Getting seriously confused about making decisions

Introduction

This discussion is motivated by a puzzle I heard which really confused me. I'll discuss it later, but here's the statement:

A tricky game

A and B are playing a game.

1. A tosses a fair coin repeatedly, until it comes out tails. Let n be the number of heads.
2. A puts 3^n cents in one box and $3^{(n+1)}$ cents in another identical box.
3. A shuffles the boxes randomly and brings them to B.
4. B picks a box and examines the contents of that box only.
5. B can choose to either take the money from the box picked, or take the money from the other box.

Question: What is B's optimal strategy? When should she stay with the opened box, and when should she switch to the other one?

A framework for making decisions

Thinking about this game got me into a tangle about how to make decisions. So here's how I think about it.

Real decisions are never simple, but we abstract what we get out of them to make it easier to think about. Let's assume a notion of *utility*, a real numbered quantity that applies to outcomes and maybe people. You make decisions now by choosing options that give higher utility. That utility real numbered suggests that we can add utilities and do various other things, which we'll discuss now.

First simple decision

You're given a choice, action A gives you utility 1, action B gives you utility 1.5. Our definition of utility says that you should take action B.

Extending this situation: Suppose action A is such that if you take twice then you get utility 2. If A can be done twice and B only once, then you'd prefer A twice to B once. Higher total utility - aha, so we can add it.

Man, I'm making a meal of this, that's because this gets tricky fast, and I want us to be ready.

Probabilistic decisions

Now suppose your choice is utility 1 with probably 0.5, or utility 1 with probability 0.6. Easy, you take the second. Higher probability of gaining utility.

What if I said utility 1 with probably 0.5, or utility 0.4 with probability 1. Can we compare these, is it part of our definition of utility?

It seems like it might be easy. The first has *expected* utility of 0.5 while the second has expected utility of 0.4, which is lower. Our notion of utility seems much more powerful if it allows us to perform these kinds of comparisons.

Crucial decision

This example is crucial to the puzzle we started with:

- Option 1: You are given utility U.
- Option 2: You get utility $\frac{U}{3}$ with probability $\frac{2}{3}$, otherwise you get utility $3U$.

Which of these is better? Crucially,

1. Does the answer depend on the value of U here (see section on unbounded utility)?
2. Do we need to know anything else before we can answer this question?

It certainly seems as if we don't need to know any more. We can calculate the expected utility in this case (as above) and discover that option 2 has higher expected utility. It's just arithmetic, because all the probabilities and values are laid out for us.

A Paradox?

Working through the problem at the top of the page we realize that, whatever value is opened in the first box, we're given a decision exactly analogous to our "crucial decision". Though we're dealing in cents instead of utility.

Applying this reasoning we see that if we played this game many times we'd always do the same thing. First we'd open a box, we'd find U cents in it, and swap to the other box. *This is clearly absurd.*

Now, you might object that actually, with enough money in the first box you actually wouldn't swap, you'd take your safe winnings, as any more wouldn't matter much, but less would matter a little. That's why I talk about utility here, and have the section on whether utility is bounded. For the purposes of this whole situation I imagine that the boxes contain "utility", though it's easier to talk about cents.

The symmetry of the boxes makes the result of the expectation calculation seem very doubtful. However, *it is correct*, it gives the true expectation for swapping, and that expectation is higher than sitting still.

On the other hand, one could calculate the completely distribution of outcomes from this game, averaged over all possible coin flips at the beginning. I think of this distribution in terms of its cumulative distribution function (CDF). Let's ask how two strategies on the extremes compare, in their total distribution of outcomes:

1. Keep the money in the box you open, unless you open 1cent, then swap.
2. Always swap boxes

Examining these strategies we find that they're almost identical in terms of outcomes. That is, the probability of getting 3^n is the same for both strategies for all n (except that the first strategy gets \$3 instead of \$1 sometimes). I want to find a way of driving this home better: it's saying that, in the long run there's no difference in what people get using strategy 1 compared to strategy 2.

We can make this more extreme. Now suppose that there's a charge of 1% of whatever you open to swap. Now the swapping strategy is actually worse than the non-swapping strategy (yet it still always gives positive expectation when you consider an individual swap). Put another way, its CDF is dominated by the non-swapping strategy. The probability that I'll get at least $\$x$ by not swapping is always the same as or greater than the probability I'll get at least $\$x$ by swapping!

Above, where I first described the 'crucial decision' it seemed very simple and unambiguous. Yet when placed in the context the puzzle at the top of this document, the answer seems seriously questionable. Note, however, that it *could* have arisen in some totally different context where it wasn't possible to make it look silly like this. The calculations woudln't change though.

Another way to decide?

Given the original puzzle we can compare the CDFs of the two different strategies and discover that not-swapping dominates always swapping. I assert that this makes it unambiguously better, it requires fewer assumptions (albeit weak) than the argument with expectation. It adds *probabilities* together to make the comparison, but that is well founded probabilistic operation.

However, in other situations one could imagine being presented a choice that gave rise to two different CDFs of utility. With neither dominating the other. We could adjust the coin flipping problem to suggest such curves, though I think it's simpler to be given a simple binary choice of utility distributions, as it removes other complications.

In this case it seems that we have no general way to decide. When expectations are finite we would usually take the distribution with the highest expectation. When one or both expectations are infinite, using expectation leads to absurdities.

Worse still, decisions that don't explicitly show unbounded or infinite expectations can be embedded in situations that do. Our 'crucial decision' has bounded utility, and is embedded in an unbounded utility situation by the original problem, casting doubt on the original conclusion.

My conclusion

This makes me suspect that this theory of utility does not, in general, work in the context of probability. It's a formalism that makes things easy to figure out sometimes, and in practical terms it's used all over the world in countless contexts to great effect.

Expected utility is great in practice. In the sitautions available to us everything is bounded, and can only exist within other bounded systems. In these cases everything seems to work out.

However in a more theoretical context, pathological cases like the top example seem to show that it doesn't hold water.

Appendix: Is utility bounded?

For the purposes of this discussion it's not really clear what utility *is*. If it were money you might argue it wasn't bounded. However you might also argue that the usefulness of money to any given person is bounded, that notion is probably closer to what we usually mean by utility. We might also think of utility as the happiness, or something like that, of a person.

Let's assume we're happy to add this over people. In principle we might consider the number of people a universe could hold to be unbounded (perhaps via some many worlds theory or something). Then in principle, total utility could be unbounded. Alternatively, perhaps the universe will not have a heat death. The prize in the boxes could be the security of a happy species that suffers no threat of extinction for 3^n years.

In any case this whole utility thing is a formalism, it's not clear that our formalism cares whether the universe is infinite or not. So I'm going to assume that utility is unbounded. I will *not* asume that we can have infinite utility.

Acknowledgemnts

Thanks to Ryan Giordano, James Martin, Owen Cotton-Barratt, and Nick Wedd for useful and interesting discussions about this.