Chapter 2

Rationality, Preferences, and Expected Utility

Before analyzing games we must be precise about what a "payoff" is and what "rational" means. This chapter builds the decision-theoretic foundation on which all of game theory rests. Readers eager for games may skim it and return later; but the ideas here—especially expected utility—are used constantly from Chapter 5 onward.

2.1 Preferences and utility

Decision theory starts not from numbers but from *preferences*. Let X be a set of outcomes. A decision-maker's preferences are captured by a binary relation $\succsim$ on X, where $x \succsim y$ reads "x is at least as good as y." We write $x \succ y$ (strict preference) when $x \succsim y$ but not $y \succsim x$, and $x \sim y$ (indifference) when both hold.

Definition 2.1: Rational preference relation

A preference relation $\succsim$ on X is *rational* if it is

- **complete**: for all $x, y \in X$, either $x \succsim y$ or $y \succsim x$ (or both); and
- **transitive**: for all $x, y, z \in X$, if $x \succsim y$ and $y \succsim z$ then $x \succsim z$.

Completeness says the decision-maker can always rank alternatives; transitivity rules out preference cycles. Neither is innocuous—real people violate both in experiments—but together they permit a dramatic simplification:

Theorem 2.1: Utility representation

If X is finite (or countable) and $\succsim$ is rational, then there exists a *utility function* $u : X \to \mathbb{R}$ such that

$$x \succsim y \iff u(x) \geq u(y).$$

Moreover, u is unique only up to strictly increasing transformations: if f is strictly increasing, $f \circ u$ represents the same preferences.

Proof. Enumerate $X = \{x_1, x_2, \dots\}$ and set $u(x) = \#\{y \in X : x \succsim y\}$ when X is finite (the countable case requires a small extra construction). Completeness and transitivity guarantee $x \succsim y \iff u(x) \geq u(y)$: transitivity implies that the set of outcomes weakly worse than y is

contained in the set weakly worse than x whenever $x \succsim y$, giving $u(x) \geq u(y)$; completeness gives the converse direction. □

Remark: Ordinal vs. cardinal utility

Theorem 2.1 delivers *ordinal* utility: only the ranking matters, not magnitudes. Saying "$u(x) = 10$ and $u(y) = 5$, so x is twice as good" is meaningless at this stage. Cardinal meaning arrives only with risk, below.

2.2 Choice under risk: lotteries

Games force players to evaluate *risky* prospects: if my opponent randomizes, my action induces a probability distribution over outcomes. Let $X = \{x_1, \dots, x_n\}$ be a finite outcome set. A *lottery* is a probability distribution $p = (p_1, \dots, p_n)$ with $p_i \geq 0$ and $\sum_i p_i = 1$. Let $\Delta(X)$ denote the set of all lotteries over X. The decision-maker now has preferences $\succsim$ over $\Delta(X)$, not just over X.

Definition 2.2: von Neumann–Morgenstern axioms

A preference relation $\succsim$ on $\Delta(X)$ satisfies the vNM axioms if it is rational (complete, transitive) and additionally:

- **Continuity**: for all p, q, r with $p \succ q \succ r$, there exist $\alpha, \beta \in (0, 1)$ such that $\alpha p + (1-\alpha) r \succ q \succ \beta p + (1-\beta) r$.
- **Independence**: for all p, q, r and $\alpha \in (0, 1]$,

$$p \succsim q \iff \alpha p + (1-\alpha) r \succsim \alpha q + (1-\alpha) r.$$

Independence is the crucial (and most debated) axiom: mixing both lotteries with a common third lottery should not reverse the ranking.

Theorem 2.2: Expected utility (von Neumann–Morgenstern, 1944)

$\succsim$ on $\Delta(X)$ satisfies the vNM axioms if and only if there exists $u : X \to \mathbb{R}$ (a *Bernoulli utility function*) such that for all lotteries p, q:

$$p \succsim q \iff \sum_{i=1}^{n} p_i \, u(x_i) \;\geq\; \sum_{i=1}^{n} q_i \, u(x_i).$$

Moreover u is unique up to positive affine transformations: $v = au + b$ with $a > 0$ represents the same preferences.

Key Idea: Why game payoffs are expected utilities

Theorem 2.2 is the license for everything we do later: when we write a payoff matrix and compute *expected* payoffs against a randomizing opponent (Chapter 5), we are implicitly assuming players are vNM expected-utility maximizers. The affine-uniqueness clause tells us which operations on payoffs are harmless: adding a constant to all of a player's payoffs, or rescaling them by a positive factor, changes nothing strategically.

2.3 Risk attitudes

Suppose outcomes are money amounts. A decision-maker with Bernoulli utility u facing a lottery with random prize W compares $\mathbb{E}[u(W)]$ with $u(\mathbb{E}[W])$.

Definition 2.3: Risk aversion

A decision-maker is *risk-averse* if she prefers the expected value of any lottery for sure to the lottery itself: $u(\mathbb{E}[W]) \geq \mathbb{E}[u(W)]$ for all W. By Jensen's inequality this holds iff u is concave. She is *risk-neutral* iff u is affine, and *risk-loving* iff u is convex.

Example 2.1: The St. Petersburg puzzle

A fair coin is flipped until the first heads; if this happens on flip k you win 2^k dollars. The expected prize is $\sum_{k=1}^{\infty} 2^{-k} \cdot 2^k = \sum_k 1 = \infty$, yet nobody would pay more than a modest amount to play. Daniel Bernoulli's 1738 resolution: people maximize expected *utility*, not expected money. With $u(x) = \log_2 x$, the certainty equivalent is finite and small: $\mathbb{E}[u] = \sum_{k \geq 1} 2^{-k} \cdot k = 2$, i.e. the gamble is worth $2^2 = 4$ dollars. This is the historical origin of expected utility theory.

2.4 Beliefs, best responses, and rationality in games

In a game, a player's optimal action depends on what the others do—which she does not observe in a simultaneous-move game. The modern view: a rational player forms a *belief* (a probability distribution over opponents' strategy choices) and chooses a *best response* to that belief.

Definition 2.4: Best response to a belief

Let player i have belief μ over the opponents' action profiles A_{-i}. Action a_i is a *best response to* μ if

$$\sum_{a_{-i}} \mu(a_{-i})\, u_i(a_i, a_{-i}) \;\geq\; \sum_{a_{-i}} \mu(a_{-i})\, u_i(a_i', a_{-i}) \quad \text{for every } a_i' \in A_i.$$

A player is *rational* if she plays a best response to *some* belief.

Rationality alone predicts little: almost any action is a best response to some belief. Predictive power comes from assumptions about what players believe about *each other*:

Definition 2.5: Common knowledge

An event E is *common knowledge* if everyone knows E, everyone knows that everyone knows E, everyone knows that everyone knows that everyone knows E, and so on ad infinitum.

The classical assumption of game theory is *common knowledge of rationality* (CKR): all players are rational, all know this, all know that all know it, etc. In Chapter 3 we will see exactly what CKR implies (iterated elimination of dominated strategies / rationalizability), and in Chapter 4 what additional assumptions Nash equilibrium requires.

Example 2.2: The email/coordinated-attack puzzle

Two generals must attack simultaneously to win. General 1 sees the enemy is weak and emails General 2: "attack at dawn." Messages fail with small probability, so General 2

replies to confirm, General 1 confirms the confirmation, *ad infinitum*. After any finite number of confirmations, "attack at dawn" is mutually known to high order but never *common* knowledge—and one can show coordinated attack cannot be guaranteed. The example (Rubinstein's email game) shows common knowledge is a genuinely stronger and more delicate notion than "everybody knows."

Exercises

Exercise 2.1: Transitivity and money pumps

A decision-maker has cyclic preferences: $x \succ y$, $y \succ z$, $z \succ x$. Show that if she is willing to pay a small amount $\varepsilon > 0$ to swap any held item for a strictly preferred one, a trader holding x, y, z can extract unbounded money from her. (This "money pump" is the classic pragmatic defense of transitivity.)

Exercise 2.2: Affine invariance

Player 1's payoffs in some game are given by u_1. Show that replacing u_1 by $v_1 = 3u_1 + 7$ leaves her set of best responses to every belief unchanged. Then give an example showing that a *nonlinear* increasing transformation (say $u \mapsto u^3$ with payoffs that can be negative) can change best responses against a *randomizing* opponent even though it preserves the ordinal ranking of pure outcomes.

Exercise 2.3: Certainty equivalents

An agent has Bernoulli utility $u(x) = \sqrt{x}$ and faces a lottery paying 100 with probability $1/2$ and 0 otherwise. Compute the expected value, the expected utility, the certainty equivalent, and the risk premium.

Exercise 2.4: Allais paradox

Consider lotteries over prizes $\{0, 1\text{M}, 5\text{M}\}$ (millions): $A = (0, 1, 0)$ (1M for sure); $B = (0.01, 0.89, 0.10)$; $C = (0.89, 0.11, 0)$; $D = (0.90, 0, 0.10)$. Most people report $A \succ B$ and $D \succ C$. Show these two rankings jointly violate the independence axiom.