



POLITECNICO
MILANO 1863



Learning from Diverse Human Feedback

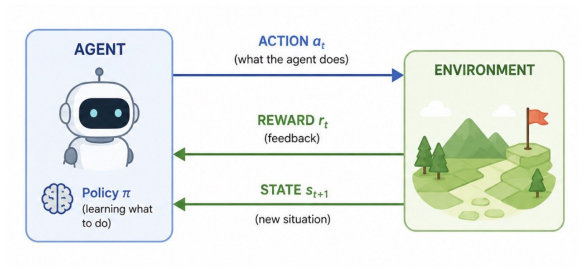
From Rewards to Comparisons and Rankings

Alberto Maria Metelli

TU Wien
28th May 2026

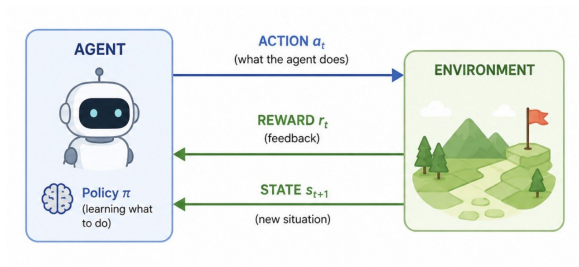
Message of the talk

- Reinforcement learning (Sutton and Barto, 2018) assumes that feedback is a **number**: a reward.
- Human feedback is often **relative**: “A is better than B”, “these are tied”, or “I cannot compare them.”
- Research question: *How can we extend reinforcement learning to non-numerical feedback?*
- We will focus on simpler setting of **multi-armed bandits** (Lattimore and Szepesvári, 2020)
 - No states, only actions



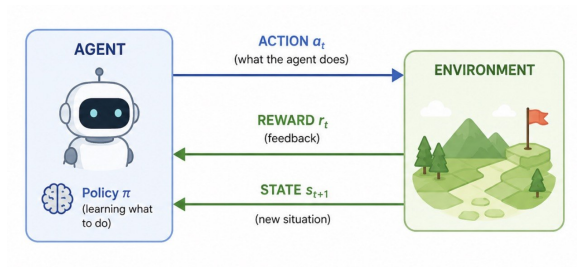
Message of the talk

- Reinforcement learning (Sutton and Barto, 2018) assumes that feedback is a **number**: a reward.
- Human feedback is often **relative**: “A is better than B”, “these are tied”, or “I cannot compare them.”
- Research question: *How can we extend reinforcement learning to non-numerical feedback?*
- We will focus on simpler setting of **multi-armed bandits** (Lattimore and Szepesvári, 2020)
 - No states, only actions



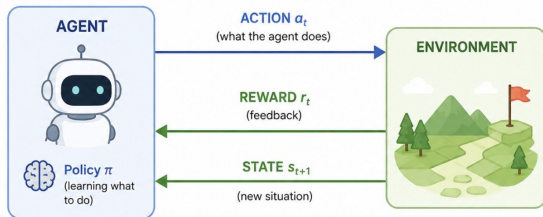
Message of the talk

- Reinforcement learning (Sutton and Barto, 2018) assumes that feedback is a **number**: a reward.
- Human feedback is often **relative**: “A is better than B”, “these are tied”, or “I cannot compare them.”
- Research question: ***How can we extend reinforcement learning to non-numerical feedback?***
- We will focus on simpler setting of **multi-armed bandits** (Lattimore and Szepesvári, 2020)
 - No states, only actions



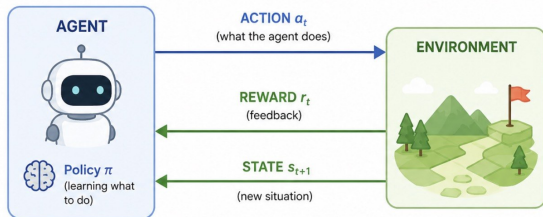
Message of the talk

- Reinforcement learning (Sutton and Barto, 2018) assumes that feedback is a **number**: a reward.
- Human feedback is often **relative**: “A is better than B”, “these are tied”, or “I cannot compare them.”
- Research question: ***How can we extend reinforcement learning to non-numerical feedback?***
- We will focus on simpler setting of **multi-armed bandits** (Lattimore and Szepesvári, 2020)
 - ▶ No states, only actions



Message of the talk

- Reinforcement learning (Sutton and Barto, 2018) assumes that feedback is a **number**: a reward.
- Human feedback is often **relative**: “A is better than B”, “these are tied”, or “I cannot compare them.”
- Research question: ***How can we extend reinforcement learning to non-numerical feedback?***
- We will focus on simpler setting of **multi-armed bandits** (Lattimore and Szepesvári, 2020)
 - ▶ No states, only actions

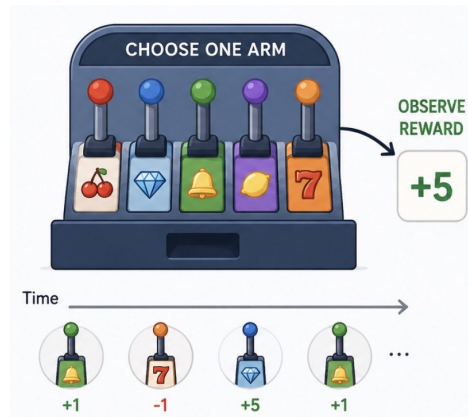


Outline

- 1 Multi-armed bandits
- 2 Dueling bandits
 - Preferences
 - Comparisons
- 3 Bandits with ranking feedback

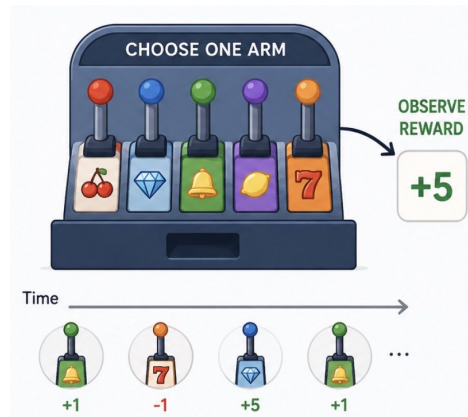
Multi-armed bandits: scalar reward feedback

- Arms (or actions): $i \in [K] := \{1, \dots, K\}$ (Lattimore and Szepesvári, 2020).
- At round $t \in [T]$:
 - the learner selects one arm $A_t \in [K]$;
 - the environment reveals only the scalar reward $X_{A_t, t}$ of that arm.



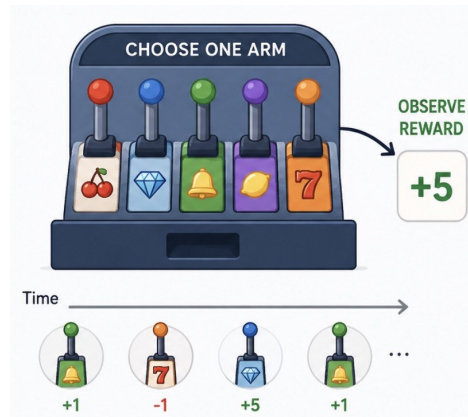
Multi-armed bandits: scalar reward feedback

- Arms (or actions): $i \in [K] := \{1, \dots, K\}$ (Lattimore and Szepesvári, 2020).
- At round $t \in [T]$:
 - 1 the learner selects one arm $A_t \in [K]$;
 - 2 the environment reveals only the scalar reward $X_{A_t, t}$ of that arm.



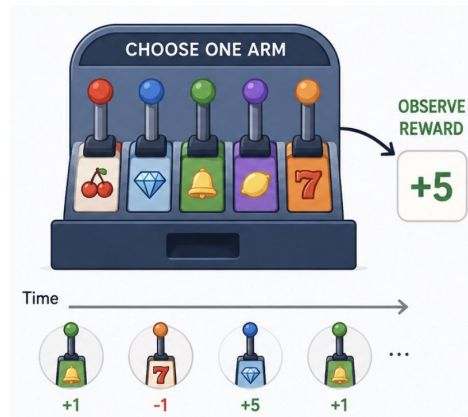
Multi-armed bandits: scalar reward feedback

- Arms (or actions): $i \in [K] := \{1, \dots, K\}$ (Lattimore and Szepesvári, 2020).
- At round $t \in [T]$:
 - the learner selects one arm $A_t \in [K]$;
 - the environment reveals only the scalar reward $X_{A_t, t}$ of that arm.



Multi-armed bandits: scalar reward feedback

- Arms (or actions): $i \in [K] := \{1, \dots, K\}$ (Lattimore and Szepesvári, 2020).
- At round $t \in [T]$:
 - the learner selects one arm $A_t \in [K]$;
 - the environment reveals only the scalar reward $X_{A_t, t}$ of that arm.



Stochastic Rewards

- Each arm has a fixed distribution ν_i
- Expected reward (or utility):

$$u_i = \mathbb{E}_{X \sim \nu_i}[X]$$

- An arm is **optimal** if it maximizes the expected reward: $i^* \in \arg \max_{i \in [K]} u_i$ with $u^* = u_{i^*}$.
- Suboptimality gap or **utility regret** $\Delta_i^u = u^* - u_i$
- We evaluate the performance of an algorithm over a learning horizon $T \in \mathbb{N}$ with the **expected cumulative regret**:

$$\text{Reg}(T) = \mathbb{E} \left[\sum_{t=1}^T \Delta_{I_t}^u \right]$$

Stochastic Rewards

- Each arm has a fixed distribution ν_i
- Expected reward (or utility):

$$u_i = \mathbb{E}_{X \sim \nu_i}[X]$$

- An arm is **optimal** if it maximizes the expected reward: $i^* \in \arg \max_{i \in [K]} u_i$ with $u^* = u_{i^*}$.
- Suboptimality gap or **utility regret** $\Delta_i^u = u^* - u_i$
- We evaluate the performance of an algorithm over a learning horizon $T \in \mathbb{N}$ with the **expected cumulative regret**:

$$\text{Reg}(T) = \mathbb{E} \left[\sum_{t=1}^T \Delta_k^u \right]$$

Stochastic Rewards

- Each arm has a fixed distribution ν_i
- Expected reward (or utility):

$$u_i = \mathbb{E}_{X \sim \nu_i}[X]$$

- An arm is **optimal** if it maximizes the expected reward: $i^* \in \arg \max_{i \in [K]} u_i$ with $u^* = u_{i^*}$.
- Suboptimality gap or **utility regret** $\Delta_i^u = u^* - u_i$
- We evaluate the performance of an algorithm over a learning horizon $T \in \mathbb{N}$ with the **expected cumulative regret**:

$$\text{Reg}(T) = \mathbb{E} \left[\sum_{t=1}^T \Delta_k^u \right]$$

Stochastic Rewards

- Each arm has a fixed distribution ν_i
- Expected reward (or utility):

$$u_i = \mathbb{E}_{X \sim \nu_i}[X]$$

- An arm is **optimal** if it maximizes the expected reward: $i^* \in \arg \max_{i \in [K]} u_i$ with $u^* = u_{i^*}$.
- Suboptimality gap or **utility regret** $\Delta_i^u = u^* - u_i$
- We evaluate the performance of an algorithm over a learning horizon $T \in \mathbb{N}$ with the **expected cumulative regret**:

$$\text{Reg}(T) = \mathbb{E} \left[\sum_{t=1}^T \Delta_k^u \right]$$

Stochastic Rewards

- Each arm has a fixed distribution ν_i
- Expected reward (or utility):

$$u_i = \mathbb{E}_{X \sim \nu_i}[X]$$

- An arm is **optimal** if it maximizes the expected reward: $i^* \in \arg \max_{i \in [K]} u_i$ with $u^* = u_{i^*}$.
- Suboptimality gap or **utility regret** $\Delta_i^u = u^* - u_i$
- We evaluate the performance of an algorithm over a learning horizon $T \in \mathbb{N}$ with the **expected cumulative regret**:

$$\text{Reg}(T) = \mathbb{E} \left[\sum_{t=1}^T \Delta_{I_t}^u \right]$$

Stochastic Rewards

- Each arm has a fixed distribution ν_i
- Expected reward (or utility):

$$u_i = \mathbb{E}_{X \sim \nu_i}[X]$$

- An arm is **optimal** if it maximizes the expected reward: $i^* \in \arg \max_{i \in [K]} u_i$ with $u^* = u_{i^*}$.
- Suboptimality gap or **utility regret** $\Delta_i^u = u^* - u_i$
- We evaluate the performance of an algorithm over a learning horizon $T \in \mathbb{N}$ with the **expected cumulative regret**:

$$\text{Reg}(T) = \mathbb{E} \left[\sum_{t=1}^T \Delta_{I_t}^u \right]$$

Adversarial rewards

- The reward vector $\mathbf{X}_{\cdot,t} \in [0, 1]^K$ may vary arbitrarily with time, i.e., it is chosen by an **adversary**.
- Compare against best **fixed** arm in hindsight: $i_T^* \in \arg \max_{i \in [K]} \sum_{t=1}^T X_{i,t}$.
- We evaluate the performance of an algorithm over a learning horizon $T \in \mathbb{N}$ with the **expected cumulative regret**:

$$\text{Reg}(T) = \max_{i \in [K]} \sum_{t=1}^T X_{i,t} - \mathbb{E} \left[\sum_{t=1}^T X_{A_t,t} \right].$$

Adversarial rewards

- The reward vector $\mathbf{X}_{\cdot,t} \in [0, 1]^K$ may vary arbitrarily with time, i.e., it is chosen by an **adversary**.
- Compare against best **fixed** arm in hindsight: $i_T^* \in \arg \max_{i \in [K]} \sum_{t=1}^T X_{i,t}$.
- We evaluate the performance of an algorithm over a learning horizon $T \in \mathbb{N}$ with the **expected cumulative regret**:

$$\text{Reg}(T) = \max_{i \in [K]} \sum_{t=1}^T X_{i,t} - \mathbb{E} \left[\sum_{t=1}^T X_{A_t,t} \right].$$

Adversarial rewards

- The reward vector $\mathbf{X}_{\cdot,t} \in [0, 1]^K$ may vary arbitrarily with time, i.e., it is chosen by an **adversary**.
- Compare against best **fixed** arm in hindsight: $i_T^* \in \arg \max_{i \in [K]} \sum_{t=1}^T X_{i,t}$.
- We evaluate the performance of an algorithm over a learning horizon $T \in \mathbb{N}$ with the **expected cumulative regret**:

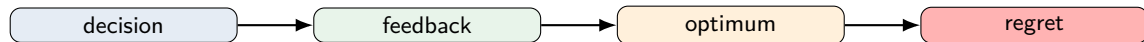
$$\text{Reg}(T) = \max_{i \in [K]} \sum_{t=1}^T X_{i,t} - \mathbb{E} \left[\sum_{t=1}^T X_{A_t,t} \right].$$

Adversarial rewards

- The reward vector $\mathbf{X}_{\cdot,t} \in [0, 1]^K$ may vary arbitrarily with time, i.e., it is chosen by an **adversary**.
- Compare against best **fixed** arm in hindsight: $i_T^* \in \arg \max_{i \in [K]} \sum_{t=1}^T X_{i,t}$.
- We evaluate the performance of an algorithm over a learning horizon $T \in \mathbb{N}$ with the **expected cumulative regret**:

$$\text{Reg}(T) = \max_{i \in [K]} \sum_{t=1}^T X_{i,t} - \mathbb{E} \left[\sum_{t=1}^T X_{A_t,t} \right].$$

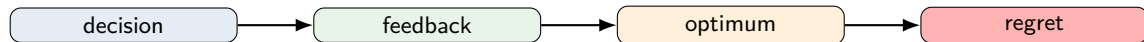
Taxonomy for Human Feedback



When moving to feedback different from **reward** we have several choices:

- What is the agent's **decision**? One arm, two arms, n arms, ...
- What is the **feedback** when executing the decision? One reward (stochastic or adversarial), a preference, a comparison, a ranking, a censored ranking, ...
- How to define a sensible notion of **optimum**? The best expected reward arm, but the reward may not exist...
- How to define a sensible notion of **regret**? Standard regret may not be achievable...

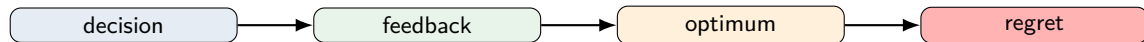
Taxonomy for Human Feedback



When moving to feedback different from **reward** we have several choices:

- What is the agent's **decision**? One arm, two arms, n arms, ...
- What is the **feedback** when executing the decision? One reward (stochastic or adversarial), a preference, a comparison, a ranking, a censored ranking, ...
- How to define a sensible notion of **optimum**? The best expected reward arm, but the reward may not exist...
- How to define a sensible notion of **regret**? Standard regret may not be achievable...

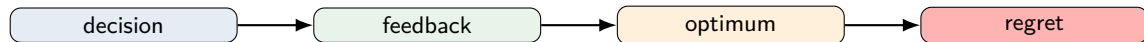
Taxonomy for Human Feedback



When moving to feedback different from **reward** we have several choices:

- What is the agent's **decision**? One arm, two arms, n arms, ...
- What is the **feedback** when executing the decision? One reward (stochastic or adversarial), a preference, a comparison, a ranking, a censored ranking, ...
- How to define a sensible notion of **optimum**? The best expected reward arm, but the reward may not exist...
- How to define a sensible notion of **regret**? Standard regret may not be achievable...

Taxonomy for Human Feedback



When moving to feedback different from **reward** we have several choices:

- What is the agent's **decision**? One arm, two arms, n arms, ...
- What is the **feedback** when executing the decision? One reward (stochastic or adversarial), a preference, a comparison, a ranking, a censored ranking, ...
- How to define a sensible notion of **optimum**? The best expected reward arm, but the reward may not exist...
- How to define a sensible notion of **regret**? Standard regret may not be achievable...

Outline

1 Multi-armed bandits

2 Dueling bandits

- Preferences
- Comparisons

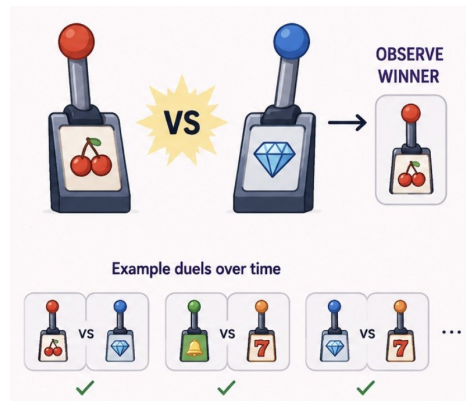
3 Bandits with ranking feedback

Outline

- 1 Multi-armed bandits
- 2 Dueling bandits
 - Preferences
 - Comparisons
- 3 Bandits with ranking feedback

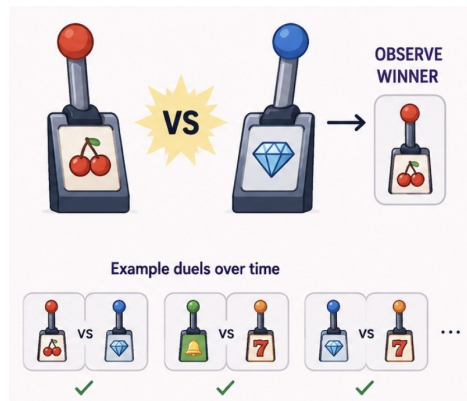
Dueling bandits

- Arms: $i \in [K]$ (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014).
- At round $t \in [T]$:
 - the learner selects a pair arms $I_t, J_t \in [K]$;
 - the environment reveals a preference $Y_{I_t, J_t, t} \in \{\succ, \prec\}$
- The feedback is binary:
 - $\succ$ means that I_t is preferred over J_t ($I_t \succ J_t$)
 - $\prec$ means that J_t is preferred over I_t ($J_t \succ I_t$)
- Often represented as the outcome of Bernoulli sampling:
1 for $\succ$ and 0 for $\prec$
- No scalar reward is directly observed.
- Optimum and regret?



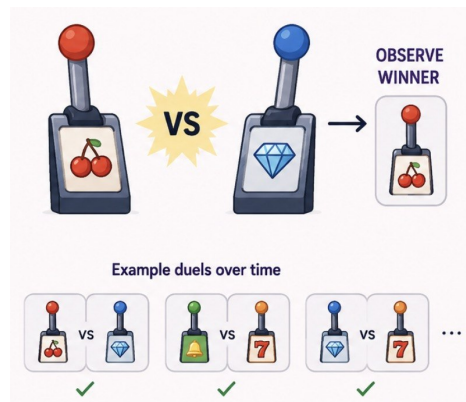
Dueling bandits

- Arms: $i \in [K]$ (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014).
- At round $t \in [T]$:
 - the learner selects a **pair** arms $I_t, J_t \in [K]$;
 - the environment reveals a preference $Y_{I_t, J_t, t} \in \{\succ, \prec\}$
- The feedback is binary:
 - $\succ$ means that I_t is preferred over J_t ($I_t \succ J_t$)
 - $\prec$ means that J_t is preferred over I_t ($J_t \succ I_t$)
- Often represented as the outcome of Bernoulli sampling:
1 for $\succ$ and 0 for $\prec$
- No scalar reward is directly observed.
- Optimum and regret?



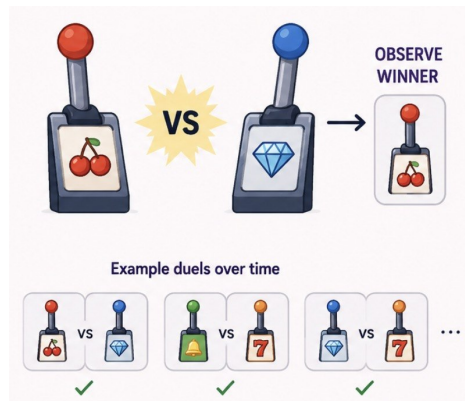
Dueling bandits

- Arms: $i \in [K]$ (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014).
- At round $t \in [T]$:
 - the learner selects a **pair** arms $I_t, J_t \in [K]$;
 - the environment reveals a preference $Y_{I_t, J_t, t} \in \{\succ, \prec\}$
- The feedback is binary:
 - $\succ$ means that I_t is preferred over J_t ($I_t \succ J_t$)
 - $\prec$ means that J_t is preferred over I_t ($J_t \succ I_t$)
- Often represented as the outcome of Bernoulli sampling:
1 for $\succ$ and 0 for $\prec$
- No scalar reward is directly observed.
- Optimum and regret?



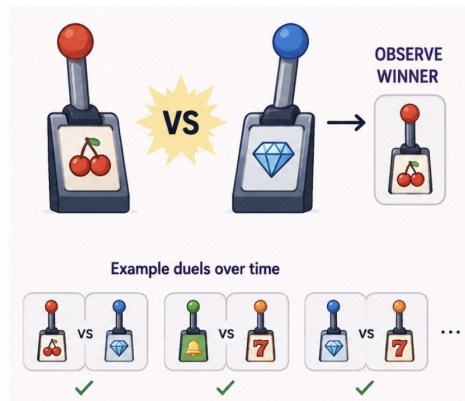
Dueling bandits

- Arms: $i \in [K]$ (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014).
- At round $t \in [T]$:
 - the learner selects a **pair** arms $I_t, J_t \in [K]$;
 - the environment reveals a preference $Y_{I_t, J_t, t} \in \{\succ, \prec\}$
- The feedback is binary:
 - $\succ$ means that I_t is preferred over J_t ($I_t \succ J_t$)
 - $\prec$ means that J_t is preferred over I_t ($J_t \succ I_t$)
- Often represented as the outcome of Bernoulli sampling:
1 for $\succ$ and 0 for $\prec$
- No scalar reward is directly observed.
- Optimum and regret?



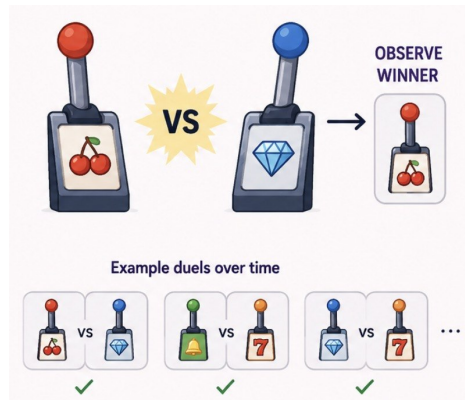
Dueling bandits

- Arms: $i \in [K]$ (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014).
- At round $t \in [T]$:
 - the learner selects a **pair** arms $I_t, J_t \in [K]$;
 - the environment reveals a preference $Y_{I_t, J_t, t} \in \{\succ, \prec\}$
- The feedback is binary:
 - $\succ$ means that I_t is preferred over J_t ($I_t \succ J_t$)
 - $\prec$ means that J_t is preferred over I_t ($J_t \succ I_t$)
- Often represented as the outcome of Bernoulli sampling:
1 for $\succ$ and 0 for $\prec$
- No scalar reward is directly observed.
- Optimum and regret?



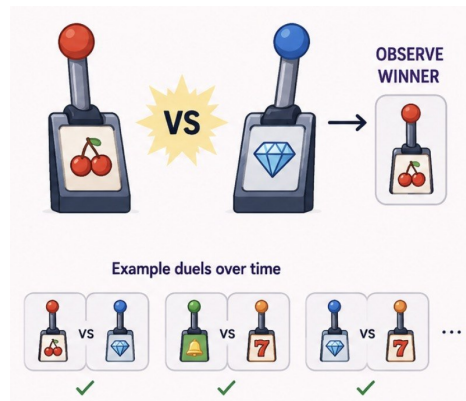
Dueling bandits

- Arms: $i \in [K]$ (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014).
- At round $t \in [T]$:
 - the learner selects a **pair** arms $I_t, J_t \in [K]$;
 - the environment reveals a preference $Y_{I_t, J_t, t} \in \{\succ, \prec\}$
- The feedback is binary:
 - $\succ$ means that I_t is preferred over J_t ($I_t \succ J_t$)
 - $\prec$ means that J_t is preferred over I_t ($J_t \succ I_t$)
- Often represented as the outcome of Bernoulli sampling:
1 for $\succ$ and 0 for $\prec$
- No scalar reward is directly observed.
- Optimum and regret?



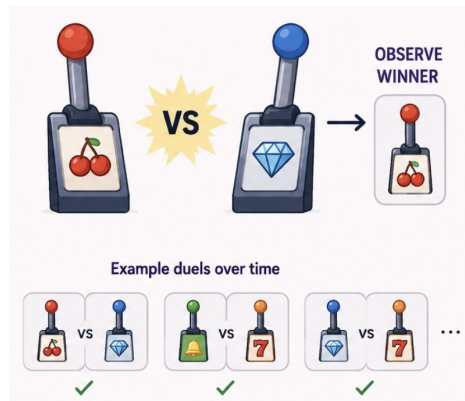
Dueling bandits

- Arms: $i \in [K]$ (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014).
- At round $t \in [T]$:
 - the learner selects a **pair** arms $I_t, J_t \in [K]$;
 - the environment reveals a preference $Y_{I_t, J_t, t} \in \{\succ, \prec\}$
- The feedback is binary:
 - $\succ$ means that I_t is preferred over J_t ($I_t \succ J_t$)
 - $\prec$ means that J_t is preferred over I_t ($J_t \succ I_t$)
- Often represented as the outcome of Bernoulli sampling:
1 for $\succ$ and 0 for $\prec$
- No scalar reward is directly observed.
- Optimum and regret?



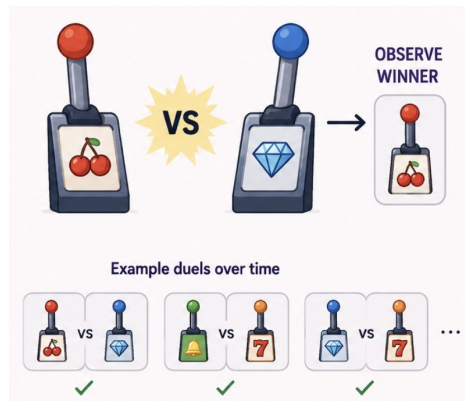
Dueling bandits

- Arms: $i \in [K]$ (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014).
- At round $t \in [T]$:
 - the learner selects a **pair** arms $I_t, J_t \in [K]$;
 - the environment reveals a preference $Y_{I_t, J_t, t} \in \{\succ, \prec\}$
- The feedback is binary:
 - $\succ$ means that I_t is preferred over J_t ($I_t \succ J_t$)
 - $\prec$ means that J_t is preferred over I_t ($J_t \succ I_t$)
- Often represented as the outcome of Bernoulli sampling:
1 for $\succ$ and 0 for $\prec$
- No scalar reward is directly observed.
- Optimum and regret?



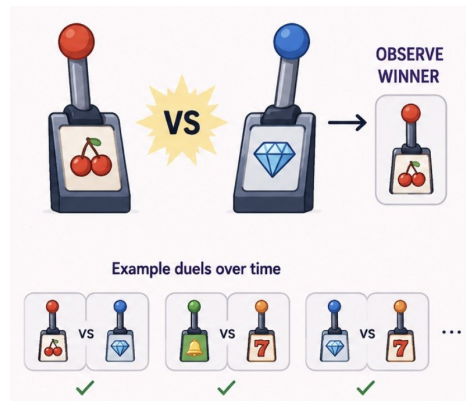
Dueling bandits

- Arms: $i \in [K]$ (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014).
- At round $t \in [T]$:
 - the learner selects a **pair** arms $I_t, J_t \in [K]$;
 - the environment reveals a preference $Y_{I_t, J_t, t} \in \{\succ, \prec\}$
- The feedback is binary:
 - $\succ$ means that I_t is preferred over J_t ($I_t \succ J_t$)
 - $\prec$ means that J_t is preferred over I_t ($J_t \succ I_t$)
- Often represented as the outcome of Bernoulli sampling:
1 for $\succ$ and 0 for $\prec$
- No scalar reward is directly observed.
- Optimum and regret?



Dueling bandits

- Arms: $i \in [K]$ (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014).
- At round $t \in [T]$:
 - the learner selects a **pair** arms $I_t, J_t \in [K]$;
 - the environment reveals a preference $Y_{I_t, J_t, t} \in \{\succ, \prec\}$
- The feedback is binary:
 - $\succ$ means that I_t is preferred over J_t ($I_t \succ J_t$)
 - $\prec$ means that J_t is preferred over I_t ($J_t \succ I_t$)
- Often represented as the outcome of Bernoulli sampling:
1 for $\succ$ and 0 for $\prec$
- No scalar reward is directly observed.
- Optimum and regret?



General probability matrix

- Preferences are generated **stochastically**: $Y_{I_t, J_t, t} \sim \text{Ber}(P_{I_t, J_t})$
- A preference matrix $\mathbf{P} \in [0, 1]^{K \times K}$ stores marginal win probabilities (Busa-Fekete and Hüllermeier, 2014):

$$P_{ij} = \mathbb{P}(i \succ j), \quad P_{ij} + P_{ji} = 1, \quad P_{ii} = \frac{1}{2}.$$

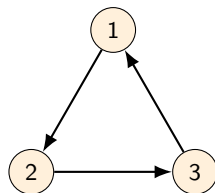
- The matrix $\mathbf{P}$ induces a relation:

$$i \succ_{\mathbf{P}} j \iff P_{ij} > \frac{1}{2} \quad (\text{also } i \succeq_{\mathbf{P}} j \iff P_{ij} \geq \frac{1}{2}).$$

- Ties are possible when $P_{ij} = 1/2$.
- The relation $\succ_{\mathbf{P}}$ can be cyclic:

$$P_{12} > \frac{1}{2}, \quad P_{23} > \frac{1}{2}, \quad P_{31} > \frac{1}{2}.$$

- Hence $\succ_{\mathbf{P}}$ may fail transitivity and is not necessarily an order.



majority cycle

General probability matrix

- Preferences are generated **stochastically**: $Y_{I_t, J_t, t} \sim \text{Ber}(P_{I_t, J_t})$
- A preference matrix $\mathbf{P} \in [0, 1]^{K \times K}$ stores marginal win probabilities (Busa-Fekete and Hüllermeier, 2014):

$$P_{ij} = \mathbb{P}(i \succ j), \quad P_{ij} + P_{ji} = 1, \quad P_{ii} = \frac{1}{2}.$$

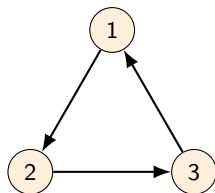
- The matrix $\mathbf{P}$ induces a relation:

$$i \succ_{\mathbf{P}} j \iff P_{ij} > \frac{1}{2} \quad (\text{also } i \succeq_{\mathbf{P}} j \iff P_{ij} \geq \frac{1}{2}).$$

- Ties are possible when $P_{ij} = 1/2$.
- The relation $\succ_{\mathbf{P}}$ can be cyclic:

$$P_{12} > \frac{1}{2}, \quad P_{23} > \frac{1}{2}, \quad P_{31} > \frac{1}{2}.$$

- Hence $\succ_{\mathbf{P}}$ may fail transitivity and is not necessarily an order.



majority cycle

General probability matrix

- Preferences are generated **stochastically**: $Y_{I_t, J_t, t} \sim \text{Ber}(P_{I_t, J_t})$
- A preference matrix $\mathbf{P} \in [0, 1]^{K \times K}$ stores marginal win probabilities (Busa-Fekete and Hüllermeier, 2014):

$$P_{ij} = \mathbb{P}(i \succ j), \quad P_{ij} + P_{ji} = 1, \quad P_{ii} = \frac{1}{2}.$$

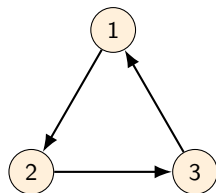
- The matrix $\mathbf{P}$ induces a relation:

$$i \succ_{\mathbf{P}} j \iff P_{ij} > \frac{1}{2} \quad (\text{also } i \succeq_{\mathbf{P}} j \iff P_{ij} \geq \frac{1}{2}).$$

- Ties are possible when $P_{ij} = 1/2$.
- The relation $\succ_{\mathbf{P}}$ can be cyclic:

$$P_{12} > \frac{1}{2}, \quad P_{23} > \frac{1}{2}, \quad P_{31} > \frac{1}{2}.$$

- Hence $\succ_{\mathbf{P}}$ may fail transitivity and is not necessarily an order.



majority cycle

General probability matrix

- Preferences are generated **stochastically**: $Y_{I_t, J_t, t} \sim \text{Ber}(P_{I_t, J_t})$
- A preference matrix $\mathbf{P} \in [0, 1]^{K \times K}$ stores marginal win probabilities (Busa-Fekete and Hüllermeier, 2014):

$$P_{ij} = \mathbb{P}(i \succ j), \quad P_{ij} + P_{ji} = 1, \quad P_{ii} = \frac{1}{2}.$$

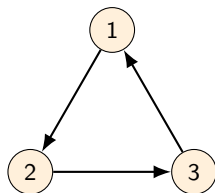
- The matrix $\mathbf{P}$ induces a relation:

$$i \succ_{\mathbf{P}} j \iff P_{ij} > \frac{1}{2} \quad (\text{also } i \succeq_{\mathbf{P}} j \iff P_{ij} \geq \frac{1}{2}).$$

- Ties are possible when $P_{ij} = 1/2$.
- The relation $\succ_{\mathbf{P}}$ can be cyclic:

$$P_{12} > \frac{1}{2}, \quad P_{23} > \frac{1}{2}, \quad P_{31} > \frac{1}{2}.$$

- Hence $\succ_{\mathbf{P}}$ may fail transitivity and is not necessarily an order.



majority cycle

General probability matrix

- Preferences are generated **stochastically**: $Y_{I_t, J_t, t} \sim \text{Ber}(P_{I_t, J_t})$
- A preference matrix $\mathbf{P} \in [0, 1]^{K \times K}$ stores marginal win probabilities (Busa-Fekete and Hüllermeier, 2014):

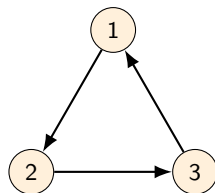
$$P_{ij} = \mathbb{P}(i \succ j), \quad P_{ij} + P_{ji} = 1, \quad P_{ii} = \frac{1}{2}.$$

- The matrix $\mathbf{P}$ induces a relation:

$$i \succ_{\mathbf{P}} j \iff P_{ij} > \frac{1}{2} \quad (\text{also } i \succeq_{\mathbf{P}} j \iff P_{ij} \geq \frac{1}{2}).$$

- Ties are possible when $P_{ij} = 1/2$.
- The relation $\succ_{\mathbf{P}}$ can be cyclic:

$$P_{12} > \frac{1}{2}, \quad P_{23} > \frac{1}{2}, \quad P_{31} > \frac{1}{2}.$$



majority cycle

- Hence $\succ_{\mathbf{P}}$ may fail transitivity and is not necessarily an order.

General probability matrix

- Preferences are generated **stochastically**: $Y_{I_t, J_t, t} \sim \text{Ber}(P_{I_t, J_t})$
- A preference matrix $\mathbf{P} \in [0, 1]^{K \times K}$ stores marginal win probabilities (Busa-Fekete and Hüllermeier, 2014):

$$P_{ij} = \mathbb{P}(i \succ j), \quad P_{ij} + P_{ji} = 1, \quad P_{ii} = \frac{1}{2}.$$

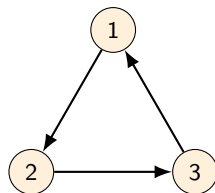
- The matrix $\mathbf{P}$ induces a relation:

$$i \succ_{\mathbf{P}} j \iff P_{ij} > \frac{1}{2} \quad (\text{also } i \succeq_{\mathbf{P}} j \iff P_{ij} \geq \frac{1}{2}).$$

- Ties are possible when $P_{ij} = 1/2$.
- The relation $\succ_{\mathbf{P}}$ can be cyclic:

$$P_{12} > \frac{1}{2}, \quad P_{23} > \frac{1}{2}, \quad P_{31} > \frac{1}{2}.$$

- Hence $\succ_{\mathbf{P}}$ may fail transitivity and is not necessarily an order.



majority cycle

General probability matrix: Optimum and regret

- Without an order, “best arm” is defined by a solution concept (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014):

Solution cept	con-	Definition	Comment
Condorcet		$P_{i^*j} > 1/2$ for all $j \neq i^*$	may not exist, may be not unique
Borda		$i_B^* \in \arg \max_{i \in [K]} s_i^B = \frac{1}{K-1} \sum_{j \neq i} P_{ij}$	averages pairwise wins
Copeland		$i_C^* \in \arg \max_{i \in [K]} s_i^C = \sum_{j \neq i} \mathbb{1}\{P_{ij} > 1/2\}$	counts majority victories

General probability matrix: Regret

- For Condorcet winner, define the suboptimality gap in probability $\Delta_i^C = P_{i^*i} - 1/2$ and
 - ▶ Strong regret

$$r_t^{C, \text{strong}} := \frac{\Delta_{i_t}^C + \Delta_{j_t}^C}{2}$$

- ▶ Weak regret

$$r_t^{C, \text{weak}} := \min\{\Delta_{i_t}^C, \Delta_{j_t}^C\} \leq r_t^{C, \text{strong}}$$

- For Borda and Copeland, define the regret as:

$$r_t^s = s_{i_t}^o - \frac{s_{i_t}^o + s_{j_t}^o}{2}, \quad i_t^* \in \arg \max_i s_i^o.$$

with $o \in \{B, C\}$.

General probability matrix: Regret

- For Condorcet winner, define the suboptimality gap in probability $\Delta_i^C = P_{i*i} - 1/2$ and
 - ▶ Strong regret

$$r_t^{C, \text{strong}} := \frac{\Delta_{I_t}^C + \Delta_{J_t}^C}{2}$$

- ▶ Weak regret

$$r_t^{C, \text{weak}} := \min\{\Delta_{I_t}^C, \Delta_{J_t}^C\} \leq r_t^{C, \text{strong}}$$

- For Borda and Copeland, define the regret as:

$$r_t^s = s_{I_t}^o - \frac{s_{I_t}^o + s_{J_t}^o}{2}, \quad i_s^* \in \arg \max_i s_i^o.$$

with $o \in \{B, C\}$.

General probability matrix: Regret

- For Condorcet winner, define the suboptimality gap in probability $\Delta_i^C = P_{i^*i} - 1/2$ and
 - ▶ Strong regret

$$r_t^{C,\text{strong}} := \frac{\Delta_{I_t}^C + \Delta_{J_t}^C}{2}$$

- ▶ Weak regret

$$r_t^{C,\text{weak}} := \min\{\Delta_{I_t}^C, \Delta_{J_t}^C\} \leq r_t^{C,\text{strong}}$$

- For Borda and Copeland, define the regret as:

$$r_t^s = s_{I_t}^o - \frac{s_{I_t}^o + s_{J_t}^o}{2}, \quad i_s^* \in \arg \max_i s_i^o.$$

with $o \in \{B, C\}$.

General probability matrix: Regret

- For Condorcet winner, define the suboptimality gap in probability $\Delta_i^C = P_{i^*i} - 1/2$ and
 - ▶ Strong regret

$$r_t^{C,\text{strong}} := \frac{\Delta_{I_t}^C + \Delta_{J_t}^C}{2}$$

- ▶ Weak regret

$$r_t^{C,\text{weak}} := \min\{\Delta_{I_t}^C, \Delta_{J_t}^C\} \leq r_t^{C,\text{strong}}$$

- For Borda and Copeland, define the regret as:

$$r_t^s = s_{I_t}^o - \frac{s_{I_t}^o + s_{J_t}^o}{2}, \quad I_t^s \in \arg \max_i s_i^o.$$

with $o \in \{B, C\}$.

General probability matrix: Regret

- For Condorcet winner, define the suboptimality gap in probability $\Delta_i^C = P_{i^*i} - 1/2$ and
 - ▶ Strong regret

$$r_t^{C,\text{strong}} := \frac{\Delta_{I_t}^C + \Delta_{J_t}^C}{2}$$

- ▶ Weak regret

$$r_t^{C,\text{weak}} := \min\{\Delta_{I_t}^C, \Delta_{J_t}^C\} \leq r_t^{C,\text{strong}}$$

- For Borda and Copeland, define the regret as:

$$r_t^s = s_{i_t^s}^o - \frac{s_{I_t}^o + s_{J_t}^o}{2}, \quad i_t^s \in \arg \max_i s_i^o.$$

with $o \in \{B, C\}$.

General probability matrix: Regret

- For Condorcet winner, define the suboptimality gap in probability $\Delta_i^C = P_{i^*i} - 1/2$ and
 - ▶ Strong regret

$$r_t^{C,\text{strong}} := \frac{\Delta_{I_t}^C + \Delta_{J_t}^C}{2}$$

- ▶ Weak regret

$$r_t^{C,\text{weak}} := \min\{\Delta_{I_t}^C, \Delta_{J_t}^C\} \leq r_t^{C,\text{strong}}$$

- For Borda and Copeland, define the regret as:

$$r_t^s = s_{i_s^*}^{\circ} - \frac{s_{I_t}^{\circ} + s_{J_t}^{\circ}}{2}, \quad i_s^* \in \arg \max_i s_i^{\circ}.$$

with $\circ \in \{B, C\}$.

General probability matrix: Regret

- For Condorcet winner, define the suboptimality gap in probability $\Delta_i^C = P_{i^*i} - 1/2$ and
 - ▶ Strong regret

$$r_t^{C,\text{strong}} := \frac{\Delta_{I_t}^C + \Delta_{J_t}^C}{2}$$

- ▶ Weak regret

$$r_t^{C,\text{weak}} := \min\{\Delta_{I_t}^C, \Delta_{J_t}^C\} \leq r_t^{C,\text{strong}}$$

- For Borda and Copeland, define the regret as:

$$r_t^s = s_{i_s^*}^{\circ} - \frac{s_{I_t}^{\circ} + s_{J_t}^{\circ}}{2}, \quad i_s^* \in \arg \max_i s_i^{\circ}.$$

with $\circ \in \{B, C\}$.

Orders

- Suppose the induced relation $\succ$ is assumed to be an **preorder** (Drago et al., 2025):
 - ▶ Reflexive: $i \succeq i$
 - ▶ Transitive $i \succeq j \wedge j \succeq l \implies i \succeq l$
- A preorder allows for **indifferent** ($\asymp$) pairs, i.e., $i \asymp j \iff i \succeq j \wedge j \succeq i$
- A (pre)order allows for **incomparable** ($\parallel$) pairs, i.e., $i \parallel j \iff \neg(i \succeq j \vee j \succeq i)$
- The relationship is an **order** if, in addition:
 - ▶ Antisymmetric: $i \succeq j \wedge j \succeq i \implies j = i$
- An order is **total** if no incomparable pairs exists, i.e., $i \succeq j \vee j \succeq i$
- Here $i \succ j \iff i \succeq j \wedge \neg(j \succeq i)$
- A sensible solution concept is the **maximal set** (i.e., the set of Condorcet winners):

$$\{i \in [K] : \nexists j \in [K] \text{ such that } j \succ i\}.$$

- Still very far from a reward

Orders

- Suppose the induced relation $\succ$ is assumed to be an **preorder** (Drago et al., 2025):
 - ▶ Reflexive: $i \succeq i$
 - ▶ Transitive $i \succeq j \wedge j \succeq l \implies i \succeq l$
- A preorder allows for **indifferent** ($\asymp$) pairs, i.e., $i \asymp j \iff i \succeq j \wedge j \succeq i$
- A (pre)order allows for **incomparable** ($\parallel$) pairs, i.e., $i \parallel j \iff \neg(i \succeq j \vee j \succeq i)$
- The relationship is an **order** if, in addition:
 - ▶ Antisymmetric: $i \succeq j \wedge j \succeq i \implies j = i$
- An order is **total** if no incomparable pairs exists, i.e., $i \succeq j \vee j \succeq i$
- Here $i \succ j \iff i \succeq j \wedge \neg(j \succeq i)$
- A sensible solution concept is the **maximal set** (i.e., the set of Condorcet winners):

$$\{i \in [K] : \nexists j \in [K] \text{ such that } j \succ i\}.$$

- Still very far from a reward

Orders

- Suppose the induced relation $\succ$ is assumed to be an **preorder** (Drago et al., 2025):
 - ▶ Reflexive: $i \succeq i$
 - ▶ Transitive $i \succeq j \wedge j \succeq l \implies i \succeq l$
- A preorder allows for **indifferent** ($\asymp$) pairs, i.e., $i \asymp j \iff i \succeq j \wedge j \succeq i$
- A (pre)order allows for **incomparable** ($\parallel$) pairs, i.e., $i \parallel j \iff \neg(i \succeq j \vee j \succeq i)$
- The relationship is an **order** if, in addition:
 - ▶ Antisymmetric: $i \succeq j \wedge j \succeq i \implies j = i$
- An order is **total** if no incomparable pairs exists, i.e., $i \succeq j \vee j \succeq i$
- Here $i \succ j \iff i \succeq j \wedge \neg(j \succeq i)$
- A sensible solution concept is the **maximal set** (i.e., the set of Condorcet winners):

$$\{i \in [K] : \nexists j \in [K] \text{ such that } j \succ i\}.$$

- Still very far from a reward

Orders

- Suppose the induced relation $\succ$ is assumed to be an **preorder** (Drago et al., 2025):
 - ▶ Reflexive: $i \succeq i$
 - ▶ Transitive $i \succeq j \wedge j \succeq l \implies i \succeq l$
- A preorder allows for **indifferent** ($\asymp$) pairs, i.e., $i \asymp j \iff i \succeq j \wedge j \succeq i$
- A (pre)order allows for **incomparable** ($\parallel$) pairs, i.e., $i \parallel j \iff \neg(i \succeq j \vee j \succeq i)$
- The relationship is an **order** if, in addition:
 - ▶ Antisymmetric: $i \succeq j \wedge j \succeq i \implies j = i$
- An order is **total** if no incomparable pairs exists, i.e., $i \succeq j \vee j \succeq i$
- Here $i \succ j \iff i \succeq j \wedge \neg(j \succeq i)$
- A sensible solution concept is the **maximal set** (i.e., the set of Condorcet winners):

$$\{i \in [K] : \nexists j \in [K] \text{ such that } j \succ i\}.$$

- Still very far from a reward

Orders

- Suppose the induced relation $\succ$ is assumed to be an **preorder** (Drago et al., 2025):
 - ▶ Reflexive: $i \succeq i$
 - ▶ Transitive $i \succeq j \wedge j \succeq l \implies i \succeq l$
- A preorder allows for **indifferent** ($\asymp$) pairs, i.e., $i \asymp j \iff i \succeq j \wedge j \succeq i$
- A (pre)order allows for **incomparable** ($\parallel$) pairs, i.e., $i \parallel j \iff \neg(i \succeq j \vee j \succeq i)$
- The relationship is an **order** if, in addition:
 - ▶ Antisymmetric: $i \succeq j \wedge j \succeq i \implies j = i$
- An order is **total** if no incomparable pairs exists, i.e., $i \succeq j \vee j \succeq i$
- Here $i \succ j \iff i \succeq j \wedge \neg(j \succeq i)$
- A sensible solution concept is the **maximal set** (i.e., the set of Condorcet winners):

$$\{i \in [K] : \nexists j \in [K] \text{ such that } j \succ i\}.$$

- Still very far from a reward

Orders

- Suppose the induced relation $\succ$ is assumed to be an **preorder** (Drago et al., 2025):
 - ▶ Reflexive: $i \succeq i$
 - ▶ Transitive $i \succeq j \wedge j \succeq l \implies i \succeq l$
- A preorder allows for **indifferent** ($\asymp$) pairs, i.e., $i \asymp j \iff i \succeq j \wedge j \succeq i$
- A (pre)order allows for **incomparable** ($\parallel$) pairs, i.e., $i \parallel j \iff \neg(i \succeq j \vee j \succeq i)$
- The relationship is an **order** if, in addition:
 - ▶ Antisymmetric: $i \succeq j \wedge j \succeq i \implies j = i$
- An order is **total** if no incomparable pairs exists, i.e., $i \succeq j \vee j \succeq i$
- Here $i \succ j \iff i \succeq j \wedge \neg(j \succeq i)$
- A sensible solution concept is the **maximal set** (i.e., the set of Condorcet winners):

$$\{i \in [K] : \nexists j \in [K] \text{ such that } j \succ i\}.$$

- Still very far from a reward

Orders

- Suppose the induced relation $\succ$ is assumed to be an **preorder** (Drago et al., 2025):
 - ▶ Reflexive: $i \succeq i$
 - ▶ Transitive $i \succeq j \wedge j \succeq l \implies i \succeq l$
- A preorder allows for **indifferent** ($\asymp$) pairs, i.e., $i \asymp j \iff i \succeq j \wedge j \succeq i$
- A (pre)order allows for **incomparable** ($\parallel$) pairs, i.e., $i \parallel j \iff \neg(i \succeq j \vee j \succeq i)$
- The relationship is an **order** if, in addition:
 - ▶ Antisymmetric: $i \succeq j \wedge j \succeq i \implies j = i$
- An order is **total** if no incomparable pairs exists, i.e., $i \succeq j \vee j \succeq i$
- Here $i \succ j \iff i \succeq j \wedge \neg(j \succeq i)$
- A sensible solution concept is the **maximal set** (i.e., the set of Condorcet winners):

$$\{i \in [K] : \nexists j \in [K] \text{ such that } j \succ i\}.$$

- Still very far from a reward

Orders

- Suppose the induced relation $\succ$ is assumed to be an **preorder** (Drago et al., 2025):
 - ▶ Reflexive: $i \succeq i$
 - ▶ Transitive $i \succeq j \wedge j \succeq l \implies i \succeq l$
- A preorder allows for **indifferent** ($\asymp$) pairs, i.e., $i \asymp j \iff i \succeq j \wedge j \succeq i$
- A (pre)order allows for **incomparable** ($\parallel$) pairs, i.e., $i \parallel j \iff \neg(i \succeq j \vee j \succeq i)$
- The relationship is an **order** if, in addition:
 - ▶ Antisymmetric: $i \succeq j \wedge j \succeq i \implies j = i$
- An order is **total** if no incomparable pairs exists, i.e., $i \succeq j \vee j \succeq i$
- Here $i \succ j \iff i \succeq j \wedge \neg(j \succeq i)$
- A sensible solution concept is the **maximal set** (i.e., the set of Condorcet winners):

$$\{i \in [K] : \nexists j \in [K] \text{ such that } j \succ i\}.$$

- Still very far from a reward

Orders

- Suppose the induced relation $\succ$ is assumed to be an **preorder** (Drago et al., 2025):
 - ▶ Reflexive: $i \succeq i$
 - ▶ Transitive $i \succeq j \wedge j \succeq l \implies i \succeq l$
- A preorder allows for **indifferent** ($\asymp$) pairs, i.e., $i \asymp j \iff i \succeq j \wedge j \succeq i$
- A (pre)order allows for **incomparable** ($\parallel$) pairs, i.e., $i \parallel j \iff \neg(i \succeq j \vee j \succeq i)$
- The relationship is an **order** if, in addition:
 - ▶ Antisymmetric: $i \succeq j \wedge j \succeq i \implies j = i$
- An order is **total** if no incomparable pairs exists, i.e., $i \succeq j \vee j \succeq i$
- Here $i \succ j \iff i \succeq j \wedge \neg(j \succeq i)$
- A sensible solution concept is the **maximal set** (i.e., the set of Condorcet winners):

$$\{i \in [K] : \nexists j \in [K] \text{ such that } j \succ i\}.$$

- Still very far from a reward

Orders

- Suppose the induced relation $\succ$ is assumed to be an **preorder** (Drago et al., 2025):
 - ▶ Reflexive: $i \succeq i$
 - ▶ Transitive $i \succeq j \wedge j \succeq l \implies i \succeq l$
- A preorder allows for **indifferent** ($\asymp$) pairs, i.e., $i \asymp j \iff i \succeq j \wedge j \succeq i$
- A (pre)order allows for **incomparable** ($\parallel$) pairs, i.e., $i \parallel j \iff \neg(i \succeq j \vee j \succeq i)$
- The relationship is an **order** if, in addition:
 - ▶ Antisymmetric: $i \succeq j \wedge j \succeq i \implies j = i$
- An order is **total** if no incomparable pairs exists, i.e., $i \succeq j \vee j \succeq i$
- Here $i \succ j \iff i \succeq j \wedge \neg(j \succeq i)$
- A sensible solution concept is the **maximal set** (i.e., the set of Condorcet winners):

$$\{i \in [K] : \nexists j \in [K] \text{ such that } j \succ i\}.$$

- Still very far from a reward

Orders

- Suppose the induced relation $\succ$ is assumed to be an **preorder** (Drago et al., 2025):
 - ▶ Reflexive: $i \succeq i$
 - ▶ Transitive $i \succeq j \wedge j \succeq l \implies i \succeq l$
- A preorder allows for **indifferent** ($\asymp$) pairs, i.e., $i \asymp j \iff i \succeq j \wedge j \succeq i$
- A (pre)order allows for **incomparable** ($\parallel$) pairs, i.e., $i \parallel j \iff \neg(i \succeq j \vee j \succeq i)$
- The relationship is an **order** if, in addition:
 - ▶ Antisymmetric: $i \succeq j \wedge j \succeq i \implies j = i$
- An order is **total** if no incomparable pairs exists, i.e., $i \succeq j \vee j \succeq i$
- Here $i \succ j \iff i \succeq j \wedge \neg(j \succeq i)$
- A sensible solution concept is the **maximal set** (i.e., the set of Condorcet winners):

$$\{i \in [K] : \nexists j \in [K] \text{ such that } j \succ i\}.$$

- Still very far from a reward

Random Utility Model

- Assume each arm has a latent (unobserved) scalar **utility** $u_i \in \mathbb{R}$ (Bradley and Terry, 1952; Thurstone, 1927; Luce, 1959).
- Consider a noise distribution Ψ
- Pairwise feedback is generated as:
 - ▶ Sample ϵ_i and ϵ_j independently from Ψ
 - ▶ Define the noisy utilities $U_i = u_i + \epsilon_i$ and $U_j = u_j + \epsilon_j$
 - ▶ Generate feedback as

$$Y_{ij} = \mathbb{1}\{U_i > U_j\}$$

Random Utility Model

- Assume each arm has a latent (unobserved) scalar **utility** $u_i \in \mathbb{R}$ (Bradley and Terry, 1952; Thurstone, 1927; Luce, 1959).
- Consider a noise distribution Ψ
- Pairwise feedback is generated as:
 - ▶ Sample ϵ_i and ϵ_j independently from Ψ
 - ▶ Define the noisy utilities $U_i = u_i + \epsilon_i$ and $U_j = u_j + \epsilon_j$
 - ▶ Generate feedback as

$$Y_{ij} = \mathbb{1}\{U_i > U_j\}$$

Random Utility Model

- Assume each arm has a latent (unobserved) scalar **utility** $u_i \in \mathbb{R}$ (Bradley and Terry, 1952; Thurstone, 1927; Luce, 1959).
- Consider a noise distribution Ψ
- Pairwise feedback is generated as:
 - ▶ Sample ϵ_i and ϵ_j independently from Ψ
 - ▶ Define the noisy utilities $U_i = u_i + \epsilon_i$ and $U_j = u_j + \epsilon_j$
 - ▶ Generate feedback as

$$Y_{i,j} = \mathbb{1}\{U_i > U_j\}$$

Random Utility Model

- Assume each arm has a latent (unobserved) scalar **utility** $u_i \in \mathbb{R}$ (Bradley and Terry, 1952; Thurstone, 1927; Luce, 1959).
- Consider a noise distribution Ψ
- Pairwise feedback is generated as:
 - ▶ Sample ϵ_i and ϵ_j independently from Ψ
 - ▶ Define the noisy utilities $U_i = u_i + \epsilon_i$ and $U_j = u_j + \epsilon_j$
 - ▶ Generate feedback as

$$Y_{i,j} = \mathbb{1}\{U_i > U_j\}$$

Random Utility Model

- Assume each arm has a latent (unobserved) scalar **utility** $u_i \in \mathbb{R}$ (Bradley and Terry, 1952; Thurstone, 1927; Luce, 1959).
- Consider a noise distribution Ψ
- Pairwise feedback is generated as:
 - ▶ Sample ϵ_i and ϵ_j independently from Ψ
 - ▶ Define the noisy utilities $U_i = u_i + \epsilon_i$ and $U_j = u_j + \epsilon_j$
 - ▶ Generate feedback as

$$Y_{i,j} = \mathbb{1}\{U_i > U_j\}$$

Random Utility Model

- Assume each arm has a latent (unobserved) scalar **utility** $u_i \in \mathbb{R}$ (Bradley and Terry, 1952; Thurstone, 1927; Luce, 1959).
- Consider a noise distribution Ψ
- Pairwise feedback is generated as:
 - ▶ Sample ϵ_i and ϵ_j independently from Ψ
 - ▶ Define the noisy utilities $U_i = u_i + \epsilon_i$ and $U_j = u_j + \epsilon_j$
 - ▶ Generate feedback as

$$Y_{i,j} = \mathbb{1}\{U_i > U_j\}$$

Random Utility Model

- Assume each arm has a latent (unobserved) scalar **utility** $u_i \in \mathbb{R}$ (Bradley and Terry, 1952; Thurstone, 1927; Luce, 1959).
- Consider a noise distribution Ψ
- Pairwise feedback is generated as:
 - ▶ Sample ϵ_i and ϵ_j independently from Ψ
 - ▶ Define the noisy utilities $U_i = u_i + \epsilon_i$ and $U_j = u_j + \epsilon_j$
 - ▶ Generate feedback as

$$Y_{i,j} = \mathbb{1}\{U_i > U_j\}$$

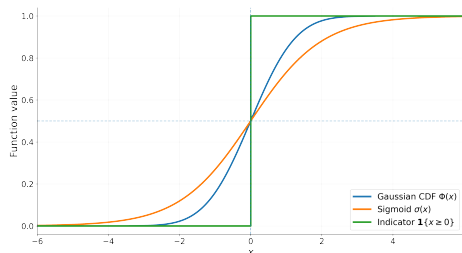
Random Utility Model

- Some observations:

- ▶ $U_i - U_j = u_i - u_j + (\epsilon_i - \epsilon_j)$
- ▶ The random variable $\epsilon_i - \epsilon_j$ is symmetric and has zero mean
- ▶ We denote its CDF as $\psi(\cdot)$ and we have that $\psi(-x) = 1 - \psi(x)$ by symmetry and, consequently, $\psi(0) = 1/2$.

- The feedback $Y_{i,j} \sim \text{Ber}(P_{ij})$ is Bernoulli distributed with parameter

$$P_{ij} = \mathbb{P}(Y_{i,j} = 1) = \mathbb{E}_{\epsilon_i, \epsilon_j \sim \nu}[\mathbf{1}\{U_i > U_j\}] = \mathbb{P}(\epsilon_i - \epsilon_j > u_j - u_i) = 1 - \psi(u_j - u_i) = \psi(u_i - u_j).$$



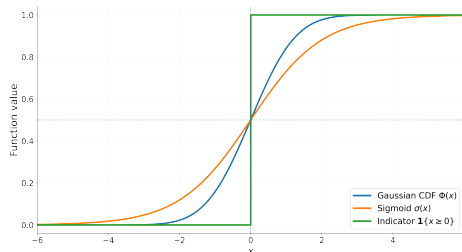
Random Utility Model

- Some observations:

- ▶ $U_i - U_j = u_i - u_j + (\epsilon_i - \epsilon_j)$
- ▶ The random variable $\epsilon_i - \epsilon_j$ is symmetric and has zero mean
- ▶ We denote its CDF as $\psi(\cdot)$ and we have that $\psi(-x) = 1 - \psi(x)$ by symmetry and, consequently, $\psi(0) = 1/2$.

- The feedback $Y_{i,j} \sim \text{Ber}(P_{ij})$ is Bernoulli distributed with parameter

$$P_{ij} = \mathbb{P}(Y_{i,j} = 1) = \mathbb{E}_{\epsilon_i, \epsilon_j \sim \nu}[\mathbf{1}\{U_i > U_j\}] = \mathbb{P}(\epsilon_i - \epsilon_j > u_j - u_i) = 1 - \psi(u_j - u_i) = \psi(u_i - u_j).$$



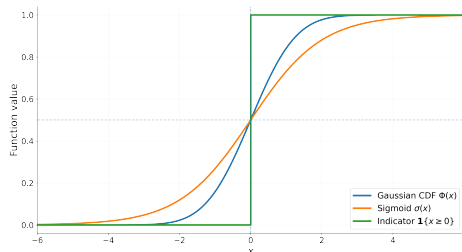
Random Utility Model

- Some observations:

- ▶ $U_i - U_j = u_i - u_j + (\epsilon_i - \epsilon_j)$
- ▶ The random variable $\epsilon_i - \epsilon_j$ is symmetric and has zero mean
- ▶ We denote its CDF as $\psi(\cdot)$ and we have that $\psi(-x) = 1 - \psi(x)$ by symmetry and, consequently, $\psi(0) = 1/2$.

- The feedback $Y_{i,j} \sim \text{Ber}(P_{ij})$ is Bernoulli distributed with parameter

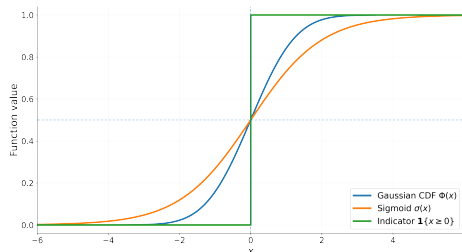
$$P_{ij} = \mathbb{P}(Y_{i,j} = 1) = \mathbb{E}_{\epsilon_i, \epsilon_j \sim \nu}[\mathbf{1}\{U_i > U_j\}] = \mathbb{P}(\epsilon_i - \epsilon_j > u_j - u_i) = 1 - \psi(u_j - u_i) = \psi(u_i - u_j).$$



Random Utility Model

- Some observations:

- ▶ $U_i - U_j = u_i - u_j + (\epsilon_i - \epsilon_j)$
- ▶ The random variable $\epsilon_i - \epsilon_j$ is symmetric and has zero mean
- ▶ We denote its CDF as $\psi(\cdot)$ and we have that $\psi(-x) = 1 - \psi(x)$ by symmetry and, consequently, $\psi(0) = 1/2$.



- The feedback $Y_{i,j} \sim \text{Ber}(P_{ij})$ is Bernoulli distributed with parameter

$$P_{ij} = \mathbb{P}(Y_{i,j} = 1) = \mathbb{E}_{\epsilon_i, \epsilon_j \sim \nu}[\mathbf{1}\{U_i > U_j\}] = \mathbb{P}(\epsilon_i - \epsilon_j > u_j - u_i) = 1 - \psi(u_j - u_i) = \psi(u_i - u_j).$$

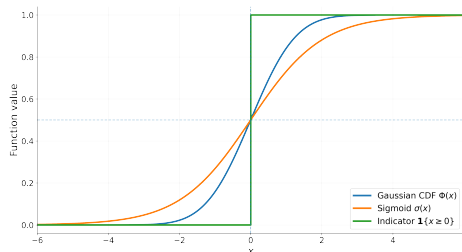
Random Utility Model

- Some observations:

- ▶ $U_i - U_j = u_i - u_j + (\epsilon_i - \epsilon_j)$
- ▶ The random variable $\epsilon_i - \epsilon_j$ is symmetric and has zero mean
- ▶ We denote its CDF as $\psi(\cdot)$ and we have that $\psi(-x) = 1 - \psi(x)$ by symmetry and, consequently, $\psi(0) = 1/2$.

- The feedback $Y_{i,j} \sim \text{Ber}(P_{ij})$ is Bernoulli distributed with parameter

$$P_{ij} = \mathbb{P}(Y_{i,j} = 1) = \mathbb{E}_{\epsilon_i, \epsilon_j \sim \nu}[\mathbb{1}\{U_i > U_j\}] = \mathbb{P}(\epsilon_i - \epsilon_j > u_j - u_i) = 1 - \psi(u_j - u_i) = \psi(u_i - u_j).$$



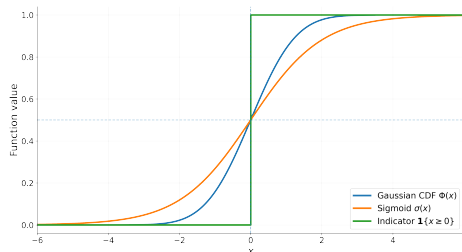
Random Utility Model

- Some observations:

- ▶ $U_i - U_j = u_i - u_j + (\epsilon_i - \epsilon_j)$
- ▶ The random variable $\epsilon_i - \epsilon_j$ is symmetric and has zero mean
- ▶ We denote its CDF as $\psi(\cdot)$ and we have that $\psi(-x) = 1 - \psi(x)$ by symmetry and, consequently, $\psi(0) = 1/2$.

- The feedback $Y_{i,j} \sim \text{Ber}(P_{ij})$ is Bernoulli distributed with parameter

$$P_{ij} = \mathbb{P}(Y_{i,j} = 1) = \mathbb{E}_{\epsilon_i, \epsilon_j \sim \nu}[\mathbb{1}\{U_i > U_j\}] = \mathbb{P}(\epsilon_i - \epsilon_j > u_j - u_i) = 1 - \psi(u_j - u_i) = \psi(u_i - u_j).$$



Random Utility Model

- Thus, we can relate win probabilities with utilities:

$$u_i > u_j \iff P_{ij} = \psi(u_i - u_j) > 1/2.$$

- Clear definition of optimum, compliant with standard rewards:

$$i^* \in \arg \max_{i \in [K]} u_i.$$

- The induced relation $\succ_P$ is a total preorder (no incomparabilities)

Random Utility Model

- Thus, we can relate win probabilities with utilities:

$$u_i > u_j \iff P_{ij} = \psi(u_i - u_j) > 1/2.$$

- Clear definition of optimum, compliant with standard rewards:

$$i^* \in \arg \max_{i \in [K]} u_i.$$

- The induced relation $\succ_P$ is a total preorder (no incomparabilities)

Random Utility Model

- Thus, we can relate win probabilities with utilities:

$$u_i > u_j \iff P_{ij} = \psi(u_i - u_j) > 1/2.$$

- Clear definition of optimum, compliant with standard rewards:

$$i^* \in \arg \max_{i \in [K]} u_i.$$

- The induced relation $\succ_P$ is a total preorder (no incomparabilities)

Random Utility Model

- Thus, we can relate win probabilities with utilities:

$$u_i > u_j \iff P_{ij} = \psi(u_i - u_j) > 1/2.$$

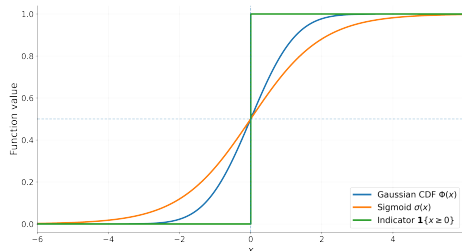
- Clear definition of optimum, compliant with standard rewards:

$$i^* \in \arg \max_{i \in [K]} u_i.$$

- The induced relation $\succ_{\mathbf{P}}$ is a total preorder (no incomparabilities)

Link functions

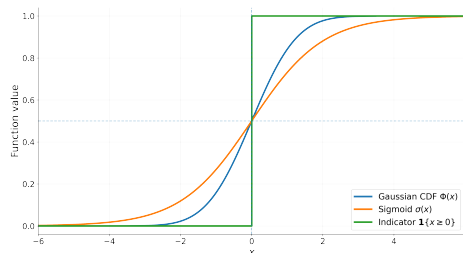
- One may also directly define a link function $\psi : \mathbb{R} \rightarrow [0, 1]$ (Drago et al., 2026b):
 - ▶ **monotonicity**: $x < y \implies \psi(x) \leq \psi(y)$
 - ▶ **symmetry**: $\psi(x) + \psi(-x) = 1$;
- This way, ψ is the CDF of a symmetric zero-mean distribution



Model	Link	Invertible	Random-utility model
Bradley–Terry	$\psi_{\text{BT}}(x) = \frac{1}{1+e^{-x}}$	Yes	Gumbel noise
Thurstone / probit	$\psi_{\text{Th}}(x) = \Phi(x/s)$	Yes	Gaussian noise
Hard threshold	$\psi(x) = \mathbb{1}\{x > 0\}$	No	deterministic

Link functions

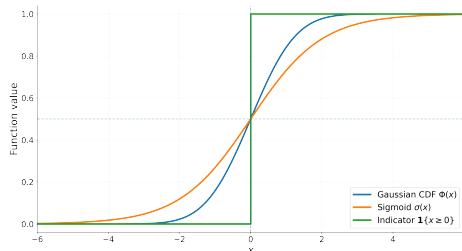
- One may also directly define a link function $\psi : \mathbb{R} \rightarrow [0, 1]$ (Drago et al., 2026b):
 - ▶ **monotonicity**: $x < y \implies \psi(x) \leq \psi(y)$
 - ▶ **symmetry**: $\psi(x) + \psi(-x) = 1$;
- This way, ψ is the CDF of a symmetric zero-mean distribution



Model	Link	Invertible	Random-utility model
Bradley–Terry	$\psi_{\text{BT}}(x) = \frac{1}{1+e^{-x}}$	Yes	Gumbel noise
Thurstone / probit	$\psi_{\text{Th}}(x) = \Phi(x/s)$	Yes	Gaussian noise
Hard threshold	$\psi(x) = \mathbb{1}\{x > 0\}$	No	deterministic

Link functions

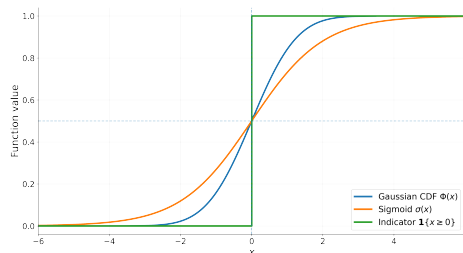
- One may also directly define a link function $\psi : \mathbb{R} \rightarrow [0, 1]$ (Drago et al., 2026b):
 - ▶ **monotonicity**: $x < y \implies \psi(x) \leq \psi(y)$
 - ▶ **symmetry**: $\psi(x) + \psi(-x) = 1$;
- This way, ψ is the CDF of a symmetric zero-mean distribution



Model	Link	Invertible	Random-utility model
Bradley–Terry	$\psi_{\text{BT}}(x) = \frac{1}{1+e^{-x}}$	Yes	Gumbel noise
Thurstone / probit	$\psi_{\text{Th}}(x) = \Phi(x/s)$	Yes	Gaussian noise
Hard threshold	$\psi(x) = \mathbb{1}\{x > 0\}$	No	deterministic

Link functions

- One may also directly define a link function $\psi : \mathbb{R} \rightarrow [0, 1]$ (Drago et al., 2026b):
 - ▶ **monotonicity**: $x < y \implies \psi(x) \leq \psi(y)$
 - ▶ **symmetry**: $\psi(x) + \psi(-x) = 1$;
- This way, ψ is the CDF of a symmetric zero-mean distribution



Model	Link	Invertible	Random-utility model
Bradley–Terry	$\psi_{\text{BT}}(x) = \frac{1}{1+e^{-x}}$	Yes	Gumbel noise
Thurstone / probit	$\psi_{\text{Th}}(x) = \Phi(x/s)$	Yes	Gaussian noise
Hard threshold	$\psi(x) = \mathbb{1}\{x > 0\}$	No	deterministic

Utility regret

- **Utility regret** for duel (i, j) :

$$r_t^u = u_{i^*} - \frac{u_i + u_j}{2} = \frac{\Delta_i^u + \Delta_j^u}{2}, \quad \Delta_i^u = u_{i^*} - u_i.$$

- This is exactly the regret of a standard bandit that receives the average utility of the two selected arms.
- If ψ is differentiable with $0 < \dot{\psi}_{\min} \leq \dot{\psi}(x) \leq \dot{\psi}_{\max} < +\infty$:

$$P_{i^*i} - 1/2 = \psi(u_{i^*} - u_i) - \psi(0) \leq \dot{\psi}_{\max}(u_{i^*} - u_i) = \dot{\psi}_{\max} \Delta_i^u$$

$$P_{i^*i} - 1/2 = \psi(u_{i^*} - u_i) - \psi(0) \geq \dot{\psi}_{\min}(u_{i^*} - u_i) = \dot{\psi}_{\min} \Delta_i^u$$

- Thus, we can control the utility regret with the Condorcet (strong) regret
- $\dot{\psi}$ is an index of the "rationality" of the human

Utility regret

- **Utility regret** for duel (i, j) :

$$r_t^u = u_{i^*} - \frac{u_i + u_j}{2} = \frac{\Delta_i^u + \Delta_j^u}{2}, \quad \Delta_i^u = u_{i^*} - u_i.$$

- This is exactly the regret of a standard bandit that receives the average utility of the two selected arms.
- If ψ is differentiable with $0 < \dot{\psi}_{\min} \leq \dot{\psi}(x) \leq \dot{\psi}_{\max} < +\infty$:

$$P_{i^*i} - 1/2 = \psi(u_{i^*} - u_i) - \psi(0) \leq \dot{\psi}_{\max}(u_{i^*} - u_i) = \dot{\psi}_{\max} \Delta_i^u$$

$$P_{i^*i} - 1/2 = \psi(u_{i^*} - u_i) - \psi(0) \geq \dot{\psi}_{\min}(u_{i^*} - u_i) = \dot{\psi}_{\min} \Delta_i^u$$

- Thus, we can control the utility regret with the Condorcet (strong) regret
- $\dot{\psi}$ is an index of the "rationality" of the human

Utility regret

- **Utility regret** for duel (i, j) :

$$r_t^u = u_{i^*} - \frac{u_i + u_j}{2} = \frac{\Delta_i^u + \Delta_j^u}{2}, \quad \Delta_i^u = u_{i^*} - u_i.$$

- This is exactly the regret of a standard bandit that receives the average utility of the two selected arms.
- If ψ is differentiable with $0 < \dot{\psi}_{\min} \leq \dot{\psi}(x) \leq \dot{\psi}_{\max} < +\infty$:

$$P_{i^*i} - 1/2 = \psi(u_{i^*} - u_i) - \psi(0) \leq \dot{\psi}_{\max}(u_{i^*} - u_i) = \dot{\psi}_{\max} \Delta_i^u$$

$$P_{i^*i} - 1/2 = \psi(u_{i^*} - u_i) - \psi(0) \geq \dot{\psi}_{\min}(u_{i^*} - u_i) = \dot{\psi}_{\min} \Delta_i^u$$

- Thus, we can control the utility regret with the Condorcet (strong) regret
- $\dot{\psi}$ is an index of the "rationality" of the human

Utility regret

- **Utility regret** for duel (i, j) :

$$r_t^u = u_{i^*} - \frac{u_i + u_j}{2} = \frac{\Delta_i^u + \Delta_j^u}{2}, \quad \Delta_i^u = u_{i^*} - u_i.$$

- This is exactly the regret of a standard bandit that receives the average utility of the two selected arms.
- If ψ is differentiable with $0 < \dot{\psi}_{\min} \leq \dot{\psi}(x) \leq \dot{\psi}_{\max} < +\infty$:

$$P_{i^*i} - 1/2 = \psi(u_{i^*} - u_i) - \psi(0) \leq \dot{\psi}_{\max}(u_{i^*} - u_i) = \dot{\psi}_{\max} \Delta_i^u$$

$$P_{i^*i} - 1/2 = \psi(u_{i^*} - u_i) - \psi(0) \geq \dot{\psi}_{\min}(u_{i^*} - u_i) = \dot{\psi}_{\min} \Delta_i^u$$

- Thus, we can control the utility regret with the Condorcet (strong) regret
- $\dot{\psi}$ is an index of the "rationality" of the human

Utility regret

- **Utility regret** for duel (i, j) :

$$r_t^u = u_{i^*} - \frac{u_i + u_j}{2} = \frac{\Delta_i^u + \Delta_j^u}{2}, \quad \Delta_i^u = u_{i^*} - u_i.$$

- This is exactly the regret of a standard bandit that receives the average utility of the two selected arms.
- If ψ is differentiable with $0 < \dot{\psi}_{\min} \leq \dot{\psi}(x) \leq \dot{\psi}_{\max} < +\infty$:

$$P_{i^*i} - 1/2 = \psi(u_{i^*} - u_i) - \psi(0) \leq \dot{\psi}_{\max}(u_{i^*} - u_i) = \dot{\psi}_{\max} \Delta_i^u$$

$$P_{i^*i} - 1/2 = \psi(u_{i^*} - u_i) - \psi(0) \geq \dot{\psi}_{\min}(u_{i^*} - u_i) = \dot{\psi}_{\min} \Delta_i^u$$

- Thus, we can control the utility regret with the Condorcet (strong) regret
- $\dot{\psi}$ is an index of the "rationality" of the human

Utility regret

- **Utility regret** for duel (i, j) :

$$r_t^u = u_{i^*} - \frac{u_i + u_j}{2} = \frac{\Delta_i^u + \Delta_j^u}{2}, \quad \Delta_i^u = u_{i^*} - u_i.$$

- This is exactly the regret of a standard bandit that receives the average utility of the two selected arms.
- If ψ is differentiable with $0 < \dot{\psi}_{\min} \leq \dot{\psi}(x) \leq \dot{\psi}_{\max} < +\infty$:

$$P_{i^*i} - 1/2 = \psi(u_{i^*} - u_i) - \psi(0) \leq \dot{\psi}_{\max}(u_{i^*} - u_i) = \dot{\psi}_{\max}\Delta_i^u$$

$$P_{i^*i} - 1/2 = \psi(u_{i^*} - u_i) - \psi(0) \geq \dot{\psi}_{\min}(u_{i^*} - u_i) = \dot{\psi}_{\min}\Delta_i^u$$

- Thus, we can control the utility regret with the Condorcet (strong) regret
- $\dot{\psi}$ is an index of the "rationality" of the human

Open questions

- What if ψ is non-differentiable or $\dot{\psi}_{\min} = 0$? Like for $\mathbb{1}$.
 - ▶ Still it is simple to identify the best arm!
- What if noises ϵ_i or utilities U_i are **adversarial**?
 - ▶ No longer true that $u_{i^*} - u_i \geq 0$, where i^* is the best in hindsight
 - ▶ Currently, no-regret possible only if ψ linear (Gajane et al., 2015)

Open questions

- What if ψ is non-differentiable or $\dot{\psi}_{\min} = 0$? Like for $\mathbb{1}$.
 - ▶ Still it is simple to identify the best arm!
- What if noises ϵ_i or utilities U_i are **adversarial**?
 - ▶ No longer true that $u_{i^*} - u_i \geq 0$, where i^* is the best in hindsight
 - ▶ Currently, no-regret possible only if ψ linear (Gajane et al., 2015)

Open questions

- What if ψ is non-differentiable or $\dot{\psi}_{\min} = 0$? Like for $\mathbb{1}$.
 - ▶ Still it is simple to identify the best arm!
- What if noises ϵ_i or utilities U_i are **adversarial**?
 - ▶ No longer true that $u_{i^*} - u_i \geq 0$, where i^* is the best in hindsight
 - ▶ Currently, no-regret possible only if ψ linear (Gajane et al., 2015)

Open questions

- What if ψ is non-differentiable or $\dot{\psi}_{\min} = 0$? Like for $\mathbb{1}$.
 - ▶ Still it is simple to identify the best arm!
- What if noises ϵ_i or utilities U_i are **adversarial**?
 - ▶ No longer true that $u_{i^*} - u_i \geq 0$, where i^* is the best in hindsight
 - ▶ Currently, no-regret possible only if ψ linear (Gajane et al., 2015)

Open questions

- What if ψ is non-differentiable or $\dot{\psi}_{\min} = 0$? Like for $\mathbb{1}$.
 - ▶ Still it is simple to identify the best arm!
- What if noises ϵ_i or utilities U_i are **adversarial**?
 - ▶ No longer true that $u_{i^*} - u_i \geq 0$, where i^* is the best in hindsight
 - ▶ Currently, no-regret possible only if ψ linear (Gajane et al., 2015)

Outline

- 1 Multi-armed bandits
- 2 Dueling bandits
 - Preferences
 - Comparisons
- 3 Bandits with ranking feedback

Comparisons

- Binary dueling forces every queried pair into one of two outcomes $\{\succ, \prec\}$, i.e., **preferences**.
- Human rational behavior may lead to four outcomes, i.e., **comparisons** (Rao and Kupper, 1967; Davidson, 1970)

$$\mathcal{O} = \{\succ, \prec, \asymp, \parallel\}.$$

- For arms i, j , define comparison probabilities

$$Q_{ij}^o = \mathbb{P}(i \circ j), \quad o \in \mathcal{O}, \quad \sum_{o \in \mathcal{O}} Q_{ij}^o = 1.$$

- Interpretations:

- ▶ $i \succ j$: i is preferred to j ;
- ▶ $i \asymp j$: i and j are equivalent;
- ▶ $i \parallel j$: i and j are incomparable.

- Difficult (and not particularly interesting) defining solution concepts like Borda, Copeland

Comparisons

- Binary dueling forces every queried pair into one of two outcomes $\{\succ, \prec\}$, i.e., **preferences**.
- Human rational behavior may lead to four outcomes, i.e., **comparisons** (Rao and Kupper, 1967; Davidson, 1970)

$$\mathcal{O} = \{\succ, \prec, \asymp, \parallel\}.$$

- For arms i, j , define comparison probabilities

$$Q_{ij}^o = \mathbb{P}(i \circ j), \quad o \in \mathcal{O}, \quad \sum_{o \in \mathcal{O}} Q_{ij}^o = 1.$$

- Interpretations:

- ▶ $i \succ j$: i is preferred to j ;
- ▶ $i \asymp j$: i and j are equivalent;
- ▶ $i \parallel j$: i and j are incomparable.

- Difficult (and not particularly interesting) defining solution concepts like Borda, Copeland

Comparisons

- Binary dueling forces every queried pair into one of two outcomes $\{\succ, \prec\}$, i.e., **preferences**.
- Human rational behavior may lead to four outcomes, i.e., **comparisons** (Rao and Kupper, 1967; Davidson, 1970)

$$\mathcal{O} = \{\succ, \prec, \asymp, \parallel\}.$$

- For arms i, j , define comparison probabilities

$$Q_{ij}^o = \mathbb{P}(i \circ j), \quad o \in \mathcal{O}, \quad \sum_{o \in \mathcal{O}} Q_{ij}^o = 1.$$

- Interpretations:

- ▶ $i \succ j$: i is preferred to j ;
- ▶ $i \asymp j$: i and j are equivalent;
- ▶ $i \parallel j$: i and j are incomparable.

- Difficult (and not particularly interesting) defining solution concepts like Borda, Copeland

Comparisons

- Binary dueling forces every queried pair into one of two outcomes $\{\succ, \prec\}$, i.e., **preferences**.
- Human rational behavior may lead to four outcomes, i.e., **comparisons** (Rao and Kupper, 1967; Davidson, 1970)

$$\mathcal{O} = \{\succ, \prec, \asymp, \parallel\}.$$

- For arms i, j , define comparison probabilities

$$Q_{ij}^o = \mathbb{P}(i \circ j), \quad o \in \mathcal{O}, \quad \sum_{o \in \mathcal{O}} Q_{ij}^o = 1.$$

- Interpretations:

- ▶ $i \succ j$: i is preferred to j ;
- ▶ $i \asymp j$: i and j are equivalent;
- ▶ $i \parallel j$: i and j are incomparable.

- Difficult (and not particularly interesting) defining solution concepts like Borda, Copeland

Comparisons

- Binary dueling forces every queried pair into one of two outcomes $\{\succ, \prec\}$, i.e., **preferences**.
- Human rational behavior may lead to four outcomes, i.e., **comparisons** (Rao and Kupper, 1967; Davidson, 1970)

$$\mathcal{O} = \{\succ, \prec, \asymp, \parallel\}.$$

- For arms i, j , define comparison probabilities

$$Q_{ij}^o = \mathbb{P}(i \circ j), \quad o \in \mathcal{O}, \quad \sum_{o \in \mathcal{O}} Q_{ij}^o = 1.$$

- Interpretations:

- ▶ $i \succ j$: i is preferred to j ;
- ▶ $i \asymp j$: i and j are equivalent;
- ▶ $i \parallel j$: i and j are incomparable.

- Difficult (and not particularly interesting) defining solution concepts like Borda, Copeland

Comparisons

- Binary dueling forces every queried pair into one of two outcomes $\{\succ, \prec\}$, i.e., **preferences**.
- Human rational behavior may lead to four outcomes, i.e., **comparisons** (Rao and Kupper, 1967; Davidson, 1970)

$$\mathcal{O} = \{\succ, \prec, \asymp, \parallel\}.$$

- For arms i, j , define comparison probabilities

$$Q_{ij}^o = \mathbb{P}(i \circ j), \quad o \in \mathcal{O}, \quad \sum_{o \in \mathcal{O}} Q_{ij}^o = 1.$$

- Interpretations:

- ▶ $i \succ j$: i is preferred to j ;
- ▶ $i \asymp j$: i and j are equivalent;
- ▶ $i \parallel j$: i and j are incomparable.

- Difficult (and not particularly interesting) defining solution concepts like Borda, Copeland

Comparisons

- Binary dueling forces every queried pair into one of two outcomes $\{\succ, \prec\}$, i.e., **preferences**.
- Human rational behavior may lead to four outcomes, i.e., **comparisons** (Rao and Kupper, 1967; Davidson, 1970)

$$\mathcal{O} = \{\succ, \prec, \asymp, \parallel\}.$$

- For arms i, j , define comparison probabilities

$$Q_{ij}^o = \mathbb{P}(i \circ j), \quad o \in \mathcal{O}, \quad \sum_{o \in \mathcal{O}} Q_{ij}^o = 1.$$

- Interpretations:

- ▶ $i \succ j$: i is preferred to j ;
- ▶ $i \asymp j$: i and j are equivalent;
- ▶ $i \parallel j$: i and j are incomparable.

- Difficult (and not particularly interesting) defining solution concepts like Borda, Copeland

Comparisons

- Binary dueling forces every queried pair into one of two outcomes $\{\succ, \prec\}$, i.e., **preferences**.
- Human rational behavior may lead to four outcomes, i.e., **comparisons** (Rao and Kupper, 1967; Davidson, 1970)

$$\mathcal{O} = \{\succ, \prec, \asymp, \parallel\}.$$

- For arms i, j , define comparison probabilities

$$Q_{ij}^o = \mathbb{P}(i \circ j), \quad o \in \mathcal{O}, \quad \sum_{o \in \mathcal{O}} Q_{ij}^o = 1.$$

- Interpretations:

- ▶ $i \succ j$: i is preferred to j ;
- ▶ $i \asymp j$: i and j are equivalent;
- ▶ $i \parallel j$: i and j are incomparable.

- Difficult (and not particularly interesting) defining solution concepts like Borda, Copeland

Comparisons

- Binary dueling forces every queried pair into one of two outcomes $\{\succ, \prec\}$, i.e., **preferences**.
- Human rational behavior may lead to four outcomes, i.e., **comparisons** (Rao and Kupper, 1967; Davidson, 1970)

$$\mathcal{O} = \{\succ, \prec, \asymp, \parallel\}.$$

- For arms i, j , define comparison probabilities

$$Q_{ij}^o = \mathbb{P}(i \circ j), \quad o \in \mathcal{O}, \quad \sum_{o \in \mathcal{O}} Q_{ij}^o = 1.$$

- Interpretations:

- ▶ $i \succ j$: i is preferred to j ;
- ▶ $i \asymp j$: i and j are equivalent;
- ▶ $i \parallel j$: i and j are incomparable.

- Difficult (and not particularly interesting) defining solution concepts like Borda, Copeland

Vector utility

- Incomparability $\parallel$ is rational and manifestation of **multiple objectives**.
- Let each arm have a latent vector utility (Drago et al., 2025; 2026a)

$$\mathbf{u}_i = (u_{i,1}, \dots, u_{i,d}) \in \mathbb{R}^d.$$

- Pairwise utility gap:

$$\delta_{ij} = \mathbf{u}_i - \mathbf{u}_j.$$

- Pareto dominance:

$$i \succeq_P j \iff u_{i,\ell} \geq u_{j,\ell} \quad \forall \ell \in [d].$$

- Strict dominance $i \succ_P j$ holds when $i \succeq_P j$ and $\mathbf{u}_i \neq \mathbf{u}_j$.
- Incomparability arises naturally when each arm is better on a different objective.

Vector utility

- Incomparability $\parallel$ is rational and manifestation of **multiple objectives**.
- Let each arm have a latent vector utility (Drago et al., 2025; 2026a)

$$\mathbf{u}_i = (u_{i,1}, \dots, u_{i,d}) \in \mathbb{R}^d.$$

- Pairwise utility gap:

$$\delta_{ij} = \mathbf{u}_i - \mathbf{u}_j.$$

- Pareto dominance:

$$i \succeq_P j \iff u_{i,\ell} \geq u_{j,\ell} \quad \forall \ell \in [d].$$

- Strict dominance $i \succ_P j$ holds when $i \succeq_P j$ and $\mathbf{u}_i \neq \mathbf{u}_j$.
- Incomparability arises naturally when each arm is better on a different objective.

Vector utility

- Incomparability $\parallel$ is rational and manifestation of **multiple objectives**.
- Let each arm have a latent vector utility (Drago et al., 2025; 2026a)

$$\mathbf{u}_i = (u_{i,1}, \dots, u_{i,d}) \in \mathbb{R}^d.$$

- Pairwise utility gap:

$$\delta_{ij} = \mathbf{u}_i - \mathbf{u}_j.$$

- Pareto dominance:

$$i \succeq_P j \iff u_{i,\ell} \geq u_{j,\ell} \quad \forall \ell \in [d].$$

- Strict dominance $i \succ_P j$ holds when $i \succeq_P j$ and $\mathbf{u}_i \neq \mathbf{u}_j$.
- Incomparability arises naturally when each arm is better on a different objective.

Vector utility

- Incomparability $\parallel$ is rational and manifestation of **multiple objectives**.
- Let each arm have a latent vector utility (Drago et al., 2025; 2026a)

$$\mathbf{u}_i = (u_{i,1}, \dots, u_{i,d}) \in \mathbb{R}^d.$$

- Pairwise utility gap:

$$\delta_{ij} = \mathbf{u}_i - \mathbf{u}_j.$$

- Pareto dominance:

$$i \succeq_P j \iff u_{i,\ell} \geq u_{j,\ell} \quad \forall \ell \in [d].$$

- Strict dominance $i \succ_P j$ holds when $i \succeq_P j$ and $\mathbf{u}_i \neq \mathbf{u}_j$.
- Incomparability arises naturally when each arm is better on a different objective.

Vector utility

- Incomparability $\parallel$ is rational and manifestation of **multiple objectives**.
- Let each arm have a latent vector utility (Drago et al., 2025; 2026a)

$$\mathbf{u}_i = (u_{i,1}, \dots, u_{i,d}) \in \mathbb{R}^d.$$

- Pairwise utility gap:

$$\delta_{ij} = \mathbf{u}_i - \mathbf{u}_j.$$

- Pareto dominance:

$$i \succeq_P j \iff u_{i,\ell} \geq u_{j,\ell} \quad \forall \ell \in [d].$$

- Strict dominance $i \succ_P j$ holds when $i \succeq_P j$ and $\mathbf{u}_i \neq \mathbf{u}_j$.
- Incomparability arises naturally when each arm is better on a different objective.

Vector utility

- Incomparability $\parallel$ is rational and manifestation of **multiple objectives**.
- Let each arm have a latent vector utility (Drago et al., 2025; 2026a)

$$\mathbf{u}_i = (u_{i,1}, \dots, u_{i,d}) \in \mathbb{R}^d.$$

- Pairwise utility gap:

$$\delta_{ij} = \mathbf{u}_i - \mathbf{u}_j.$$

- Pareto dominance:

$$i \succeq_P j \iff u_{i,\ell} \geq u_{j,\ell} \quad \forall \ell \in [d].$$

- Strict dominance $i \succ_P j$ holds when $i \succeq_P j$ and $\mathbf{u}_i \neq \mathbf{u}_j$.
- Incomparability arises naturally when each arm is better on a different objective.

Four outcomes from vector gaps

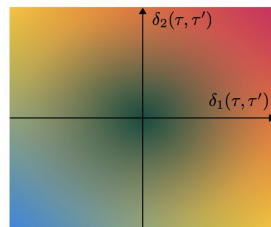
- A natural interpretation is:

$$\begin{aligned} i \succ j &\Leftarrow \delta_{ij} \in \mathbb{R}_+^d \setminus \{0\}, & \text{all objectives favor } i, \\ i \prec j &\Leftarrow -\delta_{ij} \in \mathbb{R}_+^d \setminus \{0\}, & \text{all objectives favor } j, \\ i \asymp j &\Leftarrow \delta_{ij} \approx 0, & \text{no meaningful difference,} \\ i \parallel j &\Leftarrow \text{mixed signs in } \delta_{ij}, & \text{objective conflict.} \end{aligned}$$

- With noisy human feedback, replace these regions by probabilities

$$f(o \mid \delta_{ij}), \quad o \in \{\succ, \prec, \asymp, \parallel\}.$$

- The modeling challenge is to make $\asymp$ likely near zero and $\parallel$ likely when objectives conflict, without losing statistical tractability.
- We want to generalize BT



■ $\succ$ (Direct Preference) ■ $\asymp$ (Indifference)
■ $\prec$ (Inverse Preference) ■ $\parallel$ (Incomparability)

Four outcomes from vector gaps

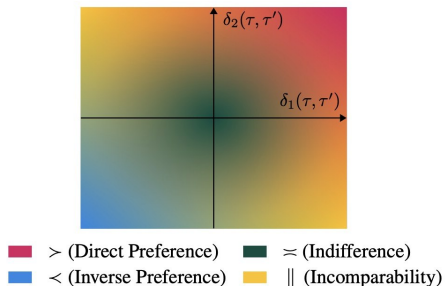
- A natural interpretation is:

$$\begin{aligned} i \succ j &\Leftarrow \delta_{ij} \in \mathbb{R}_+^d \setminus \{0\}, & \text{all objectives favor } i, \\ i \prec j &\Leftarrow -\delta_{ij} \in \mathbb{R}_+^d \setminus \{0\}, & \text{all objectives favor } j, \\ i \asymp j &\Leftarrow \delta_{ij} \approx 0, & \text{no meaningful difference,} \\ i \parallel j &\Leftarrow \text{mixed signs in } \delta_{ij}, & \text{objective conflict.} \end{aligned}$$

- With noisy human feedback, replace these regions by probabilities

$$f(o \mid \delta_{ij}), \quad o \in \{\succ, \prec, \asymp, \parallel\}.$$

- The modeling challenge is to make $\asymp$ likely near zero and $\parallel$ likely when objectives conflict, without losing statistical tractability.
- We want to generalize BT



Four outcomes from vector gaps

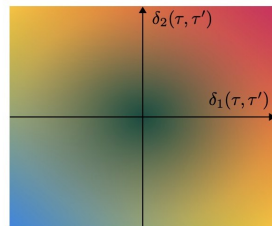
- A natural interpretation is:

$$\begin{aligned} i \succ j &\Leftarrow \delta_{ij} \in \mathbb{R}_+^d \setminus \{0\}, & \text{all objectives favor } i, \\ i \prec j &\Leftarrow -\delta_{ij} \in \mathbb{R}_+^d \setminus \{0\}, & \text{all objectives favor } j, \\ i \asymp j &\Leftarrow \delta_{ij} \approx 0, & \text{no meaningful difference,} \\ i \parallel j &\Leftarrow \text{mixed signs in } \delta_{ij}, & \text{objective conflict.} \end{aligned}$$

- With noisy human feedback, replace these regions by probabilities

$$f(o \mid \delta_{ij}), \quad o \in \{\succ, \prec, \asymp, \parallel\}.$$

- The modeling challenge is to make $\asymp$ likely near zero and $\parallel$ likely when objectives conflict, without losing statistical tractability.
- We want to generalize BT



■ $\succ$ (Direct Preference) ■ $\asymp$ (Indifference)
■ $\prec$ (Inverse Preference) ■ $\parallel$ (Incomparability)

Four outcomes from vector gaps

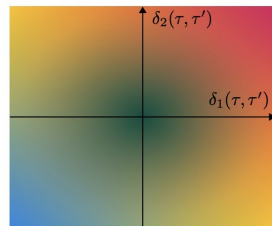
- A natural interpretation is:

$$\begin{aligned} i \succ j &\Leftarrow \delta_{ij} \in \mathbb{R}_+^d \setminus \{0\}, & \text{all objectives favor } i, \\ i \prec j &\Leftarrow -\delta_{ij} \in \mathbb{R}_+^d \setminus \{0\}, & \text{all objectives favor } j, \\ i \asymp j &\Leftarrow \delta_{ij} \approx 0, & \text{no meaningful difference,} \\ i \parallel j &\Leftarrow \text{mixed signs in } \delta_{ij}, & \text{objective conflict.} \end{aligned}$$

- With noisy human feedback, replace these regions by probabilities

$$f(o \mid \delta_{ij}), \quad o \in \{\succ, \prec, \asymp, \parallel\}.$$

- The modeling challenge is to make $\asymp$ likely near zero and $\parallel$ likely when objectives conflict, without losing statistical tractability.
- We want to generalize BT



■ $\succ$ (Direct Preference) ■ $\asymp$ (Indifference)
■ $\prec$ (Inverse Preference) ■ $\parallel$ (Incomparability)

Negative result

- Let $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{O})$ map a vector gap δ to outcome probabilities (Drago et al., 2026a).

- The per-sample negative log-likelihood is

$$\ell_o(\delta) = -\log f(o \mid \delta), \quad o \in \mathcal{O}.$$

- Natural desiderata require (Drago et al., 2026a):

- ▶ $\succ$ dominates along $a\mathbf{1}$ as $a \rightarrow +\infty$;
- ▶ $\prec$ dominates along $a\mathbf{1}$ as $a \rightarrow -\infty$;
- ▶ $\asymp$ dominates near $\mathbf{0}$;
- ▶ $\parallel$ dominates along conflicting directions such as $(a, -a, 0, \dots, 0)$.

Prop. No four-outcome link satisfying these rationality requirements has ℓ_o convex in δ for every outcome o .

- Comparison learning cannot inherit the full convex convenience of scalar BT/probit models.

Negative result

- Let $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{O})$ map a vector gap δ to outcome probabilities (Drago et al., 2026a).
- The per-sample negative log-likelihood is

$$\ell_o(\delta) = -\log f(o \mid \delta), \quad o \in \mathcal{O}.$$

- Natural desiderata require (Drago et al., 2026a):
 - ▶ $\succ$ dominates along $a\mathbf{1}$ as $a \rightarrow +\infty$;
 - ▶ $\prec$ dominates along $a\mathbf{1}$ as $a \rightarrow -\infty$;
 - ▶ $\asymp$ dominates near $\mathbf{0}$;
 - ▶ $\parallel$ dominates along conflicting directions such as $(a, -a, 0, \dots, 0)$.

Prop. No four-outcome link satisfying these rationality requirements has ℓ_o convex in δ for every outcome o .

- Comparison learning cannot inherit the full convex convenience of scalar BT/probit models.

Negative result

- Let $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{O})$ map a vector gap δ to outcome probabilities (Drago et al., 2026a).
- The per-sample negative log-likelihood is

$$\ell_o(\delta) = -\log f(o \mid \delta), \quad o \in \mathcal{O}.$$

- Natural desiderata require (Drago et al., 2026a):
 - ▶ $\succ$ dominates along $a\mathbf{1}$ as $a \rightarrow +\infty$;
 - ▶ $\prec$ dominates along $a\mathbf{1}$ as $a \rightarrow -\infty$;
 - ▶ $\asymp$ dominates near $\mathbf{0}$;
 - ▶ $\parallel$ dominates along conflicting directions such as $(a, -a, 0, \dots, 0)$.

Prop. No four-outcome link satisfying these rationality requirements has ℓ_o convex in δ for every outcome o .

- Comparison learning cannot inherit the full convex convenience of scalar BT/probit models.

Negative result

- Let $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{O})$ map a vector gap δ to outcome probabilities (Drago et al., 2026a).
- The per-sample negative log-likelihood is

$$\ell_o(\delta) = -\log f(o \mid \delta), \quad o \in \mathcal{O}.$$

- Natural desiderata require (Drago et al., 2026a):
 - ▶ $\succ$ dominates along $a\mathbf{1}$ as $a \rightarrow +\infty$;
 - ▶ $\prec$ dominates along $a\mathbf{1}$ as $a \rightarrow -\infty$;
 - ▶ $\asymp$ dominates near $\mathbf{0}$;
 - ▶ $\parallel$ dominates along conflicting directions such as $(a, -a, 0, \dots, 0)$.

Prop. No four-outcome link satisfying these rationality requirements has ℓ_o convex in δ for every outcome o .

- Comparison learning cannot inherit the full convex convenience of scalar BT/probit models.

Negative result

- Let $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{O})$ map a vector gap δ to outcome probabilities (Drago et al., 2026a).
- The per-sample negative log-likelihood is

$$\ell_o(\delta) = -\log f(o \mid \delta), \quad o \in \mathcal{O}.$$

- Natural desiderata require (Drago et al., 2026a):
 - ▶ $\succ$ dominates along $a\mathbf{1}$ as $a \rightarrow +\infty$;
 - ▶ $\prec$ dominates along $a\mathbf{1}$ as $a \rightarrow -\infty$;
 - ▶ $\asymp$ dominates near $\mathbf{0}$;
 - ▶ $\parallel$ dominates along conflicting directions such as $(a, -a, 0, \dots, 0)$.

Prop. No four-outcome link satisfying these rationality requirements has ℓ_o convex in δ for every outcome o .

- Comparison learning cannot inherit the full convex convenience of scalar BT/probit models.

Negative result

- Let $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{O})$ map a vector gap δ to outcome probabilities (Drago et al., 2026a).
- The per-sample negative log-likelihood is

$$\ell_o(\delta) = -\log f(o \mid \delta), \quad o \in \mathcal{O}.$$

- Natural desiderata require (Drago et al., 2026a):
 - ▶ $\succ$ dominates along $a\mathbf{1}$ as $a \rightarrow +\infty$;
 - ▶ $\prec$ dominates along $a\mathbf{1}$ as $a \rightarrow -\infty$;
 - ▶ $\asymp$ dominates near $\mathbf{0}$;
 - ▶ $\parallel$ dominates along conflicting directions such as $(a, -a, 0, \dots, 0)$.

Prop. No four-outcome link satisfying these rationality requirements has ℓ_o convex in δ for every outcome o .

- Comparison learning cannot inherit the full convex convenience of scalar BT/probit models.

Negative result

- Let $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{O})$ map a vector gap δ to outcome probabilities (Drago et al., 2026a).
- The per-sample negative log-likelihood is

$$\ell_o(\delta) = -\log f(o \mid \delta), \quad o \in \mathcal{O}.$$

- Natural desiderata require (Drago et al., 2026a):
 - ▶ $\succ$ dominates along $a\mathbf{1}$ as $a \rightarrow +\infty$;
 - ▶ $\prec$ dominates along $a\mathbf{1}$ as $a \rightarrow -\infty$;
 - ▶ $\asymp$ dominates near $\mathbf{0}$;
 - ▶ $\parallel$ dominates along conflicting directions such as $(a, -a, 0, \dots, 0)$.

Prop. No four-outcome link satisfying these rationality requirements has ℓ_o convex in δ for every outcome o .

- Comparison learning cannot inherit the full convex convenience of scalar BT/probit models.

Negative result

- Let $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{O})$ map a vector gap δ to outcome probabilities (Drago et al., 2026a).
- The per-sample negative log-likelihood is

$$\ell_o(\delta) = -\log f(o \mid \delta), \quad o \in \mathcal{O}.$$

- Natural desiderata require (Drago et al., 2026a):
 - ▶ $\succ$ dominates along $a\mathbf{1}$ as $a \rightarrow +\infty$;
 - ▶ $\prec$ dominates along $a\mathbf{1}$ as $a \rightarrow -\infty$;
 - ▶ $\asymp$ dominates near $\mathbf{0}$;
 - ▶ $\parallel$ dominates along conflicting directions such as $(a, -a, 0, \dots, 0)$.

Prop. No four-outcome link satisfying these rationality requirements has ℓ_o convex in δ for every outcome o .

- Comparison learning cannot inherit the full convex convenience of scalar BT/probit models.

Negative result

- Let $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{O})$ map a vector gap δ to outcome probabilities (Drago et al., 2026a).
- The per-sample negative log-likelihood is

$$\ell_o(\delta) = -\log f(o \mid \delta), \quad o \in \mathcal{O}.$$

- Natural desiderata require (Drago et al., 2026a):
 - ▶ $\succ$ dominates along $a\mathbf{1}$ as $a \rightarrow +\infty$;
 - ▶ $\prec$ dominates along $a\mathbf{1}$ as $a \rightarrow -\infty$;
 - ▶ $\asymp$ dominates near $\mathbf{0}$;
 - ▶ $\parallel$ dominates along conflicting directions such as $(a, -a, 0, \dots, 0)$.

Prop. No four-outcome link satisfying these rationality requirements has ℓ_o convex in δ for every outcome o .

- Comparison learning cannot inherit the full convex convenience of scalar BT/probit models.

MOBT: a multi-objective Bradley–Terry proposal

- Let

$$\text{mean}(\delta) = \frac{1}{d} \mathbf{1}^\top \delta, \quad \text{std}(\delta) = \|\delta - \bar{\delta} \mathbf{1}\|_2.$$

- Define four scores (Drago et al., 2026a):

$$\begin{aligned} h_{\succ}(\delta) &= \text{mean}(\delta), & h_{\prec}(\delta) &= -\text{mean}(\delta), \\ h_{\asymp}(\delta) &= \alpha, & h_{\parallel}(\delta) &= \beta + \sqrt{d} \text{std}(\delta). \end{aligned}$$

- Outcome probabilities are softmax scores:

$$f(o \mid \delta) = \frac{\exp(h_o(\delta))}{\sum_{o' \in \mathcal{O}} \exp(h_{o'}(\delta))}.$$

- Interpretation:

- ▶ the mean gap $\text{mean}(\delta)$ drives preference direction;
- ▶ high dispersion $\text{std}(\delta)$ across objectives favors incomparability.
- ▶ small norm favors indifference;

- Reduces to BT when $\beta = -\infty$ and $\alpha = -\infty$.

MOBT: a multi-objective Bradley–Terry proposal

- Let

$$\text{mean}(\delta) = \frac{1}{d} \mathbf{1}^\top \delta, \quad \text{std}(\delta) = \|\delta - \bar{\delta} \mathbf{1}\|_2.$$

- Define four scores (Drago et al., 2026a):

$$\begin{aligned} h_{\succ}(\delta) &= \text{mean}(\delta), & h_{\prec}(\delta) &= -\text{mean}(\delta), \\ h_{\asymp}(\delta) &= \alpha, & h_{\parallel}(\delta) &= \beta + \sqrt{d} \text{std}(\delta). \end{aligned}$$

- Outcome probabilities are softmax scores:

$$f(o \mid \delta) = \frac{\exp(h_o(\delta))}{\sum_{o' \in \mathcal{O}} \exp(h_{o'}(\delta))}.$$

- Interpretation:

- ▶ the mean gap $\text{mean}(\delta)$ drives preference direction;
- ▶ high dispersion $\text{std}(\delta)$ across objectives favors incomparability.
- ▶ small norm favors indifference;

- Reduces to BT when $\beta = -\infty$ and $\alpha = -\infty$.

MOBT: a multi-objective Bradley–Terry proposal

- Let

$$\text{mean}(\delta) = \frac{1}{d} \mathbf{1}^\top \delta, \quad \text{std}(\delta) = \|\delta - \bar{\delta} \mathbf{1}\|_2.$$

- Define four scores (Drago et al., 2026a):

$$\begin{aligned} h_{\succ}(\delta) &= \text{mean}(\delta), & h_{\prec}(\delta) &= -\text{mean}(\delta), \\ h_{\asymp}(\delta) &= \alpha, & h_{\parallel}(\delta) &= \beta + \sqrt{d} \text{std}(\delta). \end{aligned}$$

- Outcome probabilities are softmax scores:

$$f(o \mid \delta) = \frac{\exp(h_o(\delta))}{\sum_{o' \in \mathcal{O}} \exp(h_{o'}(\delta))}.$$

- Interpretation:

- ▶ the mean gap $\text{mean}(\delta)$ drives preference direction;
- ▶ high dispersion $\text{std}(\delta)$ across objectives favors incomparability.
- ▶ small norm favors indifference;

- Reduces to BT when $\beta = -\infty$ and $\alpha = -\infty$.

MOBT: a multi-objective Bradley–Terry proposal

- Let

$$\text{mean}(\delta) = \frac{1}{d} \mathbf{1}^\top \delta, \quad \text{std}(\delta) = \|\delta - \bar{\delta} \mathbf{1}\|_2.$$

- Define four scores (Drago et al., 2026a):

$$\begin{aligned} h_{\succ}(\delta) &= \text{mean}(\delta), & h_{\prec}(\delta) &= -\text{mean}(\delta), \\ h_{\asymp}(\delta) &= \alpha, & h_{\parallel}(\delta) &= \beta + \sqrt{d} \text{std}(\delta). \end{aligned}$$

- Outcome probabilities are softmax scores:

$$f(o \mid \delta) = \frac{\exp(h_o(\delta))}{\sum_{o' \in \mathcal{O}} \exp(h_{o'}(\delta))}.$$

- Interpretation:

- ▶ the mean gap $\text{mean}(\delta)$ drives preference direction;
- ▶ high dispersion $\text{std}(\delta)$ across objectives favors incomparability.
- ▶ small norm favors indifference;

- Reduces to BT when $\beta = -\infty$ and $\alpha = -\infty$.

MOBT: a multi-objective Bradley–Terry proposal

- Let

$$\text{mean}(\delta) = \frac{1}{d} \mathbf{1}^\top \delta, \quad \text{std}(\delta) = \|\delta - \bar{\delta} \mathbf{1}\|_2.$$

- Define four scores (Drago et al., 2026a):

$$\begin{aligned} h_{\succ}(\delta) &= \text{mean}(\delta), & h_{\prec}(\delta) &= -\text{mean}(\delta), \\ h_{\asymp}(\delta) &= \alpha, & h_{\parallel}(\delta) &= \beta + \sqrt{d} \text{std}(\delta). \end{aligned}$$

- Outcome probabilities are softmax scores:

$$f(o \mid \delta) = \frac{\exp(h_o(\delta))}{\sum_{o' \in \mathcal{O}} \exp(h_{o'}(\delta))}.$$

- Interpretation:

- ▶ the mean gap $\text{mean}(\delta)$ drives preference direction;
- ▶ high dispersion $\text{std}(\delta)$ across objectives favors incomparability.
- ▶ small norm favors indifference;

- Reduces to BT when $\beta = -\infty$ and $\alpha = -\infty$.

MOBT: a multi-objective Bradley–Terry proposal

- Let

$$\text{mean}(\delta) = \frac{1}{d} \mathbf{1}^\top \delta, \quad \text{std}(\delta) = \|\delta - \bar{\delta} \mathbf{1}\|_2.$$

- Define four scores (Drago et al., 2026a):

$$\begin{aligned} h_{\succ}(\delta) &= \text{mean}(\delta), & h_{\prec}(\delta) &= -\text{mean}(\delta), \\ h_{\asymp}(\delta) &= \alpha, & h_{\parallel}(\delta) &= \beta + \sqrt{d} \text{std}(\delta). \end{aligned}$$

- Outcome probabilities are softmax scores:

$$f(o \mid \delta) = \frac{\exp(h_o(\delta))}{\sum_{o' \in \mathcal{O}} \exp(h_{o'}(\delta))}.$$

- Interpretation:

- ▶ the mean gap $\text{mean}(\delta)$ drives preference direction;
- ▶ high dispersion $\text{std}(\delta)$ across objectives favors incomparability.
- ▶ small norm favors indifference;

- Reduces to BT when $\beta = -\infty$ and $\alpha = -\infty$.

MOBT: a multi-objective Bradley–Terry proposal

- Let

$$\text{mean}(\delta) = \frac{1}{d} \mathbf{1}^\top \delta, \quad \text{std}(\delta) = \|\delta - \bar{\delta} \mathbf{1}\|_2.$$

- Define four scores (Drago et al., 2026a):

$$\begin{aligned} h_{\succ}(\delta) &= \text{mean}(\delta), & h_{\prec}(\delta) &= -\text{mean}(\delta), \\ h_{\asymp}(\delta) &= \alpha, & h_{\parallel}(\delta) &= \beta + \sqrt{d} \text{std}(\delta). \end{aligned}$$

- Outcome probabilities are softmax scores:

$$f(o \mid \delta) = \frac{\exp(h_o(\delta))}{\sum_{o' \in \mathcal{O}} \exp(h_{o'}(\delta))}.$$

- Interpretation:

- ▶ the mean gap $\text{mean}(\delta)$ drives preference direction;
- ▶ high dispersion $\text{std}(\delta)$ across objectives favors incomparability.
- ▶ small norm favors indifference;

- Reduces to BT when $\beta = -\infty$ and $\alpha = -\infty$.

MOBT: a multi-objective Bradley–Terry proposal

- Let

$$\text{mean}(\delta) = \frac{1}{d} \mathbf{1}^\top \delta, \quad \text{std}(\delta) = \|\delta - \bar{\delta} \mathbf{1}\|_2.$$

- Define four scores (Drago et al., 2026a):

$$\begin{aligned} h_{\succ}(\delta) &= \text{mean}(\delta), & h_{\prec}(\delta) &= -\text{mean}(\delta), \\ h_{\asymp}(\delta) &= \alpha, & h_{\parallel}(\delta) &= \beta + \sqrt{d} \text{std}(\delta). \end{aligned}$$

- Outcome probabilities are softmax scores:

$$f(o \mid \delta) = \frac{\exp(h_o(\delta))}{\sum_{o' \in \mathcal{O}} \exp(h_{o'}(\delta))}.$$

- Interpretation:

- ▶ the mean gap $\text{mean}(\delta)$ drives preference direction;
- ▶ high dispersion $\text{std}(\delta)$ across objectives favors incomparability.
- ▶ small norm favors indifference;

- Reduces to BT when $\beta = -\infty$ and $\alpha = -\infty$.

Open Questions

- Defying a sensible notion of **optimum** and **regret**
 - ▶ Quantify the distance from the **Pareto front**
 - ▶ Focus on **scalarizations**?
- Identifiability issues (|| includes many scenarios)
- - ▶ Cannot recover all the objectives

Open Questions

- Defying a sensible notion of **optimum** and **regret**
 - ▶ Quantify the distance from the **Pareto front**
 - ▶ Focus on **scalarizations**?
- Identifiability issues ($\|\cdot\|$ includes many scenarios)
- - ▶ Cannot recover all the objectives

Open Questions

- Defying a sensible notion of **optimum** and **regret**
 - ▶ Quantify the distance from the **Pareto front**
 - ▶ Focus on **scalarizations**?
- Identifiability issues (|| includes many scenarios)
- - ▶ Cannot recover all the objectives

Open Questions

- Defying a sensible notion of **optimum** and **regret**
 - ▶ Quantify the distance from the **Pareto front**
 - ▶ Focus on **scalarizations**?
- Identifiability issues (|| includes many scenarios)
 - ▶ Cannot recover all the objectives

Open Questions

- Defying a sensible notion of **optimum** and **regret**
 - ▶ Quantify the distance from the **Pareto front**
 - ▶ Focus on **scalarizations**?
- Identifiability issues (|| includes many scenarios)
- ▶ Cannot recover all the objectives

Open Questions

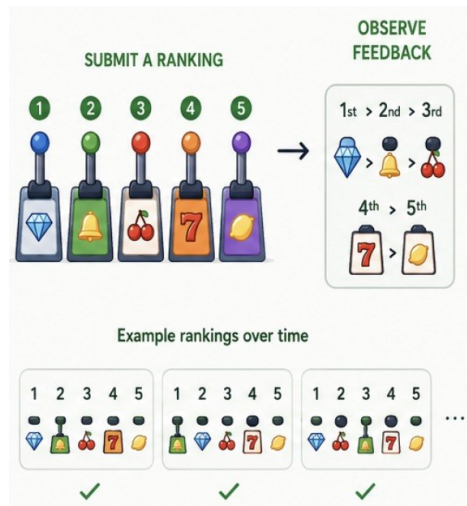
- Defying a sensible notion of **optimum** and **regret**
 - ▶ Quantify the distance from the **Pareto front**
 - ▶ Focus on **scalarizations**?
- Identifiability issues (|| includes many scenarios)
- - ▶ Cannot recover all the objectives

Outline

- 1 Multi-armed bandits
- 2 Dueling bandits
 - Preferences
 - Comparisons
- 3 Bandits with ranking feedback

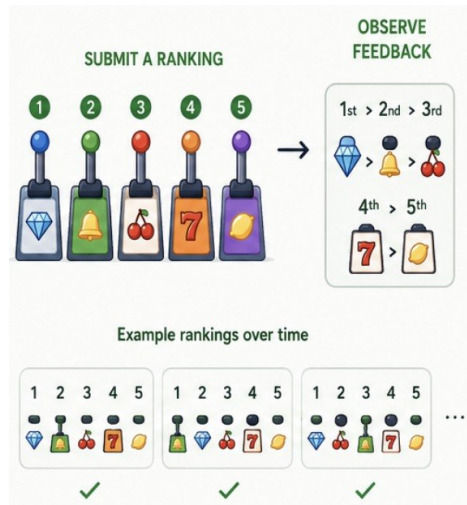
Why ranking feedback?

- Pairwise comparisons ask for one relative judgement at a time.
- If a human sees several alternatives, a **ranking** can be more informative (Saha and Gopalan, 2019a;b; Migali et al., 2026b;a).
- At round $t \in [T]$:
 - the learner chooses a (multi)set (Saha and Gopalan, 2019b; Migali et al., 2026a)
 - a latent ranking $R_t \in \mathbb{S}(S_t)$ of the items in S_t is sampled.
 - the learner observes a censored version $C_m(R_t)$.
- $\mathbb{S}(S_t)$ is the set of permutations of the S_t elements



Why ranking feedback?

- Pairwise comparisons ask for one relative judgement at a time.
- If a human sees several alternatives, a **ranking** can be more informative (Saha and Gopalan, 2019a;b; Migali et al., 2026b;a).
- At round $t \in [T]$:
 - the learner chooses a (multi)set (Saha and Gopalan, 2019b; Migali et al., 2026a)
 - $S_t \subseteq [k], \quad |S_t| = n \leq k.$
 - a latent ranking $R_t \in \mathbb{S}(S_t)$ of the items in S_t is sampled.
 - the learner observes a censored version $C_m(R_t)$.
- $\mathbb{S}(S_t)$ is the set of permutations of the S_t elements

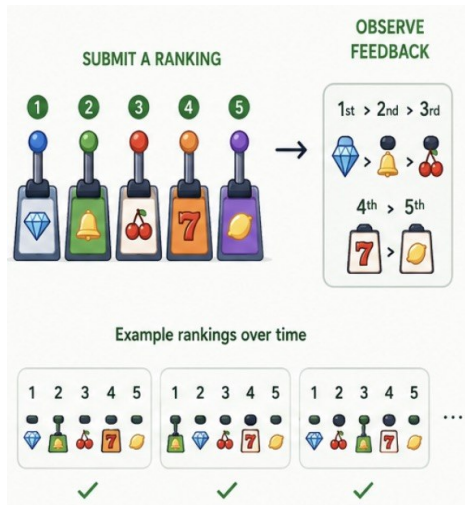


Why ranking feedback?

- Pairwise comparisons ask for one relative judgement at a time.
- If a human sees several alternatives, a **ranking** can be more informative (Saha and Gopalan, 2019a;b; Migali et al., 2026b;a).
- At round $t \in [T]$:
 - 1 the learner chooses a (multi)set (Saha and Gopalan, 2019b; Migali et al., 2026a)

$$S_t \subseteq [k], \quad |S_t| = n \leq k.$$

- 2 a latent ranking $R_t \in \mathbb{S}(S_t)$ of the items in S_t is sampled.
 - 3 the learner observes a censored version $C_m(R_t)$.
- $\mathbb{S}(S_t)$ is the set of permutations of the S_t elements

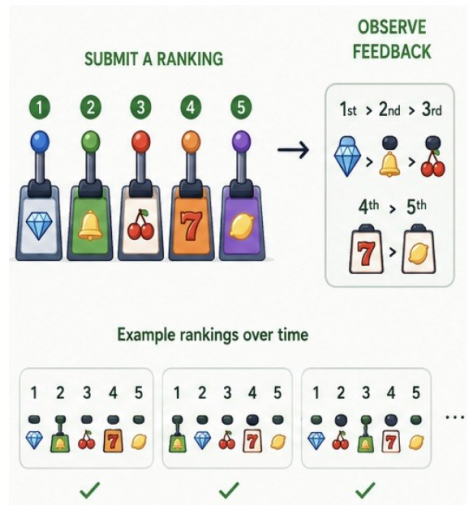


Why ranking feedback?

- Pairwise comparisons ask for one relative judgement at a time.
- If a human sees several alternatives, a **ranking** can be more informative (Saha and Gopalan, 2019a;b; Migali et al., 2026b;a).
- At round $t \in [T]$:
 - 1 the learner chooses a (multi)set (Saha and Gopalan, 2019b; Migali et al., 2026a)

$$S_t \subseteq [k], \quad |S_t| = n \leq k.$$

- 2 a latent ranking $R_t \in \mathbb{S}(S_t)$ of the items in S_t is sampled.
 - 3 the learner observes a censored version $C_m(R_t)$.
- $\mathbb{S}(S_t)$ is the set of permutations of the S_t elements

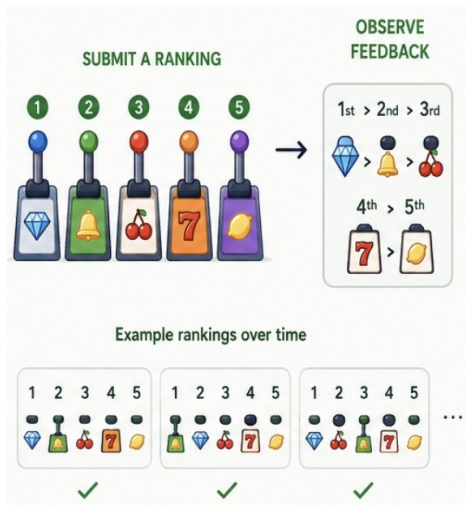


Why ranking feedback?

- Pairwise comparisons ask for one relative judgement at a time.
- If a human sees several alternatives, a **ranking** can be more informative (Saha and Gopalan, 2019a;b; Migali et al., 2026b;a).
- At round $t \in [T]$:
 - 1 the learner chooses a (multi)set (Saha and Gopalan, 2019b; Migali et al., 2026a)

$$S_t \subseteq [k], \quad |S_t| = n \leq k.$$

- 2 a latent ranking $R_t \in \mathbb{S}(S_t)$ of the items in S_t is sampled.
 - 3 the learner observes a censored version $C_m(R_t)$.
- $\mathbb{S}(S_t)$ is the set of permutations of the S_t elements

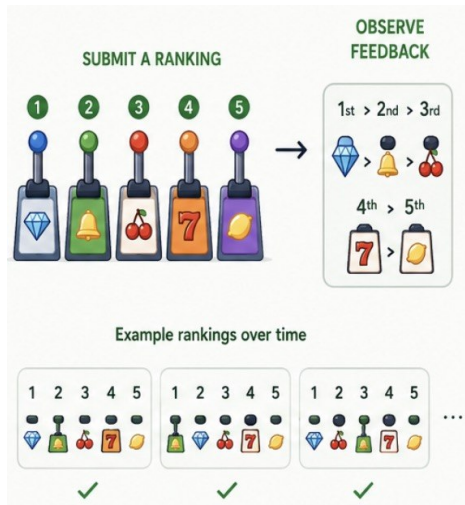


Why ranking feedback?

- Pairwise comparisons ask for one relative judgement at a time.
- If a human sees several alternatives, a **ranking** can be more informative (Saha and Gopalan, 2019a;b; Migali et al., 2026b;a).
- At round $t \in [T]$:
 - 1 the learner chooses a (multi)set (Saha and Gopalan, 2019b; Migali et al., 2026a)

$$S_t \subseteq [k], \quad |S_t| = n \leq k.$$

- 2 a latent ranking $R_t \in \mathbb{S}(S_t)$ of the items in S_t is sampled.
 - 3 the learner observes a censored version $C_m(R_t)$.
- $\mathbb{S}(S_t)$ is the set of permutations of the S_t elements

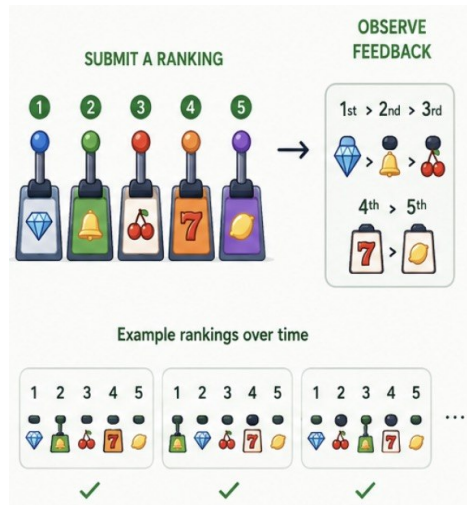


Why ranking feedback?

- Pairwise comparisons ask for one relative judgement at a time.
- If a human sees several alternatives, a **ranking** can be more informative (Saha and Gopalan, 2019a;b; Migali et al., 2026b;a).
- At round $t \in [T]$:
 - 1 the learner chooses a (multi)set (Saha and Gopalan, 2019b; Migali et al., 2026a)

$$S_t \subseteq [k], \quad |S_t| = n \leq k.$$

- 2 a latent ranking $R_t \in \mathbb{S}(S_t)$ of the items in S_t is sampled.
 - 3 the learner observes a censored version $C_m(R_t)$.
- $\mathbb{S}(S_t)$ is the set of permutations of the S_t elements



Why ranking feedback?

- Important cases:

$m = 1$: winner feedback,
 $1 < m < k$: top- m feedback,
 $m = k$: full ranking of the displayed set.

- Dueling bandits are recovered when $k = 2$ and $m = 1$.

Why ranking feedback?

- Important cases:

$m = 1$: winner feedback,
 $1 < m < k$: top- m feedback,
 $m = k$: full ranking of the displayed set.

- Dueling bandits are recovered when $k = 2$ and $m = 1$.

Random Utility Model

- Infeasible to define a general probability over $\mathbb{S}(S_t)$ since $|\mathbb{S}(S_t)| = n!$
- Assume a scalar utility $u_i \in \mathbb{R}$ for every arm i
- Sample noise $\epsilon_i \sim \Psi$ for every $i \in S_t$
- Define the utilities $U_i = u_i + \epsilon_i$
- Ranking R_t is such that:

$$U_{R_t^{-1}(1)} > U_{R_t^{-1}(2)} > \dots U_{R_t^{-1}(n)}$$

Random Utility Model

- Infeasible to define a general probability over $\mathbb{S}(S_t)$ since $|\mathbb{S}(S_t)| = n!$
- Assume a scalar utility $u_i \in \mathbb{R}$ for every arm i
- Sample noise $\epsilon_i \sim \Psi$ for every $i \in S_t$
- Define the utilities $U_i = u_i + \epsilon_i$
- Ranking R_t is such that:

$$U_{R_t^{-1}(1)} > U_{R_t^{-1}(2)} > \dots U_{R_t^{-1}(n)}$$

Random Utility Model

- Infeasible to define a general probability over $\mathbb{S}(S_t)$ since $|\mathbb{S}(S_t)| = n!$
- Assume a scalar utility $u_i \in \mathbb{R}$ for every arm i
- Sample noise $\epsilon_i \sim \Psi$ for every $i \in S_t$
- Define the utilities $U_i = u_i + \epsilon_i$
- Ranking R_t is such that:

$$U_{R_t^{-1}(1)} > U_{R_t^{-1}(2)} > \dots U_{R_t^{-1}(n)}$$

Random Utility Model

- Infeasible to define a general probability over $\mathbb{S}(S_t)$ since $|\mathbb{S}(S_t)| = n!$
- Assume a scalar utility $u_i \in \mathbb{R}$ for every arm i
- Sample noise $\epsilon_i \sim \Psi$ for every $i \in S_t$
- Define the utilities $U_i = u_i + \epsilon_i$
- Ranking R_t is such that:

$$U_{R_t^{-1}(1)} > U_{R_t^{-1}(2)} > \dots U_{R_t^{-1}(n)}$$

Random Utility Model

- Infeasible to define a general probability over $\mathbb{S}(S_t)$ since $|\mathbb{S}(S_t)| = n!$
- Assume a scalar utility $u_i \in \mathbb{R}$ for every arm i
- Sample noise $\epsilon_i \sim \Psi$ for every $i \in S_t$
- Define the utilities $U_i = u_i + \epsilon_i$
- Ranking R_t is such that:

$$U_{R_t^{-1}(1)} > U_{R_t^{-1}(2)} > \dots U_{R_t^{-1}(n)}$$

Random Utility Model

- Infeasible to define a general probability over $\mathbb{S}(S_t)$ since $|\mathbb{S}(S_t)| = n!$
- Assume a scalar utility $u_i \in \mathbb{R}$ for every arm i
- Sample noise $\epsilon_i \sim \Psi$ for every $i \in S_t$
- Define the utilities $U_i = u_i + \epsilon_i$
- Ranking R_t is such that:

$$U_{R_t^{-1}(1)} > U_{R_t^{-1}(2)} > \dots U_{R_t^{-1}(n)}$$

Plackett–Luce

- Each arm has a positive parameter $\theta_i > 0$, often $\theta_i = e^{u_i}$ (Luce, 1959; Plackett, 1975; Khetan and Oh, 2016).
- For a set S_t of size n , Plackett–Luce samples a ranking $r = (r_1, \dots, r_k)$ with

$$\mathbb{P}(R = r \mid S) = \prod_{a=1}^k \frac{\theta_{r_a}}{\sum_{b=a}^k \theta_{r_b}}.$$

- Pairwise marginal:

$$\mathbb{P}(i \text{ is ranked above } j) = \frac{\theta_i}{\theta_i + \theta_j}.$$

- This corresponds to the BT model when $k = 2$ and $m = 1$

Censoring mechanism

- Let R be a latent full ranking of the set S (Migali et al., 2026b)
- A **censoring mechanism** is a relation over positions,

$$\preceq_{\text{pos}} \subseteq [k] \times [k].$$

- The observed relation over items is the pullback:

$$i \preceq_{\text{obs}} j \iff \text{rank}_R(i) \preceq_{\text{pos}} \text{rank}_R(j).$$

- Examples:
 - ▶ full ranking: all position pairs are ordered;
 - ▶ winner: only position 1 is ordered above the rest;
 - ▶ top- m : positions $1, \dots, m$ are ordered and dominate unranked positions.

Censoring mechanism

- Let R be a latent full ranking of the set S (Migali et al., 2026b)
- A **censoring mechanism** is a relation over positions,

$$\preceq_{\text{pos}} \subseteq [k] \times [k].$$

- The observed relation over items is the pullback:

$$i \preceq_{\text{obs}} j \iff \text{rank}_R(i) \preceq_{\text{pos}} \text{rank}_R(j).$$

- Examples:
 - ▶ full ranking: all position pairs are ordered;
 - ▶ winner: only position 1 is ordered above the rest;
 - ▶ top- m : positions $1, \dots, m$ are ordered and dominate unranked positions.

Censoring mechanism

- Let R be a latent full ranking of the set S (Migali et al., 2026b)
- A **censoring mechanism** is a relation over positions,

$$\preceq_{\text{pos}} \subseteq [k] \times [k].$$

- The observed relation over items is the pullback:

$$i \preceq_{\text{obs}} j \iff \text{rank}_R(i) \preceq_{\text{pos}} \text{rank}_R(j).$$

- Examples:

- ▶ full ranking: all position pairs are ordered;
- ▶ winner: only position 1 is ordered above the rest;
- ▶ top- m : positions $1, \dots, m$ are ordered and dominate unranked positions.

Censoring mechanism

- Let R be a latent full ranking of the set S (Migali et al., 2026b)
- A **censoring mechanism** is a relation over positions,

$$\preceq_{\text{pos}} \subseteq [k] \times [k].$$

- The observed relation over items is the pullback:

$$i \preceq_{\text{obs}} j \iff \text{rank}_R(i) \preceq_{\text{pos}} \text{rank}_R(j).$$

- Examples:
 - ▶ full ranking: all position pairs are ordered;
 - ▶ winner: only position 1 is ordered above the rest;
 - ▶ top- m : positions $1, \dots, m$ are ordered and dominate unranked positions.

Censoring mechanism

- Let R be a latent full ranking of the set S (Migali et al., 2026b)
- A **censoring mechanism** is a relation over positions,

$$\preceq_{\text{pos}} \subseteq [k] \times [k].$$

- The observed relation over items is the pullback:

$$i \preceq_{\text{obs}} j \iff \text{rank}_R(i) \preceq_{\text{pos}} \text{rank}_R(j).$$

- Examples:
 - ▶ full ranking: all position pairs are ordered;
 - ▶ winner: only position 1 is ordered above the rest;
 - ▶ top- m : positions $1, \dots, m$ are ordered and dominate unranked positions.

Censoring mechanism

- Let R be a latent full ranking of the set S (Migali et al., 2026b)
- A **censoring mechanism** is a relation over positions,

$$\preceq_{\text{pos}} \subseteq [k] \times [k].$$

- The observed relation over items is the pullback:

$$i \preceq_{\text{obs}} j \iff \text{rank}_R(i) \preceq_{\text{pos}} \text{rank}_R(j).$$

- Examples:
 - ▶ full ranking: all position pairs are ordered;
 - ▶ winner: only position 1 is ordered above the rest;
 - ▶ top- m : positions $1, \dots, m$ are ordered and dominate unranked positions.

Censoring mechanism

- Let R be a latent full ranking of the set S (Migali et al., 2026b)
- A **censoring mechanism** is a relation over positions,

$$\preceq_{\text{pos}} \subseteq [k] \times [k].$$

- The observed relation over items is the pullback:

$$i \preceq_{\text{obs}} j \iff \text{rank}_R(i) \preceq_{\text{pos}} \text{rank}_R(j).$$

- Examples:
 - ▶ full ranking: all position pairs are ordered;
 - ▶ winner: only position 1 is ordered above the rest;
 - ▶ top- m : positions $1, \dots, m$ are ordered and dominate unranked positions.

Open Questions

- Does it make sense/Is it feasible to consider ranking **without utilities**?
- Optimum and regret definitions:

► If S_t is a set, we have the optimum (like combinatorial bandits):

$$S^* \in \arg \max_{S \in \mathcal{S}} \sum_{i \in S} \mu_i$$

► If S_t is a multiset, proposal of optimum:

$$S^* = \{i_1^*, \dots, i_n^*\} \quad \text{where } i_j^* \in \arg \max_{i \in \mathcal{I}} \mu_i$$

Open Questions

- Does it make sense/Is it feasible to consider ranking **without utilities**?
- Optimum and regret definitions:
 - ▶ If S_t is a set, we have the optimum (like combinatorial bandits):

$$S^* \in \arg \max_{S \subseteq [k], |S|=n} \sum_{i \in S} u_i$$

- ▶ If S_t is a **multiset**, proposal of optimum:

$$S^* = \{i^*, \dots, i^*\} \quad \text{where} \quad i^* \in \arg \max_{i \in [k]} u_i$$

Open Questions

- Does it make sense/Is it feasible to consider ranking **without utilities**?
- Optimum and regret definitions:
 - ▶ If S_t is a set, we have the optimum (like combinatorial bandits):

$$S^* \in \arg \max_{S \subseteq [k], |S|=n} \sum_{i \in S} u_i$$

- ▶ If S_t is a **multiset**, proposal of optimum:

$$S^* = \{i^*, \dots, i^*\} \quad \text{where} \quad i^* \in \arg \max_{i \in [k]} u_i$$

Open Questions

- Does it make sense/Is it feasible to consider ranking **without utilities**?
- Optimum and regret definitions:
 - ▶ If S_t is a set, we have the optimum (like combinatorial bandits):

$$S^* \in \arg \max_{S \subseteq [k], |S|=n} \sum_{i \in S} u_i$$

- ▶ If S_t is a **multiset**, proposal of optimum:

$$S^* = \{i^*, \dots, i^*\} \quad \text{where} \quad i^* \in \arg \max_{i \in [k]} u_i$$

Open Questions

- Does it make sense/Is it feasible to consider ranking **without utilities**?
- Optimum and regret definitions:
 - ▶ If S_t is a set, we have the optimum (like combinatorial bandits):

$$S^* \in \arg \max_{S \subseteq [k], |S|=n} \sum_{i \in S} u_i$$

- ▶ If S_t is a **multiset**, proposal of optimum:

$$S^* = \{i^*, \dots i^*\} \quad \text{where} \quad i^* \in \arg \max_{i \in [k]} u_i$$

Open Questions

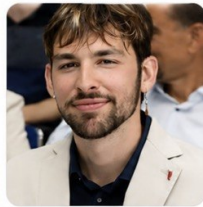
- Does it make sense/Is it feasible to consider ranking **without utilities**?
- Optimum and regret definitions:
 - ▶ If S_t is a set, we have the optimum (like combinatorial bandits):

$$S^* \in \arg \max_{S \subseteq [k], |S|=n} \sum_{i \in S} u_i$$

- ▶ If S_t is a **multiset**, proposal of optimum:

$$S^* = \{i^*, \dots i^*\} \quad \text{where} \quad i^* \in \arg \max_{i \in [k]} u_i$$

Thank You for Your Attention!



References I

- Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
- Róbert Busa-Fekete and Eyke Hüllermeier. A survey of preference-based online learning with bandit algorithms. In *Proceedings of the 25th International Conference on Algorithmic Learning Theory*, pages 18–39, 2014.
- Roger R. Davidson. On extending the Bradley–Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970.
- Simone Drago, Marco Mussi, and Alberto Maria Metelli. Towards theoretical understanding of sequential decision making with preference feedback. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=SqnVi0BHP0>.
- Simone Drago, Marco Mussi, Leonardo Bianconi, and Alberto Maria Metelli. Modeling incomparability: A new rationality paradigm for preference-based reinforcement learning. In *2nd Workshop on Epistemic Intelligence in Machine Learning (EIML@ICML 2026)*, 2026a.
- Simone Drago, Marco Mussi, and Alberto Maria Metelli. Online compatible reward identification from preference feedback. In *International Conference on Machine Learning*, 2026b. To appear.
- Pratik Gajane, Tanguy Urvoy, and Fabrice Clérot. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 218–227, 2015.
- Ashish Khetan and Sewoong Oh. Data-driven rank breaking for efficient rank aggregation. *Journal of Machine Learning Research*, 17(193):1–54, 2016.
- Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.
- R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.
- Cristiano Migali, Gianmarco Genalti, Alberto Maria Metelli, and Marco Mussi. Worst-case regret bounds for combinatorial bandits with ranking feedback, 2026a. Under review.

References II

- Cristiano Migali, Gianmarco Genalti, Alberto Maria Metelli, and Marco Mussi. Learning to rank from incomplete rankings. In *International Conference on Machine Learning*, 2026b. To appear.
- Robin L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.
- P. V. Rao and Lawrence L. Kupper. Ties in paired-comparison experiments: A generalization of the Bradley–Terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.
- Aadirupa Saha and Aditya Gopalan. PAC battling bandits in the Plackett–Luce model. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 700–737, 2019a.
- Aadirupa Saha and Aditya Gopalan. Combinatorial bandits with relative feedback. In *Advances in Neural Information Processing Systems*, volume 32, 2019b.
- Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.
- Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.
- Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.