Basic Ideas and Applications of Smooth Infinitesimal Analysis

John L. Bell

In Smooth Infinitesimal Analysis (**SIA**) we are given a subset Δ of the set $\mathbf{R}$ of real numbers called the *domain of infinitesimals.* Δ is assumed to satisfy the following conditions:

(1) $0 \in \Delta$

(2) *Any multiple of an infinitesimal is an infinitesimal,* that is, for any $a \in \mathbf{R}$**,** $x \in \Delta \Rightarrow ax \in \Delta$. It follows immediately from this that Δ is *symmetric around* 0, that is, $\forall x\ (x \in \Delta \Leftrightarrow -x \in \Delta)$.

(3) The ***Affineness Principle*** or the ***Kock- Lawvere axiom.*** This asserts that all maps $\Delta \to \mathbf{R}$ are *affine* in the following strong sense:

for any $f : \Delta \to \mathbf{R}$ there is a *unique* $a \in \mathbf{R}$ such that, for all $\varepsilon \in \Delta$, $f(\varepsilon) = f(0) + \varepsilon a$.[1] This real number a is called the *slope* of f, and is written slp(f). Thus, for any $f : \Delta \to \mathbf{R}$ and all $\varepsilon \in \Delta$, $f(\varepsilon) = f(0) + \varepsilon\ \mathrm{slp}(f)$.

Given $f : \Delta \to \mathbf{R}$, let $f^* : \Delta \to \mathbf{R} \times \mathbf{R}$ defined by $f^*(\varepsilon) = (\varepsilon, f(\varepsilon))$. f^* may be thought of as the graph of the map f n the plane. The Affineness Principle can be seen as asserting that this graph is a straight line with slope slp(f) passing through $(0, f(0))$. Thus the effect of any map on Δ is to translate and rotate it; in effect Δ behaves like a short "rigid rod", just long enough to have a slope, but too short to bend. It is, as it were, a geometric object possessing location and direction but lacking extension.

The Affineness Principle has two immediate consequences:

[1] We shall use symbols ε, η to denote arbitrary elements ofΔ.

- **The nondegeneracy of Δ :** $\Delta \neq \{0\}$. For suppose $\Delta = \{0,$ and let *a, b* be any unequal real numbers. Then the two maps $\varepsilon \mapsto \varepsilon a$ and $\varepsilon \mapsto \varepsilon b$ would both be identically 0, contradicting the Affineness Principle.
- **The Cancellation Principle:** for any $a, b \in \mathbf{R}$, if $\varepsilon a = \varepsilon b$ for all $\varepsilon \in \Delta$, then $a = b$. This follows immediately from the uniqueness condition in the Affineness Principle.

We can now show that Δ consists of *nilsquare* quantities, that is,

$$\Delta \subseteq \{x \in \mathbf{R} \colon x^2 = 0\}.$$

To prove this, write $s \colon \Delta \to \mathbf{R}$ for the map $\varepsilon \mapsto \varepsilon^2$. Then we have, for $\varepsilon \in \Delta$,

$$\varepsilon^2 = s(\varepsilon) = s(0) + \varepsilon\ \mathrm{slp}(s) = \varepsilon\ \mathrm{slp}(s)$$

and, since $-\varepsilon \in \Delta$,

$$\varepsilon^2 = (-\varepsilon)^2 = s(-\varepsilon) = s(0) - \varepsilon\ \mathrm{slp}(s) = -\varepsilon\ \mathrm{slp}(s).$$

Hence $\varepsilon^2 = -\varepsilon^2$, so that $\varepsilon^2 = 0$.

Notice that we do not claim that Δ comprises *all* nilsquare quantities. This, although usually assumed in presentations of **SIA**, is not needed.

Here is a further fact that may be of interest. Without assuming the symmetry of Δ around 0, the following conditions are equivalent:

(i) $\Delta \subseteq \{x \in \mathbf{R} \colon x^2 = 0\}$

(ii) $\mathrm{slp}(s) = 0$.

(i) ⇒ (ii). Assuming **(i),** we have, for all $\varepsilon \in \Delta$, $0 = \varepsilon^2 = s(\varepsilon) = s(0) + \varepsilon\ \mathrm{slp}(s) = \quad \varepsilon.\ \mathrm{slp}(s)$. The Cancellation Principle yields $\mathrm{slp}(s) = 0$.

(ii) ⇒ (i) Assuming **(ii),** we have, for all $\varepsilon \in \Delta$, $\varepsilon^2 = s(\varepsilon) = s(0) + \varepsilon\ \mathrm{slp}(s) = 0 + 0 = 0$. Hence **(i).**

Let us call a *real function* any real-valued function defined on an interval in to **R.** The derivative of an arbitrary real function can now be introduced. Given an interval[2] **I** in **R** and a function $f: \mathbf{I} \to \mathbf{R}$, for each $x \in \mathbf{I}$ define the function $f_x: \Delta \to \mathbf{R}$ by $f_x(\varepsilon) = f(x + \varepsilon)$. The *derivative* $f': \mathbf{I} \to \mathbf{R}$ of f is defined by $f'(x) = \text{slp}\ (f_x)$. It follows easily that

$$f(x + \varepsilon) = f(x) + \varepsilon f'(x).$$

This is the *Fundamental Equation of the Differential Calculus* in **SIA**. The quantity $f'(x)$ is the slope at x of the curve determined by f and the infinitesimal

$$\varepsilon f'(x) = f(x + \varepsilon) - f(x)$$

is the infinitesimal change or *increment* in the value of f on passing from x to $x + \varepsilon$.

Derivatives of elementary functions are easily calculated in **SIA** using the Cancellation Principle. For example, here is the calculation of the derivative of the function x^n:

$$\varepsilon(x^n)' = (x + \varepsilon)^n - x^n = \varepsilon n\ x^{n-1} + \text{terms in } \varepsilon^{2\text{-}} \text{ and higher powers} = \varepsilon n\ x^{n-1}$$

Hence, by the Cancellation Principle,

$$(x^n)' = n\ x^{n-1}.$$

And here is the calculation of the derivative of the function $1/x$ (for $x > 0$):

$$\varepsilon(1/x)' = 1/x{+}\varepsilon - 1/x = -\varepsilon/x(x{+}\varepsilon) = -\varepsilon(x{-}\varepsilon)/x(x{+}\varepsilon)(x{-}\varepsilon)$$

$$= -\varepsilon x{+}\varepsilon^2/x(x^2{-}\varepsilon^2)$$

$$= -\varepsilon x/x^3$$

$$= -\varepsilon/x^2.$$

Cancelling ε on both sides of the equation gives

$$(1/x)' = -1/x^2.$$

[2] We assume that intervals are closed under the addition of infinitesimals.

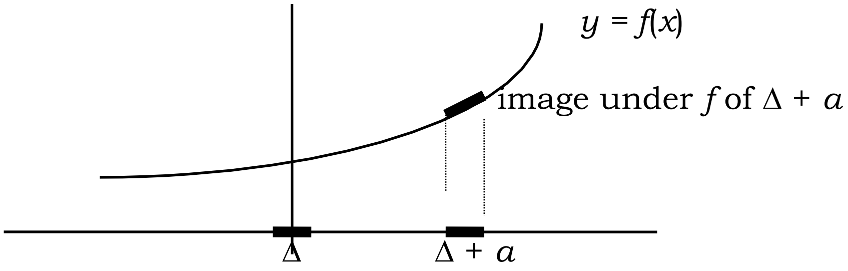
From the Fundamental Equation a version of the *Principle of Contintity* can be deduced, namely, that in **SIA** *all real functions are continuous*, in the sense of *sending neighbouring points to neighbouring points.* (Here two points x, y on **R** are said to be *neighbours* if $x - y$ is in **Δ**, that is, if x and y differ by an infinitesimal.) To see this, given a real function f and neighbouring points x, y, note that $y = x + \varepsilon$ with ε in **Δ** , so that

$$f(y) - f(x) = f(x + \varepsilon) - f(x) = \varepsilon f'(x).$$

Since $\varepsilon f'(x)$ is infinitesimal, the result follows.

From this we see that in **SIA** *every real function is differentiable, hence continuous.* In particular the derivative of an arbitrary real function is itself differentiable, so that any real function is arbitrarily many times differentiable. This fact justifies the use of the term "smooth".

If we think of a real function f as defining a curve, then, for any a, the image under f of the "infinitesimal interval" Δ + a obtained by translating Δ to a is straight and coincides with the tangent to the curve at $x = a$. Thus each real function has the effect of "bringing" Δ into coincidence" with the tangent vector to the curve associated with the function at any point on it. In this sense, then, Δ plays the role of a *generic tangent vector.* Also, since the image of Δ under a function is necessarily a straight line and a part of the associated curve, it follows that each point on a curve is contained in a nondegenerate infinitesimal



straight segment of the curve. In other words, in **SIA** *curves are infinitesimally straight.* This is the **Principle of Infinitesimal Straightness.**

In **SIA** there is a sense in which *everything is generated by the domain of infinitesimals.* For consider the set Δ^Δ of all maps $\Delta \rightarrow \Delta$. It follows from the Affineness Principle that **R** can be identified as the subset of Δ^Δ consisting of all maps vanishing at 0. In this sense **R** is "generated" by Δ [3]. Once one has **R**, Euclidean spaces of all dimensions may be obtained as powers of **R**, and arbitrary Riemannian manifolds may be obtained by patching together subspaces of these.

We observe that the postulates of **SIA** are *incompatible with the Law of Excluded Middle of classical logic* (LEM)– the assertion that, for any proposition *p,* either *p* holds or not *p* holds. This incompatibility can be demonstrated in two ways, one informal and the other rigorous. First the informal argument. Consider the function *f* defined for real numbers *x* by $f(x) = 1$ if $x = 0$ and $f(x) = 0$ whenever $x \neq 0$. If LEM held, each real number would then be either equal or unequal to 0, so that the function *f* would be defined on the whole of **R.** But, considered as a function with domain **R,** *f* is clearly discontinuous. Since, as we know, in **SIA** every function on **R** is continuous, *f* cannot have domain **R** there[4]. So LEM fails in **SIA**. To put it succinctly, *universal continuity implies the failure of the Law of Excluded Middle.*

Here now is the rigorous argument. We derive the failure of LEM from the Cancellation Principle. To begin with, if $x \neq 0$, then $x^2 \neq 0$, so that, if $x^2 = 0$, then necessarily not $x \neq 0$.

[3] Explicitly, Δ^Δ is a monoid under composition which may be regarded as acting on Δ by composition: for $f \in \Delta^\Delta$, $f \cdot \varepsilon = f(\varepsilon)$. The subset *V* consisting of all maps vanishing at 0 is a submonoid naturally identified as the set of *ratios of infinitesimals.* The identification of **R** and *V* made possible by the principle of infinitesimal affineness thus leads to the characterization of **R** itself as the set of ratios of infinitesimals. This was essentially the view of Euler, who regarded infinitesimals as formal zeros and real numbers as representing the possible values of 0/0. For this reason Lawvere has suggested that **R** in **SIA** should be called the space of *Euler reals.*

[4] The domain of *f* is in fact $(\mathbf{R} - \{0\}) \cup \{0\}$, which, because of the failure of the law of excluded middle in SIA, is provably unequal to **R.**

This means that

$$\textit{for all infinitesimal } \varepsilon, \textit{ not } \varepsilon \neq 0. \qquad (*)$$

Now suppose that LEM were to hold. Then we would have, for any ε, either ε = 0 or ε ≠ 0. But (*) allows us to eliminate the second alternative, and we infer that, for all ε, ε = 0. This may be written

$$\textit{for all } \varepsilon,\ \ \varepsilon.1 = \varepsilon.0,$$

from which we derive by Cancellation Principle the falsehood 1 = 0. So again LEM must fail.

The "internal" logic of **SIA** is accordingly not full classical logic. It is, instead, *intuitionistic* logic, that is, the logic derived from the constructive interpretation of mathematical assertions. In practice when working in **SIA** one does not notice this "change of logic" because, like much of elementary mathematics, the topics discussed there are naturally treated by constructive means such as direct computation.

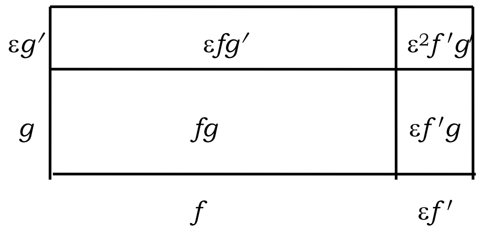
To illustrate this point, let us derive in **SIA** a basic law of the differential calculus, the *product rule:*

$$(fg)' = f'g + fg'.$$

To do this we compute

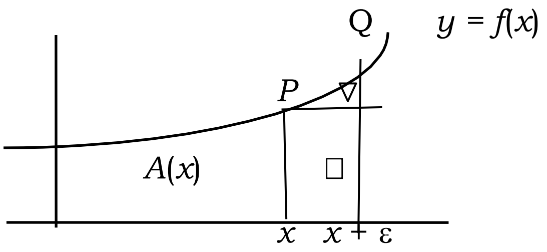
$$\begin{aligned}
(fg)(x + \varepsilon) &= (fg)(x) + (fg)'(x) = f(x)g(x) + (fg)'(x),\\
(fg)(x + \varepsilon) &= f(x + \varepsilon)g(x + \varepsilon) = [f(x) + f'(x)].[g(x) + g'(x)]\\
&= f(x)g(x) + \varepsilon(f'g + fg') + \varepsilon^2 f'g'\\
&= f(x)g(x) + \varepsilon(f'g + fg'),
\end{aligned}$$

since $\varepsilon^2 = 0$. Therefore $\varepsilon(fg)' = \varepsilon(f\,'g + fg')$, and the result follows by the Cancellation Principle. This calculation is depicted in the diagram below.



Infinitesimals in **SIA** have two fundamental aspects, *algebraic* and *geometric.* Algebraically, they are real numbers whose squares vanish; geometrically, they are straight segments of curves. Both of these aspects are used in applications.

The interplay of these aspects is well illustrated by the derivation in **SIA** of the *fundamental theorem of the calculus.* To this end, let **I** be a closed interval $\{x\colon a \leq x \leq b\}$ in **R** — or **R** itself—and let $f\colon \mathbf{I} \to \mathbf{R}$; also let $A(x)$ be the area under the curve $y = f(x)$ as indicated in the figure:



Then

$$\varepsilon A'(x) = A(x + \varepsilon\ ) - A(x) = \square + \triangledown\ = \varepsilon f(x) + \triangledown.$$

Now, by the Principle of Infinitesimal Straightness, the arc PQ is a straight line; accordingly $\triangledown$ is a triangle of area $\frac{1}{2}\,\varepsilon.\varepsilon\, f\,'(x) = 0$. It follows that $\varepsilon A'(x) = \varepsilon f(x)$, and the Cancellation Principle gives

$$A'\,(x) = f(x).$$

Thus, if we regard $A(x)$ as the integral of $f(x)$, the above equation asserts that differentiation is the inverse of integration – the fundamental theorem of the calculus.

This derivation, which is typical of such arguments in **SIA,** displays the following pattern. Suppose we are investigating the behavior of some variable quantity $F(x)$ (in the above derivation F is A). The approach taken in **SIA**, as in the differential calculus, is to begin the investigation by confining it initially to the infinitesimal world. Life in the infinitesimal world is beautifully simple: curves are just straight lines, and the squares of incremental changes vanish. This makes the determination of infinitesimal increments equally simple, enabling the increment $\varepsilon F'(x)$ in $F(x)$ to be presented in the form $\varepsilon k(x)$, where $k(x)$ is some explicit function whose form has been obtained by "infinitesimal" analysis. Thus we obtain an "infinitesimal" equation of the form $\varepsilon F'(x) = \varepsilon k(x)$. Applying the Cancellation Principle in turn yields the "differential" equation

$$(*) \qquad F'(x) = k(x)$$

which holds in the world "in the large".

The Cancellation Principle thus provides a formal, astonishingly simple link between the infinitesimal world and the world "in the large". The idea of a linkage between the two worlds exists was the animating principle behind applications of the calculus throughout the 17th and 18th centuries.

In practice, of course, the equation (*), while of fundamental importance, is only the first step in determining the explicit form of the function F. For this, it is necessary to "integrate" k, that is, to provide k with an *antiderivative,* an explicit function G such that $G' = k$. It will then follow that $F' = G'$, from which we will be able to conclude that $F = G$[5].

[5] Strictly speaking, F and G may differ by a constant function but we shall ignore this.

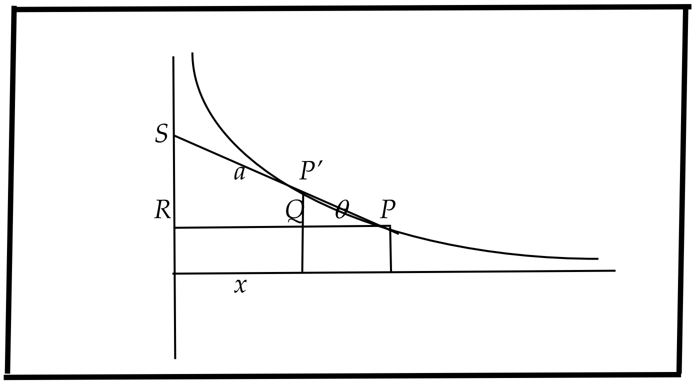
To carry out this procedure in **SIA** we need to introduce an additional postulate. Let us define a *stationary point* of a function f: $\mathbf{I} \to \mathbf{R}$ is defined to be a point $a \in \mathbf{I}$ in whose vicinity "infinitesimal changes in the value of the argument fail to change the value of f", that is, for which $f(a + \varepsilon) = f(a)$ for all ε. This means that $f(a) + \varepsilon f'(a) = f(a)$, so that $\varepsilon f'(a) = 0$ for all ε, from which it follows by the Cancellation Principle that $f'(a) = 0$. Thus a stationary point of a function is precisely a point at which the derivative of the function vanishes.

In classical analysis, if the derivative of a function is identically zero, the function is constant. This fact is the source of the following postulate concerning stationary points adopted in **SIA**:

Constancy Principle. If every point in an interval J is a stationary point of f: $J \to \mathbf{R}$ (that is, if $f'$is identically 0), then f is constant.

It follows from the Constancy Principle that two functions with identical derivatives differ by at most a constant. Using this, we can formulate simple derivations of basic equations of mathematics and physics. We illustrate the method by deriving *the equation of a tractrix*.

The tractrix has the property that the length of the tangent to the curve from an arbitrary point on it to the y-axis is constant. It is the curve traced out by an object dragged, under the influence of friction, by a string attached to a pulling point that moves at a right angle to the initial line between the object and the puller.



Referring to the figure above, let P, P' be two neighbouring points on the tractrix curve $y = y(x)$ with coordinates (x, y) and $(x - \varepsilon, y(x - \varepsilon))$ respectively. Let a be the constant length of the tangent from a point of the curve to the y-axis. Then by the Principle of Infinitesimal Straightness the tangent PS to the curve passes through P'. Write $\underline{L}$ for the length of a line segment L. Then we have

(*) $$\underline{P'Q} = y(x - \varepsilon) - y(x) = -\varepsilon y'(\mathrm{x}).$$

But also, writing θ for the angle RPS, we have

$$\underline{P'Q} = \underline{PQ}\tan\theta = \varepsilon\tan\theta = \varepsilon\underline{RS}/\underline{PR} = \varepsilon(a^2 - x^2)^{1/2}/x.$$

Equating this with (*) gives

$$\varepsilon y'(\mathrm{x}) = -\varepsilon(a^2 - x^2)^{1/2}/x.$$

Using the Cancellation Principle to cancel ε on both sides of this equation, we get

(**) $$y'(\mathrm{x}) = -(a^2 - x^2)^{1/2}/x.$$

Accordingly $y(x)$ is the antiderivative of the function $-(a^2 - x^2)^{1/2}/x$. But the usual computation in the differential calculus (which can be carried out in **SIA**) shows that antiderivative to be the function $a\log[(a + (a^2 - x^2)^{1/2})/x)] - (a^2 - x^2)^{1/2}$. Since, by the Constancy Principle, antiderivatives are unique up to a constant (which we shall set to 0), it follows that

$$y = a\log[(a + (a^2 - x^2)^{1/2})/x)] - (a^2 - x^2)^{1/2}$$

This is the equation of the tractrix.

Put succinctly, the Constancy Principle asserts that "universal infinitesimal (or "local") constancy implies global constancy", or "infinitesimal behaviour determines global behaviour" The Constancy Principle brings into sharp focus the difference in **SIA** between points and infinitesimals. For if in the Constancy Principle one replaces "infinitesimal constancy" by "constancy at a point" the resulting "Principle" is false because *any function whatsoever* is constant at every point. But since in **SIA** all functions on **R** are smooth, the Constancy Principle embodies the idea that for such functions local constancy is sufficient for global constancy, that a nonconstant smooth function must be somewhere nonconstant over arbitrarily small intervals.

The Constancy Principle provides another bridge between the infinitesimal world and the world "in the large", the lack of which Hermann Weyl believed doomed the idea of infinitesimal, and led to its inevitable replacement by the limit concept. In his *Philosophy of Mathematics and Natural Science* he saya:

> [In its struggle with the infinitely small] *the limiting process was victorious. For the* limit *is an indispensable concept, whose importance is not affected by the acceptance or rejection of the infinitely small. But once the limit concept has been grasped, it is seen to render the infinitely small superfluous. Infinitesimal analysis proposes to draw conclusions by integration from the behavior in the infinitely small, which is governed by elementary laws, to the behavior in the large; for instance, from the universal law of attraction for two material "volume elements" to the magnitude of attraction between two arbitrarily shaped bodies with homogeneous or non-homogeneous mass distribution. If the infinitely small is not interpreted 'potentially' here, in the sense of the limiting process, then the one has nothing to do with the other, the process in infinitesimal and finite dimensions become independent of each other, the tie which binds them together is cut.*

In **SIA** the Constancy Principle reconnects the infinitesimal and the extended. Behaviour "in the large" is completely determined by behaviour "in the infinitely small".