\sqrt{2})$ 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 (3.49).

Example 3.25 The maximal solutions of the differential equation $\frac{dy}{dx} = \sec^2 x$ are the functions ϕ defined by

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

(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.^{27,28} 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 3.21 is this:

Definition 3.26 For a given G , the *general solution* of the DE (3.47) is the collection of all maximal solutions of (3.47).

²⁷Said another way, every solution can be extended to *at least one* maximal solution. Maximal extensions always exist, but they are not always unique.

²⁸*Note to instructors:* Existence of a maximal extension is not so obvious, absent any conditions that ensure local uniqueness of solutions. However, it 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.

(This definition supersedes Definition 3.21.)

Example 3.24 demonstrates, we hope, the economy gained by including the word “maximal” in this definition. The student will probably agree that, even prior to writing down Definition 3.26, 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 3.27 The general solution of $\frac{dy}{dx} = x$ may be written in short-hand as

$$\left\{ y = \frac{1}{2}x^2 + C \right\}. \quad (3.50)$$

In this context equation (3.50) 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 (3.27); they tell us we’re talking about a *set* of objects of the form within the braces.) We allow ourselves to write (3.50) as short-hand for “the collection of functions $\{\phi_C \mid C \in \mathbf{R}\}$, where $\phi_C(x) = \frac{1}{2}x^2 + C$ ”.²⁹

Example 3.28

- The general solution of

$$\frac{dy}{dx} = -x^{-2}, \quad x > 0 \quad (3.51)$$

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

$$\left\{ y = \frac{1}{x} + C \right\}, \quad x > 0, \quad (3.52)$$

a one-parameter family of maximal solutions. Because the restriction $x > 0$ is stated explicitly in (3.51), it is permissible to leave out the “ $x > 0$ ” when writing the general solution; we may simply write the general solution as

$$\left\{ y = \frac{1}{x} + C \right\} \quad (3.53)$$

- The general solution of

$$\frac{dy}{dx} = -x^{-2}, \quad (3.54)$$

²⁹Students in my own classes are permitted to omit the curly braces in (3.50), but I am trying to maintain certain notational consistency across different sections of these notes.

with no interval specified, may also be written as (3.53)—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 (3.54) was written down, we must consider all possible intervals. Therefore, in this context, equation (3.53) does *not* represent a one-parameter family of maximal solutions; it represents *two* one-parameter families of maximal solutions³⁰. Equation (3.53) is acceptable short-hand for

$$\left. \begin{array}{l} \text{the union of the two families of functions} \\ \\ \text{where} \\ \text{and} \end{array} \right\} \begin{array}{l} \{\phi_C \mid C \in \mathbf{R}\}, \quad \{\psi_C \mid C \in \mathbf{R}\} \\ \\ \phi_C(x) = \frac{1}{x} + C, \quad x > 0 \\ \\ \psi_C(x) = \frac{1}{x} + C, \quad x < 0. \end{array} \quad (3.55)$$

(The *union* of the two families means the collection of functions that are in one family or the other.³¹) The solution $y = \frac{1}{x} + 6$ on $\{x < 0\}$ (the function ψ_6 in the notation of (3.55)) 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

$$\left\{y = \frac{1}{x} + C, x > 0\right\} \quad \text{and} \quad \left\{y = \frac{1}{x} + C, x < 0\right\} \quad (3.56)$$

³⁰ Many calculus textbooks, and especially integral tables, foster a misunderstanding of the indefinite integral. *By definition*, for functions f that are continuous on an open interval or a union of disjoint open intervals, “ $\int f(x)dx$ ” means “the collection of all antiderivatives of f ”. If the implied domain of f is an open interval, then this collection is the same as the general solution of $dy/dx = f(x)$. But we must be careful not to interpret formulas such as “ $\int x^{-2} dx = -x^{-1} + C$ ” or “ $\int \sec^2 x dx = \tan x + C$ ” as saying that every antiderivative of x^{-2} is of the form $x^{-1} + C$ *on the whole implied domain of the integrand* x^{-2} , or that every antiderivative of $\sec^2 x$ is of the form $\tan x + C$ *on the whole implied domain of the integrand* $\sec^2 x$.

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

³¹*Note to instructors:* In these notes I have opted not to use the symbol \cup for union of sets of functions, out of concern that this might confuse some students. You will notice later on, e.g. in (3.56), 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 (3.56) and similar expressions this way. I felt that using the word “or” in these expressions would be confusing to students.

and

$$\left\{y = \frac{1}{x} + C_1, x > 0\right\} \text{ and } \left\{y = \frac{1}{x} + C_2, x < 0\right\}.$$
 ³² (3.57)

In (3.56), 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 (3.57), 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 (3.53), (3.56), and (3.57) 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.

Example 3.29 The general solution of $\frac{dy}{dx} = \sec^2 x$ may be written as

$$\{y = \tan x + C\},$$
 (3.58)

or as

$$\left\{y = \tan x + C, \left(n - \frac{1}{2}\right)\pi < x < \left(n + \frac{1}{2}\right)\pi, \text{ } n \text{ an integer}\right\},$$
 (3.59)

or as

$$\left\{y = \tan x + C_n, \left(n - \frac{1}{2}\right)\pi < x < \left(n + \frac{1}{2}\right)\pi, \text{ } n \text{ an integer}\right\},$$
 (3.60)

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 (3.58)–(3.60) (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 $\left(n - \frac{1}{2}\right)\pi < x < \left(n + \frac{1}{2}\right)\pi$ (where n is an arbitrary integer).

Example 3.30 The general solution of the separable equation

$$\frac{dy}{dx} = -y^2$$
 (3.61)

may be written as

³²In both (3.56) and (3.57), where we combine two or more sets of solutions using the word “and”, the union-symbol \cup would be more precise.

$$\left\{ y = \frac{1}{x - C} \right\} \text{ and } \{y = 0\}, \quad (3.62)$$

or in various other ways that impart the same information³³. In the given context, the solution that is the constant function 0 may be written as “ $y = 0$ ”, as in (3.62) or as “ $y \equiv 0$ ” (which, in this context, is read “ y identically zero”). Since a solution of (3.61), expressed in terms of the variables in (3.61), is function of x , the only correct interpretation of “ $y = 0$ ” in (3.62) is “ y is the constant function whose value is zero for all x ”, *not* “ y is a real number, specifically the number 0”. An instructor may sometimes write a constant function using the identically-equal-to symbol “ \equiv ”, especially in the early weeks of a DE course, to make sure that students are absolutely clear what is meant; at other times, when there is little possibility of confusion, (s)he may just use the ordinary “ $=$ ” symbol.

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

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 (3.61):

1. The domain of every maximal solution is different from the domain of every other.
2. For every interval of the form (a, ∞) there is a maximal solution whose domain is that interval, namely $y = \frac{1}{x-a}$.
3. For every interval of the form $(-\infty, a)$ there is a maximal solution whose domain is that interval, namely $y = \frac{1}{x-a}$. (The *formula* is the same as for solution on (a, ∞) mentioned above, but we stress again that the fact that *as solutions of a differential equation*, “ $y = \frac{1}{x-a}$, $x > a$ ” and “ $y = \frac{1}{x-a}$, $x < a$ ” are *completely unrelated* to each other.)
4. There is one maximal solution whose domain includes the domain of every other, namely $y \equiv 0$.

The general solution of (3.61) also exhibits another interesting phenomenon. The way we have written the general solution in (3.62) isolates the maximal solution $y \equiv 0$ as not belonging to what appears to be a single nice family into which the other maximal solutions fall (there is no value of C for which the formula “ $y = \frac{1}{x-C}$ ” produces the constant function 0). But for $C \neq 0$, writing $K = \frac{1}{C}$,

³³We do not discuss here how to *figure out* the general solution of this DE, since that is adequately covered outside these notes.

$$\frac{1}{x - C} = \frac{C^{-1}}{C^{-1}x - 1} = \frac{K}{Kx - 1} . \quad (3.63)$$

In the right-most formula in (3.63), 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 (3.62). Thus, we can rewrite the general solution (3.62) as

$$\left\{ y = \frac{K}{Kx - 1} \right\} \quad \text{and} \quad \left\{ y = \frac{1}{x} \right\} . \quad (3.64)$$

Here, K is an arbitrary constant, allowed to assume all real values, just as C was allowed to in (3.62); we could just as well use the letter C for it. Writing the general solution as in (3.64), the two solutions with formula $y = \frac{1}{x}$ (one for $x > 0$, one for $x < 0$) may be viewed as the exceptional ones, with all the others—including the constant function 0—falling into the “ $\frac{K}{Kx-1}$ ” family. This illustrates that there be more than one way of expressing the collection of all maximal solutions as what looks like a “nice family” containing most of the maximal solutions, plus one or more maximal solutions that don’t fall into the family. This illustrates that “*falling into a family*” can be in the eye of the beholder, and not something intrinsic to a solution of a DE.

But this example also provides another instance of a theme to which we keep returning: how easy it is to mis-identify a family of *formulas* with a family of *solutions of a DE*. The maximal solutions described by $\{y = \frac{1}{x-C}\}$ in (3.62) do not form *one* one-parameter family of solutions; they form *two*.³⁴ Every value of C corresponds to two maximal solutions, one defined to the left of C and one defined to the right³⁵. In (3.64), 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 (3.62) is preferable to (3.64) 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

³⁴This 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.

³⁵*Note to instructors:* Of course, the constant solution 0 may be viewed as the “ $C = \infty$ ” case of “ $y = \frac{1}{x-C}$ ”, and you may even wish to tell your students that. However, this does *not* mean that the general solution is a one-parameter family parametrized by the one-point compactification of \mathbf{R} , i.e. the circle. Such a conclusion would be fine if we were talking the one-parameter family of *rational functions* defined by “ $y = \frac{1}{x-C}$ ”, but we are not; we are talking about solutions of an ODE, for which the *only* sensible domain is a connected one.

two maximal solutions, while in the second collection there is a value of the arbitrary constant, namely 0, for which the given formula defines only one maximal solution. However, in the next example, we will see two different ways of writing the general solution, neither of which can be preferred over the other by any such considerations.

Example 3.31 The general solution of the separable equation

$$\frac{dy}{dx} = y(1 - y) \quad (3.65)$$

may be written as

$$\left\{ y = \frac{C}{e^{-x} + C} \right\} \quad \text{and} \quad \{y \equiv 1\}. \quad (3.66)$$

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

$$\left\{ y = \frac{1}{Ce^{-x} + 1} \right\} \quad \text{and} \quad \{y \equiv 0\}. \quad (3.67)$$

(Here, the analog of the previous example's K has been renamed to C .) In each case, in the family in curly braces, the formula giving $y(x)$ yields two maximal solutions for $C < 0$ and one maximal solution for $C \geq 0$. The $C = 0$ solution in (3.66) is the constant function 0, which is the “exceptional” solution in (3.67). The $C = 0$ solution in (3.67) is the constant function 1, which is the “exceptional” solution in (3.66). The situation is completely symmetric; neither of (3.66) and (3.67) 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 “nice” linear DEs (the meaning of “nice” is not important right now), and not to define the term at all for nonlinear DEs.³⁶

³⁶*Note to instructors:* I feel, however, that too much is lost this way. It is important for students to be able to know when they've found all (maximal) solutions, whether expressed explicitly or implicitly. 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,

3.1.5 General and implicit solutions on a region in \mathbf{R}^2

For a derivative-form DE, so far we have defined “general solution on an interval I ” and “general solution” with no interval specified. There is another type of general solution (of a derivative-form DE) that will be used later in these notes. To talk about it, we first must define what an *open set* is in the xy plane. A subset R of \mathbf{R}^2 is *open* if for every point (x_0, y_0) in R , there is some open rectangle that contains (x_0, y_0) and is contained in R . Another term we will use for “open subset of \mathbf{R}^2 ” is *region*³⁷.

In these notes, we will be making quite a bit of use of the fundamental Existence and Uniqueness Theorem for first-order initial-value problems (Theorem 5.1) and some of its corollaries. **Henceforth, to avoid long wording, “the Fundamental Theorem” will mean Theorem 5.1**, unless followed by the words “of Calculus”. (We will sometimes mention the Fundamental Theorem of Calculus, which you learned in Calculus 1, but when we say just “the Fundamental Theorem”, we mean Theorem 5.1.)

The derivative-form DEs about which we can draw the most systematic conclusions are those that satisfy the conditions of the Fundamental Theorem 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 $\{(x, y) \in \mathbf{R}^2 \mid y > 0\}$ and on the region $\{(x, y) \in \mathbf{R}^2 \mid y < 0\}$, but not on all of \mathbf{R}^2 , or on any region that includes a point of the form $(x, 0)$.) For this reason we define *solution in a region*, and some related objects, below.

Definition 3.32 (Solution, and solution curve, in a region) Let R be a region in the xy plane. Given a differential equation $G(x, y, \frac{dy}{dx}) = 0$, 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 $y = \phi(x), x \in I$) is contained in R . We call the graph of a solution in R is called a *solution curve in R* .

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

Of course, you (the instructor) may have a different convention that you prefer for use of the term “general solution”. One convention I advise against, however, is to use “general solution” to refer to a *non-exhaustive* collection of solutions (or for a “typical” element of such a collection) for which (s)he has produced a nicely-parametrized family of formulas. As the simple examples 3.30 and 3.31 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.

³⁷*Note to instructors:* I am taking some liberties here. The usual definition of “region” is *connected* non-empty open subset. I did not want to distract the student with a definition of *connected*, and felt that the student would understand from context that when “an open set in \mathbf{R}^2 ” is referred to in these notes, it is understood that the set is non-empty.

The next definition generalizes Definition 3.23. If you find Definition 3.33 (and, consequently, Definition 3.37) hard to understand, just **skim from here through Definition 3.37 for now**, and re-read this material after you've read Section 3.1.6.

Definition 3.33 (Maximal solution, and maximal solution curve, in a region)

Let R be a region in the xy plane. Given a differential equation $G(x, y, \frac{dy}{dx}) = 0$, 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 . To avoid some 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 . For the same reason, we use the term *maximal solution curve in R* for the graph of a maximal solution in R .

The Fundamental Theorem has some very important and useful corollaries that can be phrased very simply in terms of maximal solutions, and maximal solution curves, in a region. Three of these, very closely related, are as follows.

Corollary 3.34 *Let f be a function of two variables (x, y) , and assume that f and $\frac{\partial f}{\partial y}$ are continuous on an open set R in \mathbf{R}^2 . Then for every (x_0, y_0) in R , the initial-value problem*

$$\frac{dy}{dx} = f(x, y), \quad y(x_0) = y_0 \tag{3.68}$$

has a unique solution that is maximal in R . This maximal solution ϕ_{\max} has the property that every solution of the initial-value problem (3.68) in R is a restriction of ϕ_{\max} . ■

Corollary 3.35 *Hypotheses as in Corollary 3.34. Then every point (x_0, y_0) in R lies on a unique maximal solution curve in R . ■*

Corollary 3.36 *Hypotheses as in Corollary 3.34. Then no two (distinct) maximal solution curves in R can intersect. ■*

The essence of all of these corollaries is that, under the given hypotheses, if two (not necessarily maximal) solution curves in R intersect, say at (x_0, y_0) , we would obtain a contradiction to the Fundamental Theorem unless both of these curves are part of the same maximal solution curve in R (in which case the curves don't simply *intersect*, they *overlap*). Said another way: if ϕ_1, ϕ_2 are two solutions of the IVP (5.1) in R , with ϕ_1 having domain I_1 and with ϕ_2 having domain I_2 , then ϕ_1 and ϕ_2 are identically equal on the intersection of I_1 and I_2 .

Definition 3.37 (General solution in a region) Let R be a region in the xy plane. The *general solution, in R* , of a differential equation $G(x, y, \frac{dy}{dx}) = 0$, is the collection of all maximal solutions in R . ■

Thus the general solution of a DE, as defined in Definition 3.26, is the same as the general solution of that DE in the region \mathbf{R}^2 .

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. (3.69)

Statement (3.69) 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 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 f_y in Theorem 5.1 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.

Sometimes we can write down an explicit formula for every solution of a DE. In this case, we usually express the general solution as a collection of “explicit solutions” (see Definitions 3.3 and 3.2), as we did in Examples 3.27–3.31. (We did not fully *justify* that we'd found all the solutions; at that point we simply wanted to illustrate the concept of “general solution” in several examples with which students were likely to be familiar.) There are other times when we can find all the solutions in implicit form, but don't have explicit formulas for them. 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 given implicitly by the equation $F(x, y) = C$ for some constant C , and conversely for which every such equation 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.

In all cases, when we want to *write down* a general solution of a DE, we almost always do so (or attempt to do so) by writing down a *collection of equations in x*

and y .³⁸ 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 , one of which may be zero or some other constant. 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 $F = F_1 - F_2$.

Definition 3.38 (General solution on a region, in implicit form)³⁹ For a given three-variable function \mathbf{G} , consider the derivative-form DE

$$\mathbf{G}\left(x, y, \frac{dy}{dx}\right) = 0. \quad (3.70)$$

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 (3.70) 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 (3.70) (see Definitions 3.8 and 3.13).
- (ii) For every maximal solution ϕ of (3.70) in R , the relation $y = \phi(x)$ satisfies *one and only one* of the equations in \mathcal{E} .⁴⁰ Equivalently, every maximal solution curve of (3.70) in R lies in the graph of one and only one equation in \mathcal{E} .

Alternatively, we refer to such a collection \mathcal{E} as *an implicit form of the general solution of (3.70) on R* .

If each of the equations in \mathcal{E} is, more strongly, an *explicit* solution (recall observation (3.29)), we may call \mathcal{E} *an explicit form of the general solution of (3.70)*. However, in accordance with convention (3.3), we will usually call \mathcal{E} simply *the general solution of (3.70) in R* , understanding that the meaning of “general solution” in this phrase is not identical to the meaning in Definition 3.37.

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 equation (3.70) makes sense: the set $\{(x, y) \in \mathbf{R}^2 \mid \text{the expression } \mathbf{G}(x, y, z) \text{ is defined for some real number } z\}$.⁴¹ ■

³⁸A *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.

³⁹The terminology in this definition was invented purely for these notes; it is not standard.

⁴⁰See Definition 3.5 for this meaning of “satisfies”.

⁴¹For 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 \mathbf{G} , this set might not be open, in which case it would not fit our definition of *region*. However, under extremely mild conditions on \mathbf{G} , satisfied by all DEs we consider in these notes, this set *will* be open.

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 3.38. 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 “on a region R ”). However, there are several 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 is *not* the general solution (in implicit form), but for which the full general solution, in implicit form, can be constructed in a systematic way. We will illustrate this for separable DEs later.

A cautionary note: Do not be misled by the terminology “the general solution of (3.70) in R , *in implicit form*.” While there is only one general solution of (3.70) 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 $\{F(x, y) - C = 0\}$, 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 *explicit* ways of expressing general solutions in Examples 3.30 and 3.31.

Remark 3.39 Suppose that \mathcal{E} is an implicit form of the general solution of a given DE. Condition (ii) in Definition 3.38 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} . ■

3.1.6 Algebraic equivalence of derivative-form DEs

Some algebraic manipulations that help us to 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.⁴² In this section of the notes, we discuss how to be aware of and deal with this problem.

Definition 3.40 We say that two derivative-form differential equations, with independent variable x and dependent variable y , are *algebraically equivalent on a region* R if one equation can be obtained from the other by the operations of (i) adding to both sides of the equation an expression that is defined for all $(x, y) \in R$ ⁴³, 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 3.40, even though they are not mentioned explicitly.

Example 3.41 The differential equations

$$\frac{dy}{dx} = y(1 - y) \tag{3.71}$$

and

$$\frac{1}{y(1 - y)} \frac{dy}{dx} = 1 \tag{3.72}$$

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 3.42 The differential equations

$$(x + y) \frac{dy}{dx} = 4x - 2y \tag{3.73}$$

⁴²Unfortunately, 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.

⁴³*Note to students:* The expression is allowed to involve $\frac{dy}{dx}$ —i.e. it could be of the form $G(x, y, \frac{dy}{dx})$ for some three-variable function G —which is why we did not say “function of x and y ” here. If the expression is $G(x, y, \frac{dy}{dx})$, our requirement that it be “defined for all $(x, y) \in \mathbf{R}^2$ ” is short-hand for: for each $(x, y) \in R$ there is *some* real number z such that $G(x, y, z)$ is defined.

and

$$\frac{dy}{dx} = \frac{4x - 2y}{x + y} \quad (3.74)$$

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.⁴⁴ This happens, for example, if we start with the equation $x^3 - 3x^2 = -2x$ and divide by x , obtaining $x^2 - 3x = -2$. In this example, we lose the solution 0. (The solution set of the first equation is $\{0, 1, 2\}$, while the solution set of the second is just $\{1, 2\}$.) For another example, if start with the equation $\sqrt{x+4} = -3$, and square both sides, we obtain $x+4 = 9$, and hence $x = 5$. But 5 is not a solution of the original equation; $\sqrt{5+4}$ is 3, not -3 . Our manipulation has introduced a “spurious solution”, a value of x that is a solution of the post-manipulation equation that we may mistakenly *think* is a solution of the original equation, when in fact it is not.

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

$$\left. \begin{array}{l} \text{If two differential equations are algebraically equivalent on a} \\ \text{region } R, \text{ then they have the same general solution on } R. \end{array} \right\} \quad (3.75)$$

We may restate (3.75) 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 (3.75).

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.

⁴⁴Usually 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.

But usually, we maintain algebraic equivalence on regions that fill out most of the xy plane, as in Examples 3.41 and 3.42 above.

To see why statement (3.75) is true, let us check that operation (ii) in Definition 3.40 does not change the set of solutions whose graphs lie in R . Let us suppose we start with a (first-order) derivative-form DE of the most general possible form:

$$\mathbf{G}_1\left(x, y, \frac{dy}{dx}\right) = \mathbf{G}_2\left(x, y, \frac{dy}{dx}\right). \quad (3.76)$$

(Of course, by subtracting $\mathbf{G}_2\left(x, y, \frac{dy}{dx}\right)$ from both sides, we can put this in the simpler form $\mathbf{G}\left(x, y, \frac{dy}{dx}\right) = 0$, but since we often perform manipulations on equations without first putting them in the simple form (3.2), we will illustrate the solution-set-doesn't-change principle for DEs that have not been put in that form.) The equation obtained by multiplying both sides of (3.76) by a function h that is defined at every point of R and is nonzero on R is

$$h(x, y)\mathbf{G}_1\left(x, y, \frac{dy}{dx}\right) = h(x, y)\mathbf{G}_2\left(x, y, \frac{dy}{dx}\right). \quad (3.77)$$

Suppose that ϕ is a solution of (3.76). Then for all x in the domain of ϕ ,

$$\mathbf{G}_1\left(x, \phi(x), \phi'(x)\right) = \mathbf{G}_2\left(x, \phi(x), \phi'(x)\right). \quad (3.78)$$

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 (3.78) by this number. Therefore

$$h(x, \phi(x))\mathbf{G}_1\left(x, \phi(x), \phi'(x)\right) = h(x, \phi(x))\mathbf{G}_2\left(x, \phi(x), \phi'(x)\right) \quad (3.79)$$

for all x in the domain of ϕ . Hence ϕ is a solution of (3.77). Thus every solution of (3.76) whose graph lies in R is also a solution of (3.77) whose graph lies in R .

Conversely, suppose that ϕ is a solution of (3.77) whose graph lies in R . Then (3.79) 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 (3.79) by this number. Therefore (3.78) is satisfied for all x in the domain of ϕ , so ϕ is a solution of (3.76). Thus every solution of (3.77) whose graph lies in R is also a solution of (3.76) whose graph lies in R .

This completes the argument that multiplying by h has not changed the set of solutions whose graphs lie in R . The argument that operation (i) in Definition 3.40 does not change this set of solutions is similar, and is left to the student.

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 3.40 does not *always* change the set of

solutions.⁴⁵ But because they *might* change the set of solutions, any time we perform one of these “unsafe” operations we must check, by some other method, that we properly account for any lost solutions or spurious solutions.

Students should already be familiar with this fact from their experience with separable equations. For example, in passing from equation (3.71) to (3.72), we potentially lose any solution whose graph intersects the horizontal line $\{y = 0\}$ or the horizontal line $\{y = 1\}$. Are there any such solutions? Yes: the two constant solutions $y \equiv 0$ and $y \equiv 1$, whose graphs happen to be exactly these two horizontal lines.

When we are dealing with separable equations $\frac{dy}{dx} = g(x)p(y)$, and there is any number r for which $p(r) = 0$, when we separate variables we don’t just *potentially* lose solutions, we *always* lose solutions (unless we make an error later in the process). For every number r for which $p(r) = 0$, the constant function given by the equation $y = r$ is a solution that separation of variables, carried out with no errors, *cannot* find⁴⁶. But fortunately, it finds all the others (in implicit form).

We will establish this general fact in the next (sub)section, but we can already see why it’s true for the DE (3.71). The right-hand side of (3.71) is a function of $(x$ and) y that is continuous everywhere and whose partial derivative with respect to y is continuous everywhere. Therefore for *every* initial-condition point (x_0, y_0) in the xy plane, the Fundamental Theorem (Theorem 5.1) applies, and so through each such point there is the graph of one and only one maximal solution. If there were a non-constant solution of (3.71) whose graph intersected the graph of the constant solution $y \equiv 1$ (the line $\{y = 1\}$), say at the point $(x_0, 1)$, we would have a contradiction to uniqueness of the solution of the IVP with differential equation (3.71) and with initial condition $y(x_0) = 1$. Similarly, no non-constant solution of (3.71) can have a graph that intersects the graph of the constant solution $y \equiv 0$ (the line $\{y = 0\}$). Therefore the graph of every non-constant solution lies entirely in one of the three regions mentioned in Example 3.41. Since equations (3.71) and (3.72) are algebraically equivalent on each of these three regions, the general solution of (3.72) is precisely the set of all solutions of (3.71) other than the two constant solutions that we have already accounted for.

Thus, if we manage to solve (3.72)—which we leave the student to do—and then add to its general solution the two constant functions $y \equiv 0$ and $y \equiv 1$, we obtain all solutions of (3.71).

⁴⁵For example, it can be shown that the DEs (3.73) and (3.74) have the same general solution, of which an implicit form is the family of equations $\{(y + 4x)^3(y - x)^2 = C\}$.

⁴⁶Unfortunately it is quite common, even in textbooks, to make a pair of canceling errors in solving separable equations, leading to the false impression that the separation-of-variables procedure of variables may lose only *some* of the constant solutions, when in fact it *always* loses *all* of them if no mistakes are made.

3.1.7 General solutions of separable DEs

Consider any separable DE

$$\frac{dy}{dx} = g(x)p(y), \quad (3.80)$$

for which

$$\left. \begin{array}{l} g \text{ is continuous on some open interval } I, \text{ and} \\ g \text{ is not identically zero on } I, \text{ and} \\ p \text{ and } p' \text{ are continuous on some open set } D \text{ in } \mathbf{R}. \end{array} \right\} \quad (3.81)$$

There is a redundancy in the third line of (3.81): 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 possible about solutions of the DE (3.80) under hypotheses of the form (3.81). 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 3.43 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 3.34 apply to the initial-value problem for (3.80) with initial condition $y(x_0) = y_0$, and Corollaries 3.35 and 3.36 apply to the DE (3.80) 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 $p(\phi(x))g(x) = p(r)g(x) = 0 \times g(x) = 0$ for all $x \in I$. Hence the constant function ϕ , with domain I , is a solution of equation (3.80) 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 3.36) no other maximal solution curve in R can intersect. Therefore if $y_0 \neq r$, and ϕ a solution of the IVP for (3.80) 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 (3.80) on I .

Combining the preceding facts:

- For each r in D , the equation $y = r$ is a (constant) solution of (3.80) on I if $p(r) = 0$, and is not a solution of (3.80) on I if $p(r) \neq 0$ (cf. Remark 3.4).

- If ϕ is a *non-constant* solution of (3.80) 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.

Now let Z denote the set $\{r \in D \mid p(r) = 0\}$ (the set of *zeroes* of p ; if p is a polynomial these numbers are also called *roots* of p), and 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 $p(x_0) \neq 0$, then $p(x) \neq 0$ for all x sufficiently close to x_0 .

On the domain D_1 , define a function h by $h(y) = \frac{1}{p(y)}$. Then 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.⁴⁷ Let H be any fixed antiderivative of h on D_1 , and let G be any fixed antiderivative of g on I .

Theorem 3.43 *Assume the hypotheses (3.81) are satisfied. Then the general solution of (3.80) on the region $R = I \times D$, in implicit form, is the collection*

$$\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2, \quad (3.82)$$

where

$$\mathcal{E}_1 = \{H(y) = G(x) + C \mid C \in \mathbf{R}\} \quad \text{and} \quad \mathcal{E}_2 = \{y = r \mid r \in Z\}. \quad (3.83)$$

(Note that the set Z may be empty, and therefore so may the collection \mathcal{E}_2 .) The collection \mathcal{E}_1 is precisely the set of all non-constant solutions, in implicit form, of (3.80), 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 “ \cup ” in (3.82) denotes *union*: the collection \mathcal{E} consists of equations that lie either in the collection \mathcal{E}_1 or the collection \mathcal{E}_2 .

Proof of Theorem 3.43. We have already established that \mathcal{E}_2 is the set of constant maximal solutions of (3.80) on I , and that if ϕ is any non-constant solution

⁴⁷The 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.

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 Z is in the domain of H .

Let ϕ be a non-constant solution of equation (3.80) in R , with domain I_1 . 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

$$\frac{1}{p(\phi(x))} \phi'(x) = g(x) \quad (3.84)$$

Hence, on the interval I_1 we have

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

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

$$H(\phi(x)) = G(x) + C, \quad (3.85)$$

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

$$H(y) = G(x) + C. \quad (3.86)$$

This establishes that

$$\left. \begin{array}{l} \text{the graph of every non-constant solution} \\ \text{of (3.80) in } R \text{ lies in the graph of one} \\ \text{of the equations in the collection } \mathcal{E}_1. \end{array} \right\} \quad (3.87)$$

Next, we claim that

$$\left. \begin{array}{l} \text{for each } C \in \mathbf{R} \text{ for which the graph of equation (3.86) in } R \\ \text{contains at least one point, there is an implicitly defined} \\ \text{function of } x \text{ determined by this equation (see Definition 3.7).} \end{array} \right\} \quad (3.88)$$

To see this, note that the graph of equation (3.86) 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 .

Conversely, suppose ϕ is a differentiable function of x , with domain an open interval I_1 , that is determined implicitly by one of the equations in \mathcal{E}_1 . Then, for

some constant C , (3.85) 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 (3.85) with respect to x , we obtain $H'(\phi(x))\phi'(x) = G'(x)$, implying that equation (3.84) holds on I_1 , hence that $\phi'(x) = p(\phi(x))g(x)$ on I_1 . Thus ϕ is a solution of (3.80). This establishes that

$$\left. \begin{array}{l} \text{every differentiable function of } x \text{ implicitly determined} \\ \text{by equation (3.86) is a solution of the DE (3.80).} \end{array} \right\} \quad (3.89)$$

Together, facts (3.88) and (3.89) imply that for each $C \in \mathbf{R}$ for which the graph of equation (3.86) in R has any points, equation (3.86) is an implicit solution of (3.80) (see Definitions 3.8 and 3.13). Combining this with fact (3.87) 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 (3.80) on R , in implicit form (Definition 3.38).

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 the graphs of two equations in \mathcal{E}_2 can't intersect each other. Thus, by Remark 3.39, every solution curve of (3.80) in R , whether maximal or not, lies in the graph of a unique equation in \mathcal{E} . This completes the proof of Theorem 3.43. ■

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 (3.80) is *not* algebraically equivalent, on R , to the DE

$$\frac{1}{p(y)} \frac{dy}{dx} = g(x) \quad (3.90)$$

that arises in the process of separating variables in equation (3.80). 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 of (3.80) and (3.90) on $I \times D_1$. Thus, assuming the conditions (3.81) are met, separation of variables always finds every non-constant solution (in implicit form), and yields no “spurious solutions”—equations that are not solutions (implicit or explicit)—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 3.43 more closely, looking to see where all the assumptions in (3.81) 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 3.43, a

fact that we used the hypotheses (3.81) 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 assumption on p' in (3.81) somewhat without altering the conclusion of Theorem 3.43.

But suppose we weaken the conditions (3.81) 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 3.43 no longer is valid, but the first bullet-point is, and the only parts of the proof of Theorem 3.43 that become invalid are those that relied on that second bullet-point stated. Thus, our argument proves the following more general theorem:

Theorem 3.44 *In the setting of the DE (3.80), assume that the first two hypotheses in (3.81) are met, and assume that the function p is continuous on some open set D . Again let $Z = \{r \in D \mid p(r) = 0\}$, and $D_1 = \{y \in D \mid p(y) \neq 0\}$. With all other notation as in Theorem 3.43, the following are true:*

1. \mathcal{E}_1 is the general solution of the DE (3.80) 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 (3.80) in $I \times D$.
4. Every solution curve of (3.80) on $I \times D$ is contained in a union of graphs of equations in the collection $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$.



What Theorem 3.44 does *not* assert, unlike Theorem 3.43, 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 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 3.45 Consider the DE

$$\frac{dy}{dx} = 6x(y - 2)^{2/3}. \quad (3.91)$$

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 $D_1 = \{y \in \mathbf{R} \mid y \neq 0\} = (-\infty, 2) \cup (2, \infty)$; in particular, p' is not continuous on \mathbf{R} .

Observe that in $\mathbf{R} \times D_1$, equation (3.91) is algebraically equivalent to the DE

$$(y - 2)^{-2/3} \frac{dy}{dx} = 6x \quad (3.92)$$

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 3.44 that the general solution of (3.92) on $\mathbf{R} \times D_1$, in implicit form, is

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

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 ϕ_1, \dots, ϕ_7 defined below are solutions of equation (3.92), but only for ϕ_1 does the solution-curve lie in the graph of an equation in $\mathcal{E}_1 \cup \mathcal{E}_2$; each of the other solution-curves lies only in a union of two or more such graphs (see Figures 4 and 5).

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

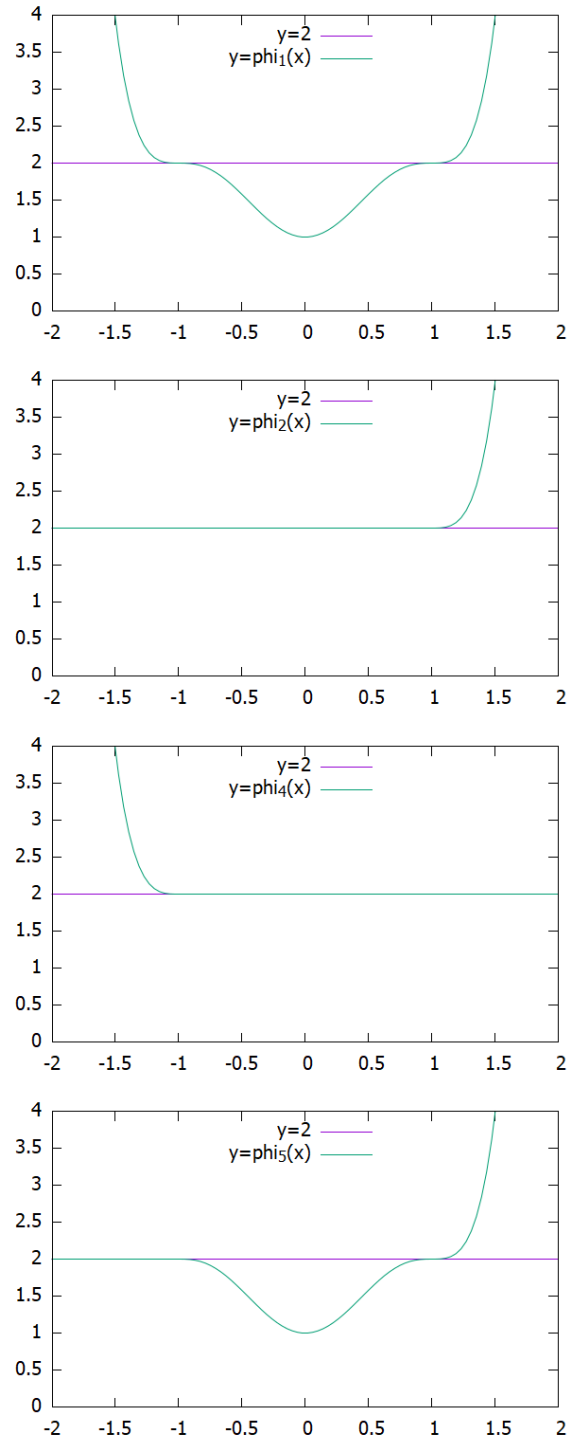


Figure 4: For Example 3.45: The solutions ϕ_1, ϕ_2, ϕ_4 , and ϕ_5 , plotted simultaneously with the constant solution 2. The coordinate axes (not shown) are the usual x and y axes. The graph of ϕ_1 intersects the graph of the constant solution, but does not overlap it. The graphs of ϕ_2, ϕ_4 , and ϕ_5 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 .

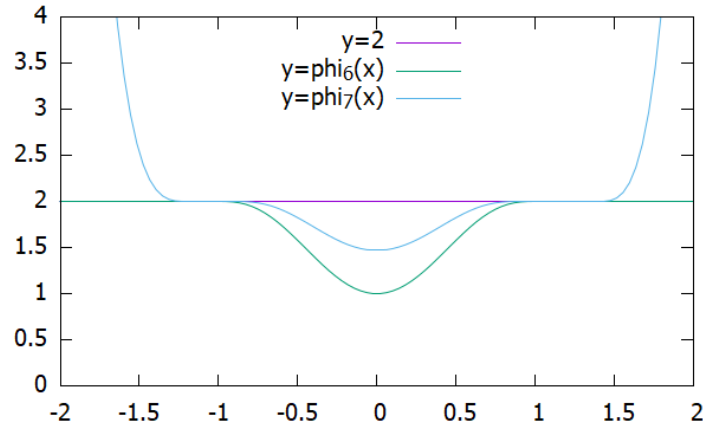


Figure 5: For Example 3.45: The solutions ϕ_6 and ϕ_7 , plotted simultaneously. The two graphs overlap for $-1.2 \leq x \leq -1$ and for $1 \leq x \leq 1.4$. The solution curve $y = 2$ (not shown) overlaps the graph of ϕ_6 for $x \leq -1$ and for $x \geq 1$, and overlaps the graph of ϕ_7 for $-1.2 \leq x \leq -0.9$ and for $0.9 \leq x \leq 1.4$.

The solution $y = 2$ of the DE (3.91) 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 all examples discussed previously, there were no singular solutions. *For separable DEs, the conditions (3.81) 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 (3.91) 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. **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 hypotheses on f in the Fundamental Theorem’s hypotheses are met. This is one reason that the Fundamental Theorem is so important. ■

In Example 3.45, 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 “ $2, x \leq -0.9$ ” or replace the bottom two lines by the single line “ $2, x \geq 0.9$ ”), or four (as exemplified by ϕ_7). The bookkeeping is laborious, but it can be done. The 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. We will put a name to this⁴⁸:

Definition 3.46 For a given three-variable function G , and region R in the xy plane, we call a collection \mathcal{E} of algebraic equations in x and y a *piecewise generating set of equations for the general solution of $G(x, y, \frac{dy}{dx}) = 0$ in R* if the following two conditions are satisfied:

- (i) Each equation in \mathcal{E} , restricted to R , is an implicit solution of $G(x, y, \frac{dy}{dx}) = 0$.
- (ii) For every solution ϕ of $G(x, y, \frac{dy}{dx}) = 0$ in R , and every point x_0 in the domain of ϕ , there is a closed interval I_{x_0} containing x_0 such that on the interval I_{x_0} , the relation $y = \phi(x)$ satisfies one and only one of the equations in \mathcal{E} .

If each of the equations in \mathcal{E} is an *explicit* solution, we may call \mathcal{E} a *piecewise generating set of equations for the general solution of $G(x, y, \frac{dy}{dx}) = 0$ on R* .

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 equation $G(x, y, \frac{dy}{dx}) = 0$ makes sense (see the end of Definition 3.38). ■

Example 3.47 Under the hypotheses of Theorem 3.44, the collection $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$, where \mathcal{E}_1 is the set of implicit solutions found by separating variables and \mathcal{E}_2 is the set of constant solutions, is piecewise generating set of equations for the general solution of $y' = g(x)p(y)$ on the region $I \times D$. If p' is continuous on D , then \mathcal{E} itself is an implicit form of the general solution on $I \times D$. ■

Separable DEs do not always come to us in the standard form (3.80):

Example 3.48 Consider the differential equation

$$x \frac{dy}{dx} = \sin y. \quad (3.93)$$

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 3.43 and 3.44 apply.

⁴⁸The terminology in Definition 3.46 is not standard; it was invented for these notes.

However, on the regions $R_1 = (0, \infty) \times \mathbf{R} = \{(x, y) \in \mathbf{R} \mid x > 0\}$ and $R_2 = (-\infty, 0) \times \mathbf{R} = \{(x, y) \in \mathbf{R} \mid x < 0\}$ equation (3.93) and the DE

$$\frac{dy}{dx} = \frac{\sin y}{x}. \quad (3.94)$$

are algebraically equivalent on the regions $R_1 = (0, \infty) \times \mathbf{R} = \{(x, y) \in \mathbf{R} \mid x > 0\}$ and $R_2 = (-\infty, 0) \times \mathbf{R} = \{(x, y) \in \mathbf{R} \mid x < 0\}$, hence have the same general solution on each of these regions. But Theorem 3.43 *does* apply to the DE (3.94), on R_1 and R_2 , regions that together comprise almost the whole xy plane (everything but the y -axis). Hence we can solve (3.93) in R_1 and R_2 , and then see if we can infer from our answer whether there are solutions (3.93) that are not confined to R_1 or R_2 (and if so, what these solutions are).

Even if you are skipping the optional blue-text reading below, you should look at Figure 6 to see what the set of solution-curves of (3.93) turns out to look like.

First consider (3.94) 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 (3.94) on R_1 , in implicit form, is

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

and the set of constant solutions of equation (3.94) on R_1 is

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

By Theorem 3.43, every solution of (3.94) satisfies exactly *one* of the equations in $\mathcal{E}_1 \cup \mathcal{E}_2$, 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 ; 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 (3.94) on R_1 , in explicit form, is

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

Since the DEs (3.93) and (3.94) are equivalent on R_1 , they have the same general solution in this region. Hence the general solution of the DE (3.93) on R_1 is $\mathcal{E} =$

$\mathcal{E}'_1 \cup \mathcal{E}_2$. 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 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 (3.93) on R_1 is

$$\begin{aligned} \mathcal{E} &= \{y = 2 \tan^{-1}(Cx) + 2m\pi \mid m \text{ is an integer and } C \in \mathbf{R}\} \\ &\quad \text{and} \\ &\{y = (2m + 1)\pi \mid m \text{ is an integer}\} \end{aligned} \tag{3.95}$$

Similar analysis on R_2 reveals that the general solution of (3.93) 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 (3.93) in R_1 and in R_2 , but what we are looking for are solutions of (3.93) 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 (3.93) on $(-\infty, 0)$ and on $(0, \infty)$. At $x = 0$ we have $x\phi'(x) = 0 \times \phi'(0) = 0$, and $\sin(\phi(x)) = \sin(\phi(0)) = \sin(m\pi) = 0$. Hence ϕ is a solution of (3.93) on the whole real line (and is therefore a maximal solution). 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

$$\begin{aligned} \phi_+(x) &= 2 \tan^{-1}(C_1x) + 2m_1\pi \\ &\quad \text{or} \\ \phi_+(x) &= (2m_1 + 1)\pi. \end{aligned} \tag{3.96}$$

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 $\phi(0) = \lim_{x \rightarrow 0} \phi(x) = \lim_{x \rightarrow 0^+} \phi_+(x)$, which has the value $2m_1\pi$ or $(2m_1 + 1)\pi$ accordingly

as ϕ_+ is given by the top or bottom line of (3.95). 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 (3.95), 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 $\phi'(0) = \lim_{x \rightarrow 0} \frac{\phi(x) - \phi(0)}{x - 0} = \lim_{x \rightarrow 0^+} \frac{\phi_+(x) - \phi_+(0)}{x} = \phi'_+(0) = 2C_1$. 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 (3.95). Therefore the general solution of equation (3.93) is (3.95), with x now running over $(-\infty, \infty)$.

Note that every solution $y(x)$ of the DE (3.93) 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 $\{(x, y) \in \mathbf{R}^2 \mid x \neq 0\}$, 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. ■

3.1.8 Algebraic equivalence and general solutions of linear DEs

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

Example 3.49 The equations

$$\frac{dy}{dx} + 3y = \sin x \quad (3.97)$$

and

$$e^{3x} \frac{dy}{dx} + 3e^{3x}y = e^{3x} \sin x \quad (3.98)$$

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. ■

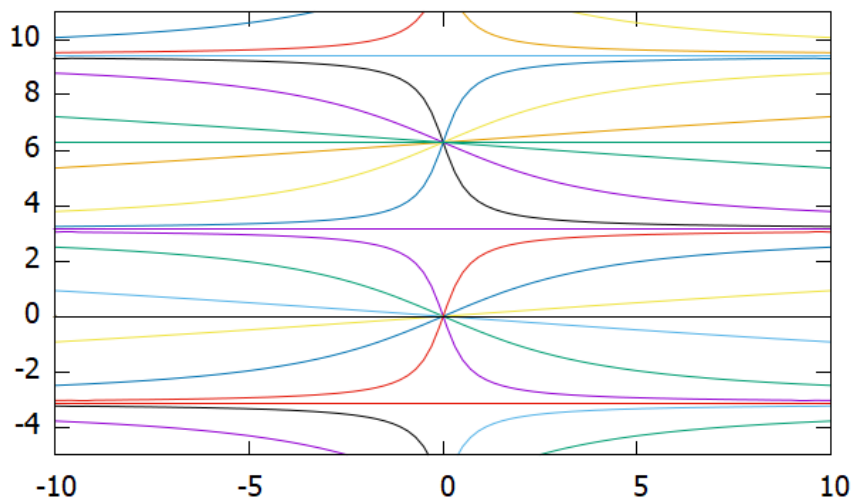


Figure 6: For Example 3.48: Several solutions of $x \frac{dy}{dx} = \sin y$, plotted in the rectangle $-10 \leq x \leq 10$, $-5 \leq y \leq 11$. (The displayed solution curves are maximal in this region.) The horizontal lines in the figure are the constant solutions $y = -\pi$, $y = 0$, $y = \pi$, $y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

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 3.40 to have language better suited to linear equations:

Definition 3.50 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 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 3.51 The equations

$$x \frac{dy}{dx} - 2y = 0 \tag{3.99}$$

and

$$x^3 \frac{dy}{dx} - 2x^2 y = 0 \quad (3.100)$$

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 3.40, 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 3.52 The equations

$$x \frac{dy}{dx} - 2y = 0 \quad (3.101)$$

(the same equation as (3.99) and

$$x^{-2} \frac{dy}{dx} - 2x^{-3} y = 0 \quad (3.102)$$

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

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

In the context of linear DEs, fact (3.75) reduces to the following simpler statement:

$$\begin{aligned} &\text{Two linear DEs that are algebraically equivalent} \\ &\text{on an interval } I \text{ have exactly the same solutions on } I. \end{aligned} \quad (3.103)$$

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 3.53 Find the general solution of

$$x \frac{dy}{dx} - 2y = 0 \quad (3.104)$$

(the same equation as (3.101) and (3.99)).

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

$$\frac{dy}{dx} - \frac{2}{x} y = 0. \quad (3.105)$$

However, (3.104) and (3.105) are not algebraically equivalent on the whole real line, but only on $(-\infty, 0)$ and $(0, \infty)$. Equation (3.105) does not even make sense at $x = 0$, while (3.104) makes perfectly good sense there.⁴⁹

As the student may verify, equation (3.105) has an integrating factor $\mu(x) = x^{-2}$. Putting our brains on auto-pilot, we multiply through by x^{-2} , and write

$$\begin{aligned} (x^{-2}y)' &= 0, \\ \Rightarrow \int (x^{-2}y)' dx &= \int 0 dx, \\ \Rightarrow x^{-2}y &= C, \\ \Rightarrow y &= Cx^2. \end{aligned} \quad (3.106)$$

(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 (3.106).)

Neither in the original DE (3.104) nor in (3.106) 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 (3.106) 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 (3.106) 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 (3.104). 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.

⁴⁹ Standard terminology related to this problem is *singular point*. Roughly speaking, a first-order linear DE does not “behave well” on an interval I if, when the DE is put in standard linear form $\frac{dy}{dx} + p(x)y = g(x)$, there is a point $x_0 \in I$ for which $\lim_{x \rightarrow x_0^+} |p(x)| = \infty$ or $\lim_{x \rightarrow x_0^-} |p(x)| = \infty$. Such points x_0 are called *singular points* of the linear DE. The point $x = 0$ is a singular point of both (3.104) and (3.105).

Let us go back to square-one and correct our work. The transition from equation (3.104) to (3.105) 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 (3.104) might include solutions on intervals that contain 0, while the general solution of (3.105) cannot.

We can still use the basic procedure that led us to (3.106); we just have to be more careful with it. Auto-pilot will not work.

Because (3.105) makes no sense at $x = 0$, we must solve it separately on $(-\infty, 0)$ and $(0, \infty)$. We can do the work for both of these intervals simultaneously, as long as we keep track of the fact that that's what we're doing.

So suppose ϕ is a differentiable function on *either* on $I = (0, \infty)$ or on $I = (-\infty, 0)$, and let $y = \phi(x)$. On I , x^{-2} is an integrating factor. Multiplying both sides of our equation on I by x^{-2} , we find that ϕ is a solution of (3.105) 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 (3.105) 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

$$\phi(x) = Cx^2. \tag{3.107}$$

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

Thus the general solution of (3.105) on $(0, \infty)$ is

$$y = Cx^2, \quad x > 0, \tag{3.108}$$

while the general solution of (3.105) on $(-\infty, 0)$ is

$$y = Cx^2, \quad x < 0. \tag{3.109}$$

Now return to the equation we originally were asked to solve, (3.104), 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 (3.104) and (3.105) are algebraically equivalent on $(0, \infty)$, ϕ_1 must be one of the solutions given by (3.108). 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 (3.109), 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 $\phi(0) = \lim_{x \rightarrow 0} \phi(x)$. 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 $\phi(0) = 0$.⁵⁰ 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

$$\phi(x) = \begin{cases} C_1x^2 & \text{if } x \geq 0, \\ C_2x^2 & \text{if } x < 0, \end{cases} \quad (3.110)$$

(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 (3.110) is a solution of (3.104) on $(-\infty, \infty)$. Therefore the general solution of (3.104) on $(-\infty, \infty)$ is the *two-parameter* family of functions given by (3.110), with C_1 and C_2 arbitrary constants⁵¹. 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 (3.110) to the interval I . (For the student who read and understood the material on maximal solutions: the two-parameter family (3.110) is the general solution of (3.104) as defined in Definition 3.26.) ■

Remark 3.54 In Example 3.53, we saw that $\{y = Cx^2\}$ is not a general solution is not the general solution of $xy' - 2y = 0$. However, in the terminology of Definition 3.46, $\{y = Cx^2\}$ is a *piecewise generating set of equations* for the general solution.

⁵⁰Another 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 (3.104), 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 (3.104), it does not work for (3.100)—which the student will later be asked to solve—but the first method we presented does.

⁵¹Some authors, with a different definition of “general solution”, would say that the first-order linear equation (3.104) *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 .



You should not draw the wrong impression from Example 3.53. 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 3.53 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 (3.99) and (3.100) have the same set of solutions.

3.2 First-order equations in differential form

Definition 3.55 A *differential* in the variables (x, y) is an expression of the form

$$M(x, y)dx + N(x, y)dy \quad (3.111)$$

where M and N are functions defined on some region in \mathbf{R}^2 . We often abbreviate (3.111) as just

$$Mdx + Ndy, \quad (3.112)$$

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 (3.111) and (3.112) are called the *coefficients* of dx and dy in these expressions. ■

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

Definition 3.56 (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

$$dF = \frac{\partial F}{\partial x} dx + \frac{\partial F}{\partial y} dy. \quad (3.113)$$

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

Remember that for a function of two variables, “continuously differentiable (on R)” means that each of the two partial derivatives 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 (3.111) and (3.112). Similarly, so far the expression “ $Mdx + Ndy$ ” is just *notation*; its information-content is just the pair of functions M, N (plus the knowledge of which function is the coefficient of dx and which is the coefficient of dy).

You (the student) may have come across the noun “differential” in your previous calculus courses. The sense in which we use this noun in these notes is more sophisticated than the notion used in Calculus 1-2-3. (For interested students, Section 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 $Mdx + Ndy$ is a differential on a region R , and (x_0, y_0) is a point in R , we call the expression $M(x_0, y_0)dx + N(x_0, y_0)dy$ the *value* of the differential $Mdx + Ndy$ at (x_0, y_0) . However, this “value” is not a real number; so far it is only a piece of notation of the form “(real number times dx) + (real number times dy)”, and we still have attached no meaning to “ dx ” and “ dy ”. The value of a differential at a point is actually a certain type of *vector*, but not the type you learned about in Calculus 3. (The type of vector that it *is* will not be described in these notes; the necessary concepts require a great deal of mathematical sophistication to appreciate, and are usually not introduced at the undergraduate level.⁵²)

We next define rules for algebraic operations involving differentials. These definitions are necessary, rather than being “obvious facts”, because so far differentials are just pieces of notation to which we have attached no meaning. **However, in an introductory course on DEs, it is generally permissible for students to treat the rules in Definition 3.57 as “obvious facts”.**⁵³ If you have trouble understanding why Definition 3.57 is necessary, don’t worry about it; just make sure that the way you manipulate differentials agrees with these rules.

Definition 3.57 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 .

⁵²However, 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.

⁵³This tends to be what DE textbooks do: the algebraic rules in Definition 3.57 are *used* without ever stating them.

(Thus $Mdx + Ndy$, $M_1dx + N_1dy$, and $M_2dx + N_2dy$ are differentials on R .) Then we make the following definitions for differentials in (x, y) .

1. Equality of differentials: $M_1dx + N_1dy = M_2dx + N_2dy$ on R if and only if $M_1(x, y) = M_2(x, y)$ and $N_1(x, y) = N_2(x, y)$ for all $(x, y) \in R$.
2. Abbreviation by omitting terms with coefficient zero:

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

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

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

4. Insensitivity to which term is written first:

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

5. Addition of differentials:

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

6. Subtraction of differentials:

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

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

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

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

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

right-hand side is to be interpreted as the zero differential⁵⁴. ■

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 $\frac{dy}{dx}$. (When we do relate them later, $\frac{dy}{dx}$ still will not be the quotient of two differentials.)

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

$$\text{one differential in } (x, y) = \text{another differential in } (x, y). \quad (3.114)$$

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 “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 one side or the other. The equation “ $dF = \frac{\partial F}{\partial x} dx + \frac{\partial F}{\partial y} dy$ ” is an example of a *definitional* equation.

Example 3.59 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

$$y dy = e^x dx. \quad (3.115)$$

■

⁵⁴As 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} ,” for the zero number, zero function, and zero differential respectively.

Note that when we write equation (3.115), 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 3.2.2 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. 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

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

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

Here is the differential-form analog of Definition 3.40:

Definition 3.60 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 $\{(x, y) \mid x \neq 0\}$ these equations are also algebraically equivalent to

$$x(2x^2y dx - \tan(x + y) dy) = 0, \quad (3.116)$$

but are *not* algebraically equivalent to (3.116) 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 (3.114) from both sides of the equation, we obtain an algebraically equivalent equation of the form

$$M dx + N dy = 0. \quad (3.117)$$

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 (3.117). (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 (3.115) 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 before 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.

3.2.1 Curves, parametrized curves, and smooth curves

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

Definition 3.61 A *parametrized curve* or *curve-parametrization* in \mathbf{R}^2 is an ordered pair of continuous real-valued functions (f, g) defined on an interval⁵⁵. The set

$$\{(f(t), g(t)) \mid t \in I\} \quad (3.118)$$

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

⁵⁵In these notes, intervals are required to have more than one point; we never mean “degenerate intervals” $[a, a]$.

A *curve* in \mathbf{R}^2 is a set $\mathcal{C} \subset \mathbf{R}^2$ that is the image of some parametrized curve and has more than one point.^{56 57}

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: the range of (f, g) , thought of as a single \mathbf{R}^2 -valued function γ (defined by $\gamma(t) = (f(t), g(t))$) rather than as a pair of real-valued functions. (In case the 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”.)

Note that in Definition 3.61, we do not require the interval I to be open. This is so that we can present certain examples below simply, without bringing in too many concepts at once that may be new to the student. Eventually, we will want to consider

⁵⁶The “ \mathcal{C} ” used in these notes for a curve is in a different font from the C that we use for a constant.

⁵⁷If the functions f, g in (3.118) 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 it is essential to allow this.

only parametrized curves that have an open domain-interval, but we will not impose that requirement just yet.

Example 3.62 Let $x(t) = 2 \cos t, y(t) = 2 \sin t, t \in [0, 2\pi]$. Then for all t we have $x(t)^2 + y(t)^2 = 4$, so the 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 parametrized curve, so the *curve* traced out by the parametrized curve $t \mapsto (x(t), y(t)), t \in [0, 2\pi]$, is the whole circle $x^2 + y^2 = 4$. Had we used the same formulas for $x(t)$ and $y(t)$, but restricted t to the interval $[0, \pi]$, the range would still have lain along the circle $x^2 + y^2 = 4$, but would have been only a semicircle. Had we used the same formulas, but used a slightly larger, open interval, say $(-0.1, 2\pi + 0.1)$, then we would have obtained the whole circle again, with some small arcs traced-out twice. ■

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

Definition 3.63 A curve-parametrization $(x(t), y(t)), t \in I$ is called

- *differentiable* if the derivatives $x'(t), y'(t)$ exist⁵⁸ for all $t \in I$;
- *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 $x'(t_0) = 0 = y'(t_0)$);⁵⁹ and
- *regular* if it is continuously differentiable and non-stop.

■

Definition 3.64 A curve \mathcal{C} in \mathbf{R}^2 is *smooth* if for every point (x_0, y_0) on the curve, there is a number $\epsilon_0 > 0$ such that for all positive $\epsilon < \epsilon_0$, the portion of \mathcal{C} lying inside the open square of side-length ϵ centered at (x_0, y_0) admits a regular parametrization,

⁵⁸When 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. Thus, if I contains a left endpoint a , then what we mean by “ $x'(a)$ ”, or “ $\frac{dx}{dt}$ at a ”, is $\lim_{t \rightarrow a^+} \frac{x(t) - x(a)}{t - a}$. Similarly if I contains a right endpoint b , then what we mean by “ $x'(b)$ ”, or “ $\frac{dx}{dt}$ at b ”, is $\lim_{t \rightarrow b^-} \frac{x(t) - x(b)}{t - b}$.

⁵⁹“Differentiable” and “continuously differentiable” are completely standard terminology; “non-stop” is not. See the “Warning about terminology” coming up soon.

with domain an open interval.⁶⁰ ■

“Admits”, as used in Definition 3.64, 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 3.64 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 3.62 to convince him/herself that a circle meets our definition of “smooth curve”.

Observe that Definition 3.64 uses a “windowing” idea similar to the one that we used to talk about implicitly-defined functions in Section 3.1.2. We will later give an equivalent definition of “smooth curve” that is even more reminiscent of that earlier discussion.

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

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

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

⁶⁰*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 a 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.

- 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 small ϵ , the portion of the curve that lies in the square of side ϵ centered at (x_0, y_0) is essentially an X, and has the same problem that the X did.

Warning about terminology. Many calculus textbooks refer to a 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.⁶¹ “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).

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.**

Definition 3.65 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.⁶² ■

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 .

⁶¹*Note to instructors:* in differential topology and differential geometry, “smooth parametrization” simply means “ C^k map” (from an open interval to \mathbf{R}^2 , in the setting of these notes) for some pre-specified k , usually 1 or ∞ . There is no requirement that the parametrization be non-stop to be called smooth. Even *constant* maps, whose images are a single point, are considered smooth parametrized curves—and are indispensable to the definition of “tangent space”.

⁶²The 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.

3.2.2 Solution curves for DEs in differential form

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

Definition 3.66 A *solution curve*⁶³ of a differential equation

$$M(x, y) dx + N(x, y) dy = 0 \quad (3.119)$$

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

$$M(x(t), y(t)) \frac{dx}{dt} + N(x(t), y(t)) \frac{dy}{dt} = 0 \quad (3.120)$$

for all t in the domain-interval I of the parametrization. In this context, we will call γ a *parametric solution* of (3.119) (in R).⁶⁴

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 domain⁶⁵.



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 3.2.4.

As we noted previously, in a differential-form DE (3.119) there is neither an independent nor a dependent variable; x and y are treated symmetrically. This symmetry is preserved in (3.120), but in a surprising way: in (3.120), *both* x and y are dependent variables! The independent variable is t —a variable that is not even visible in (3.119). (Of course, in place of t we could have used any other letter not appearing elsewhere in Definition 3.66.)

[Definition 3.66 implies more about solution curves and parametric solutions than is obvious just from reading the definition.](#)

⁶³It 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 3.72 later in these notes.

⁶⁴The terminology “parametric solution” for a DE in differential form was invented for these notes; it is not standard.

⁶⁵*Note to instructor:* I have avoided giving a careful definition of “boundary” here, and therefore of “interior”, to avoid distracting the student.

To start with, equation (3.120) has a geometric interpretation.⁶⁶ 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 (3.120) using the dot-product you learned in Calculus 3:

$$[M(x(t), y(t))\mathbf{i} + N(x(t), y(t))\mathbf{j}] \cdot \mathbf{v}(t) = 0. \quad (3.121)$$

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 $\gamma_2(t)$ (with t -domain I_2) for the second. At a given point $(x_0, y_0) = \gamma_1(t_1) = \gamma_2(t_2)$ on \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 (3.120).⁶⁷ Thus if one regular parametrization of \mathcal{C} satisfies (3.120), so does every other.

Therefore, even though Definition 3.66 requires only that *some* regular parametrization of \mathcal{C} satisfy (3.120), once we know that even *one* regular parametrization of \mathcal{C} satisfies (3.120), we know that they *all* do. Said another way:

$$\left. \begin{array}{l} \textit{Every} \text{ regular parametrization of a solution curve} \\ \text{of a differential equation } M dx + N dy = 0 \\ \text{is a parametric solution of this equation.} \end{array} \right\} \quad (3.122)$$

This gets back to what we said just before to Definition 3.66: that a DE in differential form is looking for a *curve*. We carefully did not say “*parametrized curve*”.

⁶⁶If you have not yet taken Calculus 3, either (i) look up “standard basis vectors” and “dot product” before proceeding, or (ii) skip to statement (3.122) below, taking it on faith that statement (3.122) was established in the part you skipped over.

⁶⁷This 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 79.)

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”.

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 (3.122) is saying is that the property “I am a solution curve of this differential-form DE” is an intrinsic, geometric property.

Although the concepts of “solution of a DE in derivative form” and “solution curve of a DE in differential form” are fundamentally different—the former is a function (of one variable); the latter is a *geometric object*—they are still related to each other. We will see the precise relation in Section 3.3. 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 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 desirable⁶⁸. This is another fundamental difference between derivative-form DEs and differential-form DEs.

Example 3.67 Consider the equation

$$x dx + y dy = 0. \tag{3.123}$$

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

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

⁶⁸This “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.

$t \in [0, 2\pi]$, 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 (3.123) 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. ■

3.2.3 Existence/uniqueness theorem for DEs in differential form

Recall that an initial-value problem, with dependent variable y and independent variable x , consists of a derivative-form differential equation together with an initial condition of the form $y(x_0) = y_0$. The differential-form analog of an initial-value problem is a differential-form DE together with a point (x_0, y_0) ; 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 (3.121). 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 (3.121) 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 (3.121) is satisfied at $t = t_0$, and so is (3.120). 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 3.68 A point (x_0, y_0) is a *singular point* of the differential $M dx + N dy$ if $M(x_0, y_0) = 0 = N(x_0, y_0)$.⁶⁹ ■

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

$$\frac{dy}{dx} = f(x, y). \quad (3.124)$$

If the graph of a solution of (3.124) 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

$$\frac{dy}{dx} = f(x, y), \quad y(x_0) = y_0 \quad (3.125)$$

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 (3.125).⁷⁰ 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 3.69 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 3.65).

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

⁶⁹ “Singular point” here does not mean the same thing as in footnote 49.

⁷⁰ 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.

that we will not discuss here (but one of which is a phenomenon exhibited later in Example 3.82).⁷¹

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

Theorem 3.70 *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 every point $(x_0, y_0) \in R$, there exists a unique solution curve of $M dx + N dy = 0$ 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 3.71 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 $(x_0, y_0) \in R$,

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

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

3.2.4 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'

⁷¹*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.

definitions of “solution of a differential-form DE” tends to look very similar to their definitions of “solution of a derivative-form DE”. Usually this is accomplished by saying, early on, “We’re going to use the word ‘solution’ to refer to both ‘explicit’ and implicit solutions (of derivative-form DEs),” and then effectively taking the definition of “solution of a DE in differential form” to be “implicit solution of a related derivative-form DE”.⁷² Since this is what students are most likely to see in a textbook, we make here a similar definition of “implicit solution of a DE in differential form” that is consistent with current textbooks’ treatment of this concept. But we relate it carefully to the concepts we have developed earlier, and provide alternative terminology that we will generally use instead.

Definition 3.72 An equation

$$F(x, y) = 0 \quad (\text{or } F(x, y) = \text{any fixed real number}) \quad (3.126)$$

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

$$M(x, y) dx + N(x, y) dy = 0 \quad (3.127)$$

on a region R if

- (i) the graph of (3.126) contains a smooth curve in R , and
- (ii) every smooth curve in R contained in the graph of (3.126) is a solution curve of (3.127).⁷³

If $R = \mathbf{R}^2$ then we usually omit mention of the region, and say just that (3.126) is an *implicit solution*, *algebraic solution*, or *non-parametric solution*, of (3.127). ■

Remark 3.73 Note that we have not yet defined the term “*solution of a DE in differential form*”. 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

⁷²*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*.

⁷³*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.

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 (3.126) is in relation to (3.127). Because a curve is a point-set in the plane, an equation of the form $F(x, y) = 0$ is actually a very *explicit* description of a curve \mathcal{C} (when this equation does define a curve): a point (x, y) is on \mathcal{C} if and only if $F(x, y) = 0$. “Implicitly defined *function*” 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 (3.126) *itself* is not a solution of equation (3.127)—rather, the solutions of equation (3.127) related to (3.126) are smooth curves contained in the *graph* of equation (3.126). (The solutions of equation (3.127) won’t *all* be related to (3.126), even if $dF = M dx + N dy$, because we chose a specific value for the constant on the right-hand side in (3.126); Definition (3.72) started with the words “*An equation*”.) For these reasons, in the context of Definition 3.72 I find it preferable to call equation (3.126) an *algebraic* or *non-parametric* solution of equation (3.127), rather than an *implicit* solution. 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 “implicit”. ■

Because the “implicit” in Definition 3.72 is both cumbersome and misleading, we will avoid it as much as possible in this context, and use “algebraic” instead. However, we will also allow ourselves an abuse of terminology similarly in spirit to what we allow for solutions of derivative-form equations (saying, e.g., that the equation “ $y = x^2$ ” is a solution of $\frac{dy}{dx} = 2x$ ” even though “ $y = x^2$ ” *can’t* literally be a solution [because an equation isn’t a function]):

Definition 3.74 In the setting of Definition 3.72, we also call an algebraic solution of equation (3.127) (on R) simply a *solution* of equation (3.127) (on R).

Example 3.75 The equation

$$xy = 1$$

is a solution of

$$y dx + x dy = 0. \tag{3.128}$$

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 these 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 DE (3.128). For most C , the graph is a hyperbola, but the case $C = 0$ is exceptional. The graph of

$$xy = 0 \tag{3.129}$$

is a pair of crossed lines, the x - and y -axes. Note that this graph is not a smooth curve, nor is it the *disjoint* union of two smooth curves the way a hyperbola is (“disjoint” meaning that the two curves have no points in common). We can verify that (3.129) is indeed a solution of (3.128) 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 (3.128)—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 $y dx + x dy$. (The only singular point is $(0, 0)$.) ■

Remark 3.76 (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 3.4). By contrast, a differential-form DE

$$M dx + N dy = 0 \tag{3.130}$$

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 (3.128) 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 (3.130) 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 (3.120) 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 (3.130). Conversely, these are the *only* vertical lines that are solution curves of (3.130) 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 0* 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 0 on horizontal curves, and allow ourselves to say more loosely that dy “is” zero on horizontal lines. ■

Remark 3.77 You may wonder to what extent criterion (i) in Definition 3.72 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 3.72. ■

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 3.73). *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 3.38:

Definition 3.78 (General solution of a differential-form DE on a region)

For a given differential-form DE

$$Mdx + Ndy = 0 \tag{3.131}$$

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 (3.131) on R* if

- (i) each equation in the collection \mathcal{E} is a solution of (3.131) on R (see Definitions 3.74 and 3.72), and
- (ii) every solution curve of (3.131) 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 (3.131) 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 . ■

Note that in (ii), we do not require that every solution curve be contained in *only* one graph of an equation in \mathcal{E} , in contrast to the analogous criterion in Definition 3.38. The reason is that we have not defined *maximal* solution curve except in regions containing no singular points of $M dx + N dy$. Were we to remove the word “maximal” in Definition 3.38's criterion (ii) but retain the “one and only one”, we would have a less satisfactory definition in any region in which the hypotheses of the Fundamental Theorem are not satisfied. For example, the DE 3.38 would have *no* implicit or explicit form of a general solution.

3.2.5 Exact equations

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

Example 3.79 Suppose $M dx + N dy$ is an exact differential on a region R (see Definition 3.56), and let F be a differentiable function on R for which $M dx + N dy = dF$. Then (3.119) becomes

$$\frac{\partial F}{\partial x} dx + \frac{\partial F}{\partial y} dy = 0. \quad (3.132)$$

Suppose that \mathcal{C} is a solution curve of (3.132), and that $t \mapsto (x(t), y(t))$, $t \in I$, is a continuously differentiable parametrization of \mathcal{C} . Then (3.120) says

$$\frac{\partial F}{\partial x}(x(t), y(t)) \frac{dx}{dt} + \frac{\partial F}{\partial y}(x(t), y(t)) \frac{dy}{dt} = 0. \quad (3.133)$$

By the multivariable Chain Rule (learned in Calculus 3), the left-hand side of (3.133) is just $\frac{d}{dt}F(x(t), y(t))$. Thus equation (3.120) simplifies, in this case, to

$$\frac{d}{dt}F(x(t), y(t)) = 0 \quad \text{for all } t \in I. \quad (3.134)$$

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

$$F(x(t), y(t)) = c_0 \quad (3.135)$$

for all $t \in I$. This implies that *every solution curve of (3.132) in R is contained in the graph of (3.135) for some value of the constant c_0 .*

Now, fix a number c_0 , and consider the equation

$$F(x, y) = c_0. \quad (3.136)$$

Is this equation a solution of (3.132) in R , according to Definition 3.74? The answer is yes, provided that the graph of (3.126) contains a smooth curve in R (criterion (i) of Definition 3.72). If this criterion is met, let \mathcal{C} be a smooth curve in R that is contained in the graph of (3.136). 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 (3.136), equation (3.135) is satisfied for all $t \in I$. Differentiating both sides of (3.135) with respect to t , we find that equation (3.134) is satisfied. But, by the Chain Rule, the left-hand side of (3.134) is exactly the left-hand side of (3.133), so equation (3.133) is satisfied. Therefore \mathcal{C} is a solution curve of the differential equation (3.132). Hence criterion (ii) of Definition 3.72 is met, so (3.136) is a solution of the DE (3.132) in R . ■

Example 3.80 (General solution of an exact equation) Suppose we are given a differential-form equation

$$M dx + N dy = 0 \tag{3.137}$$

that is exact on a region R , and F is a function for which $M dx + N dy = dF$ on R . Then Example 3.79 shows that **one algebraic form of the general solution of (3.137) on R is the collection of equations**

$$\{F(x, y) = C\}, \tag{3.138}$$

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 (3.138), and to refer to “ $\{F(x, y) = C\}$ ” as (an algebraic form of) the general solution of (3.137) 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 (3.138).

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 (3.137) and function F above, we can say something stronger than that (3.138) 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 $\{F(x, y) = C\}$. (See the paragraph after Definition 3.78.) ■

For any real number C and (two-variable) function F , the graph of $F(x, y) = C$ is called a *level-set* of F .⁷⁴ 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 3.81 (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

$$y^3 - |x|^3 = 0, \tag{3.139}$$

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

⁷⁴The same terminology is used for functions of any number of variables.

Equation (3.139) is a solution of

$$y^2 dy + \left\{ \begin{array}{l} -3x^2, \quad x \geq 0 \\ 3x^2, \quad x \leq 0 \end{array} \right\} dx = 0; \quad (3.140)$$

it meets both criteria in Definition 3.72. But as seen above, the graph of (3.139) contains a point, $(0, 0)$, that is not on any solution *curve* of (3.140) (see Definitions 3.66 and 3.64). 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 3.81 was a critical point of F (hence a singular point of the differential dF). In the absence of singular points, we can be much more concrete about the general solution of an exact equation:

$$\left. \begin{array}{l} \text{If a differential } M dx + N dy \text{ is equal to } dF \text{ on a} \\ \text{region } R, \text{ and has no singular points in } R, \text{ then} \\ \text{the set of } C\text{'s allowed in (3.138) is simply the range} \\ \text{of } F \text{ on the region } C, \text{ and every point in } R \text{ is contained} \\ \text{in a unique solution curve that is maximal in } R. \end{array} \right\} \quad (3.141)$$

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

3.2.6 Algebraic equivalence of DEs in differential form

Algebraic equivalence (see Definition 3.60) 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 3.72 to this analysis, we have the following:

If two differential-form DEs are algebraically equivalent on a region R , then in R they have exactly the same solution curves, exactly the same parametric solutions, and exactly the same solutions. } (3.142)

Combining this fact with Example 3.80, we have the following:

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 (an algebraic form of) the general solution of $Mdx + Ndy = 0$ on R . The same understanding concerning the allowed values of the constant C in “ $\{F(x, y) = C\}$ ” applies as in Example 3.80. } (3.143)

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 3.69 we required $M dx + N dy$ to have no singular points in R). 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 3.82 Consider the DE

$$2xy dx + (y^2 - x^2)dy = 0. \tag{3.144}$$

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 (3.128), we have

a differential with one singular point, which happens to be the origin⁷⁵. Letting $R = \{\mathbf{R}^2 \text{ minus the origin}\}$, Theorem 3.70 guarantees us that through each point $(x_0, y_0) \neq (0, 0)$, there exists a unique solution curve of (3.144) that is maximal in R .

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

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

Similarly, the negative x -axis is a solution curve. The uniqueness statement in Theorem 3.70 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 (3.144) is algebraically equivalent to

$$\frac{1}{y^2} (2xy \, dx + (y^2 - x^2) \, dy) = 0. \quad (3.145)$$

But as the student may verify,

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

So on R_+ , and also on R_- , the left-hand side of (3.145) is exact; it is dF , where $F(x, y) = \frac{x^2 + y^2}{y}$. Hence one form of the general solution of (3.145), in either of these regions, is

$$\left\{ \frac{x^2 + y^2}{y} = C \right\}, \quad (3.146)$$

where, from fact (3.141), 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

⁷⁵In 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.

see that these are the only restrictions on C , just observe that from the definition of F , we have $F(0, C) = C$.

Now for some algebraic rearrangement. Let us write $C = 2b$ in (3.146). 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,

$$\begin{aligned}
 & \frac{x^2 + y^2}{y} = 2b \\
 \iff & x^2 + y^2 = 2by, \\
 \iff & x^2 + y^2 - 2by = 0, \\
 \iff & x^2 + y^2 - 2by + b^2 = b^2, \\
 \iff & x^2 + (y - b)^2 = b^2.
 \end{aligned} \tag{3.147}$$

The graph of (3.147) 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 (3.145) on R_+ and on R_- . But since equations (3.145) and (3.144) are algebraically equivalent on these regions, the same curves are all the maximal solution curves of the DE (3.144) in these regions.

We have now found all the solution curves of (3.144) 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 $R = \{\mathbf{R}^2 \text{ minus the origin}\}$. If we now re-include the origin, we see that the origin lies on every one of the circles described by (3.147), 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 (3.144). We leave the student to check that each full circle (3.147), with the origin included, is also a solution curve of (3.144).

Thus, *among* the solution curves of (3.144) are all circles centered on the y axis, and an “exceptional” curve, the x -axis. A collection of algebraic solutions of (3.144) corresponding exactly to this set of solution curves is

$$\{x^2 + (y - b)^2 = b^2 \mid b \neq 0\} \quad \text{and} \quad \{y = 0\}. \tag{3.148}$$

From the foregoing analysis, it may appear that the set of all solution curves of (3.144) 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 (3.144) is (3.148). But both of these conclusions 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 y -axis as the first circle or on the opposite side, whatever you feel like. Stop at a point Q before you reach the origin again. Erase the endpoints P and Q (see the second paragraph after Definition 3.64), and you have a

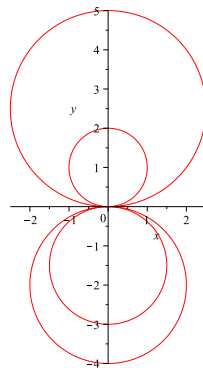


Figure 7: Some solution curves of $2xy dx + (y^2 - x^2)dy = 0$. (The graphing utility used to render this diagram does not do a good job near the origin; there should be no gap in any of the circles.)

perfectly good, smooth, solution curve that is not contained in any circle or in the x -axis.

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

Thus there are solution curves of (3.145) that are not contained in any of the “circles plus one straight line” family given by (3.148). 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 (3.148).

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 (3.148), 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 (3.144) that do not lie in the graph of a *single* equation in the collection (3.148). Thus (3.148) is (one algebraic form of) the general solution of the DE (3.144). ■

In the example above, an alternative way of expressing the collection (3.148) is as follows. In (3.146), 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 (3.146) 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 (3.148) in a unified way as

$$\{y = C(x^2 + y^2) \mid C \in \mathbf{R}\}. \quad (3.149)$$

(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 (3.148) and (3.149) are algebraic forms of the general solution of (3.144). 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.

We can also use Example 3.82 to exhibit one of the reasons it is difficult to give a satisfactory, useful, general definition of “maximal solution curve” of $Mdx + Ndy = 0$ in a region that includes singular points of $Mdx + Ndy$. 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 (3.144) 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.

In Example 3.82, all the solution curves in \mathbf{R}^2 intersected at the origin (a singular point of $Mdx + Ndy$) once extended far enough, but all had the same slope there (zero). Next we give an example of a very simple equation of the form $Mdx + Ndy = 0$ in which all the solution curves in \mathbf{R}^2 intersect at a singular point of $Mdx + Ndy$, but with all different slopes—in fact, with every possible slope.

Example 3.83 Consider the DE

$$x dy - y dx = 0. \tag{3.150}$$

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 $R = \{\mathbf{R}^2 \text{ minus the origin}\}$, 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 (3.150) 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 through 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 (3.150) 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. ■

3.3 Relation between differential form and derivative form

Definition 3.84 Let M, N be functions on a region R in \mathbf{R}^2 . Consider the equations

$$M(x, y) dx + N(x, y) dy = 0, \quad (3.151)$$

$$M(x, y) + N(x, y) \frac{dy}{dx} = 0, \quad (3.152)$$

$$M(x, y) \frac{dx}{dy} + N(x, y) = 0. \quad (3.153)$$

We call equations (3.152) and (3.153) the *derivative-form DEs associated with the differential-form DE* (3.151). Similarly, we call equation (3.151) the *differential-form DE associated with the derivative-form DE* (3.152), and also the *differential-form DE associated with the derivative-form DE* (3.153).

More generally, if a derivative-form equation is algebraically equivalent to (3.152) or (3.153) on a region R , we call the equation a *derivative form of* (3.151) on R . Similarly, if a differential-form equation is algebraically equivalent to (3.151) on a region R , we call the equation a *differential form of* (3.152) and (3.153) on R .⁷⁶ ■

Remark 3.85 (Constant solutions of differential-form DEs) Note that equation (3.151) can conceivably have solutions of the form “ $x = \text{constant}$ ”, and/or solutions of the form “ $y = \text{constant}$ ”. As discussed in Remark 3.76, these correspond to

⁷⁶The 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.

solution curves of (3.151) that are vertical and horizontal lines, respectively. However, equation (3.152) has *no* vertical solution curves, and equation (3.153) has no horizontal solution curves.

If equation (3.151) has any solutions of the form $x = \text{constant}$, *we are guaranteed to lose these solutions if we replace equation (3.151) by the associated derivative-form equation (3.152)*. The notation in (3.151)–(3.152) 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 3.76), and pretend that (3.152) is obtained from (3.151) by the (nonsensical) operation of “dividing by dx ”, then we can think of the loss of solutions $x = \text{constant}$ in passing from (3.151) to (3.152) as begin a result of dividing by zero. Similar comments apply to the relationship among equation (3.151), equation (3.153), and solutions of (3.151) of the form $y = \text{constant}$. ■

As observed in Remark 3.85, 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 (3.151) by dx or dy to obtain the associated equation (3.152) or (3.153), or formally “multiply” equation (3.152) or (3.153) by dx or dy to obtain the associated equation (3.151). *This is an extremely useful memory-device, and the student should not hesitate to use it, but mathematically it is garbage.*⁷⁷ 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 .

We will see next just how and why equations (3.151)–(3.153) *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 .⁷⁸ (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.)

⁷⁷Unfortunately, most DE textbooks do not mention that this way of viewing the relations among (3.151), (3.152), and (3.153) 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.

⁷⁸This 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 1⁷⁹ 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 :

$$\gamma_{\text{new}}(t_{\text{new}}) = (t_{\text{new}}, \phi(t_{\text{new}})). \quad (3.154)$$

The variable in (3.154) 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

$$\gamma_{\text{new}}(x) = (x, \phi(x)). \quad (3.155)$$

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

⁷⁹This 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} .

$$M(x, y) dx + N(x, y) dy = 0. \quad (3.156)$$

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

$$M(x(t), y(t)) \frac{dx}{dt} + N(x(t), y(t)) \frac{dy}{dt} = 0 \quad (3.157)$$

In particular this is true for the parametrization (3.155), in which the parameter t is x itself, and in which we have $y(t) = \phi(t) = \phi(x) = y(x)$. Therefore, for all $x \in I_2$,

$$\begin{aligned} 0 &= M(x, \phi(x)) \frac{dx}{dx} + N(x, \phi(x)) \phi'(x) \\ &= M(x, \phi(x)) + N(x, \phi(x)) \phi'(x). \end{aligned} \quad (3.158)$$

The right-hand side of (3.158) 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

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

Therefore the portion \mathcal{C}_1 of \mathcal{C} is the graph of a solution (namely ϕ) of the derivative-form differential equation (3.159). The argument above also gives us the following an important fact to which we will want to refer later:

$$\left. \begin{array}{l} \text{If a solution curve of the differential-form equation} \\ M dx + N dy = 0 \text{ can be parametrized by } \gamma(x) = (x, \phi(x)), \\ \text{where } \phi \text{ is a differentiable function, then } \phi \text{ is a solution} \\ \text{of the associated derivative-form equation } M + N \frac{dy}{dx} = 0. \end{array} \right\} \quad (3.160)$$

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

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

Therefore:

$$\left. \begin{array}{l} \text{Every solution curve of the differential-form equation (3.151)} \\ \text{is a union of solution curves of the derivative-form} \\ \text{equations (3.152) and (3.153).} \end{array} \right\} \quad (3.162)$$

Example 3.67 provides an illustration of fact (3.162). 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 circle. *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 (3.162) 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.)

Now compare (3.159) with the general first-order derivative-form DE with independent variable x and dependent variable y ,

$$\mathbf{G}(x, y, \frac{dy}{dx}) = 0. \quad (3.163)$$

Equation (3.159) is a special case of (3.163), in which the dependence of \mathbf{G} on its third variable is very simple. If we use a third letter z for the third variable of \mathbf{G} , then (3.159) corresponds to taking $\mathbf{G}(x, y, z) = M(x, y) + N(x, y)z$, a function that can depend in any conceivable way on x and y , but is linear separately in z . In general, (3.163) could be a much more complicated equation, such as

$$\left(\frac{dy}{dx}\right)^3 + (x + y) \sin\left(\frac{dy}{dx}\right) + xe^y = 0. \quad (3.164)$$

Solving equations such as the one above is *much* harder than is solving equations of the simpler form (3.159). For certain functions \mathbf{G} that are more complicated than (3.159), but much less complicated than (3.164), methods of solution are known⁸⁰. But the general theory and techniques for working with equation (3.163) for general \mathbf{G} 's are much less highly developed than they are for equations in the standard form (3.166) or in the form (3.159).

⁸⁰One such type equation is a *Clairaut equation* $y = x \frac{dy}{dx} + g(\frac{dy}{dx})$, which is equivalent to (3.163) with $\mathbf{G}(x, y, z) = xz + g(z) - y$. 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.

One of the features of (3.159) that makes it so special is that on any region on which $N(x, y) \neq 0$, (3.159) is algebraically equivalent to

$$\frac{dy}{dx} = -\frac{M(x, y)}{N(x, y)}, \quad (3.165)$$

which is of form

$$\frac{dy}{dx} = f(x, y). \quad (3.166)$$

Recall that equation (3.166) 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.

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 (3.163). These facts apply just as well to nasty DEs like (3.164) as they do to (relatively) nice ones like (3.166). However, in all of our *examples*, we used equations that were algebraically equivalent to (3.152) (hence also to (3.166)) on some region. The reason is that although the concept of “the set of all solutions” makes perfectly good sense for the general equation (3.163), the author kept to examples in which he could show the student easily that the set of all solutions had actually been found.

Nowadays, students in an introductory DE course rarely see any first-order derivative-form equations that are not algebraically equivalent, on some region, to a DE in the standard form (3.166). 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 (3.166) on some region.** The two types of equations, in full generality, are not merely two sides of the same coin.

However, for derivative-form DEs that can be “put into standard form”—which are exactly those that are algebraically equivalent to a DE of the form (3.152) 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 (3.162) gives one such relation.

To have a name for equations that are explicitly of the form (3.152) or (3.153), let us say that a derivative-form equation, with independent variable x and dependent variable y , is in “almost-standard form”⁸¹ if it is in the form (3.152), 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 (3.162), you will see that it also

⁸¹This is another bit of terminology invented only for these notes, just to have a name to distinguish (3.152) from (3.165) on regions in which $N(x, y)$ may be zero somewhere.

shows that every solution curve of (3.152) or (3.153) is a solution curve of (3.151). Thus:

$$\left. \begin{array}{l} \text{Every solution curve of a derivative-form DE} \\ \text{in almost-standard form is a solution curve} \\ \text{of the associated differential-form equation.} \end{array} \right\} \quad (3.167)$$

Combining (3.162) and (3.167), we conclude the following:

$$\left. \begin{array}{l} \text{A smooth curve } \mathcal{C} \text{ is a solution curve of a DE} \\ \text{in differential form if and only if } \mathcal{C} \text{ is a union of} \\ \text{solution curves of the associated derivative-form} \\ \text{equations.} \end{array} \right\} \quad (3.168)$$

We emphasize that in deriving these relations, the transition from the differential-form DE (3.156) to the derivative-form DEs (3.159) and (3.161) 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 $Mdx + Ndy = 0$ are curves whose parametrizations satisfy (3.157), and that for particular choices of the parameter on the intervals that we called “ x -monotone” or “ y -monotone”, (3.157) reduces to (3.152) or (3.153).

Similarly, transitions from derivative form to differential form are NOT achieved by the nonsensical process of “multiplying by dx ” or “multiplying by dy ”. The benefit of the Leibniz notation “ $\frac{dy}{dx}$ ” for derivatives is that it can be used to help remember many true statements by *pretending, momentarily*, that you can multiply or divide by a differential just as if it were a real number⁸². In particular, we can use this principle help us easily *remember* that the differential-form equation (3.151) is related to (but not the same as!) the derivative-form equations (3.152) and (3.153). But this notational trick doesn’t tell us everything, such as the *precise relationship* among these equations, which is statement (3.167) (of which statement (3.162) is the “only if” half).

3.4 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,

⁸²Simultaneously, 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.

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 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 an example that illustrates one of the problems with the procedure as outlined.

Example 3.86 Solve the differential equation

$$(10xy^9 + 2xy)\frac{dy}{dx} = -(3x^2 + 1 + y^{10} + y^2). \quad (3.169)$$

(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 is⁸³

$$(3x^2 + 1 + y^{10} + y^2)dx + (10xy^9 + 2xy)dy = 0. \quad (3.170)$$

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 (3.170) is dF . Using our usual method, we find that an F with this property is

$$F(x, y) = x^3 + x + xy^{10} + xy^2. \quad (3.171)$$

From Example 3.80, we know that the general solution of (3.170) is

$$\{x^3 + x + xy^{10} + xy^2 = C\}, \quad (3.172)$$

where C is (at worst) a semi-arbitrary constant. Fact (3.141) shows that the set of allowed values of C is simply the range of F , provided that $M dx + N dy$ has no singular points. Looking at $M(x, y)$, we observe that x^2, y^{10} , and y^2 are all ≥ 0 , so $M(x, y) \geq 1$. Therefore $M(x, y)$ is nowhere zero, so $M dx + N dy$ has no singular points. So fact (3.141) 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 (3.171) and observe that $\lim_{x \rightarrow \infty} F(x, 0) = \infty$ and $\lim_{x \rightarrow -\infty} F(x, 0) = -\infty$.)

Therefore the general solution of (3.170) is the family of equations (3.172), with C a *completely* arbitrary constant; all real values are allowed.

But the equation we wanted to solve was the derivative-form equation (3.169), not the differential-form equation (3.170), so we ask: is this same family (3.172) the general solution of (3.169), in implicit form? The answer is no.

To see why, consider the equation in \mathcal{E} corresponding to $C = 0$:

$$x^3 + x + xy^{10} + xy^2 = 0. \quad (3.173)$$

(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

$$x(x^2 + 1 + y^{10} + y^2) = 0. \quad (3.174)$$

⁸³More 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 (3.169) from the left-hand side; the other is obtained by first subtracting the left-hand side of (3.169) from the right-hand side. Each of these equations is just the other with both sides multiplied by -1 .

The quantity inside parentheses is strictly positive, so (3.174) is equivalent to just $x = 0$. The graph of (3.174) is simply the y -axis, a perfectly nice smooth curve, and a perfectly good solution curve of (3.170), 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 3.169, so equation (3.174) is not an implicit solution of (3.169). Therefore \mathcal{E} does not meet Definition 3.38's first criterion for "general solution, in implicit form, of a derivative-form DE".

This demonstrates the main point of this example:

$$\left. \begin{array}{l} \text{A collection of equations can be an implicit form of} \\ \text{the general solution of an almost-standard-form} \\ \text{derivative-form DE, yet not an algebraic} \\ \text{form of the general solution of the associated} \\ \text{differential-form DE, and vice-versa.} \end{array} \right\} \quad (3.175)$$

However, while the two general solutions in (3.175) need not be *identical*, our conclusions in the previous section show that they are closely related. A consequence of fact (3.167) is the following:

$$\left. \begin{array}{l} \text{If a collection } \mathcal{E} \text{ of algebraic equations is the general solution} \\ \text{of a DE } M(x, y) dx + N(x, y) dy = 0, \text{ and the graphs of no two} \\ \text{equations in } \mathcal{E} \text{ overlap (i.e., if the graphs intersect at all,} \\ \text{there is no } \textit{curve} \text{ contained in the intersection), then } \mathcal{E} \\ \textit{contains} \text{ an implicit form of the general solution of} \\ M(x, y) + N(x, y) \frac{dy}{dx} = 0. \text{ Thus, if we are trying to obtain the} \\ \text{general solution of } M(x, y) + N(x, y) \frac{dy}{dx} = 0 \text{ from having found} \\ \mathcal{E} \text{ as the general solution of } M(x, y) dx + N(x, y) dy = 0, \text{ we} \\ \text{need only worry whether } \mathcal{E} \text{ has any equations that are } \textit{not} \\ \text{implicit solutions of the derivative-form equation, or has} \\ \text{any equations whose graphs overlap each other.} \end{array} \right\} \quad (3.176)$$

The reason for the no-overlap restriction in fact (3.176) is that in Definition 3.38, 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 3.78 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 $Mdx + Ndy$ is *exact*, we can simplify fact (3.176) a great deal. As noted in Example 3.80, 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 an 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. (3.177)

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 (3.169)? The answer is no. (This can be shown using the Implicit Function Theorem, but in the interests of brevity, and since demonstrating (3.175) was the main point of the current example, we will omit the argument.) Thus the general solution of (3.169), in implicit form, is

$$\{x^3 + x + xy^{10} + xy^2 = C, \quad C \neq 0\}. \tag{3.178}$$



In the example above, facts 3.175, 3.176, and 3.177 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 3.86 shows is that if you try to solve a differential equation by mindlessly pushing differentials around the page as if they were numbers, the answer you wind up with may not be the set of solutions to the equation you were trying to solve. In fact, when you realize how dissimilar differentials and numbers are, it should initially strike you as miraculous that you can even get *close* to the correct set of solutions by such manipulations.

One 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 (3.175) 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 3.86 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 $M(x, y) dx + N(x, y) dy = 0$, the same collection \mathcal{E} will also be an implicit form of the general solution of the associated derivative-form DE $M(x, y) + N(x, y) \frac{dy}{dx} = 0$.** In “most” of the exceptions to this rule, we need only delete one or a few of the equations from \mathcal{E} to obtain the general solution of the derivative-form DE (in implicit form).

The simplest of these exceptions are equations that are explicitly of the form “ $x = \text{some specific constant}$ ”, or are equivalent to an equation of this form, such as equation (3.174). It is obvious that equations *written this way* are not implicit solutions of a derivative-form DE with x as independent variable, but when a whole *family* of equations is given, such as $x^3 + x + xy^{10} + xy^2 = C$ (equation (3.172)), it may take some work and cleverness to determine whether there are members of this family that are equivalent to “ $x = \text{specific constant}$ ”.

The next example involves simpler differential equations than Example 3.86, but a more complicated “spurious solution”.

Example 3.87 Consider the differential-form DE

$$(y^2 + 1) \cos x \, dx + 2y \sin x \, dy = 0 \quad (3.179)$$

and the associated derivative-form DE

$$(y^2 + 1) \cos x + 2y \sin x \frac{dy}{dx} = 0. \quad (3.180)$$

Equation (3.179) is exact. Its general solution is

$$\{(y^2 + 1) \sin x = C\} \quad (3.181)$$

where C is an arbitrary constant. For $C \neq 0$, every point (x, y) in the graph of (3.181) 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 (3.180); the general solution of (3.180), in implicit form, can be written as

$$\{y^2 + 1 = C \csc x, \quad C \neq 0\} \quad (3.182)$$

or as

$$\{(y^2 + 1) \sin x = C, \quad C \neq 0\}. \quad (3.183)$$

However, for $C = 0$, equation (3.181) 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 (3.180).

So in this example, we again need to throw away only one equation from the algebraic form (3.181) 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 $M(x, y) dx + N(x, y) dy = 0$ but 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 a collection \mathcal{E} of equations that is the general solution of the differential-form DE, and we remove from this collection all equations whose graphs have the two properties above, then the remaining collection of equations is an implicit form of the general solution of the associated derivative-form DE. “Most of the time”, there will be *no* such equations in our original collection \mathcal{E} , in which case the same collection \mathcal{E} serves as both 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

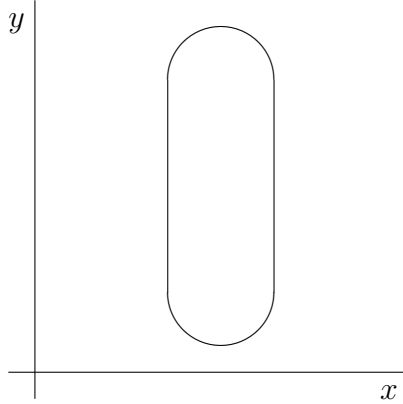


Figure 8:

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 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

$$\frac{dy}{dx} = f(x, y) \quad (3.184)$$

or, more generally, an “almost-standard form” DE

$$f_1(x, y) \frac{dy}{dx} = f_2(x, y) \quad (3.185)$$

If (3.184) or (3.185) 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

$$-f(x, y) dx + dy = 0 \quad (3.186)$$

and

$$-f_2(x, y) dx + f_1(x, y) dy = 0. \quad (3.187)$$

If we are extremely lucky, then (3.186) or (3.187) will be exact.

In the case of (3.186), 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 (3.184) was already of the form $\frac{dy}{dx} = g(x)$, solvable just by integrating g ; there is no need even to look at equation (3.186).

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

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

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

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

↓ subtract $2y^3 \sin x \cos x$ (yielding an algebraically equivalent DE)

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

↓ 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

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

↓ subtract $\frac{2y^3 \sin x \cos x}{3y^2 \cos^2 x + 1}$
(yields an algebraically equivalent equation)

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

↓ write the associated differential-form DE

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

↓ 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

$$\{y + y^3 \cos^2 x = C\} \tag{3.188}$$

(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 (3.177) guarantees us that the family of equations (3.188) *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 (3.188) contains a general solution, in implicit form, of our original derivative-form DE. The only question is whether the family (3.188) 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 (3.188) has such a graph⁸⁴. Therefore (3.188) is the general solution of the derivative-form equation that we started with, $\frac{dy}{dx} = \frac{2y^3 \sin x \cos x}{3y^2 \cos^2 x + 1}$.

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

1. Perform any algebraic manipulations that may be necessary to put the DE into “almost-standard” form $f_1(x, y) \frac{dy}{dx} = f_2(x, y)$ or $-f_2(x, y) + f_1(x, y) \frac{dy}{dx} = 0$. Each time we perform an algebraic manipulation, keep track of the region(s) on which the manipulation gives us an algebraically equivalent DE.
2. Write down the differential-form DE associated with our last derivative-form DE. If this DE does not pass the test for exactness, look for additional algebraic manipulations that may yield an exact DE (we may not find any). Again, keep track of the region(s) on which any algebraic manipulations we use give us an algebraically equivalent DE.
3. Assuming we have now produced an exact DE on some region(s) R_1, R_2, \dots , 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

⁸⁴One 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 $1 + 3y^2 \cos^2 x_0 \geq 1$.

\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.

The last step in the procedure above is not one for which we will try to state general rules; instead, we will illustrate with an example the sort of work that must be done.

Example 3.88 Solve the differential equation

$$\frac{dy}{dx} = -\frac{2x + 2y}{2x + 3y^2}. \quad (3.189)$$

First we observe that since the right-hand side of (3.189) is not defined when $2x + 3y^2 = 0$, the only regions in which “solution of (3.189)” 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, (3.189) is algebraically equivalent to

$$(2x + 2y) + (2x + 3y^2)\frac{dy}{dx} = 0, \quad (3.190)$$

whose associated differential-form equation is

$$(2x + 2y)dx + (2x + 3y^2)dy = 0. \quad (3.191)$$

Equation (3.191) 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 (3.191) 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 (3.189) is guaranteed to be a solution of (3.190), so in passing from (3.189) to (3.190) 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 (3.190) to (3.191). The latter possibility is easy to rule out: it is easy to see that (3.191) has no solutions of the form $x = \text{constant}$. (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 (3.191) is a solution curve of (3.190)

To see whether the graph of $x^2 + 2xy + y^3 = C$, for a given C , is an implicit solution of (3.189) 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 (3.141) assures us that the general solution of (3.191) on $\{\mathbf{R}^2$ minus the origin $\}$ is $x^2 + 2xy + y^3 = C$, where C can be any value in the range of F on this domain. By holding x fixed (say $x = 1$) and letting y vary over \mathbf{R} , we see that the range of F on this domain is the set of all real numbers. Therefore the general solution of (3.191) 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 (3.189) to (3.190) introduced any spurious solutions: equations $x^2 + 2xy + y^3 = C$ that are not implicit solutions of (3.189). 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 $(-\frac{3}{2}y^2)^2 + 2(-\frac{3}{2}y^2) + y^3 = C$, the latter equation simplifying to $\frac{9}{4}y^4 - 2y^3 = C$. No matter what the value of C is, there are at most four numbers y for which $\frac{9}{4}y^4 - 2y^3 = C$, so the graph of $2x + 3y^2 = 0$ intersects the graph of $x^2 + 2xy + y^3 = C$ in at most four points. But the portion of the graph of $x^2 + 2xy + y^3 = C$ that lies in $\{\mathbf{R}^2$ minus the origin $\}$ —the whole graph unless $C = 0$ —is a smooth curve \mathcal{C} . Deleting from \mathcal{C} the at-most-four points of \mathcal{C} for which $2x + 3y^2 = 0$, what remains is one or more curves each of which lies entirely in R_1 or R_2 , and hence is a solution-curve of (3.189). Therefore there are no values of C that we need to exclude, and no spurious solutions. The general solution of (3.189) 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 (3.189).)

■

In the preceding example, the passage from (3.189) to (3.191) did not introduce any spurious solutions. In the earlier examples, whenever spurious solutions were introduced, they were of the form $x = \text{constant}$. 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 $x = \text{constant}$. The answer is yes. Failure to preserve algebraic equivalence can lead to spurious solutions not of the form “one variable = constant” whether we are working with derivative-form or differential-form DEs. The next example could have been presented before we ever talked about differential form, but we have placed it in this section of the notes as a reminder.

Example 3.89 (A spurious solution not of the form $x = \text{constant}$) 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

$$\frac{dy}{dx} = f(x, y), \quad y(0) = 0, \tag{3.192}$$

has a unique maximal solution.

If we substitute the definition of $f(x, y)$ into (3.192), the DE becomes

$$\frac{dy}{dx} = \begin{cases} \frac{e^y - e^x}{y - x} & \text{if } y \neq x, \\ e^x & \text{if } y = x. \end{cases} \tag{3.193}$$

This equation is neither linear nor separable, so in an attempt to solve we might write down the associated differential-form equation, which is

$$- \left\{ \begin{array}{ll} \frac{e^y - e^x}{y - x} & \text{if } y \neq x \\ e^x & \text{if } y = x \end{array} \right\} dx + dy = 0. \tag{3.194}$$

It is natural to try to rewrite (3.194) 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 (3.194) by $y - x$ we obtain

$$- (e^y - e^x)dx + (y - x)dy = 0, \tag{3.195}$$

which certainly *looks* much simpler than (3.194). 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 (3.192)?

The answer is a resounding “No!”. If we define $\phi(x) = x$, and substitute $y = \phi(x)$ into “ $\frac{dy}{dx} = f(x, y)$ ”, then the left-hand side is identically 1, while the right-hand side is e^x . There is no x -interval on which $e^x \equiv 1$. The function ϕ is not a solution of $\frac{dy}{dx} = f(x, y)$.

It is easy to see what went wrong if, instead of writing (3.194) with the two-line formula for f , we write it as

$$- f(x, y)dx + dy = 0 \tag{3.196}$$

and if, when we multiply through by $y - x$, we write the result as

$$-(y-x)f(x,y)dx + (y-x)dy = 0 \quad (3.197)$$

rather than in the “simpler” form (3.195). It is obvious that $y = x$ is a solution of (3.197), whether or not it is a solution of (3.196). Less obvious, but true, is what we checked above: that $y = x$ is *definitely not* a solution of (3.192), hence not a solution of (3.196).

In this example, the general solution of (3.197) consists of the general solution of (3.196) *plus* the straight line $\{y = x\}$. The equation (3.196) has no solutions of the form $x = \text{constant}$, so any algebraic form of the general solution of (3.196) 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 (3.195), 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 (3.194) and (3.195) 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. ■

3.5 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 3.90 Consider the equation

$$(2y + 3x)dy + 5dx = 0. \quad (3.198)$$

This equation is not exact. The simplest associated derivative-form DEs are

$$(2y + 3x)\frac{dy}{dx} + 5 = 0 \quad (3.199)$$

and

$$2y + 3x + 5\frac{dx}{dy} = 0. \quad (3.200)$$

Equation (3.199) is nonlinear, but (3.200) is linear. We can solve the latter DE by our usual method for linear DEs (remembering that in equation (3.200), y is the independent variable and x is the dependent variable, not the other way around), obtaining the general solution

$$\left\{ x = -\frac{2}{3}y + \frac{10}{9} + Ce^{-\frac{3}{5}y} \right\}. \quad (3.201)$$

From the discussion in Section 3.3, the graph of each equation in the collection (3.201) is a solution-curve of the differential-form DE (3.198), and the only *potential* solution curves of (3.198) not represented in (3.201) 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 (3.198) to (3.200).) But since $5 \neq 0$, there are no values of c for which “ $y = c$ ” is a solution of (3.198) (see Remarks 3.76 and 3.85). Thus the collection (3.201) is (an algebraic form of) the general solution of equation (3.198).

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 3.1.8 and 3.3 are involved.

Example 3.91 We will solve the equation

$$(2y - xe^x)dx + x dy = 0. \quad (3.202)$$

Equation (3.202) is not exact on any region in the xy plane, but the associated derivative-form DE

$$2y - xe^x + x \frac{dy}{dx} = 0 \quad (3.203)$$

is linear, so we can obtain *most* (if not all) of the solutions of equation (3.202) by solving equation (3.203).

To solve (3.203) by our usual integrating-factor method, we divide-through by x and rewrite the new equation as

$$\frac{dy}{dx} + \frac{2}{x}y = e^x. \quad (3.204)$$

However, because of the division by x , the linear equations (3.203) and (3.204) 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 (3.203) whose domain includes $x = 0$.

Solving equation (3.204) by the integrating-factor method yields the collection of equations

$$\{y = e^x(1 - 2x^{-1} + 2x^{-2}) + Cx^{-2}\}, \quad (3.205)$$

representing one 1-parameter family of solutions of (3.204) on $(-\infty, 0)$ and another on $(0, \infty)$.

To see if we lost any solutions of (3.203) when passing to (3.204)—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 (3.204), 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 $\lim_{x \rightarrow 0^+} \phi(x)$ exists if and only if $C_1 = -2$, and $\lim_{x \rightarrow 0^-} \phi(x)$ 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 $(-\infty, \infty)$,

$$\phi_{\text{special}}(x) = \begin{cases} e^x(1 - 2x^{-1} + 2x^{-2}) - 2x^{-2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases} \quad (3.206)$$

is differentiable at $x = 0$. Thus, substituting “ $y = \phi_{\text{special}}(x)$ ” into equation (3.203) 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 (3.203) 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 (3.203) consists of (i) the graphs of all the equations in (3.205) on the x -interval $(-\infty, 0)$, (ii) the graphs of all the equations in (3.205) 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 (3.203), but must determine whether any solution-curves of (3.202) were lost when we passed from (3.202) to (3.203). From the discussion in Section 3.3, the only potential solution-curves of (3.202) that are not solution-curves of (3.203) are vertical lines, graphs of equations of the form $x = c$ (where c is a constant). Plugging into (3.202), there is one and only one value of c , namely 0, for which $x = c$ is a solution of (3.202). (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 3.76 and 3.85).

Thus (an algebraic form of) the general solution of equation (3.202) is

$$\{y = e^x(1 - 2x^{-1} + 2x^{-2}) + Cx^{-2}\} \cup \{y = \phi_{\text{special}}(x)\} \cup \{x = 0\}.$$
⁸⁵

⁸⁵ 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 $x \rightarrow 0$.



Remark 3.92 The differential on the left-hand side of equation (3.202) 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 3.82. In Example 3.91, 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 (3.202) 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 3.91 did indeed find all the solution curves of (3.202).

4 Optional Reading

4.1 The meaning of a differential

For the interested student, in this section we ascribe meaning to a differential.⁸⁶ 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 $Mdx + Ndy$ is a machine with an input and an output. What it takes as input is a (differentiably) parametrized curve γ . What it then outputs is a *function*, defined on the same interval I as γ . If we write $\gamma(t) = (x(t), y(t))$, then the output is the function whose value at $t \in I$ is $M(x(t), y(t))\frac{dx}{dt} + N(x(t), y(t))\frac{dy}{dt}$.

We use the language “ $Mdx + Ndy$ acts on γ ” to refer to the fact that the differential takes γ as an input and then “processes” it to produce some output. Notation we will use for the output function is $(Mdx + Ndy)[\gamma]$. This is the same function that we expressed in terms of t in the previous paragraph:

⁸⁶Differentials 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.

the function obtained
when the differential
acts on γ

$$\underbrace{(M dx + N dy)[\gamma]}_{\substack{\text{value of the function} \\ (M dx + N dy)[\gamma] \\ \text{at } t}}(t) = M(x(t), y(t)) \frac{dx}{dt} + N(x(t), y(t)) \frac{dy}{dt}. \quad (4.1)$$

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.

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

$$\int_{\mathcal{C}} M(x, y) dx + N(x, y) dy. \quad (4.2)$$

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 $\gamma(t) = (x(t), y(t))$, $t \in [a, b]$.⁸⁷ 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

$$\int_a^b \left\{ M(x(t), y(t)) \frac{dx}{dt} + N(x(t), y(t)) \frac{dy}{dt} \right\} dt. \quad (4.3)$$

3. Compute the integral (4.3). The definition of (4.2) is the value of (4.3):

⁸⁷The parametrization should also be 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 are unimportant here; the author is trying only to jog the student's memory, not to review line integrals thoroughly.

$$\int_{\mathcal{C}} M(x, y)dx + N(x, y)dy = \int_a^b \left\{ M(x(t), y(t))\frac{dx}{dt} + N(x(t), y(t))\frac{dy}{dt} \right\} dt. \quad (4.4)$$

(You also learn in Calculus 3 that this definition is self-consistent: no matter what regular parametrization of \mathcal{C} you choose⁸⁸, you get the same answer.)

A casual glance at (4.4) suggests that we have used the following misleading equality:

$$“M(x, y)dx + N(x, y)dy = \left\{ M(x(t), y(t))\frac{dx}{dt} + N(x(t), y(t))\frac{dy}{dt} \right\} dt.” \quad (4.5)$$

But that is not quite right. The left-hand side and right-hand side are not the same object. Only *after we are given a parametrized curve* γ can we produce, from the object on the left-hand side, the function of t in braces on the right-hand side.

In addition, in constructing the integral on the right-hand side of (4.4), we did not confine our substitutions to the *integrand* of the integral on the left-hand side. We made the substitution “ $\int_{\mathcal{C}} \rightarrow \int_a^b$ ” 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 $(Mdx + N dy)[\gamma](t)$ 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 $(Mdx + N dy)[\gamma]$ 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

⁸⁸Subject to the other conditions in the previous footnote.

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 3.80, assume that $Mdx + Ndy$ 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 $Mdx + Ndy = 0$ 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 (3.138) 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 3.6) 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 (3.136) 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

$$\left\{ \begin{array}{l} x(t) = t \\ y(t) = \phi(t) \end{array} \right\}, \quad t \in I_1.$$

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 (3.138). 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 (3.136) 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

$$\left\{ \begin{array}{l} x(t) = \phi(t) \\ y(t) = t \end{array} \right\}, \quad t \in J_1.$$

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 (3.138), 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 $R = \{\mathbf{R}^2 \text{ minus the origin}\}$, 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 3.80, we view as $\{C \in \mathbf{R} \mid C > 0\}$. Therefore the general solution of $x \, dx + y \, dy = 0$ in R is $\{\frac{1}{2}(x^2 + y^2) = C \mid C > 0\}$, which, by renaming the constant, we can write more simply as

$$\{x^2 + y^2 = C \mid C > 0\}. \tag{4.7}$$

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 3.80), 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 3.80 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 3.75,

$$y \, dx + x \, dy = 0. \tag{4.8}$$

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 $R = \{\mathbf{R}^2 \text{ minus the origin}\}$, 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 3.75, 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.⁸⁹) 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

⁸⁹*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.

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 3.80 and 4.1 we can deduce the following:

$$\left. \begin{array}{l} \text{If } F \text{ has continuous second partial derivatives in the region} \\ R, \text{ then the set of solution curves of } dF = 0 \text{ on } R \text{ is the set} \\ \text{of smooth curves in } R \text{ that are contained in level-sets of } F. \end{array} \right\} \quad (4.10)$$

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 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 3.81 elsewhere in these notes.

4.3 One-parameter families of equations

Suppose that G is a three-variable function with the property that

$$\begin{array}{l} \text{for each } z \in \mathbf{R}, \text{ there is } \textit{some} \text{ point } (x, y) \in \mathbf{R}^2 \\ \text{for which } (x, y, z) \text{ is in the domain of } G. \end{array} \quad (4.11)$$

Then for each $C \in \mathbf{R}$, the equation $G(x, y, C) = 0$ is an algebraic equation in the variables x and y .⁹⁰ The collection of equations

$$\{G(x, y, C) = 0 \mid C \in \mathbf{R}\} \quad (4.12)$$

is an example of a *one-parameter family of (algebraic) equations* in variables x and y .⁹¹ (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

$$\{G_1(x, y, C) = G_2(x, y, C) \mid C \in \mathbf{R}\} \quad (4.13)$$

⁹⁰For 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.

⁹¹This G should not be confused with the G in equation (3.2), which is being used to describe a *single, differential* equation, not a collection of algebraic equations.

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 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 equations⁹², 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**

⁹²“Ease”, when the DE falls into one of the categories for which systematic methods of solution are taught in an introductory course in DEs.

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 each equation in the family is an implicit solution of the DE.⁹³ ■

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, $\{y + \frac{1}{x-C} = 0\}$.) 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).

⁹³*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 $C \rightarrow \infty$ 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.

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.⁹⁴ ■

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 3.80). 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 (3.81). For simplicity, let us assume that both the interval I and set D in (3.81) 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 $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$

⁹⁴*Note to instructors:* It may seem odd that I am using this terminology here, when I argued in Remark 3.73 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 3.73 we were talking, simultaneously, about *all* explicit algebraic equations in x and y . In such a discussion there is no reasonable way to exclude 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.

defined in (3.82)–(3.83). The collection \mathcal{E}_1 is a one-parameter family of equations; the collection \mathcal{E} is not.

But, in the notation of Theorem 3.43, 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 3.43, since $H(y) - G(x) - C = 0$ is equivalent to $H(y) = G(x) + C$). Thus, for each C , the 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 3.38. ■

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 The Fundamental Theorem of ODEs

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. The first-order case is the theorem below. In this theorem, “ $\frac{\partial f}{\partial y}$ ” denotes the partial derivative of f with respect to its second variable.

Theorem 5.1 (FTODE) *Let f be a function of two variables, and consider the initial-value problem*

$$\frac{dy}{dx} = f(x, y), \quad y(x_0) = y_0. \quad (5.1)$$

Assume that f and $\frac{\partial f}{\partial y}$ are continuous on some open set R (see Section 3.1.5) containing the point (x_0, y_0) . Then there exists a number $\delta > 0$ such that for every interval I_1 that contains x_0 and is contained in $(x_0 - \delta, x_0 + \delta)$, the initial-value problem (5.1) has a unique solution on I_1 .

Most textbooks (including [1], [3], and [4]), state a version of this theorem that is far too weak to be useful, effectively replacing the last sentence with, “Then there exists a number $\delta > 0$ such that the initial-value problem (5.1) has a unique solution on $(x_0 - \delta, x_0 + \delta)$.” This weaker statement allows the possibility that (5.1) has a unique solution on $(x_0 - \delta, x_0 + \delta)$, but has more than one solution on a smaller interval, e.g. $(x_0 - \frac{\delta}{2}, x_0 + \frac{\delta}{2})$, a phenomenon ruled out by the more-carefully stated theorem above.⁹⁵ Textbooks that state the weaker theorem sometimes implicitly *use* Theorem 5.1, without observing it is not implied by the weaker version.

A rather unimportant difference between the statement of Theorem 5.1 and the version found in most textbooks is that we have said “open set” where most textbooks say “open rectangle” (or a phrase that has the same meaning as “open rectangle”). It is easy to show that the theorem above is equivalent to the one we would obtain by replacing “open set” with “open rectangle”. We have stated the theorem using “open set” since that wording makes the theorem easier to apply in this notes. Presumably, the reason that most textbooks state only the “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 enormous.

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.

⁹⁵*Note to instructors:* From Theorem 5.1, but not from the weaker version, the following can be deduced: If I_1, I_2 are intervals contained in (a, b) and containing x_0 , and ϕ_1 and ϕ_2 are solutions of (5.1) on I_1 and I_2 , respectively, then ϕ_1 and ϕ_2 agree on $I_1 \cap I_2$. This fact is of critical importance to the notion of “maximal domain of a unique solution”. The only textbook I’ve looked at recently that states this useful a form of the FTODE is [2].

- [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] Zill and Wright, *Differential Equations with Boundary Value Problems*, 8th edition, Brooks/Cole, 2013.