Maximal Lotteries: Minimax and Self-Play

Aaron Roth

University of Pennsylvania

Fall 2026

Overview

- In a strict majority cycle, every deterministic output is defeated head-to-head by another completion.

Overview

- In a strict majority cycle, every deterministic output is defeated head-to-head by another completion.
- Randomization circumvents this: there always exists a lottery that defeats or ties every possible challenger.

Overview

- In a strict majority cycle, every deterministic output is defeated head-to-head by another completion.
- Randomization circumvents this: there always exists a lottery that defeats or ties every possible challenger.
- No-regret learning proves the minimax theorem and gives a way to compute a maximal lottery.

Overview

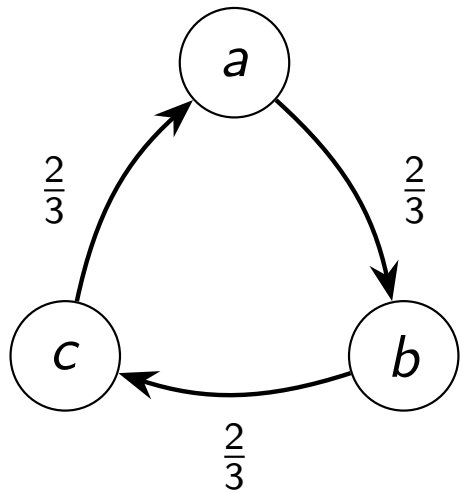
- In a strict majority cycle, every deterministic output is defeated head-to-head by another completion.
- Randomization circumvents this: there always exists a lottery that defeats or ties every possible challenger.
- No-regret learning proves the minimax theorem and gives a way to compute a maximal lottery.
- Antisymmetry lets us simplify computation by having one copy of a no regret learner play against itself.

Every Single Completion Can Be Beaten

The population contains three equally common preference types:

$$a \succ_1 b \succ_1 c, \qquad b \succ_2 c \succ_2 a, \qquad c \succ_3 a \succ_3 b.$$

Consequently, $P_{ab} = P_{bc} = P_{ca} = \frac{2}{3}$:

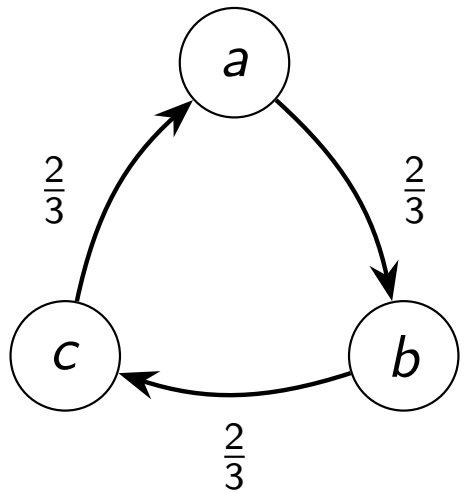


Every Single Completion Can Be Beaten

The population contains three equally common preference types:

$$a \succ_1 b \succ_1 c, \qquad b \succ_2 c \succ_2 a, \qquad c \succ_3 a \succ_3 b.$$

Consequently, $P_{ab} = P_{bc} = P_{ca} = \frac{2}{3}$:



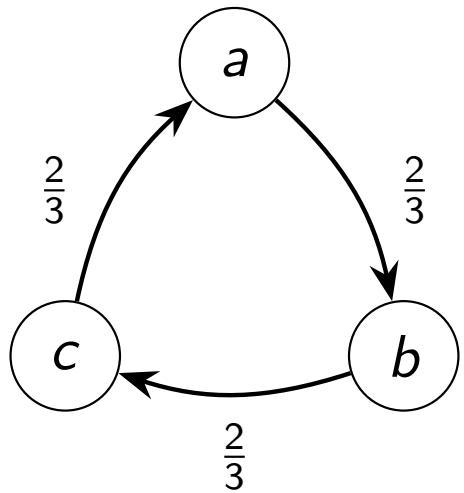
Each completion is defeated with probability 2/3 by the one behind it in the cycle: select a and challenger c wins; select b and a wins; select c and b wins.

Every Single Completion Can Be Beaten

The population contains three equally common preference types:

$$a \succ_1 b \succ_1 c, \qquad b \succ_2 c \succ_2 a, \qquad c \succ_3 a \succ_3 b.$$

Consequently, $P_{ab} = P_{bc} = P_{ca} = \frac{2}{3}$:



Each completion is defeated with probability 2/3 by the one behind it in the cycle: select a and challenger c wins; select b and a wins; select c and b wins.

The obstacle is not the selection rule: this profile contains no single completion that is robust to head-to-head challenge.

Randomization Can Balance the Cycle

Convention: a completion ties itself, so we set $P_{aa} = \frac{1}{2}$ for every a.

Randomization Can Balance the Cycle

Convention: a completion ties itself, so we set $P_{aa} = \frac{1}{2}$ for every a. Consider the uniform lottery

$$p = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right).$$

Against pure challenger a, its probability of winning is

$$\frac{1}{3}\left(P_{aa} + P_{ba} + P_{ca}\right) = \frac{1}{3}\left(\frac{1}{2} + \frac{1}{3} + \frac{2}{3}\right) = \frac{1}{2}.$$

Randomization Can Balance the Cycle

Convention: a completion ties itself, so we set $P_{aa} = \frac{1}{2}$ for every a. Consider the uniform lottery

$$p = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right).$$

Against pure challenger a, its probability of winning is

$$\frac{1}{3}\left(P_{aa} + P_{ba} + P_{ca}\right) = \frac{1}{3}\left(\frac{1}{2} + \frac{1}{3} + \frac{2}{3}\right) = \frac{1}{2}.$$

By symmetry, p also ties challengers b and c. Win probability is linear in the challenger's distribution, so p ties every randomized challenger.

Randomization Can Balance the Cycle

Convention: a completion ties itself, so we set $P_{aa} = \frac{1}{2}$ for every a. Consider the uniform lottery

$$p = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right).$$

Against pure challenger a, its probability of winning is

$$\frac{1}{3}\left(P_{aa} + P_{ba} + P_{ca}\right) = \frac{1}{3}\left(\frac{1}{2} + \frac{1}{3} + \frac{2}{3}\right) = \frac{1}{2}.$$

By symmetry, p also ties challengers b and c. Win probability is linear in the challenger's distribution, so p ties every randomized challenger.

Does every finite comparison problem admit some robust lottery?

Pairwise Margins Put a Tie at Zero

Score a comparison as $+1$ when i wins and -1 when j wins. Its expected score is

$$M_{ij} := (+1)P_{ij} + (-1)P_{ji} = P_{ij} - P_{ji} = 2P_{ij} - 1.$$

Pairwise Margins Put a Tie at Zero

Score a comparison as $+1$ when i wins and -1 when j wins. Its expected score is

$$M_{ij} := (+1)P_{ij} + (-1)P_{ji} = P_{ij} - P_{ji} = 2P_{ij} - 1.$$

Thus $M_{ij} > 0$ means that i wins more often than it loses, while $M_{ij} = 0$ means a 50–50 tie.

Pairwise Margins Put a Tie at Zero

Score a comparison as $+1$ when i wins and -1 when j wins. Its expected score is

$$M_{ij} := (+1)P_{ij} + (-1)P_{ji} = P_{ij} - P_{ji} = 2P_{ij} - 1.$$

Thus $M_{ij} > 0$ means that i wins more often than it loses, while $M_{ij} = 0$ means a 50–50 tie.
Reversing the two completions reverses the margin:

$$M_{ji} = -M_{ij}, \qquad M = -M^{\top}.$$

In particular, every completion has zero margin against itself: $M_{ii} = 0$.

General Policies Can Be Compared Pairwise

Let $p, q \in \Delta(C)$ be policies. Draw $i \sim p$ and $j \sim q$. Their expected win-minus-loss margin is

$$\mathbb{E}_{i \sim p,\, j \sim q}[M_{ij}] = \sum_{i,j \in C} p_i q_j M_{ij} = p^\top M q.$$

General Policies Can Be Compared Pairwise

Let $p, q \in \Delta(C)$ be policies. Draw $i \sim p$ and $j \sim q$. Their expected win-minus-loss margin is

$$\mathbb{E}_{i \sim p,\, j \sim q}[M_{ij}] = \sum_{i,j \in C} p_i q_j M_{ij} = p^\top M q.$$

Since $P_{ij} = (1 + M_{ij})/2$, the probability that p wins is

$$\sum_{i,j \in C} p_i q_j P_{ij} = \frac{1}{2} + \frac{1}{2} p^\top M q.$$

Thus p defeats or ties q exactly when $p^\top M q \geq 0$.

A Robust Policy is Evaluated Against the Strongest Challenger

Call the policy chooser *Max* and the challenger *Min*.

A Robust Policy is Evaluated Against the Strongest Challenger

Call the policy chooser *Max* and the challenger *Min*.
If Max commits to p, Min chooses the most damaging challenger q:

$$\min_{q \in \Delta(C)} p^\top M q.$$

Max therefore seeks the policy with the best worst-case margin:

$$\max_{p \in \Delta(C)} \min_{q \in \Delta(C)} p^\top M q.$$

A Robust Policy is Evaluated Against the Strongest Challenger

Call the policy chooser *Max* and the challenger *Min*.
If Max commits to p, Min chooses the most damaging challenger q:

$$\min_{q \in \Delta(C)} p^\top M q.$$

Max therefore seeks the policy with the best worst-case margin:

$$\max_{p \in \Delta(C)} \min_{q \in \Delta(C)} p^\top M q.$$

The opponent distribution is not fixed in advance; it responds strategically to the proposed policy.

Zero Is the Best Guarantee We Could Hope For

A lottery always ties itself:

$$p^\top M p = 0 \qquad \text{for every } p \in \Delta(C).$$

Zero Is the Best Guarantee We Could Hope For

A lottery always ties itself:

$$p^\top M p = 0 \qquad \text{for every } p \in \Delta(C).$$

Whatever policy p Max chooses, Min can copy it. Therefore

$$\min_q p^\top M q \leq p^\top M p = 0.$$

Zero Is the Best Guarantee We Could Hope For

A lottery always ties itself:

$$p^\top M p = 0 \qquad \text{for every } p \in \Delta(C).$$

Whatever policy p Max chooses, Min can copy it. Therefore

$$\min_q p^\top M q \le p^\top M p = 0.$$

Max can never guarantee a strictly positive margin.

But does some policy guarantee margin zero against *every* challenger?

The Robustness Question Is a Zero-Sum Game

Let $G \in \mathbb{R}^{d_{\max} \times d_{\min}}$ be a payoff matrix.

The Robustness Question Is a Zero-Sum Game

Let $G \in \mathbb{R}^{d_{\max} \times d_{\min}}$ be a payoff matrix.

- Max chooses $p \in \Delta_{d_{\max}}$ and wants the payoff large.
- Min chooses $q \in \Delta_{d_{\min}}$ and wants the payoff small.

The Robustness Question Is a Zero-Sum Game

Let $G \in \mathbb{R}^{d_{\max} \times d_{\min}}$ be a payoff matrix.

- Max chooses $p \in \Delta_{d_{\max}}$ and wants the payoff large.
- Min chooses $q \in \Delta_{d_{\min}}$ and wants the payoff small.

Their expected payoff is

$$u(p, q) = p^\top G q.$$

The Robustness Question Is a Zero-Sum Game

Let $G \in \mathbb{R}^{d_{\max} \times d_{\min}}$ be a payoff matrix.

- Max chooses $p \in \Delta_{d_{\max}}$ and wants the payoff large.
- Min chooses $q \in \Delta_{d_{\min}}$ and wants the payoff small.

Their expected payoff is

$$u(p, q) = p^\top G q.$$

For the preference game, $G = M$ and $d_{\max} = d_{\min} = |C|$, so $\Delta_{d_{\max}}$ and $\Delta_{d_{\min}}$ are both $\Delta(C)$. We develop the theory for general G and then return to M.

Commitment Order Defines Two Values

If Max commits first, Max guarantees the *maximin value*

$$v_{\text{maximin}} := \max_{p} \min_{q} p^{\top} G q.$$

Commitment Order Defines Two Values

If Max commits first, Max guarantees the *maximin value*

$$v_{\text{maximin}} := \max_{p} \min_{q} p^\top G q.$$

If Min commits first, Min can hold Max to the *minimax value*

$$v_{\text{minimax}} := \min_{q} \max_{p} p^\top G q.$$

Commitment Order Defines Two Values

If Max commits first, Max guarantees the *maximin value*

$$v_{\text{maximin}} := \max_{p} \min_{q} p^\top G q.$$

If Min commits first, Min can hold Max to the *minimax value*

$$v_{\text{minimax}} := \min_{q} \max_{p} p^\top G q.$$

The responding player sees the commitment before choosing a best response. Should committing first be costly?

Committing First Cannot Be an Advantage

For every fixed pair p, q,

$$\min_{q'} p^\top G q' \le p^\top G q \le \max_{p'} (p')^\top G q.$$

Committing First Cannot Be an Advantage

For every fixed pair p, q,

$$\min_{q'} p^\top G q' \le p^\top G q \le \max_{p'} (p')^\top G q.$$

Maximize the left side over p, then minimize the right side over q:

$$\max_{p} \min_{q} p^\top G q \le \min_{q} \max_{p} p^\top G q.$$

Committing First Cannot Be an Advantage

For every fixed pair p, q,

$$\min_{q'} p^\top G q' \le p^\top G q \le \max_{p'} (p')^\top G q.$$

Maximize the left side over p, then minimize the right side over q:

$$\max_{p} \min_{q} p^\top G q \le \min_{q} \max_{p} p^\top G q.$$

Thus

$$v_{\text{maximin}} \le v_{\text{minimax}}.$$

This is *weak minimax*. The question is whether the gap can be positive.

With Randomization, Commitment Order Does Not Matter

Theorem (von Neumann's minimax theorem)

For every finite payoff matrix $G \in \mathbb{R}^{d_{\max} \times d_{\min}}$,

$$\max_{p \in \Delta_{d_{\max}}} \min_{q \in \Delta_{d_{\min}}} p^\top G q = \min_{q \in \Delta_{d_{\min}}} \max_{p \in \Delta_{d_{\max}}} p^\top G q.$$

With Randomization, Commitment Order Does Not Matter

Theorem (von Neumann's minimax theorem)

For every finite payoff matrix $G \in \mathbb{R}^{d_{\max} \times d_{\min}}$,

$$\max_{p \in \Delta_{d_{\max}}} \min_{q \in \Delta_{d_{\min}}} p^\top G q = \min_{q \in \Delta_{d_{\min}}} \max_{p \in \Delta_{d_{\max}}} p^\top G q.$$

For the proof, divide every payoff by a positive constant large enough that $|G_{ij}| \leq 1$. This changes neither best responses nor whether the two values are equal.

With Randomization, Commitment Order Does Not Matter

Theorem (von Neumann's minimax theorem)

For every finite payoff matrix $G \in \mathbb{R}^{d_{\max} \times d_{\min}}$,

$$\max_{p \in \Delta_{d_{\max}}} \min_{q \in \Delta_{d_{\min}}} p^\top G q = \min_{q \in \Delta_{d_{\min}}} \max_{p \in \Delta_{d_{\max}}} p^\top G q.$$

For the proof, divide every payoff by a positive constant large enough that $|G_{ij}| \leq 1$. This changes neither best responses nor whether the two values are equal. Thus, once both players may randomize, neither loses from committing first.

With Randomization, Commitment Order Does Not Matter

Theorem (von Neumann's minimax theorem)

For every finite payoff matrix $G \in \mathbb{R}^{d_{\max} \times d_{\min}}$,

$$\max_{p \in \Delta_{d_{\max}}} \min_{q \in \Delta_{d_{\min}}} p^\top G q = \min_{q \in \Delta_{d_{\min}}} \max_{p \in \Delta_{d_{\max}}} p^\top G q.$$

For the proof, divide every payoff by a positive constant large enough that $|G_{ij}| \leq 1$. This changes neither best responses nor whether the two values are equal. Thus, once both players may randomize, neither loses from committing first.

We will prove this by imagining repeated play: Min learns from experience, while Max best-responds to Min's current strategy.

Proof Roadmap

1. Define *Hedge* and prove that its average regret vanishes against every adaptive loss sequence.

Proof Roadmap

1. Define *Hedge* and prove that its average regret vanishes against every adaptive loss sequence.
2. Let Min run Hedge while Max best-responds to Min's current strategy.

Proof Roadmap

1. Define *Hedge* and prove that its average regret vanishes against every adaptive loss sequence.
2. Let Min run Hedge while Max best-responds to Min's current strategy.
3. Bound their average payoff below by v_{minimax} and above by $v_{\text{maximin}} + \text{Reg}_T / T$. Vanishing regret closes the commitment gap.

The same repeated play will also *compute* the robust lottery.

An Online Learner Faces an Adversary

In rounds $t = 1, \ldots, T$:

1. the learner chooses a distribution $z_t \in \Delta_d$ over d actions;

An Online Learner Faces an Adversary

In rounds $t = 1, \ldots, T$:

1. the learner chooses a distribution $z_t \in \Delta_d$ over d actions;
2. an adversary chooses a loss vector $\ell_t \in [-1, 1]^d$, possibly after seeing z_t;

An Online Learner Faces an Adversary

In rounds $t = 1, \ldots, T$:

1. the learner chooses a distribution $z_t \in \Delta_d$ over d actions;
2. an adversary chooses a loss vector $\ell_t \in [-1, 1]^d$, possibly after seeing z_t;
3. the learner incurs expected loss $z_t^\top \ell_t$ and observes all of ℓ_t.

An Online Learner Faces an Adversary

In rounds $t = 1, \ldots, T$:

1. the learner chooses a distribution $z_t \in \Delta_d$ over d actions;
2. an adversary chooses a loss vector $\ell_t \in [-1, 1]^d$, possibly after seeing z_t;
3. the learner incurs expected loss $z_t^\top \ell_t$ and observes all of ℓ_t.

The adversary is unrestricted: the loss sequence may depend on everything the learner has done. What guarantee could possibly survive that?

External Regret Compares Against the Best Fixed Action

The learner's *external regret* is

$$\mathrm{Reg}_T := \sum_{t=1}^{T} z_t^\top \ell_t - \min_{z \in \Delta_d} \sum_{t=1}^{T} z^\top \ell_t.$$

External Regret Compares Against the Best Fixed Action

The learner's *external regret* is

$$\mathrm{Reg}_T := \sum_{t=1}^{T} z_t^\top \ell_t - \min_{z \in \Delta_d} \sum_{t=1}^{T} z^\top \ell_t.$$

The comparator is chosen after seeing all T loss vectors, but it must use the same distribution z on every round. Because the objective is linear in z, some pure action is an optimal comparator.

External Regret Compares Against the Best Fixed Action

The learner's *external regret* is

$$\mathrm{Reg}_T := \sum_{t=1}^{T} z_t^\top \ell_t - \min_{z \in \Delta_d} \sum_{t=1}^{T} z^\top \ell_t.$$

The comparator is chosen after seeing all T loss vectors, but it must use the same distribution z on every round. Because the objective is linear in z, some pure action is an optimal comparator. The learner is *no regret* if $\mathrm{Reg}_T / T \to 0$ for every loss sequence: on average, it does nearly as well as the best single action chosen in hindsight.

The Hedge Algorithm

Fix a learning rate $\eta > 0$ and initialize $h_{1,a} = 1$ for every action a.
In each round t:

1. Play the distribution

$$z_{t,a} := \frac{h_{t,a}}{\sum_b h_{t,b}}.$$

The Hedge Algorithm

Fix a learning rate $\eta > 0$ and initialize $h_{1,a} = 1$ for every action a.
In each round t:

1. Play the distribution

$$z_{t,a} := \frac{h_{t,a}}{\sum_b h_{t,b}}.$$

2. Observe $\ell_t \in [-1, 1]^d$ and incur expected loss $z_t^\top \ell_t$.

The Hedge Algorithm

Fix a learning rate $\eta > 0$ and initialize $h_{1,a} = 1$ for every action a.
In each round t:

1. Play the distribution

$$z_{t,a} := \frac{h_{t,a}}{\sum_b h_{t,b}}.$$

2. Observe $\ell_t \in [-1, 1]^d$ and incur expected loss $z_t^\top \ell_t$.
3. Update every action's weight:

$$h_{t+1,a} := h_{t,a} e^{-\eta \ell_{t,a}}.$$

The Hedge Algorithm

Fix a learning rate $\eta > 0$ and initialize $h_{1,a} = 1$ for every action a. In each round t:

1. Play the distribution

$$z_{t,a} := \frac{h_{t,a}}{\sum_b h_{t,b}}.$$

2. Observe $\ell_t \in [-1, 1]^d$ and incur expected loss $z_t^\top \ell_t$.
3. Update every action's weight:

$$h_{t+1,a} := h_{t,a} e^{-\eta \ell_{t,a}}.$$

Lower-loss actions retain more relative weight; η controls how aggressively the distribution responds.

Hedge Competes with Every Fixed Action

Theorem (Hedge regret bound)

For every sequence $\ell_1, \dots, \ell_T \in [-1, 1]^d$ *and every* $\eta > 0$, *Hedge satisfies*

$$\mathrm{Reg}_T \le \frac{\log d}{\eta} + \frac{\eta T}{2}.$$

Hedge Competes with Every Fixed Action

Theorem (Hedge regret bound)

For every sequence $\ell_1, \dots, \ell_T \in [-1, 1]^d$ *and every* $\eta > 0$, *Hedge satisfies*

$$\mathrm{Reg}_T \le \frac{\log d}{\eta} + \frac{\eta T}{2}.$$

The loss vector may be chosen adaptively after seeing the learner's current distribution. Thus an opponent's play can generate the loss vectors.

Hedge Competes with Every Fixed Action

Theorem (Hedge regret bound)

For every sequence $\ell_1, \dots, \ell_T \in [-1, 1]^d$ *and every* $\eta > 0$, *Hedge satisfies*

$$\mathrm{Reg}_T \le \frac{\log d}{\eta} + \frac{\eta T}{2}.$$

The loss vector may be chosen adaptively after seeing the learner's current distribution. Thus an opponent's play can generate the loss vectors.

Choosing η of order $1/\sqrt{T}$ makes Reg_T / T vanish. We now prove the bound.

The Analysis Tracks Cumulative Weight

Let

$$Z_t := \sum_a h_{t,a}, \qquad \widehat{L}_T := \sum_{t=1}^{T} z_t^\top \ell_t, \qquad L_{T,a} := \sum_{t=1}^{T} \ell_{t,a}.$$

The Analysis Tracks Cumulative Weight

Let

$$Z_t := \sum_a h_{t,a}, \qquad \widehat{L}_T := \sum_{t=1}^{T} z_t^\top \ell_t, \qquad L_{T,a} := \sum_{t=1}^{T} \ell_{t,a}.$$

We will bound the same final quantity $\log Z_{T+1}$ in two ways:

1. the one-round changes in total weight give an upper bound in terms of the learner's loss $\widehat{L}_T$;

The Analysis Tracks Cumulative Weight

Let

$$Z_t := \sum_a h_{t,a}, \qquad \widehat{L}_T := \sum_{t=1}^{T} z_t^\top \ell_t, \qquad L_{T,a} := \sum_{t=1}^{T} \ell_{t,a}.$$

We will bound the same final quantity $\log Z_{T+1}$ in two ways:

1. the one-round changes in total weight give an upper bound in terms of the learner's loss $\widehat{L}_T$;
2. the final weight of action a gives a lower bound in terms of that action's loss $L_{T,a}$.

The Analysis Tracks Cumulative Weight

Let

$$Z_t := \sum_a h_{t,a}, \qquad \widehat{L}_T := \sum_{t=1}^{T} z_t^\top \ell_t, \qquad L_{T,a} := \sum_{t=1}^{T} \ell_{t,a}.$$

We will bound the same final quantity $\log Z_{T+1}$ in two ways:

1. the one-round changes in total weight give an upper bound in terms of the learner's loss $\widehat{L}_T$;
2. the final weight of action a gives a lower bound in terms of that action's loss $L_{T,a}$.

Comparing the bounds will compare the learner with every fixed action.

The Weight Ratio Is an Exponential Moment

The Hedge update gives

$$\begin{aligned} Z_{t+1} &= \sum_a h_{t,a} e^{-\eta \ell_{t,a}} \\ &= Z_t \sum_a z_{t,a} e^{-\eta \ell_{t,a}}. \end{aligned}$$

Hence

$$\frac{Z_{t+1}}{Z_t} = \sum_a z_{t,a} e^{-\eta \ell_{t,a}}.$$

The Weight Ratio Is an Exponential Moment

The Hedge update gives

$$\begin{aligned} Z_{t+1} &= \sum_a h_{t,a} e^{-\eta \ell_{t,a}} \\ &= Z_t \sum_a z_{t,a} e^{-\eta \ell_{t,a}}. \end{aligned}$$

Hence

$$\frac{Z_{t+1}}{Z_t} = \sum_a z_{t,a} e^{-\eta \ell_{t,a}}.$$

If $A \sim z_t$ and $Y = -\ell_{t,A}$, then

$$\mathbb{E}[e^{\eta Y}] = \sum_a z_{t,a} e^{-\eta \ell_{t,a}} = \frac{Z_{t+1}}{Z_t}, \qquad \mathbb{E}[Y] = -z_t^\top \ell_t.$$

The Weight Ratio Is an Exponential Moment

The Hedge update gives

$$\begin{aligned} Z_{t+1} &= \sum_a h_{t,a} e^{-\eta \ell_{t,a}} \\ &= Z_t \sum_a z_{t,a} e^{-\eta \ell_{t,a}}. \end{aligned}$$

Hence

$$\frac{Z_{t+1}}{Z_t} = \sum_a z_{t,a} e^{-\eta \ell_{t,a}}.$$

If $A \sim z_t$ and $Y = -\ell_{t,A}$, then

$$\mathbb{E}[e^{\eta Y}] = \sum_a z_{t,a} e^{-\eta \ell_{t,a}} = \frac{Z_{t+1}}{Z_t}, \qquad \mathbb{E}[Y] = -z_t^\top \ell_t.$$

So the one-round growth of total weight is exactly an exponential moment of a bounded random variable. We need a tool that bounds such moments.

Hoeffding's Lemma Bounds the Exponential Moment

Lemma (Hoeffding's lemma)

If $Y \in [-1, 1]$, *then for every* $\eta > 0$,

$$\log \mathbb{E}[e^{\eta Y}] \leq \eta \mathbb{E}[Y] + \frac{\eta^2}{2}.$$

Hoeffding's Lemma Bounds the Exponential Moment

Lemma (Hoeffding's lemma)

If $Y \in [-1, 1]$, *then for every* $\eta > 0$,

$$\log \mathbb{E}[e^{\eta Y}] \leq \eta \mathbb{E}[Y] + \frac{\eta^2}{2}.$$

In words: an exponential moment of a bounded variable exceeds the exponential of its mean by at most a factor $e^{\eta^2/2}$.

Hoeffding's Lemma Bounds the Exponential Moment

Lemma (Hoeffding's lemma)

If $Y \in [-1, 1]$, *then for every* $\eta > 0$,

$$\log \mathbb{E}[e^{\eta Y}] \leq \eta \mathbb{E}[Y] + \frac{\eta^2}{2}.$$

In words: an exponential moment of a bounded variable exceeds the exponential of its mean by at most a factor $e^{\eta^2/2}$.
This is exactly the form we need: on the previous slide, $Y = -\ell_{t,A}$ takes values in $[-1, 1]$ and its exponential moment is the weight ratio Z_{t+1}/Z_t.

Learner Loss Upper-Bounds Final Total Weight

Substituting the two identities into Hoeffding's lemma gives

$$\log \frac{Z_{t+1}}{Z_t} \leq -\eta z_t^\top \ell_t + \frac{\eta^2}{2}.$$

Learner Loss Upper-Bounds Final Total Weight

Substituting the two identities into Hoeffding's lemma gives

$$\log \frac{Z_{t+1}}{Z_t} \leq -\eta z_t^\top \ell_t + \frac{\eta^2}{2}.$$

Summing over t telescopes the logarithms:

$$\begin{aligned} \log Z_{T+1} - \log Z_1 &\leq -\eta \sum_{t=1}^{T} z_t^\top \ell_t + \frac{\eta^2 T}{2} \\ &= -\eta \widehat{L}_T + \frac{\eta^2 T}{2}. \end{aligned}$$

Learner Loss Upper-Bounds Final Total Weight

Substituting the two identities into Hoeffding's lemma gives

$$\log \frac{Z_{t+1}}{Z_t} \leq -\eta z_t^\top \ell_t + \frac{\eta^2}{2}.$$

Summing over t telescopes the logarithms:

$$\begin{aligned} \log Z_{T+1} - \log Z_1 &\leq -\eta \sum_{t=1}^{T} z_t^\top \ell_t + \frac{\eta^2 T}{2} \\ &= -\eta \widehat{L}_T + \frac{\eta^2 T}{2}. \end{aligned}$$

Since $Z_1 = d$,

$$\log Z_{T+1} \leq \log d - \eta \widehat{L}_T + \frac{\eta^2 T}{2}.$$

Any Fixed Action Lower-Bounds Final Total Weight

Fix an action a. Iterating its Hedge update gives

$$\begin{aligned} h_{T+1,a} &= \prod_{t=1}^{T} e^{-\eta \ell_{t,a}} \\ &= e^{-\eta L_{T,a}}. \end{aligned}$$

Any Fixed Action Lower-Bounds Final Total Weight

Fix an action a. Iterating its Hedge update gives

$$\begin{aligned} h_{T+1,a} &= \prod_{t=1}^{T} e^{-\eta \ell_{t,a}} \\ &= e^{-\eta L_{T,a}}. \end{aligned}$$

Since Z_{T+1} is the sum of all final weights,

$$Z_{T+1} \geq h_{T+1,a}.$$

Taking logarithms therefore gives

$$\log Z_{T+1} \geq -\eta L_{T,a}.$$

The Two Weight Bounds Give Low Regret

Combining the upper and lower bounds gives

$$-\eta L_{T,a} \leq \log d - \eta \widehat{L}_T + \frac{\eta^2 T}{2}.$$

The Two Weight Bounds Give Low Regret

Combining the upper and lower bounds gives

$$-\eta L_{T,a} \leq \log d - \eta \widehat{L}_T + \frac{\eta^2 T}{2}.$$

Rearrange and divide by $\eta > 0$:

$$\widehat{L}_T - L_{T,a} \leq \frac{\log d}{\eta} + \frac{\eta T}{2}.$$

The Two Weight Bounds Give Low Regret

Combining the upper and lower bounds gives

$$-\eta L_{T,a} \leq \log d - \eta \widehat{L}_T + \frac{\eta^2 T}{2}.$$

Rearrange and divide by $\eta > 0$:

$$\widehat{L}_T - L_{T,a} \leq \frac{\log d}{\eta} + \frac{\eta T}{2}.$$

This holds for every action a, so

$$\boxed{\mathrm{Reg}_T \leq \frac{\log d}{\eta} + \frac{\eta T}{2}.}$$

Balancing the Two Terms Gives Vanishing Average Regret

Choose

$$\eta = \sqrt{\frac{2 \log d}{T}}.$$

Then

$$\frac{\log d}{\eta} = \frac{\eta T}{2} = \sqrt{\frac{T \log d}{2}}.$$

Balancing the Two Terms Gives Vanishing Average Regret

Choose

$$\eta = \sqrt{\frac{2 \log d}{T}}.$$

Then

$$\frac{\log d}{\eta} = \frac{\eta T}{2} = \sqrt{\frac{T \log d}{2}}.$$

Therefore

$$\mathrm{Reg}_T \leq \sqrt{2 T \log d}, \qquad \frac{\mathrm{Reg}_T}{T} \leq \sqrt{\frac{2 \log d}{T}} \longrightarrow 0.$$

Hedge is a no-regret algorithm.

A Thought Experiment: Min Learns, Max Best-Responds

Repeat the game for T rounds:

1. Min plays a distribution q_t using Hedge.

A Thought Experiment: Min Learns, Max Best-Responds

Repeat the game for T rounds:

1. Min plays a distribution q_t using Hedge.
2. After seeing q_t, Max chooses a best response

$$p_t \in \arg\max_p p^\top G q_t.$$

A Thought Experiment: Min Learns, Max Best-Responds

Repeat the game for T rounds:

1. Min plays a distribution q_t using Hedge.
2. After seeing q_t, Max chooses a best response

$$p_t \in \arg\max_p p^\top G q_t.$$

3. Min receives the loss vector $\ell_t = G^\top p_t \in [-1, 1]^{d_{\min}}$; hence Min's expected loss is $q_t^\top \ell_t = p_t^\top G q_t$.

A Thought Experiment: Min Learns, Max Best-Responds

Repeat the game for T rounds:

1. Min plays a distribution q_t using Hedge.
2. After seeing q_t, Max chooses a best response

$$p_t \in \arg\max_p p^\top G q_t.$$

3. Min receives the loss vector $\ell_t = G^\top p_t \in [-1, 1]^{d_{\min}}$; hence Min's expected loss is $q_t^\top \ell_t = p_t^\top G q_t$.

Write $u_t := p_t^\top G q_t$ for this payoff to Max. We will bound the same average payoff from below and above.

Best Responses Keep Payoff Above the Minimax Value

Because p_t is a best response to q_t,

$$u_t = \max_p p^\top G q_t.$$

Best Responses Keep Payoff Above the Minimax Value

Because p_t is a best response to q_t,

$$u_t = \max_p p^\top G q_t.$$

The minimax value is the smallest such best-response payoff over all of Min's strategies. Therefore, on every round,

$$u_t = \max_p p^\top G q_t \geq \min_q \max_p p^\top G q = v_{\text{minimax}}.$$

Best Responses Keep Payoff Above the Minimax Value

Because p_t is a best response to q_t,

$$u_t = \max_p p^\top G q_t.$$

The minimax value is the smallest such best-response payoff over all of Min's strategies. Therefore, on every round,

$$u_t = \max_p p^\top G q_t \geq \min_q \max_p p^\top G q = v_{\text{minimax}}.$$

Consequently,

$$\frac{1}{T} \sum_{t=1}^{T} u_t \geq v_{\text{minimax}}.$$

Min's Regret Keeps Payoff Near the Maximin Value

Let $\mathrm{Reg}_T^{\mathrm{Min}}$ denote Min's external regret on the loss vectors $\ell_t = G^\top p_t$. The regret guarantee says

$$\sum_{t=1}^{T} u_t - \min_q \sum_{t=1}^{T} p_t^\top G q \le \mathrm{Reg}_T^{\mathrm{Min}} .$$

Min's Regret Keeps Payoff Near the Maximin Value

Let $\mathrm{Reg}_T^{\mathrm{Min}}$ denote Min's external regret on the loss vectors $\ell_t = G^\top p_t$. The regret guarantee says

$$\sum_{t=1}^{T} u_t - \min_q \sum_{t=1}^{T} p_t^\top G q \le \mathrm{Reg}_T^{\mathrm{Min}}.$$

Let $\bar{p}_T := \frac{1}{T}\sum_t p_t$. Divide by T and use linearity:

$$\frac{1}{T}\sum_{t=1}^{T} u_t \le \min_q \bar{p}_T^\top G q + \frac{\mathrm{Reg}_T^{\mathrm{Min}}}{T}.$$

Min's Regret Keeps Payoff Near the Maximin Value

Let $\mathrm{Reg}_T^{\mathrm{Min}}$ denote Min's external regret on the loss vectors $\ell_t = G^\top p_t$. The regret guarantee says

$$\sum_{t=1}^{T} u_t - \min_q \sum_{t=1}^{T} p_t^\top G q \le \mathrm{Reg}_T^{\mathrm{Min}}.$$

Let $\bar{p}_T := \frac{1}{T}\sum_t p_t$. Divide by T and use linearity:

$$\frac{1}{T}\sum_{t=1}^{T} u_t \le \min_q \bar{p}_T^\top G q + \frac{\mathrm{Reg}_T^{\mathrm{Min}}}{T}.$$

The first term is the worst-case payoff of the specific policy $\bar{p}_T$, and v_{maximin} is by definition the largest such worst-case payoff over all policies:

$$\frac{1}{T}\sum_{t=1}^{T} u_t \le v_{\mathrm{maximin}} + \frac{\mathrm{Reg}_T^{\mathrm{Min}}}{T}.$$

Vanishing Regret Closes the Commitment Gap

The lower and upper bounds on the average payoff give

$$v_{\text{minimax}} \leq \frac{1}{T}\sum_{t=1}^{T} u_t \leq v_{\text{maximin}} + \frac{\text{Reg}_T^{\text{Min}}}{T}.$$

Vanishing Regret Closes the Commitment Gap

The lower and upper bounds on the average payoff give

$$v_{\text{minimax}} \leq \frac{1}{T}\sum_{t=1}^{T} u_t \leq v_{\text{maximin}} + \frac{\text{Reg}_T^{\text{Min}}}{T}.$$

Together with weak minimax,

$$0 \leq v_{\text{minimax}} - v_{\text{maximin}} \leq \frac{\text{Reg}_T^{\text{Min}}}{T}.$$

Vanishing Regret Closes the Commitment Gap

The lower and upper bounds on the average payoff give

$$v_{\text{minimax}} \leq \frac{1}{T}\sum_{t=1}^{T} u_t \leq v_{\text{maximin}} + \frac{\text{Reg}_T^{\text{Min}}}{T}.$$

Together with weak minimax,

$$0 \leq v_{\text{minimax}} - v_{\text{maximin}} \leq \frac{\text{Reg}_T^{\text{Min}}}{T}.$$

Hedge gives $\text{Reg}_T^{\text{Min}} / T \to 0$. Letting $T \to \infty$ proves

$$\boxed{v_{\text{maximin}} = v_{\text{minimax}}.}$$

Antisymmetry Establishes the Value of the Preference Game

Return to $G = M = -M^\top$.

Antisymmetry Establishes the Value of the Preference Game

Return to $G = M = -M^\top$. As we saw earlier, for every p, Min may copy p, so

$$\max_p \min_q p^\top M q \leq 0.$$

Antisymmetry Establishes the Value of the Preference Game

Return to $G = M = -M^\top$. As we saw earlier, for every p, Min may copy p, so

$$\max_p \min_q p^\top M q \leq 0.$$

Symmetrically — and this direction is new — for every q, Max may copy q, so

$$\min_q \max_p p^\top M q \geq 0.$$

Antisymmetry Establishes the Value of the Preference Game

Return to $G = M = -M^\top$. As we saw earlier, for every p, Min may copy p, so

$$\max_p \min_q p^\top M q \leq 0.$$

Symmetrically — and this direction is new — for every q, Max may copy q, so

$$\min_q \max_p p^\top M q \geq 0.$$

Minimax says these two values are equal. Therefore both equal zero:

$$\boxed{\max_p \min_q p^\top M q = 0.}$$

A Maximal Lottery Always Exists

Because the value is zero, some $p^\star \in \Delta(C)$ satisfies

$$\min_q (p^\star)^\top M q = 0.$$

Equivalently,

$$(p^\star)^\top M q \geq 0 \qquad \text{for every challenger } q.$$

A Maximal Lottery Always Exists

Because the value is zero, some $p^\star \in \Delta(C)$ satisfies

$$\min_q (p^\star)^\top M q = 0.$$

Equivalently,

$$(p^\star)^\top M q \geq 0 \qquad \text{for every challenger } q.$$

Such a policy is called a *maximal lottery*. Its probability of winning against any challenger is at least

$$\frac{1}{2} + \frac{1}{2}(p^\star)^\top M q \geq \frac{1}{2}.$$

The Minimax Proof Uses a Best-Response Oracle

The proof suggests an algorithm: Min learns with Hedge, while Max returns an exact best response p_t on every round. Averaging those responses gives

$$\min_q \bar{p}_T^\top M q \geq -\frac{\mathrm{Reg}_T^{\mathrm{Min}}}{T}, \qquad \bar{p}_T := \frac{1}{T}\sum_{t=1}^{T} p_t.$$

The Minimax Proof Uses a Best-Response Oracle

The proof suggests an algorithm: Min learns with Hedge, while Max returns an exact best response p_t on every round. Averaging those responses gives

$$\min_q \bar{p}_T^\top M q \geq -\frac{\mathrm{Reg}_T^{\mathrm{Min}}}{T}, \qquad \bar{p}_T := \frac{1}{T}\sum_{t=1}^{T} p_t.$$

For language models training however, computing a globally best completion or policy after every update is an unrealistic optimization problem.

The Minimax Proof Uses a Best-Response Oracle

The proof suggests an algorithm: Min learns with Hedge, while Max returns an exact best response p_t on every round. Averaging those responses gives

$$\min_q \bar{p}_T^\top M q \geq -\frac{\mathrm{Reg}_T^{\mathrm{Min}}}{T}, \qquad \bar{p}_T := \frac{1}{T}\sum_{t=1}^{T} p_t.$$

For language models training however, computing a globally best completion or policy after every update is an unrealistic optimization problem.

Can antisymmetry eliminate the best-response oracle?

Self-Play Eliminates the Best-Response Oracle

On round t, one learner chooses $p_t \in \Delta(C)$. Instead of constructing a separate best response, evaluate every action against opponents drawn from p_t itself by giving the learner the loss vector

$$\ell_t := -Mp_t.$$

Thus action a has loss $\ell_{t,a} = -(Mp_t)_a$: actions that tend to defeat the current policy have lower loss.

Self-Play Eliminates the Best-Response Oracle

On round t, one learner chooses $p_t \in \Delta(C)$. Instead of constructing a separate best response, evaluate every action against opponents drawn from p_t itself by giving the learner the loss vector

$$\ell_t := -Mp_t.$$

Thus action a has loss $\ell_{t,a} = -(Mp_t)_a$: actions that tend to defeat the current policy have lower loss.
Antisymmetry gives $p_t^\top Mp_t = 0$, and hence

$$\boxed{p_t^\top \ell_t = 0 \quad \text{on every round.}}$$

Regret will measure whether some fixed challenger could have systematically beaten the learner's changing policies.

External Regret Bounds Every Fixed Challenger

$$\underbrace{\sum_{t=1}^{T} p_t^\top \ell_t}_{\text{learner's cumulative loss}} - \underbrace{\min_{q \in \Delta(C)} \sum_{t=1}^{T} q^\top \ell_t}_{\text{best fixed challenger's cumulative loss}} \leq \mathrm{Reg}_T \,.$$

External Regret Bounds Every Fixed Challenger

$$\underbrace{\sum_{t=1}^{T} p_t^\top \ell_t}_{\text{learner's cumulative loss}} \quad - \quad \underbrace{\min_{q \in \Delta(C)} \sum_{t=1}^{T} q^\top \ell_t}_{\text{best fixed challenger's cumulative loss}} \quad \leq \mathrm{Reg}_T \,.$$

The self-tie identity makes the learner's cumulative loss zero. Substituting $\ell_t = -Mp_t$ therefore gives

$$\mathrm{Reg}_T \geq -\min_q \sum_{t=1}^{T} q^\top (-Mp_t) = \max_q \sum_{t=1}^{T} q^\top Mp_t,$$

where the last equality uses $-\min_q f(q) = \max_q [-f(q)]$.

A Fixed Challenger Sees Only the Average Policy

For any fixed challenger q, linearity gives

$$\sum_{t=1}^{T} q^\top M p_t = q^\top M \left(\sum_{t=1}^{T} p_t \right) = T q^\top M \bar{p}_T, \qquad \bar{p}_T := \frac{1}{T} \sum_{t=1}^{T} p_t.$$

A Fixed Challenger Sees Only the Average Policy

For any fixed challenger q, linearity gives

$$\sum_{t=1}^{T} q^\top M p_t = q^\top M \left(\sum_{t=1}^{T} p_t \right) = T q^\top M \bar{p}_T, \qquad \bar{p}_T := \frac{1}{T} \sum_{t=1}^{T} p_t.$$

Taking the maximum over q in the preceding regret bound gives

$$\mathrm{Reg}_T \geq T \max_{q \in \Delta(C)} q^\top M \bar{p}_T.$$

A Fixed Challenger Sees Only the Average Policy

For any fixed challenger q, linearity gives

$$\sum_{t=1}^{T} q^\top M p_t = q^\top M \left(\sum_{t=1}^{T} p_t \right) = T q^\top M \bar{p}_T, \qquad \bar{p}_T := \frac{1}{T} \sum_{t=1}^{T} p_t.$$

Taking the maximum over q in the preceding regret bound gives

$$\mathrm{Reg}_T \geq T \max_{q \in \Delta(C)} q^\top M \bar{p}_T.$$

Define the *exploitability* of p by

$$\mathrm{Expl}(p) := \max_{q \in \Delta(C)} q^\top M p.$$

A policy is a maximal lottery exactly when $\mathrm{Expl}(p) = 0$.

$$\boxed{\mathrm{Expl}(\bar{p}_T) \leq \frac{\mathrm{Reg}_T}{T}.}$$

Hedge Is an Abstraction of Policy Training

The self-play theorem is exact for a finite, explicitly represented set:

$$p_t \in \Delta(C), \qquad \ell_t(y) = -(Mp_t)_y.$$

Tabular Hedge stores one weight and constructs one loss coordinate for every $y \in C$. When C is the set of language-model completions, this is not a scalable algorithm.

Hedge Is an Abstraction of Policy Training

The self-play theorem is exact for a finite, explicitly represented set:

$$p_t \in \Delta(C), \qquad \ell_t(y) = -(Mp_t)_y.$$

Tabular Hedge stores one weight and constructs one loss coordinate for every $y \in C$. When C is the set of language-model completions, this is not a scalable algorithm.
Self-play has nevertheless made an important simplification:

one learner facing itself replaces
a learner plus a best-response oracle

To obtain a scalable procedure, we next replace the tabular learner by a sampled, parameterized policy optimizer.

Sampled Policy Updates Implement the Self-Play Idea

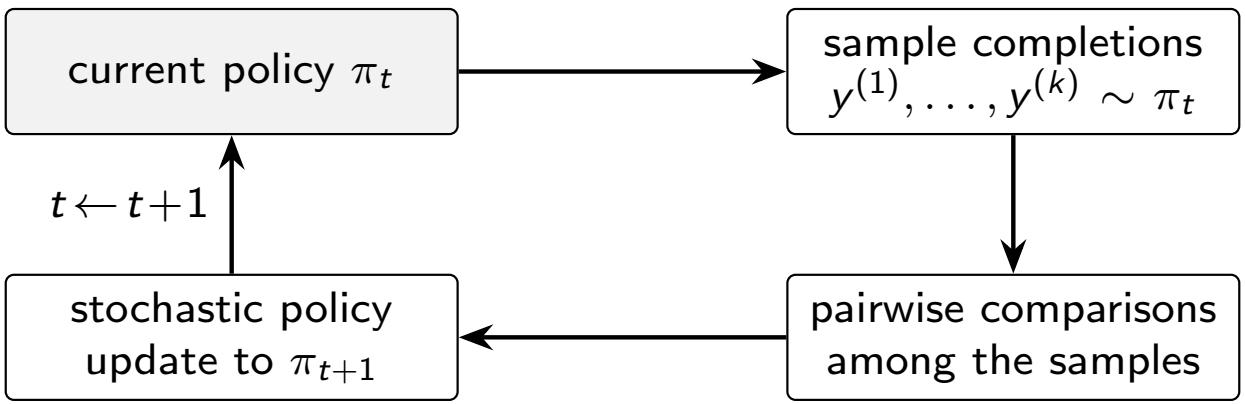
For a prompt x, sample completions from a parameterized policy $\pi_t(\cdot \mid x)$. Pairwise comparisons among the samples estimate the on-policy reward

$$r_t(x, y) := \mathbb{E}_{y' \sim \pi_t(\cdot|x)}[M_x(y, y')].$$

Sampled Policy Updates Implement the Self-Play Idea

For a prompt x, sample completions from a parameterized policy $\pi_t(\cdot \mid x)$. Pairwise comparisons among the samples estimate the on-policy reward

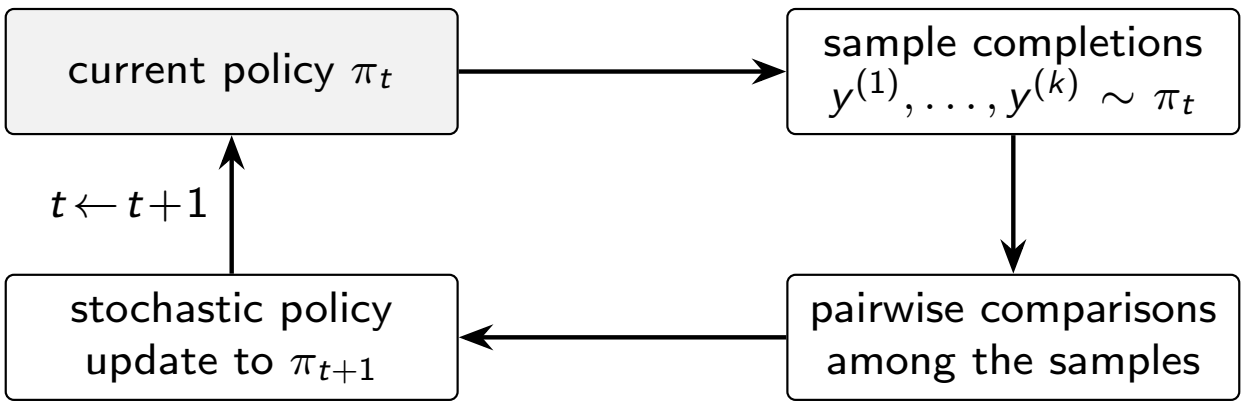
$$r_t(x, y) := \mathbb{E}_{y' \sim \pi_t(\cdot|x)}[M_x(y, y')].$$



Sampled Policy Updates Implement the Self-Play Idea

For a prompt x, sample completions from a parameterized policy $\pi_t(\cdot \mid x)$. Pairwise comparisons among the samples estimate the on-policy reward

$$r_t(x, y) := \mathbb{E}_{y' \sim \pi_t(\cdot|x)}[M_x(y, y')].$$



Sampling and parameter sharing avoid explicit enumeration. The robustness theorem is conditional on the policy optimizer achieving low regret on these changing objectives.

Summary

- A strict cycle makes every single completion vulnerable, while a lottery can balance the competing head-to-head defeats.

Summary

- A strict cycle makes every single completion vulnerable, while a lottery can balance the competing head-to-head defeats.
- Hedge achieves $O(\sqrt{T \log d})$ external regret by comparing total weight with the weight of each fixed action.

Summary

- A strict cycle makes every single completion vulnerable, while a lottery can balance the competing head-to-head defeats.
- Hedge achieves $O(\sqrt{T \log d})$ external regret by comparing total weight with the weight of each fixed action.
- If Min learns while Max best-responds, their average payoff is trapped between the minimax and maximin values up to Min's regret.

Summary

- A strict cycle makes every single completion vulnerable, while a lottery can balance the competing head-to-head defeats.
- Hedge achieves $O(\sqrt{T \log d})$ external regret by comparing total weight with the weight of each fixed action.
- If Min learns while Max best-responds, their average payoff is trapped between the minimax and maximin values up to Min's regret.
- Minimax proves that every finite comparison problem admits a maximal lottery.

Summary

- A strict cycle makes every single completion vulnerable, while a lottery can balance the competing head-to-head defeats.
- Hedge achieves $O(\sqrt{T \log d})$ external regret by comparing total weight with the weight of each fixed action.
- If Min learns while Max best-responds, their average payoff is trapped between the minimax and maximin values up to Min's regret.
- Minimax proves that every finite comparison problem admits a maximal lottery.
- Antisymmetry permits self-play: one no-regret learner satisfies $\mathrm{Expl}(\bar{p}_T) \leq \mathrm{Reg}_T / T$ without an exact best-response oracle.

Thanks!

See you next class!

References

- John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*, 1944.
- Yoav Freund and Robert E. Schapire. "Adaptive Game Playing Using Multiplicative Weights." *Games and Economic Behavior*, 1999.
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*, 2006.
- Peter C. Fishburn. "Probabilistic Social Choice Based on Simple Voting Comparisons." *Review of Economic Studies*, 1984.
- Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. "A Minimaximalist Approach to Reinforcement Learning from Human Feedback." *ICML*, 2024.