Lexicographic order

These notes are about an example that isn't directly relevant to the material in Rudin, but which seems to me to shed some light on the "least upper bound" property of the real numbers. You might read it as a "lite" version of the construction of the real numbers in the appendix to chapter 1. A side benefit is that some of the ideas are well-known to you but difficult to state formally; so you'll get some practice in interpreting formally complicated statements.

Definition 1. A *word* is a non-empty string of letters (like *supercalifragilisticexpialidocious* or *mpqskhdg* or *r*). A little more formally, a word is a symbol

$$X = x_1 x_2 \cdots x_r,$$

where r is a positive integer and each x_i is an element of the set $\{a, b, c, \ldots, x, y, z\}$. The integer r is called the *length* of the word X.

The world of words is a fascinating one, but we're going to look only at one small part of it.

Definition 2. The *lexicographic order* on words is the relation defined by $X < Y$ if X comes (strictly) before Y in the dictionary. Here's a more formal definition. First we order the alphabet in the obvious way:

$$a < b < c < d < \cdots < w < x < y < z.$$

Now suppose that

$$X = x_1 \cdots x_r, \qquad Y = y_1 \cdots y_s$$

are two words. Then $X < Y$ if and only if there is a non-negative integer t with the following properties:

i) $t \leq r$ and $t \leq s$;
ii) for every positive integer $i \leq t$, $x_i = y_i$; and
iii) either $x_{t+1} < y_{t+1}$, or $t = r$ and $t < s$.

Normal people would call this relation "alphabetical order." Mathematicians prefer the Greek origins of lexicographic to the Anglo-Saxon alphabetical. Whatever it is called, some examples of the relation are

$$a < aa < aaa < ab < aba.$$

Proposition 3. The lexicographic order is an order relation on words.

Proof. According to the definition of order relation in Rudin, there are two things we need to prove. The first is that if X and Y are two distinct words, then either $X < Y$ or $Y < X$ but not both. The second is that if $X < Y$ and $Y < Z$, then $X < Z$. I will sketch the proof of the first fact only.

Fix distinct words

$$X = x_1 \cdots x_r, \qquad Y = y_1 \cdots y_s.$$

Let t be the largest non-negative integer satisfying conditions i) and ii) of Definition 2. (The integer 0 satisfies the conditions, and no integer larger than $r + s$ satisfies

them; so there must be a largest t.) We now divide the proof into four cases; it will be clear that exactly one of these four cases must hold.

Case 1. $r = s = t$. In this case condition ii) implies that $x_i = y_i$ for all i, so $X = Y$. This contradicts the assumption that X and Y are distinct; so Case 1 is impossible.

Case 2. $r = t$, $s > t$. In this case, according to Definition 2, $X < Y$ is true and $Y < X$ is false; so exactly one of the relations is true, as we wished to show.

Case 3. $r > t$, $s = t$. In this case, according to Definition 2, $X < Y$ is false and $Y < X$ is true; so exactly one of the relations is true, as we wished to show.

Case 4. $r > t$ and $s > t$. In this case, by the choice of t, $t + 1$ cannot satisfy condition i) of Definition 2. This means that the letters x_{t+1} and y_{t+1} must be distinct. Because the ordering on the alphabet is an order relation, it follows that exactly one of the relations $x_{t+1} < y_{t+1}$ and $x_{t+1} > y_{t+1}$ must be true. By Definition 2, this means that exactly one of the relations $X < Y$ and $Y < X$ is true, as we wished to show.

We see that in every case exactly one of the relations $X < Y$ and $Y < X$ is true, so the lexicographic order satisfies Rudin's first axiom. I'll leave to you the problem of proving the second axiom. Q.E.D.

The order relations that you ought to be familiar with are the standard ones on the sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$, and perhaps those on some other subsets of $\mathbb{R}$ like the unit interval $[0, 1]$. Here are some facts to show that the lexicographic order on words is different from all of these. (You should think about which of these facts are true for which of the standard examples above.)

1. There is a first word a, but no last word.
2. Every word has an immediate successor; for example, the immediate successor of $mqwm$ is $mqwma$.
3. Not every word has an immediate predecessor; for example, $mqwm$ has no immediate predecessor. (For every predecessor of $mqwm$, like $mqwl$, we can find another word between the two; in this case

$$mqwl < mqwld < mqwm,$$

for example.)

You should try to decide exactly which words *do* have an immediate predecessor. (One example is $mqwma$, whose immediate predecessor is $mqwm$.)

The main point of looking at words in these notes was to look at the idea of least upper bound. One bounded set of words is

$$E = \text{all words beginning with } c \text{ or } d.$$

One upper bound for E is the word e. In fact it's not hard to check (from Definition 2) that the upper bounds for E are precisely the words beginning with one of the letters $e, f, g, \dots, z$. Once you know that, you can see that e is the least upper bound of E: $e = \sup E$.

Here's a more exotic example. Let F be the set of all words with no double letters. You should convince yourself that F is bounded, and that zz is a least upper bound. (What about the set of words with at most one set of double letters?)

Here at last is an example of a bounded set of words with no least upper bound. Let

$$G = \{a, ab, aba, abab, \dots\},$$

the set of all words beginning with a and consisting of an alternating string of as and bs. One upper bound for G is b, but we can do better: ac is a smaller upper bound, and abb is smaller than that, and $abac$ is smaller than that.

Here is a description of all the upper bounds of G. (The point of stating it is to give you some practice at saying "obvious" but complicated things.)

Proposition 4. *Suppose that $n = 2m + 1$ is a positive odd integer. Then the set of all upper bounds of G having length $2m + 1$ consists of all words X of length $2m + 1$ such that*

$$X \geq abab \cdots abb.$$

Here the word on the right consists of m ab's followed by b.

Suppose that $n = 2m + 2$ is a positive odd integer. Then the set of all upper bounds of G having length $2m + 2$ consists of all words X of length $2m + 2$ such that

$$X \geq abab \cdots abac.$$

Here the word on the right consists of m ab's followed by ac.

So G has a smallest 1-letter upper bound, a smallest 2-letter upper bound, a smallest 3-letter upper bound, and so on; but each of these is greater than the next, so none of them is a least upper bound. (By contrast, the smallest n-letter upper bound of E is $ba \cdots a$ (b followed by $n-1$ as; the first of these is the smallest, and so is the least upper bound of E.)

Corollary 5. *The set G of words is bounded, but has no least upper bound.*

The other point of this example was to illuminate the construction of the real numbers from the rational numbers. The problem with the rational numbers for doing analysis is that there are "approximate computations" with rational numbers that don't lead to a rational answer. An example is the old-fashioned square root algorithm that produces the sequence of approximate square roots of 2

$$1, 1.4, 1.41, 1.414, 1.4142, \ldots .$$

These rational numbers don't approach any rational number. One way to talk about this is to say that the set of rational numbers

$$\{1, 1.4, 1.41, 1.414, 1.4142, \ldots\}$$

is bounded, but has no least upper bound. Real numbers are supposed to fix that problem: the central analytic property of the real numbers is that any bounded set of real numbers has a least upper bound.

A problem that Rudin relegates to an appendix (and we'll mostly omit) is this: why does there exist an ordered field with the least upper bound property? Let's look at the corresponding problem for the lexicographic order. Can we enlarge the ordered set of words in some way to get the least upper bound property? Start with the set G above. You might think of the words appearing in Proposition 4 as "approximate" least upper bounds. The word they're trying to approximate is the "infinite word" $ababab \cdots$. So here's a way to proceed.

Definition 6. A *long word* is a non-empty string of letters that may or may not be infinite. A little more formally, a word is a symbol

$$X = x_1 x_2 x_3 \cdots ,$$

where each x_i is either a blank or an element of the set $\{a, b, c, \ldots, x, y, z\}$. We require in addition that blanks appear only at the end of the word. Formally,

i) x_1 is not blank; and
ii) If x_n is blank, then x_m is blank for all $m \geq n$.

The word is called *finite* if some x_n is blank. In that case there is a last non-blank letter x_r, and the integer r is called the *length* of the word. The word is called *infinite* if no x_n is blank.

Definition 7. The *lexicographic order* on long words is the relation defined as follows. Suppose that

$$X = x_1 x_2 \cdots, \qquad Y = y_1 y_2 \cdots$$

are two long words. Then $X < Y$ if and only if there is a non-negative integer t with the following properties:

i) x_t and y_t are not blank.
ii) for every positive integer $i \leq t$, $x_i = y_i$; and
iii) either $x_{t+1} < y_{t+1}$, or x_{t+1} is blank and and y_{t+1} is not.

Proposition 8. *The lexicographic order is an order relation on long words. Its restriction to finite words agrees with the lexicographic order already defined there.*

You should think carefully about how to modify the proof of Proposition 3 to prove this.

Here are some facts about the order on long words.

1. There is a first long word a, and a last long word $zzz\cdots$.
2. Every finite word has an immediate successor (obtained by adding an a at the end). The infinite words having an immediate successor are those ending in an infinite string of z's, except for the last word $zzz\cdots$. (The immediate successor is obtained by removing the final string of z's, and increasing the last letter by one. For example, the immediate successor of $caszzz\cdots$ is cat.)
3. The long words having an immediate predecessor are the finite words, except for the first word a. (If the finite word ends in a, the immediate predecessor is obtained by removing it. If the word ends in a letter after a, the predecessor is obtained by lowering the last letter by one, and adding an infinite string of z's. For example, the immediate predecessor of b is $azzz\cdots$.)

Fact 1 means in particular that every set of long words is bounded. The big theorem is

Proposition 9. *The lexicographic order on long words has the least upper bound property. In fact every long word is the least upper bound of some non-empty set of finite words.*

Proof. Let E be a non-empty set of long words (see Definition 6). E is automatically bounded above by the last long word $zzz\cdots$; so we must prove that E has a least upper bound. We are going to construct a long word

$$\alpha = a_1 a_2 a_3 \cdots,$$

and then prove that α is a least upper bound. We'll construct the word α by constructing its letters one at a time.

First of all, define

$$L(E) = \text{set of all first letters of words in } E \subset \{a, b, \dots, y, z\}.$$

Because E is not empty, $L(E)$ is a non-empty set of letters. Define

$$a_1 = \text{largest letter in } L(E).$$

That is, a_1 is the (alphabetically) last letter that occurs as the first letter of some word in E.

Next, we're going to define an entirely new set of long words. Define

$$E_2 = \text{set of all long words } X \text{ such that } a_1X \text{ belongs to } E.$$

There are two possibilities. If E_2 is empty, that means that the only word in E beginning with a_1 is the one-letter word a_1. In that case we define a_i to be a blank for all $i \geq 2$. It's not hard to see that the one-letter word a_1 is a least upper bound for E.

The interesting case is when E_2 is *not* empty, so that E has words of at least two letters beginning with a_1. We proceed just as we did for E, looking at the non-empty set $L(E_2)$ of first letters of words in E_2, and defining

$$a_2 = \text{largest letter in } L(E_2).$$

This means that a_1a_2 is alphabetically the highest possible beginning for a word in E.

From here on we just continue in the same way. Define

$$E_3 = \text{set of all long words } X \text{ such that } a_2X \text{ belongs to } E_2.$$

If E_3 is empty, that means that the only word in E beginning with a_1a_2 is the two-letter word a_1a_2. In that case we define a_i to be a blank for all $i \geq 3$, and the two-letter word a_1a_2 is a least upper bound for E.

When E_3 is *not* empty, define

$$a_3 = \text{largest letter in } L(E_3).$$

Continuing in this way, we get a long word

$$\alpha = a_1a_2a_3\cdots,$$

which may be finite. I'll leave to you the task of proving that α is a least upper bound of E.

There was one more assertion in Proposition 9: that every long word is the least upper bound of a non-empty set of finite words. So fix a long word

$$\alpha = a_1a_2a_3\cdots.$$

If α is finite, then it is the least upper bound of the one-element set $E = \{\alpha\}$, and we are done. So suppose α is infinite; that is, that no a_i is blank. Define finite words

$$e_1 = a_1, \quad e_2 = a_1a_2, \quad e_3 = a_1a_2a_3, \dots,$$

and define $E = \{e_1, e_2, e_3, \dots\}$. Again I leave to you the proof that α is a least upper bound of E. Q.E.D.