Chapter 10 Sure Thing Principle

10.1 Generalising Dominance

The maximise expected utility rule also supports a more general version of dominance. We'll state the version of dominance using an example, then spend some time going over how we know maximise expected utility satisfies that version.

The original dominance principle said that if *A* is better than *B* in every state, then *A* is simply better than *B* simply. But we don't have to just compare choices in individual states, we can also compare them across any number of states. So imagine that we have to choose between *A* and *B* and we know that one of four states obtains. The utility of each choice in each state is given as follows.

	S_1	S_2	S_3	S_4
A	10	9	9	0
B	8	3	3	3

And imagine we're using the maximin rule. Then the rule says that *A* does better than *B* in S_1, while *B* does better than *A* in S_4. The rule also says that *B* does better than *A* overall, since it's worst case scenario is 3, while *A*'s worst case scenario is 0. But we can also compare *A* and *B* with respect to pairs of states. So conditional on us just being in S_1 or S_2, then *A* is better. Because between those two states, its worst case is 9, while *B*'s worst case is 3.

Now imagine we've given up on maximin, and are applying a new rule we'll call maxiaverage. The maxiaverage rule tells us make the choice that has the highest (or **maxi**mum) average of best case and worst case scenarios. The rule says that *B* is better overall, since it has a best case of 8 and a worst case of 3 for an average of 5.5, while *A* has a best case of 10 and a worst case of 0, for an average of 5.

But if we just know we're in S_1 or S_2, then the rule recommends *A* over *B*. That's because among those two states, *A* has a maximum of 10 and a minimum of 9, for an average of 9.5, while *B* has a maximum of 8 and a minimum of 3 for an average of 5.5.

And if we just know we're in S_3 or S_4, then the rule also recommends *A* over *B*. That's because among those two states, *A* has a maximum of 9 and a minimum of 0, for an average of 4.5, while *B* has a maximum of 3 and a minimum of 3 for an average of 3.

This is a fairly odd result. We know that either we're in one of S_1 or S_2, or that we're in one of S_3 or S_4. And the rule tells us that if we find out which, i.e. if we find out we're in S_1 or S_2, or we find out we're in S_3 or S_4, either way we should choose *A*. But before we find this out, we should choose *B*.

Here then is a more general version of dominance. Assume our initial states are $\{S_1, S_2, ..., S_n\}$. Call this set *S*. A binary partition of*S* is a pair of sets of states, call them T_1 and T_2, such that every state in *S* is in exactly one of T_1 and T_2. (We're simplifying a little here - generally a partition is any way of dividing a collection up into parts such that every member of the original collection is in one of the ‘parts’. But we’ll only be interested in cases where we divide the original states in two, i.e. into a *binary* partition.) Then the generalised version of dominance says that if A is better than B among the states in T_1, and it is better than B among the states in T_2, where T_1 and T_2 provide a partition of S, then it is better than B among the states in S. That’s the principle that maxiaverage violates. A is better than B among the states $\{S_1, S_2\}$. And it is better than B among the states $\{S_3, S_4\}$. But it isn’t better than B among the states $\{S_1, S_2, S_3, S_4\}$. That is, it isn’t better than B among the states generally.

We’ll be interested in this principle of dominance because, unlike perhaps dominance itself, there are some cases where it leads to slightly counterintuitive results. For this reason some theorists have been interested in theories which, although they satisfy dominance, do not satisfy this general version of dominance.

On the other hand, maximise expected utility does respect this principle. In fact, it respects an even stronger principle, one that we’ll state using the notion of **conditional expected utility**. Recall that as well as probabilities, we defined conditional probabilities above. Well conditional expected utilities are just the expectations of the utility function with respect to a conditional probability. More formally, if there are states $S_1, S_2, ..., S_n$, then the expected utility of A conditional on E, which we’ll write $Exp(U(A|E)$, is

$$Exp(U(A|E)) = Pr(S_1|E)U(S_1|A) + Pr(S_2|E)U(S_2|A) + ... + Pr(S_n|E)U(S_n|A)$$

That is, we just replace the probabilities in the definition of expected utility with conditional probabilities. (You might wonder why we didn’t also replace the utilities with conditional utilities. That’s because we’re assuming that states are defined so that given an action, the state has a fixed utility. If we didn’t make this simplifying assumption, we’d have to be more careful here.) Now we can prove the following theorem.

- If $Exp(U(A|E)) > Exp(U(B|E))$, and $Exp(U(B|\neg E)) > Exp(U(B|\neg E))$, then $Exp(U(A)) > Exp(U(B))$.

We’ll prove this by proving something else that will be useful in many contexts.

- $Exp(U(A)) = Exp(U(A|E))Pr(E) + Exp(U(A|\neg E))Pr(\neg E)$

To see this, note the following

$$\begin{aligned}
Pr(S_i) &= Pr((S_i \wedge E) \vee (S_i \wedge \neg E)) \\
&= Pr(S_i \wedge E) + Pr(S_i \wedge \neg E) \\
&= Pr(S_i|E)Pr(E) + Pr(S_i|\neg E)Pr(\neg E)
\end{aligned}$$

And now we’ll use this when we’re expanding $Exp(U(A|E))Pr(E)$.

$$\begin{aligned}
Exp(U(A|E))Pr(E) &= Pr(E)[Pr(S_1|E)U(S_1|A) + Pr(S_2|E)U(S_2|A) + ... + Pr(S_n|E)U(S_n|A)] \\
&= Pr(E)Pr(S_1|E)U(S_1|A) + Pr(E)Pr(S_2|E)U(S_2|A) + ... + Pr(E)Pr(S_n|E)U(S_n|A) \\
Exp(U(A|\neg E))Pr(\neg E) &= Pr(\neg E)[Pr(S_1|\neg E)U(S_1|A) + Pr(S_2|\neg E)U(S_2|A) + ... + Pr(S_n|\neg E)U(S_n|A)] \\
&= Pr(\neg E)Pr(S_1|\neg E)U(S_1|A) + Pr(\neg E)Pr(S_2 \neg|E)U(S_2|A) + ... + Pr(\neg E)Pr(S_n|\neg E)U(S_n|A)
\end{aligned}$$

Putting those two together, we get

$$\begin{aligned}
&Exp(U(A|E))Pr(E) + Exp(U(A|\neg E))Pr(\neg E) \\
&\quad = Pr(E)Pr(S_1|E)U(S_1|A) + ... + Pr(E)Pr(S_n|E)U(S_n|A) + \\
&\quad + Pr(\neg E)Pr(S_1|\neg E)U(S_1|A) + ... + Pr(\neg E)Pr(S_n|\neg E)U(S_n|A) \\
&\quad = (Pr(E)Pr(S_1|E) + Pr(\neg E)Pr(S_1|\neg E))U(S_1|A) + ... + (Pr(E)Pr(S_n|E) + Pr(\neg E)Pr(S_n|\neg E))U(S_n|A) \\
&\quad = Pr(S_1)U(S_1|A) + Pr(S_2)U(S_2|A) + ...Pr(S_n)U(S_n|A) \\
&\quad = Exp(U(A))
\end{aligned}$$

Now if $Exp(U(A|E)) > Exp(U(B|E))$, and $Exp(U(B|\neg E)) > Exp(U(B|\neg E))$, then the following two inequalities hold.

$$Exp(U(A|E))Pr(E) \geq Exp(U(B|E))Pr(E)$$
$$Exp(U(A|\neg E))Pr(\neg E) \geq Exp(U(B|\neg E))Pr(\neg E)$$

In each case we have equality only if the probability in question ($Pr(E)$ in the first line, $Pr(\neg E)$ in the second) is zero. Since not both $Pr(E)$ and $Pr(\neg E)$ are zero, one of those is a strict inequality. (That is, the left hand side is greater than, not merely greater than or equal to, the right hand side.) So adding up the two lines, and using the fact that in one case we have a strict inequality, we get.

$$Exp(U(A|E))Pr(E) + Exp(U(A|\neg E))Pr(\neg E) \geq Exp(U(B|E))Pr(E) + Exp(U(B|\neg E))Pr(\neg E)\text{i.e. } Exp(U(A)) \quad > Exp(U(B))$$

That is, if A is better than B conditional on E, and it is better than B conditional on $\neg E$, then it is simply better than B.

10.2 Sure Thing Principle

The result we just proved is very similar to a famous principle of decision theory, the Sure Thing Principle. The Sure Thing Principle is usually stated in terms of one option being at least as good as another, rather than one option being better than another, as follows.

Sure Thing Principle If $AE \succeq BE$ and $A\neg E \succeq B\neg E$, then $A \succeq B$.

The terminology there could use some spelling out. By $A \succ B$ we mean that A is preferred to B. By $A \succeq B$ we mean that A is regarded as at least as good as B. The relation between $\succ$ and $\succeq$ is like the relation between ¿ and $\geq$. In each case the line at the bottom means that we're allowing equality between the values on either side.

The odd thing here is using $AE \succeq BE$ rather than something that's explicitly conditional. We should read the terms on each side of the inequality sign as *conjunctions*. It means that A *and* E is regarded as at least as good an outcome as B and E. But that sounds like something that's true just in case the agent prefers A to B conditional on E obtaining. So we can use preferences over conjunctions like AE as proxy for conditional preferences.

So we can read the Sure Thing Principle as saying that if A is at least as good as B conditional on E, and conditional on $\neg E$, then it really is at least as good as B. Again, this looks fairly plausible in the abstract, though we'll soon see some reasons to worry about it.

Expected Utility maximisation satisfies the Sure Thing Principle. I won't go over the proof here because it's really just the same as the proof from the previous section with $>$ replaced by $\geq$ in a lot of places. But if we regard the Sure

Thing Principle as a plausible principle of decision making, then it is a good feature of Expected Utility maximisation that it satisfies it.

It is tempting to think of the Sure Thing Principle as a generalisation of a principle of logical implication we all learned in propositional logic. The principle in question said that from $X \rightarrow Z$, and $Y \rightarrow Z$, and $X \vee Y$, we can infer C. If we let Z be that A is better than B, let X be E, and Y be $\neg E$, it looks like we have all the premises, and the reasoning looks intuitively right. But this analogy is misleading for two reasons.

First, for technical reasons we can't get into in depth here, preferring A to B conditional on E isn't the same as it being true that if E is true you prefer A to B. To see some problems with this, think about cases where you don't know E is true, and A is something quite horrible that mitigates the effects of the unpleasant E. In this case you do prefer AE to BE, and E is true, but you don't prefer A to B. But we'll set this question, which is largely a logical question about the nature of conditionals, to one side.

The bigger problem is that the analogy with logic would suggest that the following generalisation of the Sure Thing Principle will hold.

Disjunction Principle If $AE_1 \succeq BE_1$ and $AE_2 \succeq BE_2$, and $Pr(E_1 \vee E_2) = 1$ then $A \succeq B$.

But this "Disjunction Principle" seems no good in cases like the following. I'm going to toss two coins. Let p be the proposition that they will land differently, i.e. one heads and one tails. I offer you a bet that pays you \$2 if p, and costs you \$3 if $\neg p$. This looks like a bad bet, since $Pr(p)$ = 0.5, and losing \$3 is worse than gaining \$2. But consider the following argument.

Let E_1 be that at least one of the coins landing heads. It isn't too hard to show that $Pr(p|E_1) = \frac{2}{3}$. So conditional on E_1, the expected return of the bet is $\frac{2}{3} \times 2 - \frac{1}{3} \times 3 = \frac{4}{3} - 1 = \frac{1}{3}$. That's a positive return. So if we let A be taking the bet, and B be declining the bet, then conditional on E_1, A is better than B, because the expected return is positive.

Let E_2 be that at least one of the coins landing tails. It isn't too hard to show that $Pr(p|E_1) = \frac{2}{3}$. So conditional on E_2, the expected return of the bet is $\frac{2}{3} \times 2 - \frac{1}{3} \times 3 = \frac{4}{3} - 1 = \frac{1}{3}$. That's a positive return. So if we let A be taking the bet, and B be declining the bet, then conditional on E_2, A is better than B, because the expected return is positive.

Now if E_1 fails, then both of the coins lands tails. That means that at least one of the coins lands tails. That means that E_2 is true. So if $E1$ fails $E2$ is true. So one of $E1$ and $E2$ has to be true, i.e. $Pr(E_1 \vee E_2) = 1$. And $AE_1 \succeq BE_1$ and $AE_2 \succeq BE_2$. Indeed $AE_1 \succ BE_1$ and $AE_2 \succ BE_2$. But $B \succ A$. So the disjunction principle isn't in general true.

It's a deep philosophical question how seriously we should worry about this. If the Sure Thing Principle isn't any more plausible intuitively than the Disjunction Principle, and the Disjunction Principle seems false, does that mean we should be sceptical of the Sure Thing Principle? As I said, that's a very hard question, and it's one we'll return to a few times in what follows.

10.3 Allais Paradox

The Sure Thing Principle is one of the more controversial principles in decision theory because there seem to be cases where it gives the wrong answer. The most famous of these is the Allais paradox, first discovered by the French economist (and Nobel Laureate) Maurice Allais. In this paradox, the subject is first offered the following choice between A and B. The results of their choice will depend on the drawing of a coloured ball from an urn. The urn contains 10 white balls, 1 yellow ball, and 89 black balls, and assume the balls are all randomly distributed so the probability of drawing each is identical.

	White	Yellow	Black
A	\$1,000,000	\$1,000,000	\$0
B	\$5,000,000	\$0	\$0

That is, they are offered a choice between an 11% shot at \$1,000,000, and a 10% shot at \$5,000,000. Second, the subjects are offered the following choice between *C* and *D*, which are dependent on drawings from a similarly constructed urn.

	White	Yellow	Black
C	\$1,000,000	\$1,000,000	\$1,000,000
D	\$5,000,000	\$0	\$1,000,000

That is, they are offered a choice between \$1,000,000 for sure, and a complex bet that gives them a 10% shot at \$5,000,000, an 89% shot at \$1,000,000, and a 1% chance of striking out and getting nothing.

Now if we were trying to maximise expected *dollars*, then we'd have to choose both *B* and *D*. But, and this is an important point that we'll come back to, dollars aren't utilities. Getting \$2,000,000 isn't twice as good as getting \$1,000,000. Pretty clearly if you were offered a million dollars or a 50% chance at two million dollars you would, and should, take the million for sure. That's because the two million isn't twice as useful to you as the million. Without a way of figuring out the utility of \$1,000,000 versus the utility of \$5,000,000, we can't say whether *A* is better than *B*. But we can say one thing. You can't consistently hold the following three views.

- $B \succ A$
- $C \succ D$
- The Sure Thing Principle holds

This is relevant because a lot of people think $B \succ A$ and $C \succ D$. Let's work through the proof of this to finish with.

Let E be that either a white or yellow ball is drawn. So $\neg E$ is that a black ball is drawn. Now note that $A\neg E$ is identical to $B\neg E$. In either case you get nothing. So $A\neg E \succeq B\neg E$. So if $AE \succeq BE$ then, by Sure Thing, $A \succeq B$. Equivalently, if $B \succ A$, then $BE \succ AE$. Since we've assumed $B \succ A$, then $BE \succ AE$.

Also note that $C\neg E$ is identical to $D\neg E$. In either case you get a million dollars. So $D\neg E \succeq C\neg E$. So if $DE \succeq CE$ then, by Sure Thing, $D \succeq C$. Equivalently, if $C \succ D$, then $CE \succ DE$. Since we've assumed $C \succ D$, then $CE \succ DE$.

But now we have a problem, since $BE = DE$, and $AE = CE$. Given E, then choice between A and B just is the choice between C and D. So holding simultaneously that $BE \succ AE$ and $CE \succ DE$ is incoherent.

It's hard to say for sure just what's going on here. Part of what's going on is that we have a 'certainty premium'. We prefer options like *C* that guarantee a positive result. Now having a certainly good result is a kind of holistic property of *C*. The Sure Thing Principle in effect rules out assigning value to holistic properties like that. The value of the whole need not be *identical* to the value of the parts, but any comparisons between the values of the parts has to be reflected in the value of the whole. Some theorists have thought that a lesson of the Allais paradox is that this is a mistake.

We won't be looking in this course at theories which violate the Sure Thing Principle, but we will be looking at justifications of the Sure Thing Principle, so it is worth thinking about reasons you might have for rejecting it.

10.4 Exercises

10.4.1 Calculate Expected Utilities

In the following example $Pr(S_1) = 0.4$, $Pr(S_2) = 0.3$, $Pr(S_3) = 0.2$ and $Pr(S_4) = 0.1$. The table gives the utility of each of the possible actions (A, B, C, D and E) in each state. What is the expected utility of each action?

	S_1	S_2	S_3	S_4
A	0	2	10	2
B	6	2	1	7
C	1	8	9	7
D	3	1	8	6
E	4	7	1	4

10.4.2 Conditional Choices

In the previous example, C is the best thing to do conditional on S_2. It has expected utility 8 in that case, and all the others are lower. It is also the best thing to do conditional on $S_2 \vee S_3$. It has expected utility 8.4 if we conditionalise on $S_2 \vee S_3$, and again all the others are lower.

For each of the actions A, B, C, D and E, find a proposition such that conditional on that proposition, the action in question has the highest expected utility.

10.4.3 Generalised Dominance

Does the maximax decision rule satisfy the generalised dominance principle we discussed in the text? That principle says that if the initial range of states is S, and T_1 and T_2 form a partition of S, and if A is a better choice than B conditional on being in T_1, and A is also a better choice than B conditional on being in T_2, then A is simply a better choice than B. Does this principle hold for the maximax decision rule?

10.4.4 Sure Thing Principle

Assume we're using the 'Maximise Expected Utility' rule. And assume that B is not the best choice out of our available choices conditional on E. Assume also that B is not the best choice out of our available choices conditional on $\neg E$. Does it follow that B is not the best available choice? If so, provide an argument that this is the case. If not, provide a counterexample, i.e. a case where B is not the best choice conditional on E, not the best choice conditional on $\neg E$, but the best choice overall.