One Policy, Many Groups

Multigroup Alignment from Pairwise Feedback

Aaron Roth

University of Pennsylvania

Fall 2026

Overview

Lecture 9 used context-sensitive checks to make one predictor useful to many users. Today we revisit preference alignment and apply that idea to produce a personalized policy that chooses responses from preference feedback and user attributes.

- Pooled robustness can hide systematic failures within groups.

Overview

Lecture 9 used context-sensitive checks to make one predictor useful to many users. Today we revisit preference alignment and apply that idea to produce a personalized policy that chooses responses from preference feedback and user attributes.

- Pooled robustness can hide systematic failures within groups.
- We will seek one contextual policy that is robust within every listed group, including groups that overlap.

Overview

Lecture 9 used context-sensitive checks to make one predictor useful to many users. Today we revisit preference alignment and apply that idea to produce a personalized policy that chooses responses from preference feedback and user attributes.

- Pooled robustness can hide systematic failures within groups.
- We will seek one contextual policy that is robust within every listed group, including groups that overlap.
- Self-play and approachability turn pairwise feedback into a policy that meets these requirements.

Overview

Lecture 9 used context-sensitive checks to make one predictor useful to many users. Today we revisit preference alignment and apply that idea to produce a personalized policy that chooses responses from preference feedback and user attributes.

- Pooled robustness can hide systematic failures within groups.
- We will seek one contextual policy that is robust within every listed group, including groups that overlap.
- Self-play and approachability turn pairwise feedback into a policy that meets these requirements.
- The same policy then inherits Lecture 6's welfare guarantee separately within each group.

Recall: Preference Feedback and Maximal Lotteries

Fix a prompt and a finite completion set $C = \{1, \ldots, m\}$, with $m \geq 2$. Users may prefer different completions.
For distinct $a, b \in C$, let P_{ab} be the probability that a randomly sampled user chooses a over b. Set $P_{aa} = 1/2$ for a self-comparison.
A policy $p \in \Delta(C)$ is a lottery over completions.

Recall: Preference Feedback and Maximal Lotteries

Fix a prompt and a finite completion set $C = \{1, \ldots, m\}$, with $m \geq 2$. Users may prefer different completions.
For distinct $a, b \in C$, let P_{ab} be the probability that a randomly sampled user chooses a over b. Set $P_{aa} = 1/2$ for a self-comparison.
A policy $p \in \Delta(C)$ is a lottery over completions.
A *maximal lottery* satisfies

$$\mathbb{E}_{a \sim p}[P_{ab}] \geq \frac{1}{2} \qquad \text{for every fixed challenger } b \in C.$$

We proved that maximal lotteries always exist.

Recall: Preference Feedback and Maximal Lotteries

Fix a prompt and a finite completion set $C = \{1, \ldots, m\}$, with $m \geq 2$. Users may prefer different completions.
For distinct $a, b \in C$, let P_{ab} be the probability that a randomly sampled user chooses a over b. Set $P_{aa} = 1/2$ for a self-comparison.
A policy $p \in \Delta(C)$ is a lottery over completions.
A *maximal lottery* satisfies

$$\mathbb{E}_{a \sim p}[P_{ab}] \geq \frac{1}{2} \qquad \text{for every fixed challenger } b \in C.$$

We proved that maximal lotteries always exist.
This guarantee averages over the entire population. Today we ask for the same guarantee within every listed group.

Pooled Comparisons Can Hide Opposite Preferences

Two equally likely groups compare the same completions a and b:

population	probability a defeats b	probability b defeats a
group 1	$3/4$	$1/4$
group 2	$1/4$	$3/4$
pooled	$1/2$	$1/2$

Pooled Comparisons Can Hide Opposite Preferences

Two equally likely groups compare the same completions a and b:

population	probability a defeats b	probability b defeats a
group 1	$3/4$	$1/4$
group 2	$1/4$	$3/4$
pooled	$1/2$	$1/2$

In the pooled comparison, neither completion beats the other. Every lottery is therefore maximal for the pooled population.

Pooled Comparisons Can Hide Opposite Preferences

Two equally likely groups compare the same completions a and b:

population	probability a defeats b	probability b defeats a
group 1	$3/4$	$1/4$
group 2	$1/4$	$3/4$
pooled	$1/2$	$1/2$

In the pooled comparison, neither completion beats the other. Every lottery is therefore maximal for the pooled population. But always choosing a loses to b three quarters of the time in group 2. Pooled robustness has hidden a systematic group-level failure.

One Policy Must Be Allowed to Use Context

Recall: group 1 prefers a with probability $3/4$, while group 2 prefers b with probability $3/4$.
A lottery that chooses a with probability θ has win probabilities

$$\text{against } a \text{ in group 1: } \theta \cdot \frac{1}{2} + (1-\theta) \cdot \frac{1}{4},$$
$$\text{against } b \text{ in group 2: } \theta \cdot \frac{1}{4} + (1-\theta) \cdot \frac{1}{2}.$$

The first is at least $1/2$ only if $\theta = 1$; the second only if $\theta = 0$.

One Policy Must Be Allowed to Use Context

Recall: group 1 prefers a with probability $3/4$, while group 2 prefers b with probability $3/4$.
A lottery that chooses a with probability θ has win probabilities

$$\text{against } a \text{ in group 1: } \theta \cdot \frac{1}{2} + (1-\theta) \cdot \frac{1}{4},$$
$$\text{against } b \text{ in group 2: } \theta \cdot \frac{1}{4} + (1-\theta) \cdot \frac{1}{2}.$$

The first is at least $1/2$ only if $\theta = 1$; the second only if $\theta = 0$.
No single context-independent lottery is maximal within both groups. A contextual policy can instead choose a for group 1 and b for group 2.
Today, *one policy* means one rule that can use observable group membership when choosing its response.

Disjoint Groups Have a Simple Solution

Suppose each user belongs to exactly one group, with observable membership.

1. Learn a maximal lottery separately from each group's comparisons.
2. Serve each user the lottery for the group they belong to.

Within each group, the response then wins with probability at least $1/2$ against every fixed challenger.

Disjoint Groups Have a Simple Solution

Suppose each user belongs to exactly one group, with observable membership.

1. Learn a maximal lottery separately from each group's comparisons.
2. Serve each user the lottery for the group they belong to.

Within each group, the response then wins with probability at least $1/2$ against every fixed challenger.
If groups overlap, their separately learned lotteries need not agree.

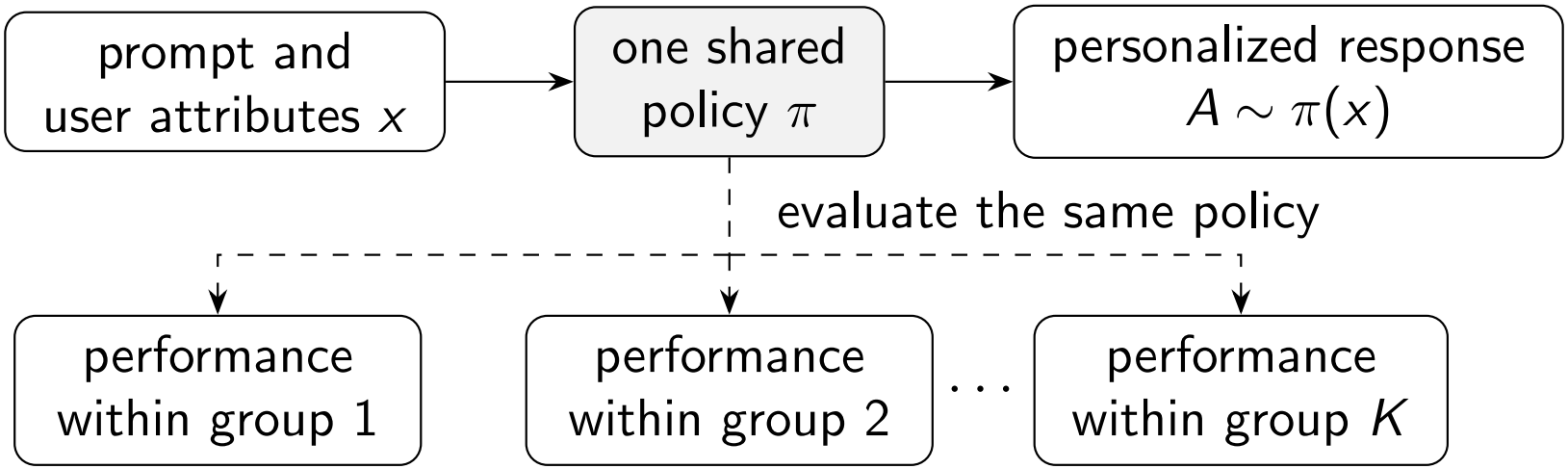
The obstacle

What should we serve someone who belongs to several groups, each with a different lottery?

Selecting one group's lottery does not automatically preserve the guarantees for the other groups.

One Policy Serves Many Potentially Overlapping Groups

We specify K groups of users, possibly overlapping. The same policy chooses responses for everyone, using the prompt and observable user attributes.



The goal

One shared policy retains, within every listed group, the robustness guarantee of a separately learned maximal lottery, up to learning error (even though implementing this directly is impossible because of intersections).

A user's interaction contributes to the evaluation of every group they belong to.

Contexts, Policies, and Group Membership

We keep the completion set C fixed and let the context X record observable user information.
An indicator $G_k(x)$ is 1 when context x belongs to group k, and 0 otherwise:

$$G_k : \mathcal{X} \to \{0, 1\}, \qquad p_k := \mathbb{P}(G_k(X) = 1) > 0.$$

A contextual policy assigns a lottery to each context:

$$\pi : \mathcal{X} \to \Delta(C).$$

Context-Dependent Comparison Margins

To measure how much a response wins or loses by at context x, use

$$M^{\star}_{ab}(x) := 2\mathbb{P}(a \succ b \mid X = x) - 1.$$

This is the win probability minus the loss probability, conditional on $X = x$. Reversing the pair reverses the sign, so $M^{\star}(x) = -M^{\star}(x)^{\top}$; self-comparisons give a zero diagonal.

Context-Dependent Comparison Margins

To measure how much a response wins or loses by at context x, use

$$M^{\star}_{ab}(x) := 2\mathbb{P}(a \succ b \mid X = x) - 1.$$

This is the win probability minus the loss probability, conditional on $X = x$. Reversing the pair reverses the sign, so $M^{\star}(x) = -M^{\star}(x)^{\top}$; self-comparisons give a zero diagonal.
For a policy lottery p and challenger lottery q, the expected margin is

$$p^{\top} M^{\star}(x) q = \mathbb{E}_{a\sim p,\, b\sim q}[M^{\star}_{ab}(x)].$$

Antisymmetry gives the *self-tie identity*

$$p^{\top} M^{\star}(x) p = 0.$$

Two independent responses from the same lottery have expected comparison margin zero.

Require Robustness Within Every Listed Group

Let e_b be the point mass on completion b.

Definition

A policy π is ε-*groupwise maximal* if, for every k, b,

$$\mathbb{E}\Big[\pi(X)^\top M^\star(X) e_b \mid G_k(X) = 1\Big] \geq -\varepsilon.$$

Its win probability against b within group k is at least $1/2 - \varepsilon/2$.

Require Robustness Within Every Listed Group

Let e_b be the point mass on completion b.

Definition

A policy π is ε-*groupwise maximal* if, for every k, b,

$$\mathbb{E}\Big[\pi(X)^\top M^\star(X) e_b \mid G_k(X) = 1\Big] \geq -\varepsilon.$$

Its win probability against b within group k is at least $1/2 - \varepsilon/2$.
Because a fixed lottery q is a mixture of the e_b's, the same bound holds against every fixed $q \in \Delta(C)$.
Each group gets its own choice of challenger. The challenger is fixed within that group; the policy $\pi(X)$ may adapt to the context.

Recall How Comparisons Relate to Welfare

Use Lecture 6's model. A sampled user has a latent utility vector $U \in [0,1]^C$ on a common cardinal scale, and

$$\mathbb{P}(a \succ b \mid X, U) = \sigma\big(\beta(U_a - U_b)\big), \qquad \sigma(z) = \frac{1}{1+e^{-z}}, \quad \beta > 0.$$

The policy observes X, not U. For $A \sim \pi(X)$,

$$U^\top \pi(X) = \mathbb{E}[U_A \mid X, U]$$

is that user's expected utility from its randomized response.

Recall How Comparisons Relate to Welfare

Use Lecture 6's model. A sampled user has a latent utility vector $U \in [0,1]^C$ on a common cardinal scale, and

$$\mathbb{P}(a \succ b \mid X, U) = \sigma\big(\beta(U_a - U_b)\big), \qquad \sigma(z) = \frac{1}{1+e^{-z}}, \quad \beta > 0.$$

The policy observes X, not U. For $A \sim \pi(X)$,

$$U^\top \pi(X) = \mathbb{E}[U_A \mid X, U]$$

is that user's expected utility from its randomized response. Define group welfare and its best fixed-lottery benchmark by

$$W_k(\pi) := \mathbb{E}[U^\top \pi(X) \mid G_k(X) = 1], \qquad \mathrm{OPT}_k := \max_{q \in \Delta(C)} W_k(q).$$

Here $W_k(q) = \mathbb{E}[U^\top q \mid G_k(X) = 1]$ for a fixed lottery q.

The Welfare Inequality Survives Conditioning

Lecture 6 proved a pointwise inequality. Writing

$$\Gamma_\beta := \frac{\beta}{4(\sigma(\beta) - 1/2)},$$

that inequality, in our margin normalization, is

$$2\sigma\big(\beta(U_a - U_b)\big) - 1 \le \frac{\beta}{2}\left(U_a - \frac{U_b}{\Gamma_\beta}\right).$$

The Welfare Inequality Survives Conditioning

Lecture 6 proved a pointwise inequality. Writing

$$\Gamma_\beta := \frac{\beta}{4(\sigma(\beta) - 1/2)},$$

that inequality, in our margin normalization, is

$$2\sigma\big(\beta(U_a - U_b)\big) - 1 \le \frac{\beta}{2}\left(U_a - \frac{U_b}{\Gamma_\beta}\right).$$

It holds for every user and pair of completions. Average over $a \sim \pi(X)$, $b \sim q$, and users and contexts *within group* k:

$$\mathbb{E}[\pi(X)^\top M^\star(X) q \mid G_k(X) = 1] \le \frac{\beta}{2}\left(W_k(\pi) - \frac{W_k(q)}{\Gamma_\beta}\right).$$

Conditioning on a group changes the average, not the pointwise inequality.

Groupwise Maximality Gives Groupwise Welfare

Theorem

Under the comparison model, an ε-groupwise maximal policy satisfies

$$\boxed{W_k(\pi) \geq \frac{\mathrm{OPT}_k}{\Gamma_\beta} - \frac{2\varepsilon}{\beta}} \qquad \textit{for every listed group } k.$$

Groupwise Maximality Gives Groupwise Welfare

Theorem

Under the comparison model, an ε-groupwise maximal policy satisfies

$$\boxed{W_k(\pi) \geq \frac{\mathrm{OPT}_k}{\Gamma_\beta} - \frac{2\varepsilon}{\beta}} \qquad \textit{for every listed group } k.$$

For any fixed challenger q, combine maximality with the preceding bound:

$$-\varepsilon \leq \mathbb{E}[\pi(X)^\top M^\star(X) q \mid G_k(X) = 1] \leq \frac{\beta}{2}\left(W_k(\pi) - \frac{W_k(q)}{\Gamma_\beta}\right).$$

Rearrange and choose q maximizing $W_k(q)$. □

Groupwise Maximality Gives Groupwise Welfare

Theorem

Under the comparison model, an ε-groupwise maximal policy satisfies

$$\boxed{W_k(\pi) \geq \frac{\mathrm{OPT}_k}{\Gamma_\beta} - \frac{2\varepsilon}{\beta}} \qquad \textit{for every listed group } k.$$

For any fixed challenger q, combine maximality with the preceding bound:

$$-\varepsilon \leq \mathbb{E}[\pi(X)^\top M^\star(X) q \mid G_k(X) = 1] \leq \frac{\beta}{2}\left(W_k(\pi) - \frac{W_k(q)}{\Gamma_\beta}\right).$$

Rearrange and choose q maximizing $W_k(q)$. □

Can we learn such a policy from pairwise labels without estimating a preference matrix at every context?

Recall Lecture 8: Response Satisfiability

In Lecture 8, learner actions $x \in \mathcal{X}$ and environment actions $y \in \mathcal{Y}$ produced vectors $g(x, y)$. Both action sets were finite. For mixed actions ρ and Q, write

$$g(\rho, Q) := \mathbb{E}_{x \sim \rho,\, y \sim Q}[g(x, y)].$$

Recall Lecture 8: Response Satisfiability

In Lecture 8, learner actions $x \in \mathcal{X}$ and environment actions $y \in \mathcal{Y}$ produced vectors $g(x, y)$. Both action sets were finite. For mixed actions ρ and Q, write

$$g(\rho, Q) := \mathbb{E}_{x \sim \rho,\, y \sim Q}[g(x, y)].$$

Let $\mathcal{K}$ be a closed convex target containing 0, as in Lecture 8. It is *response satisfiable* if

$$\boxed{\forall Q \in \Delta(\mathcal{Y}), \quad \exists \rho \in \Delta(\mathcal{X}) : \quad g(\rho, Q) \in \mathcal{K}.}$$

If we knew the environment's distribution Q, we could choose a response whose expected vector already lies in the target.

Recall Lecture 8: Response Satisfiability

In Lecture 8, learner actions $x \in \mathcal{X}$ and environment actions $y \in \mathcal{Y}$ produced vectors $g(x, y)$. Both action sets were finite. For mixed actions ρ and Q, write

$$g(\rho, Q) := \mathbb{E}_{x \sim \rho,\, y \sim Q}[g(x, y)].$$

Let $\mathcal{K}$ be a closed convex target containing 0, as in Lecture 8. It is *response satisfiable* if

$$\boxed{\forall Q \in \Delta(\mathcal{Y}), \quad \exists \rho \in \Delta(\mathcal{X}) : \quad g(\rho, Q) \in \mathcal{K}.}$$

If we knew the environment's distribution Q, we could choose a response whose expected vector already lies in the target.
But the learner must choose ρ before the environment commits to its action. Lecture 8 used projection and minimax to turn response satisfiability into a strategy that respects this order.

Recall Lecture 8: The Action Supplied by Minimax

Write $\bar{g}_t = t^{-1} \sum_{\tau=1}^{t} g(x_\tau, y_\tau)$, with $\bar{g}_0 = 0$. The current projection c and *outward direction* s are

$$c = \Pi_{\mathcal{K}}(\bar{g}_{t-1}), \qquad s = \bar{g}_{t-1} - c.$$

Recall Lecture 8: The Action Supplied by Minimax

Write $\bar{g}_t = t^{-1} \sum_{\tau=1}^{t} g(x_\tau, y_\tau)$, with $\bar{g}_0 = 0$. The current projection c and *outward direction* s are

$$c = \Pi_{\mathcal{K}}(\bar{g}_{t-1}), \qquad s = \bar{g}_{t-1} - c.$$

Projection puts $\mathcal{K}$ inside the halfspace $\{v : \langle s, v - c \rangle \leq 0\}$. Response satisfiability and minimax then give one action distribution ρ_t such that

$$\boxed{\mathbb{E}_{x \sim \rho_t}[\langle s, g(x, y) - c \rangle] \leq 0 \quad \text{for every } y.}$$

This is the *halfspace response*. We choose ρ_t, let the environment commit to y_t, and then draw $x_t \sim \rho_t$.

Recall Lecture 8: The Action Supplied by Minimax

Write $\bar{g}_t = t^{-1} \sum_{\tau=1}^{t} g(x_\tau, y_\tau)$, with $\bar{g}_0 = 0$. The current projection c and *outward direction* s are

$$c = \Pi_{\mathcal{K}}(\bar{g}_{t-1}), \qquad s = \bar{g}_{t-1} - c.$$

Projection puts $\mathcal{K}$ inside the halfspace $\{v : \langle s, v - c \rangle \leq 0\}$. Response satisfiability and minimax then give one action distribution ρ_t such that

$$\boxed{\mathbb{E}_{x \sim \rho_t}[\langle s, g(x, y) - c \rangle] \leq 0 \quad \text{for every } y.}$$

This is the *halfspace response*. We choose ρ_t, let the environment commit to y_t, and then draw $x_t \sim \rho_t$.

Blackwell's guarantee. If $\|g(x, y)\|_2 \leq G$, enforcing this condition every round gives

$$\mathbb{E}[\mathrm{dist}(\bar{g}_T, \mathcal{K})^2] \leq \frac{4G^2}{T}.$$

Plan: Groupwise Maximality as an Approachability Problem

For groupwise alignment, we will construct the halfspace response directly from the constraint weights using antisymmetry. We can then apply the Blackwell Approachability guarantee (having circumvented a minimax computation)

1. **Coordinates:** one violation per (group, challenger) pair, a vector in $\mathbb{R}^{Km}$.

Plan: Groupwise Maximality as an Approachability Problem

For groupwise alignment, we will construct the halfspace response directly from the constraint weights using antisymmetry. We can then apply the Blackwell Approachability guarantee (having circumvented a minimax computation)

1. **Coordinates:** one violation per (group, challenger) pair, a vector in $\mathbb{R}^{Km}$.
2. **Target:** the negative orthant, where every violation is ≤ 0. Its projection has a closed form.

Plan: Groupwise Maximality as an Approachability Problem

For groupwise alignment, we will construct the halfspace response directly from the constraint weights using antisymmetry. We can then apply the Blackwell Approachability guarantee (having circumvented a minimax computation)

1. **Coordinates:** one violation per (group, challenger) pair, a vector in $\mathbb{R}^{Km}$.
2. **Target:** the negative orthant, where every violation is ≤ 0. Its projection has a closed form.
3. **Response:** a policy whose weighted violation is ≤ 0 for every antisymmetric comparison matrix. Antisymmetry gives it in closed form.

Plan: Groupwise Maximality as an Approachability Problem

For groupwise alignment, we will construct the halfspace response directly from the constraint weights using antisymmetry. We can then apply the Blackwell Approachability guarantee (having circumvented a minimax computation)

1. **Coordinates:** one violation per (group, challenger) pair, a vector in $\mathbb{R}^{Km}$.
2. **Target:** the negative orthant, where every violation is ≤ 0. Its projection has a closed form.
3. **Response:** a policy whose weighted violation is ≤ 0 for every antisymmetric comparison matrix. Antisymmetry gives it in closed form.
4. **Data:** estimate the comparison matrix from one observed comparison per round, preserving antisymmetry.

Plan: Groupwise Maximality as an Approachability Problem

For groupwise alignment, we will construct the halfspace response directly from the constraint weights using antisymmetry. We can then apply the Blackwell Approachability guarantee (having circumvented a minimax computation)

1. **Coordinates:** one violation per (group, challenger) pair, a vector in $\mathbb{R}^{Km}$.
2. **Target:** the negative orthant, where every violation is ≤ 0. Its projection has a closed form.
3. **Response:** a policy whose weighted violation is ≤ 0 for every antisymmetric comparison matrix. Antisymmetry gives it in closed form.
4. **Data:** estimate the comparison matrix from one observed comparison per round, preserving antisymmetry.
5. **Guarantee:** From groupwise maximality on the sampled sequence to guarantees on the distribution.

Turn Group Failures into Vector Coordinates

For a context x, a lottery p, and an antisymmetric matrix M, define

$$r_{k,b}(x, p, M) := -G_k(x)p^\top Me_b.$$

A positive value means challenger b beats the policy on an active group constraint. Its population average is

$$\nu_{k,b}(\pi) := \mathbb{E}[r_{k,b}(X, \pi(X), M^\star(X))].$$

Turn Group Failures into Vector Coordinates

For a context x, a lottery p, and an antisymmetric matrix M, define

$$r_{k,b}(x, p, M) := -G_k(x)p^\top Me_b.$$

A positive value means challenger b beats the policy on an active group constraint. Its population average is

$$\nu_{k,b}(\pi) := \mathbb{E}[r_{k,b}(X, \pi(X), M^\star(X))].$$

Since G_k is an indicator,

$$\nu_{k,b}(\pi) = -p_k\mathbb{E}[\pi(X)^\top M^\star(X)e_b \mid G_k(X) = 1].$$

Exact groupwise maximality asks for $\nu(\pi)$ to lie in the *negative orthant*

$$\mathcal{K} := \{v \in \mathbb{R}^{Km} : v_{k,b} \le 0 \text{ for all } k, b\}.$$

Blackwell Focuses on the Positive Violations

Following the template, we need the projection onto the target and the outward direction. For the negative orthant both are simple. Let R be the running total of the violation vectors after $t \geq 1$ rounds. Define its positive and negative parts coordinatewise:

$$(R_+)_{k,b} := \max\{R_{k,b}, 0\}, \qquad (R_-)_{k,b} := \min\{R_{k,b}, 0\}.$$

The target $\mathcal{K}$ requires every coordinate to be nonpositive. Its closest point to R keeps negative coordinates and replaces positive coordinates by zero: it is R_-.

Blackwell Focuses on the Positive Violations

Following the template, we need the projection onto the target and the outward direction. For the negative orthant both are simple. Let R be the running total of the violation vectors after $t \geq 1$ rounds. Define its positive and negative parts coordinatewise:

$$(R_+)_{k,b} := \max\{R_{k,b}, 0\}, \qquad (R_-)_{k,b} := \min\{R_{k,b}, 0\}.$$

The target $\mathcal{K}$ requires every coordinate to be nonpositive. Its closest point to R keeps negative coordinates and replaces positive coordinates by zero: it is R_-.
For the current average R/t, the projection and outward direction are

$$c = \frac{R_-}{t}, \qquad s = \frac{R}{t} - c = \frac{R - R_-}{t} = \frac{R_+}{t}.$$

Only positive accumulated violations contribute. Dividing by $t > 0$ preserves all coordinate signs and scales the distance to $\mathcal{K}$ by $1/t$, so we may track the total R.

The Outward Direction Becomes Constraint Weights

Use the coordinates of R_+ as weights:

$$\lambda_{k,b} := \max\{R_{k,b}, 0\}.$$

This puts more weight on larger positive accumulated violations.

The Outward Direction Becomes Constraint Weights

Use the coordinates of R_+ as weights:

$$\lambda_{k,b} := \max\{R_{k,b}, 0\}.$$

This puts more weight on larger positive accumulated violations. Let v be the next violation vector's expectation given the past. A halfspace response makes $\langle s, v - c\rangle \leq 0$, with $s = R_+/t$ and $c = R_-/t$. Since each coordinate has at most one nonzero part, $\langle R_+, R_-\rangle = 0$. Therefore

$$\langle s, v - c\rangle = \frac{1}{t}\left\langle R_+, v - \frac{R_-}{t}\right\rangle = \frac{1}{t}\sum_{k,b}\lambda_{k,b}v_{k,b}.$$

The Outward Direction Becomes Constraint Weights

Use the coordinates of R_+ as weights:

$$\lambda_{k,b} := \max\{R_{k,b}, 0\}.$$

This puts more weight on larger positive accumulated violations. Let v be the next violation vector's expectation given the past. A halfspace response makes $\langle s, v - c\rangle \leq 0$, with $s = R_+/t$ and $c = R_-/t$. Since each coordinate has at most one nonzero part, $\langle R_+, R_-\rangle = 0$. Therefore

$$\langle s, v - c\rangle = \frac{1}{t}\left\langle R_+, v - \frac{R_-}{t}\right\rangle = \frac{1}{t}\sum_{k,b}\lambda_{k,b}v_{k,b}.$$

To make this nonpositive, we will choose a response p at each context x such that, for every antisymmetric M,

$$\sum_{k,b}\lambda_{k,b}r_{k,b}(x, p, M) \leq 0.$$

This makes the expected weighted violation nonpositive.

Active Groups Define a Challenger Distribution

Fix nonnegative weights $\lambda_{k,b}$. At context x, add the weights of the groups containing x:

$$s_b(x) := \sum_{k=1}^{K} \lambda_{k,b} G_k(x), \qquad Z(x) := \sum_b s_b(x).$$

Here $s_b(x)$ is the total weight placed on challenger b by the active groups.

Active Groups Define a Challenger Distribution

Fix nonnegative weights $\lambda_{k,b}$. At context x, add the weights of the groups containing x:

$$s_b(x) := \sum_{k=1}^{K} \lambda_{k,b} G_k(x), \qquad Z(x) := \sum_b s_b(x).$$

Here $s_b(x)$ is the total weight placed on challenger b by the active groups.
When $Z(x) > 0$, respond with the distribution:

$$\boxed{\pi_b(x) = \frac{s_b(x)}{Z(x)}.}$$

When $Z(x) = 0$, use the uniform lottery.

Self-Play Makes the Weighted Violation Zero

For $Z(x) > 0$, our choice satisfies $s(x) = Z(x)\pi(x)$. Therefore

$$\begin{aligned}\sum_{k,b} \lambda_{k,b} r_{k,b}(x, \pi(x), M) &= -\sum_{b} \left(\sum_{k} \lambda_{k,b} G_k(x) \right) \pi(x)^\top M e_b \\ &= -\pi(x)^\top M s(x) \\ &= -Z(x)\pi(x)^\top M \pi(x) \\ &= 0.\end{aligned}$$

The last equality is the self-tie identity for an antisymmetric matrix. If $Z(x) = 0$, all the weighted active constraints are zero already.

Self-Play Makes the Weighted Violation Zero

For $Z(x) > 0$, our choice satisfies $s(x) = Z(x)\pi(x)$. Therefore

$$\begin{aligned}\sum_{k,b} \lambda_{k,b} r_{k,b}(x, \pi(x), M) &= -\sum_{b} \left(\sum_{k} \lambda_{k,b} G_k(x) \right) \pi(x)^\top M e_b \\ &= -\pi(x)^\top M s(x) \\ &= -Z(x)\pi(x)^\top M \pi(x) \\ &= 0.\end{aligned}$$

The last equality is the self-tie identity for an antisymmetric matrix. If $Z(x) = 0$, all the weighted active constraints are zero already. The same response works for *every* antisymmetric matrix M. Thus we can meet the halfspace condition without knowing $M^\star(x)$ or solving a minimax problem.

The Data Are Ordinary Pairwise Comparisons

Training uses independent examples from the deployment population. On round t:

1. Draw a fresh context and user (X_t, U_t).

The Data Are Ordinary Pairwise Comparisons

Training uses independent examples from the deployment population. On round t:

1. Draw a fresh context and user (X_t, U_t).
2. Independently select one of the $n := \binom{m}{2}$ unordered pairs uniformly. Write it as $\{i, j\}$ with $i < j$.

The Data Are Ordinary Pairwise Comparisons

Training uses independent examples from the deployment population. On round t:

1. Draw a fresh context and user (X_t, U_t).
2. Independently select one of the $n := \binom{m}{2}$ unordered pairs uniformly. Write it as $\{i, j\}$ with $i < j$.
3. Observe a label $S_t \in \{-1, +1\}$: $+1$ means i wins, and -1 means j wins.

The Data Are Ordinary Pairwise Comparisons

Training uses independent examples from the deployment population. On round t:

1. Draw a fresh context and user (X_t, U_t).
2. Independently select one of the $n := \binom{m}{2}$ unordered pairs uniformly. Write it as $\{i, j\}$ with $i < j$.
3. Observe a label $S_t \in \{-1, +1\}$: $+1$ means i wins, and -1 means j wins.

Then

$$\mathbb{E}[S_t \mid X_t, \{i, j\}, \text{past}] = M^\star_{ij}(X_t).$$

The learner sees X_t, the pair, and its label, but not U_t or $M^\star(X_t)$.

One Comparison Gives an Antisymmetric Matrix Estimate

Recall that $n = \binom{m}{2}$ is the number of unordered completion pairs, so each pair is sampled with probability $1/n$. For the observed pair $i < j$, put its signed label into two matrix entries:

$$\widehat{M}_{t,ij} = nS_t, \qquad \widehat{M}_{t,ji} = -nS_t,$$

with every other entry zero. Equivalently,

$$\widehat{M}_t = nS_t(e_i e_j^\top - e_j e_i^\top).$$

Here $e_i e_j^\top$ has a one in entry (i, j) and zeros elsewhere.

One Comparison Gives an Antisymmetric Matrix Estimate

Recall that $n = \binom{m}{2}$ is the number of unordered completion pairs, so each pair is sampled with probability $1/n$. For the observed pair $i < j$, put its signed label into two matrix entries:

$$\widehat{M}_{t,ij} = nS_t, \qquad \widehat{M}_{t,ji} = -nS_t,$$

with every other entry zero. Equivalently,

$$\widehat{M}_t = nS_t(e_i e_j^\top - e_j e_i^\top).$$

Here $e_i e_j^\top$ has a one in entry (i, j) and zeros elsewhere. Multiplying by n compensates for observing each pair only with probability $1/n$. For any fixed $a < b$,

$$\mathbb{E}[\widehat{M}_{t,ab} \mid X_t, \text{past}] = \frac{1}{n} \cdot n\, M^\star_{ab}(X_t) = M^\star_{ab}(X_t).$$

The reverse entry is its negative and the diagonal is zero. Thus $\mathbb{E}[\widehat{M}_t \mid X_t, \text{past}] = M^\star(X_t)$.

The Weighted Violation Is Zero on Every Sample

Fix weights λ_t determined by past observations, and their contextual rule π_t. Evaluate the previously defined violation vector r using the sampled matrix estimate:

$$z_{t,k,b} := r_{k,b}\big(X_t, \pi_t(X_t), \widehat{M}_t\big) = -G_k(X_t)\pi_t(X_t)^\top \widehat{M}_t e_b.$$

Matrix unbiasedness gives

$$\mathbb{E}[z_{t,k,b} \mid X_t, \text{past}] = -G_k(X_t)\pi_t(X_t)^\top M^\star(X_t) e_b.$$

The Weighted Violation Is Zero on Every Sample

Fix weights λ_t determined by past observations, and their contextual rule π_t. Evaluate the previously defined violation vector r using the sampled matrix estimate:

$$z_{t,k,b} := r_{k,b}\big(X_t, \pi_t(X_t), \widehat{M}_t\big) = -G_k(X_t)\pi_t(X_t)^\top \widehat{M}_t e_b.$$

Matrix unbiasedness gives

$$\mathbb{E}[z_{t,k,b} \mid X_t, \text{past}] = -G_k(X_t)\pi_t(X_t)^\top M^\star(X_t) e_b.$$

More importantly, $\widehat{M}_t$ is antisymmetric on *every* sample, so the self-play identity applies with $M = \widehat{M}_t$ and $\lambda = \lambda_t$. For our group-weighted response,

$$\boxed{\lambda_t^\top z_t = 0 \quad \text{on every round.}}$$

No expectation is needed for this step: it holds sample by sample.

A Direct Multigroup Learning Algorithm

Initialize the cumulative violation vector $R_0 = 0$. On round t:

1. Set the constraint weights $\lambda_t = (R_{t-1})_+$.

A Direct Multigroup Learning Algorithm

Initialize the cumulative violation vector $R_0 = 0$. On round t:

1. Set the constraint weights $\lambda_t = (R_{t-1})_+$.
2. These weights define the contextual lottery:

$$\pi_{t,b}(x) = \frac{\sum_k \lambda_{t,k,b} G_k(x)}{\sum_{k,a} \lambda_{t,k,a} G_k(x)},$$

using the uniform lottery if the denominator is zero.

A Direct Multigroup Learning Algorithm

Initialize the cumulative violation vector $R_0 = 0$. On round t:

1. Set the constraint weights $\lambda_t = (R_{t-1})_+$.
2. These weights define the contextual lottery:

$$\pi_{t,b}(x) = \frac{\sum_k \lambda_{t,k,b} G_k(x)}{\sum_{k,a} \lambda_{t,k,a} G_k(x)},$$

   using the uniform lottery if the denominator is zero.
3. Observe a fresh context, uniformly sampled pair, and label. Construct $\widehat{M}_t$ and the sampled violations z_t.

A Direct Multigroup Learning Algorithm

Initialize the cumulative violation vector $R_0 = 0$. On round t:

1. Set the constraint weights $\lambda_t = (R_{t-1})_+$.
2. These weights define the contextual lottery:

$$\pi_{t,b}(x) = \frac{\sum_k \lambda_{t,k,b} G_k(x)}{\sum_{k,a} \lambda_{t,k,a} G_k(x)},$$

   using the uniform lottery if the denominator is zero.
3. Observe a fresh context, uniformly sampled pair, and label. Construct $\widehat{M}_t$ and the sampled violations z_t.
4. Update $R_t = R_{t-1} + z_t$.

A Direct Multigroup Learning Algorithm

Initialize the cumulative violation vector $R_0 = 0$. On round t:

1. Set the constraint weights $\lambda_t = (R_{t-1})_+$.
2. These weights define the contextual lottery:

$$\pi_{t,b}(x) = \frac{\sum_k \lambda_{t,k,b} G_k(x)}{\sum_{k,a} \lambda_{t,k,a} G_k(x)},$$

   using the uniform lottery if the denominator is zero.
3. Observe a fresh context, uniformly sampled pair, and label. Construct $\widehat{M}_t$ and the sampled violations z_t.
4. Update $R_t = R_{t-1} + z_t$.

Return the average contextual rule

$$\widehat{\pi}_T(x) := \frac{1}{T} \sum_{t=1}^{T} \pi_t(x).$$

For a new context x, choose a round uniformly from $1, \ldots, T$, then draw a completion from that round's lottery $\pi_t(x)$. This implements $\widehat{\pi}_T(x)$.

The Worst Group Violation Vanishes

Write $\varepsilon(\pi)$ for the smallest nonnegative error for which π is groupwise maximal. Equivalently,

$$\varepsilon(\pi) = \max_{k,b} \Big(-\mathbb{E}[\pi(X)^\top M^\star(X) e_b \mid G_k(X) = 1] \Big)_+ .$$

Here $(v)_+ = \max\{v, 0\}$. Set $B := n\sqrt{K}$, where $n = \binom{m}{2}$ and K is the number of groups. We will show $\|z_t\|_2 \le B$.

Theorem

If every group has probability $p_k \ge \gamma > 0$, the averaged policy satisfies

$$\boxed{\mathbb{E}_{\text{training}}[\varepsilon(\widehat{\pi}_T)] \le \frac{4B}{\gamma\sqrt{T}}.}$$

The Worst Group Violation Vanishes

Write $\varepsilon(\pi)$ for the smallest nonnegative error for which π is groupwise maximal. Equivalently,

$$\varepsilon(\pi) = \max_{k,b} \Big(-\mathbb{E}[\pi(X)^\top M^\star(X) e_b \mid G_k(X) = 1] \Big)_+ .$$

Here $(v)_+ = \max\{v, 0\}$. Set $B := n\sqrt{K}$, where $n = \binom{m}{2}$ and K is the number of groups. We will show $\|z_t\|_2 \le B$.

Theorem

If every group has probability $p_k \ge \gamma > 0$, the averaged policy satisfies

$$\boxed{\mathbb{E}_{\text{training}}[\varepsilon(\widehat{\pi}_T)] \le \frac{4B}{\gamma\sqrt{T}}.}$$

The maximum is *inside* the expectation: we control the worst group violation of the learned policy.
The proof has two steps: control violations along the training run, then transfer that control to the output policy on fresh users.

Each Sample Gives a Bounded Violation Vector

Lecture 8's bound needs $\|z_t\|_2 \le B = n\sqrt{K}$. Write $p_t := \pi_t(X_t)$ for the lottery at the observed context. Only columns i, j of $\widehat{M}_t$ are nonzero, so

$$z_{t,k,i} = nS_t G_k(X_t) p_{t,j}, \qquad z_{t,k,j} = -nS_t G_k(X_t) p_{t,i},$$

and $z_{t,k,b} = 0$ for $b \notin \{i, j\}$.

Each Sample Gives a Bounded Violation Vector

Lecture 8's bound needs $\|z_t\|_2 \le B = n\sqrt{K}$. Write $p_t := \pi_t(X_t)$ for the lottery at the observed context. Only columns i, j of $\widehat{M}_t$ are nonzero, so

$$z_{t,k,i} = nS_t G_k(X_t) p_{t,j}, \qquad z_{t,k,j} = -nS_t G_k(X_t) p_{t,i},$$

and $z_{t,k,b} = 0$ for $b \notin \{i, j\}$.
Since $S_t^2 = 1$ and each $G_k(X_t)$ is either zero or one,

$$\begin{aligned} \|z_t\|_2^2 &= n^2 \sum_{k=1}^{K} G_k(X_t)^2 (p_{t,i}^2 + p_{t,j}^2) \\ &\le Kn^2 (p_{t,i} + p_{t,j})^2 \\ &\le Kn^2 = B^2. \end{aligned}$$

The last inequality uses $p_{t,i} + p_{t,j} \le 1$.

Blackwell Controls the Sampled Violations

We have verified the two conditions used in Lecture 8's distance proof: the halfspace response, from $\lambda_t^\top z_t = 0$, and the bound $\|z_t\|_2 \le B$. For $\bar{z}_T := T^{-1}\sum_t z_t$ and $\mathcal{K} = \mathbb{R}^{Km}_-$, that proof gives

$$\mathbb{E}[\mathrm{dist}(\bar{z}_T, \mathcal{K})^2] \le \frac{4B^2}{T}.$$

Blackwell Controls the Sampled Violations

We have verified the two conditions used in Lecture 8's distance proof: the halfspace response, from $\lambda_t^\top z_t = 0$, and the bound $\|z_t\|_2 \le B$. For $\bar{z}_T := T^{-1}\sum_t z_t$ and $\mathcal{K} = \mathbb{R}^{Km}_-$, that proof gives

$$\mathbb{E}[\mathrm{dist}(\bar{z}_T, \mathcal{K})^2] \le \frac{4B^2}{T}.$$

The distance to the negative orthant is the norm of the positive part. Thus Cauchy–Schwarz gives

$$\mathbb{E}\|(\bar{z}_T)_+\|_2 \le \sqrt{\mathbb{E}\|(\bar{z}_T)_+\|_2^2} \le \boxed{\frac{2B}{\sqrt{T}}}.$$

Blackwell Controls the Sampled Violations

We have verified the two conditions used in Lecture 8's distance proof: the halfspace response, from $\lambda_t^\top z_t = 0$, and the bound $\|z_t\|_2 \le B$. For $\bar{z}_T := T^{-1}\sum_t z_t$ and $\mathcal{K} = \mathbb{R}_-^{Km}$, that proof gives

$$\mathbb{E}[\mathrm{dist}(\bar{z}_T, \mathcal{K})^2] \le \frac{4B^2}{T}.$$

The distance to the negative orthant is the norm of the positive part. Thus Cauchy–Schwarz gives

$$\mathbb{E}\|(\bar{z}_T)_+\|_2 \le \sqrt{\mathbb{E}\|(\bar{z}_T)_+\|_2^2} \le \boxed{\frac{2B}{\sqrt{T}}}.$$

This controls the changing policies π_t, each evaluated on its own sample. We still need a guarantee for the output policy $\widehat{\pi}_T$ on fresh users.

Each Fresh Sample Estimates Population Violations

Before drawing round t's example, freeze the past observations. They fix the policy π_t. Recall its *population violation* for group k and challenger b:

$$\nu_{k,b}(\pi_t) = \mathbb{E}_X[-G_k(X)\pi_t(X)^\top M^\star(X)e_b].$$

Each Fresh Sample Estimates Population Violations

Before drawing round t's example, freeze the past observations. They fix the policy π_t. Recall its *population violation* for group k and challenger b:

$$\nu_{k,b}(\pi_t) = \mathbb{E}_X[-G_k(X)\pi_t(X)^\top M^\star(X)e_b].$$

On a fresh training example, the *observed violation* is

$$z_{t,k,b} = -G_k(X_t)\pi_t(X_t)^\top \widehat{M}_t e_b.$$

Each Fresh Sample Estimates Population Violations

Before drawing round t's example, freeze the past observations. They fix the policy π_t. Recall its *population violation* for group k and challenger b:

$$\nu_{k,b}(\pi_t) = \mathbb{E}_X[-G_k(X)\pi_t(X)^\top M^\star(X)e_b].$$

On a fresh training example, the *observed violation* is

$$z_{t,k,b} = -G_k(X_t)\pi_t(X_t)^\top \widehat{M}_t e_b.$$

Averaging over the sampled pair and label replaces $\widehat{M}_t$ by $M^\star(X_t)$. Then averaging over the fresh context gives

$$\boxed{\mathbb{E}[z_{t,k,b} \mid \text{past}] = \nu_{k,b}(\pi_t).}$$

For brevity, write $\mu_t := \nu(\pi_t)$. We now bound the accumulated difference between the sampled violations z_t and their population means μ_t.

Sampling Errors Are Bounded and Have Mean Zero

Let $D_t := z_t - \mu_t$ be the difference between the sampled and population violations. Since $\mu_t = \mathbb{E}[z_t \mid \text{past}]$,

$$\mathbb{E}[D_t \mid \text{past}] = \mathbb{E}[z_t \mid \text{past}] - \mu_t = 0.$$

Sampling Errors Are Bounded and Have Mean Zero

Let $D_t := z_t - \mu_t$ be the difference between the sampled and population violations. Since $\mu_t = \mathbb{E}[z_t \mid \text{past}]$,

$$\mathbb{E}[D_t \mid \text{past}] = \mathbb{E}[z_t \mid \text{past}] - \mu_t = 0.$$

A mean of vectors of norm at most B also has norm at most B:

$$\|\mu_t\|_2 = \|\mathbb{E}[z_t \mid \text{past}]\|_2 \leq \mathbb{E}[\|z_t\|_2 \mid \text{past}] \leq B.$$

The triangle inequality therefore gives

$$\|D_t\|_2 \leq \|z_t\|_2 + \|\mu_t\|_2 \leq 2B.$$

These two facts suffice to show that the average sampling error becomes small.

Sampling Noise Averages Out

Adaptive policies can make the errors D_t dependent. However, $\mathbb{E}[D_t \mid \text{past}] = 0$, and for $s < t$ the past determines D_s. Therefore

$$\mathbb{E}[D_s^\top D_t] = \mathbb{E}\Big[D_s^\top \mathbb{E}[D_t \mid \text{past}]\Big] = 0.$$

Sampling Noise Averages Out

Adaptive policies can make the errors D_t dependent. However, $\mathbb{E}[D_t \mid \text{past}] = 0$, and for $s < t$ the past determines D_s. Therefore

$$\mathbb{E}[D_s^\top D_t] = \mathbb{E}\Big[D_s^\top \mathbb{E}[D_t \mid \text{past}]\Big] = 0.$$

Consequently,

$$\begin{aligned}\mathbb{E}\left\|\sum_t D_t\right\|_2^2 &= \sum_t \mathbb{E}\|D_t\|_2^2 + 2\sum_{s<t} \mathbb{E}[D_s^\top D_t] \\ &= \sum_t \mathbb{E}\|D_t\|_2^2 \le 4TB^2.\end{aligned}$$

Sampling Noise Averages Out

Adaptive policies can make the errors D_t dependent. However, $\mathbb{E}[D_t \mid \text{past}] = 0$, and for $s < t$ the past determines D_s. Therefore

$$\mathbb{E}[D_s^\top D_t] = \mathbb{E}\Big[D_s^\top \mathbb{E}[D_t \mid \text{past}]\Big] = 0.$$

Consequently,

$$\begin{aligned}\mathbb{E}\left\|\sum_t D_t\right\|_2^2 &= \sum_t \mathbb{E}\|D_t\|_2^2 + 2\sum_{s<t} \mathbb{E}[D_s^\top D_t] \\ &= \sum_t \mathbb{E}\|D_t\|_2^2 \le 4TB^2.\end{aligned}$$

Write $\bar{\mu}_T := T^{-1}\sum_t \mu_t$. Cauchy–Schwarz gives

$$\mathbb{E}\|\bar{z}_T - \bar{\mu}_T\|_2 = \frac{1}{T}\mathbb{E}\left\|\sum_t D_t\right\|_2 \le \frac{1}{T}\sqrt{\mathbb{E}\left\|\sum_t D_t\right\|_2^2} \le \frac{2B}{\sqrt{T}}.$$

The Population Guarantee Holds Within Every Group

Linearity in the policy gives

$$\nu(\widehat{\pi}_T) = \frac{1}{T}\sum_t \nu(\pi_t) = \bar{\mu}_T.$$

The Population Guarantee Holds Within Every Group

Linearity in the policy gives

$$\nu(\widehat{\pi}_T) = \frac{1}{T}\sum_t \nu(\pi_t) = \bar{\mu}_T.$$

Conditioning on group k divides its violation by $p_k = \mathbb{P}(G_k(X) = 1) \geq \gamma$. For every k, b,

$$\begin{aligned}\frac{(\bar{\mu}_{T,k,b})_+}{p_k} &\leq \frac{1}{\gamma}\left((\bar{z}_{T,k,b})_+ + |\bar{\mu}_{T,k,b} - \bar{z}_{T,k,b}|\right) \\ &\leq \frac{1}{\gamma}\left(\|(\bar{z}_T)_+\|_2 + \|\bar{\mu}_T - \bar{z}_T\|_2\right).\end{aligned}$$

The last step bounds each coordinate by its vector's norm.

The Population Guarantee Holds Within Every Group

Linearity in the policy gives

$$\nu(\widehat{\pi}_T) = \frac{1}{T}\sum_t \nu(\pi_t) = \bar{\mu}_T.$$

Conditioning on group k divides its violation by $p_k = \mathbb{P}(G_k(X) = 1) \geq \gamma$. For every k, b,

$$\begin{aligned}\frac{(\bar{\mu}_{T,k,b})_+}{p_k} &\leq \frac{1}{\gamma}\left((\bar{z}_{T,k,b})_+ + |\bar{\mu}_{T,k,b} - \bar{z}_{T,k,b}|\right) \\ &\leq \frac{1}{\gamma}\left(\|(\bar{z}_T)_+\|_2 + \|\bar{\mu}_T - \bar{z}_T\|_2\right).\end{aligned}$$

The last step bounds each coordinate by its vector's norm. Maximizing the left side over k, b gives $\varepsilon(\widehat{\pi}_T)$. Taking expectations and applying the two bounds gives

$$\mathbb{E}_{\text{training}}[\varepsilon(\widehat{\pi}_T)] \leq \frac{1}{\gamma}\left(\frac{2B}{\sqrt{T}} + \frac{2B}{\sqrt{T}}\right) = \boxed{\frac{4B}{\gamma\sqrt{T}}}.$$

One Learned Policy Has Welfare Guarantees for All Groups

For every realized training run, our earlier welfare theorem gives

$$W_k(\widehat{\pi}_T) \geq \frac{\mathrm{OPT}_k}{\Gamma_\beta} - \frac{2\varepsilon(\widehat{\pi}_T)}{\beta} \qquad \text{for every } k.$$

The *same* error controls every listed group, and its expected value vanishes as training grows.

One Learned Policy Has Welfare Guarantees for All Groups

For every realized training run, our earlier welfare theorem gives

$$W_k(\widehat{\pi}_T) \geq \frac{\mathrm{OPT}_k}{\Gamma_\beta} - \frac{2\varepsilon(\widehat{\pi}_T)}{\beta} \qquad \text{for every } k.$$

The *same* error controls every listed group, and its expected value vanishes as training grows.
One group-aware policy has been learned from pairwise labels, without observing utilities or fitting a separate preference model on every intersection.

Summary

- A pooled alignment guarantee can hide systematic failures within groups. We can instead require robustness separately for each group.

Summary

- A pooled alignment guarantee can hide systematic failures within groups. We can instead require robustness separately for each group.
- One contextual policy can meet overlapping group requirements: group-weighted self-play supplies the Blackwell response.

Summary

- A pooled alignment guarantee can hide systematic failures within groups. We can instead require robustness separately for each group.
- One contextual policy can meet overlapping group requirements: group-weighted self-play supplies the Blackwell response.
- Ordinary pairwise comparisons suffice to learn the policy, without observing utilities or estimating context-specific preference matrices.

Summary

- A pooled alignment guarantee can hide systematic failures within groups. We can instead require robustness separately for each group.
- One contextual policy can meet overlapping group requirements: group-weighted self-play supplies the Blackwell response.
- Ordinary pairwise comparisons suffice to learn the policy, without observing utilities or estimating context-specific preference matrices.
- The resulting groupwise robustness carries a welfare guarantee for every listed group under the comparison model.

Thanks!

See you next class!

References

- Jacob Brodkey, Aaron Roth, and Jesus Tamez Villarreal. *Multigroup Optimal Distortion of AI Alignment*. Unpublished manuscript, 2026.
- Paul Gölz, Nika Haghtalab, and Kunhe Yang. "Distortion of AI Alignment: Does Preference Optimization Optimize for Preferences?" arXiv:2505.23749, 2025.
- David Blackwell. "An Analog of the Minimax Theorem for Vector Payoffs." *Pacific Journal of Mathematics*, 1956.
- Peter C. Fishburn. "Probabilistic Social Choice Based on Simple Voting Comparisons." *Review of Economic Studies*, 1984.