Expected Utility and Harsanyi Aggregation

Aaron Roth

University of Pennsylvania

Fall 2026

Overview

- A policy is a lottery over completions, so evaluating policies requires preferences over lotteries.

Overview

- A policy is a lottery over completions, so evaluating policies requires preferences over lotteries.
- The von Neumann–Morgenstern (VNM) axioms characterize exactly when an agent can be modeled as maximizing expected utility.

Overview

- A policy is a lottery over completions, so evaluating policies requires preferences over lotteries.
- The von Neumann–Morgenstern (VNM) axioms characterize exactly when an agent can be modeled as maximizing expected utility.
- Harsanyi's theorem characterizes the aggregate expected-utility preferences that satisfy the Pareto property (weighted social welfare).

Richer Information Changes the Aggregation Question

The Arrow lecture began with ordinal rankings of deterministic completions.

Richer Information Changes the Aggregation Question

The Arrow lecture began with ordinal rankings of deterministic completions.

	Arrow	Today
Objects ranked	completions	lotteries over completions

Richer Information Changes the Aggregation Question

The Arrow lecture began with ordinal rankings of deterministic completions.

	Arrow	Today
Objects ranked	completions	lotteries over completions
Individual input	strict order	ordinal preference over lotteries

Richer Information Changes the Aggregation Question

The Arrow lecture began with ordinal rankings of deterministic completions.

	Arrow	Today
Objects ranked	completions	lotteries over completions
Individual input	strict order	ordinal preference over lotteries
Aggregate output	strict order	cardinal preference over lotteries

Richer Information Changes the Aggregation Question

The Arrow lecture began with ordinal rankings of deterministic completions.

	Arrow	Today
Objects ranked	completions	lotteries over completions
Individual input	strict order	ordinal preference over lotteries
Aggregate output	strict order	cardinal preference over lotteries

The added structure will permit positive aggregation results—but it also requires information that ordinal comparisons on individual completions do not contain.

A Language-Model Policy Is a Lottery

Fix one prompt and a finite set C of candidate completions.

A Language-Model Policy Is a Lottery

Fix one prompt and a finite set C of candidate completions.

Definition

The set of *lotteries* over C is

$$\Delta(C) := \left\{ p \in \mathbb{R}_+^C : \sum_{j \in C} p_j = 1 \right\}.$$

A Language-Model Policy Is a Lottery

Fix one prompt and a finite set C of candidate completions.

Definition

The set of *lotteries* over C is

$$\Delta(C) := \left\{ p \in \mathbb{R}_+^C : \sum_{j \in C} p_j = 1 \right\}.$$

The unit vector e_j is the *pure lottery* that selects completion j with probability one.

A Language-Model Policy Is a Lottery

Fix one prompt and a finite set C of candidate completions.

Definition

The set of *lotteries* over C is

$$\Delta(C) := \left\{ p \in \mathbb{R}_+^C : \sum_{j \in C} p_j = 1 \right\}.$$

The unit vector e_j is the *pure lottery* that selects completion j with probability one.

- A completion is an element $j \in C$.
- A policy is a lottery $p \in \Delta(C)$.

Rankings of Completions Do Not Rank Every Policy

Suppose one person ranks three completions

$$a \succ b \succ c.$$

Rankings of Completions Do Not Rank Every Policy

Suppose one person ranks three completions

$$a \succ b \succ c.$$

This tells us how the person compares the pure lotteries e_a, e_b, e_c.

Rankings of Completions Do Not Rank Every Policy

Suppose one person ranks three completions

$$a \succ b \succ c.$$

This tells us how the person compares the pure lotteries e_a, e_b, e_c.
It does not tell us whether

$$e_b \quad \text{or} \quad \frac{1}{2}e_a + \frac{1}{2}e_c$$

is preferred.

Rankings of Completions Do Not Rank Every Policy

Suppose one person ranks three completions

$$a \succ b \succ c.$$

This tells us how the person compares the pure lotteries e_a, e_b, e_c. It does not tell us whether

$$e_b \quad \text{or} \quad \frac{1}{2}e_a + \frac{1}{2}e_c$$

is preferred.

To evaluate randomized policies, we need preferences defined on all of $\Delta(C)$.

Weak Preferences Allow Indifference

Let $\succeq$ be a preference relation on $\Delta(C)$.

Weak Preferences Allow Indifference

Let $\succeq$ be a preference relation on $\Delta(C)$.

- $p \succeq q$ means that p is weakly preferred to q.
- $p \sim q$ means $p \succeq q$ and $q \succeq p$.
- $p \succ q$ means $p \succeq q$ but not $q \succeq p$.

Weak Preferences Allow Indifference

Let $\succeq$ be a preference relation on $\Delta(C)$.

- $p \succeq q$ means that p is weakly preferred to q.
- $p \sim q$ means $p \succeq q$ and $q \succeq p$.
- $p \succ q$ means $p \succeq q$ but not $q \succeq p$.

The first two VNM axioms require $\succeq$ to be a weak order:

1. *Completeness:* for every p, q, either $p \succeq q$ or $q \succeq p$.
2. *Transitivity:* $p \succeq q$ and $q \succeq r$ imply $p \succeq r$.

Continuity Interpolates via Randomization

Definition

The preference satisfies *continuity* if

$$p \succeq q \succeq r$$

implies that, for some $\alpha \in [0, 1]$,

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

Continuity Interpolates via Randomization

Definition

The preference satisfies *continuity* if

$$p \succeq q \succeq r$$

implies that, for some $\alpha \in [0, 1]$,

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

As the probability assigned to the better lottery p rises from zero to one, the resulting mixture cannot pass discontinuously from below q to above q: at some intermediate probability, it must be indifferent to q.

Independence Makes Common Mixtures Cancel

Definition

The preference satisfies *independence* if, for every p, q, r and every $\alpha \in (0,1)$,

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

Independence Makes Common Mixtures Cancel

Definition

The preference satisfies *independence* if, for every p, q, r and every $\alpha \in (0, 1)$,

$$p \succeq q \quad \Longleftrightarrow \quad \alpha p + (1-\alpha) r \succeq \alpha q + (1-\alpha) r.$$

Mixing both options with the same third lottery cannot reverse their order.

VNM Characterizes Expected-Utility Maximization

Theorem (von Neumann–Morgenstern)

A preference on $\Delta(C)$ satisfies completeness, transitivity, continuity, and independence if and only if there is a utility vector $u \in \mathbb{R}^C$ such that

$$p \succeq q \quad \iff \quad u^\top p \geq u^\top q.$$

Here

$$u^\top p = \sum_{j \in C} p_j u(j) = \mathbb{E}_{J \sim p}[u(J)].$$

Thus choosing a most-preferred available lottery is exactly expected-utility maximization. Conversely, every expected-utility preference satisfies the four axioms.

Expected Utility Calibrates Completions Against Lotteries

Suppose a is best, c is worst, and

$$e_b \sim 0.7 e_a + 0.3 e_c.$$

Expected Utility Calibrates Completions Against Lotteries

Suppose a is best, c is worst, and

$$e_b \sim 0.7e_a + 0.3e_c.$$

Normalize $u(a) = 1$ and $u(c) = 0$. The indifference forces

$$u(b) = 0.7.$$

Expected Utility Calibrates Completions Against Lotteries

Suppose a is best, c is worst, and

$$e_b \sim 0.7e_a + 0.3e_c.$$

Normalize $u(a) = 1$ and $u(c) = 0$. The indifference forces

$$u(b) = 0.7.$$

A policy $p = (0.2, 0.5, 0.3)$ over (a, b, c) then has expected utility

$$u^\top p = \mathbb{E}_{J \sim p}[u(J)] = 0.2(1) + 0.5(0.7) + 0.3(0) = 0.55.$$

Expected Utility $\Longrightarrow$ the VNM Axioms

Suppose $p \succeq q$ exactly when $u^\top p \geq u^\top q$.

Expected Utility $\implies$ the VNM Axioms

Suppose $p \succeq q$ exactly when $u^\top p \geq u^\top q$. Completeness and transitivity follow from the usual order on $\mathbb{R}$. Now suppose $p \succeq q \succeq r$ and write

$$U_p = u^\top p, \qquad U_q = u^\top q, \qquad U_r = u^\top r.$$

Expected Utility $\implies$ the VNM Axioms

Suppose $p \succeq q$ exactly when $u^\top p \geq u^\top q$. Completeness and transitivity follow from the usual order on $\mathbb{R}$. Now suppose $p \succeq q \succeq r$ and write

$$U_p = u^\top p, \qquad U_q = u^\top q, \qquad U_r = u^\top r.$$

If $U_p = U_r$, then all three are equal and we may take $\alpha = 0$. Otherwise, $U_p > U_r$. To make the mixture indifferent to q, its expected utility must satisfy

$$\alpha U_p + (1 - \alpha) U_r = U_q.$$

Expected Utility $\implies$ the VNM Axioms

Suppose $p \succeq q$ exactly when $u^\top p \geq u^\top q$. Completeness and transitivity follow from the usual order on $\mathbb{R}$. Now suppose $p \succeq q \succeq r$ and write

$$U_p = u^\top p, \qquad U_q = u^\top q, \qquad U_r = u^\top r.$$

If $U_p = U_r$, then all three are equal and we may take $\alpha = 0$. Otherwise, $U_p > U_r$. To make the mixture indifferent to q, its expected utility must satisfy

$$\alpha U_p + (1-\alpha) U_r = U_q.$$

Solving,

$$\alpha = \frac{U_q - U_r}{U_p - U_r}.$$

Since $U_p \geq U_q \geq U_r$, this α lies in $[0, 1]$. By construction, the mixture has expected utility $U_q = u^\top q$, so $q \sim \alpha p + (1-\alpha) r$.

Expected Utility Implies Independence

For every $\alpha \in (0, 1)$,

$$\begin{aligned} u^\top\big(\alpha p + (1-\alpha) r\big) - u^\top\big(\alpha q + (1-\alpha) r\big) \\ = \alpha u^\top (p - q). \end{aligned}$$

Expected Utility Implies Independence

For every $\alpha \in (0, 1)$,

$$\begin{aligned} u^\top\big(\alpha p + (1-\alpha)r\big) - u^\top\big(\alpha q + (1-\alpha)r\big) \\ = \alpha u^\top(p - q). \end{aligned}$$

Because $\alpha > 0$, the mixed-lottery difference has the same sign as the original utility difference.

Expected Utility Implies Independence

For every $\alpha \in (0, 1)$,

$$u^\top \big(\alpha p + (1-\alpha) r\big) - u^\top \big(\alpha q + (1-\alpha) r\big) \\ = \alpha u^\top (p - q).$$

Because $\alpha > 0$, the mixed-lottery difference has the same sign as the original utility difference. Therefore

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

The VNM Axioms $\Longrightarrow$ Expected Utility

Now assume that $\succeq$ satisfies the four VNM axioms. We will construct a representing utility vector u:

1. Choose a best pure completion j^+ and a worst one j^-.

The VNM Axioms $\Longrightarrow$ Expected Utility

Now assume that $\succeq$ satisfies the four VNM axioms. We will construct a representing utility vector u:

1. Choose a best pure completion j^+ and a worst one j^-.
2. Show that the probability of j^+ strictly orders best–worst lotteries.

The VNM Axioms $\Longrightarrow$ Expected Utility

Now assume that $\succeq$ satisfies the four VNM axioms. We will construct a representing utility vector u:

1. Choose a best pure completion j^+ and a worst one j^-.
2. Show that the probability of j^+ strictly orders best–worst lotteries.
3. Use continuity to match every pure completion to a unique best–worst lottery.

The VNM Axioms $\Longrightarrow$ Expected Utility

Now assume that $\succeq$ satisfies the four VNM axioms. We will construct a representing utility vector u:

1. Choose a best pure completion j^+ and a worst one j^-.
2. Show that the probability of j^+ strictly orders best–worst lotteries.
3. Use continuity to match every pure completion to a unique best–worst lottery.
4. Use independence to substitute these equivalent lotteries inside any policy.

The VNM Axioms $\Longrightarrow$ Expected Utility

Now assume that $\succeq$ satisfies the four VNM axioms. We will construct a representing utility vector u:

1. Choose a best pure completion j^+ and a worst one j^-.
2. Show that the probability of j^+ strictly orders best–worst lotteries.
3. Use continuity to match every pure completion to a unique best–worst lottery.
4. Use independence to substitute these equivalent lotteries inside any policy.
5. Read off an expected-utility representation and prove its uniqueness.

Choose Best and Worst Pure Completions

If every lottery is indifferent, $u = 0$ represents the preference.

Choose Best and Worst Pure Completions

If every lottery is indifferent, $u = 0$ represents the preference. Otherwise, some pure completions are strictly ordered. Indeed, if all pure completions were indifferent, independence would let us substitute one pure component for another, making all lotteries indifferent.

Choose Best and Worst Pure Completions

If every lottery is indifferent, $u = 0$ represents the preference. Otherwise, some pure completions are strictly ordered. Indeed, if all pure completions were indifferent, independence would let us substitute one pure component for another, making all lotteries indifferent. Because C is finite, choose a best completion j^+ and a worst completion j^-, with

$$e_{j^+} \succ e_{j^-}.$$

Choose Best and Worst Pure Completions

If every lottery is indifferent, $u = 0$ represents the preference. Otherwise, some pure completions are strictly ordered. Indeed, if all pure completions were indifferent, independence would let us substitute one pure component for another, making all lotteries indifferent. Because C is finite, choose a best completion j^+ and a worst completion j^-, with

$$e_{j^+} \succ e_{j^-}.$$

Define the best–worst lotteries

$$\ell(\alpha) := \alpha e_{j^+} + (1-\alpha) e_{j^-}, \qquad \alpha \in [0,1].$$

Independence Also Preserves Strict Preference

Suppose $p \succ q$. Independence first gives

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

Independence Also Preserves Strict Preference

Suppose $p \succ q$. Independence first gives

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

Could the reverse weak comparison also hold?

Independence Also Preserves Strict Preference

Suppose $p \succ q$. Independence first gives

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

Could the reverse weak comparison also hold? If it did, independence applied to the comparison with p and q interchanged would imply $q \succeq p$.

Independence Also Preserves Strict Preference

Suppose $p \succ q$. Independence first gives

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

Could the reverse weak comparison also hold? If it did, independence applied to the comparison with p and q interchanged would imply $q \succeq p$. That contradicts $p \succ q$. Therefore

$$p \succ q \quad \Longrightarrow \quad \alpha p + (1-\alpha) r \succ \alpha q + (1-\alpha) r.$$

More Weight on the Best Completion Is Strictly Better

Fix two mixing probabilities $0 \leq \gamma < \alpha \leq 1$. We will show that

$$\ell(\alpha) \succ \ell(\gamma).$$

For $0 < \alpha < 1$, strict-preference preservation gives

$$\ell(\alpha) = \alpha e_{j^+} + (1 - \alpha) e_{j^-} \succ e_{j^-};$$

when $\alpha = 1$, the same comparison holds directly.

More Weight on the Best Completion Is Strictly Better

Fix two mixing probabilities $0 \leq \gamma < \alpha \leq 1$. We will show that

$$\ell(\alpha) \succ \ell(\gamma).$$

For $0 < \alpha < 1$, strict-preference preservation gives

$$\ell(\alpha) = \alpha e_{j^+} + (1 - \alpha) e_{j^-} \succ e_{j^-};$$

when $\alpha = 1$, the same comparison holds directly. If $\gamma = 0$, then $\ell(\gamma) = e_{j^-}$, and we are done. Otherwise, set $\lambda = \gamma/\alpha \in (0, 1)$. Direct calculation gives

$$\ell(\gamma) = \lambda \ell(\alpha) + (1 - \lambda) e_{j^-}.$$

More Weight on the Best Completion Is Strictly Better

Fix two mixing probabilities $0 \leq \gamma < \alpha \leq 1$. We will show that

$$\ell(\alpha) \succ \ell(\gamma).$$

For $0 < \alpha < 1$, strict-preference preservation gives

$$\ell(\alpha) = \alpha e_{j^+} + (1-\alpha)e_{j^-} \succ e_{j^-};$$

when $\alpha = 1$, the same comparison holds directly. If $\gamma = 0$, then $\ell(\gamma) = e_{j^-}$, and we are done. Otherwise, set $\lambda = \gamma/\alpha \in (0,1)$. Direct calculation gives

$$\ell(\gamma) = \lambda \ell(\alpha) + (1-\lambda)e_{j^-}.$$

Strict-preference preservation therefore gives

$$\ell(\alpha) = \lambda \ell(\alpha) + (1-\lambda)\ell(\alpha) \succ \lambda \ell(\alpha) + (1-\lambda)e_{j^-} = \ell(\gamma).$$

More Weight on the Best Completion Is Strictly Better

Fix two mixing probabilities $0 \leq \gamma < \alpha \leq 1$. We will show that

$$\ell(\alpha) \succ \ell(\gamma).$$

For $0 < \alpha < 1$, strict-preference preservation gives

$$\ell(\alpha) = \alpha e_{j^+} + (1-\alpha)e_{j^-} \succ e_{j^-};$$

when $\alpha = 1$, the same comparison holds directly. If $\gamma = 0$, then $\ell(\gamma) = e_{j^-}$, and we are done. Otherwise, set $\lambda = \gamma/\alpha \in (0,1)$. Direct calculation gives

$$\ell(\gamma) = \lambda\ell(\alpha) + (1-\lambda)e_{j^-}.$$

Strict-preference preservation therefore gives

$$\ell(\alpha) = \lambda\ell(\alpha) + (1-\lambda)\ell(\alpha) \succ \lambda\ell(\alpha) + (1-\lambda)e_{j^-} = \ell(\gamma).$$

Thus $\ell(\alpha) \succ \ell(\gamma)$ whenever $0 \leq \gamma < \alpha \leq 1$: more weight on the best completion is strictly better.

Continuity Matches Each Completion to a Lottery

Every pure completion j lies between the best and worst:

$$e_{j^+} \succeq e_j \succeq e_{j^-}.$$

Continuity Matches Each Completion to a Lottery

Every pure completion j lies between the best and worst:

$$e_{j^+} \succeq e_j \succeq e_{j^-}.$$

The continuity axiom applies directly to these three lotteries. For some $\alpha_j \in [0, 1]$,

$$e_j \sim \alpha_j e_{j^+} + (1 - \alpha_j) e_{j^-} = \ell(\alpha_j).$$

Continuity Matches Each Completion to a Lottery

Every pure completion j lies between the best and worst:

$$e_{j^+} \succeq e_j \succeq e_{j^-}.$$

The continuity axiom applies directly to these three lotteries. For some $\alpha_j \in [0, 1]$,

$$e_j \sim \alpha_j e_{j^+} + (1 - \alpha_j) e_{j^-} = \ell(\alpha_j).$$

Strict ordering of the $\ell(\alpha)$ makes α_j unique.

Continuity Matches Each Completion to a Lottery

Every pure completion j lies between the best and worst:

$$e_{j^+} \succeq e_j \succeq e_{j^-}.$$

The continuity axiom applies directly to these three lotteries. For some $\alpha_j \in [0, 1]$,

$$e_j \sim \alpha_j e_{j^+} + (1 - \alpha_j) e_{j^-} = \ell(\alpha_j).$$

Strict ordering of the $\ell(\alpha)$ makes α_j unique.

This mixing probability will become the utility of completion j.

Independence Preserves Indifference

The relation $e_j \sim \ell(\alpha_j)$ means that both

$$e_j \succeq \ell(\alpha_j) \qquad \text{and} \qquad \ell(\alpha_j) \succeq e_j.$$

Independence Preserves Indifference

The relation $e_j \sim \ell(\alpha_j)$ means that both

$$e_j \succeq \ell(\alpha_j) \qquad \text{and} \qquad \ell(\alpha_j) \succeq e_j.$$

For any lottery r and any $\lambda \in (0,1)$, independence applied to these two comparisons gives

$$\begin{aligned} \lambda e_j + (1-\lambda)r &\succeq \lambda\ell(\alpha_j) + (1-\lambda)r, \\ \lambda\ell(\alpha_j) + (1-\lambda)r &\succeq \lambda e_j + (1-\lambda)r. \end{aligned}$$

Independence Preserves Indifference

The relation $e_j \sim \ell(\alpha_j)$ means that both

$$e_j \succeq \ell(\alpha_j) \qquad \text{and} \qquad \ell(\alpha_j) \succeq e_j.$$

For any lottery r and any $\lambda \in (0,1)$, independence applied to these two comparisons gives

$$\begin{aligned} \lambda e_j + (1-\lambda) r &\succeq \lambda \ell(\alpha_j) + (1-\lambda) r, \\ \lambda \ell(\alpha_j) + (1-\lambda) r &\succeq \lambda e_j + (1-\lambda) r. \end{aligned}$$

The two mixed lotteries are therefore indifferent.

Substitute Equivalent Components One at a Time

Independence therefore lets us replace one component of a mixture by an indifferent lottery without changing the mixture's value.

Substitute Equivalent Components One at a Time

Independence therefore lets us replace one component of a mixture by an indifferent lottery without changing the mixture's value. Replace the pure components of p one at a time. A finite induction gives

$$\begin{aligned} p = \sum_j p_j e_j &\sim \sum_j p_j \ell(\alpha_j) \\ &= \ell\left(\sum_j p_j \alpha_j\right). \end{aligned}$$

Substitute Equivalent Components One at a Time

Independence therefore lets us replace one component of a mixture by an indifferent lottery without changing the mixture's value. Replace the pure components of p one at a time. A finite induction gives

$$\begin{aligned} p = \sum_j p_j e_j \sim \sum_j p_j \ell(\alpha_j) \\ = \ell\left(\sum_j p_j \alpha_j\right). \end{aligned}$$

Every lottery is therefore indifferent to one best–worst lottery.

The Calibration Numbers Are Utilities

Define

$$u(j) := \alpha_j.$$

The Calibration Numbers Are Utilities

Define

$$u(j) := \alpha_j.$$

The substitution step gives

$$p \sim \ell(u^\top p), \qquad q \sim \ell(u^\top q).$$

The Calibration Numbers Are Utilities

Define

$$u(j) := \alpha_j.$$

The substitution step gives

$$p \sim \ell(u^\top p), \qquad q \sim \ell(u^\top q).$$

Because the parameter of $\ell(\alpha)$ strictly orders these lotteries,

$$p \succeq q \iff u^\top p \geq u^\top q.$$

The Calibration Numbers Are Utilities

Define

$$u(j) := \alpha_j.$$

The substitution step gives

$$p \sim \ell(u^\top p), \qquad q \sim \ell(u^\top q).$$

Because the parameter of $\ell(\alpha)$ strictly orders these lotteries,

$$p \succeq q \quad \Longleftrightarrow \quad u^\top p \geq u^\top q.$$

This proves the expected-utility representation.

The Scale Is Cardinal but Not Interpersonal

Our construction normalized

$$u(j^+) = 1, \qquad u(j^-) = 0.$$

Under this normalization, each indifference $e_j \sim \ell(\alpha_j)$ forces $u(j) = \alpha_j$.

The Scale Is Cardinal but Not Interpersonal

Our construction normalized

$$u(j^+) = 1, \qquad u(j^-) = 0.$$

Under this normalization, each indifference $e_j \sim \ell(\alpha_j)$ forces $u(j) = \alpha_j$. Hence every normalized representation is identical. Undoing the normalization gives exactly

$$\widetilde{u} = au + b\mathbf{1}, \qquad a > 0.$$

The Scale Is Cardinal but Not Interpersonal

Our construction normalized

$$u(j^+) = 1, \qquad u(j^-) = 0.$$

Under this normalization, each indifference $e_j \sim \ell(\alpha_j)$ forces $u(j) = \alpha_j$. Hence every normalized representation is identical. Undoing the normalization gives exactly

$$\widetilde{u} = au + b\mathbf{1}, \qquad a > 0.$$

Ratios of utility differences now have meaning within one person's scale. Absolute levels and units do not, and the theorem does not equate one person's utility unit with another person's.

Now Consider the Aggregation Question

Fix n people with VNM preferences represented by

$$u_1, \ldots, u_n \in \mathbb{R}^C.$$

Now Consider the Aggregation Question

Fix n people with VNM preferences represented by

$$u_1, \ldots, u_n \in \mathbb{R}^C.$$

At this fixed preference profile, an aggregation procedure must output some preference over lotteries.

Now Consider the Aggregation Question

Fix n people with VNM preferences represented by

$$u_1, \ldots, u_n \in \mathbb{R}^C.$$

At this fixed preference profile, an aggregation procedure must output some preference over lotteries. Call a possible output a *candidate aggregate preference*. Require it to satisfy the VNM axioms, and let u_0 be one of its representations.

Weak Pareto Compares Unanimous Lottery Improvements

Definition

The candidate aggregate satisfies the *weak Pareto axiom* if, for every $p, q \in \Delta(C)$,

$$u_i^\top p \geq u_i^\top q \quad \text{for every } i \quad \Longrightarrow \quad u_0^\top p \geq u_0^\top q.$$

Weak Pareto Compares Unanimous Lottery Improvements

Definition

The candidate aggregate satisfies the *weak Pareto axiom* if, for every $p, q \in \Delta(C)$,

$$u_i^\top p \geq u_i^\top q \quad \text{for every } i \quad \Longrightarrow \quad u_0^\top p \geq u_0^\top q.$$

Whenever every person weakly prefers p to q, the aggregate may not rank q strictly above p.

Weak Pareto Compares Unanimous Lottery Improvements

Definition

The candidate aggregate satisfies the *weak Pareto axiom* if, for every $p, q \in \Delta(C)$,

$$u_i^\top p \geq u_i^\top q \quad \text{for every } i \quad \Longrightarrow \quad u_0^\top p \geq u_0^\top q.$$

Whenever every person weakly prefers p to q, the aggregate may not rank q strictly above p. The axiom does not specify what happens when people disagree.

Nonnegative Weighted Sums Work

Suppose

$$u_0 = c\mathbf{1} + \sum_{i=1}^{n} w_i u_i, \qquad w_i \geq 0.$$

Nonnegative Weighted Sums Work

Suppose

$$u_0 = c\mathbf{1} + \sum_{i=1}^{n} w_i u_i, \qquad w_i \geq 0.$$

If every person weakly prefers p to q, then

$$\begin{aligned} u_0^\top (p - q) &= c\mathbf{1}^\top (p - q) + \sum_i w_i u_i^\top (p - q) \\ &= \sum_i w_i u_i^\top (p - q) \geq 0, \end{aligned}$$

because $\mathbf{1}^\top p = \mathbf{1}^\top q = 1$.

Nonnegative Weighted Sums Work

Suppose

$$u_0 = c\mathbf{1} + \sum_{i=1}^{n} w_i u_i, \qquad w_i \geq 0.$$

If every person weakly prefers p to q, then

$$\begin{aligned} u_0^\top (p-q) &= c\mathbf{1}^\top (p-q) + \sum_i w_i u_i^\top (p-q) \\ &= \sum_i w_i u_i^\top (p-q) \geq 0, \end{aligned}$$

because $\mathbf{1}^\top p = \mathbf{1}^\top q = 1$. Thus such aggregates exist—for example, choose every $w_i = 1$.

Harsanyi's Aggregation Theorem

Theorem (Harsanyi)

Fix individual VNM preferences over $\Delta(C)$, represented by $u_1, \ldots, u_n$. A candidate aggregate VNM preference represented by u_0 satisfies the weak Pareto axiom if and only if

$$u_0 = c\mathbf{1} + \sum_{i=1}^{n} w_i u_i$$

for some constant $c \in \mathbb{R}$ and weights $w_i \geq 0$.

Harsanyi's Aggregation Theorem

Theorem (Harsanyi)

Fix individual VNM preferences over $\Delta(C)$*, represented by* $u_1, \ldots, u_n$*. A candidate aggregate VNM preference represented by* u_0 *satisfies the weak Pareto axiom if and only if*

$$u_0 = c\mathbf{1} + \sum_{i=1}^{n} w_i u_i$$

for some constant $c \in \mathbb{R}$ *and weights* $w_i \geq 0$*.*

The easy direction showed that every nonnegative weighted sum works. The converse says that weak Pareto permits no other VNM aggregate.

Proof Roadmap for the Converse

1. Replace a pair of lotteries p, q by their difference $z = p - q$.

Proof Roadmap for the Converse

1. Replace a pair of lotteries p, q by their difference $z = p - q$.
2. Define the scalar-valued linear functions $\ell_i(z) = u_i^\top z$ and express weak Pareto as an implication among their signs.

Proof Roadmap for the Converse

1. Replace a pair of lotteries p, q by their difference $z = p - q$.
2. Define the scalar-valued linear functions $\ell_i(z) = u_i^\top z$ and express weak Pareto as an implication among their signs.
3. Form an LP that searches for a lottery-difference direction that benefits every person but harms the aggregate.

Proof Roadmap for the Converse

1. Replace a pair of lotteries p, q by their difference $z = p - q$.
2. Define the scalar-valued linear functions $\ell_i(z) = u_i^\top z$ and express weak Pareto as an implication among their signs.
3. Form an LP that searches for a lottery-difference direction that benefits every person but harms the aggregate.
4. Read the dual solution as nonnegative welfare weights plus one additive constant.

Lottery Differences Form a Vector Space

If $p, q \in \Delta(C)$, then $z = p - q$ satisfies

$$\mathbf{1}^\top z = 0.$$

Lottery Differences Form a Vector Space

If $p, q \in \Delta(C)$, then $z = p - q$ satisfies

$$\mathbf{1}^\top z = 0.$$

Define the zero-sum space

$$V := \{z \in \mathbb{R}^C : \mathbf{1}^\top z = 0\}.$$

Lottery Differences Form a Vector Space

If $p, q \in \Delta(C)$, then $z = p - q$ satisfies

$$\mathbf{1}^\top z = 0.$$

Define the zero-sum space

$$V := \{z \in \mathbb{R}^C : \mathbf{1}^\top z = 0\}.$$

Conversely, every $z \in V$ is proportional to a lottery difference. Let $\bar{p} = \mathbf{1}/|C|$. For sufficiently small $t > 0$,

$$p := \bar{p} + tz \in \Delta(C), \qquad q := \bar{p} \in \Delta(C),$$

and therefore $p - q = tz$.

Weak Pareto Becomes a Linear Implication

A linear map from V to $\mathbb{R}$ is called a *linear functional*. For $i = 0, 1, \ldots, n$, define

$$\ell_i : V \to \mathbb{R}, \qquad \ell_i(z) := u_i^\top z.$$

Weak Pareto Becomes a Linear Implication

A linear map from V to $\mathbb{R}$ is called a *linear functional*. For $i = 0, 1, \ldots, n$, define

$$\ell_i : V \to \mathbb{R}, \qquad \ell_i(z) := u_i^\top z.$$

Fix $z \in V$ with $\ell_i(z) \geq 0$ for every person $i = 1, \ldots, n$, and use the lotteries satisfying $p - q = tz$ from the previous slide.

Weak Pareto Becomes a Linear Implication

A linear map from V to $\mathbb{R}$ is called a *linear functional*. For $i = 0, 1, \ldots, n$, define

$$\ell_i : V \to \mathbb{R}, \qquad \ell_i(z) := u_i^\top z.$$

Fix $z \in V$ with $\ell_i(z) \geq 0$ for every person $i = 1, \ldots, n$, and use the lotteries satisfying $p - q = tz$ from the previous slide. Every person weakly prefers p to q, so weak Pareto gives

$$0 \leq u_0^\top (p - q) = t\ell_0(z).$$

Weak Pareto Becomes a Linear Implication

A linear map from V to $\mathbb{R}$ is called a *linear functional*. For $i = 0, 1, \ldots, n$, define

$$\ell_i : V \to \mathbb{R}, \qquad \ell_i(z) := u_i^\top z.$$

Fix $z \in V$ with $\ell_i(z) \geq 0$ for every person $i = 1, \ldots, n$, and use the lotteries satisfying $p - q = tz$ from the previous slide. Every person weakly prefers p to q, so weak Pareto gives

$$0 \leq u_0^\top (p - q) = t\ell_0(z).$$

Because $t > 0$,

$$\ell_i(z) \geq 0 \ \forall i \quad \Longrightarrow \quad \ell_0(z) \geq 0.$$

What Remains to Prove

We have converted the weak Pareto axiom into the implication

$$\ell_i(z) \geq 0 \ \forall i \quad \Longrightarrow \quad \ell_0(z) \geq 0$$

for every $z \in V$.

What Remains to Prove

We have converted the weak Pareto axiom into the implication

$$\ell_i(z) \geq 0 \ \forall i \quad \Longrightarrow \quad \ell_0(z) \geq 0$$

for every $z \in V$. To prove the converse in Harsanyi's theorem, we must show that this implication forces

$$u_0 = c\mathbf{1} + \sum_{i=1}^{n} w_i u_i, \qquad w_i \geq 0.$$

What Remains to Prove

We have converted the weak Pareto axiom into the implication

$$\ell_i(z) \geq 0 \ \forall i \quad \Longrightarrow \quad \ell_0(z) \geq 0$$

for every $z \in V$. To prove the converse in Harsanyi's theorem, we must show that this implication forces

$$u_0 = c\mathbf{1} + \sum_{i=1}^{n} w_i u_i, \qquad w_i \geq 0.$$

An LP will search for a violation of the implication; its dual will instead produce the coefficients $w_1, \ldots, w_n$ and c.

Recall a Standard Form of LP Duality

For a primal variable z whose coordinates may have either sign,

$$
\begin{array}{llll}
\text{(P)} & \max\limits_{z} & g^\top z & \text{subject to } Gz \le h, \quad Ez = f, \\
\text{(D)} & \min\limits_{w,\lambda} & h^\top w + f^\top \lambda & \text{subject to } G^\top w + E^\top \lambda = g.
\end{array}
$$

Recall a Standard Form of LP Duality

For a primal variable z whose coordinates may have either sign,

$$
\begin{array}{llll}
\text{(P)} & \max\limits_{z} & g^\top z & \text{subject to } Gz \le h, \quad Ez = f, \\
\text{(D)} & \min\limits_{w,\lambda} & h^\top w + f^\top \lambda & \text{subject to } G^\top w + E^\top \lambda = g.
\end{array}
$$

Inequality multipliers satisfy $w \ge 0$; equality multipliers λ may have either sign.

Recall a Standard Form of LP Duality

For a primal variable z whose coordinates may have either sign,

$$\begin{array}{llll} \text{(P)} & \max\limits_{z} & g^\top z & \text{subject to } Gz \le h, \quad Ez = f, \\ \text{(D)} & \min\limits_{w,\lambda} & h^\top w + f^\top \lambda & \text{subject to } G^\top w + E^\top \lambda = g. \end{array}$$

Inequality multipliers satisfy $w \ge 0$; equality multipliers λ may have either sign. If the primal is feasible and bounded above, the dual has a solution and the two optimal values are equal.

The Primal Searches for a Pareto Violation

Collect the individual utility vectors as

$$U = [u_1 \ \cdots \ u_n], \qquad U^\top z = (\ell_1(z), \ldots, \ell_n(z))^\top.$$

The Primal Searches for a Pareto Violation

Collect the individual utility vectors as

$$U = [u_1 \;\; \cdots \;\; u_n], \qquad U^\top z = (\ell_1(z), \ldots, \ell_n(z))^\top.$$

Consider

$$\begin{aligned} \max_z \quad & -u_0^\top z \\ \text{subject to} \quad & U^\top z \geq 0, \\ & \mathbf{1}^\top z = 0. \end{aligned}$$

The equality keeps z in V; the inequality says every person weakly gains. The objective is positive exactly when the aggregate loses.

The Primal Searches for a Pareto Violation

Collect the individual utility vectors as

$$U = [u_1 \ \cdots \ u_n], \qquad U^\top z = (\ell_1(z), \ldots, \ell_n(z))^\top.$$

Consider

$$\begin{aligned} \max_{z} \quad & -u_0^\top z \\ \text{subject to} \quad & U^\top z \geq 0, \\ & \mathbf{1}^\top z = 0. \end{aligned}$$

The equality keeps z in V; the inequality says every person weakly gains. The objective is positive exactly when the aggregate loses. Weak Pareto bounds the objective above by zero. Since $z = 0$ is feasible, the primal optimal value is zero.

The Equality Multiplier Is the Constant

Write $U^\top z \geq 0$ as $-U^\top z \leq 0$. Give it a multiplier $w \geq 0$, and give $\mathbf{1}^\top z = 0$ an unrestricted multiplier γ. The dual is

$$\begin{aligned} \min_{w \geq 0,\, \gamma \in \mathbb{R}} \quad & 0 \\ \text{subject to} \quad & -Uw + \gamma \mathbf{1} = -u_0. \end{aligned}$$

The Equality Multiplier Is the Constant

Write $U^\top z \geq 0$ as $-U^\top z \leq 0$. Give it a multiplier $w \geq 0$, and give $\mathbf{1}^\top z = 0$ an unrestricted multiplier γ. The dual is

$$\begin{aligned} \min_{w \geq 0,\, \gamma \in \mathbb{R}} \quad & 0 \\ \text{subject to} \quad & -Uw + \gamma \mathbf{1} = -u_0. \end{aligned}$$

The primal is feasible and bounded above, so LP duality says that the dual has a feasible solution.

The Equality Multiplier Is the Constant

Write $U^\top z \geq 0$ as $-U^\top z \leq 0$. Give it a multiplier $w \geq 0$, and give $\mathbf{1}^\top z = 0$ an unrestricted multiplier γ. The dual is

$$\begin{aligned} \min_{w \geq 0,\, \gamma \in \mathbb{R}} \quad & 0 \\ \text{subject to} \quad & -Uw + \gamma \mathbf{1} = -u_0. \end{aligned}$$

The primal is feasible and bounded above, so LP duality says that the dual has a feasible solution. Rearranging its equality and writing $c = -\gamma$ gives

$$u_0 = Uw + c\mathbf{1} = c\mathbf{1} + \sum_{i=1}^{n} w_i u_i, \qquad w_i \geq 0.$$

The Converse Completes the Theorem

Before stating the theorem, we proved the easy direction:

$$u_0 = c\mathbf{1} + \sum_i w_i u_i \text{ with } w_i \geq 0 \quad \Longrightarrow \quad \text{weak Pareto.}$$

The Converse Completes the Theorem

Before stating the theorem, we proved the easy direction:

$$u_0 = c\mathbf{1} + \sum_i w_i u_i \text{ with } w_i \geq 0 \quad \Longrightarrow \quad \text{weak Pareto.}$$

The LP argument has now proved the converse:

$$\text{weak Pareto} \quad \Longrightarrow \quad u_0 = c\mathbf{1} + \sum_i w_i u_i \text{ with } w_i \geq 0.$$

The Converse Completes the Theorem

Before stating the theorem, we proved the easy direction:

$$u_0 = c\mathbf{1} + \sum_i w_i u_i \text{ with } w_i \geq 0 \quad \Longrightarrow \quad \text{weak Pareto.}$$

The LP argument has now proved the converse:

$$\text{weak Pareto} \quad \Longrightarrow \quad u_0 = c\mathbf{1} + \sum_i w_i u_i \text{ with } w_i \geq 0.$$

Together, the two implications prove Harsanyi's characterization. $\square$

Harsanyi Does Not Choose the Weights

Harsanyi characterizes the form of every weak-Pareto VNM aggregate:

$$u_0 = c\mathbf{1} + \sum_{i=1}^{n} w_i u_i, \qquad w_i \geq 0.$$

Harsanyi Does Not Choose the Weights

Harsanyi characterizes the form of every weak-Pareto VNM aggregate:

$$u_0 = c\mathbf{1} + \sum_{i=1}^{n} w_i u_i, \qquad w_i \geq 0.$$

The theorem does not choose the weights; weak Pareto even permits $w_i = 0$.

Harsanyi Does Not Choose the Weights

Harsanyi characterizes the form of every weak-Pareto VNM aggregate:

$$u_0 = c\mathbf{1} + \sum_{i=1}^{n} w_i u_i, \qquad w_i \geq 0.$$

The theorem does not choose the weights; weak Pareto even permits $w_i = 0$. Suppose u_i measures a person's utility in "dollars." The equivalent representation $100u_i$ measures it in "cents" and induces exactly the same preferences over lotteries. Yet

$$w_i u_i = \frac{w_i}{100}(100u_i).$$

Harsanyi Does Not Choose the Weights

Harsanyi characterizes the form of every weak-Pareto VNM aggregate:

$$u_0 = c\mathbf{1} + \sum_{i=1}^{n} w_i u_i, \qquad w_i \geq 0.$$

The theorem does not choose the weights; weak Pareto even permits $w_i = 0$. Suppose u_i measures a person's utility in "dollars." The equivalent representation $100u_i$ measures it in "cents" and induces exactly the same preferences over lotteries. Yet

$$w_i u_i = \frac{w_i}{100}(100u_i).$$

The same observed preferences support different numerical weights. Interpersonal utility scales are not identified from data.

What to Carry Forward

- The VNM axioms characterize when preferences over lotteries can be represented by expected utility.

What to Carry Forward

- The VNM axioms characterize when preferences over lotteries can be represented by expected utility.
- Once interpersonal utility scales are chosen, Harsanyi characterizes the weak-Pareto aggregates as nonnegative weighted sums.

What to Carry Forward

- The VNM axioms characterize when preferences over lotteries can be represented by expected utility.
- Once interpersonal utility scales are chosen, Harsanyi characterizes the weak-Pareto aggregates as nonnegative weighted sums.
- Choosing the scales and welfare weights is an additional normative decision; it is not determined by observed preferences alone.

Thanks!

See you next class!