Part III
CHOICE UNDER OBJECTIVE RISK

7 Modelling Risk

Choice under uncertainty is a topic of fundamental interest to economists, since most economic decisions are made in the face of uncertainty. For instance, firms have to make decisions regarding prices and production, and investors in the stock market have to decide whether to buy or sell stocks, and if so, then how many, etc. Insurance is a huge industry in developed countries, and it exists only because people are averse to the uncertainty that pervades their everyday lives. But in order to rigorously study the economics of uncertainty, one first needs a formal model of how agents behave in the face of uncertainty, which we develop in this topic.

When we turn our attention to subjective uncertainty in later chapters, it will become clear that there are in fact different types of uncertainty. We begin by studying the most basic type of uncertainty, which we refer to as *risk*. This is the kind of uncertainty we face in the casino while playing the slot machine: there is uncertainty about whether we will win a prize or not, but we know enough about this uncertainty that we can compute the exact probability of winning. In later chapters this will be contrasted with the uncertainty one may feel when buying stocks: this uncertainty relies on fine details of the economy that may not even be able to conceptualize, and as a result, it does not lend itself to calculating the probabilities of outcomes, at least not in the same way as in the casino.

Before we can write down a formal model of how people choose among uncertain alternatives, we first need to find a way to formally describe uncertain alternatives.

7.1 Choice Domain: Lotteries

We use the terms "gamble" or "lottery" or "uncertain prospects" for any uncertain alternatives for which the probability of each outcome is known. A prospect of getting an outcome x with probability 1 is referred to as a *degenerate* lottery. In general, an outcome could be anything – it could be money, or a trip to Vegas, etc. When the outcomes of a lottery are money, we call it a *monetary* lottery.

A lottery can be viewed as a *probability tree* with a final outcome at

each terminal node. For instance, suppose there is a gamble where a fair coin is flipped twice. It yields \$10 if two heads come up, and \$0 otherwise. Then there are two possible final outcomes, namely \$10 and \$0, and the probabilities of obtaining each are 0.25 and 0.75, respectively.

7.2 Reduced Form of a Lottery

Notice that three things go into describing the lottery in the preceding example:

(i) the possible outcomes (\$10 and \$0),

(ii) their associated probabilities (0.25 and 0.75, respectively), and

(iii) the structure of the lottery (the coin is flipped at most two times, that is, the uncertainty resolves in up to two stages).

The *reduced form* of a lottery specifies just the first two and leaves out the third. In the current example, we would write the reduced form as $(\frac{1}{4}, \$10; \frac{3}{4}\$0)$. In general, the reduced form of a lottery p is denoted by

$$(p_1, x_1; p_2, x_2; .....; p_n, x_n),$$

or alternatively, when we need clearer exposition we denote it by

$$\begin{bmatrix} p_1 & x_1 \\ \vdots & \vdots \\ p_n & x_n \end{bmatrix}.$$

Notice that two different lotteries can have the same reduced form. The above lottery involving two coin flips is different from one that involves a biased coin that comes up heads (resp. tails) with probability 0.25 (resp. 0.75) and yields \$10 if heads and \$0 if tails. Yet both have the same reduced form.

A degenerate lottery that yields x with probability 1 is written $(1, x)$.

7.3 Mixtures of Lotteries

Take any two lotteries p, q, each of which involves only one stage of uncertainty:

$$p = \begin{bmatrix} p_1 & x_1 \\ \vdots & \vdots \\ p_n & x_n \end{bmatrix} \quad \text{and} \quad q = \begin{bmatrix} q_1 & x_1 \\ \vdots & \vdots \\ q_n & x_n \end{bmatrix}$$

Now consider another lottery that involves two stages of uncertainty. Specifically, suppose that in the first stage, the lottery p is realized with probability α and lottery q is realized with probability $1-\alpha$, and in the second stage the realized lottery is played out. Thus, in the first stage we learn whether we obtain lottery p or q, and in the second stage, the outcome of the obtained lottery is received. This "mixture of lotteries" or "compound lottery" can be denoted:[24]

$$(\alpha, p; 1-\alpha, q) = \begin{bmatrix} \alpha & \begin{bmatrix} p_1 & x_1 \\ \vdots & \vdots \\ p_n & x_n \end{bmatrix} \\ 1-\alpha & \begin{bmatrix} q_1 & x_1 \\ \vdots & \vdots \\ q_n & x_n \end{bmatrix} \end{bmatrix}.$$

In order to specify the probabilities α and $(1-\alpha)$ by which p and q are being mixed, we call it an α-mixture of p and q.

The reduced forms of p, q are, of course,

$$\begin{aligned} p &= (p_1, x_1; p_2, x_2; .....; p_n, x_n), \\ q &= (q_1, x_1; q_2, x_2; .....; q_n, x_n). \end{aligned}$$

As a trivial exercise that just requires you to apply definitions and use elementary algebra, you are asked to:

Exercise 5 *Show that the reduced form of the compound lottery* $(\alpha, p; 1-\alpha, q)$ *is*

$$(\alpha, p; 1-\alpha, q) = (\alpha p_1 + (1-\alpha)q_1, x_1; .....; \alpha p_n + (1-\alpha)q_n, x_n).$$

[24] In general, a compound lottery could be of the form $(\alpha_1, p_1; \alpha_2, p_2; ..; \alpha_n, p_n)$ where there are many possibly outcomes of the first stage, not just two. But we will not be needing this much generality in what follows.

8 Expected Value Theory

The standard theory of choice under risk in economics is Expected Utility theory, or EU theory for short. We first present, however, the earliest version of that theory in order to introduce all the basic ideas before introducing the full details of EU theory.

8.1 Model

Expected Value Theory (EV for short) posits that the agent has a preference $\succsim$ over some set of alternatives A, and that choice maximizes preference. What makes it a theory of choice under uncertainty is that A is not just any set of alternatives, but rather a set of lotteries. Thus, the theory is one of agents who choose between lotteries.

The primitive of the theory is a preferences $\succsim$ over monetary lotteries. The hypothesis about the preference $\succsim$ is that it admits a utility representation EV where the utility of any lottery is the expected value of its reduced form:

$$EV(p_1, x_1; p_2, x_2; .....; p_n, x_n) = p_1x_1 + p_2x_2 + ...p_nx_n.$$

The theory says that when faced with a lottery, the agent only cares about its reduced form, and moreover, ranks lotteries according to their expected value.

The theory has some very nice features. First, the fact that only the reduced form matters to the agent can be viewed as "rational" – to the extent that all that matters is where we get to at the end (as opposed to how we get there) it makes sense that she should concerns herself only with the overall probabilities of possible outcomes, as opposed to the structure of the lottery. Second, the model captures an intuitive idea that if a lottery gives better outcomes with higher probabilities, then it will be more attractive. Indeed, this holds in the model because lotteries that give higher outcomes with higher probabilities will also have higher expected value, and thus the agent would prefer them. Finally, the elegance of the model is to be appreciated. It captures in a simple and highly compact manner some of the essential considerations that one might like from a theory of choice under uncertainty.

However, simplicity usually comes at the cost of sacrificing realism. As we will show now, the cost associated with the simplicity of the EV theory is too high.

8.2 Evidence

First a quick review of definitions:

Definition 1 *A preference $\succsim$ over lotteries is said to be risk averse toward a lottery of the form $p = (\alpha, x; (1-\alpha), y)$ if:*

$$p \prec (1, EV(p)).$$

Similarly it is said to be risk loving (respectively, risk neutral) if the above expression holds with $\succ$ (respectively, $\sim$).

To explain these definitions, consider a lottery p; for concreteness let $p = (\frac{1}{2}, 100; \frac{1}{2}, 0)$. The expected outcome of this lottery is $EV(p) = \$50$.[25] Although the expected outcome of p and $(1, EV(p))$ is identical, $(1, EV(p))$ gives the expected outcome for *sure* whereas p yields it with *risk*. Thus, an agent's preference between p and $(1, EV(p))$ comes down to how he feels about risk vs certainty. Risk averse agents will prefer a sure \$50 over a lottery that gives an expected \$50. Similarly for risk loving and risk neutral agents.[26]

Another definition:

Definition 2 *The certainty equivalent for a lottery p is the sure sum of money, denoted $CE(p)$, such that*

$$(1, CE(p)) \sim p.$$

[25] The expected outcome is defined as the average of the outcomes you'd get *if* you played the lottery repeatedly. Don't let this confuse you: the lottery is actually being played only once.

[26] Note that risk attitude (that is, aversion, affinity, or neutrality toward risk) is really a *psychological* notion, but we have defined it in terms of *behavior*. Risk aversion is properly defined in terms of a distaste for risk. We don't observe distaste directly, and thus this intuitive definition is useless for scientific purposes. However, by identifying the behavioral expression of distaste for risk, we are able to provide an empirical means of determining an agent's risk attitude.

The certainty equivalent is a measure of how much you like or dislike the lottery. If you say that playing the lottery $p = (\frac{1}{2}, 100; \frac{1}{2}, 0)$ is just as good as receiving \$10, then your certainty equivalent for the lottery is \$10. The low value of the certainty equivalent (relative to the expected outcome of \$50) suggests that the agent doesn't find himself very drawn to playing the lottery.[27]

Let us turn now to the case against EV theory. Write down your responses to the following two questions.

(A) What is your preference between the lottery $(\frac{1}{2}, 1000; \frac{1}{2}, -1000)$ and the sure (zero) outcome $(1, 0)$? Put differently, how do you feel about playing this lottery vs not playing it?

(B) What is your certainty equivalent for the following lottery? Suppose that an unbiased coin is tossed again and again until it lands on tails, and then you are paid an amount that depends on how many tosses it took for the coin to land on tails. Specifically, the payment rule is that you receive $\$2^n$ if the coin lands on tails in the n^{th} toss. Thus, you get \$2 if it lands on tails in the first toss, \$4 if it lands on heads the first time and tails the second, \$8 if it lands on heads the first two times and tails the third, etc. Yes, you can potentially win billions of dollars if n is large enough. Note that the probability of getting tails in the n^{th} toss is $\frac{1}{2^n}$. Thus this lottery can be written as $(\frac{1}{2}, 2; \frac{1}{4}, 4; \frac{1}{8}, 8; ...; \frac{1}{2^n}, 2^n; .....)$.

Most people would rather not play the lottery in (A) for the simple reason that uncertainty makes them uncomfortable as it is, and facing the possibility of losing \$1000 makes them even more uncomfortable. Since you can never lose any money with the lottery in (B), and you only stand to gain, the certainty equivalent will be strictly positive for any reasonable agent. Experiments report that typical certainty equivalents are a few dollars.

The following propositions establish that such responses contradict the EV theory, and thus that the EV theory is not a good descriptive theory

[27] Indeed, this suggests risk aversion – the agent values the lottery less than its expected outcome. Assuming that more money is preferred to less, it is not a surprise that $p \sim (1, \$10) \prec (1, \$50)$, that is, $p \prec (1, EV(p))$!

of choice under uncertainty. The propositions derive testable implications of the theory.

Proposition 6 *If an agent's preferences $\succsim$ over lotteries respects EV theory, then he must be risk neutral.*

Proof. Risk neutrality is defined by the indifference:

$$p \sim (1, EV(p)),$$

for any $p = (\alpha, x; (1-\alpha), y)$. To see that this indifference must always hold in EV theory, compute the utilities:

$$\begin{aligned} EV(p) &= \alpha x + (1-\alpha)y, \\ EV(1, EV(p)) &= \alpha x + (1-\alpha)y. \end{aligned}$$

Therefore, $EV(p) = EV(1, EV(p))$, and consequently $p \sim (1, EV(p))$. ■

Thus an EV agent is not swayed by the uncertainty he faces. All he cares about is the expected outcome of the lottery. In particular, he must be indifferent between playing $(\frac{1}{2}, 1000; \frac{1}{2}, -1000)$ and not playing it. What this tells us is that the EV theory cannot capture such attitudes as aversion or affinity toward uncertainty. To the extent that few people are neutral toward uncertainty (look at the prevalence of the demand for insurance), the EV theory is an inadequate descriptive theory of choice under uncertainty. Indeed this inadequacy is at a fundamental level, because the main interest in choice under risk comes from the observation that people are typically sensitive (in fact averse) to risk.

Turning to question (B), regardless of how much money served as your certainty equivalent for the lottery, your response rejected the EV theory by an unbelievable margin:

Proposition 7 *If an agent's preferences $\succsim$ over lotteries respects EV theory, then for **any** sure sum of money x (no matter how large),*

$$(\frac{1}{2}, 2; \frac{1}{4}, 4; \frac{1}{8}, 8; ...; \frac{1}{2^n}, 2^n; .....) \succ (1, x).$$

Proof. Denote $(\frac{1}{2}, 2; \frac{1}{4}, 4; \frac{1}{8}, 8; ...; \frac{1}{2^n}, 2^n; .....)$ by q to ease notation. Compute that

$$\begin{aligned} EV(q) &= EV(\tfrac{1}{2}, 2; \tfrac{1}{4}, 4; \tfrac{1}{8}, 8; ...; \tfrac{1}{2^n}, 2^n; .....) \\ &= \tfrac{1}{2} \times 2 + \tfrac{1}{4} \times 4 + \tfrac{1}{8} \times 8 + ... + \tfrac{1}{2^n} \times 2^n + ... \\ &= 1 + 1 + 1 + ... + 1 + ... \\ &= \infty. \end{aligned}$$

On the other hand, the utility of any sure sum x is $EV(1, x) = x$. Since x is finite, it follows that $EV(q) > EV(1, x)$, and thus, $q \succ (1, x)$, as desired. ■

Exercise 6 *Compute the certainty equivalent of the lottery* $q = (\frac{1}{2}, 2; \frac{1}{4}, 4; \frac{1}{8}, 8; ...; \frac{1}{2^n}, 2^n; .....)$.

The proposition reflects that the EV agent cares more about small probability outcomes than the typical person. When you were determining your certainty equivalent for the lottery $(\frac{1}{2}, 2; \frac{1}{4}, 4; \frac{1}{8}, 8; ...; \frac{1}{2^n}, 2^n; .....)$, you most likely were not bothered by the fact that the lottery could make you insanely rich. Most likely, the odds of the coin turning up heads say 19 times in a row before turning up tails (which would earn you more than one million dollars) are too small (about 10^{-6}) for you to give it any consideration. Yet, for the EV agent, it matters enough to weigh in on how he feels toward the lottery. A different (but more standard) interpretation of the proposition concerns how the EV agent feels about an additional dollar at different wealth levels. For the EV agent, an additional dollar adds one unit to utility *regardless* of how much money he already has (this is reflected in Proposition **??** below). This underlies the fact that for the EV agent, decreasing the odds and increasing a reward *by the same proportion* would leave utility unchanged. For instance, the EV agent is indifferent between $(\frac{1}{2}, 2; \frac{1}{2}, 0)$ and $(\frac{1}{1000}, 1000; \frac{999}{1000}, 0)$ since $EV(\frac{1}{2}, 2; \frac{1}{2}, 0) = EV(\frac{1}{1000}, 1000; \frac{999}{1000}, 0) = 1$. If you go back to the proof you will see that this is what leads to an infinite utility. However, it is more realistic for people to care less for an additional dollar as they get wealthier. In this case, decreasing the odds and increasing a reward by the same proportion would lead to a reduction in utility. For instance, $(\frac{1}{2}, 2; \frac{1}{2}, 0)$ would be better than $(\frac{1}{1000}, 1000; \frac{999}{1000}, 0)$. As we will see later, allowing for such behavior can give rise to a finite certainty equivalent for the lottery $(\frac{1}{2}, 2; \frac{1}{4}, 4; \frac{1}{8}, 8; ...; \frac{1}{2^n}, 2^n; .....)$.

This concludes our case against EV theory. A historical note on the experiment involving the lottery $(\frac{1}{2}, 2; \frac{1}{4}, 4; \frac{1}{8}, 8; ...; \frac{1}{2^n}, 2^n; .....)$: the experiment is called the St. Petersburg experiment, and the finding that people are willing to pay only a finite amount of money for the lottery (in contrast to the prediction of the theory) is called the St. Petersburg Paradox. The case against EV theory using this paradox was made in 1738 by Daniel Bernoulli, who then suggested Expected Utility theory.

9 Expected Utility Theory

9.1 Model

We now formulate Expected Utility theory for objective risk. It will be referred to as *Objective* or *von Neuman-Morgenstern* Expected Utility theory and denoted as EU.

Primitives: *A preference* $\succsim$ *over the set* A *of all lotteries, given some set* X *of outcomes*

Hypothesis for $\succsim$: *The preference* $\succsim$ *admits a utility representation of the form*

$$EU(p) = p_1u(x_1) + p_2u(x_2) + ...p_nu(x_n),$$

where $(p_1, x_1; p_2, x_2; .....; p_n, x_n)$ *is the reduced form of* p*, and* u *is a utility index.*

Hypothesis for Choice: *Preference maximization.*

The model states that when agents face a risky option $(p_1, x_1; p_2, x_2; .....; p_n, x_n)$ they consider the utility of each possible outcome, and then weight them by the corresponding probability.

9.2 Some Features of EU

9.2.1 Risk Attitudes

To illustrate some highly desirable properties of this model, we consider the fact that people's decisions under uncertain display different *risk attitudes*. That is, some people may be risk averse, others risk loving, and some may be blind to the risk altogether. A basic question of interest is in how the theory embodies various risk attitudes. In order to answer this we must first confirm that the primitives of the model are rich enough to allow us to behaviorally express risk attitudes:

Definition 3 *A preference $\succsim$ over lotteries is said to be risk averse toward a lottery of the form $p = (\alpha, x; (1-\alpha), y)$ if:*

$$p \prec (1, EV(p)).$$

Similarly it is said to be risk loving (respectively, risk neutral) if the above expression holds with $\succ$ (respectively, $\sim$).

To explain these definitions, consider a lottery p; for concreteness let $p = (\frac{1}{2}, 100; \frac{1}{2}, 0)$. The expected outcome of this lottery is $EV(p) = \$50$.[28] Although the expected outcome of p and $(1, EV(p))$ is identical, $(1, EV(p))$ gives the expected outcome for *sure* whereas p yields it with *risk*. Thus, an agent's preference between p and $(1, EV(p))$ comes down to how he feels about risk vs certainty. Risk averse agents will prefer a sure \$50 over a lottery that gives an expected \$50. Similarly for risk loving and risk neutral agents.[29]

There is also a second way of describing risk attitudes. Consider:

Definition 4 *The certainty equivalent for a lottery p is the sure sum of money, denoted $CE(p)$, such that*

$$(1, CE(p)) \sim p.$$

The certainty equivalent is a measure of how much you like or dislike the lottery. If you say that playing the lottery $p = (\frac{1}{2}, 100; \frac{1}{2}, 0)$ is just as good as receiving \$10, then your certainty equivalent for the lottery is \$10. The low value of the certainty equivalent (relative to the expected outcome of \$50) suggests that the agent doesn't find himself very drawn to playing the lottery. Indeed, this suggests risk aversion – the agent values the lottery less than its

[28] The expected outcome is defined as the average of the outcomes you'd get *if* you played the lottery repeatedly. Don't let this confuse you: the lottery is actually being played only once.

[29] Note that risk attitude (that is, aversion, affinity, or neutrality toward risk) is really a *psychological* notion, but we have defined it in terms of *behavior*. Risk aversion is properly defined in terms of a distaste for risk. We don't observe distaste directly, and thus this intuitive definition is useless for scientific purposes. However, by identifying the behavioral expression of distaste for risk, we are able to provide an empirical means of determining an agent's risk attitude.

expected outcome. Assuming that more money is preferred to less, it is not a surprise that $p \sim (1, \$10) \prec (1, \$50)$, that is, $p \prec (1, EV(p))$! Indeed, for agents who prefer more money to less, risk aversion (resp affinity, neutrality) towards a lottery p is equivalent to the condition that $CE(p) < EV(p)$ (resp, $CE(p) > EV(p)$, $CE(p) = EV(p)$).

As it turns out, EU theory captures risk attitudes in a very elegant way: the curvature of u fully describes the agent's risk attitude. Recall that concave functions are bowed upwards and convex functions are bowed downwards. The formal definitions are as follows. A function u is *concave* if $u(\alpha x + (1-\alpha)y) > \alpha u(x) + (1-\alpha)u(y)$ for all x, y and $0 < \alpha < 1$. A function u is *convex* if $u(\alpha x + (1-\alpha)y) < \alpha u(x) + (1-\alpha)u(y)$ for all x, y and $0 < \alpha < 1$.[30]

Proposition 8 *If an agent's preference $\succsim$ over monetary lotteries respects the EU theory with a strictly increasing u, then the following statements hold.*

(i) if u is affine, then the agent exhibits risk neutrality.

(ii) if u is concave, then the agent exhibits risk aversion.

(ii) if u is convex, then the agent exhibits risk affinity.

Proof. We need to show that there exists an example of the EU theory (that is, a specification of the utility index) for each of the cases. Consider each in turn:

(a) Risk neutrality.

Take any lottery $p = (\alpha, x; (1-\alpha), y)$ with $0 < \alpha < 1$. Let u be a strictly increasing concave utility index. Then

$EU(p)$

$= \alpha u(x) + (1-\alpha)u(y)$ (by definition of EU)

$= u(\alpha x + (1-\alpha)y)$ (by definition of affinity)

$= EU(1, \alpha x + (1-\alpha)y)$ (by definition of EU)

$= EU(1, EV(p))$ (by definition of EV).

That is, $EU(p) = EU(1, EV(p))$. Since EU is a representation, it follows that $p \sim (1, EV(p))$, that is, the agent is risk neutral, as desired.

[30] An example of a concave (respectively convex) function is $u(x) = \sqrt{x}$ (respectively $u(x) = x^2$).

(b) Risk aversion.

Take any lottery $p = (\alpha, x; (1-\alpha), y)$ with $0 < \alpha < 1$. Let u be a strictly increasing concave utility index. Then

$EU(p)$
$= \alpha u(x) + (1-\alpha)u(y)$ (by definition of EU)
$< u(\alpha x + (1-\alpha)y)$ (by definition of concavity)
$= EU(1, \alpha x + (1-\alpha)y)$ (by definition of EU)
$= EU(1, EV(p))$ (by definition of EV).

That is, $EU(p) < EU(1, EV(p))$. Since EU is a representation, it follows that $p \prec (1, EV(p))$, as desired.

(c) Risk affinity

Exercise. ■

9.2.2 Identification and Marginal Utility

Having analyzed what the model implies for behavior, we ask: Can we identify her utility index u from her choice behavior? The answer is yes.

Suppose outcomes are money amounts between \$0 and \$100. Suppose that for any such money amount m we ask the agent to reveal the probability α_m such that

$$(\alpha_m, 100; 1-\alpha_m, 0) \sim (1, m).$$

Note that the answer α_m is obtained from the agent's *behavior*, and in particular we are not observing her utility function. Then assuming that the agent is indeed EU with some utility index u and assuming further for simplicity that $u(0) = 0$ and $u(100) = 1$, we get that

$(1, m) \sim (\alpha_m, 100; 1-\alpha_m, 0)$
$\Longrightarrow EU(1, m) = EU(\alpha_m, 100; 1-\alpha_m, 0)$
$\Longrightarrow u(m) = \alpha_m u(100) + (1-\alpha_m)u(0)$
$\Longrightarrow u(m) = \alpha_m.$

That is, we have found out the exact utility $u(m)$ that the agent must be using. Repeating the above procedure for various m allows us to map out her entire u over the \$0 to \$100 range. Thus, by asking the agent a series of questions and observing her answers (that is, her behavior) we can identify u (after fixing the values of $u(0)$ and $u(100)$).

While it is interesting that an unobservable object like utility can nevertheless be indirectly observed through behavior, it is also interesting that the law of diminishing marginal utility has meaning in this model whereas in the more abstract setting it did not (recall our discussion in Chapter 2.3). The curvature of u was not restricted by any observables there, while here the entire function u and consequently its curvature is tightly connected with behavior, and indeed is entirely pinned down by it. Another expression of this is in Proposition 8, where we saw that the behavior exhibited by the EU agent is intimately connected with the curvature of u. Thus the law of diminishing marginal utility (which corresponds to concavity) has behavioral content, and is therefore empirically meaningful.

Since risk aversion is pervasive in the real world, economists typically assume that u is strictly concave.

10 Testable Implications of EU

In this chapter we derive two key testable implications of the model.

Proposition 9 *If $\succsim$ respects EU theory, then $\succsim$ satisfies Reduction of Compound Lotteries: for any gambles $p, q \in A$, if p and q have the same reduced form then $p \sim q$.*

Proof. If p and q have the same reduced form, then they must have the same Expected Utility, since the EU calculation is done on the reduced form of a gamble alone. Hence the EU agent must be indifferent between them. ■

Reduction of Compound Lotteries, or Reduction for short, may seem harmless enough as a behavioral prediction, but it is certainly possible to imagine that people's choices may not respect it. A nervous fellow may like all the uncertainty to be resolved sooner, and thus may prefer the single draw lottery to a multi draw lottery with the same reduced form, for instance. Framing effects may also come into play via the details of how the uncertainty resolves.

The second testable implication requires a bit of preparation. Recall the notion of a *mixture of lotteries*. In particular, recall that for any two lotteries p, q with reduced forms given by:

$$\begin{aligned} p &= (p_1, x_1; p_2, x_2; .....; p_n, x_n), \\ q &= (q_1, x_1; q_2, x_2; .....; q_n, x_n), \end{aligned}$$

the α-mixture of p and q is the lottery $(\alpha, p; 1-\alpha, q)$ with reduced form

$$(\alpha, p; 1-\alpha, q) = (\alpha p_1 + (1-\alpha)q_1, x_1; .....; \alpha p_n + (1-\alpha)q_n, x_n).$$

A very convenient property of EU which pertains to mixtures of lotteries is "mixture linearity": the expected utility of a mixture of lotteries equals the mixture of the expected utility of the gambles.

Exercise 7 *Consider the above three lotteries.*

(i) Using this reduced form of $(\alpha, p; 1-\alpha, q)$ *and the definition of EU, write out the expression for its expected utility* $EU(\alpha, p; (1-\alpha)q)$.

(ii) Show (by simple algebraic manipulations) that EU is ***mixture linear*** *in the sense that for any lotteries* $p, q \in A$ *and* α *between 0 and 1,*

$$EU(\alpha, p; (1-\alpha)q) = \alpha EU(p) + (1-\alpha)EU(q).$$

Proposition. *If* $\succsim$ *respects EU theory, then it satisfies Independence: for any* $0 < \alpha < 1$ *and any lotteries* $p, q, r \in A$,[31]

$$p \succsim q \Longleftrightarrow (\alpha, p; (1-\alpha), r) \succsim (\alpha, q; (1-\alpha), r).$$

Proof. We begin by proving that

$$p \succ q \Longrightarrow (\alpha, p; (1-\alpha), r) \succ (\alpha, q; (1-\alpha), r).$$

Take any α, p, q, r as in the statement of the proposition. Then,
$p \succ q$
$\Longrightarrow EU(p) > EU(q)$
$\Longrightarrow \alpha EU(p) > \alpha EU(q)$
$\Longrightarrow \alpha EU(p) + (1-\alpha)EU(r) > \alpha EU(q) + (1-\alpha)EU(r)$
since $(1-\alpha)EU(r)$ is just some number
$\Longrightarrow EU(\alpha, p; (1-\alpha), r) > EU(\alpha, q; (1-\alpha), r)$
by mixture linearity of EU
$\Longrightarrow (\alpha, p; (1-\alpha), r) \succ (\alpha, q; (1-\alpha), r)$,

[31] In the expanded notation, Independence states that

$$\begin{bmatrix} p_1 & x_1 \\ \vdots & \vdots \\ p_n & x_n \end{bmatrix} \succsim \begin{bmatrix} q_1 & x_1 \\ \vdots & \vdots \\ q_n & x_n \end{bmatrix} \Longleftrightarrow \begin{bmatrix} \alpha & \begin{bmatrix} p_1 & x_1 \\ \vdots & \vdots \\ p_n & x_n \end{bmatrix} \\ 1-\alpha & \begin{bmatrix} r_1 & x_1 \\ \vdots & \vdots \\ r_n & x_n \end{bmatrix} \end{bmatrix} \succsim \begin{bmatrix} \alpha & \begin{bmatrix} q_1 & x_1 \\ \vdots & \vdots \\ q_n & x_n \end{bmatrix} \\ 1-\alpha & \begin{bmatrix} r_1 & x_1 \\ \vdots & \vdots \\ r_n & x_n \end{bmatrix} \end{bmatrix}.$$

which proves

$$p \succ q \Longrightarrow (\alpha, p; (1-\alpha), r) \succ (\alpha, q; (1-\alpha), r).$$

It remains to prove the following three statements in order to establish Independence:

$$\begin{aligned} p &\succ q \Longleftarrow (\alpha, p; (1-\alpha), r) \succ (\alpha, q; (1-\alpha), r) \\ p &\sim q \Longrightarrow (\alpha, p; (1-\alpha), r) \sim (\alpha, q; (1-\alpha), r) \\ p &\sim q \Longleftarrow (\alpha, p; (1-\alpha), r) \sim (\alpha, q; (1-\alpha), r). \end{aligned}$$

However, the proof for these is analogous to the above argument. ■

Independence implies that the agent bases her preferences over lotteries on what is different between the lotteries and not what is common.

11 Psychology of Choice Under Risk

In this chapter we review evidence from economics and psychology on how people make choices when faced with uncertainty. Along the way we will evaluate the descriptive validity of EU theory.

11.1 Common Consequence Effect

Violations of Expected Utility theory have been observed in many experiments. One of the earliest findings is the so-called Allais' Paradox (also called the Common Consequence Effect) which involves the following typical preferences:

$$\begin{bmatrix} 1 & \$1 \text{ million} \end{bmatrix} \succ \begin{bmatrix} 0.1 & \$5 \text{ million} \\ 0.89 & \$1 \text{ million} \\ 0.01 & \$0 \end{bmatrix}$$

$$\begin{bmatrix} 0.11 & \$1 \text{ million} \\ 0.89 & \$0 \end{bmatrix} \prec \begin{bmatrix} 0.10 & \$5 \text{ million} \\ 0.90 & \$0 \end{bmatrix}$$

The first preference seems to be swayed by a desire for certainty, whereas the second by the magnitude of the higher reward. We show that these preferences violate Expected Utility theory.

If an agent was an Expected Utility agent, then

$$\begin{bmatrix} 1 & \$1 \text{ million} \end{bmatrix} \succ \begin{bmatrix} 0.1 & \$5 \text{ million} \\ 0.89 & \$1 \text{ million} \\ 0.01 & \$0 \end{bmatrix}$$

$$\Longrightarrow \begin{bmatrix} 0.11 & \begin{bmatrix} 1 & \$1 \text{ million} \end{bmatrix} \\ 0.89 & \begin{bmatrix} 1 & \$1 \text{ million} \end{bmatrix} \end{bmatrix} \succ \begin{bmatrix} 0.11 & \begin{bmatrix} 10/11 & \$5 \text{ million} \\ 1/11 & \$0 \end{bmatrix} \\ 0.89 & \begin{bmatrix} 1 & \$1 \text{ million} \end{bmatrix} \end{bmatrix} \text{ (by Reduction)}$$

$$\Longrightarrow \begin{bmatrix} 1 & \$1 \text{ million} \end{bmatrix} \succ \begin{bmatrix} 10/11 & \$5 \text{ million} \\ 1/11 & \$0 \end{bmatrix} \text{ (by Independence)}$$

$$\Longrightarrow \begin{bmatrix} 0.11 & \begin{bmatrix} 1 & \$1 \text{ million} \end{bmatrix} \\ 0.89 & \begin{bmatrix} 1 & \$0 \end{bmatrix} \end{bmatrix} \succ \begin{bmatrix} 0.11 & \begin{bmatrix} 10/11 & \$5 \text{ million} \\ 1/11 & \$0 \end{bmatrix} \\ 0.89 & \begin{bmatrix} 1 & \$0 \end{bmatrix} \end{bmatrix} \text{ (by Independence)}$$

$$\Longrightarrow \begin{bmatrix} 0.11 & \$1 \text{ million} \\ 0.89 & \$0 \end{bmatrix} \succ \begin{bmatrix} 0.10 & \$5 \text{ million} \\ 0.90 & \$0 \end{bmatrix} \text{ (by Reduction).}$$

That is, Expected Utility theory is inconsistent with the Allais Paradox.

By the way, the "paradox" in the Allais Paradox is that EU theory is inconsistent with real world behavior. That is, the paradox is for the theory, not for us.

11.2 Common Ratio Effect

The 'Common Ratio Effect' or 'Certainty Effect' is demonstrated by Kahneman and Tversky (1979) through the following typical preferences in their experiment:

$$\begin{bmatrix} 1 & \$3000 \end{bmatrix} \succ \begin{bmatrix} 0.80 & \$4000 \\ 0.20 & \$0 \end{bmatrix}$$

$$\begin{bmatrix} 0.25 & \$3000 \\ 0.75 & \$0 \end{bmatrix} \prec \begin{bmatrix} 0.20 & \$4000 \\ 0.80 & \$0 \end{bmatrix}.$$

We show that that is inconsistent with Expected Utility. Indeed, for an Expected Utility agent,

$$\begin{bmatrix} 1 & \$3000 \end{bmatrix} \succ \begin{bmatrix} 0.80 & \$4000 \\ 0.20 & \$0 \end{bmatrix}$$

$$\Longrightarrow \begin{bmatrix} 0.25 & \begin{bmatrix} 1 & \$3000 \end{bmatrix} \\ 0.75 & \begin{bmatrix} 1 & \$0 \end{bmatrix} \end{bmatrix} \succ \begin{bmatrix} 0.25 & \begin{bmatrix} 0.80 & \$4000 \\ 0.20 & \$0 \end{bmatrix} \\ 0.75 & \begin{bmatrix} 1 & \$0 \end{bmatrix} \end{bmatrix} \text{ (by Independence)}$$

$$\Longrightarrow \begin{bmatrix} 0.25 & \$3000 \\ 0.75 & \$0 \end{bmatrix} \succ \begin{bmatrix} 0.20 & \$4000 \\ 0.80 & \$0 \end{bmatrix} \text{ (by Reduction).}$$

Therefore, Expected Utility theory is inconsistent with the Certainty Effect, and is therefore refuted by it.

11.3 Isolation Effect

Kahneman and Tversky posit that agents often disregard the common components of two alternatives and isolate the differences in order to simplify the choice between the alternatives – this is called the Isolation Effect. We saw

that a version of this very same idea seemed to underlie in Independence. However, the Isolation Effect applies to 'obvious' differences whereas Independence applies the idea even when the differences are not obvious, such as when the differences are evident only after applying Reduction.

The Isolation Effect generates a violation of Reduction. In an experiment involving a two-stage lottery, a reward of 0 is received with probability 0.75, and with 0.25 probability the subject receives either

$$\begin{bmatrix} 1 & \$3000 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0.80 & \$4000 \\ 0.20 & \$0 \end{bmatrix}$$

depending on which of the two he chose *prior* to the first stage. Most chose the first option, thereby revealing a preference:

$$\begin{bmatrix} 0.25 & \begin{bmatrix} 1 & \$3000 \end{bmatrix} \\ 0.75 & \$0 \end{bmatrix} \succ \begin{bmatrix} 0.25 & \begin{bmatrix} 0.80 & \$4000 \\ 0.20 & \$0 \end{bmatrix} \\ 0.75 & \$0 \end{bmatrix}$$

But Reduction implies that

$$\begin{bmatrix} 0.25 & \$3000 \\ 0.75 & \$0 \end{bmatrix} \succ \begin{bmatrix} 0.20 & \$4000 \\ 0.80 & \$0 \end{bmatrix},$$

whereas in the Certainty Effect experiment above, the subjects exhibited the opposite preference. This implies that subjects violate Reduction.

What is your take on Reduction? If you were asked directly to compare the following reduction-equivalent one-stage and two-stage lotteries,

$$\begin{bmatrix} 0.20 & \$4000 \\ 0.80 & \$0 \end{bmatrix} \text{ and } \begin{bmatrix} 0.25 & \begin{bmatrix} 0.80 & \$4000 \\ 0.20 & \$0 \end{bmatrix} \\ 0.75 & \$0 \end{bmatrix}$$

would you have a strict preference or would you be indifferent? Why might some people not be indifferent?

11.4 Reference Dependence

Kahneman and Tversky (1979) ask subjects two questions:

– In addition to whatever you own, you have been given \$1000. Choose between a 50:50 chance of gaining \$1000 or 0, and sure gain of \$500.

– In addition to whatever you own, you have been given \$2000. Choose between a 50:50 chance of losing \$1000 or 0, and sure loss of \$500.

Observe that both problems yield *identical distributions over final wealth*. Standard economics posits that people only care about how their final wealth is influenced. Consequently, standard theory would predict that choices in either problem will be the same. However, the typical response involved a choice of the sure gain in the first problem, but a choice for the gamble in the second.

A similar finding involving non-monetary outcomes employs the following thought-experiment. Suppose the U.S. is preparing for the outbreak of a disease which is expected to kill 600 people. They have a choice of adopting one of two programs:

– If program A is adopted, 200 people will be saved.

– If program B is adopted, then there is $\frac{1}{3}$ probability that 600 people will be saved, and $\frac{2}{3}$ probability that no people will be saved.

Among subjects, the typical choice was Program A. However, subjects typically chose Program B when the following equivalent description of the two programs was given:

– If program A is adopted, 400 people will die.

– If program B is adopted, then there is $\frac{1}{3}$ probability that nobody will die, and $\frac{2}{3}$ probability that 600 people will die.

These findings are examples of a broad phenomenon known as the *framing effect*: the response to a question depends on how the question is framed. The framing effect in these examples can be explained through the idea of *reference dependence*. Note that in the first experiment, preferences change as the frame changes from one involving gains in wealth to one involving losses in wealth. Indeed, it appears that utility is derived from changes in wealth rather than final wealth, that is, people evaluate alternatives in terms of changes from some *reference level*, such as wealth. In the disease

experiment, preferences changed as alternatives were framed as lives saved vs lives lost. With this interpretation, the reference point is 'no one saved' in the first frame, and 'no one dies' in the second ('200 saved' is a gain only if the reference point is 'no one saved' – if the reference is 'all saved' then '200 saved' is in fact a loss).

11.5 Reflection Effect and the Fourfold Pattern

Note that in the reference dependence experiments above, subjects actually went from being risk averse when alternatives were framed as gains to being risk loving when alternatives were framed as losses. This is an example of what is called a *reflection effect*: the property that risk attitudes flip as rewards flip from being gains to being losses. A direct demonstration of the reflection effect is the typical preference in Kahneman and Tversky (1979):

$$\begin{aligned}(1, 3000) &\succ (0.8, 4000;\ \ 0.2, 0)\\ (1, -3000) &\prec (0.8, -4000;\ \ 0.2, 0).\end{aligned}$$

Here the agents were risk averse for gains but risk loving for losses. Kahneman and Tversky find a reflection effect in the opposite direction in the following typical preferences:

$$\begin{aligned}(0.001, 5000;\ \ 0.999, 0) &\succ (1, 5)\\ (0.001, -5000;\ \ 0.999, 0) &\prec (1, -5),\end{aligned}$$

that is, risk affinity for gains and risk aversion for losses.

Based on such observations, Kahneman and Tversky hypothesize a *fourfold pattern of risk attitudes*: with large probabilities, subjects are risk averse for gains and risk loving for losses, but with small probabilities, subjects are risk loving for gains and risk averse for losses:

	gains	losses
large prob	risk averse	risk loving
small prob	risk loving	risk averse

Note that the change in risk attitudes in the rows are what we were referring to as the reflection effect.

11.6 Loss Aversion

Another observation by Kahneman and Tversky is *loss aversion*: losses loom larger than gains. Intuition supports that for $x > y \geq 0$,

$$(\frac{1}{2}, y; \frac{1}{2}, -y) \succ (\frac{1}{2}, x; \frac{1}{2} - x).$$

So if losing \$1 hurts more than it feels good to gain \$1, then if you had to choose between a 50-50 chance of winning or losing \$10,000 vs a 50-50 chance of winning or losing \$100, you may exhibit a preference for the smaller-stakes lottery.

12 Prospect Theory

A major alternative to EU in the behavioral economics literature is Prospect Theory, which is due to Kahneman and Tversky. As we have seen, these researchers conducted experiments to explore properties of people's choices under uncertainty. They used their findings to construct Prospect Theory.

12.1 Model

Prospect Theory posits the existence of a utility representation for preferences over lotteries that can accommodate a variety of experimental findings on risk preference. It takes some work to describe the theory, and indeed it is not as elegant (meaning simple to describe, highly tractable, easy to put into applications) as EU theory. This reflects a general feature of theories: the more elegant ones tend to be less realistic and the more realistic ones tend to be less elegant. The real breakthrough is when one finds a theory that is both elegant and more realistic than the incumbent theory. The Prospect theory does not fit the bill in this regard, and it is no surprise that it has not been adopted in mainstream economics. Nevertheless, it is the standard model in the field of behavioral economics.

Prospect theory states that subjects have a preference $\succsim$ over lotteries that is represented by a utility function of a particular form. This utility function is hard to describe, and understanding it fully requires understanding a different theory of choice that was developed after the EU theory (called Rank-Dependent Expected Utility). Therefore we will focus on a special case of the Prospect Theory, which nevertheless captures its essential features. Specifically, instead of considering a preference $\succsim$ over all kinds of lotteries, we will consider a preference $\succsim$ over monetary lotteries of the form

$$(p_1, x; p_2, y; p_3, 0)$$

where $x \geq 0 \geq y$. That is, the lotteries we consider have at most one positive outcome x and at most one negative outcome y. To ease notation somewhat, we write this lottery more simply as

$$(p_1, x; p_2, y).$$

Of course, if $p_1+p_2<1$, then you should infer that the remaining probability $1-p_1-p_2$ is associated with 0. For most purposes, we will use lotteries that yield a positive outcome with some probability and 0 otherwise.

Now we present the model.

An agent's preference $\succsim$ over the set of lotteries of the form $(p_1,x;p_2,y)$ respects Prospect theory if his preference $\succsim$ has a utility representation of the form

$$V(p_1,x;p_2,y)=\varphi(p_1)v(x)+\varphi(p_2)v(y),$$

where the v is a utility function over outcomes (called the 'value function') and φ is a function over the interval $[0,1]$ (called the 'probability weighting function'); both these functions satisfy the particular properties stated below.

As you can see, the utility of a lottery is an average, like in the EU theory. The major differences (discussed below) are that, first, utility v is defined for gains and losses rather than for absolute wealth, and second, the agent is assumed to use 'decision weights' $\varphi(p_i)$ (which are in fact distorted probabilities) rather than actual probabilities.

Properties of the Value Function v

(i) *Reference Dependence*: v is defined for gains and losses with respect to some underlying reference point, and $v(0)=0$. For instance, if you are told that the agent has initial wealth w, faces a gamble that leaves him with $w+100$ with probability 0.5 and $w-100$ with probability 0.5, and that current wealth w is the agent's reference point, then you should compute utility as $\varphi(0.5)v(100)+\varphi(0.5)v(-100)$ rather than $\varphi(0.5)v(w+100)+\varphi(0.5)v(w-100)$. That is, the relevant outcomes of the gamble are not $w+100$ and $w-100$ but rather the *change* with respect to the reference point, that is, $+100$ and -100. Note that the only difference between the value function v and the utility index u used in the EU theory is that the former is defined only over gains and losses whereas the latter may be defined for absolute outcomes.

(ii) *Diminishing Sensitivity*: v exhibits diminishing marginal utility for gains and diminishing marginal disutility for losses. On a graph, v has an S-shape and passes through the origin.

(iii) *Loss Aversion*: $v(x) < -v(-x)$ for all positive x. Typically it is presumed that v is kinked at the origin. This accentuates how quickly disutility increases for losses. Loss aversion is usually captured in the following way. Given a function $v(x)$ defined only on gains $x \geq 0$, the corresponding function for losses is defined by specifying for all $x < 0$

$$v(x) = -\kappa v(-x).$$

You can check that a function so defined exhibits $v(x) < -v(-x)$ for all positive x when $\kappa > 1$.

The functional form for v proposed by Kahneman and Tversky is

$$v(x) = \begin{cases} \frac{x^\alpha}{\alpha} & x \geq 0 \\ -\kappa \frac{(-x)^\alpha}{\alpha} & x < 0 \end{cases}, \tag{2}$$

and based on evidence, they suggest $\alpha = .88$ and $\kappa = 2.25$.

Properties of the Probability Weighting Function φ

(i) *Basic Properties*: For each $0 \leq \alpha \leq 1$, we have $0 \leq \varphi(\alpha) \leq 1$, that is, the decision weight corresponding to α is a number $\varphi(\alpha)$ between 0 and 1. Moreover, $\varphi(0) = 0$, $\varphi(1) = 1$ and $\varphi(\alpha)$ increases strictly with α.

(ii) *Overweighting-Underweighting*: Small probabilities are overweighted (that is, $\varphi(\alpha) > \alpha$ for small α) and large probabilities are underweighted (that is, $\varphi(\alpha) < \alpha$ for large α). Some studies have tried to estimate the weighting function and have found that $\varphi(\alpha) = \alpha$ approximately around $\alpha = \frac{1}{3}$.

The functional form proposed by Kahneman and Tversky is

$$\varphi(\alpha) = \frac{\alpha^\delta}{\left(\alpha^\delta + (1-\alpha)^\delta\right)^{1/\delta}}, \quad \text{where } \delta = 0.65.$$

12.2 Accommodating the Evidence

We now show briefly and informally how the Prospect Theory can accommodate some of the findings we have already discussed. In each case, pay very careful attention to precisely what property of the theory is determining

the preference. You should not be surprised that the Prospect Theory can accommodate these findings: it was constructed to fit these choice patterns.

– Certainty Effect
Recall the following choices that violate the Independence axiom:

$$\begin{aligned} (1, 3000) &\succ (0.8, 4000;\ \ 0.2, 0) \\ (0.25, 3000;\ \ 0.75, 0) &\prec (0.2, 4000;\ \ 0.8, 0). \end{aligned}$$

The key observation here is that the \$4000 didn't matter as much when the probability of receiving it was high (that is, 0.8). This is captured by underweighting of large probabilities. The agent gave it less importance when the probabilities were high (first comparison) than when they were low (second comparison).

– Fourfold Pattern of Risk Attitudes
Recall the preferences for the following large-probability lotteries:

$$\begin{aligned} (1, 3000) &\succ (0.8, 4000;\ \ 0.2, 0) \\ (1, -3000) &\prec (0.8, -4000;\ \ 0.2, 0). \end{aligned}$$

Underweighting of large probabilities means that agents did not place too much weight on the chance of getting \$4000 in the first comparison, and on the chance of losing \$4000 in the second. This led to risk aversion in the first comparison but risk affinity in the second. Similarly, overweighting of small probabilities can make sense of the following small-probability lotteries:

$$\begin{aligned} (0.001, 5000;\ \ 0.999, 0) &\succ (1, 5) \\ (0.001, -5000;\ \ 0.999, 0) &\prec (1, -5). \end{aligned}$$

This is an elegant explanation for the four-fold pattern.

- Loss Aversion.
Consider the following instance of loss aversion:

$$(\frac{1}{2}, 100; \frac{1}{2}, -100) \succ (\frac{1}{2}, 1000; \frac{1}{2}, -1000).$$

This behavior is consistent with the properties of the value function. For a simple example, consider the Prospect Theory with a value function that is given by $v(x) = \sqrt{x}$ and $v(-x) = -\lambda\sqrt{x}$ for all positive x where $\lambda > 1$. You can check that this satisfies all the properties of a value function. Now observe that

$V(\frac{1}{2}\delta_{100} + \frac{1}{2}\delta_{-100}) = \varphi(\frac{1}{2})v(100) + \varphi(\frac{1}{2})v(-100) = -\varphi(\frac{1}{2})(\lambda - 1)\sqrt{100}$, and

$V(\frac{1}{2}\delta_{1000} + \frac{1}{2}\delta_{-1000}) = \varphi(\frac{1}{2})v(1000) + \varphi(\frac{1}{2})v(-1000) = -\varphi(\frac{1}{2})(\lambda - 1)\sqrt{1000}$.

It is evident that $V(\frac{1}{2}\delta_{100} + \frac{1}{2}\delta_{-100}) > V(\frac{1}{2}\delta_{1000} + \frac{1}{2}\delta_{-1000})$, which is consistent with the preferences. The key feature here is the loss aversion property of the value function: $v(x) < -v(-x)$ for all positive x.

Part IV
CHOICE UNDER SUBJECTIVE UNCERTAINTY

13 Modelling Uncertainty

EU theory presumes that risky options come with probabilities specified. Probabilities of outcomes can be calculated in the casino, but there are few other settings where this is possible. What is the probability of the stock market going up tomorrow? You may try to arrive at a probability, say by quantifying the feeling about the likelihood that you get from reading the news, or by asking a financial expert to use past data to suggest a probability. But such probabilities are one's personal guess, that is, they are subjective probabilities. They are not objective probabilities in the sense that one can compute the probability of winning at the slot machine.

The theory of Expected Utility studied in earlier chapterns is a theory of agents choosing between objectively risky actions – it is a theory of how people choose among lotteries, and lotteries are risky actions with known probabilities. In this section we outline a version of EU theory built for the more natural setting where the risk exists but the probabilities are not objectively computable. It is called *Subjective Expected Utility* (SEU) theory and it is due to Leonard Savage [CITE].[32]

The first step is to formally describe an "uncertain prospect", which we shall call an *act*, or a *Savage act*.

13.1 States of the World

To define an act, we first need to specify two sets, denoted X and S. The set X is the set of outcomes, also called consequences. For instance, if outcomes involve money values, then X would be the set of real numbers. The second set S is the set of possible *states of the world*, also called states of nature. A state of the world is a complete description of the aspects of the world that are relevant to the agent's decisions. To illustrate, suppose you are considering whether or not to enter a bet with your friend, that the Red Sox will win the world series – you get \$20 from him if you win, and you give him \$20 otherwise. The aspect of the world that is relevant for you is the

[32] On a historical note: von Neuman-Morgenstern Expected Utility predates Savage Expected Utility theory. In keeping with the economic theory literature, we refer to the former as EU and the latter as SEU.

result of the world series. The set of states of the world is the set of *relevant* possible scenarios, that is, $S = \{\text{Red Sox win, Red Sox lose}\}$.

13.2 Choice Domain: Acts

Now we can define an act: an act is a *specification* of what outcome would arise in every state of the world. For instance, the above bet is an act. The outcomes are monetary (so X can be taken as the set of real numbers) and the set of states is $S = \{\text{Red Sox win, Red Sox lose}\}$. The bet is given by the act:

$$(\text{Red Sox win}, w + 20; \text{Red Sox Lose}, w - 20),$$

where w is your initial wealth. Note that this act specifies the outcome in each state of the world: if the state "win" obtains, then the outcome is $w + 20$ dollars, and if state "lose" obtains, then the outcome is $w - 20$. More generally, for some set of consequences X, and some set of states of the world $S = \{s_1, s_2, ..., s_n\}$, an act is denoted

$$(s_1, x; s_2, y; .....; s_n, z),$$

where $x, y, ..z$ are consequences (elements of X). This act specifies that the agent will get outcome x if state s_1 occurs, y if s_2 occurs, etc. Put differently, it maps S into X. An act can also be written as

$$\begin{bmatrix} s_1 & x \\ s_2 & y \\ .. & .. \\ s_n & z \end{bmatrix}.$$

Denote by A the set of all acts that map S into X is denoted.

14 Subjective Expected Utility Theory

Subjective Expected Utility, due to Savage [CITE], can be described as follows.

Primitives:

i) *A set X of consequences.*

ii) *A set S of states of the world (or a state space, for short).*

iii) *A preference $\succsim$ over the set of acts A (that map S into X).*

Hypothesis for $\succsim$:

The preference $\succsim$ admits a utility representation of the form

$$SEU_{p,u}(s_1, x_1; s_2, x_2; .....; s_n, x_n) = p(s_1)u(x_1) + p(s_2)u(x_2) + ...p(s_n)u(x_n),$$

where p is a probability measure[33] *over S (called the prior beliefs) and u is the agent's utility from outcomes (called the utility index).*

Hypothesis for Choice: *Preference maximization.*

That is, the Subjective Expected Utility (SEU) theory hypothesizes that, when agents are uncertain about which state of the world will obtain, *they form beliefs about S*; they assign probabilities to each state of the world. These probabilities are subjective in the sense that they are in his head, and not necessarily given to him objectively. When faced with some act $(s_1, x; s_2, y; .....; s_n, z)$, the agent uses his beliefs over S to view the act as a "lottery" that yields outcome x_i with probability $p(s_i)$. He then computes a belief-weighted average of the utility from possible outcomes $p(s_1)u(x) + p(s_2)u(y) + ...p(s_n)u(z)$. Finally, he maximizes this utility to determine his choice from a menu.

14.1 Choices Reveal Beliefs

Beliefs are unobservable – you cannot "see" another person's beliefs. But beliefs can be identified from behavior in SEU theory. The basic idea is that

[33] We say that p is a probability measure over S if it specifies a number $0 \leq p(s) \leq 1$ for each state $s \in S$, and if these numbers sum to one $\sum_{s \in S} p(s) = 1$.

an agent considers state s more likely than another state s' *if he would bet on* s *rather than* s'. This is very intuitive: would you bet on something that you consider to be unlikely? To see how SEU theory incorporates this insight, suppose there are (say) three states of the world, $S = \{s_1, s_2, s_3\}$, and you want to know whether a person believes that state s_1 orstate s_2 is more likely. SEU theory would suggest that you ask this person to choose between the following two acts:

$$\begin{bmatrix} s_1 & 100 \\ s_2 & 0 \\ s_3 & 0 \end{bmatrix} \text{ vs } \begin{bmatrix} s_1 & 0 \\ s_2 & 100 \\ s_3 & 0 \end{bmatrix}.$$

The first act is a bet on s_1 – you get \$100 if s_1 obtains, and 0 otherwise. Similarly, the second act is a bet on s_2. Suppose the agent chooses the second act. Then, assuming that the agent prefers more money to less, SEU theory tells us that:

$$\begin{bmatrix} s_1 & 0 \\ s_2 & 100 \\ s_3 & 0 \end{bmatrix} \succ \begin{bmatrix} s_1 & 100 \\ s_2 & 0 \\ s_3 & 0 \end{bmatrix}$$

$$\iff \quad p(s_1)u(0) + p(s_2)u(100) + p(s_3)u(0) > p(s_1)u(100) + p(s_2)u(0) + p(s_3)u(0)$$

$$\iff \quad p(s_1)u(0) + p(s_2)u(100) > p(s_1)u(100) + p(s_2)u(0)$$

$$\iff \quad p(s_2)u(100) - p(s_2)u(0) > p(s_1)u(100) - p(s_1)u(0)$$

$$\iff \quad p(s_2)\left[u(100) - u(0)\right] > p(s_1)\left[u(100) - u(0)\right]$$

$$\iff \quad p(s_2) > p(s_1).^{34}$$

That is, via betting preference we can identify that the agent believes that s_2 is more likely than s_1. (We did not assume $u(0) = 0$ here since it was not necessary to get the result, but you can see that the proof becomes substantially simpler if we make that assumption).

To be clear, in the preceding we did not talk about identifying p, but rather only judgements about relative likelihoods underlying p. This is not enough to identify the exact probability $p(s)$ of each state. For instance,

[34] This last implication follows because we are assuming that $u(100) - u(0) > 0$, that is, more money makes the agent happier. We can always use the agent's choices to check if he satisfies this. See exercise 1.

if there are only two states and s_1 is identified to be more likely than s_2 then there are infinitely many p's that are consistent with this. Identifying p requires the data to be richer and more besides. We eschew a discussion of this.

14.2 A Testable Implication

Suppose that tomorrow it can either be sunny (s), rainy (r) or cloudy but dry (c). Suppose that you are given a choice between the following acts:

$$\begin{bmatrix} s & \text{vanilla ice cream} \\ r & \$100 \\ c & 0 \end{bmatrix} \quad \text{vs} \quad \begin{bmatrix} s & \text{vanilla ice cream} \\ r & \$0 \\ c & \$100 \end{bmatrix}.$$

Notice that if the state is S then you get vanilla ice cream in either act. Otherwise, the choice effectively comes down to whether you want to bet on rain or no-rain in the event that it is cloudy tomorrow.

The question for you is: whatever your choice is, would it change if you got chocolate ice cream in each act rather than vanilla? If you say that the flavor of the ice cream does not matter for your choice, then you are behaving in accordance to the "Sure-Thing Principle":

Proposition 10 *If a preference $\succsim$ over acts satisfies SEU theory, then it satisfies the Sure-Thing Principle: for any outcomes $x'..x'', y'..y''$ and z, w,*

$$\begin{bmatrix} s & z \\ s' & x' \\ \vdots & \vdots \\ s'' & x'' \end{bmatrix} \succ \begin{bmatrix} s & z \\ s' & y' \\ \vdots & \vdots \\ s'' & y'' \end{bmatrix} \Longrightarrow \begin{bmatrix} s & w \\ s' & x' \\ \vdots & \vdots \\ s'' & x'' \end{bmatrix} \succ \begin{bmatrix} s & w \\ s' & y' \\ \vdots & \vdots \\ s'' & y'' \end{bmatrix},$$

and similarly with indifference.

Proof. Exercise. ■

The Sure-Thing Principle (STP for short) states that if two acts yield a common consequence in some state (for instance z in s), then the preference does not change if that common consequence is changed to some other

common consequence (for instance w in s). The fact that the preference does not change with the common consequence means simply that the agent essentially ignores common consequences when evaluating preferences.[35]

[Add illustrative example on STP, counterfactuals can matter).

[35] The reader will note that STP is the counterpart of Expected Utility's Common Consequence Independence translated to the subjective setting.

15 Ambiguity Aversion

15.1 Risk vs Ambiguity

Daniel Ellsberg conducted the following famous experiment. After describing the experiment and its results, we will relate it to the SEU theory.

Subjects were told that there is an urn containing 90 balls, all identical except for color. Furthermore, they were told that there are exactly 30 red balls in the urn, and the remaining balls are black or yellow – the exact proportion was not specified, so there could be anywhere from 60 black and 0 yellow, to 0 black and 60 yellow balls. One ball was going to be randomly selected from the urn. The subjects were asked to choose between:

(i) receiving \$100 if the ball is red, \$0 otherwise.

(ii) receiving \$100 if the ball is black, \$0 otherwise.

Next they were to choose between:

(iii) receiving \$100 if the ball is either red or yellow, \$0 otherwise.

(iv) receiving \$100 if the ball is either black or yellow, \$0 otherwise.

The typical preferences were

$$\begin{aligned} (i) &\succ (ii) \\ (iii) &\prec (iv). \end{aligned}$$

This preference pattern is inconsistent with the the SEU theory because it violates the STP. To see this, note first that the state space corresponds to the possible colors of the selected ball, that is, $S = \{R, B, Y\}$, where R denotes the state of the world in which the selected ball is Red, etc. With this understanding, each of the bets respectively define the following acts:

$$\begin{bmatrix} R & 100 \\ B & 0 \\ Y & 0 \end{bmatrix}, \begin{bmatrix} R & 0 \\ B & 100 \\ Y & 0 \end{bmatrix}, \begin{bmatrix} R & 100 \\ B & 0 \\ Y & 100 \end{bmatrix}, \begin{bmatrix} R & 0 \\ B & 100 \\ Y & 100 \end{bmatrix}.$$

It is then evident that the preference pattern violates the STP. This preference pattern is known as the Ellsberg Paradox.

The Ellsberg Paradox is actually pretty bad news, because it contradicts not only a tractable model of choice under risk, but it also contradicts the

hypothesis that *people assess likelihoods in the form of probability judgments.* Recall from our earlier discussion on eliciting beliefs that beliefs underlie choices between bets. Note that the preference

$$\begin{bmatrix} R & 100 \\ B & 0 \\ Y & 0 \end{bmatrix} \succ \begin{bmatrix} R & 0 \\ B & 100 \\ Y & 0 \end{bmatrix}$$

implies that the agent believes it more likely that the ball with be R than B. On the other hand, the preference

$$\begin{bmatrix} R & 100 \\ B & 0 \\ Y & 100 \end{bmatrix} \prec \begin{bmatrix} R & 0 \\ B & 100 \\ Y & 100 \end{bmatrix}$$

implies that the agent believes it more likely that the ball will be B or Y than R or Y. Now, *if* the agent has a probabilistic belief p over S, these likelihood assessments would imply

$$p(R) > p(B) \text{ and } p(B) + p(Y) > p(R) + p(Y).$$

However, this is impossible – there is no probability measure that can satisfy these inequalities.

The Ellsberg Paradox demonstrates that the existence of different orders of uncertainty. Facing uncertainty simply means that the outcome of an act is uncertain, that is, the outcome depends on a state of the world that occurs with probability less than 1. Notice that the probability of state R is certain: it is known to be $1/3$ for sure. We say that R is a *risky* state of the world. But notice that the probability of state Y is *not* certain: it can be between 0 and $2/3$. The probability of a state already encodes uncertainty, but when *this probability is itself uncertain*, then we are talking about a higher degree of uncertainty than just risk. In this situation we say that there exists *ambiguity*, and that Y is an ambiguous state of the world. You can imagine even higher orders of uncertainty. For instance, imagine a version of the Ellsberg experiment where the ball will either be drawn out of the above urn, or out of an urn where there are 30 yellow balls the remaining are red

or blue, but you do not know anything about which urn will be used. In this case, the probability of Y can either be ambiguous (between 0 and 2/3), or it can be unambiguous (exactly 1/3), but there is ambiguity about which of these scenarios holds. That is, there is ambiguity about the ambiguity about the probability of yellow.

The uncertainty that one is faced with in the casino is uncertainty that is amenable to calculation of probabilities, and is therefore known or at least in principle knowable. Thus people face risky choices in the casino. Then there is the stock market where the sources of uncertainty are so vast that one may not even be able to conceive all of them. In the stock market, analysts only make guesses about the probability of the stock index going up or down in the future. They may *assume* that past frequencies will predict future frequencies, but no one can be fully confident about the estimated probability, the way one can be about the probabilities in a casino. Thus people face ambiguity in the stock market. When it is business as usual in the economy, then this ambiguity may be small, but when there is, say, political instability, then this ambiguity magnifies.

It is instructive at this point to ask "why exactly did SEU fail in the Ellsberg experiment"? The answer is that the SEU agent is one who is certain about her beliefs: she will give you a sharp number if asked what she thinks is the probability of a given state. With her certainty about her beliefs, all the uncertainty she faces collapses to just being risk. People do not usually experience that kind of certainty about their beliefs in the absence of complete information, as in the Ellsberg experiment. Unlike the SEU agent, they do not carry around that much confidence in things that they do not know so well.

15.2 Max-Min Expected Utility Theory

The Ellsberg Paradox gave birth to a literature on ambiguity in economics. Clearly, the distinctive feature of the Ellsberg experiment is that there is not enough information about the contents of the urn for the agent to be able to form a coherent belief – the subjects were faced with ambiguity in that likelihoods could not be captured or quantified by means of a probability

measure. We present here a model of choice under ambiguity.

15.2.1 Model

The primitive is a preference $\succsim$ over acts, as in the SEU theory. Recall that SEU is computed by the formula

$$SEU_{p,u}(s_1, x; s_2, y; .....; s_n, z) = p(s_1)u(x) + p(s_2)u(y) + ...p(s_n)u(z).$$

SEU is indexed by the prior p and utility u that is used in the formula. The MaxMin Expected Utility (MEU) model of ambiguity defined below is an extension of the SEU model. While in the SEU model, the agent has one utility index u and one prior p, in the MEU model the agent has one u but *multiple priors*, that is, a set of priors. The fact that the agent admits more than one prior as a possible expresses that she perceives ambiguity: she does not have a precise conception of likelihoods. How does she rank acts with multiple priors? The MEU theory says that for any given act, she computes $SEU_{p,u}$ with respect to the *most pessimistic prior* she admits as possible. That is, the MEU from an act is the minimum value of $SEU_{p,u}$ that is attainable by varying p over her set of priors.

Formally, the theory is given by:

Primitives:

i) *A finite set X of consequences.*

ii) *A finite set S of states of the world (or a state space, for short).*

iii) *A preference $\succsim$ over the set of acts A (that map S into X).*

Hypothesis for $\succsim$:

An agent's preference $\succsim$ over the set of acts A respects the MaxMin Expected Utility theory if the preference $\succsim$ has a utility representation of the form

$$MEU(s_1, x; s_2, y; .....; s_n, z) = \min_{p \in P}\{SEU_{p,u}(s_1, x; s_2, y; .....; s_n, z)\},$$

where P is a set of probability measures over S, and u is the agent's utility index.

Hypothesis for Choice: *Preference maximization.*

15.2.2 Example

We illustrate the model by showing how it accommodates the Ellsberg paradox. Consider the Ellsberg experiment: suppose u is strictly increasing with $u(0) = 0$, and that P is the set of all probability measures over $\{R, B, Y\}$ that satisfy

$$p(R) = \frac{1}{3} \text{ and } p(B) + p(Y) = \frac{2}{3}.$$

That is, P consists of all probability measures that are consistent with the information available to the agent in the experiment. We compute the MEU of each act in the experiment.

Compute that

$$\begin{aligned} SEU_{p,u}(R, 0; B, 100; Y, 0) &= 0 + p(B)u(100) + 0 \\ &= p(B)u(100). \end{aligned}$$

Indeed, the SEU of $(R, 0; B, 100; Y, 0)$ depends on which p is used. Observe that $p(B)$ ranges between 0 and $\frac{2}{3}$ across all the priors p in P. Indeed, the minimum possible value of $SEU_{p,u}(R, 0; B, 100; Y, 0)$ is 0:

$$\begin{aligned} MEU(R, 0; B, 100; Y, 0) &= \min_{p \in P} \{SEU_{p,u}(R, 0; B, 100; Y, 0)\} \\ &= \min_{p \in P} \{p(B)u(100)\} \\ &= 0 \cdot u(100) = 0. \end{aligned}$$

Similarly, compute that

$$\begin{aligned} SEU_{p,u}(R, 100; B, 0; Y, 0) &= p(R)u(100) \\ &= \frac{1}{3}u(100). \end{aligned}$$

Observe that the SEU of this act is the same regardless of what p in P is used. Intuitively, this is because there is no ambiguity about the probability of R. Technically this is because all p in P assign the same probability $p(R)$

to R. Thus,

$$\begin{aligned} MEU(R,100;B,0;Y,0) &= \min_{p\in P}\{SEU_{p,u}(R,100;B,0;Y,0)\} \\ &= \min_{p\in P}\{\frac{1}{3}u(100)\} \\ &= \frac{1}{3}u(100). \end{aligned}$$

Observe that our calculations yield that $MEU(R,100;B,0;Y,0) > MEU(R,0;B,100;Y,0)$, and thus

$$\begin{bmatrix} R & 100 \\ B & 0 \\ Y & 0 \end{bmatrix} \succ \begin{bmatrix} R & 0 \\ B & 100 \\ Y & 0 \end{bmatrix},$$

as observed in the experiment. Intuitively, the agent knows how to compute the SEU of the bet on R, but assumes that the worst outcome will obtain if she bets on the ambiguous B – she will assume that there are no B colored balls in the urn. Consequently, she will prefer to bet on R than on B.

Determining the other preference in the Ellsberg experiment is left to you.

Exercise: To make sure you understand the model, compute $MEU(R,100;B,75;Y,25)$. The correct answer is not $[0.3\times 100]+[0\times 75]+[0\times 25]$, which is clear from the fact that the probabilities satisfy $0.3+0+0\neq 1$.

16 Case-Based Decision Theory

The outcome of an action depends on the 'state of the world.' For instance, the action of 'going to school without an umbrella' will lead the outcome 'dry' if the state of the world is 'sunny,' or the outcome 'wet' if the state of the world is 'rainy.' In standard economics, an agent will first work out a probability (the prior) over the different possible states of the world, and then compute an expected utility of the possible outcomes with respect to this probability. This is called the Subjective Expected Utility theory – it differs from the Expected Utility theory only in that probabilities are subjectively formed, whereas in the Expected Utility theory probabilities are objectively known, since actions involve a choice of lotteries. However, in the real world, the agent may never have a clear idea of what all the possible states of the world are (e.g. it can be complicated to work out all the possible scenarios that will determine the outcome of buying a house). Sometimes they may not even have a clear idea of what the possible outcomes of an action might be. Neither may they sit and try to form a prior about the relevant uncertainty. The Case-Based Decision Theory (CBDT) is a theory of how people choose actions in such situations. The basic idea is that people use their knowledge of past cases/experiences (their own or others') to construct a utility over actions, given the problem at hand.

16.1 Model

A *case* is defined as a triple (P, a, r) where P denotes a *problem*, a an *action*, and r an *outcome* of taking action a in problem P. Denote the set of actions by A. The primitives of the model are (i) a preference $\succsim_P$ over actions A *for each problem* P, and (ii) a *set* of cases $M = \{(P_1, a_1, r_1), (P_2, a_2, r_2), ...\}$ that consist of the agent's *memory*, that is, all the cases that he knows about.

The Model: The agent has a utility u over outcomes. He also possesses a *similarity function* s that assigns a number $s(P, Q)$ between 0 and 1 to each pair of problems P and Q. The higher the value of $s(P, Q)$, the greater the similarity between the two problems. These are used to construct, for each problem P, a utility representation over actions for preference $\succsim_P$. The

utility he constructs for some action a given problem P is as follows:

$$U(a|P) = s(P, Q_1)u(r_1) + s(P, Q_2)u(r_2) + ...s(P, Q_n)u(r_n),$$

which is calculated on the basis of all the *relevant* cases in his memory, that is, all the cases in his memory M of the form (Q_i, a, r_i) in which he took action a. He compares the problem in each case, for instance case Q_i, with the current problem P, assesses the similarity $s(P, Q_i)$ and then weights the utility $u(r_i)$ of the outcome r_i with $s(P, Q_i)$. Then, he takes the sum of the weighted utilities over all the relevant cases to compute $U(a|P)$. If there are no relevant cases, then he sets $U(a|P) = 0$.

16.2 Example

A guy is faced with the problem of whether or not to ask a girl, called P, out on a date. His actions are a (ask) and d (don't ask). The possible outcomes are x (he goes out on a date), y (he spends the evening playing video games), and z (he spends the evening playing video games and feeling like a nerd). His utility u from the outcomes is given by

$$u(x) = 100,\ u(y) = 0 \text{ and } u(z) = -10.$$

His memory consists of the following cases:

$$M = \{(Q_1, d, y), (Q_2, d, y), (Q_3, a, z), (Q_4, a, x)\}.$$

So, for instance, in the case (Q_3, a, z) he remembers asking out a girl called Q_3, and getting rejected. In addition to his memory, he can compare types of girls by means of a similarity function s. According to this function, each of the girls in his memory compares to P in the following way:

$$\begin{aligned} s(P, Q_1) &= 1 \\ s(P, Q_2) &= 0.5 \\ s(P, Q_3) &= 0.25 \\ s(P, Q_4) &= 0. \end{aligned}$$

Thus Q_1 is most similar to P and Q_4 is least similar.

Does he ask P out? According to the Case-Based Decision theory, the agent computes his utilities over actions as follows:

$$\begin{aligned} U(a|P) &= s(P,Q_3)u(z) + s(P,Q_4)u(x) = 0.25 \times -10 + 0 \times 100 = -2.5 \\ U(d|P) &= s(P,Q_1)u(y) + s(P,Q_2)u(y) = 1 \times 0 + 0.5 \times 0 = 0. \end{aligned}$$

Thus, he chooses action d since $U(d|P) > U(a|P)$.

His experiences of not asking out Q_1 and Q_2 lead to some level of utility $U(d|P)$ from option d. But, the bad experience of asking out Q_3 (and the irrelevance of the great experience with Q_4) makes a a relatively unattractive option, even though P is very dissimilar to Q_3 and Q_4.

Observe that this model views agents as essentially backward-looking. This is in contrast with the SEU model, where agents are forward-looking in that they think about all the relevant contingencies and form expectations about them.

Part V
JUDGEMENT UNDER UNCERTAINTY

18 Modelling Uncertainty and Information

SEU theory expresses the idea that agents in an uncertain context will construct beliefs about the uncertainty to guide their choices. The topic "Judgement under Uncertainty" is concerned with understanding the properties of beliefs that underlie people's choices. There are two key questions. First, what are the basic properties of likelihood judgements under uncertainty at one point in time? Second, if the agent receives some information about the uncertainty she is facing, then how will she update her likelihood judgements?

Before we study the nature of an agent's likelihood judgements, we must first be clear how to mathematically capture the environment independently of the agent. Specifically, how to we model uncertainty and how do we model information? The first question has already been answered when we discussed SEU but we repeat the answer here and then move on to modelling information.

18.1 Modelling Uncertainty

The uncertainty in the outcome of our choice ultimately arises due to uncertainty that, from our perspective, exists in the world. For instance, if you choose to leave your apartment on an overcast day without an umbrella, your outcome – whether you remain dry or get wet on your way to school – depends on whether it rains or not. Or, if you are meeting a friend for lunch, whether you have a good time or not depends on how good your and your friend's mood is. Or for a more standard economic example, a firm's future profits depend on factors such as market demand. In any of these cases, the uncertainty being faced is rooted in uncertainty in your environment. The uncertainty about your own mood is in some sense part of the environment, as it belongs to a realm that is outside your full control.

Modelling uncertainty comes down to describing the possible ways that "the world" can look, where we limit attention to only that part of the world that is relevant. For instance, in the first example above, the world could be one where it rains today, or it could be one where it does not rain today. A description is called a *state of the world* or *state of nature*. The *state space* is the set of *mutually exclusive and collectively exhaustive descriptions of the*

world. In general, the state space is given by some abstract set

$$S = \{s, s', ...\}.$$

We will restrict attention to finite state spaces throughout for simplicity.

Very often in Economics we run into multi-dimensional state spaces. For instance, a firm's profits may depend on the market price and the wage rate in the labor market. Then any description of the world therefore has two descriptions in it: one of market price and one of wage rate. We write such a 2-dimensional description as a 2-dimensional vector, such as $(high_{price}, high_{wage})$. This describes one state of the world. To specify the state space, we first need to specify the possible descriptions in each dimension. For instance:

$$S_{price} = \{high_{price}, low_{price}\}, \quad S_{wage} = \{high_{wage}, low_{wage}\}.$$

Then the state space is the set of all such vectors formed from these:

$$S = \{(high_{price}, high_{wage}), (high_{price}, low_{wage})(low_{price}, high_{wage})(low_{price}, low_{wage})\}.$$

A set of vectors composed in such a way is called a product set, and is denoted

$$S = S_{price} \times S_{wage}.$$

In general, if there are $n > 1$ dimensions, and the uncertainty pertaining to each dimension i is described by some set S_i, then the state space is the product set:

$$S = \prod_{i=1}^{n} S_i = S_1 \times .... \times S_n,$$

and a generic state is denoted $s = (s_1, ..., s_n)$.

18.2 Partitional Information Structures

The first kind of information that we consider is of the simplest form: the information that only *rules out* states.

18.2.1 Definition

As a running example, suppose an agent wakes up in the middle of the night and wonders if the moon is full (f), partial (h, for half) or not out (n) tonight. These mutually exclusive descriptions of the relevant uncertainty give us a state space

$$S = \{f, h, n\}.$$

These are the states of the world, but only one of them is the *true state*, and the agent really wants to get as close as possible to knowing which one it is.

Suppose the agent lives near a forest where there is a wolf that howls if and only if the moon is full. So, if the agent hears a howl, she can *rule out* the possibility that the state is h or n, and therefore deduce that the true state is f. The howl can be denoted in terms of its conclusion about the state, namely that the true state lies in $\{f\}$. If the wolf does not howl, the agent can rule out the possibility of f, and deduce that the true state is either h or n. Therefore, "not howling" corresponds to learning that the true state is in $\{h, n\}$. By howling or not howling, the wolf splits up the state space into two mutually exclusive subsets, which we will write as

$$\mathcal{I}_W = \{\{f\}, \{h, n\}\}.$$

The set $\mathcal{I}_W$ specifies what all the agent could *possibly* learn from the wolf.

More formally, refer to any subset $E \subset S$ as an *event in* S, and interpret it as the event that states outside E have been ruled out, or equivalently, the true state lies in E. A *partitional information structure*, or *information structure* for short, is a set

$$\mathcal{I} = \{E_1, E_2, ...E_m\},$$

consisting of events which *partition* the state space S in the sense that (i) they cover all of S, that is, $E_1 \cup ... \cup E_m = S$, and (ii) they are mutually exclusive, that is, pair-wise disjoint in that $E_i \cap E_j = \phi$ for each i, j.

Exercise: What would an uninformative (that is, completely useless) information structure look like?

18.2.2 Combining Information

Suppose, in the same forest, there is also an owl that hoots if and only if the moon is out (regardless of whether it is full or partial). The owl is therefore chattier than the wolf since it hoots when the moon is full or partial, while the wolf howls only when it is full. By checking if the owl is hooting or not, the agent obtains the partitional information structure:

$$\mathcal{I}_O = \{\{f, h\}, \{n\}\}.$$

Must we treat the agent as if she has two separate information structures, $\mathcal{I}_W$ and $\mathcal{I}_O$? No, not if the agent is smart and can draw inferences correctly (and awake enough in this example). By combining the information she gets from both wolves, the agent generates for herself an information structure that is "finer" than both $\mathcal{I}_W$ and $\mathcal{I}_O$:

$$\mathcal{I}_{WO} = \{\{f\}, \{h\}, \{n\}\}.$$

Why? Well, if the true state is f, then the wolf will howl and she will learn $\{f\}$. If the true state is h, then the owl will hoot but the wolf will be silent, and the agent can infer that the true state must lie in$\{h\}$: the owl's hoot will rule out n and the wolf's silence will rule out f, leaving h as the only remaining possibility. Finally, if the true state is n, it will be a quiet night, and in particular the owl's silence will confirm that the true state lies in $\{n\}$. This confirms that regardless of what the true state is, the agent will be able to infer it. In this example, the agent obtains a perfect partition of the state space when she combines her sources of information, but that does not always have to be the case. It is necessarily true that the agent's combined information is (weakly) finer than her individual sources. But if there is no source that can separate two states, then combining the information will not yield a perfect partition.

We should make clear that the preceding relies on an independence assumption: the wolf's howling is in no way related to the owl's hooting, except indirectly through the moon. The picture would be different if the wolf howled if, for instance, the moon was full or if the owl hooted. In this case the two partitional information structures are not independent.

Maintaining the assumption of rational processing of information by agents, economists just care about the finest partition that the agent has constructed, ignoring the details of how she has constructed it. We will do the same.

18.3 Signal Structures

In the real world, information is rarely as stark as a perfectly confident "these states are ruled out". Whether it is in the form of a weather forecast, or of a financial expert telling us what stocks are likely to go up, we often receive signals that provide us with confidence in one state or the other, but without necessarily ruling out any state. In Economics, information is often modelled as a *signal structure*, also known as an *information structure,* or an *experiment*, or a *Blackwell experiment*.

18.3.1 Definition

To illustrate, recall the agent who wakes up in the middle of the night wondering how the moon looks. Suppose that the information provided by the wolf is no longer partitional: the wolf may howl or it may not howl on any given night, regardless of the moon. However, we will suppose that the wolf is more likely to how when the moon is out, and that it most likely to howl when the moon is full. In such a situation – where the likelihood of a howl varies with the state of the moon – the howl is informative. A howl is more indicative of a full moon than of a partial or no moon. Silence is more indicative of no moon than of a full or partial moon. The agent does not acquire any kind of certainty about the state of the moon, but she has received some indication – a *signal* – that may help shape her beliefs. Below we will describe the structure, but will leave an analysis of how signals shape beliefs for a later chapter.

Formally, let the state space be given by some S, and let M denote the set of signals (also called messages). A *signal structure* σ specifies, for each $s \in S$, a probability distribution $\sigma(\cdot|s)$ over M. The following table provides an example of the wolf as a signal structure, where $S = \{f, h, n\}$

and $M = \{howl, silent\}$:

$state \backslash signal$	$howl$	$silent$
$full\ moon$	0.8	0.2
$partial\ moon$	0.5	0.5
$no\ moon$	0.2	0.8

The probability distribution over signals conditional on a state (conditional probability distribution for short) is defined by a horizontal entry of numbers (for instance, $\sigma(howl|full) = 0.8$ and $\sigma(silence|full) = 0.2$). Observe that the more full the moon is, the higher the conditional probability of a howl.

As noted already, unlike partitional information structures, signal structures do not typically rule out states. If the wolf howls, then we cannot conclude that the moon is full, since the wolf howl with positive probability in the other two states as well. The howl just provides a signal that the agent may use to sharpen her beliefs about the moon. That said, partitional information structures are a *special case* of signal structures. The following signal structure describes the earlier wolf that howls if and only if there is a full moon. In this case, the agent can infer when the moon is full, but can never distinguish between a partial moon or no moon.

$state \backslash signal$	$howl$	$silent$
$full\ moon$	1	0
$partial\ moon$	0	1
$no\ moon$	0	1

Exercise: Imagine that on your way to the gym, you always forget whether to turn left or to turn right at a particular intersection. You know that your friend is very often wrong when giving directions (so much so, that you always do the opposite of his recommendation). Model the friend as a signal structure.

Exercise: Model someone who has a perfect poker face.

Exercise: Suppose an Urn contains 4 chips, either red or black, but in unknown proportions. Drawing a sample (with replacement) of, say, 2 chips tells you something about the composition of the urn (that is, it generates a signal). What is the state space? What is the signal structure generated by sampling 2 chips (with replacement).

18.3.2 Combining Signal Structures

As in the case of partitional information structures, signal structures can be combined. Suppose the wolf and the owl are given by the following signal structures respectively

*state**signal*	*howl*	*silent*
full moon	0.8	0.2
partial moon	0.5	0.5
no moon	0.2	0.8

and

*state**signal*	*hoot*	*silent*
full moon	0.6	0.4
partial moon	0.6	0.4
no moon	0	1

.

In order to combine these signal structures, we need to know how they are related, and we will assume that they are independent. That means that, say if there's a full moon, then the probability of a howl and a hoot is 0.8×0.6, the probability of a howl and the owl's silence is 0.8×0.4, and so on. Here the new message space is two dimensional: each signal now takes the form (m_w, m_o), where w and o refer to wolf and owl, respectively. The combined signal structures will therefore look like this:

*state**signal*	$(howl_w, hoot_o)$	$(howl_w, silence_o)$	$(silence_w, hoot_o)$	$(silence_w, silence_o)$
full moon	0.48	0.32	0.12	0.08
partial moon	0.3	0.2	0.3	0.2
no moon	0	0.2	0	0.8

.

More generally, if σ_1 is a signal structure with messages in M_1 and σ_2 is one with messages in M_2, and if both are independent, then the agent can always combine these by defining a new message space $M = M_1 \times M_2$ and a new signal structure σ by

$$\sigma(m_1, m_2|s) = \sigma_1(m_1|s)\sigma_2(m_2|s),$$

for each state s and message $(m_1, m_2) \in M$.

19 The Bayesian Model of Beliefs

19.1 Beliefs as Likelihood Relations

Before we start thinking of beliefs in terms of probabilities, let us think of beliefs in terms of judgements of relative likelihood. To be concrete, suppose that the value of a stock can either be high, medium or low, so that the state space is $S = \{h, m, l\}$. People can express beliefs about relative likelihood of states, such as "it is more likely that the state will be high (h) than low (l)". They may be able to do this even if they find it hard to assign precise probabilities.

People can also express beliefs about the relative likelihood of *events*. We have already seen that an event can be modelled as a set of states $E \subseteq S$, with the understanding that the true state lies in E, or equivalently, that the states not in E are ruled out. People can express statements such as "it is more likely that the state will at least be medium (the event $\{h, m\}$) than take extreme values (the event $\{h, l\}$)".

To describe beliefs, we need to fix a state space S, but also a space of events. Let us take this to be the space of all possible events: the set of all possible subsets of S, which is known as the *power set* and is typically denoted by 2^S.[41] For instance, if $S = \{s_1, s_2, s_3\}$ then its power set is

$$2^S = \{\phi, \{s_1, s_2, s_3\}, \{s_1\}, \{s_2\}, \{s_3\}, \{s_1, s_2\}, \{s_1, s_3\}, \{s_2, s_3\}\},$$

where it should be noted that "all subsets" include not just the singleton and binary subsets of S but also the empty set ϕ and the full space $S = \{s_1, s_2, s_3\}$.

Just as we modelled preferences as rankings in Chapter 1, we can model such beliefs about relative likelihoods as rankings as well:

Definition 9 (Likelihood Relation) *A likelihood relation $\succsim_\ell$ over $(S, 2^S)$ ranks every pair of events in 2^S and requires each $E \in 2^S$ to be as likely as itself.*

[41] This strange notation indicates the total number of possible subsets you can generate from S.

Note well that the *domain* of beliefs (the set of objects being compared) is the space of events 2^S. Analogous to the ordinal preference rankings in Chapter 1, we use the notation $\succ_\ell$, $\sim_\ell$ and $\bowtie_\ell$ to denote ordinal likelihood rankings (or their absence).To illustrate, take $S = \{s, s'\}$ so that the space of events is $2^S = \{\phi, \{s, s'\}, \{s\}, \{s'\}\}$ and consider an *incomplete* likelihood relation defined by:

$$\{s, s'\} \succ_\ell \{s\} \succ_\ell \phi \text{ and } \{s, s'\} \succ_\ell \{s'\} \succ_\ell \phi \text{ but } \{s\} \bowtie_\ell \{s'\}$$

and of course the reflexive rankings: $\{s, s'\} \sim_\ell \{s, s'\}$, $\{s\} \sim_\ell \{s\}$, $\{s'\} \sim_\ell \{s'\}$ and $\phi \sim_\ell \phi$. A likelihood relation is just "like" a preference, except that it ranks events rather than alternatives, and moreover, it is to be interpreted as a ranking in terms of likelihood rather than preference.

Henceforth, we imagine that in its most basic form, a belief is a likelihood relation $\succsim_\ell$ over $(S, 2^S)$. The standard model that we describe below answers two kinds of questions. The *static* version of the model provides structure on beliefs in any given period. The *dynamic* version provides structure on how beliefs change across two periods, in response to information.

19.2 Static Bayesian Model

Fix a point in time and take a likelihood relation $\succsim_\ell$ over $(S, 2^S)$ as the primitive. The static theory hypothesizes simply that likelihood relations can be "expressed" in terms of probability assessments. Formally, a *probability measure* $p(\cdot)$ *on* $(S, 2^S)$ assigns a number $0 \leq p(E) \leq 1$ to each event $E \in 2^S$ that satisfy the following two requirements, where we denote $p(\{s\})$ by $p(s)$:

(i) $\sum_{s \in S} p(s) = 1$

(ii) (Additivity) $p(E) = \sum_{s \in E} p(s)$ for any event $E \in 2^S$.

In Chapter 14 we had defined a probability p over S just as an assignment of a number $0 \leq p(s) \leq 1$ to each $s \in S$ such that $\sum_{s \in S} p(s) = 1$.[42] That formulation did not make clear what probability the agent assigns to events. The above formulation basically adds to the original definition the

[42]It is worth noting that $p(S) = 1$ says that the probability that the true state lies in S is 1. We as analysts picked S, and the property tells us that we have an adequate state space: if the agent was also imagining states outside S this property would fail.

property that the probability of an event is obtained simply by summing the probability of the states in it, that is, $p(E) = \sum_{s \in E} p(s)$. We adopt the rule that the sum over an empty set is always 0. Therefore we have that the null event $\phi \in 2^S$ (which means "no state occurs") has probability $p(\phi) = 0$.

Note that in the above definition, strictly speaking, $p(\cdot)$ on $(S, 2^S)$ assigns probabilities to events, not states. However, a singleton event $\{s\} \in 2^S$ corresponds to the event that state s has occurred. So we can read $p(\{s\})$ as the probability of state s, and just write it as $p(s)$ instead.

Given a primitive likelihood relation $\succsim_\ell$ over $(S, 2^S)$, the static Bayesian model hypothesizes that: *there exists a probability measure that represents* $\succsim_\ell$, *in the sense that for any events* $E, F \in 2^S$,

$$E \succsim_\ell F \Longleftrightarrow p(E) \geq p(F).$$

While a probability representation might seem natural enough for a theory of beliefs, we will see evidence from psychology suggesting that beliefs do not behave like probability measures at all.

Exercise: Based on what we know from previous chapters, deduce that in the static Bayesian model of beliefs, likelihood relations must be complete and transitive.

Exercise: Say that an event $E \in 2^S$ is *non-null* with respect to $\succsim_\ell$, or $\succsim_\ell$*-non-null* for short, if

$$E \succ_\ell \phi.$$

Show that, in terms of the probability representation, an event E is non-null if and only if $p(E) > 0$.

Exercise: Prove that static Bayesian beliefs $\succsim_\ell$ are *monotone* in the sense that for any $A, B \in 2^S$,

$$B \subseteq A \Longrightarrow A \succsim_\ell B,$$

that is, larger events are considered more likely.

19.3 Dynamic Bayesian Model

The primitive in the static model was a single likelihood relation $\succsim_\ell$ over $(S, 2^S)$, understood to reflect the agent's beliefs at any given point in time.

The *dynamic* model describes how people update beliefs in response to information. Fixing some time 0, it involves a belief $\succsim_\ell$ over $(S, 2^S)$, called the *prior belief*. It also specifies the updated belief $\succsim_\ell^E$ over $(S, 2^S)$ that the agent adopts at time 1 if she learns that the event $E \in 2^S$ is true – we refer to it as the *posterior belief*, or *conditional belief,* or if we want to emphasize the event, we refer to it as the *E-conditional belief*.

Not all conceivable information will be considered. Instead we only consider beliefs formed after receiving information that the agent considered *possible* at time 0, that is, only $\succsim_\ell$-non-null events.[43] The primitive of the dynamic model therefore consists of a prior belief $\succsim_\ell$ over $(S, 2^S)$ and the family of posterior beliefs $\succsim_\ell^E$ corresponding to $\succsim_\ell$-non-null events E.

The dynamic model posits that the prior likelihood relation $\succsim_\ell$ is represented by some probability measure $p(\cdot)$, and that for each non-null event E, the conditional likelihood relation $\succsim_\ell^E$ is represented by $p(\cdot|E)$. That is, each belief satisfies the static Bayesian model. The main question is: how are all these static beliefs related to each other? The defining feature of the dynamic model is the hypothesis, for any non-null event $E \in 2^S$, the posterior belief is related to the prior belief by the Bayesian conditioning formula: for any event $A \in 2^S$,

$$p(A|E) = \frac{p(A \cap E)}{p(E)}.$$

Observe that the left-hand side expresses a posterior belief while the right-hand side involves only prior beliefs. Therefore posteriors are related to the prior in a very particular way. We explore the meaning of this formula in the next section.

Exercise: Suppose the state space is $S = \{s, s', s'', s'''\}$. Suppose we know that the agent's prior beliefs satisfy $p(\{s, s', s''\}) = 0.75$, $p(\{s, s'\}) = 0.5$ and $p(\{s, s', s'''\}) = 0.75$. What is the posterior probability she assigns to the event $\{s, s', s''\}$ after receiving information $\{s, s', s'''\}$? Is it higher or lower than the prior probability of $\{s, s', s''\}$?

[43] To the extent that we are gathering data on posteriors by asking a person how likely they think state s is if they imagined receiving information E, we would not know how to make sense of their answers.

Exercise: Show that $p(E|E) = 1$. Show that $p(A|E) = 0$ if $A \cap E = \phi$.

Exercise: Suppose the agent is on a game show where she has to pick one of 3 boxes. The grand prize has been placed in one of the boxes randomly. If she selects the correct box, she wins. Otherwise she leaves empty handed. The agent's uncertainty is captured by a state space $S = \{s_1, s_2, s_3\}$ where s_i is the state of the world where box i contains the grand prize.

(i) What "should" the prior belief over $(S, 2^S)$ be if the agent imbibes the information that the winning box was selected randomly? (It will suffice to specify just the prior beliefs over the 3 states since, by additivity, we can infer the probability for all events 2^S.)

(ii) Suppose that, before selecting the box, the agent is given some information: "the grand prize is not in box 1". Compute the agent's posterior belief over $(S, 2^S)$. Be precise in how you use Bayesian conditioning.

19.4 More on Bayesian Conditioning

We now try to get a better handle on what the Bayesian conditioning formula is saying. The formula describes beliefs about events. But it is simpler to think about beliefs about states. We show that the formula can in fact be made sense of in terms of beliefs about states. As before, a state s can be viewed as a singleton event $\{s\} \in 2^S$ and we can write $p(\{s\}|E)$ as $p(s|E)$. Then:

Proposition 13 *A family of beliefs satisfies the Bayesian conditioning formula if and only if it satisfies the following "Bayesian conditioning formula for states": for any non-null event $E \in 2^S$, the posterior belief for state $s \in S$ is given by*

$$p(s|E) = \begin{cases} \frac{p(s)}{\sum_{s' \in E} p(s')} & \text{if } s \in E \\ 0 & \text{if } s \notin E. \end{cases}$$

The proof is provided in the next section. The proposition asserts that the Bayesian conditioning formula is equivalent to a more transparent formula that applies only to states. According to this formula, if information E is provided to the agent, then she assigns $p(s|E) = 0$ to any s that is ruled out

by E, just as what one would hope. For any state s that is included in E, her updated belief $p(s|E)$ is a scaled version of her prior, obtained by dividing the prior $p(s)$ by the sum $\sum_{s' \in E} p(s')$ of prior probabilities of the states in E . The scaling ensures that the updated beliefs satisfy the requirement that probabilities sum to 1:

$$\sum_{s \in E} p(s|E) = \sum_{s \in E} \frac{p(s)}{\sum_{s' \in E} p(s')} = \frac{\sum_{s \in E} p(s)}{\sum_{s' \in E} p(s')} = 1.$$

To illustrate, consider the above example of the game show with three boxes. The state space is $S = \{s_1, s_2, s_3\}$ and the prior is uniform:

$$p(f) = p(h) = p(n) = \frac{1}{3}.$$

If the agent learns that the prize is in box 1, that is, she learns the event $\{s_1\}$, then the updated beliefs will be

$$p(s_1|\{s_1\}) = 1 \text{ and } p(s_2|\{s_1\}) = p(s_3|\{s_1\}) = 0.$$

That is, since the event $\{s_1\}$ rules out s_2 and s_3, the posterior should assign exactly 0 likelihood to them. Suppose instead that the agent learns that the prize is not in box 1. Then the event $\{s_2, s_3\}$ is learned and

$$p(s_1|\{s_2, s_3\}) = 0 \text{ and } p(s_2|\{s_2, s_3\}) = p(s_2|\{s_2, s_3\}) = \frac{1/3}{1/3 + 1/3} = \frac{1}{2},$$

that is, the posterior places 0 likelihood on s_1 and rescales the prior belief on s_2 and s_3 so that the new belief sums to 1.

One way of thinking about Bayesian conditioning is that it requires the *relative beliefs* about any two states s, s' to be *independent* of whether or not a third state s'' is ruled out. In the illustration, the prior belief regarded both states s_2 and s_3 as equally likely, and so did her posterior beliefs when s_1 was ruled out:

$$\frac{p(f|\{f, h\})}{p(h|\{f, h\})} = 1 = \frac{p(f)}{p(h)}.$$

19.5 Proof of Proposition 13

It is easy to see that the Bayesian conditioning formula implies the formula for states. Indeed, for any non-null event $E \in 2^S$ and state $s \in S$, there are two possibilities: either $s \in E$ or $s \notin E$. if $s \notin E$, then $\{s\} \cap E = \phi$ and so by Bayesian conditioning,

$$p(s|E) = \frac{p(\{s\} \cap E)}{p(E)} = 0.$$

Similarly, if $s \in E$, then $\{s\} \cap E = \{s\}$ and so by Bayesian conditioning and the additivity of probability measures,

$$p(s|E) = \frac{p(\{s\} \cap E)}{p(E)} = \frac{p(s)}{\sum_{s' \in E} p(s')},$$

as desired. Note that we implicitly made use of the notational assumption that $p(s|E)$ means $p(\{s\}|E)$.

Consider now the converse: assume that the Bayesian conditioning formula for states holds. We show that the formula for events must hold. Due to the additivity of probability measures, for any $A \in 2^S$, the prior must satisfy

$$p(A) := \sum_{s \in A} p(s),$$

and similarly any B-conditional posterior belief must satisfy

$$p(A|B) := \sum_{s \in A} p(s|B).$$

But we can also fine-tune this conditional probability by recalling that $p(s|B) = 0$ for any state s ruled out by B. In particular:

$p(A|B) = \sum_{s \in A} p(s|B)$

$= \sum_{s \text{ that is in } A \text{ and } B} p(s|B) + \sum_{s \text{ that is in } A \text{ but not in } B} p(s|B)$

$= \sum_{s \text{ that is in } A \text{ and } B} p(s|B)$ (since $p(s|B) = 0$ if $s \notin B$)

$= \sum_{s \in A \cap B} p(s|B)$ (since $s \in A \cap B$ means that s is in A and B), and so we obtain

$$p(A|B) = \sum_{s \in A \cap B} p(s|B).$$

Thus, the posterior probability of an event A is the sum of probabilities of those states s in A that are not ruled out by the information B, which is precisely the sum of probabilities of states in $A \cap B$. But let us not stop here. Observe that by the presumed conditioning formula for states,

$p(A|B) = \sum_{s \in A \cap B} p(s|B)$
$= \sum_{s \in A \cap B} \frac{p(s)}{\sum_{s' \in B} p(s')}$
$= \frac{\sum_{s \in A \cap B} p(s)}{\sum_{s' \in B} p(s')} = \frac{p(A \cap B)}{p(B)}$. That is,

$$p(A|B) = \frac{p(A \cap B)}{p(B)},$$

and we have recovered the Bayesian conditioning formula from events.

20 Properties of Bayesian Beliefs

What are the different ways that the relationship between priors and posteriors can be expressed in the Bayesian model?

20.1 Law of Total Probability

Recall from high school that for any sets A and B, we can break A into two distinct parts: the part that is in B and the part that is not in B. The first proposition uses this fact to show that the prior probability of A can be written as the sum of two terms: the probability that B happens and then A happens, and the probability that B does not happen and then A happens.

Proposition 14 *If beliefs are Bayesian then for any pair of events $A, B \subset S$ such that $p(B), p(B^c) > 0$,*

$$p(A) = p(A|B)p(B) + p(A|B^c)p(B^c).$$

Proof. Suppose that beliefs are Bayesian, that is, they satisfy Bayesian conditioning. Take any pair of events $A, B \subset S$ such that $p(B), p(B^c) > 0$. Then using basic set theory and the Bayesian conditioning formula:

$$\begin{aligned} p(A) &= \sum_{s \text{ in } A} p(s) \\ &= \sum_{s \text{ in } A \text{ and } B} p(s) + \sum_{s \text{ in } A \text{ but not in } B} p(s) \\ &= p(A \cap B) + p(A \cap B) \\ &= p(A|B)p(B) + p(A|B^c)p(B^c), \end{aligned}$$

as desired. ■

The *Law of Total Probability* is the name for the extension of this relationship to any finite number of sets $B_1, B_2, ..B_n$, that are pairwise disjoint in the sense that $B_i \cap B_j = \phi$ for all i, j. The proof is entirely similar to the above proposition.

Proposition 15 *If beliefs are Bayesian then it satisfies the Law of Total Probability: for any pair of events $A, B_1, ..B_n \subset S$ such that $B_1, ..B_n$ are pairwise disjoint and $p(B_i) > 0$ for each i,*

$$p(A) = \sum_{i=1}^{n} p(A|B_i)p(B_i).$$

Proof. Exercise. ■

20.2 Bayes' Rule

Bayes' Rule, also known as *Bayes' Law* and *Bayes' Theorem*, is an important relationship between priors and posteriors that is appreciated across various disciplines. We show that:

Proposition 16 *If beliefs are Bayesian then they respect Bayes' Rule: for any pair of events $A, B \subset S$ such that $p(B) > 0$,*

$$p(A|B) = p(A) \times \frac{p(B|A)}{p(B)}.$$

Proof. Suppose that beliefs are Bayesian, that is, they satisfy Bayesian conditioning. Take any pair of events $A, B \subset S$ such that $p(B) > 0$. If $p(A) = 0$ then by Bayesian conditioning, $p(A|B) = 0$, thereby establishing the claim for the case where $p(A) = 0$. For the case where $p(A) > 0$, note that Bayesian conditioning implies $p(A|B) = \frac{p(A \cap B)}{p(B)}$ and $p(B|A) = \frac{p(A \cap B)}{p(A)}$. Rearranging these expressions yields

$$p(A|B)p(B) = p(A \cap B) = p(B|A)p(A),$$

which rearranges in turn again to yield $p(A|B) = p(A) \times \frac{p(B|A)}{p(B)}$, as desired. ■

While Bayesian conditioning establishes a particular relationship between posteriors and the prior, Bayes' Rule highlights another relationship that arises as a consequence of it. Specifically, the posterior belief $p(A|B)$ must be an adjustment of the prior belief $p(A)$ by a factor given by

$$\frac{p(B|A)}{p(B)}.$$

This factor can be interpreted in terms of how *representative* B is about A. For instance, if $\frac{p(B|A)}{p(B)} > 1$, then the occurrence of A increases the chances of observing B (that is, is $p(B|A) > p(B)$) and it can be said that B is "favorably representative" about the occurrence of A. It is intuitive that observing such a favorably representative event B should increase our posterior assessment $p(A|B)$ relative to the prior $p(A)$. Bayes' Rule tells us that it should, and moreover, it gives us a precise calculation of the posterior as a product of $p(A)$ and $\frac{p(B|A)}{p(B)}$.

To illustrate, suppose you meet a person that you find really cool. Being a student of economics, you may wonder if this person is an economist as well. If you are Bayesian, you would embody Bayes' Rule, and your posterior $p(economist|cool)$ would depend on two things: how many economists you know of in the population (which would determine your prior $p(economist)$), how many economists you know of that are cool (which would underlie $p(cool|economists)$) and how many cool people you know of in the population (which would underlie $p(cool)$). As a Bayesian, your posterior would be given by $p(economist|cool) = p(economist) \times \frac{p(cool|economists)}{p(cool)}$. Indeed, if you think there is no such thing as a cool economist, then $p(cool|economists) = 0$ and by Bayes' Rule you'd be quite sure the cool person you met is not an economist, $p(economist|cool) = 0$.

Bayes' Rule can be written in more than one way. Here is one form in which it commonly appears:

Proposition 17 *If beliefs are Bayesian, then they must respect the following form of Bayes' Rule: for any pair of events $A, B \subset S$,*

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|A^c)p(A^c)}.$$

Proof. As before, Bayesian beliefs satisfy $p(A|B) = p(A) \times \frac{p(B|A)}{p(B)}$ for all $A, B \subset S$. By the Law of Total Probability, $p(B) = p(B|A)p(A) + p(B|A^c)p(A^c)$. This yields the alternative form. ■

Exercise: Consider a partition $\{A_1, .., A_n\}$ of S and any $B \subset S$. Show that for any $i = 1, .., n$,

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_{j=1,...,n} p(B|A_j)p(A_j)}.$$

20.3 Chain Rule

Another useful relationship between priors and posteriors is given by the *chain rule*, also called the *product rule*. To illustrate, suppose we are interested in assessing the (prior) probability that it will rain on three consecutive weekdays, say, Monday, Tuesday and Wednesday. Denote "rain on Monday" (resp. Tuesday, Wednesday) by M (resp. T, W). The intersection of all three events (our event of interest) can be written as MTW, and we can use similar notation for the intersection of any two events. The chain rule tells us that we can compute $p(MTW)$ by pulling up data that yields (a) the prior probability $p(M)$ of rain on Monday, (b) the posterior probability $p(T|M)$ of rain on Tuesday given that it rained on Monday, and (c) the posterior probability $p(W|MT)$ of rain on Wednesday given that it rained on both Monday and Tuesday. In particular:

$$p(MTW) = p(W|MT) \times p(T|M) \times p(M).$$

To see that this must be true, apply Bayesian conditioning to obtain

$$\begin{aligned} & p(W|MT) \times p(T|M) \times p(M) \\ = & \frac{p(MTW)}{p(MT)} \times \frac{p(MT)}{p(M)} \times p(M) \\ = & p(MTW). \end{aligned}$$

This observation generalizes to longer sequences:

Proposition 18 *If beliefs are Bayesian, then they must respect the chain rule: for any events $A_1, .., A_n \subset S$,*

$$p(A_1 \cap ... \cap A_n) = p(A_n|A_1 \cap .. \cap A_{n-1}) \times ... \times p(A_3|A_1 \cap A_2) \times p(A_2|A_1) \times p(A_1).$$

Proof. For any $A_1, .., A_n \subset S$, Bayesian conditioning yields $p(A_n|A_1 \cap .. \cap A_{n-1}) = \frac{p(A_1 \cap ... \cap A_n)}{p(A_1 \cap ... \cap A_{n-1})}$, that is,

$$p(A_1 \cap ... \cap A_n) = p(A_n|A_1 \cap .. \cap A_{n-1}) \times p(A_1 \cap ... \cap A_{n-1}).$$

Similarly, it must be that

$$p(A_1 \cap ... \cap A_{n-1}) = p(A_{n-1}|A_1 \cap .. \cap A_{n-2}) \times p(A_1 \cap ... \cap A_{n-2}),$$

which we can insert into the first equation to obtain

$$p(A_1 \cap ... \cap A_n) = p(A_n|A_1 \cap .. \cap A_{n-1}) \times p(A_{n-1}|A_1 \cap .. \cap A_{n-2}) \times p(A_1 \cap ... \cap A_{n-2}).$$

Continue in this fashion until the last term is $p(A_1)$. This completes the proof. ■

If $n = 2$ then the chain rule is just the Bayesian conditioning formula.

21 Bayesian Inference

Bayesian inference studies how Bayesian agents update beliefs after observing signals arising from a signal structure. Given a state space S and a signal space M, recall that a *signal structure* σ (or experiment) specifies, for each $s \in S$, a probability distribution $\sigma(\cdot|s)$ over M. Consider a Bayesian agent with prior p over S. How would she update her beliefs about the state if she received a signal m from the signal structure σ? The answer will come from Bayes' Rule.

21.1 Computing Posterior Beliefs

We will first introduce the expression for computing the Bayesian posteriors without being completely rigorous. Once we have a sense of what the expression is saying, we will go back and derive it properly.

Prior to receiving any information, the agent's belief about state s is given by $p(s)$. Conditional on receiving a signal m, the Bayesian model implies that posterior belief $p(s|m)$ must satisfy:

Proposition 19 *If beliefs are Bayesian, then the Bayesian posterior belief about s conditional on m is*

$$p(s|m) = \frac{\sigma(m|s) \times p(s)}{\sum_{s' \in S} \sigma(m|s') \times p(s')}.$$

Proof. Using the Bayesian conditioning formula:

$$p(s|m) = \frac{p(s,m)}{p(m)} = \frac{\sigma(m|s) \times p(s)}{\sum_{s' \in S} \sigma(m|s') \times p(s')},$$

as desired. ■

Let us first go over the derivation in the proof and then interpret the expression. By Bayesian conditioning,

$$p(s|m) = \frac{p(s,m)}{p(m)},$$

that is, the posterior probability $p(s|m)$ of state s given signal m equals the prior probability $p(s,m)$ of state s and signal m occurring together, divided by the prior probability $p(m)$ of ever receiving m. The prior probability $p(s,m)$ of state s and signal m occurring is

$$p(s,m) = \sigma(m|s) \times p(s),$$

that is, the prior probability $p(s)$ of s and, given s, the probability $\sigma(m|s)$ that the signal structure will generate the signal m. The prior probability of seeing signal m is

$$p(m) = \sum_{s' \in S} p(s', m) = \sum_{s' \in S} \sigma(m|s') \times p(s'),$$

that is, we see m when m is generated in state s', or m is generated in state s'', or m is generated in state s'''.... The probability of seeing m is therefore the sum of the probabilities of m being generated in each state $p(m) = \sum_{s' \in S} p(s', m)$. But we have already seen that $p(s', m) = \sigma(m|s') \times p(s')$. Therefore $p(m) = \sum_{s' \in S} \sigma(m|s') \times p(s')$.

Inserting the expression for $p(s,m)$ and $p(m)$ in the Bayesian conditioning formula $\frac{p(s,m)}{p(m)}$ yields the expression in the Proposition.

Let us now interpret the expression. It says that the posterior $p(s|m)$ revises the prior $p(s)$ by multiplying it with a factor $\frac{\sigma(m|s)}{\sum_{s' \in S} \sigma(m|s')p(s')}$. This factor compares the probability of m in state s (given by $\sigma(m|s)$) with the probability of ever receiving m (given by $\sum_{s' \in S} \sigma(m|s')p(s')$). If this ratio is greater than 1, that means that m is more likely to be generated in state s – here, the "more likely" points to the comparison between the likelihood of occurrence $\sigma(m|s)$ of signal in state s with the likelihood of occurrence $p(m) = \sum_{s' \in S} \sigma(m|s')p(s')$ of the signal m averaged across all states. Thus, the signal m is indicative or representative of state s. Indeed, in this case, the posterior about s must be higher than the prior, $p(s|m) > p(s)$. If it is less than 1, then m is relatively unlikely to happen in state s, and so observing m is a negative indication about s, and we should have $p(s|m) < p(s)$. Finally, if the ratio equals 1, then m is not an informative signal about s and our beliefs about s must not change, $p(s|m) = p(s)$.

Exercise: Recall the example of Jack and Jill. Compute Jill's posterior belief about Jack's depression conditional on seeing that he is withdrawn.

What is her posterior conditional on seeing that he is not withdrawn? In each case, compare her posterior belief with her prior belief.

21.2 Rigorous Derivation

There are many ways that the preceding discussion lacked rigor. Beliefs are defined over events, and events are sets of states. But there were no sets of states indicated anywhere, and beliefs seemed to be defined over states directly. Moreover, the Bayesian conditioning formula, as originally defined, involves the intersection of events but, again, there were no sets of states that we could takes intersections of. We make up for all this looseness by being more clear and explicit about how we derived the expression for the Bayesian posterior.

21.2.1 Extended State Space

If the agent has access to an experiment, we must recognize that the relevant uncertainty in the world is no longer just the states is S, but now also includes the uncertainty about the signals in M. Thus, we must *extend* the state space S to the space given by

$$S^* = S \times M,$$

so that an extended state of the world is now described by the pair (s, m) specifying an original state of the world s and a message m that could be received.

It should be noted that, with the extended state space, we can no longer simply make statements such as "the true state is s" or that "the message received is m" unless we are being informal (as we were in the previous section). The formal meaning of the statement "the true state is s" is that we have learned the *event* in S^* given by

$$\{s\} \times M = \{(s, m), (s, m'), (s, m''), ...\},$$

that is, the event in S^* where we rule out all extended states except the ones that include s (recall that the point of an event is that it excludes states). In particular, we rule out all states except s, but rule out no messages. Hence

the event $\{s\} \times M$ allows only state s but allows for all messages. *This event is often simply denoted s for convenience.* We will do the same, but the reader should be clear that this is a blatant abuse of notation since, by definition, an event is a subset of $S^* = S \times M$ whereas the notation s is a state in S. Similarly, the meaning of "the message received is m" is the event where we rule out all messages except m but do not rule out any state:

$$S \times \{m\} = \{(s, m), (s', m), (s'', m), ...\}.$$

This is often denoted m for convenience.

It is worth emphasizing that although we are in the world of signal structures (as opposed to partitional information), after having formulated the extended state space we are back to talking about events as in the case of partitional information!

Exercise: The event "state s occurs and signal m is received" rules out all states and signals except (s, m), and thus is of the form $\{(s, m)\}$. Verify that this event is in fact the conjunction (that is, intersection) of two events

$$\{(s, m)\} = s \cap m,$$

where s denotes the event $\{s\} \times M$ and m denotes the event $S \times \{m\}$.

Exercise: Model a state space and signal structure that captures the following story: Jill's friend, Jack, may or may not be suffering from depression. While he is very secretive about mental health issues, Jill is emotionally intelligent enough to tell when he seems withdrawn. While he is always withdrawn when depressed, he is quite moody as a person in general, and is withdrawn 40% of the time even if he is not depressed.

(a) What is the state space and the message space? What is the signal structure?

(b) Write the event that "Jack is not depressed".

(c) Write the event that "Jack is withdrawn".

(d) Is being withdrawn an informative signal about Jack's mental health?

21.2.2 Extended Prior

The prior p is defined only on S, but since we have extended the state space to S^*, we need to extend the prior to S^* as well. That is, we need to say what prior probability that agent assigns to state s occurring and message m being generated. *We will continue to use the notation p to denote an extended prior on the extended state space $S^* = S \times M$.* This is again an abuse of notation, but the hope is that the reader will understand what the notation is referring to. We will also use the convention of denoting a singleton event $\{(s, m)\}$ as (s, m). We have used this notation before (see the definition of a probability measure).

So, how do we extend p to S^*? We *assume* that the agent understands the experiment perfectly. Then, when assessing the probability of the event $\{(s, m)\}$ that state s and signal m happen, she assigns it the probability

$$p(s, m) = \sigma(m|s) \times p(s). \tag{3}$$

That is, the probability that the extended state (s, m) is realized equals the probability that s occurs multiplied by the probability of m being generated in state s.

The fact that the agent uses the correct probabilities $\sigma(m|s)$ defined by the experiment is what reflects the assumption that she understands the experiment. As another expression of her understanding of the experiment, note that the Bayesian posterior $p(m|s)$ about the likelihood of signal m conditional on state s must coincide with the probability $\sigma(m|s)$ yielded by the experiment. Using the convenient notation we have introduced,[44] this is easily seen by using Bayesian conditioning and (3) to obtain

$$p(m|s) = \frac{p(s, m)}{p(s)} = \frac{\sigma(m|s) \times p(s)}{p(s)} = \sigma(m|s).$$

Using the extended prior we can also compute the prior belief about receiving some signal m

$$p(m) = \sum_{s' \in S} p(s', m) = \sum_{s' \in S} \sigma(m|s') \times p(s').$$

[44] That is, s denotes the event $\{s\} \times M$, m denotes the event $S \times \{m\}$, and (s, m) denotes the event $\{(s, m)\}$.

Thus $p(m)$ is the probability of the event $\{(s,m),(s',m),...,(s'',m)\}$, and additivity of probability measures implies that $p(m)$ equals $\sum_{s' \in S} p(s',m)$, the sum of the probability of (s,m) over all possible s.

Exercise: Consider again the example of Jack and Jill. Suppose that Jill's prior belief about Jack's depression puts probability 0.3 on him being depressed at any point in time.

(a) What is Jill's prior probability of Jack being depressed and withdrawn?

(b) What is her prior probability of Jack being withdrawn?

21.2.3 Extended Posterior

We are finally ready to derive the Bayesian posterior. First, recognize that the posterior $p(\cdot|E)$ conditional on event $E \in 2^{S^*}$ must be a probability measure on the extended state space $(S^*, 2^{S^*})$. However, our interest only particular events: those that correspond to observing a signal m, and therefore, in posteriors of the form $p(\cdot|m)$. Moreover, since we want to know the agent's posterior beliefs about the state of the world, we are typically interested in events that correspond to a state s. Therefore we only compute the posterior $p(s|m)$ over the event corresponding to s given an event corresponding to m.

Based on all this, we can now go back and look at the proof for Proposition 19 and make better sense of it. The Bayesian conditioning formula is:

$$p(s|m) = \frac{p(s \cap m)}{p(m)},$$

where the intersection is meaningful since, as before, the notation s, m stand for events in S^*. The proof should now make rigorous sense.

An alternative proof for the same proposition uses Bayes Rule. Make sure you see that, by Bayes' Rule,

$$p(s|m) = \frac{p(m|s)p(s)}{p(m)} = \frac{\sigma(m|s) \times p(s)}{\sum_{s' \in S} \sigma(m|s') \times p(s')},$$

as desired.

21.3 Posterior Beliefs After a Sequence of Signals

We conclude by formulating the inference problem more generally. Given a signal structure σ, it may be that we observe not one, but a sequence of signals $m_1, m_2, ...m_n$. For instance, in times where there is a concern of a recession (which is a state of the world describing the economy), we obtain economic and financial news every day and continually update our beliefs about the likelihood of a recession. We have formulated the Bayesian posterior conditional on observing one signal, but what is the expression for the posterior after a sequence of signals?

The Bayesian posterior depends on how the signal structure σ generates sequences of signals. Say that signals are *conditionally independent* if, conditional on any state s, the probability of the signals $m_1, m_2, ...m_n$ is:

$$\sigma(m_1|s) \times ... \times \sigma(m_n|s).$$

Make sure to construct the proper extended prior in this environment – assume that each signal is an independent draw from the signal structure (independence means that in state s the probability of $m_1, m_2, ...m_n$ is $\sigma(m_1|s) \times ... \times \sigma(m_n|s)$).

Proposition 20 *Suppose that beliefs are Bayesian and the signal structure σ generates signals that are conditionally independent. Then the Bayesian posterior belief about s conditional on a sequence of signals $m_1, m_2, ...m_n$ m is given by*

$$p(s|m_1, m_2, ...m_n) = \frac{\sigma(m_1|s) \times ... \times \sigma(m_n|s) \times p(s)}{\sum_{s' \in S} \sigma(m_1|s') \times ... \times \sigma(m_n|s') \times p(s')}.$$

Proof. The simple way to prove this is just to imagine that the space M of signals consist of sequences $(m_1, m_2, ...m_n)$, and conditional on a state s, the probability of signal $(m_1, m_2, ...m_n)$ is given by $\sigma(m_1, m_2, ...m_n|s)$. By Proposition 19,

$$p(s|m_1, m_2, ...m_n) = \frac{\sigma(m_1, m_2, ...m_n|s) \times p(s)}{\sum_{s' \in S} \sigma(m_1, m_2, ...m_n|s') \times p(s')}.$$

But by conditional independence, $\sigma(m_1, m_2, ...m_n|s) = \sigma(m_n|s) \times ... \times \sigma(m_n|s)$. Inserting this yields the desired result. ■

22 Psychology of Beliefs

Research in psychology as early as the 1960's and 1970's has sought to understand the properties of people's beliefs. The most influential research was due to Kahneman and Tversky and their theory of beliefs (the "Heuristics and Biases" program) is the dominant paradigm in psychology today. We describe the main findings in the psychology literature and then discuss the Heuristics and Biases program.

We divide up the findings in terms of whether they pertain to static beliefs or dynamic beliefs. We present both experimental evidence, and the testable implications of the Bayesian model that are violated.

References:

Tversky, A. and Kahneman, D. (1982). "Judgments of and by representativeness". In Kahneman, D.; Slovic, P.; Tversky, A. (eds.). Judgment under uncertainty: Heuristics and biases. Cambridge, UK: Cambridge University Press. ISBN 0-521-28414-7

Tversky A, Kahneman D (1983). "Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment". Psychological Review 90:293–315.

22.1 Static Beliefs

In a famous experiment, Tversky and Kahneman (1983) present subjects with the following description of a fictitious person:

"Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations."

The subjects are asked to report which of the following events are more probable:

A. Linda is a bank teller.

B. Linda is a bank teller and is active in the feminist movement.

The majority of subjects chose B. But this response implies that subjects' likelihood judgements cannot be represented by a probability measure! The reason is as follows. If A denotes the set of all bank tellers, and C denotes the set of all people active in the feminist movement, then B is the conjunction

of these two sets, that is, $B = A \cap C$. But then B is a subset of A. The additivity property of probability measures implies that

$$p(B) \leq p(A).$$

Intuitively, larger events must be more likely. Event A consists of bank tellers who are activists and also bank tellers who are not activists. Therefore it must be at least as likely as the event B consisting of just bank tellers that are activists. But the subjects in the experiment (and subsequent replications of the experiment) reported that B was more likely. This is called the *conjunction fallacy*, and is also known as the *Linda fallacy*.

22.2 Static Beliefs about Sequences

Suppose that a fair coin is to be flipped 3 times. There is uncertainty about the realizations of these flips, and thus can be described by the state space $S = \{(H,H,H),(H,H,T),....,(T,T,T)\}$. That is, a state of the world is a sequence of realizations.

We consider properties of beliefs over such sequences. The findings presented below are typically expressed in the literature in terms of dynamic beliefs (that is, given that event E has happened, what do you think is the likelihood of event A?), but they strongly suggest a property of people's priors: people harbor a prior belief in particular patterns even when those patterns do not objectively exist. Consequently, we categorize these findings as indicating something about static beliefs.

Strictly speaking, the findings do not contradict the static Bayesian model: the model says that people have beliefs, but it is silent about where beliefs come from and what properties they must have beyond being represented by probabilities. That said, it is standard practice in economics to assume that people understand their environment. In particular, if they are told that a fair coin is being flipped, then it is presumed that they will understand that the probability of 2 heads is the same as that of 2 tails and are both given by $\frac{1}{2} \times \frac{1}{2}$, etc.

22.2.1 Gambler's Fallacy

Given a fair coin, which is more likely:

$$(\underbrace{H, H...H}_{\text{9 Heads}}, H) \text{ or } (\underbrace{H, H...H}_{\text{9 Heads}}, T)?$$

Evidence suggests that most people believe the latter to be more likely (subjects in Benjamin, Moore and Rabin 2018 believed on average that the latter sequence was twice as likely as the first). The "reasoning" is that after that many heads in a row, tails has to appear if the coin is indeed fair. This logic is incorrect, of course, since the probability of both sequences is the same, $\left(\frac{1}{2}\right)^{10}$.[45] People's beliefs are said to exhibit the *gambler's fallacy*.

There is evidence of the gambler's fallacy outside the experimental lab and in the field. For instance it is observed in horse-race bets and in roulette playing in casinos. The most famous example is the roulette at the Monte Carlo Casino on August 18, 1913, when the ball fell in black 26 times in a row. Gamblers reportedly lost millions of francs betting against black afterwards.

22.2.2 Hot Hand Effect

In basketball, a *hot hand* refers to a temporary increase in a player's ability to make her shots. Gilovich, Vallone and Tversky (1985) showed that there is no statistical evidence of a hot hand, despite much reaction by sports lovers who vowed that it must exist. Subsequent studies have replicated their result, but some controversy remains even within academia. The *hot hand effect* is the term used to describe a belief in the persistence of a streak (not necessarily in a sports context).

It appears that the gamblers fallacy and the hot hand effect co-exist. Suetend, Labo-Jorgensen and Tyran (2016) found that, in their sample, after a lottery number won once, players bet less on it (that is, they exhibited the gambler's fallacy) but when a streak of two or more wins occurred with that number, then players bet more on it the longer the streak lasted (that is, they exhibited a hot hand effect).

[45] There is a theorem in statistics known as the Law of Large Numbers that says, roughly, that in *infinitely* long sequences, the proportions of heads and tails must be equal. This does not apply to sequences of finite length.

22.3 Dynamic Beliefs

Recall that Bayes' Rule in the Bayesian inference set up takes the form

$$p(s|m) = \frac{\sigma(m|s) \times p(s)}{\sum_{s'' \in S} \sigma(m|s'') \times p(s'')}.$$

Conditional on m then posterior belief about state s' is $p(s'|m) = \frac{\sigma(m|s') \times p(s')}{\sum_{s'' \in S} \sigma(m|s'') \times p(s'')}$. Assume that $p(s|m)$ and $p(s'|m)$ are both strictly positive, that is, the message m does not rule out either state. Since the denominator in both these expressions is the same, we see that the ratio of posteriors can be written:

$$\frac{p(s|m)}{p(s'|m)} = \frac{\sigma(m|s)}{\sigma(m|s')} \times \frac{p(s)}{p(s')}.$$

This "odds-ratio" version of Bayes' Rule is useful and is what we will be focusing on. Observe that the posterior odds ratio depends on two terms. One term is the prior odds ratio, $\frac{p(s)}{p(s')}$, and the other term, $\frac{\sigma(m|s)}{\sigma(m|s')}$, tells us the odds of signal m in state s relative to state s'. Thus, conditional on signal m, the posterior odds for state s relative to s' are high if the prior odds are high and the signal is more likely in state s than s'. Let us see if people's update their beliefs in this manner.

22.3.1 Base-Rate Neglect

A group of subjects were told that a person has been drawn randomly from a set of professionals consisting of 30 engineers and 70 lawyers. When asked what was the likelihood of the person being an engineer, the subjects' answers reflected the *base rate* (that is, the proportion of engineers in the group), as one would expect. That is, writing the state space as $\{E, L\}$, the prior odds ratio was

$$\frac{p(E)}{p(L)} = \frac{3}{7}.$$

Next, the subjects were shown a brief personality description of this person: he was 30 years of age, unmarried with no children, liked by his colleagues and promised to have a successful career, etc. Denote this description by m. When subjects were now asked for their posterior beliefs, they indicated

$$\frac{p(E|m)}{p(L|m)} = 1,$$

that is, the subjects judged it equally likely that the person was an engineer or a lawyer.

Nothing here is inconsistent with Bayes' Rule. Presumably each profession leads to some distribution over descriptions, and this fits the description of a signal structure σ. Moreover, there certainly exists σ such that $\frac{p(E|m)}{p(L|m)} = \frac{\sigma(m|E)}{\sigma(m|L)} \times \frac{p(E)}{p(L)}$. The contradiction comes when subjects were given the same description m but were told that the set of professionals consisted of 70 engineers and 30 lawyers. Looking at the Bayes' Rule formula, we see that $\frac{\sigma(m|E)}{\sigma(m|L)}$ has not changed, but now $\frac{p(E)}{p(L)}$ has been increased (from $\frac{3}{7}$ to $\frac{7}{3}$).

Bayes' Rule requires that $\frac{p(E|m)}{p(L|m)}$ *must* increase. This should make sense: if there are now more engineers, then whatever the description of the person, it has become more likely that an engineer was randomly picked, and the posterior belief must increase.

But in the experiment, posterior odds continued to be $\frac{p(E|m)}{p(L|m)} = 1$! It is as if subjects completely ignored the base rates once they receive information, which is why this is called *base-rate neglect*. It is as if the posterior is determined mainly in terms of the information provided, captured by $\frac{\sigma(m|E)}{\sigma(m|L)}$ here.

22.3.2 Conservatism

In a classic experiment, Edwards (1968) told subjects that there are two urns, each with 1000 poker chips. Urn 1 has 700 red and 300 blue chips, and Urn 2 has 300 red and 700 blue. An urn is chosen at random, and a dozen chips drawn at random (with replacement) contain 8R 4B. He asks subjects: What is the probability that the chips came from urn 1?

The prior odds ratio is naturally

$$\frac{p(Urn\ 1)}{p(Urn\ 2)} = 1$$

since each urn was picked with probability 0.5. Note that the sample (8R 4B) is in fact a signal about the composition of the urn. Clearly, this signal is more likely under Urn 1 than Urn 2, since Urn 1 contains more red than

blue chips. So it is intuitive that the posterior odds favoring Urn 1 must be higher than the prior odds

$$\frac{p(Urn\ 1|8R4B)}{p(Urn\ 2|8R4B)} > \frac{p(Urn\ 1)}{p(Urn\ 2)}.$$

This is consistent with Bayes' Rule as well, and in fact most subjects updated their beliefs in the correct direction. The experiment, however, finds that subjects do not update *as much as* Bayes' Rule requires. The finding is referred to as *conservatism*. Specifically, in the experiment, subjects' posterior odds ratio was

$$\frac{p(Urn\ 1|8R4B)}{p(Urn\ 2|8R4B)} = 2.3.$$

In the following exercise, you are asked to:

Exercise: Show that the Bayesian posterior odds ratio is 29.6. [Hint: To get you started, here is how to figure out the signal structure σ. The state space is $S = \{Urn1, Urn2\}$ and a signal is defined by the number of red vs blue chips in the sample of 12 chips drawn (with replacement) from the urn. For instance, $8R4B$ is a signal. What is the probability $\sigma(8R4B|Urn1)$ of this signal conditional on the state being Urn 1? Well, in Urn 1, the probability of red is 0.7 and the probability of blue is 0.3. Therefore, any sequence with 8 reds and 4 blues has probability $0.7^8 0.3^4$, and there are $\frac{12!}{8!4!}$ such sequences. Use these details to figure out the rest.]

22.4 Other

Experiments shows that beliefs about a state s is determined not just by the frequency with which s is observed but also by factors that influence memory, specifically the easy of recall. These factors include things like salience. Tversky and Kahneman (1974) describe an experiment where subjects are briefly shown lists containing the names of well-known women and men. In some lists the women are relatively better known and in others the men are relatively better known. Subjects are asked to assess the proportion of women vs men in the lists. It turned out that subjects' assessments were driven not by the actual proportions, but by the relatively more popular names.

Finally, there is evidence that people's beliefs exhibit anchoring in a manner similar to what we saw in the evidence for abstract choice.

22.5 The Heuristics and Biases Program

The "Heuristics and Biases" program of Kahneman and Tversky (see Tversky and Kahneman 1974) is based on their extensive experimental study of beliefs. As we have seen, they find that beliefs can neither be represented by probability measures, nor is updating Bayesian. Their explanation of the findings is that people's beliefs are intuitive in that they are based on heuristics, that is, mental shortcuts or rules of thumb. Heuristics are evolutionarily adaptive: they help make quick decisions that are "right" in most situations. However, they can also give rise to systematic biases relative to the Bayesian model. The Heuristics and Bias program proposes that the study of beliefs should be directed at understanding the heuristics that people use, and the biases that these can create.

Kahneman and Tversky proposed three heuristic to organize the evidence that they collected:

– The *Representativeness heuristic:* People's update reflects the representativeness of the information rather than the Bayesian update of prior beliefs. Representativeness is based on similarity. The evidence we outlined for static and dynamic beliefs is covered by this heuristic. For instance, beliefs exhibit the gambler's fallacy because a mix of heads and tails is more representative of a fair coin. Beliefs are subject to the conjunction fallacy because "bank teller and activist" is a more representative description of Linda than just "bank teller".

– The *Availability heuristic:* People's evaluation of likelihoods is based on the number of examples they can recall. We discussed this as part of the "other" evidence above.

– The *Anchoring and Adjustment heuristic:* People's beliefs anchor on something possibly irrelevant and then adjust relative to the anchor. We noted this in the "other" evidence as well.

Part VI
CHOICE OVER TIME

23 Discounted Utility Theory

Most actions we take inevitably come with (positive and/or negative) consequences that can potentially extend far into the future. If you work hard tonight, you will be closer to reaching your future goals. If you have fun tonight, you will have more work to do tomorrow to meet your future goals. The consequences of an action could be more complicated. If you work too hard, you will get close to your future goals, but you could also lose your mental efficiency due to exhaustion and thus make it harder for yourself to make sufficient progress tomorrow. In many important economic decisions the consequences are in fact explicit. Decision such as whether to go to college or how much to save or how much to invest are clearly about incurring immediate costs (negative consequences) for future returns (positive consequences).

The point here is just that many current actions are in fact potentially the tip of an iceberg of consequences extending into the future. The theory of intertemporal choice is about how we make choices between actions that have a string of consequences following them. In particular, while we would naturally think in terms of the utility from the consequence in each future period, the main question of interest is how do we trade off consequences across different time periods. It is obvious that effort is painful, and completing a task is good, but what determines how a person chooses between working a little today and a little tomorrow versus doing no work today and doing all the work tomorrow? Intertemporal choice theory studies such questions, and more generally how people evaluate options with delayed consequences.

There are two different types of intertemporal decisions: those made at one point in time (one-shot, or static decisions), and those made across time (dynamic decisions). When you sign up for a certain major, you have made a static decision. Your dynamic choice constitutes this static choice along with your future static choices of whether to stick to this major or change it. We will study static and dynamic choice separately.

23.1 Static Discounted Utility

Most choices that people make in real life affect not just consumption in one period in the future, but consumption in different periods. For instance, the decision to go to school decreases consumption for four years, but improves life-time consumption after that. For such situations, we need a theory of how agents choose between *consumption streams*. The theory we describe below is static in the sense that it involves a one-shot choice of a consumption stream today. Suppose that there are $T+1$ periods. A consumption stream is a vector $(c_0, c_1, c_2, .., c_T)$ that specifies consumption in each of the $T+1$ periods. Consumption does not necessarily have to be in monetary terms. It could also be an action (e.g. study or play) or an object (e.g. burger or salad). Let A denote the set of all consumption streams (that are $T+1$-periods long):[44]

The static Discounted Utility theory (for consumption streams) states that the agent has a preference $\succsim$ defined over the set of consumption streams A, and that $\succsim$ is represented by a utility function of the form:

$$DU(c_0, c_1, c_2, .., c_T) = \sum_{t=0}^{T} \delta^t u(c_t),$$

where the discount factor δ satisfies $0 < \delta < 1$, and the instantaneous utility u is strictly increasing and satisfies $u(0) = 0$.

The theory is called the DU theory or the DU model for short. The theory captures the idea that all that matters to an agent when he is evaluating a consumption stream $(c_0, c_1, c_2, .., c_T)$ is the anticipated utility from consumption in each period, that is, $u(c_0), ..., u(c_T)$, and the date t (that is, temporal distance from the present) of each consumption. According to the DU theory, the agent *discounts* each $u(c_t)$ when it is in the future, and this discounting takes a very special form: utility that is t periods away is discounted by δ^t where $0 < \delta < 1$.

[44] The notation $\sum_{t=0}^{T}$ is short for "sum from 0 to T", and $\sum_{t=0}^{T} \delta^t u(c_t)$ is just a short way of writing $u(c_0) + \delta u(c_1) + \delta^2 u(c_2) + ... + \delta^T u(c_T)$. (Note that $\delta^0 = 1$).

As a function of t, δ^t defines what is called a *discount function*, that is, a specification of what factor the utility from consumption is discounted by when the delay is t. The fact that δ is less than 1 captures the idea of impatience: since δ^t decreases as t increases, the model implies that delayed consumption is worth less than earlier consumption. (We will not talk much about the model with $\delta > 1$, but suffice it to note that this would capture patience). The requirement that $u(0) = 0$ is just the statement that one gets zero utility if one consumes nothing.

The particular form δ^t is called *exponential discounting*. Note that the utility from consumption is discounted by a factor of $\delta^0 = 1$ when it is available immediately, by a factor of $\delta^1 = \delta$ when it is available in the next period, by a factor of δ^2 when it is available two periods later, etc. In particular, every time consumption is delayed by one more period, it is discounted by δ – its attractiveness decreases at a *constant* rate, and this is what defines exponential discounting. For the sake of some perspective, consider a different discount function, the *hyperbolic discount function* $\frac{1}{1+t}$. Here the utility from consumption is discounted by a factor of $\frac{1}{1+0} = 1$ when it is available immediately, by a factor of $\frac{1}{1+1} = \frac{1}{2}$ when it is available in the next period, by a factor of $\frac{1}{1+2} = \frac{1}{3}$ when it is available two periods later, etc. Note that every time the reward is delayed by one more period, its utility is *not* discounted by some constant factor. It is discounted by a factor of $\frac{1}{2}$ when going from period 0 to period 1, by $\frac{2}{3}$ when going from period 1 to period 2, by $\frac{3}{4}$ when going from period 2 to period 3, and so on and so forth. That is, as consumption is delayed by one more period, its attractiveness is reduced at a *decreasing* rate, unlike exponential discounting.

But what is Discounting?

Discounting can be understood in several ways:

1 – To quote Pigou: "our telescopic faculty is defective, and we, therefore, see future pleasures, as it were, on a diminished scale." Similary Bohm-Bawerk asserted: "[It] may be that we possess inadequate power to imagine and to abstract, or that we are not willing to put forth the necessary effort, but in any event we limn a more or less incomplete picture of our future wants and especially of the remotely distant ones." Thus, discounting can be viewed as the result of a psychological constraint: the diminished visibility of

future rewards (and their utility). It is the weakness in our ability to imagine and feel the pleasure attached to a future reward that leads it to carry less weight than immediate pleasure.

2 – Another approach views discounting as psychologically motivated rather than as a contraint. In the words of Rae: "[S]uch pleasures as may now be enjoyed generally awaken a passion strongly prompting to the partaking of them. The actual presence of the immediate object of desire in the mind by exciting the attention, seems to rouse all the faculties, as it were to fix their view on it, and leads them to a very lively conception of the enjoyments which it offers to their instant possession." Thus, even if the agent was able to clearly visualize all future pleasures, he may still discount future rewards relative to immediate rewards because of an *urge for immediate gratification*. So, in this view, discounting is just the weighting of utility in different periods in accordance with their importance to the agent in the present.

4 – Add: Multiple selves and limited social cognition.

3 – The preceding two approaches view discounting as an intrinsic feature of preference – they advocate the existence of what is called a pure time preference. The approach we discuss now views discounting as arising due to the *risk* that is intrinsic to intertemporal choice settings. If consumption c is expected at time t, then there is *always* a chance that the consumption may not be received: on the one hand, something may go wrong and c may not be provided to the agent at t, and on the other, the agent may not be around to enjoy c at t (that is, he faces a mortality risk). Thus, consumption c at time t should always be viewed as a *lottery* that pays c at t with some probability, and 0 otherwise. Then, consumption c at t is discounted in the same sense that an outcome is "discounted" by probabilities in the Expected Utility theory. The following exercise asks you to formalize this message.

Exercise: Suppose that an agent has no pure time preference (that is, $\delta = 1$), and that he makes choices under risk in accordance with the Expected Utility theory. He is offered a consumption stream $(c_0, c_1, c_2, .., c_T)$. Suppose that the only risk the agent perceives is a mortality risk, and in particular, suppose that he faces a constant mortality risk: the probability of surviving any *additional* period is always γ, except that the probability of surviving beyond time T is 0 (sorry about all the morbidity...). He views dying at t as

consuming 0 at times t, $t+1$, $t+2$....

(i) For any consumption stream $(c_0, c_1, c_2, .., c_T)$, the agent perceives different possible outcomes, depending on when he might die:

$$(c_0, 0, 0, .., 0), (c_0, c_1, 0, .., 0), ..(c_0, c_1, c_2, .., c_T).$$

What is the probability of each possible stream? To get you started, note that the probability of the first stream equals the probability of dying at $t = 1$, which is $1 - \gamma$; the probability of the second streams equals the probability of surviving at $t = 1$, but dying at $t = 2$, and so on.

(ii) What is the Expected Utility of the lottery you defined in part (i)? To an outside observer who does not take mortality risk into account, does the agent behave like a DU agent?

(iii) Redo parts (i) and (ii) assuming the mortality risk is of the following type: the probability of surviving to period $t+1$ if the agent is alive at t is $\frac{t+1}{t+2}$.

23.2 Testable Implications of Static DU

We will establish some behavioral implications of the DU model below. As before, let a consumption of 0 denote 'no consumption.' To reduce notational burden, we will denote by $[c, t]$ a consumption stream that pays c at time t and 0 otherwise. For instance, $[c, 2]$ denotes $(0, 0, c, 0, .., 0)$. Note that $DU([c, 2]) = DU(0, 0, c, 0, .., 0) = 0 + \delta 0 + \delta^2 u(c) + \delta^3 0 + .. + \delta^T 0 = \delta^2 u(c)$. Indeed, in general,

$$DU([c, t]) = \delta^t u(c).$$

A stream $[c, t]$ may be referred to as *dated consumption*.

23.2.1 Impatience

Proposition 18 *If $\succsim$ respects the DU model then $\succsim$ exhibits Impatience: for any $c > 0$,*

$$t < t' \Longrightarrow (c, t) \succ (c, t').$$

Proof. Observe that since u is strictly increasing and $u(0) = 0$, it must be that $u(c) > 0$ for any $c > 0$. Then, for any $c > 0$,

$t < t'$
$\Longrightarrow \delta^t > \delta^{t'}$ (since $0 < \delta < 1$)
$\Longrightarrow \delta^t u(c) > \delta^{t'} u(c)$ (since $u(c) > 0$)
$\Longrightarrow [c,t] \succ [c,t']$. ■

Impatience says that the agent would rather have a good reward sooner, and not later. Thus, Impatience says that waiting is undesirable for the agent. Observe the key role of $\delta < 1$ in establishing the proposition. If $\delta > 1$ then we would have the opposite result.

23.2.2 Stationarity

Proposition 19 *If $\succsim$ respects the DU model then $\succsim$ exhibits Stationarity: for any c, c', t, t' and d,*

$$[c,t] \succsim [c',t'] \Longleftrightarrow [c,t+d] \succsim [c',t'+d].$$

Proof. Observe that
$[c,t] \succsim [c',t']$
$\Longleftrightarrow \delta^t u(c) \geq \delta^{t'} u(c')$
$\Longleftrightarrow \delta^d \times \delta^t u(c) \geq \delta^d \times \delta^{t'} u(c')$
$\Longleftrightarrow \delta^{t+d} u(c) \geq \delta^{t'+d} u(c')$
$\Longleftrightarrow [c,t+d] \succsim [c',t'+d]$, as was to be shown. ■

Stationarity states that if the agent prefers having c in period t, rather than c' in t', then his preference does not change if both these dated consumptions are pushed into the future by d periods. Similarly for indifference. This is a key property of the DU model.

The proposition is primarily driven by exponential discounting. This is what makes it possible to simply multiply both sides of the relevant inequality by δ^d to get to the result. Intuitively, the 'constant impatience' feature of exponential discounting causes an irrelevance of common delays.

23.2.3 Separability

We'll consider two notions of Separability:

Proposition 20 *If $\succsim$ respects the DU model then $\succsim$ exhibits Current Separability: for any $c, c', c_1, c_2, ..c_T, c'_1, c'_2, ..c'_T$,*

$$(c, c_1, c_2, ..c_T) \succsim (c, c'_1, c'_2, ..c'_T) \Longleftrightarrow (c', c_1, c_2, ..c_T) \succsim (c', c'_1, c'_2, ..c'_T).$$

Proof. Take any $c, c', c_1, c_2, ..c_T, c'_1, c'_2, ..c'_T$. Then,

$(c, c_1, c_2, ..c_T) \succsim (c, c'_1, c'_2, ..c'_T)$

$\Longleftrightarrow DU(c, c_1, c_2, ..c_T) \geq DU(c, c'_1, c'_2, ..c'_T)$

$\Longleftrightarrow u(c) + \sum_{t=1}^{T} \delta^t u(c_t) \geq u(c) + \sum_{t=1}^{T} \delta^t u(c'_t)$

$\Longleftrightarrow \sum_{t=1}^{T} \delta^t u(c_t) \geq \sum_{t=1}^{T} \delta^t u(c'_t)$

$\Longleftrightarrow u(c') + \sum_{t=1}^{T} \delta^t u(c_t) \geq u(c') + \sum_{t=1}^{T} \delta^t u(c'_t)$

$\Longleftrightarrow DU(c', c_1, c_2, ..c_T) \geq DU(c', c'_1, c'_2, ..c'_T)$

$\Longleftrightarrow (c', c_1, c_2, ..c_T) \succsim (c', c'_1, c'_2, ..c'_T)$, as was to be shown. ■

The left-hand side of the Current Separability proposition says that when current consumption is fixed at c, then the agent prefers the "continuation stream" $c_1, c_2, ..c_T$ over $c'_1, c'_2, ..c'_T$. The right-hand side similarly makes a statement about the ranking of "continuation streams" when current consumption is fixed at c'. Current Separability says that the agent's preference over "continuation streams" is independent of current consumption. That is, what the agent consumes today does not affect how he feels about future consumption.

In a similar fashion, Forward Separability says that what the agent consumes in the future does not affect how he feels about today's consumption.

Proposition 21 *If $\succsim$ respects the DU model then $\succsim$ exhibits Forward Separability: for any $c, c', c_1, c_2, ..c_T, c'_1, c'_2, ..c'_T$,*

$$(c, c_1, c_2, ..c_T) \succsim (c', c_1, c_2, ..c_T) \Longleftrightarrow (c, c'_1, c'_2, ..c'_T) \succsim (c', c'_1, c'_2, ..c'_T).$$

Proof. Exercise. ■

Henceforth, we willl say that $\succsim$ exhibits *Separability* if it exhibits both Current and Forward Separability.

Observe that the additive feature of the DU representation plays the key role in ensuring that Separability holds. Due to additivity, consumption in one period is evaluated independently from consumption in other periods. This is what gives rise to Separability.

23.3 Evidence

Present-biased Preference Reversals:

This is a key finding that contradicts the DU model. A typical example of a present-biased preference reversal is the following:

$$\begin{aligned} [100, 0] &\succ [105, 1] \\ [100, 12] &\prec [105, 13]. \end{aligned}$$

That is, an immediate \$100 is better than \$105 after (say) one month, but preferences reverse when these alternatives are together pushed into the future by a year. This violates Stationarity. The usual interpretation in the literature is that this reflects a *desire for immediate gratification*: the fact that the smaller reward is inferior when delayed, but overwhelmingly attractive when available immediately suggests that the immediate availability appeals to our desire for immediate gratification. This interpretation was first given by psychologists.

Though this is one possible interpretation of preference reversals, this is not the only one. Can you think of others?

Non-Separability

We will present thought-experiments to convince ourselves that Separability may be violated in practice.

For an example of when Forward Separability might be violated, let p denote pizza and b denote burgers, and consider the following consumption streams:

$$\begin{aligned} &(p, b, b, ..., b) \text{ and } (b, b, b, ..., b), \\ &(p, p, p, ..., p) \text{ and } (b, p, p, ..., p). \end{aligned}$$

Most people would find the idea of having pizza everyday or burgers everyday distasteful, and would thus prefer some variety. If this is the case, then one would expect people to prefer $(p, b, b, ..., b)$ over $(b, b, b, ..., b)$ and also, $(b, p, p, ..., p)$ over $(p, p, p, ..., p)$. This violates Forward Separability.

For an example of when Current Separability might be violated, let d denote abusing drugs and b denote burgers. If a potential addict has burgers today, then he may strictly prefer to spend the rest of his life eating burgers rather than abusing drugs:

$$(b, d, d, ..., d) \prec (b, b, b, ..., b).$$

However, if he has drugs today and finds the experience sufficiently enjoyable, he may strictly prefer the opposite:[45]

$$(d, d, d, ..., d) \succ (d, b, b, ..., b).$$

This violates Current Separability.[46]

Finally, for a proper experiment that contradicts the Separability feature of the DU model, consider the following experiment (Loewenstein and Prelec (1993)). Two groups of subjects were told:

> Imagine that over the next five weekends you must decide how to spend your Saturday nights. From each pair of sequences of dinners below, circle the one you would prefer. "Fancy French" refers to a dinner at a fancy French Restaurant. "Fancy Lobster" refers to an exquisite lobster dinner at a 4 star restaurant. Ignore scheduling considerations (e.g., your current plans).

Group 1 was then offered a choice between sequences A and B below, and group 2 was offered a choice between sequences C and D. (H stands for eat at

[45] I'm assuming that the munchies are not strong enough for him to find the idea of burgers for the rest of his life attractive.

[46] The example suggests that addiction leads to violations of Current Separability. However, there are less extreme examples where Current Separability will be violated. Try constructing one involving burgers and pizza.

home, F stands for fancy french, L stands for fancy lobster). The percentage responding in favor is given in brackets at the end of each line.

options\ *weekends*	1	2	3	4	5	
A	F	H	H	H	H	[11%]
B	H	H	F	H	H	[89%]
C	F	H	H	H	L	[49%]
D	H	H	F	H	L	[51%]

These preferences can be explained by a preference for spreading consumption over time.

Exercise: Prove formally that the groups' preferences

$$A \prec B \text{ and } C \sim D$$

contradict the DU model. How is the violation related to Separability?

23.4 Dynamic Discounted Utility

23.4.1 The Model

The standard economic model of dynamic choice is basically an extension of the static DU model. Suppose that there are $T+1$ periods, $t = 0, 1, 2, ...T$. In each period, say period t, the agent has a preference $\succsim_t$ defined over the set A_t consisting of consumption streams of length $T-t+1$, that is, consumption streams of the form $(c_t, c_{t+1}, .., c_T)$.

The (dynamic) Discounted Utility theory states that in each period t the agent has a preference $\succsim_t$ defined over the set of consumption streams A_t, and that there exists a discount factor $0 < \delta < 1$ and an instantaneous utility u (that is independent of t) such that each $\succsim_t$ is represented by a utility function of the form:

$$DU_t(c_t, c_{t+1}, .., c_T) = u(c_t) + \sum_{n=t+1}^{T} \delta^{n-t} u(c_n).$$

That is, $DU_t(c_t, c_{t+1}, .., c_T) = u(c_t) + \delta u(c_{t+1}) + \delta^2 u(c_{t+2}) + ... + \delta^{T-t} u(c_T)$. The theory imposes the static DU model on every preference $\succsim_t$. The crucial point; however, is that each of the preferences $\succsim_t$ share the *same* discount factor δ and the same instantaneous utility function u.

23.4.2 A Testable Implication

We show here that the dynamic DU model has the important property of "dynamic consistency."

Definition: The preferences $\succsim_0, \succsim_1, .., \succsim_T$ exhibit *Dynamic Consistency* if for any $t < T$, $c, c_{t+1}, ..c_T, c'_{t+1}, ..c'_T$,

$$(c, c_{t+1}, c_{t+2}, ..c_T) \succ_t (c, c'_{t+1}, c'_{t+2}, ..c'_T) \Longleftrightarrow (c_{t+1}, c_{t+2}, ..c_T) \succ_{t+1} (c'_{t+1}, c'_{t+2}, ..c'_T).$$

Dynamic Consistency states that preferences do not disagree with one another. Note that, given that period t consumption is fixed at c, the left-hand side says that the preference at t prefers the continuation (i.e. period $t+1$) stream $c_{t+1}, c_{t+2}, ..c_T$ over $c'_{t+1}, c'_{t+2}, ..c'_T$. Dynamic Consistency states that the preference at t prefers one continuation stream to the other if and only if preference at $t+1$ agrees. The existence of disagreement is the definition of Dynamic *In*consistency.

Can you see why it is important for the period t consumption to be fixed?

Proposition 22 *If the preferences $\succsim_0, \succsim_1, .., \succsim_T$ respect the dynamic DU model, then they exhibit Dynamic Consistency.*

Proof. Take any $t < T$, $c, c_{t+1}, ..c_T, c'_{t+1}, ..c'_T$. Note that

$$\begin{aligned}
&(c, c_{t+1}, c_{t+2}, ..c_T) \succsim_t (c, c'_{t+1}, c'_{t+2}, ..c'_T) \\
&\Longleftrightarrow DU_t(c, c_{t+1}, c_{t+2}, ..c_T) \geq DU_t(c, c'_{t+1}, c'_{t+2}, ..c'_T) \\
&\Longleftrightarrow u(c) + \sum_{n=t+1}^{T} \delta^{n-t} u(c_n) \geq u(c) + \sum_{n=t+1}^{T} \delta^{n-t} u(c'_n) \\
&\Longleftrightarrow \sum_{n=t+1}^{T} \delta^{n-t} u(c_n) \geq \sum_{n=t+1}^{T} \delta^{n-t} u(c'_n) \\
&\Longleftrightarrow \delta u(c_{t+1}) + ... + \delta^{T-t} u(c_T) \geq \delta u(c'_{t+1}) + ... + \delta^{T-t} u(c'_T)
\end{aligned}$$

$\iff \frac{1}{\delta}\left(\delta u(c_{t+1}) + ... + \delta^{T-t}u(c_T)\right) \geq \frac{1}{\delta}\left(\delta u(c'_{t+1}) + ... + \delta^{T-t}u(c'_T)\right)$
$\iff u(c_{t+1}) + ... + \delta^{T-t-1}u(c_T) \geq u(c'_{t+1}) + ... + \delta^{T-t-1}u(c'_T)$
$\iff u(c_{t+1}) + ... + \delta^{T-(t+1)}u(c_T) \geq u(c'_{t+1}) + ... + \delta^{T-(t+1)}u(c'_T)$
$\iff u(c_{t+1}) + \sum_{n=t+2}^{T} \delta^{n-t-1}u(c_n) \geq u(c'_{t+1}) + \sum_{n=t+2}^{T} \delta^{n-t-1}u(c'_n)$
$\iff DU_{t+1}(c_{t+1}, c_{t+2}, ..c_T) \geq DU_{t+1}(c'_{t+1}, c'_{t+2}, ..c'_T)$
$\iff (c_{t+1}, c_{t+2}, ..c_T) \succsim_{t+1} (c'_{t+1}, c'_{t+2}, ..c'_T)$, and this establishes the strict preference part of Dynamic Consistency. The indifference part follows an identical argument. ■

23.5 Dynamic Choice: an Illustration

We illustrate how the dynamic DU model can be used to make predictions about what choices an agent makes over time.

Suppose that there are 3 periods, $t = 0, 1, 2$. In each of periods 0 and 1, a student can either study (denoted s) or play (denoted p). In period 2 he makes no choice, but he takes an exam and receives a score that equals the number of periods he had studied. That is, the maximum score is 2; if he does not study at all, he gets 0; if he studies for one period, he gets 50% of the points; if he studies for both periods he gets 100%. His preferences in periods 0 and 1 respect a dynamic DU model with the following specifications:

$$\delta = 0.5, u(s) = -4, u(p) = 4, u(0) = -18, u(1) = 0, u(2) = 40.$$

Thus, studying gives disutility, but so does failing the exam; and playing gives utility, but so does getting a perfect score.

In order to determine the agent's choices, we need to specify his options. In period 0, his set of options is the set of "plans" available to him: $\{(s, s, 2), (p, p, 0), (p, s, 1), (s, p, 1)\}$. Each plan specifies an action for today, an action for tomorrow, and the resulting consumption for the day after. In period 1, he can potentially reconsider his previously chosen plan. His options are $\{(s, 2), (p, 1)\}$ if he studied yesterday, and $\{(s, 1), (p, 0)\}$ if he played yesterday. There is no relevant choice for period 2, as his consumption (exam score) is fully determined by his previous choices.

The DU agent's choices maximize preference, as in the standard theory. Thus we can determine choice in each period – we find the option that maximizes the preference $\succsim_t$, or equivalently, the option that maximizes utility DU_t. So, to determine choice in period 0, we compute:[47]

$$\begin{aligned} DU_0(s,s,2) &= 4 \\ DU_0(p,p,0) &= 1.5 \\ DU_0(p,s,1) &= 2 \\ DU_0(s,p,1) &= -2, \end{aligned}$$

and observe that utility is the highest for $(s,s,2)$. Thus:

$$C_0(\{(s,s,2),(p,p,0),(p,s,1),(s,p,1)\}) = \{(s,s,2)\},$$

that is, he chooses $(s,s,2)$, and thus studies today, and plans to study tomorrow as well.

What does he do in period 1? He has the option of changing his plan if he wants, but he needs to consult his preferences. So, to determine period 1 choice, note that he studied in period 0 thereby facing the choice problem $\{(s,2),(p,1)\}$, and moreover,

$$\begin{aligned} DU_1(s,2) &= 16 \\ DU_1(p,1) &= 4. \end{aligned}$$

Thus, *given that he chose to study in the previous period*, he also finds it optimal to study today – denote this choice by:

$$C_{1,s}(\{(s,2),(p,1)\}) = \{(s,2)\}.$$

The subscript on C indicates the current period, and the consumption in the previous period.

To summarize, in period 0, the agent planned to study in both periods, and began implementing the plan by studying in that period. In period 1, after reconsidering, he decided to stick to the plan and study in the second

[47] Note that the outcome of the exam in period 3 must enter the calculations, since it is not irrelevant for his decision,

period as well. Since he studied in both periods, he received a perfect score (i.e., 2 out of 2 points) in period 2.

An important point to note is that the fact that the agent in period 1 decided to stick with the plan he made in period 0 is not a coincidence, or specific to the setting in the example. *This is a general feature of the dynamic DU model, specifically because of its property of Dynamic Consistency.* The fact that the agent went through with his initial plan is simply the statement that his preferences in different periods are in agreement with one another: the preference $\succsim_1$ agreed with the preference $\succsim_0$. This is a useful thing to know: whenever you need to solve a problem of dynamic choice involving the dynamic DU model, *all you need to do is find the optimal plan in the first period.* You do not need to waste your time checking if the agent will stick to this plan in later periods, because, given the fact that the DU model features dynamic consistency, you already know that he will follow through with his plan.

24 Psychology of Intertemporal Choice

It has long been recognized by philosophers and psychologists that people have self-control problems: they experience a 'desire for immediate gratification' and are often unable to resist it. Many psychologists and economists consider preference reversals and dynamic inconsistency to be manifestations of the desire for immediate gratification. Below, we introduce a model – an alternative to the DU model – that captures desire for immediate gratification. It is a model of an agent who has self-control problems.

24.1 The Static Beta-Delta Model

24.1.1 The Model

Suppose that there are $T+1$ periods. A consumption stream is a vector $(c_0, c_1, c_2, .., c_T)$ that specifies consumption in each of the $T+1$ periods. Let A denote the set of all consumption streams (that are $T+1$-periods long).

The (static) β-δ model states that the agent has a preference $\succsim$ defined over the set of consumption streams A, and that $\succsim$ is represented by a utility function of the form:

$$BDU(c_0, c_1, c_2, .., c_T) = u(c_0) + \beta \left(\sum_{t=1}^{T} \delta^t u(c_t) \right),$$

where β and δ are between 0 and 1, and u satisfies $u(0) = 0$.[48]

The novelty in this model is the β, which is a special discount factor applied to the entire future. This captures a desire for immediate gratification, since such a desire makes you care less about the future, which is equivalent

[48] That is,

$$\begin{aligned} BD(c_0, c_1, c_2, .., c_T) &= u(c_0) + \beta \left[\delta u(c_1) + \delta^2 u(c_2) + ... + \delta^T u(c_T) \right] \\ &= u(c_0) + \beta\delta u(c_1) + \beta\delta^2 u(c_2) + ... + \beta\delta^T u(c_T). \end{aligned}$$

to saying that it makes you care relatively more about the present. This will show up repeatedly in what follows.

Recall that the discount function in the DU model was exponential δ^t. We saw that this embodied the idea that delaying consumption leads to a *constant rate of loss of attractiveness*: each additional period's delay causes additional discounting by a constant factor δ. We also mentioned a different discount function, the *hyperbolic discount function* $\frac{1}{1+t}$ where the loss of attractiveness due to delay occurs at a *consistently decreasing rate*. The discount function in the Beta-Delta Model is an easier-to-work-with version of hyperbolic discounting, and is called *quasi-hyperbolic discounting*. Utility from consumption at $t = 0, 1, 2, .., t, ..$ is discounted by $1, \beta\delta, \beta\delta^2, .., \beta\delta^t, ..$ respectively. Observe that when immediate utility is delayed to $t = 1$, it is discounted by a factor of $\beta\delta$ $(= \frac{\beta\delta}{1})$. But, when utility is already in the future, at some time $t > 0$, then delaying it to $t+1$ causes it to be discounted by a factor of δ $(= \frac{\beta\delta^{t+1}}{\beta\delta^t})$. That is, discounting is relatively steep (at $\beta\delta$) when immediate utility is delayed by one period, but every subsequent one period delay is discounted by a constant factor of δ. This asymmetry between today-tomorrow vs any two consecutive future periods reflects the fact that a desire for immediate gratification is relevant only when today is involved.

24.1.2 Present-biased Preference Reversals

It is straightforward to prove that the static β-δ model satisfies Completeness, Transitivity, Impatience and Separability, just like the DU model. The model can violate Stationarity, however, and indeed it can accommodate *present-biased preference reversals*, which can be defined as follows: there exists at least one time period $t < T$, one delay $0 < d \leq T - t$ and pair of consequences c', c'' such that

$$[c', 0] \succ [c'', d] \text{ and } [c', t] \prec [c'', t + d].$$

We demonstrate this here.

According to the model, the comparison of $[c', 0]$ vs $[c'', d]$ depends on

$$u(c') \text{ vs } \beta\delta^d u(c''),$$

and the comparison of $[c', t]$ vs $[c'', t+d]$ depends on $\beta\delta^t u(c')$ vs $\beta\delta^{t+d}u(c'')$, or equivalently (by dividing both by $\beta\delta^t$),

$$u(c') \text{ vs } \delta^d u(c'').$$

Observe that in the two comparisons, the β is entirely irrelevant when no immediate reward is involved, whereas it becomes relevant when an immediate reward is involved. This reflects the idea that a desire for immediate gratification is irrelevant for distant comparisons, but impacts comparisons involving an immediate reward.

Indeed, because of this differential manner in which β appears, the model can produce preference reversals: if c', c'', δ, d are any values such that

$$u(c') < \delta^d u(c'')$$

then it is easy to see that for a small enough (non-zero) value for β it will be true that

$$u(c') > \beta\delta^d u(c'').$$

This shows that the model can produce the preference reversal: $[c', 0] \succ [c'', d]$ and $[c', t] \prec [c'', t+d]$ for any $t > 0$.

An Aside

We saw above that the comparison $[c', t]$ vs $[c'', t+d]$ depends on $u(c')$ vs $\delta^d u(c'')$. Notice that the former expression involves t but the latter does not. This means that in the model, for any $t, t' > 0$,

$$[c', t] \succsim [c'', t+d] \Longleftrightarrow [c', t'] \succsim [c'', t'+d].$$

In particular, there can be no preference reversals when we consider only delayed rewards. This reflects a particular feature of quasi-hyperbolic discounting: there is a one-shot decline in the agent's degree of impatience when going from today to tomorrow, but after that the impatience is as in the DU model. This is a product of the simple structure of the model, which is meant to be a more parsimonious version of the hyperbolic discounting model, which takes the form

$$U(c_0, c_1, c_2, .., c_T) = u(c_0) + \sum_{t=1}^{T} \frac{1}{1+t} u(c_t).$$

24.2 The Dynamic Beta-Delta Model

24.2.1 Model

Suppose that there are $T+1$ periods. In each period, say period t, the agent has a preference $\succsim_t$ defined over the set A_t consisting of consumptions steams of length $T-t+1$, that is, consumption streams of the form $(c_t, c_{t+1}, .., c_T)$. The agent should be viewed as being divided into separate 'selves,' one for each period. Thus each preference $\succsim_t$ reflects the preference of a self, and indeed, sometimes we will refer to a preference $\succsim_t$ as a self, or as period t self. The reason that such a division of the agent (into selves) makes sense is the 'dynamic inconsistency' that we will discuss in the next subsection.

The (dynamic) β-δ model is defined by a preference $\succsim_t$ over A_t in each period t such that there exist discount factors β and δ between 0 and 1, and an instantaneous utility u (all independent of t) such that each $\succsim_t$ is represented by a utility function of the form:

$$BDU_t(c_t, c_{t+1}, .., c_T) = u(c_t) + \beta \left(\sum_{n=t+1}^{T} \delta^{n-t} u(c_n) \right).$$

The utility function can be expanded and written as:

$$BDU_t(c_t, c_{t+1}, .., c_T) = u(c_t) + \beta\delta u(c_{t+1}) + \beta\delta^2 u(c_{t+2}) + ... + \beta\delta^{T-t} u(c_T).$$

The dynamic β-δ model imposes the static β-δ model on every preference $\succsim_t$, and each of the preferences $\succsim_t$ share the same β, δ and u.

In the dynamic DU model, the hypothesis for choice was simply preference maximization. We will see that the hypothesis for the β-δ model may be different. We address this shortly when we discuss self-awareness and dynamic choice.

24.2.2 A Testable Implication

Say that the preferences $\succsim_0, \succsim_1, .., \succsim_T$ exhibit *dynamic inconsistency* if there exists at least one $t < T$ and $c, c_{t+1}, ..c_T, c'_{t+1}, ..c'_T$ such that

$$(c, c_{t+1}, c_{t+2}, ..c_T) \succ_t (c, c'_{t+1}, c'_{t+2}, ..c'_T) \text{ and } (c_{t+1}, c_{t+2}, ..c_T) \prec_{t+1} (c'_{t+1}, c'_{t+2}, ..c'_T).$$

That is, there is Dynamic Inconsistency if there is some disagreement among the different 'selves' of the agent: some self $\succsim_t$ prefers a continuation stream $c_{t+1}, c_{t+2}, ..c_T$ over $c'_{t+1}, c'_{t+2}, ..c'_T$, but the next period self $\succsim_{t+1}$ prefers the stream $(c'_{t+1}, c'_{t+2}, ..c'_T)$ to $(c_{t+1}, c_{t+2}, ..c_T)$.

Recall that we showed that the static version of the model can exhibit preference reversals. The same kind of argument can be used to establish that for an appropriately chosen u, β and δ, and assuming $T > 1$, there can exist c', c'', d, t such that

$$[c', t] \prec_0 [c'', d + t] \text{ and } [c', t] \succ_t [c'', d + t].$$

Intuitively: at bed time, you may determine that waking up at 7am is better than waking up at 8am. These are preferences expressed about options that lie in the future, and so immediate gratification is irrelevant. But your 7am self is subject to a desire for immediate sleep, and therefore may prefer waking up at 8am to 7am.

24.3 Self-Awareness

When preferences are dynamically inconsistent, dynamic choice depends on how aware an agent is of his dynamic inconsistency – this will be demonstrated in the next section by means of an example. We will consider two levels of self-awareness: naivete and sophistication. A naive β-δ agent (also called a naif) is not aware of his dynamic inconsistency at all. He thinks that whichever plan he adopts today will be respected by future selves, that is, he thinks his future selves will follow through with the plan he adopts today. A sophisticated β-δ agent (sometimes called a sophisticate) is on the other end of the spectrum of self-awareness: he is completely aware of his dynamic inconsistency. He knows precisely what his future selves prefer, and how they will revise (if at all) whichever plan he adopts today. He realizes that in forming his plans, he must *take the behavior of future selves as his constraint*. Clearly, the behavior of a sophisticate will tend to be different, and, in particular, more *strategic* than the behavior of a naif.

Naivete and sophistication are two extreme degrees of self-awareness. One can also think about partially naive agents who have some idea of their self-

control problem (i.e., their dynamic inconsistency), but underestimate how severe the problem is.

24.4 Dynamic Choice: an Illustration

We illustrate how the naive and sophisticated dynamic β-δ models can be used to make predictions about what choices an agent makes over time. Consider the same example we studied when discussing the DU model. Just to recall, the info was:

"Suppose that there are 3 periods, $t = 0, 1, 2$. In each of periods 0 and 1, a student can either study (denoted s) or play (denoted p). In period 2 he makes no choice, but he takes an exam and receives a score that equals the number of periods he had studied. That is, the maximum score is 2: if he does not study at all, he gets 0; if he studies one period, he gets 50% of the points; and if he studies both periods he gets 100%."

Now suppose the student's preferences in periods 0 and 1 respect a dynamic β-δ model with the following specifications:

$$\beta = 0.5, \delta = 0.5, u(s) = -4, u(p) = 4, u(0) = -18, u(1) = 0, u(2) = 40.$$

Note the similarities to the DU model used to determine the students choice: this model has the same δ and u. Thus, if this agent did not have self-control problems (that is, if $\beta = 1$) then he would choose to study in both periods (this was the solution for the DU case).

Recall that the period 0 set of options/plans is $\{(s, s, 2), (p, p, 0), (p, s, 1), (s, p, 1)\}$. The period 1 set of options is $\{(s, 2), (p, 1)\}$ if he studies in period 0, and $\{(s, 1), (p, 0)\}$ if he doesn't. There is no meaningful choice for period 2.

24.4.1 Prediction of the Naive β-δ Model

Suppose the student, in period 0, naively thinks that his future selves will go through with any plan he forms today. Thus, the hypothesis for the naive β-δ model is simple '*preference maximization in every period.*' That is, at any point in time, the naive agent will simply choose the plan that he prefers most. We illustrate this next:

In period 0, the agent will select the plan that maximizes utility BDU_0. So, to determine choice in period 0, compute:

$$\begin{aligned} BDU_0(s,s,2) &= 0 \\ BDU_0(p,p,0) &= 2.75 \\ BDU_0(p,s,1) &= 3 \\ BDU_0(s,p,1) &= -3, \end{aligned}$$

and observe that utility is the highest for $(p,s,1)$. Thus:

$$C_0(\{(s,s,2),(p,p,0),(p,s,1),(s,p,1)\}) = \{(p,s,1)\},$$

that is, he chooses $(p,s,1)$, and thus plays today, but plans to study tomorrow.

What does he do in period 1? He reconsiders the plan adopted by the period 0 self. To determine period 1 choice, note that he played in period 0, so in period 1 he faces the choice problem $\{(s,1),(p,0)\}$. Moreover:

$$\begin{aligned} BDU_1(s,1) &= -4 \\ BDU_1(p,0) &= -0.5. \end{aligned}$$

Thus, *given that he chose to play in the previous period*, in period 1 he also finds it optimal to play – denote this choice by:[49]

$$C_1(\{(s,1),(p,0)\}|p) = \{(p,0)\}.$$

To summarize, in period 0, the agent planned to play today and study tomorrow, and began implementing the plan by playing. In period 1, after reconsidering, he decided to change the plan and play in that period as well. Since he played in both periods, he receives 0% (i.e., 0 out of 2) in period 2.

Recall that in the "no self-control problem" case (i.e. the DU model), the agent chose $(s,s,2)$. In the current case, self-control problems induced him to play today, and his naivete induced him to plan to study tomorrow. In the "no self-control problem" case, the agent followed through with his plan (since the DU model exhibits dynamic consistency), whereas in this case, dynamic inconsistency led him to deviate from his plan.

[49]Read $C_1(\{(s,1),(p,0)\}|p) = \{(p,0)\}$ as "the period 1 choice from $\{(s,1),(p,0)\}$, conditional on having chosen p in period 0, is $(p,0)$".

24.4.2 Prediction of the Sophisticated β-δ Model

The sophisticated student realizes that he will not always follow through with any plan he adopts today. He therefore restricts attention only to *consistent plans*, that is, plans that his future selves will actually follow through with. Then he selects the most preferred consistent plan. Thus the hypothesis for the sophisticated β-δ model is *'choice maximizes preference over the consistent plans.'* We illustrate this now:

In period 0, the agent will first try to assess how future selves will behave. Specifically, he will ask "if I study today, what will my future self do?" and "if I play today, what will my future self do?" That is, in period 1, the agent will first try to assess $C_1(\{(s,2),(p,1)\}|s)$ and $C_1(\{(s,1),(p,0)\}|p)$. Once he has figured out this information, he will know which of his available plans (from $\{(s,s,2),(p,p,0),(p,s,1),(s,p,1)\}$) are in fact consistent.

So, in period 0, the agent will compute that

$$\text{if he plays in period 0, then } BDU_1(s,1) = -4 \text{ and } BDU_1(p,0) = -0.5,$$

and

$$\text{if he studies in period 0, then } BDU_1(s,2) = 6 \text{ and } BDU_1(p,1) = 4.$$

Consequently, he will conclude that

$$\begin{aligned} C_1(\{(s,1),(p,0)\}|p) &= \{(p,0)\} \\ C_1(\{(s,2),(p,1)\}|s) &= \{(s,2)\}. \end{aligned}$$

That is, if he plays in period 0, his future self will not find it worthwhile to study and will end up playing as well. On the other hand, if he studies in period 0, his future self will be motivated to study as well, since doing so will ensure getting a perfect score on the exam, which is something that gives him a lot of utility.

Equipped with this information, he will see that he will never follow through with the plans $(p,s,1)$ and $(s,p,1)$. These plans are inconsistent in the sense that he would never follow through with them. Thus, his *set of consistent plans in period 0 is*

$$\{(p,p,0),(s,s,2)\}.$$

Evidently, the only two consistent/enforceable plans are such that his choice boils down to failing miserably or aceing the exam.

Having figured out which plans are consistent, the agent will now choose the best consistent plan, i.e., the consistent plan that gives the higher utility:

$$\begin{aligned} BDU_0(p,p,0) &= 2.75 \\ BDU_0(s,s,2) &= 0. \end{aligned}$$

Hence, he chooses the plan that leads to failure:[50]

$$C_0(\{(p,p,0),(s,s,2)\}) = \{(p,p,0)\}.$$

To summarize, self 0 will adopt the plan $(p,p,0)$, and implement the plan by playing today. Self 1 will follow through with the plan (since the plan is consistent), and play in period 1. In period 2, he will receive zero marks on his exam. Note that he could have done well if he could only get himself to study today – but his desire for immediate gratification was too strong. Note also that in the naive case, the student at least planned to study tomorrow. In the sophisticated case, the student isn't even planning to try.

Compared to the naive case, we see that sophistication did not lead to any change in the *eventual* behavior and *eventual* outcome: in both cases, the agent played in both periods and failed the exam. In this example, sophistication just made the agent realistic, without affecting eventual behavior. However, it must be stressed that this is just one example: in other examples with different u, β and δ (or in a different setting, with different rules regarding how outcomes depend on behavior), we could have different results. For instance, sophistication could lead the agent to engage in damage-control by studying today and playing tomorrow.

One of the lessons from the above example is that two people may behave the same way, but for entirely different reasons. In particular, the naive and

[50] You may feel frustrated by this choice (as would the agent's mom and dad) because you may have a sense that the agent *should* choose $(s,s,2)$. However, our analysis is purely about what the agent *would* choose, rather than what he *should* choose. Regardless of what an observer or the agent himself believes he should choose, at the end of the day his choices are driven by his urge for immediate gratification.

sophisticated students both ended up not studying and failing their exams. However, one didn't mean for things to happen this way, while the other completely embraced the fact that he was incapable of behaving differently!

24.4.3 Prediction of the Partially Naive β-δ Model

In the sophisticated case, the agent maximizes preferences over the set of plans he will carry out. We do not assume that the agent can look into the future and observe his behavior, but rather that he can *correctly anticipate* his future behavior. Partial naivete arises when the agent does not necessarily anticipate his future behavior correctly. Choice in this case is determined similarly to the sophisticated case. In the sophisticated case, the agent maximizes preferences over his set of consistent plans, whereas in the partially naive case, he maximizes preferences over the set of plans he *thinks* are consistent.

In order to determine the agent's choices, you need to know what he thinks his preferences will look like in the next period – you need to know his perceived preferences. Suppose that he overestimates just his desire for immediate gratification β, but otherwise understands his δ and u. Specifically, suppose he thinks his β in the next period will be $\widehat{\beta} = 0.95$. Determining his choices is left to you as an exercise.

24.4.4 Tie-Break Rule

In the example we've been studying, no self was ever indifferent between two options. But, what if there is an example where a self is indifferent between two options while a previous self has a strict preference? We usually adopt the following tie-break rule: *ties are broken in favor of the preferences of the previous self.* For instance, if self 0 plays today and strictly prefers $(p, s, 1)$ to $(p, p, 0)$, and if self 1 were indifferent between $(s, 1)$ and $(p, 0)$, then the tie-break rule dictates that self 1 will choose $(s, 1)$, because self 0 strictly prefers his future self to choose $(s, 1)$ rather than $(p, 0)$.

24.5 Preference for Commitment

Observe that in the sophisticated case in the previous section, the plan $(p, s, 1)$ was not a consistent one. However, suppose the agent had some means of *enforcing* this plan. That is, suppose he had available to him a device that would commit him to the plan.[51] Would he choose it? The answer is clearly yes. The utility from committing to $(p, s, 1)$ is $BDU_0(p, s, 1) = 3$, while his choice without a commitment device (as in the example) was $(p, p, 0)$, which gives just $BDU_0(p, p, 0) = 2.75$ units of utility.

This demonstrates that sophisticated (and partially naive) agents with self-control problems would desire commitment devices – a demand for commitment devices is a testable implication of dynamic inconsistency coupled with at least some sophistication. Such agents strictly prefer to tie the hands of their future selves so that (what the current self considers to be) the best plan will be enforced. In sharp constrast, DU agents and naive β-δ agents have no use for commitment devices. DU agents are dynamically consistent and thus will always follow through with the best available plan. Naive β-δ agents are not dynamically consistent, but they think that they are: they think that they will follow through with whichever plan they consider best.

It is interesting to note that in reality, there are several examples of commitment devices in the market: rehabilitation for addicts, antabuse/disulfirum for alcoholics[52], savings vehicles (such as 401(k)s and IRAs) for people trying to save for retirement, and stomach stapling for people desperate to lose

[51] For instance, the student could make a bet with his friends, telling them that he is going to study like crazy tomorrow. This bet would serve the purpose of forcing him to study tomorrow because of the fear of being ridiculed by his friends if he doesn't study. Such a bet plays the role of a commitment device. Perhaps a more effective commitment device would be giving an ex-girlfriend written permission to shoot him if he does not study tomorrow.

[52] Drinking alcohol while on disulfirum can cause serious effects that can last from 30 minutes to several hours. It produces an unpleasant reaction of flushing, headache, nausea, vomiting, dizziness, sweating, heart palpitations, and blurred vision or weakness when even small amounts of alcohol are ingested. Severe reactions can include respiratory depression, cardiovascular collapse, myocardial infarction, acute congestive heart failure, unconsciousness, arrhythmias, convulsions, and death (whoa!). These disulfiram-alcohol reactions can occur *up to two weeks after the medication has been stopped.* That is, commitment is ensured for up to two weeks when on this medication.

weight. This suggests that, indeed, people have self-control problems, and that they are also aware of them.