Math S166 Inverse Functions Background (final version)

Let f be a function. In the discussion below, I and J represent sets. In our case, these are generally sets of real numbers and, very often, intervals of real numbers but also frequently, the sets are unions of intervals of real numbers.

Definition We say "f maps I to J" or "f maps I into J" and write $f : I \to J$ if $f(x)$ is defined for each x in I and, for all such x, the value $f(x)$ is in J.

Most functions we study in calculus are real valued functions of a real variable, that is, usually $I \subset \mathbb{R}$ and $J \subset \mathbb{R}$, so we can usually write $f : I \to \mathbb{R}$. If we want to emphasize $J \neq \mathbb{R}$, then we might write, instead, $f : I \to J$ where is J explicitly specified. For example, it is correct to say for $s(x) = x^2$, the squaring function, that $s : \mathbb{R} \to \mathbb{R}$, but we might also write $s : \mathbb{R} \to [0, \infty)$ and the latter statement conveys more information.

Suppose $f : I \to J$.

Definition We say "f is one-to-one" if x_1 and x_2 are in I with $f(x_1) = f(x_2)$ can only happen when $x_1 = x_2$.

This can be described as 'f passes the horizontal line test' because it says that a horizontal line meets the graph in exactly one point.

On Friday, January 30, we outlined the following theorem that identifies the one-to-one functions defined on an interval:

Theorem *If I is an interval and $f : I \to \mathbb{R}$ is continuous and one-to-one on I, then either f is strictly increasing on I or f is strictly decreasing on I.*

Proof. Suppose x_1, x_2, and x_3 are three points of I such that $x_1 < x_2 < x_3$. Since f is one-to-one on I and x_1, x_2, and x_3 are distinct points, it must be true that $f(x_1)$, $f(x_2)$, and $f(x_3)$ are also distinct points. In particular, this means that either $f(x_1) < f(x_3)$ or $f(x_1) > f(x_3)$.

To prove that f is strictly monotonic, we need to show that for any three such points in I, that $f(x_2)$ lies between $f(x_1)$ and $f(x_3)$. Doing this requires examining several cases; we prove one of these cases and leave the rest for the reader to consider in detail because all the cases are handled in essentially the same way.

Suppose in the above that $f(x_1) < f(x_3)$ and, contrary to the desired conclusion, $f(x_2) > f(x_3)$. Choose c so that $f(x_3) < c < f(x_2)$. Since f is continuous on I, by the intermediate value theorem, there is q with $x_2 < q < x_3$ and $f(q) = c$. Also, because $f(x_1) < f(x_3)$, it is also the case that $f(x_1) < c < f(x_2)$. As before, since f is continuous on I, by the intermediate value theorem, there is p with $x_1 < p < x_2$ and $f(p) = c$.

Since $x_1 < p < x_2 < q < x_3$, it follows that $p \neq q$ which means that $f(p) = c = f(q)$ contradicts the fact that f is one-to-one. Thus, it is not possible to have $f(x_1) < f(x_3) < f(x_2)$.

In a similar way, we see that $f(x_2) < f(x_1) < f(x_3)$, $f(x_3) < f(x_1) < f(x_2)$, and $f(x_2) < f(x_3) < f(x_1)$ are also impossible.

This means that it must be the case that $f(x_1) < f(x_2) < f(x_3)$ or $f(x_3) < f(x_2) < f(x_1)$, that is, it means that for any three points $x_1 < x_2 < x_3$ in I, the value $f(x_2)$ is between the values $f(x_1)$ and $f(x_3)$. Since this is true for any three points in I, the function f must be strictly monotonic. □

Definition We say "f maps I onto J" if $f(x)$ is defined for each x in I, for all such x, the value $f(x)$ is in J, *and* for every y in J, there is x in I so that $f(x) = y$.

For example, the squaring function maps $\mathbb{R}$ onto $[0, \infty)$ but does not map $\mathbb{R}$ onto $\mathbb{R}$.

Also, on Friday, we stated and outlined the proof of a basic theorem on inverse functions:

Theorem *If $f : I \to J$ is one-to-one and maps I onto J, then there is a unique function $g : J \to I$ that maps J onto I such that $g(y) = x$ if and only if $f(x) = y$.*

The functions f and g are called inverse functions of each other. The condition defining g can also be written as $g(f(x)) = x$ for each x in I or $f(g(y)) = y$ for each y in J. We noted in class that if f is strictly increasing, then g is also strictly increasing and if f is strictly decreasing, so is g.

We are most concerned in determining the properties of the set J and the function g for the case that f is continuous, or even differentiable, and one-to-one on an interval.

- What kind of set is possible for J?
- Is g necessarily continuous?
- Is g necessarily differentiable?
- If g is differentiable, what is the derivative of g?

Theorem *If a and b are real numbers with $a < b$ and $f : [a, b] \to J$ is a continuous, one-to-one map of the closed interval $[a, b]$ onto J, then J is also a closed and bounded interval, that is there are real numbers c and d so that $J = [c, d]$. Moreover, if g is the inverse of f, then the continuity of f on $[a, b]$ implies that g is also continuous on $[c, d]$.*

Proof. When f is a continuous, one-to-one map defined on an interval, the theorem above showed that either f is strictly increasing or f is strictly decreasing.

Suppose f is strictly increasing on $[a, b]$. Since $a < b$, the fact that f is strictly increasing means that $f(a) < f(b)$; we let c denote $f(a)$ and let $d = f(b)$. Moreover, since f is strictly increasing, we also see that if $a < x < b$, then $c = f(a) < f(x) < f(b) = d$.

Now the intermediate value theorem says that, because f is continuous, if y is a number with $f(a) = c < y < d = f(b)$ then there is an x with $a < x < b$ so that $f(x) = y$. Thus, the set of values of f,

$$\{y : \text{ there is } x \text{ with } a \le x \le b \text{ and } f(x) = y\}$$

is exactly the interval $J = [c, d]$. That is, f maps the interval $[a, b]$ one-to-one and onto the interval $[c, d]$ and its inverse, g, maps $[c, d]$ one-to-one and onto $[a, b]$.

To show that g is continuous, we first show that if y_0 satisfies $c < y_0 < d$, then for any $\epsilon > 0$ there is a number δ so that when $|y - y_0| < \delta$, then $|g(y) - g(y_0)| < \epsilon$. Let $x_0 = g(y_0)$. Since g maps $[c, d]$ onto $[a, b]$, we know x_0 is in $[a, b]$ and since $c < y_0 < d$, then actually $a < x_0 < b$.

So, suppose $\epsilon > 0$ is given. Let ϵ_1 be the smallest of the three (positive) numbers ϵ, $x_0 - a$, and $b - x_0$. This means that if x is a number with $|x - x_0| < \epsilon_1$, then x is in (a, b) and $|x - x_0| < \epsilon$. Now let $y_1 = f(x_0 - \epsilon_1)$ and $y_2 = f(x_0 + \epsilon_1)$. Since f is strictly increasing $y_1 < y_0 < y_2$. We have set up the situation so that f maps the open interval $(x_0 - \epsilon_1, x_0 + \epsilon_1)$ one-to-one and onto the open interval (y_1, y_2), so this means that the function g maps the open interval (y_1, y_2) onto the open interval $(x_0 - \epsilon_1, x_0 + \epsilon_1)$.

Finally, since y_0 is in the open interval (y_1, y_2), we choose δ to be small enough so that the interval $(y_0 - \delta, y_0 + \delta)$ is contained in the interval (y_1, y_2).

Putting this together, we see that if y satisfies $|y - y_0| < \delta$, then y is in (y_1, y_2) and this means that $g(y)$ is in the interval $(x_0 - \epsilon_1, x_0 + \epsilon_1)$. That is,

$$|g(y) - g(y_0)| = |g(y) - x_0| < \epsilon_1 \le \epsilon$$

Thus, we have shown that for any positive number ϵ, we can find a number δ so that if $|y - y_0| < \delta$, then $|g(y) - g(y_0)| < \epsilon$. This says that g is continuous on the interval (c, d).

Similar arguments show that if f is strictly continuous $[a, b]$ so that g is strictly increasing on $[c, d]$, then, also, g is right-continuous at c and left-continuous at d. This will complete the proof that, in the case that g is strictly increasing, g is continuous on $[c, d]$. These arguments will be left for the reader to complete.

The proof for the case in which f is strictly decreasing is precisely analogous. We begin with the proof that when f is strictly decreasing, f maps the interval $[a, b]$ one-to-one and onto J, then J is the interval $[c, d]$ where this time, $c = f(b)$ and $d = f(a)$. As before, g maps the interval $[c, d]$ onto the interval $[a, b]$, but this time many of the inequalities need to be reversed in the proof that g is mapping the interval continuously. This proof will also be left to the reader. □

We now consider the question of differentiability: if f is differentiable on (a, b), is g differentiable on (c, d)? The answer is 'yes, almost'!

Theorem *Let a and c be real numbers or $-\infty$ and let b and d be real numbers or $+\infty$ such that $a < b$ and $c < d$. Suppose f is a continuous one-to-one map of (a, b) onto (c, d). If f is differentiable at x_0 in (a, b) and $f'(x_0) \neq 0$, then g, the inverse of f, is differentiable at $y_0 = f(x_0)$ and*

$$g'(y_0) = \frac{1}{f'(g(y_0))} = \frac{1}{f'(x_0)}$$

The exception $f'(x_0) \neq 0$ is necessary! For example, if $f(x) = x^3$, then f maps $\mathbb{R}$ one-to-one and onto $\mathbb{R}$, and the inverse of f is $g(y) = \sqrt[3]{y}$. Now, the derivative $f'(x) > 0$ except for $x_0 = 0$ and $f'(0) = 0$, so we know f is strictly increasing on $(-\infty, \infty)$. The inverse, $g(y) = \sqrt[3]{y}$ is also strictly increasing on $(-\infty, \infty)$ and g is differentiable everywhere except at $y_0 = 0 = f(0)$ where $f'(0) = 0$ and $g'(0)$ does not exist.

Proof. Suppose a, b, f, g, x_0, and y_0 are as given in the hypothesis of the Theorem above. Since f maps (a, b) one-to-one and onto (c, d) and f is continuous, we have seen above that g is continuous at y_0 and f maps open intervals containing x_0 onto open intervals containing y_0 and g maps open intervals contain y_0 onto open intervals containing x_0.

Suppose we write $f(x) = y$ for x near x_0. Then the continuity of g at y_0 implies $\lim_{y \to y_0} g(y) = g(y_0) = x_0$, so $y \to y_0$ is the same as $x \to x_0$.

To see if g is differentiable at y_0, we must check the existence of a certain limit. But we have

$$\lim_{y \to y_0} \frac{g(y) - g(y_0)}{y - y_0} = \lim_{x \to x_0} \frac{x - x_0}{f(x) - f(x_0)} = \frac{1}{f'(x_0)}$$

where the last inequality holds because, by hypothesis, $f'(x_0)$ exists and is non-zero. This means that $g'(y_0)$ exists and is equal to $1/f'(x_0) = 1/f'(g(y_0))$ as we were to prove. □

The previous three theorems, or sometimes just the last two theorems, are known as the 'inverse function theorem' and are the basis of our proofs of the properties of exp, arcsin, and arctan, based on our definition of them as the inverses of

$$\ln x = \int_1^x \frac{1}{t}\, dt$$

for $0 < x < \infty$ and of $\sin(x)$ and $\tan(x)$ for $-\pi/2 < x < \pi/2$.