Notes by David Groisser, Copyright ©2010

First-order ODEs: Derivative form, Differential Form, and Implicit Solutions

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

Introduction

First-order ODEs come in two forms: *derivative form* and *differential form*. The two forms are closely related, but differ in subtle ways not addressed adequately in most textbooks (and often overlooked entirely)[1]. This often leads to an unclear or inadequate definition of "implicit solution" to an equation in derivative form, before differential-form equations have even been introduced.

The purpose of these notes is to give a definition of "implicit solution" that is accurate, complete, and unambiguous. In order to make our presentation readable concurrently with a DE textbook whose topics appear in a traditional order, we define "implicit solutions of a DE in derivative form" before we even introduce 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 in these notes, we come back to derivative-form equations to clean up the picture.

The first section below is written for mathematicians; 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.

1 Notes for Instructors

[This section is not yet written]

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

2 Notes for Students

2.1 Review of "derivative form" and "solution"

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

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

$$\mathsf{F}(x, y, \frac{dy}{dx}) = 0, \tag{1}$$

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

Definition 2.1 For a given F, a *solution of* (1) *on an open interval* I is a real-valued differentiable function ϕ on I such that when "$y = \phi(x)$" is substituted into (1), the resulting equation is a true statement for all $x \in I$.[2]

For a given F, we call a one-variable function ϕ a *solution of* (1) (no interval mentioned) if ϕ is a solution of (1) on *some* open 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.[3]

If ϕ is a solution of a given DE (perhaps with an interval specified, perhaps not) whose dependent and independent variables are y and x respectively, we allow ourselves the freedom to say that the *equation* "$y = \phi(x)$" is a solution of the DE. This

[2] Some authors refer to what we have just defined as an *explicit solution* of (1) on I. This use of "explicit" is intended to help students understand later, by way of contrast, what an *implicit solution* is. But the author of these notes feels that the terminology "explicit solution" 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".

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

allows us the convenience of being able to say, for example, "$y = x^2$ is a solution of $\frac{dy}{dx} = 2x$" without having to introduce extra notation (e.g. the letter ϕ we have been using) for the squaring function. This is an example of "permissible abuse of terminology". An equation and a function are two different animals, and we should not forget the fact that, by definition, a solution of a DE is a one-variable *function.* But once we understand what "solution of a DE" means, we allow ourselves the luxury of saying, imprecisely, that "$y = x^2$ is a solution of $\frac{dy}{dx} = 2x$" instead of the precise but awkward, "The function ϕ defined by $\phi(x) = x^2$ is a solution of $\frac{dy}{dx} = 2x$."

2.2 Implicit solution of a derivative-form DE

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

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

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

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

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

then

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

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

$$G(x, y) = 0 \quad \text{if and only if} \quad y = \phi(x). \tag{4}$$

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

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

More generally than this example, any time (4) is true for a two-variable function G and one-variable function ϕ, we say that the equation $G(x, y) = 0$ *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 $G(x,y)=0$.

Now consider the equation

$$x^2+y^2-1=0. \tag{5}$$

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

$$y=\pm\sqrt{1-x^2}. \tag{6}$$

Looking just at (5), it is already clear that any numerical choice of x restricts the possible choices of y that will make the equation a true statement. Equation (6) tells us the only possible values for y that will work. It also tells that for $-1<x<1$ 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 (5), we see that all the values of y that we have labeled as "possible" actually do work. Thus

$$x^2+y^2-1=0 \text{ if and only if } |x|\leq 1 \text{ and either } y=\sqrt{1-x^2} \text{ or } y=-\sqrt{1-x^2}. \tag{7}$$

This is a *much* weaker statement than a statement of the form (4), because the sign in $\pm\sqrt{1-x^2}$ can be chosen independently for each x. On the domain $[-1,1]$, if we define

$$\begin{aligned}
\phi_1(x) &= \sqrt{1-x^2}, && (8)\\
\phi_2(x) &= -\sqrt{1-x^2}, && (9)\\
\phi_3(x) &= \begin{cases} \sqrt{1-x^2} & \text{if } x \text{ is a rational number,}\\ -\sqrt{1-x^2} & \text{if } x \text{ is an irrational number,}\end{cases} && (10)
\end{aligned}$$

then all three of these functions yield true statements, for all $x\in[-1,1]$, when substituted in as y in (5). In fact, since the sign "±" can be assigned randomly for each $x\in[-1,1]$, there are *infinitely many* functions ϕ that work. What distinguishes ϕ_1 and ϕ_2 from all the others is that they are *continuous.* If we restrict their domains to the open interval $(-1,1)$, then they are even differentiable.

Now consider a more complicated equation, such as

$$e^x+x+6y^5-15y^4-10y^3+30y^2+10xy^2=0. \tag{11}$$

Clearly, choosing a numerical value for x restricts the possible values for y that will make (11) a true statement. It turns out that, depending on the choice x, there can be anywhere from one to five values of y for which the pair (x,y) satisfies (11). As in the previous example, on any x-interval I for which there is more than one

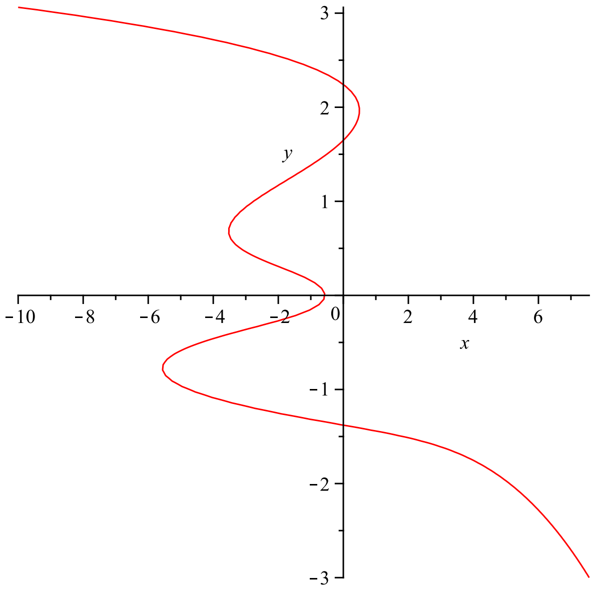

Figure 1: The graph of $e^x + x + 6y^5 - 15y^4 - 10y^3 + 30y^2 + 10xy^2 = 0$.

y-value that "works" for each x, there will be infinitely many functions ϕ for which $G(x, \phi(x)) = 0$, where $G(x, y)$ is the left-hand side of (11). However, there are not very many *continuous* ϕ's that work. In this example, whatever x-interval I we choose, there can are at most five continuous functions ϕ defined on I for which $G(x, \phi(x)) = 0$. Writing out *explicit formulas* for them, analogous to the formulas for ϕ_1 and ϕ_2 in the previous example, is a hopeless task. But these continuous functions ϕ exist nonetheless. We can see this visually in Figure 1.

Definition 2.2 Let G be a function of two variables, ϕ a function of one variable, and I an interval. We say that the equation $G(x, y) = 0$ *implicitly determines* or *implicitly defines* the function ϕ, regarded as a function of x (or whatever name is used for the first variable of G), if $G(x, \phi(x)) = 0$ for all $x \in I$.

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

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

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

There are instances in which we are interested in whether there is one-variable function ϕ such that $G(\phi(y), y) = 0$. This comes up when we think of trying to solve the equation $G(x, y) = 0$ for x in terms of y, rather than for y in terms of x. To handle this case we can give a definition analogous to Definition 2.2, replacing the phrases "regarded as a function of x" and "first variable" with "regarded as a function y and "second variable", and replacing "$G(x, \phi(x)) = 0$ with "$G(\phi(y), y) = 0$". To simplify wording below, any time we say an equation $G(x, y) = 0$ implicitly determines (or defines) a function ϕ, we mean to regard ϕ as a function of x, unless we say otherwise.

Thus:

- Equation (2) implicitly determines the function ϕ given by the formula $\phi(x) = (1 - x^2)^{1/3}$.

- Equation (5) implicitly determines the functions ϕ_1, ϕ_2, ϕ_3 defined in (8)–(10), and infinitely many others on the interval $[-1, 1]$. The only *continuous* functions that (5) determines on $[-1, 1]$ are ϕ_1 and ϕ_2.

- Equation (11) implicitly determines infinitely many functions, but only a few continuous functions. In Figure 1, if we travel along the graph by starting at the upper left and moving along the curve, we encounter vertical tangents at points A, B, C, and D (labeled in the order that we encounter them). Let x_A, x_B, x_C, and x_D denote the x coordinates of these points. Then (11) implicitly determines a continuous function of x, say ϕ_1, with domain $(-\infty, x_A]$; another continuous function of x, say ϕ_2, with domain $[x_B, x_A]$; another, say ϕ_3, with domain $[x_B, x_C]$; another, say ϕ_4, with domain $[x_D, x_C]$; and another, say ϕ_5, with domain $[x_D, \infty]$. On the interval $[-3, -2]$, the equation $G(x, y) = 0$ determines five continuous functions (the restrictions of $\phi_1, \phi_2, \phi_3, \phi_4$, and ϕ_5 to this interval). On the interval $[-5, -4]$, $G(x, y) = 0$ determines three continuous functions (the restrictions of ϕ_1, ϕ_4, and ϕ_5 to this interval).

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

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

In these notes we will call such a set a *rectangle*, even though we do not exclude the possibility that I and/or J extend(s) infinitely in one direction or both. Thus, for

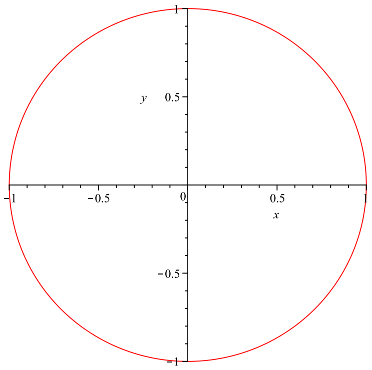

Figure 2: The graph of $x^2 + y^2 = 1$.

example, we consider the whole xy plane a rectangle; the set $[1, \infty) \times (-\infty, \infty)$ is a rectangle (consisting of all pairs (x, y) for which $x > 1$); the strip $(-\infty, \infty) \times (0, 1]$ is a rectangle (consisting of all pairs (x, y) with $0 < y \leq 1$). Of course, objects that Euclid would have called rectangles, such as $[1, 2] \times [3.1, 4.9]$, are also rectangles in our terminology. In these notes, we will be most interested in *open* rectangles, those we get by taking the intervals I and J to open.

When an equation $G(x, y) = 0$ 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 (8)–(9) ($\phi_1(x) = \sqrt{1 - x^2}$, $\phi_2(x) = -\sqrt{1 - x^2}$). For example, if $y_0 > 0$ you can take J to be any open subinterval of $(0, \infty)$ that contains y_0, and then take I to be any open interval whatsoever that contains x_0. Choose some points on the graph in Figure 2 and draw rectangles around them with the desired property.

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

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

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

The *Implicit Function Theorem* gives conditions under which the "windowing near a point (x_0, y_0)" idea works very nicely to guarantee that an equation such as "$G(x, y) = 0$" determines at least one differentiable function of x, and, if it determines more than one such function, to use (x_0, y_0) to single out one of them:

Theorem 2.3 (Implicit Function Theorem) *Let G be a two-variable function whose first partial derivatives are continuous on an open rectangle $R = I \times J$. Suppose that $(x_0, y_0) \in R$ and that $\frac{\partial G}{\partial y}(x_0, y_0) \neq 0$, where $\frac{\partial G}{\partial y}$ denotes the partial derivative of G with respect to the second variable. Let $c_0 = G(x_0, y_0)$.*

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

$$\begin{aligned} &\text{for all points } (x, y) \in I_1 \times J_1, \\ &\qquad G(x, y) = c_0 \text{ if and only if } y = \phi(x). \end{aligned} \tag{12}$$

■

Since x_0 lies in I_1, we may look at what (12) tells us when $x = x_0$. What this statement reduces to when $x = x_0$ is the following:

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

But by the definition of c_0, we have $G(x_0, y_0) = c_0$. 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 (12) says that for each $x \in I_1$, there is *one and only one* value $y \in J_1$ for which $G(x, y) = c_0$, namely the value $\phi(x)$. Thus, (12) says that within $I_1 \times J_1$, the equation $G(x_0, y_0) = 0$ determines y *uniquely* as a function of x. Not just uniquely among "nice" functions, like continuous or differentiable functions. Among *all* functions with domain I_1 and range contained in J_1, ϕ is the *only* function that satisfies $G(x, \phi(x)) = c_0$ 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 $G(x, \phi(x)) = c_0$ identically in x.

Compare statement (12) with statement (4). The only important difference is that to get the second line of (12), we had to make the windowing restriction in the first line. (The fact that we have "c_0" in (12) where we have "0" in (4) is an unimportant difference.) This is usually the best we can do; only occasionally do we have situations in which we can take the "window" to be the whole xy plane and still get a unique implicitly-defined function.

The uniqueness of a function ϕ that is guaranteed by a statement of the form (12) allows us to use terminology that is less awkward than what we used in Definition 2.2. Specifically, whenever a statement of the form (12) holds true, we can dispense with the phrase "regarded as a function of the first variable of G" in that definition, or even naming the function ϕ at all. We may simply say the following:

> Within the rectangle $I_1 \times J_1$, the equation $G(x, y) = c_0$ determines y uniquely as a function of x.

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 G); 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

$$\begin{array}{l}\text{for all points } (x, y) \in I_1 \times J_1, \\ \qquad\qquad G(x, y) = c_0 \ \text{ if and only if } \ x = \phi(y)\end{array} \tag{13}$$

then we can say simply that within the rectangle $I_1 \times J_1$, the equation $G(x, y) = c_0$ determines x uniquely as a function of y. Thus, when condition (13) is met, we do not have to write a whole new definition analogous to Definition 2.2, with "regarded as a function of the first variable" replaced with "regarded as a function of the second variable", and with "$G(x, \phi(x)) = 0$" replaced with "$G(\phi(y), y) = 0$".

When either (12) or (13) holds for some rectangle $I_1 \times J_1$, we call ϕ an *implicitly-defined function.*

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

Now, let us get back to differential equations:

Definition 2.4 (temporary) We call an equation $G(x,y) = 0$ an *implicitsolution* (one word, for now) of a differential equation

$$\mathsf{F}(x, y, \frac{dy}{dx}) = 0 \tag{14}$$

(for a given F) if

(i) the equation $G(x,y) = 0$ implicitly determines at least one function ϕ that is a solution of (14), and

(ii) *every* differentiable function ϕ determined by the equation $G(x,y) = 0$ on an open interval is a solution of (14). ■

Definition 2.5 If ϕ is a differentiable function determined implicitly by an implicitsolution $G(x,y) = 0$ of (14), then we call ϕ an *implicitly-defined* solution of (14). ■

Example 2.6 Consider the differential equation

$$x + y\frac{dy}{dx} = 0. \tag{15}$$

We claim that the equation

$$x^2 + y^2 - 1 = 0 \tag{16}$$

is an implicitsolution of (15). (Equivalently, so is the equation $x^2+y^2 = 1$.) To verify this, we check that criteria (i) and (ii) of Definition 2.4 are satisfied:

- Criterion (i). Let $\phi_1(x) = \sqrt{1-x^2}$ as in (8), but restricted to the open interval $(-1,1)$. Note that $G(x, \phi_1(x)) = 1$ for all $x \in (-1,1)$, so ϕ_1 is a function implicitly determined by the equation $G(x,y) = 1$ (the conditions of Definition 2.2) 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 (15), 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 (15). Thus criterion (i) is satisfied[4].

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

- Criterion (ii). Suppose ϕ is any differentiable function determined implicitly by (16) on some open interval I. Then we have

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

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

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

  Therefore ϕ is a solution of the equation

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

  on I. Dividing by 2 we see that ϕ is a solution of (15) on I. Therefore criterion (ii) is satisfied.

Hence (16) is an implicitsolution of (15), and the function ϕ_1 is an implicitly-defined solution of (15).

There are actually two implicitly-defined solutions in this example: ϕ_1 and $-\phi_1$ (the function that we called ϕ_2 in (9)). The first of these is the function implicitly defined by $x^2 + y^2 = 1$ on the rectangle $(-1, 1) \times (0, \infty)$; the second is the function implicitly defined by $x^2 + y^2 = 1$ on the rectangle $(-1, 1) \times (-\infty, 0)$. Both functions are solutions of (15). ■

Example 2.7 We claim that

$$(y - e^x)(x^2 + y^2 - 1) = 0 \tag{17}$$

is *not* an implicitsolution of (15). To verify this claim, it suffices to show that *at least one* of criteria (i) and (ii) in Definition 2.4 is not met. For this, we observe that if $y = e^x$, then (17) is satisfied. Thus, the function ϕ defined on any open interval I by $\phi(x) = e^x$ is a function determined implicitly by (17). However, if we substitute $y = e^x$ into (15), we get

$$x + e^{2x} = 0. \tag{18}$$

Is it possible to choose the interval I in such a way that (18) holds true for all $x \in I$? No, for if there were such an interval I, the left-hand side of (18) would be a differentiable function on I, so we could differentiate both sides of (18) and obtain

$$1 + 2e^{2x} = 0. \tag{19}$$

But there isn't even a single value of x for which this is true; $1+2e^x>0$ for all x. Thus there is no open interval I on which ϕ is a solution of (15).

Thus ϕ is a differentiable function determined implicitly by (17) that is not a solution of (15). Therefore criterion (ii) in Definition 2.4 is not met, so equation (17) is not an implicitsolution of (15). (Of course, the same reasoning shows that the equation $y-e^x=0$ is not an implicitsolution of (15).)

We mention that in this example, criterion (i) *is* met. The same function ϕ used in Example 2.6 is a solution of (15) that is defined implicitly by (17). ■

Example 2.8 The equation

$$x^2+y^2+1=0 \tag{20}$$

is *not* an implicitsolution of (15), because it fails criterion (i) of Definition 2.4. There are no real numbers x,y at all for which (20) holds, let alone an open interval I on which (20) implicitly determines a function of x. Since (20) determines no functions ϕ whatsoever on any open interval I, criterion (ii) of Definition 2.4 is moot.

Similarly, the equation

$$x^2+y^2=0 \tag{21}$$

is not an implicitsolution of (15). In this case there *is* a pair of real numbers (x,y) that satisfies (21), but there is no *open x-interval* I on which, for each $x\in I$, there is a real number y for which (21) is satisfied. ■

Now let us make an observation about implicitsolutions:

$$\textit{An implicitsolution of a DE is } \underline{\textit{not}} \textit{ a solution of that DE.} \tag{22}$$

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

However, there is an "abuse of terminology" that we have already said is permissible. When a function ϕ is a solution of a given differential equation $\mathsf{F}(x,y,\frac{dy}{dx})=0$, we have said that we would allow ourselves to call the *equation* $y=\phi(x)$ a solution of that DE. We must recognize that the *equation* $y=\phi(x)$ is *not* a function, of any number of variables. 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.

We allow ourselves to say, *technically incorrectly*, that "$y=x^2$ is a solution of $\frac{dy}{dx}=2x$", because that wording is so much less awkward than "the function ϕ

defined by $\phi(x) = x^2$ is a solution of $\frac{dy}{dx} = 2x$".[5] 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 $G(x, y) = 0$. In the same spirit, we make the following definition:

Definition 2.9 We say that an equation $G(x, y) = 0$ is an *implicit solution* (two words) of a given differential equation if it is an implicitsolution (one word) of that differential equation, as defined in Definition 2.4. ■

Combining this definition with observation (22), 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 the made-up word "implicitsolution", which the student is not likely to find outside of these notes. Most textbooks give a definition of "implicit solution" that is similar to our definition of "implicitsolution"[6].

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

Note that the terminology "implicitly-defined solution" (Definition 2.5) does not suffer from any paradox. An implicitly-defined solution of a DE *is* a solution of that DE. It meets the criteria of Definition 2.1 perfectly.

Our approach to Example 2.6 above relied on our ability to produce an explicit formula for a "candidate solution" of the given DE. What if, in place of (16), we had been given an equation so complicated that we could not solve for y and produce

[5] Only slightly more awkward than "$y = x^2$ is a solution of $\frac{dy}{dx} = 2x$" is the following type of phrasing that you may have seen instructors or textbook-authors use: "The function $\phi(x) = x^2$ is a solution of $\frac{dy}{dx} = 2x$." This phrasing is certainly much less awkward than, "The function ϕ defined by $\phi(x) = x^2$ is a solution of $\frac{dy}{dx} = 2x$." The reason we try not to use phrasing like "The function $\phi(x) = x^2$ ..." in these notes is that the function is ϕ, not $\phi(x)$. The object $\phi(x)$—a *number*—is the output of the function ϕ when the input is called x.

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

[6] Except that most neglect to include criterion (ii).

a candidate-solution ϕ to plug into the DE? This is where the Implicit Function Theorem comes to the rescue.

Example 2.10 [7] Show that the equation

$$x + y + e^{xy} = 1 \tag{23}$$

is an implicit solution of

$$(1 + xe^{xy})\frac{dy}{dx} + 1 + ye^{xy} = 0. \tag{24}$$

To show this, we start with the observation that, writing $G(x, y) = x+y+e^{xy}$, we have $G(0, 0) = 1$. So, let us check whether the Implicit Function Theorem applies to the equation $G(x, y) = 1$ near the point $(0, 0)$ (i.e. taking $(x_0, y_0) = (0, 0)$ in Theorem 2.3). We compute

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

Both of these functions are continuous on the whole xy plane, and $\frac{\partial G}{\partial y}(0, 0) = 1 \neq 0$. Thus, the hypotheses of Theorem 2.3 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 $G(x, \phi(x)) = 1$ for all $x \in I_1$.

Now we use the same method by which we checked criterion (ii) in Example 15: implicit differentiation (i.e. computing derivatives of an expression that contains an implicitly-defined function). Let us simplify the notation a little by writing $y(x) = \phi(x)$. Then

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

Therefore ϕ is a solution of (24). Thus, criterion (i) in Definition 2.4 is satisfied. The exact same implicit-differentiation argument shows that if ψ is *any* differentiable

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

function determined on an open interval by (23), then ψ is a solution of (24). Therefore criterion (ii) in Definition 2.4 is also satisfied. Hence (23) is an implicit solution of (24). ■

Looking back at Example 2.6, could we have shown that criterion (i) of Definition 2.4 is satisfied using the technique of Example 2.10, using the function $G(x, y) = x^2+y^2$? Absolutely! For (x_0, y_0) we could have taken any point of the circle $x^2+y^2 = 1$ other than $(\pm 1, 0)$. The partial derivatives are $\frac{\partial G}{\partial x}(x, y) = 2x$ and $\frac{\partial G}{\partial y}(x, y) = 2y$. As in Example 2.10, the partial derivatives of G are continuous on whole xy plane again[8], and since we are choosing a point (x_0, y_0) for which $y_0 \neq 0$, we have $\frac{\partial G}{\partial y}(x_0, y_0) \neq 0$. Thus, the Implicit Function Theorem applies, guaranteeing the existence of a differentiable, implicitly-defined function ϕ, with $\phi(x_0) = y_0$. We can then differentiate implicitly, as we did when we checked criterion (ii) in Example 2.6 (and as we did to check both criteria in Example 2.10), to show that ϕ is a solution of (15). If our point (x_0, y_0) has $y_0 > 0$, then the solution of (15) 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 (15) that we get is $-\phi_1$.

The student may wonder how we could have used the method of Example 2.10 had we not been clever (or lucky) enough to be able to find a point (x_0, y_0) that lay on the graph of our equation $G(x, y) =$ a given constant. The answer is that we could not have, unless we had some other argument showing that the graph contains at least one point, and, more restrictively, that it contains at least one point at which $\frac{\partial G}{\partial y}$ is not 0. For example, had we started with the equation

$$x + y + e^{xy} = 2 \tag{25}$$

instead of (23), we would have had a much harder time. We could show by implicit differentiation that every differentiable function determined by 25 is a solution of 24—thus, that criterion (ii) of Definition 2.4 is satisfied—but that would not tell us that there is even a single function of x defined by (25), or even that the graph of (25) contains any points whatsoever. Conceivably, we could be in the same situation as in Example 2.8, in which all differentiable functions implicitly defined by (20)—all none of them—are solutions of our differential equation.

It so happens that we *can* show that the graph of (25) 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 $G(x, y)$ is much more complicated.

Example 2.11 Show that there is a number c_0 for which the equation

[8] This does not always happen—Examples 2.6 and 2.10, and several other examples in these notes, just happen to have G's with this property.

$$e^x + x + y^5 - y^4 + y^3 + y^2 + xy^2 = c_0 \tag{26}$$

is an implicit solution of the differential equation

$$e^x + 1 + y^2 + (5y^4 - 4y^3 + 3y^2 + 2y + 2xy)\frac{dy}{dx} = 0. \tag{27}$$

To approach this problem, we start with a variation on the second step of Examples 2.6 and 2.10: we assume that there is a number c_0 for which (26) implicitly determines a differentiable function ϕ, say on an interval I. On the interval I, we may then implicitly differentiate the equation (26)—i.e. differentiate with respect to x both sides of the equation we obtain by substituting "$y = \phi(x)$" into (26). 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 (27).

Thus, all differentiable functions ϕ determined implicitly by an equation of the form (26) will be solutions of (27). Thus for any c_0 for which (26) implicitly determines a differentiable function, equation (26) will be an implicit solution of (27).

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 $G(x, y)$ denote the left-hand side of (26), we compute

$$\begin{aligned} \frac{\partial G}{\partial x}(x,y) &= e^x + 1 + y^2, & (28) \\ \frac{\partial G}{\partial y}(x,y) &= 5y^4 - 4y^3 + 3y^2 + 2y + 2xy. & (29) \end{aligned}$$

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

$$\frac{\partial G}{\partial y}(x,y) = y(5y^3 - 4y^2 + 3y + 2 + 2x). \tag{30}$$

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

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

So if we take, for example, $(x_0, y_0) = (0, 1)$, then $\frac{\partial G}{\partial y}(x_0, y_0) \neq 0$. For this choice of (x_0, y_0), we have $G(x_0, y_0) = 3$. The Implicit Function Theorem then guarantees us that on some open x-interval containing 0, the equation $G(x, y) = 3$ 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 $G(x, y) = 3$ is an implicit solution of (27). So we have found a c_0 with the desired property. ■

As you probably noticed, in this example our expressions (28)–(29) for the partial derivatives of G appeared also in (27). This is no accident. As students who have taken Calculus 3 know, the multivariable chain rule implies that if we implicitly differentiate the equation $G(x, y) = c_0$ with respect to x, we obtain the equation

$$\frac{\partial G}{\partial x} + \frac{\partial G}{\partial y}\frac{dy}{dx} = 0. \tag{31}$$

With foresight, the author chose the DE (27) to be exactly the equation (31) for $G(x, y)$ equal to the left-hand side of (26). For *most* DEs, it will *not* be true that there is a value of c_0 for which (26) is an implicit solution.

It may seem to you that the author cheated, by choosing essentially the only DE for which the fact you were instructed to establish was actually a true fact. But you will see later that equations of the form (31) actually come up a lot.

You may also have noticed, in Example 2.11, 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 (26) is an implicit solution of (27), are all the numbers $G(x_0, y_0)$ 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 G—perhaps the whole range of G. A challenging question for you to think about is this: are there any numbers c_0 for which (26) is *not* an implicit solution of (27)? Let's strip away the distracting complexity of the function G in (26) and pose the analogous question for a much simpler G, the one in Example 2.10:

Question: Are there any numbers c_0 for which the equation

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

is *not* an implicit solution of (24)? (Note that (24) is the equation (31) for the function G defined by $G(x,y) = x + y + e^{xy}$.) ■

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 G that we could put on the left-hand side of "$G(x,y) = c_0$". For example, if we take $G(x,y) = x^2 + y^2$, then only for $c_0 > 0$ is the equation $G(x,y) = c_0$ an implicit solution of (15) (which is the equation (31) for this G, simplified by dividing by 2). But if we take $G(x,y) = x + y$, then for every real number c_0 the equation $G(x,y) = c_0$ is an implicit solution of the analogous differential equation, $1 + \frac{dy}{dx} = 0$, as you can see easily by explicitly solving the equation $x + y = c_0$ for y in terms of x.

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

Example 2.12 Consider the algebraic equation

$$x^2 - y^2 = 0 \tag{32}$$

and the differential equation

$$x - y\frac{dy}{dx} = 0. \tag{33}$$

Equation (32) is equivalent to $y = \pm x$. Thus on any interval I, equation (32) implicitly determines two differentiable functions ϕ of x, namely $\phi(x) = x$ and $\phi(x) = -x$. Both of these are solutions of (33). Therefore (32) is an implicit solution of (33), and the two functions ϕ above are implicitly-defined solutions of (33), on any interval.

The point $(x,y) = (0,0)$ satisfies (32). But on no open rectangle containing the point $(0,0)$ does (32) 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 (12) holds.

Does this contradict the Implicit Function Theorem? No—the theorem says only that there are I_1 and J_1 with the property (12) *if the hypotheses of the theorem are met.* But in the current example, the function G for which (32) is of the form $G(x,y) = c_0$ is given by $G(x,y) = x^2 - y^2$. Thus $\frac{\partial G}{\partial y}(x,y) = -2y$, and if we take $(x_0,y_0) = (0,0)$ then $\frac{\partial G}{\partial y}(x_0,y_0) = 0$. One of the hypotheses of the theorem is not met, and therefore we can draw no conclusion from the theorem. The two functions ϕ above are perfectly good implicitly-defined solutions of (33); they just are not solutions that the Implicit

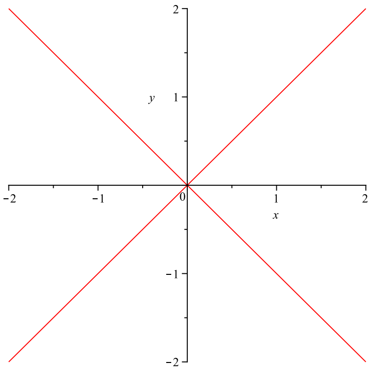

Figure 3: The graph of $x^2 - y^2 = 0$.

Function Theorem finds. ■

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

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

2.3 Maximal and general solutions of derivative-form DEs

Definition 2.13 For a given F, the *general solution* of the differential equation $\mathsf{F}(x, y, \frac{dy}{dx}) = 0$ on an interval I is the collection of *all* solutions 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 2.14 (temporary) for a given F, the *general solution* of the differential equation

$$\mathsf{F}(x, y, \frac{dy}{dx}) = 0 \tag{34}$$

is the collection of all solutions of (34), where "solution" is defined as in the second part of Definition 2.1. Said another way, the general solution of (34) is the collection of pairs (I, ϕ), where I is an open interval and ϕ is a solution of (34) 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.

The student should not overlook our careful use of the articles "a" and "the" in "a solution" (Definition 2.1) and "the general solution" (Definition 2.14). Use of the definite article "the" implies that we are talking about something that is *unique*—i.e. only one such thing exists. "The" should never be used by a writer (or speaker) unless s/he has already given enough information for the reader (or listener) to know that only one exists. Differential equations, even on a specified interval, virtually never have just one solution (although *initial-value problems* usually do). The only thing that "the solution" of a given DE can unambiguously mean is the collection of *all* solutions. Thus, to the author of these notes, "the solution of equation (1)" is *synonymous* with "the general solution of equation (1)". To avoid misinterpretation, in these notes we will not use the terminology "the solution" (of a given DE, in the absence of an initial condition); we will always say either "a solution" or "the general solution".[9]

There is a problem with Definition 2.14 that we will discuss shortly. However, in their first exposure to the subject, many students will not have the mathematical sophistication needed to understand the problem or the way to fix it. Therefore **in a first course on differential equations, it is acceptable to use Definition 2.14 as the definition of "general solution", and students in this author's course will not be penalized for doing so.** Some students, however, may recognize (eventually, if not immediately) that there is a problem. The discussion below is for those students, and any others who might be interested in what the problem is. **Students who are not interested, or have trouble understanding the discussion, may skip to Example 2.19 and simply ignore the word "maximal" wherever it appears.**

To illustrate the problem, let us suppose that we are able to show for every solution ϕ of some differential equation, there is a constant C such that

[9] Not all mathematicians are equally picky about terminology, and the author cannot guarantee that your instructor will so strictly separate the meanings of "a" and "the", or will agree that the only logically possible meaning of "the solution of a (given) DE" is the general solution of that DE.

$$\phi(x) = \frac{1}{x - C} \ . \tag{35}$$

Remembering that the domain of a solution of a DE is required to be an *interval*, we look at equation (35) and say, "Okay, for each C this formula gives two solutions, one on $(-\infty, C)$ and (C, ∞)." But even this is not technically correct. These are not the only two intervals on which equation (35) defines solutions. If ϕ is a solution on (C, ∞), then it satisfies the DE at every point of this interval. Therefore it also satisfies the DE at every point of $(C, C+1)$, at every point of $(C + 26.4, C + 93.7)$, and on any open subinterval of $(-\infty, C)$ or (C, ∞) whatsoever.

This example illustrates that the collection of pairs (I, ϕ) referred to in Definition 2.14 has a certain redundancy. There is terminology that allows us to speak more clearly about this redundancy:

Definition 2.15 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 $\underline{\text{a}}$ restriction of ϕ if it is $\underline{\text{the}}$ 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.[10]

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

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

(The symbol " $\Longleftrightarrow$ " means "if and only if".)

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

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

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

ϕ, then we know every speck of information about $\phi|_{I_1}$. Therein lies the redundancy of Definition 2.14: the definition names a much larger collection of functions than is needed to capture all the information there is to know about solutions of (34). We will see below that we can be more efficient.

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

$$\frac{dy}{dx} = -x^{-2}, \quad y(3) = \frac{1}{3}, \tag{36}$$

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

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

Example 2.16 All the functions ϕ below are different functions, even though we are using the same letter for them.

- $\phi(x) = \frac{1}{x}$, $0 < x < 5$, is a solution of $\frac{dy}{dx} = -x^{-2}$, but not a maximal solution. It is also a solution of the IVP (36).

- $\phi(x) = \frac{1}{x}$, $2.9 < x < 16.204$, is another solution of $\frac{dy}{dx} = -x^{-2}$, and of the IVP (36), but not a maximal solution.

- $\phi(x) = \frac{1}{x}$, $3.1 < x < 16.204$, is another solution of $\frac{dy}{dx} = -x^{-2}$, but it is neither a maximal solution nor a solution of the IVP (36),

- $\phi(x) = \frac{1}{x}$, $x \in (0, \infty)$ is *a* maximal solution of $\frac{dy}{dx} = -x^{-2}$, and is *the* maximal solution of the IVP (36).

- $\phi(x) = \frac{1}{x}$, $x \in (-\infty, 0)$ is a *different* maximal solution of $\frac{dy}{dx} = -x^{-2}$. It is *not* a solution of the IVP (36).

- $\phi(x) = \frac{1}{x}$, $x \in (-\infty, -\sqrt{2})$ is another non-maximal solution of $\frac{dy}{dx} = -x^{-2}$.

- $\phi(x) = \frac{1}{x} + 37,\ x \in (0, \infty)$ is yet another maximal solution of $\frac{dy}{dx} = -x^{-2}$. It is not a solution of the IVP (36).

Example 2.17 The maximal solutions of the differential equation $\frac{dy}{dx} = \sec^2 x$ are

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

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

It can be shown that every non-maximal solution of a DE is the restriction of some maximal solution of that DE.[11] Thus the collection of maximal solutions "contains" all solutions in the sense that the graph of every solution is contained in the graph of some maximal solution. So, better than Definition 2.14 is this:

Definition 2.18 For a given F, the *general solution* of (1) is the collection of all maximal solutions of (1).

(This definition supersedes Definition 2.14.)

Example 2.16 demonstrates, we hope, the economy gained by including the word "maximal" in this definition. The student will probably agree that, even prior to writing down Definition 2.18, maximal solutions are what we really would have been thinking of had we been asked what all the solutions of "$\frac{dy}{dx} = -x^{-2}$" are—we just might not have realized consciously that that's what we were thinking of.

Example 2.19 The general solution of $\frac{dy}{dx} = x$ may be written as

$$y = \frac{1}{2}x^2 + C. \tag{37}$$

In this context equation (37) 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. We allow ourselves to write (37) as short-hand for "the collection of functions $\{\phi_C \mid C \in \mathbf{R}\}$, where $\phi_C(x) = \frac{1}{2}x^2 + C$".

Example 2.20

- The general solution of

$$\frac{dy}{dx} = -x^{-2}, \quad x > 0 \tag{38}$$

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

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

$$y = \frac{1}{x} + C,\ x > 0, \tag{39}$$

a one-parameter family of maximal solutions. Because the restriction $x > 0$ is stated explicitly in (38), it is permissible to leave out the "$x > 0$" when writing the general solution; we may simply write the general solution as

$$y = \frac{1}{x} + C \tag{40}$$

- The general solution of

$$\frac{dy}{dx} = -x^{-2}, \tag{41}$$

with no interval specified, may also be written as (40)—i.e. it is *permissible* to write it this way, in the interests of saving time and space. However, because no interval was specified when the DE was written down, we must consider all possible intervals. Therefore, <u>in this context</u>, equation (40) does *not* represent a one-parameter family of maximal solutions; it represents *two* one-parameter families of maximal solutions[12]. Equation (40) is acceptable short-hand for

$$\left.\begin{array}{l}\text{the union of the two families of functions}\\[1em] \qquad \{\phi_C \mid\ C \in \mathbf{R}\},\quad \{\psi_C \mid\ C \in \mathbf{R}\}\\ \text{where}\\ \qquad \phi_C(x) = \frac{1}{x} + C, \quad x > 0\\ \text{and}\\ \qquad \psi_C(x) = \frac{1}{x} + C, \quad x < 0.\end{array}\right\} \tag{42}$$

(The *union* of the two families means the collection of functions that are in one family or the other.) The solution $y = \frac{1}{x} + 6$ on $\{x < 0\}$ (the function ψ_6 in the notation of (42)) is no more closely related to the solution $y = \frac{1}{x} + 6$ on $\{x > 0\}$

[12] 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 $\mathsf{F} + C$ 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 $f(x) = \sec^2 x$.

(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 lies in the same family as ψ_6, where as ϕ_6 does not.)

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

$$\text{“}\{y = \frac{1}{x} + C, x > 0\} \text{ and } \{y = \frac{1}{x} + C, x < 0\}\text{”} \tag{43}$$

and

$$\text{“}\{y = \frac{1}{x} + C_1, x > 0\} \text{ and } \{y = \frac{1}{x} + C_2, x < 0\}\text{”}. \tag{44}$$

In (43), 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 (44), 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 (40), (43), and (44) are acceptable ways of writing the general solution, as long as we understand what they mean, and are communicating with someone else who understands what they mean. These forms do not exhaust all permissible ways of writing the general solution; there are other notational variations on the same theme.

Example 2.21 The general solution of $\frac{dy}{dx} = \sec^2 x$ may be written as

$$y = \tan x + C, \tag{45}$$

or as

$$y = \tan x + C, \quad (n - \frac{1}{2})\pi < x < (n + \frac{1}{2})\pi, \quad n \text{ an integer}, \tag{46}$$

or as

$$y = \tan x + C_n, \quad (n - \frac{1}{2})\pi < x < (n + \frac{1}{2})\pi, \quad n \text{ an integer}, \tag{47}$$

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 (45)–(47) (or other variations on the same theme) that we choose for writing the general solution of $\frac{dy}{dx} = \sec^2 x$, we must not forget that each of these forms represents *an infinite collection of one-parameter families of maximal solutions*, one family for each interval of the form $(n - \frac{1}{2})\pi < x < (n + \frac{1}{2})\pi$ (n an integer).

Example 2.22 The general solution of the separable equation

$$\frac{dy}{dx} = -y^2 \tag{48}$$

may be written as

$$\left\{ y = \frac{1}{x - C} \right\} \text{ and } \; y \equiv 0, \tag{49}$$

or as

$$y = \frac{1}{x - C} \; \text{ or } \; y = 0, \tag{50}$$

or in various other ways that impart the same information[13]. In the given context, the solution that is the constant function 0 may be written as "$y \equiv 0$" (which, in this context, is read "y identically zero") or as $y = 0$. Since a solution of (48), expressed in terms of the variables in (48), is function of x, the only correct interpretation of "$y = 0$" in (50) 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 (48):

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.)

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

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

The general solution of (48) also exhibits another interesting phenomenon. The way we have written the general solution in (49) and (50) isolates the maximal solution $y \equiv 0$ as not belonging to what appears to be a single nice family into which the other maximal solutions fall (there is no value of C for which the formula "$y = \frac{1}{x-C}$" produces the constant function 0). But for $C \neq 0$, writing $K = \frac{1}{C}$,

$$\frac{1}{x-C} = \frac{C^{-1}}{C^{-1}x-1} = \frac{K}{Kx-1} \ . \tag{51}$$

In the right-most formula in (51), 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 (49). Thus, we can rewrite the general solution (49) as

$$\left\{ y = \frac{K}{Kx-1} \right\} \text{ and } \ y = \frac{1}{x} \ . \tag{52}$$

Here, K is an arbitrary constant, allowed to assume all real values, just as C was allowed to in (49). Writing the general solution this way, 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.

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 (49) do not form *one* one-parameter family; they form *two.* Every value of C corresponds to two maximal solutions, one defined to the left of C and one defined to the right[14]. In (52), 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.

[14] *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 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. The natural parameter-space for the collection of all maximal solutions of (48) is not a circle, but a figure-8. In our next example, a logistic equation, the natural parameter space is two simple closed curves joined along a common line segment whose endpoints correspond to the constant solutions.

In this example, one may reasonably decide that (49) is preferable to (52) as a way of writing down the general solution. The constant solution $y \equiv 0$ is distinguished from all the others not just by being constant, but by being the only solution defined on the whole real line. Furthermore, the collection of solutions described by $\{y = \frac{1}{x-C}\}$ is more "uniform" than is the collection described by $\{y = \frac{K}{Kx-1}\}$, in the sense that in the first collection, *every* value of the arbitrary constant corresponds to two maximal solutions, while in the second collection there is a value of the arbitrary constant, namely 0, for which the given formula defines only one maximal solution. However, in the next example, we will see two different ways of writing the general solution, neither of which can be preferred over the other by any such considerations.

Example 2.23 The general solution of the separable equation

$$\frac{dy}{dx} = y(1-y) \tag{53}$$

may be written as

$$\left\{ y = \frac{C}{e^{-x} + C} \right\} \quad \text{and} \quad y \equiv 1. \tag{54}$$

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

$$\left\{ y = \frac{1}{Ce^{-x} + 1} \right\} \quad \text{and} \quad y \equiv 0. \tag{55}$$

(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 (54) is the constant function 0, which is the "exceptional" solution in (55). The $C = 0$ solution in (55) is the constant function 1, which is the "exceptional" solution in (54). The situation is completely symmetric; neither of (54) and (55) can be preferred over the other.

The last example illustrates that for nonlinear DEs there may be no singled-out way to write the collection of all maximal solutions (or solutions on a specified interval) of a nonlinear equation as a one-parameter family, or as several one-parameter families, or as one or more one-parameter families of solutions plus some "exceptional" solutions. Because of this, many authors prefer to use the terminology "general solution" *only* for linear DEs, and not to define the term at all for nonlinear DEs.[15]

[15] *Note to instructors:* This author, however, feels that too much is lost this way. It is important for students to be able to know when they've found all solutions. This author has found that many textbooks that avoid defining "general solution" for nonlinear DEs do not systematically address the question "Have we found all solutions?" at all, or even make the importance of the question

2.4 Algebraic equivalence of derivative-form DEs

In these notes we have defined *open rectangles*. You may also be familiar with *open disks*: the open disk of radius $\epsilon > 0$ centered at (x_0, y_0) is the set of points (x, y) a distance less than ϵ from (x_0, y_0) (equivalently, the set of points (x, y) that satisfy the strict inequality $\sqrt{(x - x_0)^2 + (y - y_0)^2} < \epsilon$). More generally, a subset R of $\mathbf{R}^2$ is called an *open set* if for every point $(x_0, y_0) \in R$, the set R contains the open disk of *some* radius (possibly tiny), centered at (x_0, y_0). If you draw yourself a picture you should easily be able to convince yourself that "open disk" and "open rectangle" meet the definition of "open set", so our terminology is self-consistent.[16]

Another term we will use for "open subset of $\mathbf{R}^2$ " is *region*[17].

Definition 2.24 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 $(x, y) \in R$ [18] , and/or

clear. This can reinforce 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.

This author feels that it is worthwhile to give the student a name for the collection of all solutions, and to choose the name that is the most consistent with terminology that mathematicians use throughout mathematics. By this criterion, "general solution" seems best to him.

Other DE instructors may have different conventions for use of the term "general solution", but we caution the instructor to be wary of using "general solution" to refer to a non-exhaustive collection of solutions for which (s)he has produced a nicely-parametrized family of formulas. As the simple examples 2.22 and 2.23 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 method used to produce the solutions.

We mention, however, that there *is* an accepted definition of *singular solution* of an ODE. A singular solution of an ODE is one "at every point of which the uniqueness of the solution of the Cauchy problem for this equation is violated" (*Encyclopedia of Mathematics*, online edition, Springer, `http://eom.springer.de/s/s085610.htm`). This definition provides a way to canonically separate "exceptional" solutions from the rest, and some authors have used "general solution" to refer to the collection of all solutions that are not singular. This happens to reproduce what we have called the general solution in all the examples in these notes, for the simple reason that, like virtually every DE shown students in a typical first course on ODEs nowadays, the DEs in our examples have *no* singular solutions. But even for equations that do have singular solutions, it would seem preferable to use the term *generic* for the other solutions, rather than *general*.

[16] For example, if R is the open disk of radius 1 centered at $(0, 0)$, and we take $(x_0, y_0) = (0.99, 0)$, then the open disk of radius 0.005 centered at (x_0, y_0) is contained in R.

[17] The author is taking some liberties here. The usual definition of "region" is *connected* non-empty open subset. The author 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, i.e. that it has at least one point.

[18] *Note to students*: The expression is allowed to involve $\frac{dy}{dx}$, which is why we did not say "function of x and y" here. If the expression does involve $\frac{dy}{dx}$, our requirement that it be defined for all $(x, y) \in$

(ii) multiplying both sides of the equation by a function of x and y that is defined *and nonzero* at every point of R.

Note that subtraction of an expression is the same as addition of the negative of that expression, so subtraction is an operation allowed in Definition 2.24, even though it is not mentioned explicitly.

Example 2.25 The differential equations

$$\frac{dy}{dx} = y(1-y) \tag{56}$$

and

$$\frac{1}{y(1-y)}\,\frac{dy}{dx} = 1 \tag{57}$$

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 2.26 The differential equations

$$(y-x)\frac{dy}{dx} = 2y+4x \tag{58}$$

and

$$\frac{dy}{dx} = \frac{2y+4x}{y-x} \tag{59}$$

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

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

R means that it is defined whenever $(x,y)\in R$ and any real number whatsoever is substituted for $\frac{dy}{dx}$.

Note to instructors: The latter requirement is more restrictive than necessary—for example, it eliminates adding to both sides $\frac{1}{dy/dx}$, $\sqrt{1-(\frac{dy}{dx})^2}$, or an expression like $\sqrt{\frac{dy}{dx}+x+y}$ that it is hard to imagine ever arising in any DE that anyone would ever have an interest in solving

But often, when we manipulate equations in an attempt to find their solution sets, we perform a manipulation that changes the solution set.[19] This happens, for example, if we start with the equation $x^3 - 3x^2 = -2x$ and divide by x, obtaining $x^2 - 3x^2 = -2$. In this example, we lose the solution 0. (The solution set of the first equation is $\{0, 1, 2\}$, while the solution set of the second is just $\{1, 2\}$. For another example, if start with the equation $\sqrt{x+4} = -3$, and square both sides, we obtain $x + 4 = 9$, and hence $x = 5$. But 5 is not a solution of the original equation; $\sqrt{5+4}$ is 3, not -3. Our manipulation has introduced a "spurious solution", a value of x that is a solution of the post-manipulation equation that we may *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:

$$\left.\begin{array}{l}\text{If two differential equations are algebraically equivalent on a region } R,\\ \text{then the set of solutions of the first equation whose graphs are contained}\\ \text{in } R\text{, is the same as the set of solutions of the second equation whose}\\ \text{graphs are contained in } R.\end{array}\right\}\tag{60}$$

If the region R above is the whole xy plane, then the collection of *all* solutions of the first equation—hence its general solution—is the same as the general solution of the second equation. In this case, if $R = \mathbf{R}^2$ is understood, we may restate (60) more briefly as "Algebraically equivalent DEs have the same general solution," "Algebraically equivalent DEs have the same set of solutions,", or "Algebraically equivalent DEs are equivalent." But on regions that are not all of $\mathbf{R}^2$, the briefer wording must be interpreted more carefully to mean statement (60).

When we perform a sequence of algebraic operations in an attempt to solve a differential equation, especially a nonlinear one, we are rarely lucky enough to end up with a DE that is algebraically equivalent to the original one on the whole xy plane. But usually, we maintain algebraic equivalence on regions that fill out most of the xy plane, as in Examples 2.25 and 2.26 above.

To see why statement (60) is true, let us check that operation (ii) in Definition 2.24 does not change the set of solutions whose graphs lie in R. Let us suppose we start with a (first-order) derivative-form DE of the most general possible form:

$$\mathsf{F}_1(x, y, \frac{dy}{dx}) = \mathsf{F}_2(x, y, \frac{dy}{dx}).\tag{61}$$

(Of course, by subtracting $\mathsf{F}_2(x, y, \frac{dy}{dx})$ from both sides, we can put this in the simpler form $\mathsf{F}(x, y, \frac{dy}{dx}) = 0$, but since we often perform manipulations on equations without

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

first putting them in the simple form (1), we will illustrate the solution-set-doesn't-change principle for DEs that have not been put in that form.) The equation obtained by multiplying both sides of (61) by a function h that is defined at every point of R and is nonzero on R is

$$h(x,y)\mathsf{F}_1(x,y,\frac{dy}{dx}) = h(x,y)\mathsf{F}_2(x,y,\frac{dy}{dx}). \tag{62}$$

Suppose that ϕ is a solution of (61). Then for all x in the domain of ϕ,

$$\mathsf{F}_1(x,\phi(x),\phi'(x)) = \mathsf{F}_2(x,\phi(x),\phi'(x)). \tag{63}$$

If the graph of ϕ lies in R [20], then for all x in the domain of ϕ, the point $(x,\phi(x))$ lies in R, hence in the domain of h. Therefore for all x in the domain of ϕ, $h(x,\phi(x))$ is some number, and equality is maintained if we multiply both sides of (63) by this number. Therefore

$$h(x,\phi(x))\mathsf{F}_1(x,\phi(x),\phi'(x)) = h(x,\phi(x))\mathsf{F}_2(x,\phi(x),\phi'(x)) \tag{64}$$

for all x in the domain of ϕ. Hence ϕ is a solution of (62). Thus every solution of (61) whose graph lies in R is also a solution of (62) whose graph lies in R.

Conversely, suppose that ϕ is a solution of (62) whose graph lies in R. Then (64) 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 (64) by this number. Therefore (63) is satisfied for all x in the domain of ϕ, so ϕ is a solution of (61). Thus every solution of (62) whose graph lies in R is also a solution of (61) whose graph lies in R.

This completes the argument that multiplying by h has not changed the set of solutions whose graphs lie in R. The argument that operation (i) in Definition 2.24 does not change this set of solutions is similar, and is left to the student.

We mention that it is possible for two differential equations to be equivalent without being algebraically equivalent. Performing operations other than those in Definition 2.24 does not *always* change the set of solutions. But because they *might* change the set of solutions, any time we perform one of these "unsafe" operations we must check, by some other method, that we properly account for any lost solutions or spurious solutions.

[20] In this argument we are talking about *all* solutions whose graphs lie in R, not just maximal solutions whose graphs lie in R. (Students who did not read or did not understand the earlier material on maximal solutions should ignore the part of the previous sentence after the comma.) If there is a solution $\tilde{\phi}$ whose graph lies partly inside R and partly outside R, then there are x-intervals I to which we can restrict $\tilde{\phi}$ and obtain a solution whose graph lies in R. All solutions obtained this way are covered by our argument, as well as any maximal solutions whose graphs lie in R. (Students who did not read or did not understand the material on maximal solutions should replace the second half of the previous sentence with "as well as any solutions whose graphs lay entirely inside R to begin with".)

Students should already be familiar with this fact from their experience with separable equations. For example, in passing from equation (56) to (57), 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 y_0 for which $p(y_0) = 0$, 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 y_0 for which $p(y_0) = 0$, the constant function $y = y_0$ is a solution that separation of variables, carried out with no errors, *cannot* find. But fortunately, it finds all the others (in implicit form).

We can see why in the context of Example 2.25. The right-hand side of (56) is a function of y whose partial derivative with respect to y is continuous everywhere. Therefore for *every* initial-condition point (x_0, y_0) in the xy plane, the fundamental Existence and Uniqueness Theorem for initial-value problems applies, and so through each such point there is the graph of one and only one maximal solution. If there were a non-constant solution of (56) 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 (56) and with initial condition $y(x_0) = 1$. Similarly, no non-constant solution of (56) can have a graph that intersects the graph of the constant solution $y \equiv 0$ (the line $\{y = 0\}$). Therefore the graph of every non-constant solution lies entirely in one of the three regions mentioned in Example 2.25. Since equations (56) and (57) are algebraically equivalent on each of these three regions, the general solution of (57) is precisely the set of all solutions of (56) other than the two constant solutions that we have already accounted for.

Thus, if we manage to solve (57)—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 (56).

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

Example 2.27 The equations

$$\frac{dy}{dx} + 3y = \sin x \tag{65}$$

and

$$e^{3x}\frac{dy}{dx} + 3e^{3x}y = e^{3x}\sin x \tag{66}$$

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

The student familiar with integrating-factors will recognize that the e^{3x} in the example above is an integrating factor for the first equation. To solve linear DEs by the integrating-factor method, the only functions we ever need to multiply by are functions of x alone. Of course, every such function can be viewed as a function of x and y that simply happens not to depend on y. More explicitly, given a function one-variable function μ, we can define a two-variable function $\tilde{\mu}$ by $\tilde{\mu}(x,y) = \mu(x)$. If $\mu(x)$ is nonzero for every x in an interval I, then $\tilde{\mu}(x,y)$ is nonzero at every (x,y) in the region $I \times R$ (an vertical strip, infinite in the $\pm y$-directions). So we will add a bit to Definition 2.24 to have language better suited to linear equations:

Definition 2.28 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 a function of x that is defined at every point of the region $I \times \mathbf{R}$, or y times such function of x, or $\frac{dy}{dx}$ times such a function of x; and/or (ii) multiplying both sides of the equation by a function of x that is defined and nonzero at every point of the interval I.

Example 2.29 The equations

$$x\frac{dy}{dx} - 2y = 0 \qquad (67)$$

and

$$x^3\frac{dy}{dx} - 2x^2y = 0 \qquad (68)$$

are algebraically equivalent on the interval $(0,\infty)$, and also on the interval $(-\infty,0)$, but not on $(-\infty,\infty)$ or on any other interval that includes 0. (Thus, in accordance with Definition 2.24, we do not simply call them "algebraically equivalent".) The second can be obtained from the first by multiplying by x^2, which satisfies the "nowhere zero" criterion on any interval not containing 0, but violates it on any interval that includes 0.

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

Example 2.30 The equations

$$x\frac{dy}{dx} - 2y = 0 \tag{69}$$

(the same equation as (67) and

$$x^{-2}\frac{dy}{dx} - 2x^{-3}y = 0 \tag{70}$$

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, equation (60) reduces to the following simpler statement:

$$\left.\begin{array}{l}\text{Two linear DEs that are algebraically equivalent}\\ \text{on an interval } I \text{ have exactly the same solutions on } I.\end{array}\right\} \tag{71}$$

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

Example 2.31 Find the general solution of

$$x\frac{dy}{dx} - 2y = 0 \tag{72}$$

(the same equation as (69) and (67)).

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

$$\frac{dy}{dx} - \frac{2}{x}\,y = 0. \tag{73}$$

However, (72) and (73) are not algebraically equivalent on the whole real line, but only on $(-\infty, 0)$ and $(0, \infty)$. Equation (73) does not even make sense at $x = 0$, while

(72) makes perfectly good sense there.[21]

As the student may verify, equation (73) has an integrating factor $\mu(x) = x^{-2}$. Putting our brains on auto-pilot, we multiply through by x^{-2}, and write

$$\begin{aligned} (x^{-2}y)' &= 0, \\ \Rightarrow \int (x^{-2}y)' dx &= \int 0 \, dx, \\ \Rightarrow x^{-2}y &= C, \\ \Rightarrow y &= Cx^2. \end{aligned} \tag{74}$$

(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 (74).)

Neither in the original DE (72) nor in (74) do we see any of the clues we are used to seeing, such as a "$\frac{1}{x}$", that warn us that there may be a problem with (74) at $x = 0$. (There were clues in the intermediate steps, in which negative powers of x appeared, but we ignored them.) The functions given by (74) form a 1-parameter family of functions defined on the whole real line, and it is easy to check that all of them are solutions of (72). We have been taught that the general solution of a first-order linear DE is a 1-parameter family of solutions—*under certain hypotheses.* (We have ignored the fact that those hypotheses were not met, however.) Having found what we expected to find, we write "$y = Cx^2$" as our final, but wrong, answer.

Let us go back to square one and correct our work. The transition from equation (72) to (73) 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 to (72) might include solutions on intervals that contain 0, while the general solution to (73) cannot.

We can still use the basic procedure that led us to (74); we just have to be more careful with it. Auto-pilot will not work.

Because (73) 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 $y = \phi(x)$. $\underline{\text{On } I}$, x^{-2} is an integrating factor. Multiplying both

[21] Standard terminology related to this problem is *singular point.* Generally speaking, a first-order linear DE does not "behaves well" on an interval I if, when put in standard linear form $\frac{dy}{dx} + p(x)y = g(x)$, there is a point $x_0 \in I$ for which $\lim_{x \to x_0+} p(x) = \pm\infty$ or $\lim_{x \to x_0-} p(x) = \pm\infty$. Such points x_0 are called *singular points* of the linear DE. The point $x = 0$ is a singular point of both (72) and (73).

sides of our equation on I by x^{-2}, we find that ϕ is a solution of (73) 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 (73) 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
$$\phi(x) = Cx^2. \tag{75}$$
- Exactly the same conclusion holds on the interval $(-\infty, 0)$.

Thus the general solution of (73) on $(0, \infty)$ is

$$y = Cx^2, \quad x > 0, \tag{76}$$

while the general solution of (73) on $(-\infty, 0)$ is

$$y = Cx^2, \quad x < 0. \tag{77}$$

Now return to the equation we originally were asked to solve, (72), 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 interval $(0, \infty)$, and let ϕ_2 be the restriction of ϕ to the interval $(-\infty, 0)$. Since (72) and (73) are algebraically equivalent on $(0, \infty)$, ϕ_1 must be one of the solutions given by (76). 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 (77), 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 the every solution of an ODE is continuous on its domain (since, by definition, solutions are differentiable functions, and differentiable functions are continuous). Therefore $\phi(0) = \lim_{x\to 0} \phi(x)$. Whether we approach 0 from the left (using $\phi(x) = C_2x^2$) or the right (using $\phi(x) = C_1x^2$), we get the same limit, namely 0. Hence $\phi(0) = 0$.[22] Since 0 also happens to be the value of C_1x^2 at $x = 0$ (as well as the value of C_2x^2 at $x = 0$), we can write down a formula for ϕ in several equivalent ways, one of which is

$$\phi(x) = \begin{cases} C_1x^2 & \text{if } x \geq 0, \\ C_2x^2 & \text{if } x < 0, \end{cases} \tag{78}$$

[22] Another way to find the value of $\phi(0)$ in this example is as follows. Since ϕ is differentiable on its domain, the whole real line, $\phi'(0)$ is *some* real number. Whatever this value is, when we plug $x = 0$ and $y = \phi(x)$ into (72), 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 (72), it does not work for (68)—which the student will later be asked to solve—but the first method we presented does.

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

Conversely, as the student may check, every function of the form (78) is a solution of (72). Therefore the general solution of (72) on $(-\infty, \infty)$ is the *two-parameter* family of functions given by (78), with C_1 and C_2 arbitrary constants[23]. 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 (78) to the interval I. (For the student who read and understood the material on maximal solutions: the two-parameter family (78) is the general solution of (72) as defined in Definition 2.18.) ■

We do not want the student to come away from the previous example with the wrong impression. For the vast majority, if not 100%, of n^{th}-order linear DEs you are likely to encounter in your first course on DEs, you will be shown how to solve them (or asked to solve them) only on intervals for which the general solution is an n-parameter family of functions. You are unlikely to see a two-parameter family of functions as the general solution unless the equation is second-order. Example 2.31 is the exception, not the rule. But we wanted the student to see another example of the perils of what can happen when algebraic equivalence is not maintained during the manipulation of equations.

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 (67) and (68) have the same set of solutions.

2.5 First-order equations in differential form

Definition 2.32 A *differential* in the variables (x, y) is an expression of the form

$$M(x,y)dx + N(x,y)dy \tag{79}$$

where M and N are functions defined on some region in $\mathbf{R}^2$. We often abbreviate this by writing (79) as just

$$Mdx + Ndy, \tag{80}$$

[23] We warn the student that most textbooks apply the term "general solution" to the collection of all solutions of a linear first-order DE on an interval only when that collection is a one-parameter family.

leaving it understood that M and N are functions of x and y. Also, another term we will use for "open subset of $\mathbf{R}^2$ " is *region*[24] When a region R is specified, we call $Mdx + Ndy$ a *differential on R.*

The functions M, N in (79) and (80) are called the *coefficients* of dx and dy in these expressions. ■

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

Definition 2.33 (a) If F is a differentiable function on a region R, and the variables we use for $\mathbf{R}^2$ are x and y, then the *differential of F on R* is the differential dF defined by

$$dF = \frac{\partial F}{\partial x}dx + \frac{\partial F}{\partial y}dy. \tag{81}$$

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

Note that we have not yet ascribed *meaning* to "dx" or "dy"; effectively, they are just place-holders for the functions M and N in (79) and (80). 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 you probably learned there. There is a relation between the two notions, but we are not ready yet to say 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.)

[24] The author is taking some liberties here. The usual definition of "region" is *connected* non-empty open subset. The author 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, i.e. that it has at least one point.

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.

Definition 2.34 Let R be an open set in $\mathbf{R}^2$ and let M, N, M_1, M_2, N_1, N_2, and f be functions defined on R. (Thus $Mdx + Ndy, M_1dx + N_1dy$, and $M_2dx + N_2dy$ are differentials on R.) Then we make the following definitions:

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:

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

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

8. The *zero differential* on R is the differential $0dx+0dy$, which we often abbreviate just as "0". (We tell from context whether the symbol "0" is being used to denote the *real number* zero, the *constant function* whose value at every point is the real number zero, or the zero differential. In the equation "$0dx+0dy=0$", context tells us that each zero on the left-hand side of the equation is to be interpreted as *the constant function with constant value* 0, while the zero on the right-hand side is to be interpreted as the zero differential[25]. ■

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

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

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

Finally, we are ready to bring differential $\underline{\text{equations}}$ back into the picture!

Definition 2.35 A *differential equation in differential form, with variables* (x,y), is an equation of the form

$$\text{one differential in } (x,y) = \text{another differential in } (x,y). \tag{82}$$

We write such an equation only when where there is some region R on which both differentials are defined. When the region R is specified, we append "*on R*" to the phrase "DE in differential form", or insert it after "DE". ■

Example 2.36 Whenever we separate variables in a separable, derivative-form ODE, we go through a step in which we write down a differential-form ODE, such as

$$ydy=e^xdx. \tag{83}$$

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

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

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

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

Here is the differential-form analog of Definition 2.24:

Definition 2.37 We say that two DEs in differential form 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 defined at every point of R and is nowhere zero on R. ■

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

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

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

and

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

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

$$x(2x^2ydx - \tan(x+y)dy) = 0, \qquad (84)$$

but are *not* algebraically equivalent to (84) 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 (82) from both sides of the equation, we obtain an algebraically equivalent equation of the form

$$Mdx + Ndy = 0.$$

Later, after we have defined "solution of a DE in differential form", we will see that algebraically equivalent equations have the same solutions. Therefore we lose no generality, in our discussion of solutions of DEs in differential form, if we restrict attention to equations of the form (86). (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 (83) is an example.)

In our discussion of derivative-form DEs, we frequently mentioned the *graph* of a solution. The graph is an important curve. Its analog for differential-form DEs is what we call *solution curve*, and it is even more important for differential-form DEs than it is for derivative-form DEs. Below, we will define *solution curve* and *solution* for differential-form DEs. In reading this material the student should pay careful attention to whether or not the word "curve" appears after "solution", since *solution curve* and *solution* are very different gadgets, although they are related.

2.5.1 Solution curves of equations in differential form

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

Definition 2.38 A *parametrized curve* in $\mathbf{R}^2$ is an ordered pair of continuous real-valued functions (f, g) defined on an interval (the *parameter interval*) I. The set

$$\{(f(t), g(t)) \mid t \in I\} \tag{85}$$

is called the *range*, *trace*, or *image* of the parametrized curve.

A *curve* in $\mathbf{R}^2$ is a point-set $\mathcal{C} \subset \mathbf{R}^2$ that is the range of some parametrized curve[26].

Given a curve $\mathcal{C}$, if (f, g) is a parametrized curve with trace $\mathcal{C}$, 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

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

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

as t ranges over a parameter-interval; hence the terminology "trace"[27]. 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 $\mathbf{R}$-valued functions. The word *image* is often preferred by mathematicians, but it means the same thing as "range".

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

Most of the time it is simpler to write "$(x(t), y(t))$" than to introduce the extra letters f, g and write "$(f(t), g(t))$" for the point in the xy plane defined by "$x = f(t), y = g(t)$". We will often use the simpler notation $(x(t), y(t))$ when there is no danger of misinterpretation. Thus we we also sometimes write "$\gamma(t) = (x(t), y(t))$".

Note that in Definition 2.38, we do not require the interval I to be open. This is so that we can present certain examples below simply, without bringing in too many concepts at once that may be new to the student. Eventually, we will want to consider only parametrized curves that have an open domain-interval, but we will not impose that requirement just yet.

Example 2.39 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 range of this parametrized curve lies along the circle $x^2 + y^2 = 4$. It is not hard to see that every point on the circle is in the range of this parametrized curve, so the (just-plain, or unparametrized) curve associated with this parametrized curve is the whole circle $x^2 + y^2 = 4$. Had we used the same formulas for $x(t)$ and $y(t)$, but restricted t to the interval $[0, \pi]$, the range would still have lain along the circle $x^2 + y^2 = 4$, but would have been only a semicircle. Had we used the same formulas, but used a slightly larger, open interval, say $(-0.1, 2\pi + 0.1)$, then we would have obtained the whole circle again, with some small arcs traced-out twice. ■

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

Definition 2.40 A parametrization $(x(t), y(t)), t \in I$ is called

- *differentiable* if the derivatives $x'(t)$, $y'(t)$ exist[28] for all $t \in I$;

[27] The word "trace" has several different meanings in mathematics, each of them completely unrelated to the others. The author is using the word reluctantly in these footnote not yet written

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

- *continuously differentiable* if it is differentiable and $x'(t)$, $y'(t)$ are continuous in t; and
- *non-stop* if it is differentiable and $x'(t)$ and $y'(t)$ are never simultaneously zero (i.e. there is no t_0 for which $x'(t_0) = 0 = y'(t_0)$).

  ■

Definition 2.41 A curve $\mathcal{C}$ in $\mathbf{R}^2$ is *smooth* if for every point (x_0, y_0) on the curve, there is a number $\epsilon_0 > 0$ such that for all positive $\epsilon < \epsilon_0$, the portion of $\mathcal{C}$ lying inside the open square of side-length ϵ centered at (x_0, y_0) admits a continuously differentiable, nonstop parametrization, with domain an open interval. ■

"Admits", as used in Definition 2.41, is essentially another word for "has". We use the word "admits" because "has" might mislead the student into thinking that the curve has already been dropped on his/her plate with a regular parametrization; "*admits* a regular parametrization" does not lend itself to this misinterpretation.

The open-interval requirement at the end of Definition 2.41 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 convince him/herself that a circle meets our definition of "smooth curve".

Observe that Definition (2.41) uses a "windowing" idea similar to the one that we used to talk about implicitly-defined functions in Section 2.2. We will later give an equivalent definition of "smooth curve" that is even more reminiscent of that earlier discussion.

Every curve admits parametrizations that are not continuously differentiable and/or are not non-stop. Every *smooth* curve admits continuously differentiable parametrizations that do not meet the "non-stop" criterion, as well as those that do meet this criterion. But curves with corners, such as the graph of $y = |x|$, admit *no* continuously differentiable, nonstop parametrizations. We can parametrize the graph of $y = |x|$ continuously differentiably—for example, by $\gamma(t) = (t^3, |t|^3)$, 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" (you are parametrizing it using time as the parameter), but you'll find that you need to violate the "non-stop" criterion in order to do so.

- A figure-8. The whole curve does admit a continuously differentiable, non-stop parametrization, but the point (x_0, y_0) at which the curve crosses itself causes the definition of "smooth" not to be met. For small ϵ, the portion of the curve that lies in the disk of radius ϵ centered at (x_0, y_0) is essentially an X, and has the same problem that the X did.

Warning about terminology. Many calculus textbooks refer to a continously differentiable, non-stop parametrization as a *smooth* parametrization. This usage of "smooth" is unfortunate. It conflicts with the modern meaning of "smooth function" in advanced mathematics[29]. A preferable one-word term is "regular", and the only reason we are not using it in these notes is that the meaning of "regular" is not self-evident; we did not want to present the student with extra terminology to remember. "Regular" is flexible term that mathematicians use with a contextually varying meaning, which usually is "having the most common features" or "having no nasty or inconvenient features" (where the context determines what features are important). The meaning of *non-stop* is self-evident (regarding $\gamma'(t) = (x'(t), y'(t))$ as the velocity vector $\mathbf{v}(t)$ at time t associated with the parametrization, "non-stop" is the condition that the velocity vector is not the zero vector for any t), but the author of these notes has never seen it in any textbook[30].

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

Definition 2.42 A *solution curve* of a differential equation

$$M(x,y)dx + N(x,y)dy = 0 \tag{86}$$

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

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

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

$$M(x(t), y(t))\frac{dx}{dt} + N(x(t), y(t))\frac{dy}{dt} = 0 \tag{87}$$

for all t in the domain-interval I of the parametrization. In this context, we call γ a *parametrized solution* of (86).[31]

When no region R is specified, it is understood that the region of interest is the interior of the common implied domain of M and N. Here, "common implied domain" means the set of points at which both M and N are defined, and "interior" means that we don't count points that are on the boundary of the common domain[32].

■

For reasons too technical to discuss here, we will not define "maximal solution curve" for a general differential-form DE. In a later section, we will define this term under hypotheses that remove the technical difficulties.

As we noted previously, in a differential-form DE (86) there is neither an independent nor a dependent variable; x and y are treated symmetrically. This symmetry is preserved in (87), but in a surprising way: in (87), *both* x and y are dependent variables! The independent variable is t—a variable that is not even visible in (86).

Algebraic equivalence (see Definition 2.37) has the same importance for DEs in differential form that it has for DEs in derivative form. Suppose that two equations $M_1dx + N_1dy = 0$ and $M_2dx + N_2dy = 0$ are algebraically equivalent on a region R . Then there is a function f on R, nonzero at every point of R, such that $M_2 = fM_1$ and $N_2 = fN_1$. If $\mathcal{C}$ is a solution curve of $M_1dx + N_1dy = 0$ and $(x(t), y(t))$, $t \in I$, is a continuously differentiable, non-stop parametrization of $\mathcal{C}$, then

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

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

[31] The terminology "solution curve" and "parametrized solution" were invented for these notes; they are not standard.

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

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 parametrized solution of $M_2dx + N_2dy = 0$ is a solution curve or parametrized solution of $M_1dx + N_1dy = 0$. Thus:

> Two algebraically equivalent DEs in differential form have exactly the same solution curves, and exactly the same parametrized solutions.

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

Definition 2.42 implies more about solution curves and parametrized solutions than is obvious just from reading the definition.

To start with, equation (87) has a geometric interpretation. Let $(x(t), y(t))$ be a continuously differentiable, non-stop parametrization of some solution curve $\mathcal{C}$ of $Mdx + Ndy = 0$. Let $\mathbf{v}(t) = \gamma'(t) = x'(t)\mathbf{i} + y'(t)\mathbf{j}$, where $\mathbf{i}$ and $\mathbf{j}$ are the standard basis vectors in the xy plane. Then $\mathbf{v}(t)$, the velocity-vector function associated with the parametrization, is tangent to the smooth curve $\mathcal{C}$ at the point $(x(t), y(t))$. We can rewrite equation (87) using the dot-product you learned in Calculus 3:

$$(M(x(t), y(t))\mathbf{i} + N(x(t), y(t))\mathbf{j}) \cdot \mathbf{v}(t) = 0. \tag{88}$$

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

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

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

Since this holds for all points (x_0, y_0) on $\mathcal{C}$, it follows that the parametrization $x = f_2(t), y = g_2(t)$ also satisfies (87).[33] Thus if one continuously differentiable, non-stop parametrization of $\mathcal{C}$ satisfies (87), so does every other continuously differentiable, non-stop parametrization of $\mathcal{C}$. Therefore, even though Definition 2.42 requires only that there be *some* continuously differentiable, non-stop parametrization of $\mathcal{C}$ satisfying (87), once we know that even *one* continuously differentiable, non-stop parametrization of $\mathcal{C}$ has this property, they all do. Said another way:

$$\left.\begin{array}{l}\textit{Every} \text{ continuously differentiable, non-stop parametrization of a} \\ \text{solution curve of a differential equation } Mdx + Ndy = 0 \text{ is a} \\ \text{parametrized solution of this equation.}\end{array}\right\} \quad (89)$$

This gets back to the statement we made just prior to Definition 2.42: that a DE in differential form is looking for a curve. We did not say "*parametrized* curve". A curve is a geometric object, a certain type of point-set in the plane. The concept of *parametrized curve* is needed to define which point-sets are curves and which aren't. It's also needed to define many other features or properties of a curve, such as whether a curve is a solution curve of a (given) DE in differential form. Any property that is defined via parametrizations (such as being a solution curve of a DE in differential form) can potentially hold true for one parametrization but not for another. A property defined in terms of parametrizations is intrinsic to a (smooth) *curve*—the point-set traced out by any parametrization—if and only if the property holds true for *all* continuously differentiable, non-stop parametrizations of that curve. These are the properties that are truly *geometric*. What statement (89) is saying is that the property "I am a solution curve of this differential-form DE" is an intrinsic, geometric property.

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

Many smooth curves in $\mathbf{R}^2$ that are not graphs of one-variable functions can still be expressed entirely or "mostly" as a union of graphs of equations of the form "$y =$ differentiable function of x." But for many smooth curves, including those that arise as solution curves of differential equations in differential form, this is often

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

neither necessary nor desirable[34]. This is another fundamental difference between derivative-form DEs and differential-form DEs.

Example 2.43 Consider the equation

$$xdx + ydy = 0. \tag{90}$$

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

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

$t \in [0, 2\pi]$, is a parametrized solution. The solution curve it parametrizes is the circle $x^2 + y^2 = 25$, which is not the graph of a function of x. The circle is a beautiful smooth curve, and as far as the DE (90) is concerned, there is no reason to exclude any point of it.

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

2.5.2 The meaning of a differential

Now we are ready to ascribe meaning to a differential[35]. However, don't worry if you don't understand the meaning given below. Understanding it is not essential to the use of differentials in differential equations. In fact, in this section of the notes, there are no differential *equations*—just differentials.

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

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

[35] Differentials can be understood at different levels of loftiness. The level chosen for these notes is a higher level than the author has seen in Calculus 1-2-3 and introductory DE textbooks, but it is not the highest level.

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:

$$\underbrace{\overbrace{(Mdx+Ndy)[\gamma]}^{\substack{\text{the function obtained}\\ \text{when the differential}\\ \text{acts on } \gamma}}\ (t)}_{\substack{\text{value of the function}\\ (Mdx+Ndy)[\gamma]\\ \text{at } t}} = M(x(t),y(t))\frac{dx}{dt} + N(x(t),y(t))\frac{dy}{dt}\ . \tag{91}$$

The notation on the left-hand side of (91) may look intimidating and unwieldy, but it (or something like it) is a necessary evil for this section of the notes. It will not be used much outside this section.

Let us make contact between the meaning of differential given above, and what the student may have seen about differentials before. The easiest link is to differentials that arise as *notation* in the context of line integrals in Calculus 3. (Students who haven't completed Calculus 3 should skip down to the paragraph that includes equation (95), read that paragraph, and skip the rest of this section.) Recall that one notation for the line integral of a vector field $M(x,y)\mathbf{i} + N(x,y)\mathbf{j}$ over a smooth, oriented curve $\mathcal{C}$ in the xy plane is

$$\int_C M(x,y)dx + N(x,y)dy. \tag{92}$$

To see that the integrand in (92) is the same gadget we described above, let's review the rules you learned for computing such an integral:

1. Choose a continuously differentiable, nonstop parametrization γ of $\mathcal{C}$. Write this as $\gamma(t) = (x(t), y(t))$, $t \in [a,b]$.[36] 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 (92), make the following substitutions: $x = x(t), y = y(t), dx = \frac{dx}{dt}dt, dy = \frac{dy}{dt}dt$, and $\int_C = \int_a^b$. The integral obtained from these substitutions is

[36] The parametrization should also consistent with the given orientation of $\mathcal{C}$, and to be one-to-one, except that "$\gamma(a) = \gamma(b)$" is allowed in order to handle closed curves. These technicalities is unimportant here; the author is trying only to jog the student's memory, not to review line integrals thoroughly.

$$\int_a^b \left\{ M(x(t),y(t))\frac{dx}{dt} + N(x(t),y(t))\frac{dy}{dt} \right\} dt. \tag{93}$$

3. Compute the integral (93). The definition of (92) is the value of (93):

$$\int_C M(x,y)dx+N(x,y)dy = \int_a^b \left\{ M(x(t),y(t))\frac{dx}{dt} + N(x(t),y(t))\frac{dy}{dt} \right\} dt. \tag{94}$$

   (You also learn in Calculus 3 that this definition is self-consistent: no matter what continuously differentiable, non-stop parametrization of $\mathcal{C}$ you choose[37], you get the same answer.)

A casual glance at (94) suggests that we have used the following misleading equality:

$$\text{“}M(x,y)dx + N(x,y)dy = \left\{ M(x(t),y(t))\frac{dx}{dt} + N(x(t),y(t))\frac{dy}{dt} \right\} dt.\text{”} \tag{95}$$

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 (94), we did not confine our substitutions to the *integrand* of the integral on the left-hand side. We made the substitution “$\int_C \to \int_a^b$ ” as well. Attempting to equate *pieces* of the notation on the left-hand side with *pieces* of the notation on the right-hand side helps lead to a wrong impression of what is equal to what. Instead of making this fallacious attempt, understand that (94) 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 trace $\mathcal{C}$.

2. Once we choose such a γ, what is the integrand on the right-hand side? It is exactly the quantity $(Mdx+Ndy)[\gamma](t)$ in (91). The effect of the “$M(x,y)dx+N(x,y)dy$” on the left-hand side has been to produce the function $(Mdx+Ndy)[\gamma]$ when fed the parametrized curve γ.

[37] Subject to the other conditions in the previous footnote

Thus, the differential that appears as the integrand on the left-hand side is exactly the machine we described at the start of this section.

There is one other topic in Calculus 3 that makes reference to differentials (if the instructor chooses to discuss them at that time): the tangent-plane approximation of a function of two variables. The differentials you learned about in that context are not quite the same gadgets as the machines we have defined. They are related, but different. To demonstrate the precise relation, there are two things we would need to do: (1) restrict attention to exact differentials, and (2) discuss what kind of gadget the *value of a differential at a point*—an expression of the form $M(x_0, y_0)dx + N(x_0, y_0)dy$—is. This would require a digression that, in the interests of both brevity and comprehensibility, we omit.

2.5.3 Existence/uniqueness theorem for DEs in differential form

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

Just as for derivative-form IVPs, there is an Existence and Uniqueness Theorem for differential-form IVPs, which we will state shortly. To understand what's behind a restriction that will appear in the statement of this theorem, let us look again at equation (88). Suppose (x_0, y_0) lies on a smooth solution curve $\mathcal{C}$ of $Mdx + Ndy = 0$. $\underline{\text{If}}$ $M(x_0, y_0)$ 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 (88) 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 (88) is satisfied at $t = t_0$, and so is (87). There is no restriction at all on the slope!

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

Definition 2.44 A point (x_0,y_0) is a *singular point* of the differential $Mdx+Ndy$ if $M(x_0,y_0)=0=N(x_0,y_0)$. ■

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

$$\frac{dy}{dx}=f(x,y). \tag{96}$$

If the graph of a solution of (96) passes through (x_0,y_0), then the slope of the graph must be $f(x_0,y_0)$. This is true even if the IVP

$$\frac{dy}{dx}=f(x,y), \quad y(x_0)=y_0 \tag{97}$$

has more than one solution (which can happen if the hypotheses of the Existence and Uniqueness Theorem for derivative-form IVPs are not met, e.g. if $\frac{\partial f}{\partial y}$ is not continuous at (x_0,y_0)). So in some sense, a singular point (x_0,y_0) of a differential $Mdx+Ndy$ is a worse problem for the differential-form IVP "$Mdx+Ndy=0$ with initial condition (x_0,y_0)" than we ever see for the derivative-form IVP (97).

It is difficult to define "maximal solution curve" for an equation $Mdx+Ndy=0$ on a region in which $Mdx+Ndy$ has a singular point. But in regions free of singular points, there are no technical difficulties. We make the following definition[38]:

Definition 2.45 Let R be a region in which the differential $Mdx+Ndy$ has no singular points. A solution curve $\mathcal{C}$ of the equation $Mdx+Ndy=0$ is *maximal in R* if $\mathcal{C}$ is contained in R and either

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

2. $\mathcal{C}$ is an "open curve without endpoints" (i.e. $\mathcal{C}$ has a continuously differentiable, non-stop parametrization with domain an open interval,) and $\mathcal{C}$ is not a subset of another solution curve in R of the same DE. ■

[38] The terminology "solution curve that is maximal in a region" in Definition 2.45 was invented for these notes; the author does not know whether it is standard.

Less formally, a solution curve is maximal in R if it is inextendible to a larger solution curve in R. A smooth closed curve never has any directions in which it could be extended (without violating the definition of "smooth curve"), but an open curve without endpoints may or may not be extendible. For example, the graph $\mathcal{C}_1$ of $y = 1/x$ in the open first quadrant R is an open curve without endpoints that is inextendible because it already "runs off to infinity in both directions". It is a solution curve of the equation $ydx + xdy = 0$ that is maximal in R. (This differential has a singular point at the origin, but the origin is not in R, so Definition 2.45 applies.) The portion $\mathcal{C}_2$ of $\mathcal{C}_1$ for which $1 < x < 2$ is a solution curve of the same DE, but it is not maximal in R, since it can be extended to the larger solution curve $\mathcal{C}_1$ (of course, it can be extended to solution curves of intermediate size).

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

Theorem 2.46 *Suppose M and N are continuously differentiable functions on an open region R in $\mathbf{R}^2$, and that $Mdx + Ndy$ has no singular points in R. Then for every $(x_0, y_0) \in R$, there exists a unique maximal solution curve of $Mdx + Ndy = 0$ passing through (x_0, y_0).*

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

For *exact* differentials, singular points are familiar to students who've taken Calculus 3, but under another name:

Example 2.47 Suppose $Mdx + Ndy$ is exact on a region R, and let F be a function on R for which $Mdx + Ndy = dF$. Then $M = \frac{\partial F}{\partial x}$ and $N = \frac{\partial F}{\partial y}$. Hence (using the mathematician's notation " $\Longleftrightarrow$ "), for a given point $(x_0, y_0) \in R$,

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

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

2.5.4 Solutions (as opposed to "solution curves" or "parametrized solutions") of DEs in differential form

Definition 2.48 An equation

$$G(x,y)=0 \ \ \text{(or } G(x,y)\text{= any real number } c_0\text{)} \tag{98}$$

is a *solution* of a differential-form equation

$$M(x,y)dx+N(x,y)dy=0 \tag{99}$$

on a region R if

(i) the portion of the graph of (98) that lies in R contains a smooth curve, and

(ii) every smooth curve in R contained in the graph of (98) is a solution curve of (99).

If $R=\mathbf{R}^2$ then we usually omit mention of the region, and say just that (98) is a *solution* of (99).

If $Mdx+Ndy$ has no singular points in R, then a solution (98) is called *maximal* in R if its graph is a solution curve of $Mdx+Ndy=0$ that is maximal in R. ■

Observe that there is a certain structural similarity between Definition 2.4 and Definition 2.48 ("implicitsolutions", later re-named "implicit solutions" in Definition 2.5, of a derivative-form DE). In both definitions, the same object—an equation of the form (98)—is being given a solution-related name ("implicit solution" in the setting of derivative-form DEs, "solution" in the setting of differential-form DEs). In each definition there are two criteria to be met, of this form:

> (i) there is at one object with a certain property, say Property X, and
> (ii) every object with Property X also has some other property related to another type of solution.

We will elaborate on this similarity later.

Example 2.49 The circle with equation

$$x^2+y^2=53 \tag{100}$$

is a solution of

$$x\ dx+y\ dy=0. \tag{101}$$

Since the only singular point of $x\ dx+y\ dy$ is the origin, which does not lie on the graph of (100), the equation $x^2+y^2=53$ is a solution of (101) that is maximal in the region $\{\mathbf{R}^2$ minus the origin$\}$. ■

Example 2.50 The equation

$$xy = 1$$

is a solution of

$$ydx + xdy = 0. \tag{102}$$

The graph, a hyperbola, consists of two maximal solution curves that are maximal in the region $\{\mathbf{R}^2$ minus the origin$\}$. (Just as in the previous example, the origin is the only singular point of the differential.) One of the maximal solution curves admits the continuously differentiable, non-stop parametrization $x(t) = t,\ y(t) = \frac{1}{t},\ t \in (0, \infty)$, while the other admits the continuously differentiable, non-stop parametrization $x(t) = t,\ y(t) = \frac{1}{t},\ t \in (-\infty, 0)$.

More generally, for every real number C, the equation

$$xy = C$$

is a solution of the same DE (102). For most C, the graph is a hyperbola, but the case $C = 0$ is exceptional. The graph of

$$xy = 0 \tag{103}$$

is a pair of crossed lines, the x- and y-axes. Note that this graph is not a smooth curve, nor is it the *disjoint* union of two smooth curves the way a hyperbola is ("disjoint" meaning that the two curves have no points in common). We can verify that (103) is indeed a solution of (102) by observing that the parametrized curves $x(t) = t, y(t) = 0,\ t \in \mathbf{R}$ (a continuously differentiable, non-stop parametrization of the x-axis) and $x(t) = 0, y(t) = t,\ t \in \mathbf{R}$ (a continuously differentiable, non-stop parametrization of the y-axis) both satisfy

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

So we can express the graph of $xy = 0$ as the union of two solution curves of (102)—the graph of $y = 0$ and the graph of $x = 0$—but, unlike for the graph of $xy = C$, with $C \neq 0$ we cannot do it without having the two solution curves intersect. The source of this difference is that only for $C = 0$ does the graph of $xy = C$ contain $(0, 0)$, a singular point of $ydx + xdy$. ■

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

Example 2.51 (Exact equations, part 1) Suppose $Mdx + Ndy$ is an exact differential on a region R (see Definition 2.33), and let F be a differentiable function on R for which $Mdx + Ndy = dF$. Then (86) becomes

$$\frac{\partial F}{\partial x}\, dx + \frac{\partial F}{\partial y}\, dy = 0. \tag{104}$$

Suppose that $\mathcal{C}$ is a solution curve of (104), and that $g(t) = (x(t), y(t))$, $t \in I$, is a continuously differentiable parametrization of $\mathcal{C}$. Then (87) says

$$\frac{\partial F}{\partial x}(x(t), y(t))\frac{dx}{dt} + \frac{\partial F}{\partial y}(x(t), y(t))\frac{dy}{dt} = 0. \tag{105}$$

By the Chain Rule, the left-hand side of (105) is just $\frac{d}{dt}F(x(t), y(t))$. Thus, (87) simplifies, in this case, to

$$\frac{d}{dt}F(x(t), y(t)) = 0 \quad \text{for all } t \in I. \tag{106}$$

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

$$F(x(t), y(t)) = c_0 \tag{107}$$

for all $t \in I$. This implies that *every solution curve of* (104) *in* R *is contained in the graph of* (107) *for some value of the constant* c_0.

Now, fix a number c_0, and consider the equation

$$F(x, y) = c_0. \tag{108}$$

Is this equation a solution of (104) in R, according to Definition 2.48? The answer is yes, provided that criterion (i) of the definition is met. If this criterion is met, let $\mathcal{C}$ be a smooth curve in R that is contained in the graph of (108). Let γ be such a continuously differentiable parametrization of $\mathcal{C}$, and write $\gamma(t) = (x(t), y(t))$, $t \in I$. Since every point of $\mathcal{C}$ lies on the graph of (108), equation (107) is satisfied for all $t \in I$. Differentiating both sides of (107) with respect to t, we find that (106) is satisfied. But, by the Chain Rule, the left-hand side of (106) is exactly the left-hand side of (105), so (105) is satisfied. Therefore $\mathcal{C}$ is a solution curve of (104). Hence criterion (ii) of Definition 2.48 is met, so (108) is a solution of (104) in R. ■

Defining "general solution" for equations in differential form is trickier than it is for derivative form. One reason is that in differential form we have the notions both of *solution curve*—a geometric object—and *solution* (in the sense of Definition 2.48)—an algebraic equation (i.e. a non-differential equation). The other reason is that for differential-form DEs, some of the problems caused by singular points have no analog in derivative-form DEs. We will use the following definition:

Definition 2.52 [39] The *general solution* of a differential-form equation

$$Mdx + Ndy = 0 \tag{109}$$

in a region R is the collection of all solution curves in R.

We call a collection of algebraic equations in x and y the *general solution of* (109) *in* R (or *on* R), *in implicit form*, if

(i) each equation in the collection is a solution in the sense of Definition 2.48,

(ii) every solution curve of (109) in R that does not pass through a singular point of $Mdx + Ndy$ is contained in the graph of some equation in the collection, and

(iii) every solution curve of (109) in R, whether or not it passes through a singular point of $Mdx+Ndy$, is contained in the union of graphs of finitely many or countably many[40] equations in the collection.

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

We will explain the reason for criterion (iii) later.

Example 2.53 (Exact equations, part 2) Suppose we are given a differential-form equation (109) that is exact on a region R, and we have found a function F such that $Mdx + Ndy = dF$ on R. Then Example 2.51 shows that the general solution of (104) on R, in implicit form, is the collection of equations

$$F(x, y) = C, \tag{110}$$

where C is a "semi-arbitrary" constant: the allowed values of C are those for which the graph of (110) contains a smooth curve in R. ■

Above, if we assume more about the differential, it is easier to tell which C's are allowed:

[39] This definition was invented for these notes; it is not standard.

[40] The set $\mathbf{N}$ of natural numbers $\{1, 2, 3, \dots\}$ is an infinite set that is called *countable*, or *countably infinite*. More generally, the empty set and any set that can be indexed by a subset of $\mathbf{N}$ (for example, a collection of three curves $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$, or an infinite collection of curves $\{\mathcal{C}_n\}_{n=1}^{\infty}$) is called *countable*, and we say it has *countably many* elements. Every finite set is countable, so the phrase "finitely many or countably many" is redundant, but the author nonetheless wanted the student to see "finitely many" explicitly in Definition 2.52. Not every infinite set is countable; the set of all real numbers is an uncountable set.

Example 2.54 (Exact equations, part 3) In the setting of Example 2.53, assume additionally that M and N ($=\frac{\partial F}{\partial x}$ and $\frac{\partial F}{\partial y}$, respectively) are continuously differentiable in R, and that $Mdx + Ndy$ has no singular points (equivalently, F has no critical points) in R. We claim that in this case, the general solution of (104) on R, in implicit form, is (110), but where the allowed values of C are those for which the graph of (110) contains even a single *point* of R. Equivalently, *the set of allowed values of C is the range of F on the domain R.*

To see that this is the case, it suffices to show that if, for a given C, the graph of (110) contains a point (x_0, y_0) of R, then the graph contains a smooth curve in R. So, with C held fixed, assume there is such a point (x_0, y_0). 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 2.3) 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 (108) contained in $I_1 \times J_1$ is the graph of $y = \phi(x)$, i.e. the set of points $\{(x, \phi(x)) \mid x \in I_1\}$. This same set is the trace 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 trace of γ is a smooth curve contained in the graph of (110). Since $(x_0, y_0) \in R$, and R is an open set, a small enough segment of this curve, passing through (x_0, y_0), will be contained in R.

- If $\frac{\partial F}{\partial x}(x_0, y_0) \neq 0$, 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 (108) contained in $I_1 \times J_1$ is the graph of $x = \phi(y)$, i.e. the set of points $\{(\phi(y), y) \mid y \in J_1\}$. This graph is exactly the trace 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 trace of γ is again a smooth curve contained in the graph of (110), and again a small enough segment of it, passing through (x_0, y_0), will be contained in R. ■

Example 2.55 Consider again the DE from Example 2.49,

$$xdx + ydy = 0. \tag{111}$$

The left-hand side is the exact differential dF (on the whole plane $\mathbf{R}^2$), where $F(x,y) = \frac{1}{2}(x^2 + y^2)$. 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 $R = \{\mathbf{R}^2 \text{ minus the origin}\}$, there are no critical points in R, and Example 2.54 applies. For every $C > 0$, there is a point in R for which $\frac{1}{2}(x^2 + y^2) = C$. Therefore the general solution of $xdx + ydy = 0$ in R, in implicit form, is

$$\frac{1}{2}(x^2 + y^2) = C, \quad C > 0,$$

which we can write more simply as

$$x^2 + y^2 = C, \quad C > 0. \tag{112}$$

The graph of each solution is a circle. The collection of these circles is what we call the general solution of (111) in R (according to Definition 2.52), and the general solution in R fills out the region R.

If we look at (111) on the whole xy plane rather than just R, then Example 2.54 no longer applies (because of the critical point at the origin), but Example 2.53 still applies. From the 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 (111) in implicit form, with no restriction on the region, is the same as the general solution on R in implicit form, namely (112). ■

Example 2.56 Consider again the DE from Example 2.50,

$$ydx + xdy = 0. \tag{113}$$

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 2.54 applies. This time, for every $C \in \mathbf{R}$ there is a point in R for which $xy = C$. Therefore the general solution of $ydx + xdy = 0$ in R, in implicit form, is

$$xy = C, \tag{114}$$

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) that are maximal in R. For $C = 0$, there are four solution curves that are maximal in R: the positive x-axis, the negative x-axis, the positive y-axis, and the negative y-axis. The general solution of (113) (without the words "in implicit form") is the collection of all these half-hyperbolas and the four open half-axes circles is what we call (111) in R (according to Definition 2.52). The general solution in R again fills out R.

If we look at (113) on the whole xy plane rather than just R, then from the preceding, the only point we do not yet know to be on a solution curve is the origin. But, as we saw in Example 2.50, the origin *is* on two inextendible solution curves: the x-axis and the y-axis. So the general solution (without the words "in implicit form", and with no restriction on the region) is the set of the half-hyperbolas noted above, plus the x-axis and the y-axis. The general solution of (113) in implicit form, with no restriction on the region, is again (114). But in contrast to Example 2.55, this time the general solution fills out the whole plane $\mathbf{R}^2$. ■

Students who've taken Calculus 3 have studied equations taht 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.[41]) A level-set may have more than one *connected component*, such as the graph of $xy = 1$: there is no way to move along the portion of this hyperbola in the first quadrant, and reach the portion of the hyperbola in the third quadrant. Our definition of "smooth curve" prevents any level-set with more than one connected component from being called a smooth curve. However, it is often the case that a level-set is the union of several connected components, each of which is a smooth curve. From Examples 2.53 and 2.54 we can deduce the following:

[41] *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 is nowhere dense.

If F has continuous second partial derivatives in the region R, then the general solution of $dF = 0$ on R (see first sentence of Definition 2.52) **is the set of smooth curves in R that are contained in level-sets of F.**

If we additionally assume that F has no critical points in R, then the general *maximal* solution of $dF = 0$ on R—i.e. the collection of solution curves that are maximal in R—is the collection of connected components of level-sets of F in R. (See "Some cautionary notes on our terminology", item 1, later in this section.) (115)

Neither of these statements is an "if and only if". For example, the function $F(x, y) = xy$ has a critical point at the origin, but the general solution of $dF = 0$ is still the set of smooth curves in $\mathbf{R}^2$ that are contained in level-sets of F. (One of these smooth curves is the x-axis, one is the y-axis, and the others are half-hyperbolas.) For an example of a level-set that contains smooth curves, but is not a union of smooth curves (i.e. has a point that's not contained in any of the smooth curves in the level-set), see Example 2.59 later in this section.

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

Example 2.57 Consider the DE

$$2xy\ dx + (y^2 - x^2)dy = 0. \tag{116}$$

This DE is not exact on any region in the xy plane. However, the functions $M(x, y) = 2xy$ and $N(x, y) = y^2 - x^2$ are continuously differentiable on the whole plane, and the only point at which they are both zero is $(0, 0)$. So again, we have a differential with one singular point, which happens to be the origin[42] Again letting $R = \{\mathbf{R}^2 \text{ minus the origin}\}$, Theorem 2.46 guarantees us that through each point $(x_0, y_0) \neq (0, 0)$, there exists a unique solution curve of (116). (We could have used this theorem similarly in Examples 2.55 and 2.56, but there was no real need since we were able to solve these equations quickly, and just see directly that every point of R lay on a unique maximal-in-R solution curve.)

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 trace of this parametrized curve is the positive x-axis, and for all $t \in (0, \infty)$ we have

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

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

Similarly, the negative x-axis is a solution-curve. The uniqueness statement in Theorem 2.46 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 curves 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 (116) is algebraically equivalent to

$$\frac{1}{y^2}\left(2xy\ dx + (y^2 - x^2)dy\right) = 0. \tag{117}$$

But as the student may verify,

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

So on R_+, and also on R_-, the left-hand side of (117) is exact; it is dF, where $F(x, y) = \frac{x^2+y^2}{y}$. This differential has no singular points in R_+ or R_-, so Example 2.54 applies. The general solution of (117), in implicit form, on either of these regions, is

$$\frac{x^2 + y^2}{y} = C, \tag{118}$$

where set of allowed values of C is the range of F on each region. Since the sign of $\frac{x^2+y^2}{y}$ is the same as the sign of y, this means that on R_+, only positive C's will be allowed, and on R_-, only negative C's will be allowed. To see that these are the only restrictions on C, just set $x = 0$ in (117), and see that $F(0, C) = C$.

Now for some algebraic rearrangement. Let us write $C = 2b$ in (118). Then b is a semi-arbitrary constant with exactly the same restrictions as C ($b > 0$ for solution curves in R_+, $b < 0$ for solution curves in R_-). Then on each region,

$$\begin{aligned}& \frac{x^2 + y^2}{y} = 2b \\ \Longleftrightarrow\quad & x^2 + y^2 = 2by \\ \Longleftrightarrow\quad & x^2 + y^2 - 2by = 0 \\ \Longleftrightarrow\quad & x^2 + y^2 - 2by + b^2 = b^2 \\ \Longleftrightarrow\quad & x^2 + (y - b)^2 = b^2. \end{aligned} \tag{119}$$

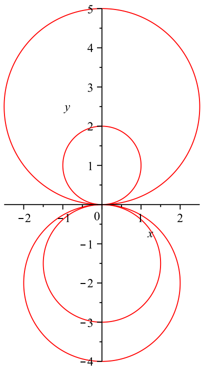

Figure 4: Some solution curves of $2xy\ dx + (y^2 - x^2)dy = 0$. (The graphing utility used does not do a good job near the origin; there should be no gap in any of the circles.)

The graph of (119) in $\mathbf{R}^2$ is a circle of radius $|b|$ centered at $(0, b)$ on the y-axis; the graph in R is the circle with the origin deleted. Thus, these circles-with-origin-deleted are the solution curves of (117) on R_+ and on R_-. But since (117) is algebraically equivalent to (116) on these regions, the same curves are all the solution curves of (116) in these regions.

We have now found all the solution curves of (116) in R that do not intersect the x-axis, as well as all those that do intersect it. So we have all the solution curves in $R = \{\mathbf{R}^2$ minus the origin$\}$. If we now re-include the origin, we see that the origin lies on every one of the circles (119), as well as on the x-axis. With the origin re-included, it is easy to see that the full x-axis is a solution curve of (116). We leave the student to check that each full circle (119), with the origin included, is also a solution curve of (116).

So it appears that the general solution of (116) consists of all circles centered on the y axis, plus one "exceptional" curve, the x-axis. We will see shortly that this does not meet our definition of "general solution", however. But what *is* correct is that the general solution of (116), in implicit form, is

$$\{x^2 + (y-b)^2 = b^2 \mid b \neq 0\} \text{ and } \{y = 0\}. \tag{120}$$

An alternative way of expressing the general solution in implicit form is as follows. In (118), C can be any nonzero constant, so we may write C as $\frac{1}{K}$, where the allowed values of K are also anything other than zero. We can then rewrite (118) as $y = K(x^2 + y^2)$. The solution curve that lie in R_+ have $K > 0$; those that lie in R_+ have $K < 0$. These give all the implicit-form solutions in the "b-family", just expressed in different-looking but algebraically equivalent way. But magically, if we now allow $K = 0$, we get the lonely $y = 0$ solution as well. So we can also write the general solution of (116), in implicit form, as

$$y = C(x^2 + y^2) \tag{121}$$

where C is a completely arbitrary constant. (We have renamed K back to C just because we felt like it.)

Now, why is it that the general solution of (116) (no "in implicit form") is not the collection of circles plus the x-axis? In Figure 4, start at a point other than the origin on any of the circles. Move along the 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 before you reach the origin again. Erase the endpoints of the curve you just drew (see the second paragraph after Definition 2.41), and you have a perfectly good, smooth, solution curve that is not contained in any circle or in the x-axis.

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

The phenomenon above is the reason we allow possibility (iii) in Definition 2.52.

■

Some cautionary notes on our terminology:

1. An extremely careful reader may have noticed that in the first part of the Definition 2.52 we do not require the solution curves to be maximal, as one might have expected from comparison with Definition 2.18 and the discussion before that definition. The reason is that we have defined maximal solution curves of $Mdx + Ndy = 0$ only in regions in which $Mdx + Ndy$ has no singular points, while Definition 2.52 allows for singular points. Because we do not insist on maximality of the curves in Definition 2.52, there is redundancy built into this definition that we were able to avoid in Definition 2.18: the general solution of (109) includes solution curves that are subsets of other solution curves.

   Example 2.57 illustrates 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 4 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 (116) 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.

2. Do not be misled by the terminology "<u>the</u> general solution of (109), in R, *in implicit form*." While there is only one general solution of (109) in R—the *collection* of all solution curves in R—there are infinitely many implicit forms of this general solution. Sometimes two different implicit forms of the same general solution in R may differ only in "trivial" ways; for example, if one implicit form of the general solution in R is a family of equations $F(x, y) = C$, then another is $2F(x, y) = C$, and another is $F(x, y)^3 = C$. But this is not the only way that the implicit forms of the same general solution can differ. We saw this in Example 116, and we see it again in the next example.

3. In Definition 2.52, the author chose to reserve the term "general solution" (with no extra words other than, perhaps, "in R") for the collection of all solution *curves*, because curves, and not functions or equations, are what a DE in differential form is looking for. An unfortunate consequence of this choice is that one must then decide what other term to use for a collection of algebraic equations whose graphs yield all the solution curves. The author's choice, "general solution in implicit form", has some definite disadvantages. Among these is the fact that the general solution in implicit form can be very explicit, as in the next example.

Example 2.58 Consider the DE

$$xdy - ydx = 0. \tag{122}$$

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

The only singular point of $xdy - ydx$ is the origin. Therefore in $R = \{\mathbf{R}^2$ minus the origin$\}$, there is a unique 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 solution curves of (122) in R. It follows that there are no inextendible solution curves in $\mathbf{R}^2$ other than what we get by re-including the origin, i.e., the family of all straight lines through the origin.

There are several ways we can write equations for this family of straight lines, i.e. write the general solution of (122) in implicit form, one of which is

$$\{y = Cx\} \ \text{ and } \ \{x = 0\}. \tag{123}$$

This grouping puts all the non-vertical lines into one family, and makes the vertical line look lonely. But another simple way of writing the general solution of (122) in implicit form is

$$\{x = Cy\} \ \text{ and } \ \{y = 0\}. \tag{124}$$

This groups all the non-horizontal lines together, and orphans the horizontal line instead. In contrast to what we saw in Example 116, in the current example there is no single family of equations, parametrized by one real-valued arbitrary (or semi-arbitrary) constant, that constitutes a general solution of (122) in implicit form. ■

Example 2.59 (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 $\frac{\partial F}{\partial x}(x,y) = \left\{ \begin{array}{cc} -3x^2, & x \geq 0 \\ 3x^2, & x \leq 0 \end{array} \right.$, so $\frac{\partial^2 F}{\partial x^2}(x,y) = \left\{ \begin{array}{cc} -6x, & x \geq 0 \\ 6x, & x \leq 0 \end{array} \right.$). It has one critical point, the origin. The level-set containing this critical point is the graph of

$$y^3 - |x|^3 = 0, \tag{125}$$

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

Equation (125) is a solution of

$$y^2 dy + \left\{ \begin{array}{cc} -3x^2, & x \geq 0 \\ 3x^2, & x \leq 0 \end{array} \right\} dx = 0; \tag{126}$$

it meets both criteria in Definition 2.48. However, the graph of (125) contains a point, $(0,0)$, that is not on any solution *curve* of (126) (see Definitions 2.42 and 2.41). 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 $dF = 0$. ■

2.6 Relation between differential form and derivative form

Suppose that $\mathcal{C}$ is smooth curve, and γ a continuously differentiable, non-stop parametrization of $\mathcal{C}$, with domain-interval I. Write $\gamma(t) = (f(t), g(t))$ (for what we are about to do, writing "$\gamma(t) = (x(t), y(t))$" would lead to confusion). Let's call 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.[43] (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.)

Since γ is a non-stop parametrization, for every $t \in I$ at least one of the two numbers $f'(t), g'(t)$ is nonzero. Hence every $t \in I$ lies in at least one of the sets $\{t \in I \mid f'(t) \neq 0\}$, $\{t \in I \mid g'(t) \neq 0\}$. It can be shown that each of these sets is a union of subintervals of I. Thus, every $t \in I$ lies in a subinterval I_1 that is either x-monotone or y-monotone.

Let I_1 be an x-monotone interval. Then $f'(t)$ not zero for any $t \in I_1$. The Inverse Function Theorem that you may have learned in Calculus 1 assures us that there is an inverse function f^{-1}, with domain an interval I_2 and with range I_1, and that f^{-1} is continuously differentiable[44]. Let $\mathcal{C}_1$ be the smooth curve parametrized by $(f(t), g(t))$ using just the x-nice open interval I_1 rather than the whole original interval I. On this domain, "$x = f(t)$" is equivalent to "$t = f^{-1}(x)$". So, temporarily writing $t_{\text{new}} = x$, for $(x, y) = (f(t), g(t)) \in \mathcal{C}_1$ we have

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

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

$$\gamma_{\text{new}}(t_{\text{new}}) = (t_{\text{new}}, \phi(t_{\text{new}})). \tag{127}$$

The variable in (127) is a "dummy variable"; we can give it any name we like. Since the x-component of $\gamma_{\text{new}}(t_{\text{new}})$ is simply the parameter t_{new} itself, we will simply use the letter x for the parameter; thus

$$\gamma_{\text{new}}(x) = (x, \phi(x)). \tag{128}$$

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 $y = \phi(x)$.

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

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

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

$$M(x,y)dx + N(x,y)dy = 0. \tag{129}$$

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

$$M(x(t),y(t))\frac{dx}{dt} + N(x(t),y(t))\frac{dy}{dt} = 0 \tag{130}$$

In particular this is true for the parametrization (128), in which the parameter t is x itself, and in which have $y(t) = \phi(t) = \phi(x) = y(x)$. Therefore, for all $x \in I_2$,

$$\begin{aligned} 0 &= M(x,\phi(x))\frac{dx}{dx} + N(x,\phi(x))\ \phi'(x) \\ &= M(x,\phi(x)) + N(x,\phi(x))\ \phi'(x). \end{aligned} \tag{131}$$

The right-hand side of (131) is exactly what we get if we substitute "$y = \phi(x)$" into $M(x,y) + N(x,y)\frac{dy}{dx}$. Hence ϕ is a solution of

$$M(x,y) + N(x,y)\frac{dy}{dx} = 0. \tag{132}$$

Therefore the portion $\mathcal{C}_1$ of $\mathcal{C}$ is the graph of a solution (namely ϕ) of the derivative-form differential equation (132).

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

$$M(x,y)\frac{dx}{dy} + N(x,y) = 0. \tag{133}$$

Therefore:

$$\left.\begin{array}{l}\textbf{Every solution curve of the differential-form equation (129)}\\ \textbf{is a union of graphs of solutions of the derivative-form}\\ \textbf{equations (132) and (133).}\end{array}\right\} \tag{134}$$

Note that the graphs mentioned in (134) will overlap, in general, since the x-monotone intervals and y-monotone intervals of a continuously differentiable, non-stop parametrization γ will usually overlap. (The only way there will not be an overlap is if $f'(t) \equiv 0$ or $g'(t) \equiv 0$, in which case $\mathcal{C}$ is a vertical or horizontal straight line, respectively, and there are, respectively, no x-monotone or y-monotone subintervals.)

We call (132) and (133) the *derivative-form equations associated with* (129). Similarly, we call (129) the differential-form equation associated with either of the equations (132), (133).

More generally, if a derivative-form equation is algebraically equivalent to (132) or (133) on a region R, we call the equation a derivative form of (129) on R. Similarly, if a differential-form equation is algebraically equivalent to (129) on a region R, we call the equation a differential form of (132) and (133) on R.[45]

Now compare (132) with the general first-order derivative-form DE with independent variable x and dependent variable y,

$$\mathsf{F}(x, y, \frac{dy}{dx}) = 0. \tag{135}$$

Equation (132) is a special case of (135), in which the dependence of F on its third variable is very simple. If we use a third letter z for the third variable of F, then (132) corresponds to taking $\mathsf{F}(x, y, z) = M(x, y) + N(x, y)z$, a function that can depend in any conceivable way on x and y, but is linear separately in z. In general, (135) could be a much more complicated equation, such as

$$\left(\frac{dy}{dx}\right)^3 + (x + y)\sin(\frac{dy}{dx}) + xe^y = 0. \tag{136}$$

Solving equations such as the one above is *much* harder than is solving equations of the simpler form (132). For certain functions F that are more complicated than (132), but much less complicated than (136), methods of solution are known. But there is not a highly-developed general theory for working with equation (135) for general F's.

One of the features of (132) that makes it so special is that on any region on which $N(x, y) \neq 0$, (132) is algebraically equivalent to

$$\frac{dy}{dx} = -\frac{M(x, y)}{N(x, y)}, \tag{137}$$

which is of form

$$\frac{dy}{dx} = f(x, y). \tag{138}$$

Recall that equation (138) is exactly the "standard form" equation that appears in the fundamental Existence and Uniqueness Theorem for initial-value problems. This

[45] This is more restrictive than the analogous statement in the textbook from which the author is currently teaching, which omits the requirement of algebraic equivalence. This textbook, and others, allow multiplication/division by functions that can be zero. But this can lead to the omission of one or more solutions of the original DE, or the inclusion of one or more spurious solutions—functions (or curves) that are not solutions (or solution curves) of the original DE—when trying to write down the general solution of the original DE.

theorem is absolutely crucial in enabling us to determine whether our techniques of finding solutions actually give us *all* solutions.

If you re-read these notes, you will see that all the *general* facts about DEs in derivative form—such as the definition of “solution”, “implicit solution”, “general solution”, and the fact that algebraically equivalent DEs have the same set of solutions—were stated for the general first-order DE (135). These facts apply just as well to nasty DEs like (136) as they do to (relatively) nice ones like (135). However, in all of our *examples*, we used equations that were algebraically equivalent to (132) on some region (hence also to (138)). The reason is that although the concept of “the set of all solutions” makes perfectly good sense for the general equation (135), the author wanted to use examples in which he could show the student easily that the set of all solutions had actually been found.

Nowadays, students in an introductory DE course rarely see any first-order derivative-form equations that are not algebraically equivalent, on some region, to a DE in the standard form (138). Because of this, it is easy to overlook a significant fact: the *only* derivative-form DEs that are related to differential-form DEs are those that are algebraically equivalent to (138) on some region. The two types of equations, in full generality, are not merely two sides of the same coin.

However, for derivative-form DEs that can be “put into standard form”—which are exactly those that are algebraically equivalent to a DE of the form (132)—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 (134) gives one such relation.

Let us say that a derivative-form equation, with independent variable x and dependent variable y, is in “almost standard form”[46] if it is in the form (132), 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 below equation (133), you will see that it also shows that the graph of every solution of (132) or (133) is a solution curve of (129). Thus:

The graph of every solution of a derivative-form equation in almost-standard form is a solution curve of the associated differential-form equation. (139)

Combining (134) and (139), we conclude the following:

A smooth curve $\mathcal{C}$ is a solution curve of an equation in differential form if and only if $\mathcal{C}$ is a union of graphs of solutions of the associated derivative-form equations. (140)

[46] Another bit of terminology invented only for these notes, just to have a name to distinguish (132) from (137) on regions in which $N(x, y)$ may be zero somewhere.

We emphasize that in deriving these relations, the transition from the differential-form DE (129) to the derivative-form DEs (132) and (133) 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 (104) is curves whose parametrizations satisfy (130), and that for particular choices of the parameter (valid on the intervals that we called "x-monotone" or "y-monotone") (130) reduces to (132) or (133).

Similarly, transitions from derivative form to differential form are NOT achieved by the nonsensical process of "multiplying by dx" or "multiplying by dy". The beauty of the Leibniz notation " $\frac{dy}{dx}$ " for derivatives is that it can be used to help remember many true statements by *pretending, momentarily*, that you can multiply or divide by a differential just as if it were a real number[47]. In particular, we can use this principle help us easily *remember* that the differential-form equation (129) is related to (but not the same as!) the derivative-form equations (132) and (133). But this notational trick doesn't tell us everything, such as the *precise relationship* among these equations, which is statement (139) (of which statement (134) is the "only if" half).

Now let us turn to the way that differential-form DEs are used to help us find solutions of almost-standard-form derivative-form DEs. In this setting, we start with an equation of the form (132) (or one that can be put in this form by subtracting one side of the equation from the other). We then look at the associated differential-form equation $M(x,y)dx + N(x,y)dy = 0$, which treats x and y symmetrically, remembering that what we want in the end are solutions that are *functions of* x. Suppose that $\mathcal{C}$ is a solution curve of $M(x,y)dx+N(x,y)dy = 0$. Then, from statement (134), every solution curve is a union of (usually overlapping) sub-curves, each of which is either a solution $y = \phi(x)$ of (132), or a solution $x = \phi(y)$ of (133). But what are looking for now is solutions only of the first type. $\mathcal{C}$ may contain a vertical line segment, but such a segment is not contained in the graph of any equation of the form $y = \phi(x)$. However, if we delete from $\mathcal{C}$ any points at which the tangent line is vertical, remains is a union of graphs of solutions of (132).

That describes the *geometric* relation between solutions of $Mdx + Ndy = 0$ and solutions of (132), but what can we say in terms of formulas? Let us suppose that (for our given M and N) we have found a solution $G(x,y) = c_0$ of $M(x,y)dx+N(x,y)dy = 0$. Referring back to (2.48), this implies that

(a) the graph of $G(x,y) = c_0$ contains a smooth curve,

and that

[47] Simultaneously, the weakness 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.

(b) any portion of this graph that's a smooth curve is a solution curve of $Mdx + Ndy = 0$.

We ask the question: is $G(x, y) = c_0$ an implicit solution of our original derivative-form equation (132)?

To answer this question, we go back to Definition 2.4. In order for $G(x, y) = c_0$ to be an implicit solution of (132), its graph must, first of all, contain the graph of some solution $y = \phi(x)$ of (132). Focusing on the fact that such a solution is a differentiable function of x, we ask: is it ever possible for a graph of $G(x, y) = c_0$ *not* to contain the graph of a differentiable function of x, on *any* interval, no matter how tiny?

The graph of $G(x, y) = c_0$ contains points of (potentially) two types: those that lie in a smooth curve contained in the graph, and those that do not. Let's suppose that $\mathcal{C}$ is a smooth curve lying in the graph of $G(x, y) = c_0$, but assume that this graph does not contain the graph of a differentiable function of x. Let $\gamma(t) = (f(t), g(t))$ be a continuously differentiable, non-stop parametrization of $\mathcal{C}$, with parameter-interval I. In the language we used in the argument leading to (134), if I contains an x-monotone interval, then that argument shows that $\mathcal{C}$ contains the graph of a differentiable function of x, which would contradict our assumption. Therefore I contains no x-monotone intervals, so $f'(t) \equiv 0$ on I. Therefore f is constant; we have $f(t) \equiv x_0$ for some x_0. Hence $\mathcal{C}$ is contained in the vertical line $\{x = x_0\}$.

This shows that if the graph of $G(x, y) = c_0$ does not contain the graph of a differentiable function of x, then the graph consists entirely of segments of vertical lines, plus any points of the graph not contained in a smooth curve.

It can be shown that if the function G is differentiable—which will usually be the case if the equation $G(x, y) = c_0$ is found by the techniques used in an introductory DE course—and the graph of $G(x, y) = c_0$ satisfies all the conditions above, then there are *no* points on this graph that do not lie on a smooth curve in the graph, and the graph consists entirely of vertical lines. From this, it can further be shown that $G(x, y)$ is a function of x alone. (For example, the equation $G(x, y) = c_0$ could be $x = 3$, whose graph in the xy plane is a single vertical line, or $x^2 - 1 = 0$, whose graph is two vertical lines; or $\sin x = 0$, whose graph is an infinite collection of vertical lines.) In this case, the solution "$G(x_0, y_0) = c_0$" of $M(x, y)dx + N(x, y)dy = 0$ is *not* an implicit solution of $M(x, y) + N(x, y)\frac{dy}{dx} = 0$.

So if $G(x, y)$ is differentiable and is not a function of x alone, then the graph of $G(x, y) = c_0$ *does* contain the graph of some differentiable function ϕ of x. The graph of $y = \phi(x)$ is a smooth curve lying in the graph of $G(x, y) = c_0$. Referring to (b) above, we see that this implies that the graph of $y = \phi(x)$ is a solution curve of $Mdx + Ndy = 0$. The argument leading from the sentence that includes (129) to the sentence that includes (132) then shows that ϕ is a solution of (132).

To recap: we have shown that if the equation $G(x, y) = c_0$ is a solution of $M(x, y)dx + N(x, y)dy = 0$, and G is differentiable, then:

either $G(x,y)$ is a function of x alone, in which case the equation $G(x,y) = c_0$ is not an implicit solution of $M(x,y) + N(x,y)\frac{dy}{dx} = 0$, or

$G(x,y)$ is not a function of x alone, in which case the equation $G(x,y) = c_0$ *is* an implicit solution of $M(x,y) + N(x,y)\frac{dy}{dx} = 0$. (141)

Since the graph of every solution of $M + N\frac{dy}{dx}$ is a solution curve of $Mdx + Ndy = 0$, (141) implies the following:

Suppose that we have a general solution, in implicit form, of a differential-form equation $Mdx + Ndy = 0$. Further suppose that each equation in the collection comprising the general solution is of the form $G(x,y) =$ constant (not necessarily the same G for all equations in the general implicit-form solution), where G is differentiable. Then the collection of equations obtained by deleting those equations for which $G(x,y)$ depends only on y, is the general solution, in implicit form, of the associated derivative -form equation $M + N\frac{dy}{dx} = 0$. (142)

Example 2.60 (Exact equations, part 4) Suppose that we wish to solve a DE of the form

$$M(x,y) + N(x,y)\frac{dy}{dx} = 0 \tag{143}$$

on a region R on which $N(x,y)$ is not identically zero (if $N(x,y)$ were identically zero, then (143) would reduce to the *algebraic* equation $M(x,y) = 0$, not a true differential equation). If the associated differential-form equation is exact on R, and we have found a function F such that $Mdx + Ndy = dF$ on R, then Example 2.53 tells us that the general solution of $Mdx + Ndy = 0$ on R, in implicit form, is the family of equations

$$F(x,y) = C \tag{144}$$

where C is a "semi-arbitrary" constant. The function F is automatically differentiable, so (142) applies: unless F is a function of x alone, each of the equations (144) is an implicit solution of (143). But if F is a function of x alone, then $N(x,y) = \frac{\partial F}{\partial y}(x,y) \equiv 0$. Therefore *if $Mdx + Ndy = dF$ on R, then* (144) *is the general solution of* (143) *in implicit form* (i.e. it is not just the general solution, in implicit form, of the associated differential-form equation). ■