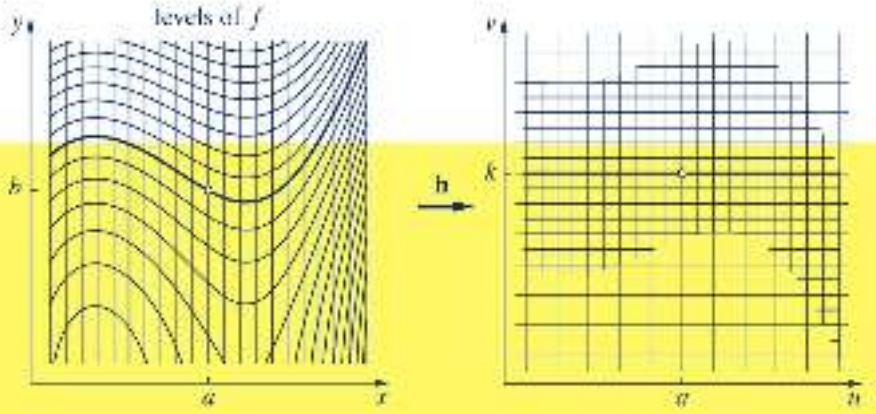
Advanced Calculus

A Geometric View



Springer

Undergraduate Texts in Mathematics

For other titles Published in this series, go to
www.springer.com/series/666

James J. Callahan

Advanced Calculus

A Geometric View

Springer

James J. Callahan
Department of Mathematics and Statistics
Smith College
Northampton, MA 01063
USA
callahan@math.smith.edu

ISSN 0172-6056
ISBN 978-1-4419-7331-3 e-ISBN 978-1-4419-7332-0
DOI 10.1007/978-1-4419-7332-0
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010935598

Mathematics Subject Classification (2010): 26-01, 26B12, 26B15, 26B10, 26B20, 26A12

To my teacher, Linus Richard Foy

Preface

A half-century ago, advanced calculus was a well-defined subject at the core of the undergraduate mathematics curriulum. The classic texts of Taylor [19], Buck [1], Widder [21], and Kaplan [9], for example, show some of the ways it was approached. Over time, certain aspects of the course came to be seen as more significant—those seen as giving a rigorous foundation to calculus—and they became the basis for a new course, an introduction to real analysis, that eventually supplanted advanced calculus in the core.

Advanced calculus did not, in the process, become less important, but its role in the curriculum changed. In fact, a bifurcation occurred. In one direction we got calculus on n-manifolds, a course beyond the practical reach of many undergraduates; in the other, we got calculus in two and three dimensions but still with the theorems of Stokes and Gauss as the goal.

The latter course is intended for everyone who has had a year-long introduction to calculus; it often has a name like *Calculus III*. In my experience, though, it does not manage to accomplish what the old advanced calculus course did. Multivariable calculus naturally splits into three parts: (1) several functions of one variable, (2) one function of several variables, and (3) several functions of several variables. The first two are well-developed in Calculus III, but the third is really too large and varied to be treated satisfactorily in the time remaining at the end of a semester. To put it another way: Green's theorem fits comfortably; Stokes' and Gauss' do not.

I believe the common view is that any such limitations of Calculus III are at worst only temporary because a student will eventually progress to the study of general k-forms on n-manifolds, the proper modern setting for advanced calculus. But in the last half-century, undergraduate mathematics has changed in many ways, not just in the flowering of rigor and abstraction. Linear algebra has been brought forward in the curriculum, and with it an introduction to important multivariable functions. Differential equations now have a larger role in the first calculus course, too; students get to see something of their power and necessity. The computer vastly expands the possibilities for computation and visualization.

The premise of this book is that these changes create the opportunity for a new geometric and visual approach to advanced calculus.

* * *

More than forty years ago—and long before the curriculum had evolved to its present state—Andrew Gleason outlined a modern geometric approach in a series of lectures, "The Geometric Content of Advanced Calculus" [8]. (In a companion piece [17], Norman Steenrod made a similar assessment of the earlier courses in the calculus sequence.) Because undergraduate analysis bifurcated around the same time, Gleason's insights have not been implemented to the extent that they might have been; nevertheless, they fit naturally into the approach I take in this book.

Let me try to describe this geometric viewpoint and to indicate how it hangs upon recent curricular and technological developments. Geometry has always been bound up with the teaching of calculus, of course. Everyone associates the derivative of a function with the slope of its graph. But when the function becomes a map $\mathbf{f} : \mathsf{R}^n \to \mathsf{R}^p$ with $n, p \geq 2$, we must ask: Where is the graph? What is its slope at a point? Even in the simplest case $n = p = 2$, the graph (a two-dimensional surface) lies in R^4 and thus cannot be visualized directly. Nevertheless, we can get a picture if we turn our attention from the graph to the image, because the image of $\mathbf{f}$ lies in the R^2 target. Computer algebra systems now make such pictures a practical possibility. For example, the *Mathematica* command `ParametricPlot` produces a nonlinear grid that is the image under a given map of a uniform coordinate grid from its source. We can train ourselves to learn as much about a map from its image grid as we learn about a function from its graph.

How do we picture the derivative in this setting? When we dealt with graphs, the derivative of a nonlinear function f at the point a was the linear function whose graph was tangent to the graph of f at a. Tangency implies that, under progressive magnification at the point $(a, f(a))$, the two graphs look more and more alike. At some stage the nonlinear function becomes indistinguishable from the linear one. There are two subtly different concepts at play here, depending on what we mean by "indistinguishable." One is *local linearity* (or *differentiability*): $f(a + \Delta x) - f(a)$ and $f'(a)\Delta x$ are indistinguishable in the technical sense that their difference vanishes to greater than first order in Δx. The other is *looking linear locally*: the graphs themselves are indistinguishable under sufficient magnification. For our function f, there is no difference: f is locally linear precisely where it looks linear locally.

There is a real and important difference, though, when we replace graphs by image grids, as we must do to visualize a map $\mathbf{f} : \mathsf{R}^2 \to \mathsf{R}^2$ and its derivative $\mathrm{d}\mathbf{f}_\mathbf{a}$. We say $\mathbf{f}$ is *locally linear* (or *differentiable*) at $\mathbf{a}$ if $\mathbf{f}(\mathbf{a} + \Delta\mathbf{x}) - \mathbf{f}(\mathbf{a})$ and $\mathrm{d}\mathbf{f}_\mathbf{a}(\Delta\mathbf{x})$ are indistinguishable in the sense that their difference vanishes to greater than first order in $\|\Delta\mathbf{x}\|$. By contrast, we say $\mathbf{f}$ *looks linear locally* at $\mathbf{a}$ if the image grid of $\mathbf{f}$ near $\mathbf{a}$ is indistinguishable from the image grid of $\mathrm{d}\mathbf{f}_\mathbf{a}$ under sufficient magnification. To make the difference clear, consider the quadratic map $\mathbf{q}$ and its derivative at a point $\mathbf{a} = (a, b)$:

$$\mathbf{q} : \begin{array}{l} u = x^2 - y^2, \\ v = 2xy; \end{array} \qquad \mathrm{d}\mathbf{q}_\mathbf{a} = \begin{bmatrix} 2a & -2b \\ 2b & 2a \end{bmatrix}.$$

Because the derivative exists everywhere, $\mathbf{q}$ is locally linear everywhere. Moreover, $\mathbf{q}$ also looks like its derivative under sufficient magnification as long as $\mathbf{a} \neq \mathbf{0}$. But

at the origin, **q** doubles angles and squares distances, and continues to do so at any magnification. No linear map does this. Thus in no open neighborhood of the origin does **q** look like any linear map, and certainly not its derivative, which is the zero map. (There is no contradiction, of course, because the difference between **q** and its derivative vanishes to second order at the origin.)

Quite generally, a locally linear map $\mathbf{f} : \mathsf{R}^n \to \mathsf{R}^n$ need not look linear at a point; however (as our example suggests), if the derivative is invertible at that point, the map *will* look linear there. In fact, this is the essential geometric content of the inverse function theorem. Here is why. By hypothesis, a linear coordinate change will transform the derivative into the identity map. The local inverse for **f** that is provided by the theorem can be viewed as another coordinate change, one that transforms **f** itself into the identity map, at least locally. Thus **f** must look like its derivative locally because a suitable (composite) coordinate change will transform one into the other. This leads us, in effect, to gather maps into geometric equivalence classes: two maps are equivalent if a coordinate change transforms one into another. In other words, a class consists of different coordinate descriptions of the same geometric action. The invertible maps together make up a single class. (Geometrically, there is only one invertible map!)

For parametrized surfaces $\mathbf{f} : \mathsf{R}^2 \to \mathsf{R}^3$, or more generally for maps in which the source and target have different dimensions, invertibility of the derivative is out of the question. The appropriate notion here is maximal rank. Then, at a point where the derivative has maximal rank, the implicit function theorem implies that the map and its derivative once again look alike in a neighborhood of that point. Coordinate changes convert both into the standard form of either a linear injection or a linear projection. For each pair of source–target dimensions, maps whose derivatives have maximal rank at a point make up a single local geometric class.

A nonlinear map can certainly have other local geometric forms; for example, a plane map can fold the plane on itself or it can wrap it doubly on itself (like **q**, above). The inverse and implicit function theorems imply that all such local geometric forms must therefore occur at points where the derivative fails to have maximal rank. Such points are said to be *singular*. The analysis of the singularities of a differentiable map is an active area of current research that was initiated by Hassler Whitney half a century ago [20] and guided to a mature form by René Thom in the following decades. Although this book is not about map singularities, its geometric approach reflects the way singularities are analyzed. There are further connections. In 1975, I wrote a survey article on singularities of plane maps [2]; one of my aims here is to provide more detailed background for that article.

We do analyze singularities in one familiar setting: a real-valued function f. The target dimension is now 1, so only the zero derivative fails to have maximal rank. This happens precisely at a critical point, where all the linear terms in the Taylor expansion of f vanish. So we turn to the quadratic terms, that is, to the quadratic form Q defined by the Hessian matrix of the function at that point. Taylor's theorem assures us that the Hessian form approximates f near the critical point (up to terms that vanish to third order). We ask: does f also look like its Hessian form near that point?

Some condition is needed; for example, $f(x,y) = x^2 - y^4$ does not look like its quadratic part $Q(x,y) = x^2$ near the origin. Morse's lemma provides the condition: f does look like Q near a critical point if the Hessian matrix has maximal rank. That is to say, a local coordinate change in a neighborhood of the critical point will transform the original function into its Hessian form, in effect, removing all higher-order terms in the Taylor expansion of f.

A nondegenerate Hessian therefore has an invariant geometric meaning, but only at a critical point. At a noncritical point, even concavity, for example, fails to be preserved under all coordinate changes. More generally, if linear terms are present and "robust" in the Taylor expansion of f at a point (i.e., they define a linear map that has maximal rank), quadratic and higher terms have no invariant geometric meaning. This is the implicit function theorem speaking once again.

By asking whether a map looks like the beginning of its Taylor series, we are led to see the underlying geometric character of the inverse and implicit function theorems and Morse's lemma. The question thus provides a way to organize and unify much of our subject and, in so doing, to bring out its simple beauty.

Let me now describe the geometric approach this book takes to another of its central themes: the change of variables formula for integrals.

To fix ideas, suppose we have a double integral, so the change of variables is an invertible map of (a portion of) the plane. Locally, that map looks linear. Each linear map has a characteristic factor by which it magnifies areas. To a nonlinear map we can therefore assign a *local* area magnification factor at each point, the area magnification factor of its local linear approximation at that point. This is the Jacobian.

In the simplest case, the integrand is identically equal to 1, and the value of the integral is just the area of the domain of integration. A change of variables maps that domain to a new one with, in general, a different area. If the map is linear, and has area multiplication factor M, the new area is just M times the old (or the integral of the constant M over the old domain). However, if the map is nonlinear, then we need to proceed in steps. First subdivide the old domain into small regions on each of which the local area magnification factor M (the Jacobian) is essentially constant. The area of the image of one small region is then approximately the product of its own area and the local value of M, and the area of the entire image is approximately the sum of those individual products. To get better approximations, make finer and finer subdivisions; in the limit, we have the area of the new region as the integral of the local area multiplication function M over the original domain. For an arbitrary integrand, transform the integral the same way: multiply the integrand by M. All of this is easily generalized from two to n variables; areas become n-volumes.

A typical proof of the change of variables formula proceeds one dimension at a time; this tends to submerge the geometric force and meaning of the Jacobian M. By contrast, my proof in Chapter 9 follows the geometric argument above. I found it in an article by Jack Schwartz ([16]), who remarks that his proof appears to be new; he could not find a similar argument in any of the standard calculus texts of the time.

* * *

One way I have chosen to stress the geometric is by concentrating on what happens in two and three dimensions, where we can construct—with the help of a computer algebra system as needed—illustrations that help us "see" theorems. And this is not a bad thing: the words *theorem* and *theatre* stem from the same Greek root $\theta\varepsilon\overline{\alpha}$, "the act of seeing." In a literal sense, a theorem is "that which is seen." But the eye, and the mind's eye not less, can play tricks. To be certain a theorem is true, we know we must test what we see. Here is where proof comes in: *to prove* means "to test." The cognate form *to probe* makes this more evident; probate tests the validity of a will. Ordinary language supports this meaning, too: yeast is "proofed" before it is used to leaven bread dough, "the proof of the pudding is in the eating," and "the exception proves the rule" because it tests how widely the rule applies.

In much of mathematical exposition, *proving* is given more weight than *seeing*. Jean Dieudonné's seminal *Foundations of Modern Analysis* [4] is a good example. In the preface he argues for the "necessity of a strict adherence to axiomatic methods, with no appeal to 'geometric intuition', at least in the formal proofs: a necessity which we have emphasized by deliberately abstaining from introducing any diagram in the book." As prevalent as it is, the axiomatic tradition is not the only one. René Thom, a contemporary of Dieudonné and Bourbaki, followed a distinctly different geometric tradition in framing the study of map singularities, a study whose outlines have guided the development of this book. Although proof may be given a different weight in the geometric tradition, it still has a crucial role. I believe that a student who sees a theorem more fully has all the more reason to test its validity.

But there is, of course, usually no reason to restrict the proofs themselves to low dimensions. For example, my proof of the inverse function theorem (Chapter 5, p. 169ff.) is for maps on $\mathbb{R}^n$. It elaborates upon Serge Lang's proof for maps on infinite-dimensional Banach spaces [10, 11]. Incidentally, Lang points out that, in finite dimensions, the inverse function theorem is often proven using the implicit function theorem, but that does not work in infinite dimensions. Lang gives the proofs the other way around, and I do the same. Furthermore, because there is so much instructive geometry associated with implicit functions, I provide not just a general proof but a sequence of more gradually complicated ones (Chapter 6) that fold in the growing geometric complexity that additional variables entail. I think the student benefits from seeing all this put together. Other important examples of n-dimensional proofs of theorems that are visualized primarily in $\mathbb{R}^2$ are Taylor's theorem (Chapter 3), the chain rule (Chapter 4), and Morse's lemma (Chapter 7). The definition of the derivative gets the same kind of treatment as the proof of the implicit function theorem, and for the same reason. Unlike the other topics, integral proofs are mainly restricted to two dimensions. One reason is that the many technical details about Jordan content are easiest to see there. Another reason is that the extension to higher dimensions is straightforward and can be carried out by the student.

At a couple of points in the text, I provide brief *Mathematica* commands that generate certain 3D images. Because programs like *Mathematica* are always being updated (and the *Mathematica 5* code I have used in the text has already been superceded), details are bound to change. My aim has simply been to indicate how

easy it is to generate useful images. I have also included a simple BASIC program that calculates a Riemann sum for a particular double integral. Again, it is not my aim to advocate for a particular computational tool. Nevertheless, I do think it is important for students to see that programs do have a role—integrals arise out of computations—and that even a simple program can increase our power to estimate the value of an integral.

To help keep the focus on geometry, I have excluded proofs of nearly all the theorems that are associated with introductory real analysis (e.g., those concerning uniform continuity, convergence of sequences of functions, or equality of mixed partial derivatives). I consider real analysis to be a different course, one that is treated thoroughly and well in a variety of texts at different levels, including the classics of Rudin [15] and Protter and Morrey [14]. To be sure, I am recalibrating the balance here between that which is seen and that which is tested.

This book does not attempt to be an exhaustive treatment of advanced calculus. Even so, it has plenty of material for a year-long course, and it can be used for a variety of semester courses. (As I was writing, it occurred to me that a course is like a walk in the woods—a personal excursion—but a text must be like a map of the whole woodland, so that others can take walks of their own choosing.) My own course goes through the basics in Chapters 2–4 and then draws mainly on Chapters 9–11. A rather different one could go from the basics to inverse and implicit functions (Chapters 5 and 6), in preparation for a study of differentiable manifolds. The pace of the book, with its numerous visual examples to introduce new ideas and topics, is particularly suited for independent study. From start to finish, illustrations carry the same weight as text and the two are thoroughly interwoven. The eye has an important role to play.

In addition to the *CUPM Proceedings* [12] that contain the lectures of Gleason and Steenrod, I have been strongly influenced by the content and tone of the beautiful three-volume *Introduction to Calculus and Analysis* [3] by Richard Courant and Fritz John. In particular, I took their approach to integration via Jordan content. At a different level of detail, I adopted their phrase *order of vanishing* as a replacement for the less apt *order of magnitude* for vanishing quantities. For the theorems connecting Riemann and Darboux integrals in Chapter 8, I relied on Protter and Morrey [14]; my own contribution was a number of figures to illustrate their proofs. It was Gleason who argued that the Morse lemma has a place in the undergraduate advanced calculus course. I was fully persuaded after my student Stephanie Jakus (Smith ’05) wrote her senior honors thesis on the subject.

The *Feynman Lectures on Physics* [6] have had a pervasive influence on this book. First of all, Feynman’s vision of his subject, and his flair for explanation, is awe-inspiring. I felt I could find no better introduction to surface integrals than the context of fluid flux. Because physics works with two-dimensional surfaces in $\mathbb{R}^3$, I also felt justified in concentrating my treatment of surface integrals on this case. I believe students will have learned all they need in order to deal with the integral of a k-form over a k-dimensional parametrized surface patch in $\mathbb{R}^n$, for arbitrary $k < n$. In providing a physical basis for the curl, the *Lectures* prodded me to try to

understand it geometrically. The result is a discussion of the curl (in Chapter 11) that—like the discussion of the Morse lemma—has not previously appeared in an advanced calculus text, as far as I am aware.

I thank my students over the last decade for their curiosity, their perseverance, their interest in the subject, and their support. I especially thank Anne Watson (Smith ’09), who worked with me to produce and check exercises. My editor at Springer, Kaitlin Leach, makes the rough places smooth; I am most fortunate to have worked with her. I am grateful to Smith College for its generous sabbatical policy; I wrote much of the book while on sabbatical during the 2005–2006 academic year. My deepest debt is to my teacher, Linus Richard Foy, who stimulated my interest in both mathematics and teaching. In his advanced calculus course, I often caught myself trying to follow him along two tracks simultaneously: what he was saying, and how he was saying it.

Amherst, MA
June 2010 *James Callahan*

Contents

Chapter 1
Starting Points

Abstract Our goal in this book is to understand and work with integrals of functions of several variables. As we show, the integrals we already know from the introductory calculus courses give us a basis for the understanding we need. The key idea for our future work is change of variables. In this chapter, we review how we use a change of variables to compute many one-variable integrals as well as path integrals and certain double integrals that can be evaluated by making a change from Cartesian to polar coordinates.

1.1 Substitution

Two kinds of substitutions

There are two kinds of integral substitutions. As an example of the first kind, consider the familiar integral

$$\int_0^b \frac{dx}{1+x^2}.$$

We know that the substitution $x = \tan s$ is helpful here because $1 + x^2 = 1 + \tan^2 s = \sec^2 s$ and $dx = \sec^2 s\, ds$. Therefore,

$$\int \frac{dx}{1+x^2} = \int \frac{\sec^2 s\, ds}{1+\tan^2 s} = \int ds = s = \arctan x,$$

and we then have

$$\int_0^b \frac{dx}{1+x^2} = \arctan x\Big|_0^b = \arctan b.$$

As an example of the second kind of substitution, take the apparently similar integral

$$\int_0^b \frac{x\,dx}{(1+x^2)^p}, \quad p \neq 1.$$

J.J. Callahan, *Advanced Calculus: A Geometric View*, Undergraduate Texts in Mathematics,
DOI 10.1007/978-1-4419-7332-0_1, © Springer Science+Business Media, LLC 2010

The factor x in the numerator suggests the substitution $u = 1 + x^2$. Then $du = 2x\,dx$ and

$$\int \frac{x\,dx}{(1+x^2)^p} = \frac{1}{2}\int \frac{du}{u^p} = \frac{1}{2}\frac{u^{-p+1}}{(-p+1)} = \frac{-1}{2(p-1)(1+x^2)^{p-1}}.$$

Thus,

$$\int_0^b \frac{x\,dx}{(1+x^2)^p} = \frac{1}{2(p-1)}\left(1 - \frac{1}{(1+b^2)^{p-1}}\right).$$

Integral as antiderivative

In these examples, integration is done with the fundamental theorem of calculus. That is, we use the fact that the indefinite integral of a given function f,

$$F = \int f(x)\,dx,$$

is an *antiderivative* of f: $F'(x) = f(x)$. However, the substitutions we used to find the two antiderivatives are different in important ways.

Pullback

We call the first an example of a **pullback** substitution, for reasons that become clear in a moment. In a pullback, we express the variable x itself as some differentiable function $x = \varphi(s)$ of a new variable s. Then $dx = \varphi'(s)\,ds$ and we get

$$\int f(x)\,dx = \int f(\varphi(s))\,\varphi'(s)\,ds = \Phi,$$

where $\Phi(s)$ is an antiderivative of $f(\varphi(s))\,\varphi'(s)$. Here the aim is to choose the function φ so the antiderivative Φ becomes evident. The indefinite integral we want is then $F(x) = \Phi(\varphi^{-1}(x))$, where $s = \varphi^{-1}(x)$ is the *inverse* of the function $x = \varphi(s)$. (In our example, φ is the tangent function and φ^{-1} is the arctangent function; $\Phi(s)$ is just s.) We also use φ^{-1} to get the upper and lower endpoints of the definite integral:

$$\int_a^b f(x)\,dx = \int_{\varphi^{-1}(a)}^{\varphi^{-1}(b)} f(\varphi(s))\,\varphi'(s)\,ds.$$

Push-forward

In the second example, a **push-forward** substitution, we replace some functional expression $g(x)$ involving x with a new variable u. As with $\varphi(s)$, it takes practice and experience to make an effective choice of $g(x)$: the aim is to be able to write

$$f(x) = G'(g(x)) \cdot g'(x) \text{ or } f(x)\,dx = G'(u)\,du$$

for a suitable function $G'(u)$. That is, $du = g'(x)\,dx$ and

$$\int f(x)\,dx = \int G'(g(x))\,g'(x)\,dx = \int G'(u)\,du = G,$$

and the antiderivative is $F(x) = G(g(x))$. In our example,

$$G'(u) = \frac{1}{2u^p} \text{ and } G(u) = \frac{-1}{2(p-1)u^{p-1}}.$$

Note that we use g (and not g^{-1}) to get the endpoints of the transformed definite integral:

$$\int_a^b f(x)\,dx = \int_{g(a)}^{g(b)} G'(u)\,du.$$

Why *pull back* and *push forward*?

To see how the substitutions using φ and g are different, and also to see how they got their names, let us think of them as maps:

$$s \xrightarrow{\varphi} x \xrightarrow{g} u$$

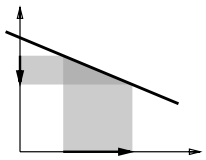
Then we can say g pushes forward information about the value of x to the variable u, and φ pulls back that information to s. Note that the pullback needs to be invertible: without a well-defined φ^{-1}, a given value of x may pull back to two or more different values of s or to none at all. This problem does not arise with g, though.

Transformation of differentials

To complete this section, let us review why the differential changes the way it does in a substitution. For example, in the pullback $x = \varphi(s)$, why is $dx = \varphi'(s)\,ds$? The answer might seem obvious: because dx/ds is just another notation for the derivative—that is, $dx/ds = \varphi'(s)$—we simply multiply by ds to get $dx = \varphi'(s)\,ds$. This is a good mnemonic; however, it is not an explanation, because the expressions dx and ds have no independent meaning, at least as far as derivatives are concerned. We must look more carefully at the link between differentials and derivatives.

Slope as length multiplier

In a linear function, $x = \varphi(s) = ms + b$, we usually interpret the coefficient m as the slope of the graph: $\Delta x/\Delta s = m$. However, if we rewrite the slope equation in the form $\Delta x = m\Delta s$, it becomes natural to interpret m instead as a *multiplier*. That is, our linear map $\varphi : s \mapsto x$ multiplies lengths by the factor m: an interval of length Δs on the s-axis is mapped to an interval of length $\Delta x = m\Delta s$ on the x-axis. Furthermore, when $m < 0$, Δs and Δx have opposite orientations, so φ also carries out a "flip." (The role of the coefficient as a multiplier rather than as a slope suggests why it is so commonly represented by the letter "m".)



When $x = \varphi(s)$ is nonlinear, the slope of the graph (or the slope of its tangent line) varies from point to point. Nevertheless, by fixing our attention on a small neighborhood of a particular point $s = s_0$, we still have a way to interpret the derivative as a multiplier. To see how this happens, recall first that we assume φ is differentiable, so

$$\varphi'(s_0) = \lim_{\Delta s \to 0} \frac{\Delta x}{\Delta s} = \lim_{s \to s_0} \frac{\varphi(s) - \varphi(s_0)}{s - s_0}.$$

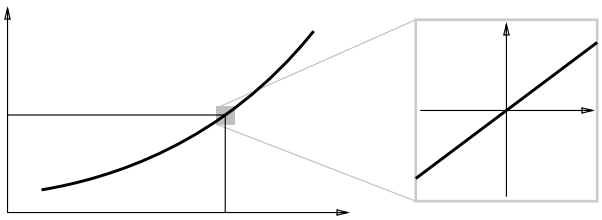
The microscope equation and linear approximations

According to the meaning of a limit, we can make $\Delta x/\Delta s$ as close to $\varphi'(s_0)$ as we wish by making $\Delta s = s - s_0$ sufficiently small; in other words,

$$\Delta x \approx \varphi'(s_0)\,\Delta s \text{ when } \Delta s \approx 0.$$

To see what this means, focus a microscope at the point (s_0, x_0) and use coordinates $\Delta s = s - s_0$ and $\Delta x = x - x_0$ centered in this window. Then, under sufficient magnification (i.e., with $\Delta s \approx 0$), φ looks like $\Delta x \approx \varphi'(s_0)\,\Delta s$. We call this the **microscope**

equation for $x = \varphi(s)$ at s_0; it is linear, and defines the **linear approximation** of the function φ at s_0.



φ' is the local length multiplier

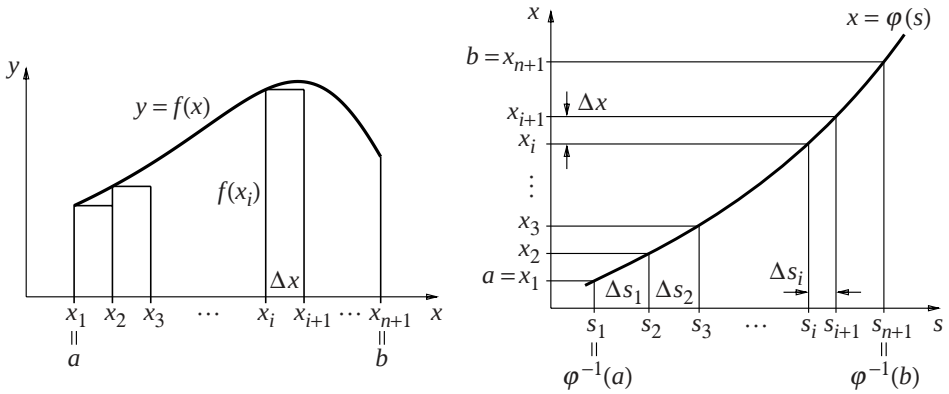
Finally, we can say that φ is *locally linear*, in the sense that $x = \varphi(s)$ comes as close as we wish to its linear approximation $\Delta x \approx \varphi'(s_0)\Delta s$ when s is restricted to a sufficiently small neighborhood of s_0. Thus, because the map $\varphi : s \to x$ is locally linear at s_0, it multiplies lengths (approximately) by $\varphi'(s_0)$ in any sufficiently small neighborhood of $s = s_0$.

Integral as a limit of Riemann sums

With the microscope equation, we can now see why the differential transforms the way it does when we make a change of variables in an integral. First of all, a definite integral is defined as a limit of Riemann sums. In the simplest case—a left-endpoint Riemann sum with n equal subintervals—we can set $\Delta x = (b - a)/n$ and $x_i = a + (i - 1)\Delta x$ and write

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)\,\Delta x.$$

We think of each term in the sum as the area of a rectangle with height $f(x_i)$ and base Δx, as in the figure at the left, below.



The pullback creates a new Riemann sum

The figure at the right shows how the substitution $x = \varphi(s)$ pulls back our partition of the interval $a \le x \le b$ to a partition of $\varphi^{-1}(a) \le s \le \varphi^{-1}(b)$. We set $s_i = \varphi^{-1}(x_i)$ $(i = 1, \ldots, n + 1)$ and $\Delta s_i = s_{i+1} - s_i$ $(i = 1, \ldots, n)$. Note that the subintervals Δs_i are generally unequal when φ is nonlinear. In fact, $\Delta s_i \approx \Delta x/\varphi'(s_i)$, by

the microscope equation. The pullback allows us to write

$$\sum_{i=1}^{n} f(x_i)\,\Delta x \approx \sum_{i=1}^{n} f(\varphi(s_i))\,\varphi'(s_i)\,\Delta s_i.$$

By choosing n sufficiently large, we can make every Δs_i arbitrarily small and thus can make these two sums arbitrarily close. Notice that the right-hand side is also a Riemann sum, in this case for the function $f(\varphi(s))\,\varphi'(s)$. Therefore, in the limit as $n \to \infty$, the Riemann sums become integrals and we have the equality

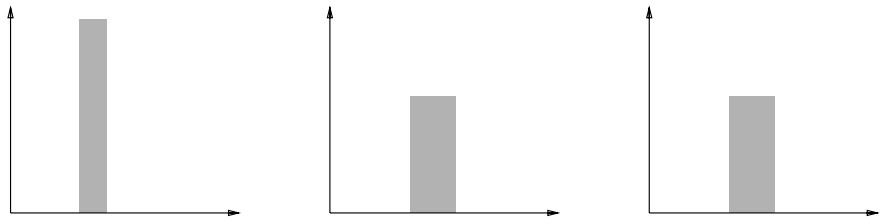
$$\int_a^b f(x)\,dx = \int_{\varphi^{-1}(a)}^{\varphi^{-1}(b)} f(\varphi(s))\,\varphi'(s)\,ds.$$

Thus we see that the justification for the transformation $dx = \varphi'(s)\,ds$ of differentials in integration lies in the transformation $\Delta x \approx \varphi'(s_i)\,\Delta s_i$ that the microscope equation provides for the Riemann sums.

$dx = \varphi'(s)\,ds$

The microscope equation $\Delta x \approx \varphi'(s_i)\,\Delta s_i$ has one further geometric consequence. In our Riemann sum for the second integral, the standard way to think about each term is as the area of a rectangle with height $f(\varphi(s_i))\,\varphi'(s_i)$ and base Δs_i. However, if we change the proportions and make the height $f(\varphi(s_i))$ and the base $\varphi'(s_i)\,\Delta s_i$, then we have a rectangle that matches (as closely as we wish) the shape of the original rectangle with height $f(x_i)$ and base Δx, because $f(x_i) = f(\varphi(s_i))$ and $\Delta x \approx \varphi'(s_i)\,\Delta s_i$.

Rectangles in the Riemann sums



To understand why differentials transform the way they do, we worked with a pullback substitution. We get the same result with a push-forward, though the details are different. Our work has led us to several questions that we ask again when we turn to more general integrals that involve functions of several variables: what different kinds of substitutions occur? What role do inverses play? What is the form of a linear approximation? What is the analogue of the local length multiplier? What are differentials and how do they transform? What is the geometric interpretation of that transformation?

Some questions raised

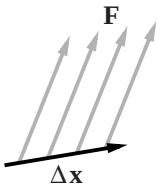
1.2 Work and path integrals

Path integrals are one of the centerpieces of the first multivariable calculus course, and they are often treated, as we do here, in the context of work.

Force, displacement, and work

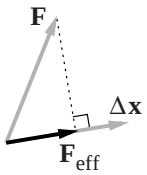
By definition, a force moving a body from one place to another produces work, and the work done is proportional to both the force applied and to the displacement caused. The simplest formula that captures this idea is

$$\text{work} = \text{force} \times \text{displacement}.$$



Although work is a scalar quantity, force and displacement are actually both vectors, and the force is a *field*, that is, a variable function of position. We must elaborate our simple formula to reflect these facts. Consider a straight-line displacement along some vector $\Delta\mathbf{x}$ and a constant force field $\mathbf{F}$ that acts the same way at every point along $\Delta\mathbf{x}$. Only the component of the force that lies in the direction of the displacement does any work; this is the *effective* force $\mathbf{F}_{\text{eff}}$. We can take all this into account in the new formula

$$\text{work} = \|\mathbf{F}_{\text{eff}}\|\,\|\Delta\mathbf{x}\|.$$



The scalar $\|\mathbf{F}_{\text{eff}}\|$ is the length of the perpendicular projection of $\mathbf{F}$ on $\Delta\mathbf{x}$. Now, in general, for arbitrary vectors $\mathbf{A}$ and $\mathbf{B} \neq \mathbf{0}$,

$$\text{length of projection of } \mathbf{A} \text{ onto } \mathbf{B} = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{B}\|}.$$

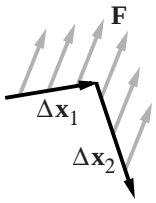
work = $\mathbf{F}\cdot\Delta\mathbf{x}$

Rewriting the length $\|\mathbf{F}_{\text{eff}}\|$ this way, we see work is still a product; it is the *dot* (or *scalar*) product of force $\mathbf{F}$ and displacement $\Delta\mathbf{x}$, now regarded as *vectors*:

$$\text{work} = W = \frac{\mathbf{F}\cdot\Delta\mathbf{x}}{\|\Delta\mathbf{x}\|}\,\|\Delta\mathbf{x}\| = \mathbf{F}\cdot\Delta\mathbf{x}.$$

Work is additive

In our new formula, W can take negative values (e.g., if $\mathbf{F}$ makes an obtuse angle with $\Delta\mathbf{x}$). To see why "negative work" must arise, consider a constant force $\mathbf{F}$ that displaces an object along a path consisting of two straight segments $\Delta\mathbf{x}_1$ and $\Delta\mathbf{x}_2$, one immediately followed by the other. We want the total work to be the sum of the work done on the separate segments:



$$\text{total work} = \mathbf{F}\cdot\Delta\mathbf{x}_1 + \mathbf{F}\cdot\Delta\mathbf{x}_2.$$

Orientation matters

We say that work is additive on displacements. In particular, if $\Delta\mathbf{x}_2 = -\Delta\mathbf{x}_1$, then the total work done is 0. Consequently, the work done by $\mathbf{F}$ along $-\Delta\mathbf{x}$ must be the negative of the work done by the same $\mathbf{F}$ along $+\Delta\mathbf{x}$. Orientation matters: reversing the displacement reverses the sign of the work done.

Components of work: $W = P\Delta x + Q\Delta y$

Let us introduce coordinates into the plane containing the vectors $\mathbf{F}$ and $\Delta\mathbf{x}$ and write $\mathbf{F} = (P, Q)$ and $\Delta\mathbf{x} = (\Delta x, \Delta y)$. Then

$$W = \mathbf{F} \cdot \Delta\mathbf{x} = P\Delta x + Q\Delta y.$$

This formula gives the **coordinate components** of work. It says that, in the x-direction, there is a force of size P acting along a displacement of size Δx, doing work $W_x = P\Delta x$. Similarly, in the y-direction the work done is $W_y = Q\Delta y$. We call W_x and W_y the *components of W in the x- and y-directions*. The following definition summarizes our observations to this point.

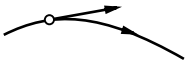
Definition 1.1 *The **work** done by the constant force $\mathbf{F} = (P, Q)$ in displacing an object along the line segment $\Delta\mathbf{x} = (\Delta x, \Delta y)$ is*

$$W = \mathbf{F} \cdot \Delta\mathbf{x} = P\Delta x + Q\Delta y = W_x + W_y.$$

Displacement along a curved path

Ultimately, we need to deal with variable forces and displacements along curved paths. The prototype is a smooth simple curve C in the plane. We say C is **smooth** if it is the image of a map (an example of a **vector-valued function**)

$$\mathbf{x} : [a, b] \to \mathbb{R}^2 : t \mapsto (x(t), y(t))$$



(a **parametrization**) whose coordinate functions $x(t)$ and $y(t)$ have continuous derivatives on $a \le t \le b$. We call t the **parameter**. In addition, C is **simple** if it has no self-intersections, that is, if $\mathbf{x}$ is 1–1. The parametrization orders the points on C in the following sense: $\mathbf{x}(t_1)$ *precedes* $\mathbf{x}(t_2)$ if $t_1 < t_2$ (i.e., t_1 precedes t_2 in $[a, b]$). The ordering gives C an **orientation**; we write $\vec{C}$ to indicate C is oriented. At any point on $\vec{C}$ where the tangent vector $\mathbf{x}'(t)$ is nonzero, it points in the direction of increasing t, and thus also indicates the orientation of $\vec{C}$. We can immediately extend these ideas to paths in $\mathbb{R}^n$.

Parametrizing a smooth simple curve

Definition 1.2 *A **smooth, simple, oriented curve** $\vec{C}$ in $\mathbb{R}^n$ is the image of a smooth 1–1 map,*

$$\mathbf{x} : [a, b] \to \mathbb{R}^n : t \mapsto \mathbf{x}(t),$$

*where $\mathbf{x}'(t) \neq 0$ for all $a < t < b$. The point $\mathbf{x}(a)$ is the **start** of $\vec{C}$ and $\mathbf{x}(b)$ is its **end**.*

Linear displacements as oriented curves

The simple formula $W = \mathbf{F} \cdot \Delta\mathbf{x}$ for work assumes that the force $\mathbf{F}$ is constant, so the location of the base point $\mathbf{a}$ of the displacement $\Delta\mathbf{x}$ is irrelevant. However, if $\mathbf{F}$ varies, then the work done will depend on $\mathbf{a}$. We must, in fact, treat a linear displacement as we would any displacement, and provide it with a parametrization. A natural one is

$$\mathbf{x}(t) = \mathbf{a} + t \cdot \Delta\mathbf{x}, \quad 0 \le t \le 1.$$

Work done by a variable force

We are now in a position to estimate the work done by a *variable* force as it displaces an object along a smooth, simple, oriented curve $\vec{C}$ in $\mathbb{R}^3$. Force is now a (continuous) **vector field**—that is, a vector-valued function $\mathbf{F}(\mathbf{x})$ that varies (continuously) with position $\mathbf{x}$. To estimate the work done, chop the curve into small pieces. When a piece is small enough, it is essentially straight and the force is essentially constant along it. On this piece, the linear formula for work (Definition 1.1) gives a good approximation. By additivity, the sum of these contributions will approximate