Latent Reward Bandits with Ordinal Feedback

Arqam Patel, Sayak Ray Chowdhury

2025

1 Introduction

We consider an ordinal variant of the multi-armed bandits problem. Instead of directly observing numerical values of the reward instances, rankings over them are obtained. We construct a U-statistic estimator for the probability of one arm beating another and examine its asymptotic statistical properties. An upper bound on its variance is also derived. Finally, we propose and empirically demonstrate strategies to identify the best arm using these probability estimates.

1.1 Motivation

This variant, like the similar qualitative and duelling settings, is motivated by the need for robust methods for eliciting and incorporating subjective, hard to quantify, domain expert knowledge. We aim to propose more robust methods than subjective scores for settings where ordinal information is more appropriate. This may for example be relevant in exploring the purchase and review of goods or food from multiple sources and establishing the best one.

1.2 Problem statement

We consider a multi-armed bandit problem where the reward for each action is not observed directly. Instead, upon choosing an action at time step t, we observe $t-1$ ordinal reward signals comparing our action's (latent) reward to those of each of the previous actions. Equivalently, we can interpret the feedback as providing a ranking over all latent reward instances so far.

Assume we have K arms, indexed by $\mathcal{K} = \{1, 2, \ldots, K\}$. For all $i \in \mathcal{K}$ let the reward distribution for the ith arm be some sub-gaussian random variable R_i, with mean Δ_i.

2 Literature review

2.1 Hypothesis testing for stochastic order using ranking information

2.1.1 Stochastic order

Given two random variables X and Y with continuous cumulative distribution functions F_X and F_Y respectively. X is said to be stochastically larger than Y if for all values of a, $F_X(a) < F_Y(a)$, which is equivalent to $\mathbb{P}(X > a) > \mathbb{P}(Y > a)$.

If X is stochastically larger than Y, it is necessarily larger in expectation.

2.1.2 Wilcoxon rank sum statistic or Mann Whitney U-statistic

This is a non-parametric version of the two-sample t-test, since we want to infer whether two samples have the same *central tendency*. Unlike the t-test, it does not make normality (or any distributional) assumptions about our data but has lower statistical power. The Wilcoxon and Mann-Whitney tests are inferentially equivalent, only differing in the test statistics used.

The U statistic [MW47] measures the number of pairs (X_i, Y_j) such that $X_i < Y_j$)

Assumptions:

1. The two samples are independent of each other.
2. The two samples have equal variance.

Procedure for computing either statistic:

1. Pool data from both groups $C = \{x_1, \ldots, x_m, y_1, \ldots, y_n\}$ and assign ranks $Z = \{z_1, \ldots, z_{m+n}\}$.
2. Let sums of ranks $R_X = \sum_{i=1}^{m} z_i$ and $R_Y = \sum_{i=m}^{m+n} z_i$. ($R_X + R_Y = \frac{(m+n)(m+n+1)}{2}$)
3. The Wilcoxon rank sum statistic for X is $W = R_X$.
4. The U-statistic for X is $U_X = R_X - \frac{m(m+1)}{2}$.

Null hypothesis: $\mathbb{P}(X > Y) = \mathbb{P}(X < Y) = 0.5$
Alternative hypothesis: $\mathbb{P}(X > Y) > 0.5$ or $\mathbb{P}(X > Y) < 0.5$

Under the null hypothesis, for large sample sizes ($m, n > 10$), $W \sim \mathcal{N}(\Delta_W, \sigma_W^2)$ approximately, where $\Delta_W = \frac{m(m+n+1)}{2}$ and $\sigma_W^2 = \frac{mn(m+n+1)}{12}$, meaning we can compute the Z-score and from it, the p-value.

Under the same conditions, $U_X \sim \mathcal{N}(\Delta_U, \sigma_U^2)$ approximately, where $\Delta_U = \frac{mn}{2}$ and $\sigma_U^2 = \frac{mn(m+n+1)}{12}$
Bayesian equivalent of this has also been proposed. [DLMW19]

2.2 Qualitative Bandits

The problem of qualitative bandits [SBWH15] studies a generalized bandit setting where we get rewards on a qualitative scale that allows comparison but not arithmetic operations like averaging of rewards. This is analogous to our setting.

Arm distributions are defined over a completely ordered set $(L, \succeq)$. The quality of an arm is expressed in terms of its τ quantile of the arm's (categorical) distribution over L.

2.3 Dueling bandits

In the dueling bandits problem[BBMH21], we have K arms, say $\{1, \ldots, K\}$, and we want to find the optimal one. Unlike the standard multi-armed bandit setting, however, we do not have access to numerical rewards on pulling a single arm- we can only get feedback in the form of noisy ordinal comparisons $\mathbb{I}(i \succ j)$ between pairs of arms. An action is thus the selection of two arms to compare.

We can characterize the feedback with pairwise preference probabilities

$$\mathbb{P}(i \succ j) = q_{i,j}$$

of observing a preference for arm i versus j. We can structure this as matrix representing a preference relation $\mathbf{Q} = [q_{i,j}]_{1 \leq i,j \leq K} \in [0,1]^{K \times K}$. Clearly, each $q_{i,j}$ specifies a Bernoulli distribution over $\mathbb{I}(i \succ j) \in \{0, 1\}$, which is assumed to be stationary and independent across arms, and time steps.

Hence, the outcomes of previous iterations (even those of (i, j) do not affect the outcome of action (i, j)- it is freshly sampled from the aforementioned Bernoulli.

We say arm i beats j if $q_{i,j} > \frac{1}{2}$ i.e. i is more likely to win in a pairwise comparison than j is.

$$\Delta_{i,j} = q_{i,j} - \frac{1}{2}$$

2.3.1 Learning tasks

1. **Best arm**: The best arm (equivalent to the Condorcet winner) is $i^* \in \mathcal{K}$ such that $\Delta_{i,j} > 0$ $\forall j \in \mathcal{K} \setminus \{i^*\}$. A Condorcet winner may not exist if preferences are cyclic.
2. Ranking of arms
3. Top-k arms
4. Estimation of the preference relation/utility function

2.3.2 Regret

The regret at time step t, if it chooses arms $i(t)$ and $j(t)$, is defined as

$$r_t = \frac{\Delta_{i^*,i(t)} + \Delta_{i^*,j(t)}}{2} = \frac{q_{i^*,i(t)} + q_{i^*,j(t)}}{2} - \frac{1}{2}$$

2.3.3 Generating process models

There can be multiple valid generative models for dueling bandits as the various comparison signals are mutually independent.

Bradley Terry model
According to the Bradley-Terry model, we directly model the probability of the outcome of pairwise comparisons as:

$$\mathbb{P}(i \succ j) = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}} = \frac{1}{1 + \exp(\beta_2 - \beta_1)}$$

where β_i and β_j are real-valued scores (possibly of utility functions) assigned to i and j respectively. Instead of the logistic, we can also use some other link function.

Latent random reward model
In this, the ith arm b_i is associated with a reward random variable R_i from which a draw is made each time to compare with the reward sampled from another arm. Assuming that $R_i \sim \mathcal{N}(\Delta_i, 1)$, we have $R_i - R_j \sim \mathcal{N}(\Delta_i - \Delta_j, 2)$

$$\mathbb{P}(i \succ j) = \mathbb{P}(R_i - R_j > 0)$$

3 Estimation of $\mathbb{P}(i \succ j)$

3.1 Two arm case

3.1.1 Estimation of $\mathbb{P}(X > Y)$

Let $\mathcal{K} = \{1, 2\}$. Let $X \sim \mathcal{N}(\Delta, 1), Y \sim \mathcal{N}(\Delta, 1)$. $X - Y \sim \mathcal{N}(\Delta, 2)$.

Assume some i.i.d. latent reward realizations $x_i \sim X$ and $y_j \sim Y$ i.i.d.

We propose the following estimator:

$$\hat{p} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}(x_i > y_j) = \frac{1}{mn} U$$

where U is the Mann-Whitney statistic (computed using ranks) for the reward samples from X and Y.

We will prove the following theorems regarding the statistical properties of this estimator.

Theorem 1.1 $\hat{p}$ is a consistent estimator of p.

Theorem 1.2 $\hat{p}$ is an unbiased estimator of p.

Theorem 2.1 The variance of $\hat{p}$ given the value of Δ is

$$V_{m,n}(\Delta) = \frac{1}{mn} \left\{ \Phi\left(\frac{\Delta}{\sqrt{2}}\right) - (m+n-1) \cdot \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right) + (m+n-2) \cdot \Psi(\Delta) \right\}$$

where $\Phi(\cdot)$ is the standard normal cumulative distirbution function and $\Psi(\Delta)$ is the CDF of a bivariate normal random variable with mean $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and covariance $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ until the point $\begin{bmatrix} \Delta \\ \Delta \end{bmatrix}$.

Theorem 2.2 The best Δ agnostic bound on the variance of $\hat{p}$ is

$$Var(\hat{p}) \leq \sup_{\Delta} V_{m,n}(\Delta) = \frac{1}{12m} + \frac{1}{12n} + \frac{1}{12mn}$$

3.1.2 Consistency

$$\begin{aligned}
\lim_{m\to\infty} \lim_{n\to\infty} \hat{p} &= \lim_{m\to\infty} \lim_{n\to\infty} \frac{1}{mn} \sum_{i\in T_1} \sum_{j\in T_2} \mathbb{I}(X > Y) \\
&= \lim_{m\to\infty} \lim_{n\to\infty} \frac{1}{mn} \sum_{i\in T_1} \sum_{j\in T_2} \mathbb{I}\left(y_j > x_i\right) \\
&= \lim_{m\to\infty} \frac{1}{m} \sum_{i\in T_1} \left[\lim_{n\to\infty} \frac{1}{n} \sum_{j\in T_2} \mathbb{I}\left(y_j > x_i\right) \right] \\
&\xrightarrow{\text{a.s.}} \lim_{m\to\infty} \frac{1}{m} \sum_{i\in T_1} \mathbb{E}\left[\mathbb{I}\left(R_2 > x_i\right)\right] \\
&= \lim_{m\to\infty} \frac{1}{m} \sum_{i\in T_1} \mathbb{P}\left(R_2 > x_i\right) \\
&= \lim_{m\to\infty} \frac{1}{m} \sum_{i\in T_1} \mathbb{P}\left(R_2 > R_1^{(i)}\right) \\
&\xrightarrow{\text{a.s.}} \mathbb{E}_{R_1}\left[\mathbb{P}(R_2 > R_1)\right] \\
&= \mathbb{P}(R_2 > R_1)
\end{aligned}$$

3.1.3 Unbiasedness

For fixed i, the inner sum $\sum_{j \in T_2} \mathbb{I}(y_j > x_i)$ is a sum of i.i.d. indicators (because we're conditioning on x_i), each with probability $\mathbb{P}(R_2 > x_i)$

$$
\begin{aligned}
\mathbb{E}[\hat{p}] &= \mathbb{E}\left[\frac{1}{mn}\sum_{i \in T_1}\sum_{j \in T_2}\mathbb{I}(y_j > x_i)\right] \\
&= \frac{1}{mn}\mathbb{E}_{R_1}\left[\sum_{i \in T_1}\mathbb{E}_{R_2}\left[\sum_{j \in T_2}\mathbb{I}(y_j > x_i) \mid x_i\right]\right] \\
&= \frac{1}{mn}\mathbb{E}_{R_1}\left[\sum_{i \in T_1}\sum_{j \in T_2}\mathbb{E}_{R_2}\left[\mathbb{I}(y_j > x_i) \mid x_i\right]\right] \quad \text{For fixed } i\text{, inner sum is sum of i.i.d. indicators} \\
&= \frac{1}{mn}\mathbb{E}_{R_1}\left[\sum_{i \in T_1} n\mathbb{P}(R_2 > x_i)\right] \\
&= \frac{1}{m}\sum_{i \in T_1}\mathbb{E}_{R_1}\left[\mathbb{P}(R_2 > x_i)\right] \\
&= \frac{1}{m}\sum_{i \in T_1}\mathbb{P}(R_2 > R_1) \\
&= \mathbb{P}(R_2 > R_1)
\end{aligned}
$$

3.1.4 Asymptotic Normality

Both the Wilcoxon and Mann-Whitney test statistics (differing only by a constant) have well-known results on asymptotic normality [Cap61]. This also holds for $\hat{p}$, which is a scaled U statistic.

3.1.5 Variance analysis of $\hat{p}$

$$
\begin{aligned}
Var[\hat{p}] &= Var\left[\frac{1}{mn}\sum_{i\in T_1}\sum_{j\in T_2}\mathbb{I}\left(y_j > x_i\right)\right] \\
&= \frac{1}{m^2n^2}Var\left[\sum_{i\in T_1}\sum_{j\in T_2}\mathbb{I}\left(y_j > x_i\right)\right] \\
&= \frac{1}{m^2n^2}\cdot\left(\sum_{i\in T_1}\sum_{j\in T_2}Var\left[\mathbb{I}\left(y_j > x_i\right)\right] + \sum_{i\in T_1}\sum_{j\in T_2}\sum_{i'\in T_1}\sum_{j'\in T_2}Cov[\mathbb{I}(y_j > x_i), \mathbb{I}(y_{j'} > x_{i'})]\right) \\
&= \frac{1}{m^2n^2}\Bigg(\sum_{i\in T_1}\sum_{j\in T_2}Var\left[\mathbb{I}\left(y_j > x_i\right)\right] \\
&\qquad + \sum_{i\in T_1}\sum_{j\in T_2}\sum_{j'\in T_2\setminus\{j\}}Cov[\mathbb{I}(y_j > x_i), \mathbb{I}(y_{j'} > x_i)] \\
&\qquad + \sum_{j\in T_2}\sum_{i\in T_1}\sum_{i'\in T_1\setminus\{i\}}Cov[\mathbb{I}(y_j > x_i), \mathbb{I}(y_j > x_{i'})]\Bigg)
\end{aligned}
$$

The last line follows because $\mathbb{I}(y_j > x_i)$ and $\mathbb{I}(y_{j'} > x_{i'})$ are independent if $i \neq i'$ and $j \neq j'$ (since x_i, x_j and y_i, y_j are all mutually independent). Thus all covariance terms not sharing a common random variable become zero.
We now consider the three types of terms one by one:

$$\mathbb{E}[\mathbb{I}(y_j > x_i)] = p \quad \forall i, j$$

$$
\begin{aligned}
Var\left[\mathbb{I}\left(y_j > x_i\right)\right] &= \mathbb{E}[(\mathbb{I}(y_j > x_i))^2] - (\mathbb{E}[\mathbb{I}(y_j > x_i)])^2 \\
&= \mathbb{E}[\mathbb{I}(y_j > x_i)] - (\mathbb{E}[\mathbb{I}(y_j > x_i)])^2 \quad \text{square of indicator is itself} \\
&= p - p^2 \\
&= \Phi\left(\frac{\Delta}{\sqrt{2}}\right) - \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right)
\end{aligned}
$$

$$
\begin{aligned}
Cov[\mathbb{I}(y_j > x_i), \mathbb{I}(y_{j'} > x_i)] &= \mathbb{E}\left[(\mathbb{I}(y_j > x_i) - p)(\mathbb{I}(y_{j'} > x_i) - p)\right] \\
&= \mathbb{E}\left[\mathbb{I}(y_j > x_i) \cdot \mathbb{I}(y_{j'} > x_i))\right] - p^2 \\
&= \mathbb{E}\left[\mathbb{E}[\mathbb{I}(y_j > x_i) \cdot \mathbb{I}(y_{j'} > x_i))|x_i]\right] - p^2 \\
&= \mathbb{E}\left[\mathbb{E}[\mathbb{I}(y_j > x_i)|x_i] \cdot \mathbb{E}[\mathbb{I}(y_{j'} > x_i))|x_i]\right] - p^2 \quad \text{conditionally independent} \\
&= \mathbb{E}\left[\mathbb{P}(Y_j > X_i|X_i = x_i) \cdot \mathbb{P}(Y_{j'} > X_i|X_i = x_i)\right] - p^2 \\
&= \mathbb{E}\left[(1 - \Phi_Y(x_i)) \cdot (1 - \Phi_Y(x_i))\right] - p^2 \\
&= \mathbb{E}_X\left[(1 - \Phi_Y(x))^2\right] - p^2 \\
&= \mathbb{E}_X\left[(1 - \Phi(x - \Delta))^2\right] - \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right) \\
&= \mathbb{E}_X\left[\mathbb{P}(v > x - \Delta, w > x - \Delta)\right] - \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right) \quad \text{iid } v, w \sim \mathcal{N}(0,1) \\
&= \mathbb{P}(x - v < \Delta, x - w < \Delta) - \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right) \\
&= \Psi\left(\mathbf{x} = \begin{bmatrix}\Delta \\ \Delta\end{bmatrix} \;\middle|\; \boldsymbol{\Delta} = \begin{bmatrix}0 \\ 0\end{bmatrix}, \Sigma = \begin{bmatrix}2 & 1 \\ 1 & 2\end{bmatrix}\right) - \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right)
\end{aligned}
$$

where Ψ is the multivariate CDF.

$$
\begin{aligned}
Cov[\mathbb{I}(y_j > x_i), \mathbb{I}(y_j > x_{i'})] &= \mathbb{E}\left[(\mathbb{I}(y_j > x_i) - p)(\mathbb{I}(y_j > x_{i'}) - p)\right] \\
&= \mathbb{E}\left[\mathbb{I}(y_j > x_i) \cdot \mathbb{I}(y_j > x_{i'}))\right] - p^2 \\
&= \mathbb{E}\left[\mathbb{E}[\mathbb{I}(y_j > x_i) \cdot \mathbb{I}(y_j > x_{i'}))|y_j]\right] - p^2 \\
&= \mathbb{E}\left[\mathbb{E}[\mathbb{I}(y_j > x_i)|y_j] \cdot \mathbb{E}[\mathbb{I}(y_j > x_{i'}))|y_j]\right] - p^2 \quad \text{conditionally independent} \\
&= \mathbb{E}\left[\mathbb{P}(Y_j > X_i|Y_j = y) \cdot \mathbb{P}(Y_j > X_{i'}|Y_j = y)\right] - p^2 \\
&= \mathbb{E}\left[\mathbb{P}(X_i < y) \cdot \mathbb{P}(X_{i'} < y)\right] - p^2 \\
&= \mathbb{E}\left[\ \Phi_X(y) \cdot \Phi_X(y)\right] - p^2 \\
&= \mathbb{E}_{Y \sim \mathcal{N}(\Delta,1)}\left[\Phi^2(y)\right] - \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right) \\
&= \int_{-\infty}^{\infty} \phi_Y(y)\Phi^2(y)dy - \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right) \\
&= \int_{-\infty}^{\infty} \phi(y - \Delta)\Phi^2(y)dy - \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right) \\
&= \int_{-\infty}^{\infty} \phi(z)\Phi^2(z + \Delta)dz - \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right) \\
&= \mathbb{E}\left(\mathbb{P}(v \leq z + \Delta, w \leq z + \Delta)\right) - \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right) \\
&= \mathbb{P}(v - z \leq \Delta, w - z \leq \Delta) - \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right) \quad \text{iid } v, w \sim \mathcal{N}(0,1) \\
&= \Psi\left(\mathbf{x} = \begin{bmatrix}\Delta \\ \Delta\end{bmatrix} \;\middle|\; \boldsymbol{\Delta} = \begin{bmatrix}0 \\ 0\end{bmatrix}, \Sigma = \begin{bmatrix}2 & 1 \\ 1 & 2\end{bmatrix}\right) - \Phi^2\left(\frac{\Delta}{\sqrt{2}}\right)
\end{aligned}
$$

This is the same as $Cov[\mathbb{I}(y_j > x_i), \mathbb{I}(y_{j'} > x_i)]$, and is numerically computable. We plot this as a function of Δ and compare it to the Monte Carlo estimate of the covariance:

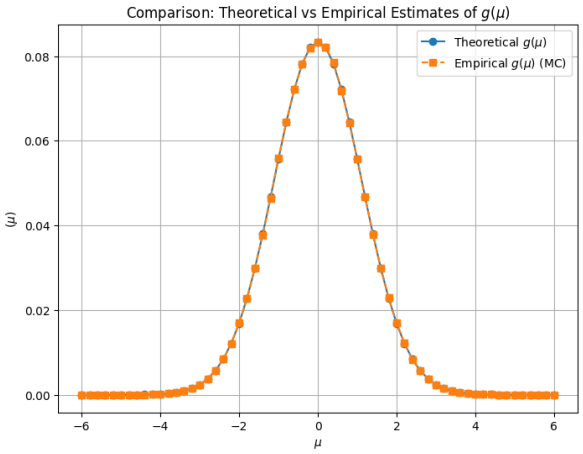

Figure 1: $g(\Delta)$ represents $Cov[\mathbb{I}(y_j > x_i), \mathbb{I}(y_j > x_{i'})]$ as a function of Δ

Visually, it's apparent that it reaches a maximum at $\Delta = 0$, where its value is $\frac{1}{12}$. Plugging these into the expression for $Var[\hat{p}]$, we obtain

$$Var[\hat{p}] = \frac{1}{mn}\left\{\Phi\left(\frac{\Delta}{\sqrt{2}}\right) - (m+n-1)\cdot\Phi^2\left(\frac{\Delta}{\sqrt{2}}\right) + (m+n-2)\cdot\Psi(\Delta)\right\}$$

Thus, we can upper-bound the variance by

$$\begin{aligned} Var[\hat{p}] &\leq \frac{1}{m^2n^2}\left[mn\cdot(p-p^2) + m(m-1)n\cdot\frac{1}{12} + mn(n-1)\cdot\frac{1}{12}\right] \\ &\leq \frac{1}{mn}\left[(p-p^2) + (m+n-2)\cdot\frac{1}{12}\right] \\ &\leq \frac{1}{mn}\left[\frac{1}{4} + \frac{(m+n-2)}{12}\right] \\ &\leq \frac{m+n+1}{12mn} \end{aligned}$$

This bound is identical to the appropriately scaled asymptotic variance of the U statistic under the null hypothesis in the literature (since $\Delta = 0$ corresponds to $p = \frac{1}{2}$).

3.1.6 Estimation of Δ from $\hat{p}$

We can see that:

$$\begin{aligned} \mathbb{P}(X > Y) &= \mathbb{P}(X - Y > 0) \\ p &= 1 - \Phi\left(-\frac{\Delta}{\sqrt{2}}\right) \\ p &= \Phi\left(\frac{\Delta}{\sqrt{2}}\right) \end{aligned}$$

We can define an estimator for Δ based on our estimate for p:

$$\hat{\Delta} = \sqrt{2}\Phi^{-1}(\hat{p})$$

3.2 Comparison with i.i.d. estimator

Define

$$\tilde{p} = \sum_{i=1}^{\min\{m,n\}} \mathbb{I}(x_i > y_i)$$

We can observe that unlike in the case of $\hat{p}$, $\tilde{p}$ is a sample mean of i.i.d. observations. Thus, it is also consistent, unbiased and asymptotically normal.

The variance of $\tilde{p}$ in terms of Δ is

$$W_{m,n}(\Delta) = \frac{1}{\min\{m,n\}} \Phi\left(\frac{\Delta}{\sqrt{2}}\right)\left(1 - \Phi\left(\frac{\Delta}{\sqrt{2}}\right)\right)$$

For all values of Δ and m, n, we have $W_{m,n}(\Delta) \geq V_{m,n}(\Delta)$. Visually, we can check this plotting W and V as functions of Δ keeping m, n fixed.

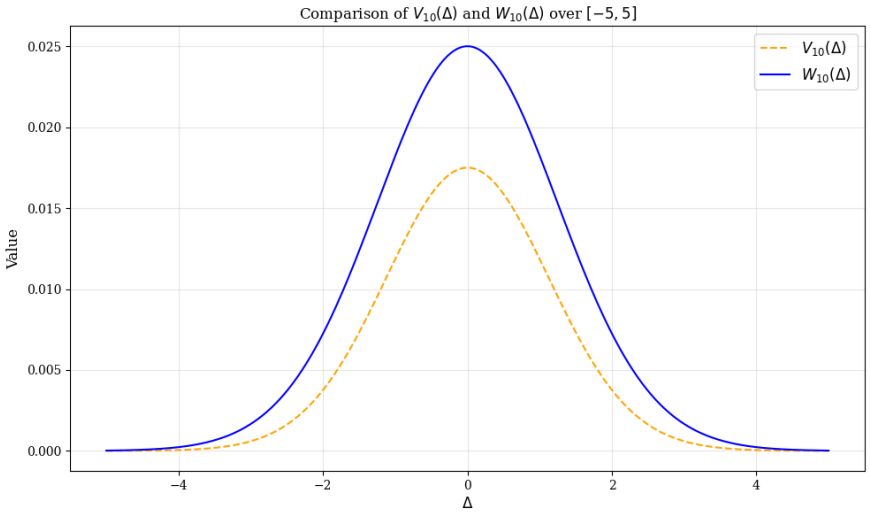

Figure 2: Variances of estimators

Additionally, the variance bound for $\hat{p}$ is also tighter than the resulting bound for $\tilde{p}$. We visualise this below. In the limit,

$$\frac{\sup_{\Delta} \text{Var}[\tilde{p}]}{\sup_{\Delta} \text{Var}[\hat{p}]} \to \frac{3}{2}$$

Thus, our proposed estimator $\hat{p}$ is better than $\tilde{p}$ in terms of utilising all available information in order to lessen uncertainty.

3.3 Extending to multi-arm case

It is non-transitive i.e. $\hat{p}_{i,j} > 1/2$ and $\hat{p}_{j,k} > 1/2$ does not imply $\hat{p}_{i,k} > 1/2$ 2. Thus, just estimating the various $p_{i,j}$s could be insufficient to determine a ranking over arms or even the best arm (corresponding to the Condorcet winner).

Assume $\mathcal{K} = \{1, 2, 3\}$. We want to draw inferences about orderings.

One simple estimator for the probability that arm 1 is the best arm is:

$$\hat{P}_1 = \frac{1}{m_1 m_2 m_3} \sum_{i \in T_1} \sum_{j \in T_2} \sum_{k \in T_3} \mathbb{I}\left(R_1^{(i)} > R_2^{(j)}\right) \mathbb{I}\left(R_1^{(i)} > R_3^{(k)}\right)$$

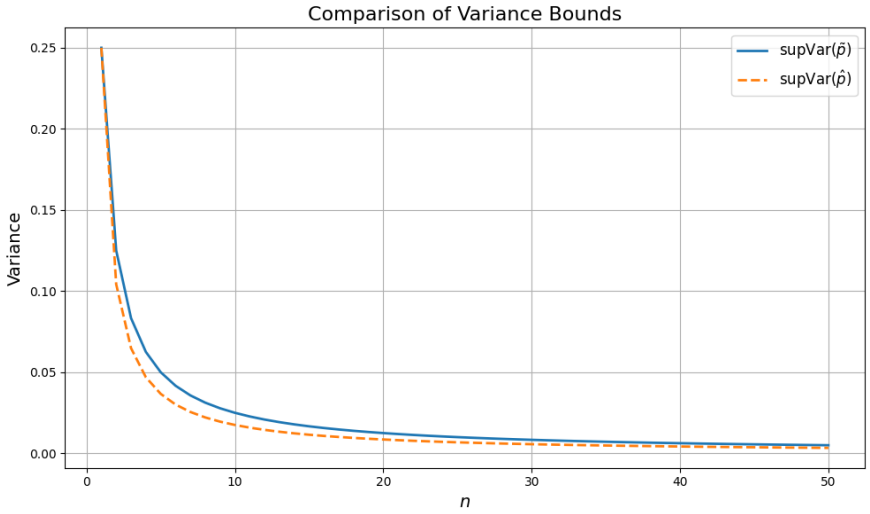

Figure 3: Variance bounds of estimators

We have insufficient information about the distribution of this estimator so far, but it could be a promising, direct direction for the multi-arm case.

Alternatively, we can use a version of the Copeland score from computational social choice theory using the pairwise probability estimates:

$$C_i = \sum_{i \neq j} \mathbb{I}\left(\hat{p}_{i,j} > \frac{1}{2}\right)$$

This could be used to rank arms in a way similar to that in [SG20] for successive elimination for a PAC best-arm identification approach.

4 Fixed Budget Algorithm

4.1 Algorithm

Algorithm 1 UNIFORM-ALLOCATION

input: Set of items: $\mathcal{A} = \{1, 2, \dots k\}$, budget Q
init: $\mathcal{R} \leftarrow \mathcal{A}$, $l \leftarrow 1$
while $|\mathcal{R}| > 1$ **do**
    for $a \in \mathcal{R}$ **do**
        Play arm a for $Q' := \frac{Q}{2^l |\mathcal{R}|}$ times
    end for
    for $i \in \mathcal{R}$ **do**
        for $j \in \mathcal{R}$ **do**
            update $\hat{p}_{ij}$
        end for
        Compute $C_i := \sum_{j \neq i} \mathbb{I}(\hat{p}_{ij} > \frac{1}{2})$
    end for
    Define $\bar{C} \leftarrow \text{Median}(\{C_i\}_{i \in R})$
    $\mathcal{R} \leftarrow \{i \in \mathcal{R} \mid C_i \geq \bar{C}\}$
    $l \leftarrow l + 1$
end while
output: The remaining item in $\mathcal{R}$

4.2 Theorem

Assuming that any estimator with m and n samples from the two distributions respectively is distributed as $\hat{p} \sim N(p, \frac{m+n+1}{12mn})$. This is only true asymptotically for large sample sizes so the bound is invalid for the small sample setting.

Let our budget be Q, number of arms be k. Define the problem complexity parameter $\Delta_{min} = \min_{i \neq i^*} \Delta_i$ where $\Delta_i = p_{i^*,i} - \frac{1}{2} = \frac{1}{2} - p_{i,i^*}$.
Then, our algorithm finds the optimal arm with probability at least

$$1 - \frac{\sqrt{k} \log_2 k}{\sqrt{\pi Q} \Delta_{\min}} \exp\left(-\frac{Q}{k} \Delta_{min}^2\right)$$

4.3 Proof

Assumption: $\hat{p}_{i,j} \sim \mathcal{N}(p_{i,j}, \sigma^2)$, 1 is the best arm

$$
\begin{aligned}
\sigma^2 &\leq \frac{2m+1}{12m^2} \\
&\leq \frac{2m+m}{12m^2} \\
&\leq \frac{1}{4m} \\
\frac{\Delta_i}{\sigma} &\geq 2\sqrt{m}\Delta_i \\
\Phi\left(\frac{\Delta_i}{\sigma}\right) &\geq \Phi\left(2\sqrt{m}\Delta_i\right) \\
1-\Phi\left(\frac{\Delta_i}{\sigma}\right) &\leq 1-\Phi\left(2\sqrt{m}\Delta_i\right)
\end{aligned}
$$

Probability of suboptimal arm $i \neq 1$ beating arm 1 according to $\hat{p}$

$$
\begin{aligned}
\mathbb{P}\left(\hat{p}_{i,1} > \frac{1}{2}\right) &= \mathbb{P}\left(\hat{p}_{i,1} - p_{i,1} > \frac{1}{2} - p_{i,1}\right) \\
&= \mathbb{P}\left(\hat{p}_{i,1} - p_{i,1} > \Delta_i\right) \\
&= \mathbb{P}\left(\frac{\hat{p}_{i,1} - p_{i,1}}{\sigma} > \frac{\Delta_i}{\sigma}\right) \\
&= 1-\Phi\left(\frac{\Delta_i}{\sigma}\right) \\
&\leq 1-\Phi\left(\frac{\Delta_i}{\sqrt{\frac{1}{4m}}}\right) \\
&\leq 1-\Phi\left(2\sqrt{m}\Delta_i\right) \\
&\leq \frac{1}{2\sqrt{2\pi m}\Delta_i}\exp\left(-2m\Delta_i^2\right)
\end{aligned}
$$

The last step is by Mill's inequality.
Define $V = \sum_{i\neq 1}\mathbb{I}\left(p_{i,1} > \frac{1}{2}\right)$

$$
\begin{aligned}
\mathbb{E}[V] &= \sum_{i\neq 1}\mathbb{P}\left(p_{i,1} > \frac{1}{2}\right) \\
&\leq (k_t - 1)\cdot\frac{1}{2\sqrt{2\pi m}\Delta_{\min}}\exp\left(-2m\Delta_{\min}^2\right) \\
\mathbb{P}\left(V \geq \frac{k_t}{2}\right) &\leq \frac{\mathbb{E}[V]}{\frac{k_t}{2}} \\
&\leq \frac{k_t - 1}{k_t\sqrt{2\pi m}\Delta_i}\exp\left(-2m\Delta_i^2\right) \\
&\leq \frac{k_t - 1}{k_t\sqrt{2\pi m}\Delta_{\min}}\cdot\exp\left(-2m\Delta_{\min}^2\right) \\
&\leq \frac{1}{\sqrt{2\pi m}\Delta_{\min}}\exp\left(-2m\Delta_{\min}^2\right)
\end{aligned}
$$

Probability of elimination throughout $\log_2(k)$ rounds:

$$
\begin{aligned}
\mathbb{P}(1 \text{ gets eliminated in one of } 1, 2, \ldots, \log_2 k) &\leq \sum_{l=1}^{\log_2 k} \mathbb{P}(1 \text{ eliminated in round } l) \\
&\leq \sum_{l=1}^{\log_2 k} \frac{1}{\sqrt{2\pi m_l}\Delta_{\min}} \exp\left(-2m_l\Delta_{min}^2\right) \\
&\leq \frac{\log_2 k}{\sqrt{2\pi \frac{Q}{2k}}\Delta_{\min}} \exp\left(-2 \cdot \frac{Q}{2k} \cdot \Delta_{min}^2\right) \\
&\leq \frac{\sqrt{k}\log_2 k}{\sqrt{\pi Q}\Delta_{\min}} \exp\left(-\frac{Q}{k}\Delta_{min}^2\right)
\end{aligned}
$$

5 Future directions

1. Obtain a practically useful concentration bound for $\hat{p}$ that lets us derive an algorithm that provably finds the best arm with any given probability (PAC setting) or given a budget, lets us find a lower bound on probability of finding best arm (fixed budget setting).
2. Study the distribution and concentration of $\hat{P}$
3. van Doorn et al (2019) propose a Bayesian variant of the Mann-Whitney test between two random variables using Gibbs sampling. This sampling procedure could be extended to the several arms case to obtain an ordinal Thomson sampling scheme.

References

[BBMH21] V. Bengs, R. Busa-Fekete, A. E. Mesaoudi-Paul, and E. Hüllermeier. *Preference-based Online Learning with Dueling Bandits: A Survey*. 2021. arXiv: 1807.11398 [cs.LG]. URL: https://arxiv.org/abs/1807.11398.

[Cap61] J. Capon. "A Note on the Asymptotic Normality of the Mann-Whitney-Wilcoxon Statistic". In: *Journal of the American Statistical Association* 56.295 (1961), pp. 687–691. ISSN: 01621459. URL: http://www.jstor.org/stable/2282089 (visited on 04/25/2025).

[DLMW19] J. van Doorn, A. Ly, M. Marsman, and E.-J. Wagenmakers. *Bayesian Rank-Based Hypothesis Testing for the Rank Sum Test, the Signed Rank Test, and Spearman's ρ*. 2019. arXiv: 1712.06941 [stat.ME]. URL: https://arxiv.org/abs/1712.06941.

[MW47] H. B. Mann and D. R. Whitney. "On a test of whether one of two random variables is stochastically larger than the other". In: *The annals of mathematical statistics* (1947), pp. 50–60.

[SBWH15] B. Szorenyi, R. Busa-Fekete, P. Weng, and E. Hüllermeier. "Qualitative Multi-Armed Bandits: A Quantile-Based Approach". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 1660–1668. URL: https://proceedings.mlr.press/v37/szorenyi15.html.

[SG20] A. Saha and A. Gopalan. *From PAC to Instance-Optimal Sample Complexity in the Plackett-Luce Model*. 2020. arXiv: 1903.00558 [cs.LG]. URL: https://arxiv.org/abs/1903.00558.