Some mistakes and misleading items in Section 1.1 of Bartle & Sherbert (4th ed.)

Below, "B&S" stands for Bartle & Sherbert (4th ed.). The book's notations $D(f)$ and $R(f)$ for the domain and range (respectively) of a function f, are also used.

Composition of Functions

Definition 1.1.12 of B&S starts with "If $f : A \to B$ and $g : B \to C$ and $R(f) \subset= D(g) = B$, then the composite function $g \circ f$ is ..." (Recall that $R(f)$ and $D(g)$ are the book's notation for the range of f and the domain of g, respectively.) The "and $R(f) \subset= D(g)$" is redundant, and should have been omitted. The range of a function is always a subset of the codomain. The notation "$f : A \to B$" tells you that the codomain of f is B, hence that $R(f) \subseteq B$. The notation "$g : B \to C$" tells you that the domain of g is B. Hence we automatically have $R(f) \subset D(g)$.

In class, in the second lecture, I talked about a situation more general than having"$f : A \to B$ and $g : B \to C$" in which the composition $g \circ f$ can be defined. (See my lecture notes for 9/2/20, p. 5, second half.) This appears to be what Bartle and Sherbert had in mind, but is not what they wrote.

In *most* higher-level math classes and books, the setup with "$f : A \to B$ and $g : B \to C$" isss always used when defining composition. I mentioned the more general situation in class *only* because your textbook mentions it (implicitly, and usually in redundant statements like the one in Definition 1.1.12); I've never used it in MAA 4211–4212 before.

Inverse Functions

Exercise 1.1/20a (which you were not assigned) says "Suppose that f is an injection. Show that $f^{-1} \circ f(x) = x$ for all $x \in D(f)$ and that $f \circ f^{-1}(y) = y$ for all $y \in R(f)$." This problem makes no sense as written, because the hypothesis that f is injective is not enough to ensure that an inverse function exists; for that, we need f to be *bi*jective. This is one reason I assigned you non-book problem B1 (on Assignment 1) rather than the book's problems 1.1/20 and 1.1/24.

What the writers were thinking of in this exercise, but expressed inconsistently with the definitions in Section 1.1, is the following. Let $f : A \to B$ be a bijection. Define a function $\hat{f} : A \to R(f)$ by $\hat{f}(x) = f(x)$ for all $x \in A$. (This is an example of "shrinking the codomain", which was discussed in class.) By the definition of the range of f, the function $\hat{f}$ is surjective. Since f is injective, so is $\hat{f}$ (why?). Thus $\hat{f}$ has an inverse function, $\hat{f}^{-1}$, and for all $x \in D(f)$ and all $y \in R(f) = \text{codomain}(\hat{f})$ we have $(\hat{f}^{-1} \circ f)(x) = (\hat{f}^{-1} \circ \hat{f})(x) = x$ and $(f \circ \hat{f}^{-1})(y) = (\hat{f} \circ \hat{f}^{-1})(y) = y$. But unless f is surjective (in which case $\hat{f} = f$), the functions f and $\hat{f}$ are *different functions*, and f is not invertible.

> **Notational remarks.** (1) Above, note that I used parentheses around the compositions in "$(\hat{f}^{-1} \circ f)(x)$, "$(\hat{f}^{-1} \circ \hat{f})(x)$, "$(f \circ \hat{f}^{-1})(y)$," and "$(\hat{f} \circ \hat{f}^{-1})(y)$", just as I recommended in class. B&S do this themselves in 1.1/24. (It may be that one of the exercises 1.1/20a, 1.1/24, was written by Bartle—the sole author of the first two

editions—and the other was later (re)written by Sherbert.) Even when f^{-1} exists, the notation used in 1.1/20a, "$f^{-1} \circ f(x)$" and "$f \circ f^{-1}(y)$", is not recommended. (2) The notation "$\hat{f}$" is not an official or standard notation for the function defined above. I'm using it just for this discussion, and will use it with different meanings in other discussions.

Definition of a function

B&S does not actually define what a *function* is; the book only defines the term "function from A to B". This can lead to some misunderstandings when comparing functions with different domains and/or codomains, and especially when trying to decide whether two functions are equal.

The *full* set-theoretic definition of "function" is this: A function f is actually an ordered triple of sets $(A, B, \bar{f})$, where A and B are sets. and where $\bar{f}$ is the set that B&S Definition 1.1.6 calls "a function from A to B" (i.e. $\bar{f}$ is a subset of $A \times B$ such that for every $a \in A$, there exists a unique $b \in B$ for which $(a, b) \in \bar{f}$.) The set $\bar{f}$ is what is called the *graph* of the function f. Thus, this careful set-theoretic definition recovers the distinction between a function and its graph that the "active" definition gave us. A function and its graph are truly *not* the same thing.

Recall that the "active", but vague, definition of *function* that I gave in class was that a function "consists of" three things: a set called the domain, a set called the codomain, and an assignment of an element of the codomain to each element of the range. Since "domain", "codomain", and "assignment of ..." are different words/phrases, the vague phrase "consists of" is easy to make precise without sacrificing the "active" nature of this definition; we can (re)define a function as an ordered triple "(domain, codomain, assignment ...)". As mentioned in class, the real problem with the "active" definition is that the meaning of "assignment" is unclear. *Given the domain and codomain*, the set-theoretic definition of "function from A to B" eliminates this problem. But this is *all* that it does.

To get a correct definition set-theoretic definition of "function" (with no "from A to B"), we still need all the data in the "active" definition. All that the "passive", set-theoretic definition of *function* does is to replace the vaguely-worded third component of the triple "(domain, codomain, assignment ...)" by something crystal clear. We can't get rid of the need to specify a domain and codomain.

In the set-theoretic definition of *function*, we obtain the wrong definition of *equality of functions* if we don't use the *full* set-theoretic definition. Two functions $f = (A, B, \hat{f})$ and $g = (C, D, \hat{g})$ are defined to be equal if and only if these ordered triples are equal, i.e. if and only if $A = C$, $B = D$,and $\hat{f} = \hat{g}$.

One instance in which this is important is the notion of "shrinking the codomain" of a function. If $f : A \to B$ is a function, and $R(f) \subseteq C \subseteq B$, we can define a function $\hat{f} : A \to C$ by $\hat{f}(x) = f(x)$ for all $x \in A$. If $C \subsetneq B$, then f *and* $\hat{f}$ *are different functions*, even though their

graphs are the same set:

$$\{(x, f(x)) \mid x \in A\} = \{(x, \hat{f}(x)) \mid x \in A\}.$$

Similarly, there is a notion of *enlarging* (or *extending*) *the codomain*. If $f : A \to B$ is a function, and $B \subseteq E$, we can define a function $\tilde{f} : A \to E$ by $\tilde{f}(x) = f(x)$ for all $x \in A$. If $B \subsetneq E$, then f *and* $\tilde{f}$ *are different functions* with the same graph. "Enlarging the codomain" comes up quite often. For example, when given an integer-valued function, we may want to "regard" it as a real-valued function. What this really means is that, using the fact that $\mathbf{Z} \subset \mathbf{R}$, we are replacing the original $\mathbf{Z}$-valued function f by the corresponding $\mathbf{R}$-valued function $\tilde{f}$.

A second instance in which the full set-theoretic definition of "function" plays a role is in the distinguishing unequal "empty" functions. An *empty* function is a function whose domain is the empty set. For any set B, we have $\emptyset \times B = \emptyset$. The empty set itself satisfies the criteria to be a function $\hat{f}$ from $\emptyset$ to B: a set of order pairs (a, b) such that for each $a \in \emptyset$, there is a unique $b \in B$ for which $(a, b) \in \emptyset$. (This criterion is satisfied *vacuously*, since there *are* no elements $a \in \emptyset$.) Moreover, $\emptyset$ is the *unique* subset of the empty set $\emptyset \times B$, so there are no other functions from $\emptyset$ to B. Thus for each set B, the ordered triple $(\emptyset, B, \emptyset)$ is the unique function from $\emptyset$ to B. We call this function *the empty function with codomain* B, or *the empty* B*-valued function*. The *graph* of every empty function is the same, namely the empty set. But if B and C are different sets, then the corresponding empty functions are different: $(\emptyset, B, \emptyset) \neq \emptyset, C, \emptyset)$.

Empty functions are not very interesting or important. The main purpose of defining them is to enable various *other* definitions, theorems, etc., to be worded without an extra statement for the special case in which some set happens to be empty. (However, if the statement of a theorem allows for some function to be an empty function, then depending on what's being stated, it may still be necessary to handle the empty-function and nonempty-function case separately in the proof.) The student may easily check the following:

1. If $A \neq \emptyset$, there does not exist any function from A to $\emptyset$. (Hence there's not a second type of "empty function" that arises from considering empty codomains.)

2. Every empty function is injective. (The injectivity criterion is met vacuously.)

3. The empty function from $\emptyset$ to itself (vacuously) satisfies the definition of "surjective function" (hence is bijective, in view of the preceding), and satisfies the definition of the identity map id_A for $A = \emptyset$. Thus, the function $\mathrm{id}_\emptyset$ could reasonably be called "the empty bijection".