Agreement on What to Do

Calibrated Recommendations for a Shared Objective

Aaron Roth

University of Pennsylvania

Fall 2026

From One Probability to Many Possible States

Lecture 11 studied conversations about a probability of a binary event. Calibrated conversation reached approximate agreement and improved aggregate accuracy, up to calibration error.

From One Probability to Many Possible States

Lecture 11 studied conversations about a probability of a binary event. Calibrated conversation reached approximate agreement and improved aggregate accuracy, up to calibration error.
A shared full forecast would also settle which downstream actions are predicted to be optimal for a shared objective. But real utility functions can depend on a high cardinality state space.

From One Probability to Many Possible States

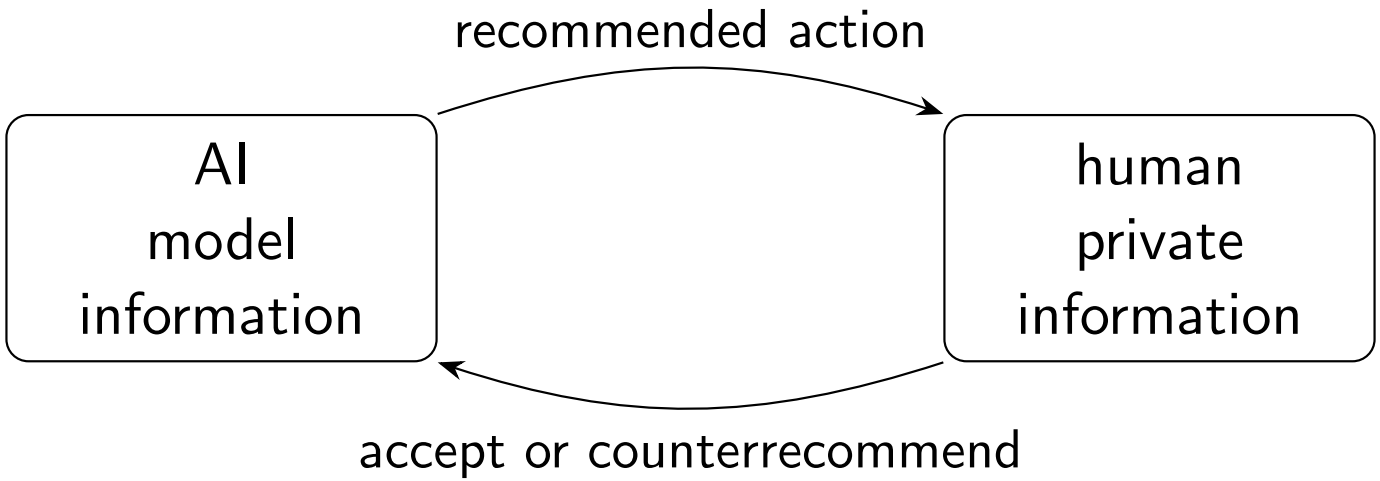
Lecture 11 studied conversations about a probability of a binary event. Calibrated conversation reached approximate agreement and improved aggregate accuracy, up to calibration error.
A shared full forecast would also settle which downstream actions are predicted to be optimal for a shared objective. But real utility functions can depend on a high cardinality state space.
The direct extension to high dimensions has two costs:

- long probability vectors for the human and AI to exchange;
- conversation-length bounds that grow with the number of coordinates.

That is a cumbersome and expensive protocol.

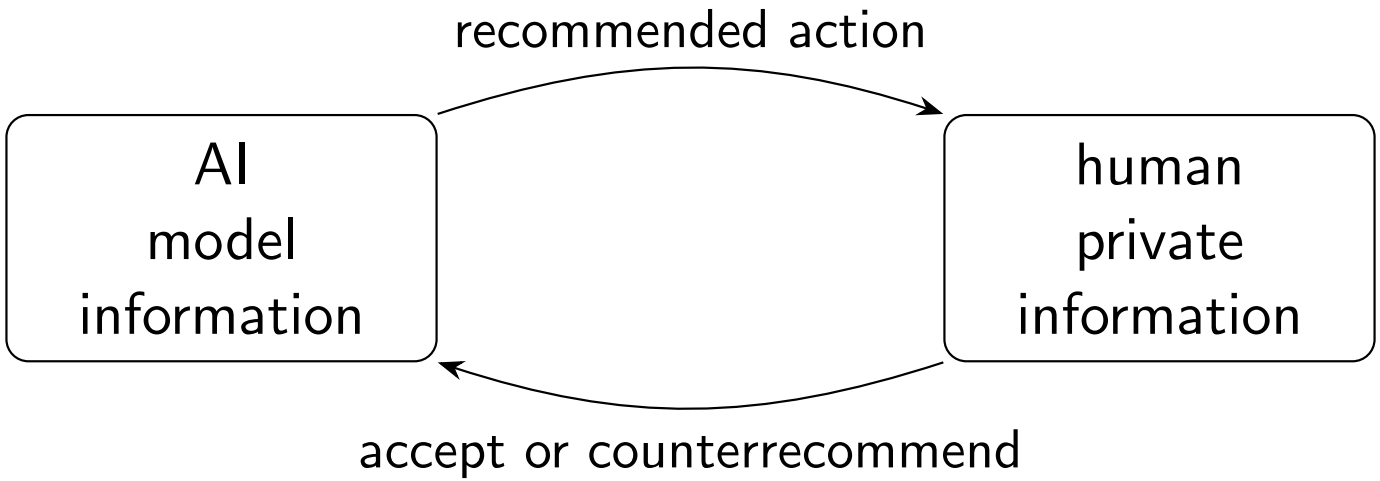
A Simpler Target: Agree on What to Do

An AI and a human must decide whether to deploy a system, hold it back, or audit it further. They want the same thing, but have different information.



A Simpler Target: Agree on What to Do

An AI and a human must decide whether to deploy a system, hold it back, or audit it further. They want the same thing, but have different information.



Instead of communicating their entire forecasts, they exchange recommended actions. Belief differences that do not affect the decision can remain unresolved.

Today's question

Can these shorter messages lead quickly to a shared decision and improve utility, without requiring agreement on every probability?

Overview

- A shared decision can need less communication than a shared forecast: exchange recommended actions.

Overview

- A shared decision can need less communication than a shared forecast: exchange recommended actions.
- Conversation decision calibration makes predicted utility gains reliable in aggregate. Continued disagreement then forces progress, so most conversations end quickly.

Overview

- A shared decision can need less communication than a shared forecast: exchange recommended actions.
- Conversation decision calibration makes predicted utility gains reliable in aggregate. Continued disagreement then forces progress, so most conversations end quickly.
- Conversation can improve on both parties' initial recommendations, with gains measured in aggregate.

Overview

- A shared decision can need less communication than a shared forecast: exchange recommended actions.
- Conversation decision calibration makes predicted utility gains reliable in aggregate. Continued disagreement then forces progress, so most conversations end quickly.
- Conversation can improve on both parties' initial recommendations, with gains measured in aggregate.
- We can learn this reliability efficiently: reuse approachability and the zero-payoff self-play argument.

The Parties Share a Utility Function

As in Lecture 7, let $\mathcal{Y} = \{1, \ldots, d\}$ be a finite outcome set, $d \geq 2$, and $\mathcal{A}$ a finite action set. Both parties use the same known utility

$$u(a, y) \in \mathbb{R}.$$

Higher is better.

The Parties Share a Utility Function

As in Lecture 7, let $\mathcal{Y} = \{1, \ldots, d\}$ be a finite outcome set, $d \geq 2$, and $\mathcal{A}$ a finite action set. Both parties use the same known utility

$$u(a, y) \in \mathbb{R}.$$

Higher is better.

A forecast $p \in \Delta(\mathcal{Y})$ is a probability distribution over outcomes.

Its implied expected utility for action a is

$$U(a, p) := \mathbb{E}_{Y \sim p}[u(a, Y)] = \sum_{y \in \mathcal{Y}} p_y \, u(a, y).$$

The Parties Share a Utility Function

As in Lecture 7, let $\mathcal{Y} = \{1, \ldots, d\}$ be a finite outcome set, $d \geq 2$, and $\mathcal{A}$ a finite action set. Both parties use the same known utility

$$u(a, y) \in \mathbb{R}.$$

Higher is better.
A forecast $p \in \Delta(\mathcal{Y})$ is a probability distribution over outcomes.
Its implied expected utility for action a is

$$U(a, p) := \mathbb{E}_{Y \sim p}[u(a, Y)] = \sum_{y \in \mathcal{Y}} p_y \, u(a, y).$$

Fix a tie-breaking rule and write

$$\mathrm{BR}(p) \in \arg\max_{a \in \mathcal{A}} U(a, p)$$

for the action that maximizes expected utility under p. The objective is shared; the forecasts need not be.

Very Different Forecasts Can Recommend the Same Action

Here is Lecture 7's user. Let r be the forecasted probability that deployment is safe.

action	safe	harmful	expected utility
deploy	1	-4	$5r - 4$
hold	0	0	0
audit	$4/5$	$-1/5$	$r - 1/5$

Very Different Forecasts Can Recommend the Same Action

Here is Lecture 7's user. Let r be the forecasted probability that deployment is safe.

action	safe	harmful	expected utility
deploy	1	-4	$5r - 4$
hold	0	0	0
audit	$4/5$	$-1/5$	$r - 1/5$

Audit beats hold when $r > 1/5$, and beats deploy when $r < 19/20$.

What Counts as Agreement on an Action?

Fix a tolerance $\varepsilon > 0$, measured in utility units. An action c is *ε-optimal under p* if

$$\max_{a \in \mathcal{A}} U(a, p) - U(c, p) \leq \varepsilon.$$

The forecaster believes no alternative improves on c by more than ε.

What Counts as Agreement on an Action?

Fix a tolerance $\varepsilon > 0$, measured in utility units. An action c is *ε-optimal under p* if

$$\max_{a \in \mathcal{A}} U(a, p) - U(c, p) \leq \varepsilon.$$

The forecaster believes no alternative improves on c by more than ε.

Definition

An action c is *mutually ε-acceptable* under forecasts p, q if

$$\max_{a} U(a, p) - U(c, p) \leq \varepsilon, \qquad \max_{a} U(a, q) - U(c, q) \leq \varepsilon.$$

What Counts as Agreement on an Action?

Fix a tolerance $\varepsilon > 0$, measured in utility units. An action c is *ε-optimal under p* if

$$\max_{a \in \mathcal{A}} U(a, p) - U(c, p) \leq \varepsilon.$$

The forecaster believes no alternative improves on c by more than ε.

Definition

An action c is *mutually ε-acceptable* under forecasts p, q if

$$\max_{a} U(a, p) - U(c, p) \leq \varepsilon, \qquad \max_{a} U(a, q) - U(c, q) \leq \varepsilon.$$

There is one common action to take. The parties need not have the same forecast, or even the same favorite action.

Two Openings, Then Accept or Counterrecommend

The AI and human first announce recommendations c^1 and c^2, each maximizing expected utility under their own internal forecast. We impose no calibration requirement on these initial forecasts. From message 3 onward, the AI and human alternate replies.

Two Openings, Then Accept or Counterrecommend

The AI and human first announce recommendations c^1 and c^2, each maximizing expected utility under their own internal forecast. We impose no calibration requirement on these initial forecasts. From message 3 onward, the AI and human alternate replies.

A reply to the current recommendation b

1. Form an internal forecast p and compute $a = \mathrm{BR}(p)$.
2. Compute the predicted utility gain

$$g = U(a, p) - U(b, p) \geq 0.$$

3. If $g \leq \varepsilon$, say *accept*: keep b and stop. Otherwise send the counterrecommendation a and continue.

Two Openings, Then Accept or Counterrecommend

The AI and human first announce recommendations c^1 and c^2, each maximizing expected utility under their own internal forecast. We impose no calibration requirement on these initial forecasts. From message 3 onward, the AI and human alternate replies.

A reply to the current recommendation b

1. Form an internal forecast p and compute $a = \mathrm{BR}(p)$.
2. Compute the predicted utility gain

$$g = U(a, p) - U(b, p) \geq 0.$$

3. If $g \leq \varepsilon$, say *accept*: keep b and stop. Otherwise send the counterrecommendation a and continue.

The first reply answers the human's opening c^2. Only actions or acceptance messages are sent; forecasts remain internal.

Acceptance Certifies a Common Decision

Suppose the preceding speaker recommended $b = \mathrm{BR}(q)$, and the current speaker accepts using forecast p.

Acceptance Certifies a Common Decision

Suppose the preceding speaker recommended $b = \mathrm{BR}(q)$, and the current speaker accepts using forecast p.

The previous speaker regards b as optimal:

$$\max_a U(a, q) - U(b, q) = 0.$$

The current speaker's acceptance says

$$\max_a U(a, p) - U(b, p) = U(\mathrm{BR}(p), p) - U(b, p) \leq \varepsilon.$$

Acceptance Certifies a Common Decision

Suppose the preceding speaker recommended $b = \mathrm{BR}(q)$, and the current speaker accepts using forecast p.
The previous speaker regards b as optimal:

$$\max_a U(a, q) - U(b, q) = 0.$$

The current speaker's acceptance says

$$\max_a U(a, p) - U(b, p) = U(\mathrm{BR}(p), p) - U(b, p) \leq \varepsilon.$$

Thus b is mutually ε-acceptable under the two forecasts used in this final exchange. This establishes agreement relative to the forecasts used by the protocol. Calibration will connect those forecasts to realized utility.

What remains to prove

Do most conversations reach acceptance quickly? How does the final decisions' aggregate utility compare with that of the opening recommendations?

The Same Decision Problem Repeats Across Days

As in Lecture 11, there are T days and at most $R \geq 3$ messages per day. The outcome y_t is fixed before the conversation and all its random draws, and revealed afterward.

The Same Decision Problem Repeats Across Days

As in Lecture 11, there are T days and at most $R \geq 3$ messages per day. The outcome y_t is fixed before the conversation and all its random draws, and revealed afterward.

Let S_k be the days that reach message k, with $S_1 = S_2 = S_3 = \{1, \ldots, T\}$. On an active day $t \in S_k$, for $k \geq 3$, let b_t^k be the previous recommendation and p_t^k the replying speaker's forecast. Write

$$a_t^k := \mathrm{BR}(p_t^k), \qquad g_t^k := U(a_t^k, p_t^k) - U(b_t^k, p_t^k).$$

The Same Decision Problem Repeats Across Days

As in Lecture 11, there are T days and at most $R \geq 3$ messages per day. The outcome y_t is fixed before the conversation and all its random draws, and revealed afterward.
Let S_k be the days that reach message k, with $S_1 = S_2 = S_3 = \{1, \ldots, T\}$. On an active day $t \in S_k$, for $k \geq 3$, let b_t^k be the previous recommendation and p_t^k the replying speaker's forecast. Write

$$a_t^k := \mathrm{BR}(p_t^k), \qquad g_t^k := U(a_t^k, p_t^k) - U(b_t^k, p_t^k).$$

The days that remain unresolved are

$$S_{k+1} = \{t \in S_k : g_t^k > \varepsilon\}.$$

At the cap, output the last recommendation. S_{R+1} records unresolved days, not an additional message. Capped conversations need not be mutually acceptable.

A Predicted Gain Need Not Be a Realized Gain

Suppose the previous recommendation is hold. The current speaker forecasts safe with probability $1/2$:

action	safe	harmful	expected utility
deploy	1	-4	$-3/2$
hold	0	0	0
audit	$4/5$	$-1/5$	$3/10$

A Predicted Gain Need Not Be a Realized Gain

Suppose the previous recommendation is hold. The current speaker forecasts safe with probability $1/2$:

action	safe	harmful	expected utility
deploy	1	-4	$-3/2$
hold	0	0	0
audit	$4/5$	$-1/5$	$3/10$

With $\varepsilon = 1/4$, the speaker counterrecommends audit because

$$g = \frac{3}{10} - 0 = \frac{3}{10} > \varepsilon.$$

A Predicted Gain Need Not Be a Realized Gain

Suppose the previous recommendation is hold. The current speaker forecasts safe with probability $1/2$:

action	safe	harmful	expected utility
deploy	1	-4	$-3/2$
hold	0	0	0
audit	$4/5$	$-1/5$	$3/10$

With $\varepsilon = 1/4$, the speaker counterrecommends audit because

$$g = \frac{3}{10} - 0 = \frac{3}{10} > \varepsilon.$$

But on a harmful outcome the realized gain is $-1/5 - 0 = -1/5$. This happens with probability $1/2$ even if the forecast is "correct"! We therefore cannot prove utility increases on every day. As in Lecture 7, we need *decision calibration* to relate predicted and realized gains in aggregate. We will adapt it to conversation.

Conversation Decision Calibration

At reply position $k \geq 3$, group the days by the proposed change $b \to a$:

$$C_{k,b,a} := \{t \in S_k : b_t^k = b,\ a_t^k = a,\ g_t^k > \varepsilon\}.$$

These cells partition S_{k+1}, the days with a counterrecommendation.

Conversation Decision Calibration

At reply position $k \geq 3$, group the days by the proposed change $b \to a$:

$$C_{k,b,a} := \{t \in S_k : b_t^k = b,\ a_t^k = a,\ g_t^k > \varepsilon\}.$$

These cells partition S_{k+1}, the days with a counterrecommendation.
For $t \in C_{k,b,a}$, recall $g_t^k = U(a, p_t^k) - U(b, p_t^k)$. Sum realized gain minus predicted gain:

$$r_{k,b,a} := \sum_{t \in C_{k,b,a}} \left[u(a, y_t) - u(b, y_t) - g_t^k\right].$$

Our *conversation decision calibration* condition requires $r_{k,b,a} = 0$ for every pair at every reply position.

Conversation Decision Calibration

At reply position $k \geq 3$, group the days by the proposed change $b \to a$:

$$C_{k,b,a} := \{t \in S_k : b_t^k = b,\ a_t^k = a,\ g_t^k > \varepsilon\}.$$

These cells partition S_{k+1}, the days with a counterrecommendation.
For $t \in C_{k,b,a}$, recall $g_t^k = U(a, p_t^k) - U(b, p_t^k)$. Sum realized gain minus predicted gain:

$$r_{k,b,a} := \sum_{t \in C_{k,b,a}} \left[u(a, y_t) - u(b, y_t) - g_t^k\right].$$

Our *conversation decision calibration* condition requires $r_{k,b,a} = 0$ for every pair at every reply position.
First we will analyze exact calibration; later we will allow small errors.

Calibrate Utility Comparisons, Not the Whole Forecast

Write $u_a = (u(a,1), \ldots, u(a,d))$, and let e_y be the point mass on y. For a proposed change $b \to a$, realized gain minus predicted gain is

$$u(a,y) - u(b,y) - \big(U(a,p) - U(b,p)\big) = \langle u_a - u_b, e_y - p \rangle.$$

Calibrate Utility Comparisons, Not the Whole Forecast

Write $u_a = (u(a,1), \ldots, u(a,d))$, and let e_y be the point mass on y. For a proposed change $b \to a$, realized gain minus predicted gain is

$$u(a,y) - u(b,y) - \big(U(a,p) - U(b,p)\big) = \langle u_a - u_b, e_y - p \rangle.$$

Thus, on the days $C_{k,b,a}$ that actually replace b by a,

$$r_{k,b,a} = \left\langle u_a - u_b, \sum_{t \in C_{k,b,a}} (e_{y_t} - p_t^k) \right\rangle.$$

Action Pairs Replace Probability Buckets

Lecture 11 grouped days by pairs of probability reports to control

$$2(y - p)(p - q).$$

The multiplier $p - q$ changes with the reports, so those reports had to be fixed within a cell.

Action Pairs Replace Probability Buckets

Lecture 11 grouped days by pairs of probability reports to control

$$2(y - p)(p - q).$$

The multiplier $p - q$ changes with the reports, so those reports had to be fixed within a cell.
Here the multiplier is $u_a - u_b$. Once the two actions are fixed, this utility difference is fixed even if the forecasts vary.

Action Pairs Replace Probability Buckets

Lecture 11 grouped days by pairs of probability reports to control

$$2(y - p)(p - q).$$

The multiplier $p - q$ changes with the reports, so those reports had to be fixed within a cell.
Here the multiplier is $u_a - u_b$. Once the two actions are fixed, this utility difference is fixed even if the forecasts vary.
There are only $|\mathcal{A}|^2$ action pairs. A grid of full probability vectors over d outcomes could require exponentially many report cells; action-pair cells avoid that grid.

Why the weaker condition helps

We control only the forecast errors that could change a utility comparison, rather than every coordinate of every possible forecast.

Calibrated Recommendations Give Short Conversations

Theorem

Suppose $u(a, y) \in [0, 1]$ *and* $r_{k,b,a} = 0$ *for every action pair at every reply position* $k = 3, \ldots, R$. *Then*

$$\boxed{\frac{|S_{R+1}|}{T} \leq \frac{1}{(R-2)\varepsilon}.}$$

Every conversation that ends by acceptance selects a mutually ε*-acceptable action.*

Calibrated Recommendations Give Short Conversations

Theorem

Suppose $u(a, y) \in [0, 1]$ *and* $r_{k,b,a} = 0$ *for every action pair at every reply position* $k = 3, \ldots, R$*. Then*

$$\boxed{\frac{|S_{R+1}|}{T} \leq \frac{1}{(R-2)\varepsilon}.}$$

Every conversation that ends by acceptance selects a mutually ε*-acceptable action.*

For any $\delta \in (0, 1)$, at least a $1 - \delta$ fraction agree within

$$R = 2 + \left\lceil \frac{1}{\delta\varepsilon} \right\rceil$$

messages: two openings followed by the indicated replies.

Why $1/\varepsilon$ Rather Than $1/\varepsilon^2$?

Under exact calibration, both proofs accumulate progress across days. The rates differ because they measure different kinds of progress.

Lecture 11: probability agreement

The squared-error argument squares the gap between successive reports. A gap exceeding ε contributes at least ε^2 to the aggregate progress bound, giving $O(1/(\delta\varepsilon^2))$ messages.

Why $1/\varepsilon$ Rather Than $1/\varepsilon^2$?

Under exact calibration, both proofs accumulate progress across days. The rates differ because they measure different kinds of progress.

Lecture 11: probability agreement

The squared-error argument squares the gap between successive reports. A gap exceeding ε contributes at least ε^2 to the aggregate progress bound, giving $O(1/(\delta\varepsilon^2))$ messages.

This lecture: action agreement

Every counterrecommendation predicts a utility gain exceeding ε. Conversation decision calibration transfers the sum of these gains to realized utility progress. Bounded utility then gives $O(1/(\delta\varepsilon))$ messages.

Why $1/\varepsilon$ Rather Than $1/\varepsilon^2$?

Under exact calibration, both proofs accumulate progress across days. The rates differ because they measure different kinds of progress.

Lecture 11: probability agreement

The squared-error argument squares the gap between successive reports. A gap exceeding ε contributes at least ε^2 to the aggregate progress bound, giving $O(1/(\delta\varepsilon^2))$ messages.

This lecture: action agreement

Every counterrecommendation predicts a utility gain exceeding ε. Conversation decision calibration transfers the sum of these gains to realized utility progress. Bounded utility then gives $O(1/(\delta\varepsilon))$ messages.

The tolerances differ: Lecture 11 measures probability disagreement; here ε measures utility loss. We gain by targeting decisions directly.

Track the Action Actually Retained

On day t, c_t^1 and c_t^2 are the AI's and human's openings. At each reply $k \geq 3$, the previous recommendation is $b_t^k = c_t^{k-1}$. The retained action is

$$c_t^k = \begin{cases} b_t^k, & \text{if the reply accepts,} \\ a_t^k, & \text{if the reply counterrecommends.} \end{cases}$$

After stopping, keep c_t^k fixed. Thus c_t^R is the actual output.

Track the Action Actually Retained

On day t, c_t^1 and c_t^2 are the AI's and human's openings. At each reply $k \geq 3$, the previous recommendation is $b_t^k = c_t^{k-1}$. The retained action is

$$c_t^k = \begin{cases} b_t^k, & \text{if the reply accepts,} \\ a_t^k, & \text{if the reply counterrecommends.} \end{cases}$$

After stopping, keep c_t^k fixed. Thus c_t^R is the actual output. Measure the total realized utility of these recommendations:

$$\Phi_k := \sum_{t=1}^{T} u(c_t^k, y_t), \qquad 0 \leq \Phi_k \leq T.$$

Φ_1 and Φ_2 evaluate the two openings. Only counterrecommendations change this total from round 3 onward.

Every Unresolved Reply Forces Utility Progress

Recall $\Phi_k = \sum_t u(c_t^k, y_t)$, the utility of the retained actions. For $k \geq 3$, actions change only on counterrecommending days S_{k+1}:

$$\Phi_k - \Phi_{k-1} = \sum_{t \in S_{k+1}} \left[u(a_t^k, y_t) - u(b_t^k, y_t) \right].$$

Every Unresolved Reply Forces Utility Progress

Recall $\Phi_k = \sum_t u(c_t^k, y_t)$, the utility of the retained actions. For $k \geq 3$, actions change only on counterrecommending days S_{k+1}:

$$\Phi_k - \Phi_{k-1} = \sum_{t \in S_{k+1}} \left[u(a_t^k, y_t) - u(b_t^k, y_t) \right].$$

Add and subtract the predicted gains g_t^k, then group their errors:

$$\Phi_k - \Phi_{k-1} = \sum_{t \in S_{k+1}} g_t^k + \sum_{b,a} r_{k,b,a}.$$

Every Unresolved Reply Forces Utility Progress

Recall $\Phi_k = \sum_t u(c_t^k, y_t)$, the utility of the retained actions. For $k \geq 3$, actions change only on counterrecommending days S_{k+1}:

$$\Phi_k - \Phi_{k-1} = \sum_{t \in S_{k+1}} \left[u(a_t^k, y_t) - u(b_t^k, y_t) \right].$$

Add and subtract the predicted gains g_t^k, then group their errors:

$$\Phi_k - \Phi_{k-1} = \sum_{t \in S_{k+1}} g_t^k + \sum_{b,a} r_{k,b,a}.$$

Exact calibration makes the second sum zero. Every remaining term exceeds ε, so

$$\boxed{\Phi_k - \Phi_{k-1} \geq \varepsilon |S_{k+1}|.}$$

Accepting replies contribute zero: they keep the previous action.

Bounded Utility Limits Unresolved Conversations

Sum the progress inequality over replies $k = 3, \ldots, R$:

$$\Phi_R - \Phi_2 \geq \varepsilon \sum_{k=3}^{R} |S_{k+1}|.$$

Bounded Utility Limits Unresolved Conversations

Sum the progress inequality over replies $k = 3, \ldots, R$:

$$\Phi_R - \Phi_2 \geq \varepsilon \sum_{k=3}^{R} |S_{k+1}|.$$

Every day still unresolved at the cap counterrecommended at all $R - 2$ reply positions. Therefore

$$\sum_{k=3}^{R} |S_{k+1}| \geq (R-2)|S_{R+1}|.$$

Bounded Utility Limits Unresolved Conversations

Sum the progress inequality over replies $k = 3, \ldots, R$:

$$\Phi_R - \Phi_2 \geq \varepsilon \sum_{k=3}^{R} |S_{k+1}|.$$

Every day still unresolved at the cap counterrecommended at all $R - 2$ reply positions. Therefore

$$\sum_{k=3}^{R} |S_{k+1}| \geq (R-2)|S_{R+1}|.$$

Since $0 \leq \Phi_k \leq T$, these inequalities give

$$T \geq \Phi_R - \Phi_2 \geq \varepsilon(R-2)|S_{R+1}|,$$

and hence

$$\boxed{\frac{|S_{R+1}|}{T} \leq \frac{1}{(R-2)\varepsilon}.}$$

This proves the theorem. □

Does Agreement Improve on Both Opening Decisions?

Progress starts from the human's opening c^2. What about the AI's opening? Suppose half the days are safe:

- The AI initially recommends deploy on safe days and hold on harmful days: $\Phi_1 = T/2$.
- The human always recommends hold: $\Phi_2 = 0$.

Does Agreement Improve on Both Opening Decisions?

Progress starts from the human's opening c^2. What about the AI's opening? Suppose half the days are safe:

- The AI initially recommends deploy on safe days and hold on harmful days: $\Phi_1 = T/2$.
- The human always recommends hold: $\Phi_2 = 0$.

The first reply instead forecasts safe with probability $1/2$, ignoring the AI's initial information. Under this forecast, hold is worth 0 and audit $3/10$. Set $\varepsilon = 2/5$: the reply accepts hold.

Does Agreement Improve on Both Opening Decisions?

Progress starts from the human's opening c^2. What about the AI's opening? Suppose half the days are safe:

- The AI initially recommends deploy on safe days and hold on harmful days: $\Phi_1 = T/2$.
- The human always recommends hold: $\Phi_2 = 0$.

The first reply instead forecasts safe with probability $1/2$, ignoring the AI's initial information. Under this forecast, hold is worth 0 and audit $3/10$. Set $\varepsilon = 2/5$: the reply accepts hold.
There are no counterrecommendations, so all their calibration checks pass. Yet the output loses $T/2$ relative to the AI's opening, more than εT.

The missing check

The first AI reply must also compare its retained action with the AI's opening, not just the human's opening.

One Extra Check Compares with the AI's Opening

On every day, compare the action c_t^3 retained after the first reply with the AI's opening c_t^1. For each pair (b, c), group the days into

$$D_{b,c} := \{t : c_t^1 = b,\ c_t^3 = c\}.$$

One Extra Check Compares with the AI's Opening

On every day, compare the action c_t^3 retained after the first reply with the AI's opening c_t^1. For each pair (b, c), group the days into

$$D_{b,c} := \{t : c_t^1 = b,\ c_t^3 = c\}.$$

Sum the error in the predicted utility difference:

$$s_{b,c} := \sum_{t \in D_{b,c}} \left[u(c, y_t) - u(b, y_t) - \left(U(c, p_t^3) - U(b, p_t^3)\right)\right].$$

Require $s_{b,c} = 0$ for every pair.

One Extra Check Compares with the AI's Opening

On every day, compare the action c_t^3 retained after the first reply with the AI's opening c_t^1. For each pair (b, c), group the days into

$$D_{b,c} := \{t : c_t^1 = b,\ c_t^3 = c\}.$$

Sum the error in the predicted utility difference:

$$s_{b,c} := \sum_{t \in D_{b,c}} \left[u(c, y_t) - u(b, y_t) - \left(U(c, p_t^3) - U(b, p_t^3)\right)\right].$$

Require $s_{b,c} = 0$ for every pair.
This is the same kind of utility-comparison check as before, but it compares with the AI's opening and includes accepting replies.
We needed no such check to prove short conversations. It strengthens the guarantee about the decision they reach.

The First Reply Matches the Better Opening, Up to ε

Our target is $\Phi_3 \geq \max\{\Phi_1, \Phi_2\} - \varepsilon T$. The progress inequality already gives $\Phi_3 \geq \Phi_2$.

The First Reply Matches the Better Opening, Up to ε

Our target is $\Phi_3 \geq \max\{\Phi_1, \Phi_2\} - \varepsilon T$. The progress inequality already gives $\Phi_3 \geq \Phi_2$.

The retained action c_t^3 is ε-optimal under p_t^3: acceptance keeps an ε-optimal action; a counterrecommendation chooses a best response. Thus

$$U(c_t^3, p_t^3) - U(c_t^1, p_t^3) \geq -\varepsilon.$$

The First Reply Matches the Better Opening, Up to ε

Our target is $\Phi_3 \geq \max\{\Phi_1, \Phi_2\} - \varepsilon T$. The progress inequality already gives $\Phi_3 \geq \Phi_2$.

The retained action c_t^3 is ε-optimal under p_t^3: acceptance keeps an ε-optimal action; a counterrecommendation chooses a best response. Thus

$$U(c_t^3, p_t^3) - U(c_t^1, p_t^3) \geq -\varepsilon.$$

The extra check transfers this comparison to realized utility:

$$\Phi_3 - \Phi_1 = \sum_t \left[U(c_t^3, p_t^3) - U(c_t^1, p_t^3)\right] + \underbrace{\sum_{b,c} s_{b,c}}_{=0} \geq -\varepsilon T.$$

The First Reply Matches the Better Opening, Up to ε

Our target is $\Phi_3 \geq \max\{\Phi_1, \Phi_2\} - \varepsilon T$. The progress inequality already gives $\Phi_3 \geq \Phi_2$.

The retained action c_t^3 is ε-optimal under p_t^3: acceptance keeps an ε-optimal action; a counterrecommendation chooses a best response. Thus

$$U(c_t^3, p_t^3) - U(c_t^1, p_t^3) \geq -\varepsilon.$$

The extra check transfers this comparison to realized utility:

$$\Phi_3 - \Phi_1 = \sum_t \left[U(c_t^3, p_t^3) - U(c_t^1, p_t^3)\right] + \underbrace{\sum_{b,c} s_{b,c}}_{=0} \geq -\varepsilon T.$$

Combining the two opening comparisons,

$$\boxed{\Phi_3 \geq \max\{\Phi_1, \Phi_2\} - \varepsilon T.}$$

The allowance εT comes from accepting an approximately optimal action.

The Same Progress Improves on the Better Opening

Let $\bar{N}$ count counterrecommendations after the first reply, averaged over days:

$$\bar{N} := \frac{1}{T} \sum_{k=4}^{R} |S_{k+1}|.$$

Our progress inequality, summed over these later replies, gives

$$\Phi_R - \Phi_3 \geq \varepsilon T \bar{N}.$$

The Same Progress Improves on the Better Opening

Let $\bar{N}$ count counterrecommendations after the first reply, averaged over days:

$$\bar{N} := \frac{1}{T} \sum_{k=4}^{R} |S_{k+1}|.$$

Our progress inequality, summed over these later replies, gives

$$\Phi_R - \Phi_3 \geq \varepsilon T \bar{N}.$$

Add the opening guarantee $\Phi_3 \geq \max\{\Phi_1, \Phi_2\} - \varepsilon T$:

$$\boxed{\frac{\Phi_R - \max\{\Phi_1, \Phi_2\}}{T} \geq \varepsilon \bar{N} - \varepsilon.}$$

The Same Progress Improves on the Better Opening

Let $\bar{N}$ count counterrecommendations after the first reply, averaged over days:

$$\bar{N} := \frac{1}{T} \sum_{k=4}^{R} |S_{k+1}|.$$

Our progress inequality, summed over these later replies, gives

$$\Phi_R - \Phi_3 \geq \varepsilon T \bar{N}.$$

Add the opening guarantee $\Phi_3 \geq \max\{\Phi_1, \Phi_2\} - \varepsilon T$:

$$\boxed{\frac{\Phi_R - \max\{\Phi_1, \Phi_2\}}{T} \geq \varepsilon \bar{N} - \varepsilon.}$$

The final decisions match the better opening's average utility up to ε and strictly improve on it when $\bar{N} > 1$. It is an across-days comparison, not a day-by-day guarantee.

Approximate Calibration Adds an Error Allowance

Exact calibration exposed the mechanism. Now allow errors, measuring the total absolute residual at reply position k by

$$\Lambda_k := \underbrace{\sum_{b,a} |r_{k,b,a}|}_{\text{counterrecommendations}} + \mathbf{1}\{k = 3\} \underbrace{\sum_{b,c} |s_{b,c}|}_{\text{comparison with AI opening}} .$$

Empty cells contribute zero. Absolute values prevent errors in different cells from cancelling.

Approximate Calibration Adds an Error Allowance

Exact calibration exposed the mechanism. Now allow errors, measuring the total absolute residual at reply position k by

$$\Lambda_k := \underbrace{\sum_{b,a} |r_{k,b,a}|}_{\text{counterrecommendations}} + \mathbf{1}\{k=3\} \underbrace{\sum_{b,c} |s_{b,c}|}_{\text{comparison with AI opening}} .$$

Empty cells contribute zero. Absolute values prevent errors in different cells from cancelling.
The same two calculations now give

$$\Phi_k - \Phi_{k-1} \geq \varepsilon |S_{k+1}| - \Lambda_k \qquad (k \geq 3),$$

$$\Phi_3 \geq \max\{\Phi_1, \Phi_2\} - \varepsilon T - \Lambda_3.$$

Each sum of residuals is bounded below by minus its sum of absolute values. The rest of the proof is unchanged.

The Same Proof Gives Approximate Guarantees

Sum the corrected progress inequalities exactly as before:

$$\boxed{\frac{|S_{R+1}|}{T} \leq \frac{1}{(R-2)\varepsilon} + \frac{\sum_{k=3}^{R} \Lambda_k}{T(R-2)\varepsilon}.}$$

The Same Proof Gives Approximate Guarantees

Sum the corrected progress inequalities exactly as before:

$$\boxed{\frac{|S_{R+1}|}{T} \leq \frac{1}{(R-2)\varepsilon} + \frac{\sum_{k=3}^{R} \Lambda_k}{T(R-2)\varepsilon}.}$$

Combining the first-reply bound with later utility progress gives

$$\boxed{\frac{\Phi_R - \max\{\Phi_1, \Phi_2\}}{T} \geq \varepsilon \bar{N} - \varepsilon - \frac{1}{T} \sum_{k=3}^{R} \Lambda_k.}$$

The Same Proof Gives Approximate Guarantees

Sum the corrected progress inequalities exactly as before:

$$\boxed{\frac{|S_{R+1}|}{T} \leq \frac{1}{(R-2)\varepsilon} + \frac{\sum_{k=3}^{R} \Lambda_k}{T(R-2)\varepsilon}.}$$

Combining the first-reply bound with later utility progress gives

$$\boxed{\frac{\Phi_R - \max\{\Phi_1, \Phi_2\}}{T} \geq \varepsilon \bar{N} - \varepsilon - \frac{1}{T} \sum_{k=3}^{R} \Lambda_k.}$$

If $\Lambda_k / T \leq \gamma$ at every reply position, calibration adds γ/ε to the unresolved fraction and at most $(R-2)\gamma$ to the utility allowance. These forms also let us use expected calibration-error bounds: take expectations on both sides.

Learning Conversation Decision Calibration Online

So far, our agreement and utility guarantees assume small conversation decision-calibration errors. We now give an efficient randomized online algorithm that chooses forecasts with small errors.

Learning Conversation Decision Calibration Online

So far, our agreement and utility guarantees assume small conversation decision-calibration errors. We now give an efficient randomized online algorithm that chooses forecasts with small errors.

The algorithm learns across days: it chooses each internal forecast before the day's outcome is revealed, then uses the outcome to update for future days.

Learning Conversation Decision Calibration Online

So far, our agreement and utility guarantees assume small conversation decision-calibration errors. We now give an efficient randomized online algorithm that chooses forecasts with small errors.

The algorithm learns across days: it chooses each internal forecast before the day's outcome is revealed, then uses the outcome to update for future days.

We will use Blackwell approachability from Lecture 8, as we did for conversation calibration in Lecture 11. This time, we want to drive the average action-pair residuals toward zero, ensuring approximate conversation decision calibration.

Learning Conversation Decision Calibration Online

So far, our agreement and utility guarantees assume small conversation decision-calibration errors. We now give an efficient randomized online algorithm that chooses forecasts with small errors.

The algorithm learns across days: it chooses each internal forecast before the day's outcome is revealed, then uses the outcome to update for future days.

We will use Blackwell approachability from Lecture 8, as we did for conversation calibration in Lecture 11. This time, we want to drive the average action-pair residuals toward zero, ensuring approximate conversation decision calibration.

The algorithmic goal

Make the expected calibration error $\mathbb{E}[\Lambda_k / T]$ small at each reply position k, so our conversation-length and utility guarantees apply to the forecasts we actually compute.

Learn Each Reply from Repeated Outcome Feedback

As in Lecture 11, a reporting assistant can enforce calibration for either speaker; an unassisted human need not satisfy it.

Learn Each Reply from Repeated Outcome Feedback

As in Lecture 11, a reporting assistant can enforce calibration for either speaker; an unassisted human need not satisfy it.
Run a separate forecasting procedure for each reply position $k \geq 3$:

1. Observe the earlier recommendation(s).
2. Compute and draw an internal forecast; accept or counterrecommend.
3. When the day's outcome is revealed, update the procedure with its utility-comparison errors. Inactive days give zero updates.

Learn Each Reply from Repeated Outcome Feedback

As in Lecture 11, a reporting assistant can enforce calibration for either speaker; an unassisted human need not satisfy it.
Run a separate forecasting procedure for each reply position $k \geq 3$:

1. Observe the earlier recommendation(s).
2. Compute and draw an internal forecast; accept or counterrecommend.
3. When the day's outcome is revealed, update the procedure with its utility-comparison errors. Inactive days give zero updates.

Keep the three timescales separate:

T days	outcome feedback for learning
up to R messages	communication within one day
J steps per reply	internal computation of a forecast

The J steps require no new outcomes or messages.

Each Utility Comparison Supplies a Pairwise Error

Put $n := |\mathcal{A}|$. For a forecast p, let $c(p)$ be the action the reply retains: the previous action b on acceptance, and $\mathrm{BR}(p)$ otherwise. For a comparison of c with b, define an n^2-coordinate vector:

$$v(b, c, p, y)_{b',c'} := \mathbf{1}\{b' = b,\ c' = c\}\langle u_c - u_b, e_y - p\rangle.$$

Each Utility Comparison Supplies a Pairwise Error

Put $n := |\mathcal{A}|$. For a forecast p, let $c(p)$ be the action the reply retains: the previous action b on acceptance, and $\mathrm{BR}(p)$ otherwise. For a comparison of c with b, define an n^2-coordinate vector:

$$v(b, c, p, y)_{b',c'} := \mathbf{1}\{b' = b,\ c' = c\}\langle u_c - u_b, e_y - p\rangle.$$

At reply position k, on an active day t, use

$$h_t(p, y) := \begin{cases} \big(v(c_t^2, c(p), p, y), v(c_t^1, c(p), p, y)\big), & k = 3, \\ v(b_t^k, c(p), p, y), & k \geq 4. \end{cases}$$

Each Utility Comparison Supplies a Pairwise Error

Put $n := |\mathcal{A}|$. For a forecast p, let $c(p)$ be the action the reply retains: the previous action b on acceptance, and $\mathrm{BR}(p)$ otherwise. For a comparison of c with b, define an n^2-coordinate vector:

$$v(b, c, p, y)_{b',c'} := \mathbf{1}\{b' = b,\ c' = c\}\langle u_c - u_b, e_y - p\rangle.$$

At reply position k, on an active day t, use

$$h_t(p, y) := \begin{cases} \big(v(c_t^2, c(p), p, y), v(c_t^1, c(p), p, y)\big), & k = 3, \\ v(b_t^k, c(p), p, y), & k \geq 4. \end{cases}$$

Set $h_t = 0$ on inactive days and average the observed vectors into $\bar{h}_T$. Its coordinates are $r_{k,b,a}/T$, with $s_{b,c}/T$ appended at $k = 3$:

$$\|\bar{h}_T\|_1 = \Lambda_k/T.$$

Approachability Leaves One Forecasting Problem

Our exact target is $\bar{h}_T = 0$: every calibration residual vanishes. As in Lecture 11, allow a small neighborhood of this zero-error target:

$$\mathcal{K}_\eta := \{x : \|x\|_1 \le 2\eta\}, \qquad \eta > 0.$$

The allowance will absorb error from approximate computation.

Approachability Leaves One Forecasting Problem

Our exact target is $\bar{h}_T = 0$: every calibration residual vanishes. As in Lecture 11, allow a small neighborhood of this zero-error target:

$$\mathcal{K}_\eta := \{x : \|x\|_1 \le 2\eta\}, \qquad \eta > 0.$$

The allowance will absorb error from approximate computation. Project the current average error $\bar{h}_{t-1}$ onto $\mathcal{K}_\eta$, obtaining z, and put $w = \bar{h}_{t-1} - z$. The approachability proof asks for a distribution ν_t over forecasts with

$$\boxed{\mathbb{E}_{p\sim\nu_t}\langle w, h_t(p, y) - z\rangle \le 0 \quad \text{for every } y \in \mathcal{Y}.}$$

If $w = 0$, any forecast works.

Approachability Leaves One Forecasting Problem

Our exact target is $\bar{h}_T = 0$: every calibration residual vanishes. As in Lecture 11, allow a small neighborhood of this zero-error target:

$$\mathcal{K}_\eta := \{x : \|x\|_1 \le 2\eta\}, \qquad \eta > 0.$$

The allowance will absorb error from approximate computation. Project the current average error $\bar{h}_{t-1}$ onto $\mathcal{K}_\eta$, obtaining z, and put $w = \bar{h}_{t-1} - z$. The approachability proof asks for a distribution ν_t over forecasts with

$$\boxed{\mathbb{E}_{p\sim\nu_t}\langle w, h_t(p, y) - z\rangle \le 0 \quad \text{for every } y \in \mathcal{Y}.}$$

If $w = 0$, any forecast works.
For $w \neq 0$, we still need to compute ν_t. We must do so without a grid over probability vectors, which would be too large to maintain efficiently.

The Enlarged Target Allows Computational Error

Rearrange approachability's condition to expose the allowed weighted error:

$$\mathbb{E}_{p\sim\nu_t}\langle w, h_t(p,y)\rangle \le \langle w, z\rangle \qquad \text{for every } y.$$

The Enlarged Target Allows Computational Error

Rearrange approachability's condition to expose the allowed weighted error:

$$\mathbb{E}_{p\sim\nu_t}\langle w, h_t(p,y)\rangle \le \langle w, z\rangle \qquad \text{for every } y.$$

The projection inequality $\langle w, x - z\rangle \le 0$ for $x \in \mathcal{K}_\eta$ says that z maximizes $\langle w, x\rangle$ over the target. Thus

$$\langle w, z\rangle = \max_{\|x\|_1 \le 2\eta} \langle w, x\rangle = 2\eta\|w\|_\infty.$$

Here $\|w\|_\infty := \max_j |w_j|$. The upper bound follows from $\langle w, x\rangle \le \|w\|_\infty \|x\|_1$. Equality is attained by putting all the allowed magnitude on a coordinate maximizing $|w_j|$, with the same sign.

The Enlarged Target Allows Computational Error

Rearrange approachability's condition to expose the allowed weighted error:

$$\mathbb{E}_{p\sim\nu_t}\langle w, h_t(p,y)\rangle \le \langle w, z\rangle \qquad \text{for every } y.$$

The projection inequality $\langle w, x - z\rangle \le 0$ for $x \in \mathcal{K}_\eta$ says that z maximizes $\langle w, x\rangle$ over the target. Thus

$$\langle w, z\rangle = \max_{\|x\|_1 \le 2\eta} \langle w, x\rangle = 2\eta\|w\|_\infty.$$

Here $\|w\|_\infty := \max_j |w_j|$. The upper bound follows from $\langle w, x\rangle \le \|w\|_\infty \|x\|_1$. Equality is attained by putting all the allowed magnitude on a coordinate maximizing $|w_j|$, with the same sign. At the zero-error vector, the weighted error is zero. For $w \neq 0$, the allowed value $2\eta\|w\|_\infty$ is positive, leaving room for approximate computation.

Using No Regret to Compute the Forecast

Use the current weighted calibration error as a zero-sum game's payoff:

$$A(p, y) := \langle w, h_t(p, y) \rangle.$$

The forecasting player chooses p to minimize this payoff. The outcome player chooses y to maximize it.

Using No Regret to Compute the Forecast

Use the current weighted calibration error as a zero-sum game's payoff:

$$A(p, y) := \langle w, h_t(p, y) \rangle.$$

The forecasting player chooses p to minimize this payoff. The outcome player chooses y to maximize it.

Approachability requires a forecast distribution ν_t with

$$\max_y \mathbb{E}_{p \sim \nu_t} A(p, y) \leq 2\eta \|w\|_\infty.$$

Using No Regret to Compute the Forecast

Use the current weighted calibration error as a zero-sum game's payoff:

$$A(p, y) := \langle w, h_t(p, y) \rangle.$$

The forecasting player chooses p to minimize this payoff. The outcome player chooses y to maximize it.
Approachability requires a forecast distribution ν_t with

$$\max_y \mathbb{E}_{p \sim \nu_t} A(p, y) \leq 2\eta \|w\|_\infty.$$

We run Hedge over outcomes. At each simulated step, we will choose a forecast that makes *Hedge's expected payoff* zero.

Using No Regret to Compute the Forecast

Use the current weighted calibration error as a zero-sum game's payoff:

$$A(p, y) := \langle w, h_t(p, y) \rangle.$$

The forecasting player chooses p to minimize this payoff. The outcome player chooses y to maximize it.
Approachability requires a forecast distribution ν_t with

$$\max_y \mathbb{E}_{p \sim \nu_t} A(p, y) \leq 2\eta \|w\|_\infty.$$

We run Hedge over outcomes. At each simulated step, we will choose a forecast that makes *Hedge's expected payoff* zero. Hedge's regret guarantee then says that every fixed outcome's average payoff is at most the average regret. Equivalently, the uniform mixture of our forecasts has payoff at most the average regret against every outcome.
We will run enough internal steps to make this regret bound at most $2\eta \|w\|_\infty$.

Copying Gives the Response Needed for Self-Play

Against the outcome learner's distribution Q, choose the forecast $p = Q$. We now check that this makes the expected calibration error vector zero, and hence makes Hedge's expected payoff zero.

Copying Gives the Response Needed for Self-Play

Against the outcome learner's distribution Q, choose the forecast $p = Q$. We now check that this makes the expected calibration error vector zero, and hence makes Hedge's expected payoff zero. Once $p = Q$ is fixed, the compared actions and cell labels are fixed too. Each error block then has mean zero:

$$\mathbb{E}_{Y\sim Q}\langle u_c - u_b, e_Y - Q\rangle = \langle u_c - u_b, Q - Q\rangle = 0.$$

Consequently,

$$\mathbb{E}_{Y\sim Q} h_t(Q, Y) = 0, \qquad \boxed{\mathbb{E}_{Y\sim Q} A(Q, Y) = 0.}$$

These identities hold for every Q.

Copying Gives the Response Needed for Self-Play

Against the outcome learner's distribution Q, choose the forecast $p = Q$. We now check that this makes the expected calibration error vector zero, and hence makes Hedge's expected payoff zero. Once $p = Q$ is fixed, the compared actions and cell labels are fixed too. Each error block then has mean zero:

$$\mathbb{E}_{Y\sim Q}\langle u_c - u_b, e_Y - Q\rangle = \langle u_c - u_b, Q - Q\rangle = 0.$$

Consequently,

$$\mathbb{E}_{Y\sim Q} h_t(Q, Y) = 0, \qquad \boxed{\mathbb{E}_{Y\sim Q} A(Q, Y) = 0.}$$

These identities hold for every Q.
Here Q is an internal distribution maintained by Hedge. Copying it requires no knowledge of the true outcome law and no search over possible forecasts.

Review: Zero Payoff and Low Regret Give Robustness

For gains $f_s(y) \in [-B, B]$, with $B > 0$, Hedge starts with $m_1(y) = 1$ and rescales gains by B:

$$Q_s(y) = \frac{m_s(y)}{\sum_{y'} m_s(y')}, \qquad m_{s+1}(y) = m_s(y) e^{\lambda f_s(y)/B}.$$

With $\lambda = \sqrt{2 \log d / J}$, the regret bound is

$$\max_y \frac{1}{J} \sum_s f_s(y) - \frac{1}{J} \sum_s \sum_y Q_s(y) f_s(y) \leq B \sqrt{\frac{2 \log d}{J}}.$$

Review: Zero Payoff and Low Regret Give Robustness

For gains $f_s(y) \in [-B, B]$, with $B > 0$, Hedge starts with $m_1(y) = 1$ and rescales gains by B:

$$Q_s(y) = \frac{m_s(y)}{\sum_{y'} m_s(y')}, \qquad m_{s+1}(y) = m_s(y) e^{\lambda f_s(y)/B}.$$

With $\lambda = \sqrt{2 \log d / J}$, the regret bound is

$$\max_y \frac{1}{J} \sum_s f_s(y) - \frac{1}{J} \sum_s \sum_y Q_s(y) f_s(y) \le B \sqrt{\frac{2 \log d}{J}}.$$

If the learner's expected gain is zero on each step, $\sum_y Q_s(y) f_s(y) = 0$, Lecture 5's self-play argument gives

$$\max_y \frac{1}{J} \sum_s f_s(y) \le B \sqrt{\frac{2 \log d}{J}}.$$

Review: Zero Payoff and Low Regret Give Robustness

For gains $f_s(y) \in [-B, B]$, with $B > 0$, Hedge starts with $m_1(y) = 1$ and rescales gains by B:

$$Q_s(y) = \frac{m_s(y)}{\sum_{y'} m_s(y')}, \qquad m_{s+1}(y) = m_s(y) e^{\lambda f_s(y)/B}.$$

With $\lambda = \sqrt{2 \log d / J}$, the regret bound is

$$\max_y \frac{1}{J} \sum_s f_s(y) - \frac{1}{J} \sum_s \sum_y Q_s(y) f_s(y) \le B \sqrt{\frac{2 \log d}{J}}.$$

If the learner's expected gain is zero on each step, $\sum_y Q_s(y) f_s(y) = 0$, Lecture 5's self-play argument gives

$$\max_y \frac{1}{J} \sum_s f_s(y) \le B \sqrt{\frac{2 \log d}{J}}.$$

Hedge allows the gains to depend on its current distribution Q_s.

Self-Play Computes the Forecasting Distribution

Run J internal Hedge steps over the d possible outcomes:

1. Let Q_s be Hedge's distribution over outcomes.
2. Choose the candidate forecast $p_s = Q_s$.
3. Update Hedge with the gains $f_s(y) = A(p_s, y)$ for every y.

Let ν_t sample uniformly from the J candidate forecasts.

Self-Play Computes the Forecasting Distribution

Run J internal Hedge steps over the d possible outcomes:

1. Let Q_s be Hedge's distribution over outcomes.
2. Choose the candidate forecast $p_s = Q_s$.
3. Update Hedge with the gains $f_s(y) = A(p_s, y)$ for every y.

Let ν_t sample uniformly from the J candidate forecasts.
The learner's expected gain is zero by copying. Since h_t has at most two nonzero entries, each of magnitude at most 2, the gain range is bounded by $B := 4\|w\|_\infty$. Thus

$$\max_y \mathbb{E}_{p\sim\nu_t}\langle w, h_t(p, y)\rangle \le 4\|w\|_\infty \sqrt{\frac{2\log d}{J}}.$$

Self-Play Computes the Forecasting Distribution

Run J internal Hedge steps over the d possible outcomes:

1. Let Q_s be Hedge's distribution over outcomes.
2. Choose the candidate forecast $p_s = Q_s$.
3. Update Hedge with the gains $f_s(y) = A(p_s, y)$ for every y.

Let ν_t sample uniformly from the J candidate forecasts.
The learner's expected gain is zero by copying. Since h_t has at most two nonzero entries, each of magnitude at most 2, the gain range is bounded by $B := 4\|w\|_\infty$. Thus

$$\max_y \mathbb{E}_{p\sim\nu_t}\langle w, h_t(p, y)\rangle \le 4\|w\|_\infty \sqrt{\frac{2\log d}{J}}.$$

Taking $J \ge 8\log d/\eta^2$ makes this at most $2\eta\|w\|_\infty$. Since $\langle w, z\rangle = 2\eta\|w\|_\infty$, for every outcome y we obtain

$$\mathbb{E}_{p\sim\nu_t}\langle w, h_t(p, y) - z\rangle \le 2\eta\|w\|_\infty - \langle w, z\rangle = 0.$$

This is the required approachability condition.

Lecture 11's Bound Now Controls Utility Errors

Each vector has at most two nonzero entries, each of magnitude at most 2. Thus $\|h_t\|_2 \le 2\sqrt{2}$, and approachability gives

$$\mathbb{E}[\mathrm{dist}(\bar{h}_T, \mathcal{K}_\eta)^2] \le \frac{32}{T}.$$

Lecture 11's Bound Now Controls Utility Errors

Each vector has at most two nonzero entries, each of magnitude at most 2. Thus $\|h_t\|_2 \le 2\sqrt{2}$, and approachability gives

$$\mathbb{E}[\mathrm{dist}(\bar{h}_T, \mathcal{K}_\eta)^2] \le \frac{32}{T}.$$

There are at most $2n^2$ coordinates. The same projection and Cauchy–Schwarz conversion as in Lecture 11 gives

$$\frac{\Lambda_k}{T} \le 2\eta + \sqrt{2}\, n\, \mathrm{dist}(\bar{h}_T, \mathcal{K}_\eta).$$

Taking expectations,

$$\boxed{\mathbb{E}[\Lambda_k / T] \le 2\eta + \frac{8n}{\sqrt{T}} =: \gamma_T.}$$

Lecture 11's Bound Now Controls Utility Errors

Each vector has at most two nonzero entries, each of magnitude at most 2. Thus $\|h_t\|_2 \le 2\sqrt{2}$, and approachability gives

$$\mathbb{E}[\mathrm{dist}(\bar{h}_T, \mathcal{K}_\eta)^2] \le \frac{32}{T}.$$

There are at most $2n^2$ coordinates. The same projection and Cauchy–Schwarz conversion as in Lecture 11 gives

$$\frac{\Lambda_k}{T} \le 2\eta + \sqrt{2}\, n\, \mathrm{dist}(\bar{h}_T, \mathcal{K}_\eta).$$

Taking expectations,

$$\boxed{\mathbb{E}[\Lambda_k / T] \le 2\eta + \frac{8n}{\sqrt{T}} =: \gamma_T.}$$

The same bound works at every reply position, and it does not depend on the number of outcomes d. The forecasting calculation is also polynomial in d, n, and $1/\eta$: Hedge supplies the per-reply response efficiently.

Learned Calibration Can Improve on Both Openings

Take expectations in our deterministic guarantees and use $\mathbb{E}[\Lambda_k/T] \le \gamma_T$ for each of the $R-2$ replies:

$$\boxed{\mathbb{E}\left[\frac{|S_{R+1}|}{T}\right] \le \frac{1}{(R-2)\varepsilon} + \frac{\gamma_T}{\varepsilon},}$$

Learned Calibration Can Improve on Both Openings

Take expectations in our deterministic guarantees and use $\mathbb{E}[\Lambda_k/T] \le \gamma_T$ for each of the $R-2$ replies:

$$\boxed{\mathbb{E}\left[\frac{|S_{R+1}|}{T}\right] \le \frac{1}{(R-2)\varepsilon} + \frac{\gamma_T}{\varepsilon},}$$

$$\boxed{\mathbb{E}\left[\frac{\Phi_R - \max\{\Phi_1, \Phi_2\}}{T}\right] \ge \varepsilon\mathbb{E}[\bar{N}] - \varepsilon - (R-2)\gamma_T.}$$

Learned Calibration Can Improve on Both Openings

Take expectations in our deterministic guarantees and use $\mathbb{E}[\Lambda_k/T] \le \gamma_T$ for each of the $R-2$ replies:

$$\boxed{\mathbb{E}\left[\frac{|S_{R+1}|}{T}\right] \le \frac{1}{(R-2)\varepsilon} + \frac{\gamma_T}{\varepsilon},}$$

$$\boxed{\mathbb{E}\left[\frac{\Phi_R - \max\{\Phi_1, \Phi_2\}}{T}\right] \ge \varepsilon\mathbb{E}[\bar{N}] - \varepsilon - (R-2)\gamma_T.}$$

Here $\bar{N}$ counts counterrecommendations after the first reply, per day. The final actions do at least as well as the better opening in aggregate, up to acceptance and calibration allowances. Further counterrecommendations yield strict improvement when their progress exceeds those allowances.

These expectations average the protocol's random draws over the T days of interaction, not a later deployment period.

Summary

- A shared decision can require much less communication than a shared forecast: exchange recommendations rather than full probability vectors.

Summary

- A shared decision can require much less communication than a shared forecast: exchange recommendations rather than full probability vectors.
- Conversation decision calibration turns predicted gains into realized utility progress. Bounded utility then makes most conversations short.

Summary

- A shared decision can require much less communication than a shared forecast: exchange recommendations rather than full probability vectors.
- Conversation decision calibration turns predicted gains into realized utility progress. Bounded utility then makes most conversations short.
- The final decisions do at least as well as the better opening, up to acceptance and calibration error; further counterrecommendations yield improvement in aggregate utility.

Summary

- A shared decision can require much less communication than a shared forecast: exchange recommendations rather than full probability vectors.
- Conversation decision calibration turns predicted gains into realized utility progress. Bounded utility then makes most conversations short.
- The final decisions do at least as well as the better opening, up to acceptance and calibration error; further counterrecommendations yield improvement in aggregate utility.
- Approachability and zero-payoff self-play make the reliability condition efficiently learnable. Conversation can then use information, but cannot manufacture missing facts—our next topic.

Thanks!

See you next class!

References

- Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth. *Tractable Agreement Protocols*. STOC, 2025.
- Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. *High-Dimensional Prediction for Sequential Decision Making*. ICML, 2025.
- David Blackwell. *An Analog of the Minimax Theorem for Vector Payoffs*. Pacific Journal of Mathematics, 1956.

The lecture uses a simplified common-action protocol and calibrates the utility gains of counterrecommendations; the source paper develops more general agreement protocols. The inner Hedge construction specializes Noarov et al.'s efficient minimax method to a finite outcome set.