Leveraging observations in bandits: Between risks and benefits Supplementary

Andrei Lupu, Audrey Durand and Doina Precup

November 13, 2018

1 Regret upper-bound

We can express the cumulative regret (Eq. 1) as $\mathfrak{R}(T) = \sum_{\substack{a\in\mathcal{A}\\ a\neq\star}} \Delta_a N_{a,T}$. This quantity can be bounded by controlling sub-optimal plays:

$$N_{a,T} = 1 + \sum_{t=A+1}^{T-1} \mathbb{I}\{a_t = a\}.$$

Also let us introduce the following events to characterize the concentration of the empirical means.

Definition 1 (Events E_t^a and $E_t^\star$)**.** *Let these events respectively denote the situations where*

$$m_{a,s} - \mu_a \leqslant \sqrt{\frac{3\ln t}{2N_{a,s}}} \quad \textit{and} \quad \mu_\star - m_{\star,s} \leqslant \sqrt{\frac{3\ln t}{2N_{\star,s}}}$$

simultaneously for all $s \leqslant t$.

The idea is to decompose

$$\begin{aligned} N_{a,T} &\leqslant 1 + \sum_{t=A+1}^{T-1} \mathbb{I}\{a_t = a, E_t^a, E_t^\star\} + \sum_{t=A+1}^{T-1} \mathbb{I}\{\bar{E}_t^a \cup \bar{E}_t^\star\} \\ &\leqslant \ell + \sum_{t=\ell}^{T-1} \mathbb{I}\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell\} + \sum_{t=A+1}^{T-1} \mathbb{I}\{\bar{E}_t^a \cup \bar{E}_t^\star\}. \end{aligned} \tag{A.1}$$

and control the two sums separately. Then, the general idea of the proof will be to pick ℓ such that the first sum is controlled and to show that the second sum is controlled by definition of events. Let us begin by bounding the first sum, that is the sum of sub-optimal plays under the occurrence of events E_t^a and $E_t^\star$. We can decompose

$$\begin{aligned} \sum_{t=\ell}^{T-1} \mathbb{I}\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell\} \leqslant &\sum_{t=\ell}^{T-1} \mathbb{I}\left\{\frac{N_{a_t}}{\tilde{N}_{a_t}} < 1\right\} \mathbb{I}\left\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell, \frac{N_{a_t}}{\tilde{N}_{a_t}} < 1\right\} \\ &+ \sum_{t=\ell}^{T-1} \mathbb{I}\left\{\frac{N_{a_t}}{\tilde{N}_{a_t}} \geqslant 1\right\} \mathbb{I}\left\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell, \frac{N_{a_t}}{\tilde{N}_{a_t}} \geqslant 1\right\} \end{aligned}$$

The first part refers to Target-UCB being *better* than the target policy with respect to action a, while the second part refers to Target-UCB being *worse*. Also recall that playing sub-optimal action a requires Equation 4 to be satisfied. This will be useful in the following developments.

1.1 Sub-optimal plays when worse

If $N_{a,t}/\tilde{N}_{a,t} \geqslant 1$ for sub-optimal action a, then we can say that Target-UCB has played worse than (or equal to) its target up to time t, with respect to action a. We want to bound the selection of action a in this situation, where Equation 4 simplifies to

$$\Delta_a \leqslant \sqrt{\frac{3\ln t}{2N_{a,t}}} + \sqrt{\frac{3\ln t}{2N_{\star,t}}} - \sqrt{\frac{C\ln t}{N_{\star,t}}}\sqrt{1-\left(\frac{N_{\star,t}}{\tilde{N}_{\star,t}} \wedge 1\right)}. \tag{A.2}$$

We consider two situations A) $N_{\star,t}/\tilde{N}_{\star,t} \leqslant 1 - \frac{3}{2C}$ or B) $N_{\star,t}/\tilde{N}_{\star,t} > 1 - \frac{3}{2C}$.

1.1.1 Case A)

When $N_{\star,t}/\tilde{N}_{\star,t} \leqslant 1 - \frac{3}{2C}$, Equation A.2 further simplifies to

$$\Delta_a \leqslant \sqrt{\frac{3\ln t}{2N_{a,t}}}.$$

Therefore, in order for

$$\mathbb{I}\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell, N_{a,t}/\tilde{N}_{a,t} \geqslant 1\}$$

to happen when $N_{\star,t}/\tilde{N}_{\star,t} \leqslant 1 - \frac{3}{2C}$, the following condition must hold:

$$N_{a,t} \leqslant \frac{3\ln t}{2\Delta_a^2}. \tag{A.3}$$

1.1.2 Case B)

When $N_{\star,t}/\tilde{N}_{\star,t} > 1 - \frac{3}{2C}$, Equation A.2 further simplifies to

$$\Delta_a \leqslant \sqrt{\frac{3\ln t}{2N_{a,t}}} + \sqrt{\frac{3\ln t}{2N_{\star,t}}}.$$

Now, either $N_{a,t} > N_{\star,t}$ or $N_{a,t} \leqslant N_{\star,t}$. The first situation requires that

$$N_{\star,t} \leqslant \frac{6\ln t}{\Delta_a^2}$$

to hold. Thanks to Assumption 1, this is not possible for $N_{\star,t}/\tilde{N}_{\star,t} > 1 - \frac{3}{2C}$. Therefore, we must be in the situation where $N_{a,t} \leqslant N_{\star,t}$, such that, in order for

$$\mathbb{I}\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell, N_{a,t}/\tilde{N}_{a,t} \geqslant 1\}$$

to happen when $N_{\star,t}/\tilde{N}_{\star,t} > 1 - \frac{3}{2C}$, the following condition must hold:

$$N_{a,t} \leqslant \frac{6\ln t}{\Delta_a^2}. \tag{A.4}$$

1.1.3 Summary

Recall that being worse than (or equal to) the target implicitly requires that $N_{a,t} \geqslant \tilde{N}_{a,t}$. Therefore, Target-UCB cannot be worse than its target with respect to sub-optimal action a if $\tilde{N}_{a,t} > \frac{6\ln t}{\Delta_a^2}$.

1.2 Sub-optimal plays when better

If $N_{a,t}/\tilde{N}_{a,t} < 1$ for sub-optimal action a, then we can say that Target-UCB has played better than its target up to time t, with respect to action a. We want to bound the selection of action a in this situation. Again, we consider that either A) $N_{\star,t}/\tilde{N}_{\star,t} \leqslant 1 - \frac{3}{2C}$ or B) $N_{\star,t}/\tilde{N}_{\star,t} > 1 - \frac{3}{2C}$.

1.2.1 Case A)

When $N_{\star,t}/\tilde{N}_{\star,t} \leqslant 1 - \frac{3}{2C}$, Equation 4 simplifies to

$$\Delta_a \leqslant \sqrt{\frac{3\ln t}{2N_{a,t}}} + \sqrt{\frac{C\ln t}{N_{a,t}}}.$$

Therefore, in order for

$$\mathbb{I}\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell, N_{a,t}/\tilde{N}_{a,t} < 1\}$$

to happen when $N_{\star,t}/\tilde{N}_{\star,t} \leqslant 1 - \frac{3}{2C}$, the following condition must hold:

$$N_{a,t} \leqslant \frac{(\sqrt{3/2} + \sqrt{C})^2 \ln t}{2\Delta_a^2}. \tag{A.5}$$

1.2.2 Case B)

When $N_{\star,t}/\tilde{N}_{\star,t} > 1 - \frac{3}{2C}$, either 1) $\frac{N_{a,t}}{\tilde{N}_{a,t}} \leqslant \frac{N_{\star,t}}{\tilde{N}_{\star,t}}$ or 2) $\frac{N_{a,t}}{\tilde{N}_{a,t}} > \frac{N_{\star,t}}{\tilde{N}_{\star,t}}$.

Case B.1) When $\frac{N_{a,t}}{\tilde{N}_{a,t}} \leqslant \frac{N_{\star,t}}{\tilde{N}_{\star,t}}$, the condition given by Equation 4 simplifies to

$$\Delta_a \leqslant \sqrt{\frac{3\ln t}{2N_{a,t}}} + \sqrt{\frac{C\ln t}{N_{a,t}}} + \sqrt{\frac{3\ln t}{2N_{\star,t}}}.$$

If $N_{a,t} \leqslant N_{\star,t}$, we have that in order for

$$\mathbb{I}\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell, N_{a,t}/\tilde{N}_{a,t} < 1\}$$

to happen when $N_{\star,t}/\tilde{N}_{\star,t} > 1 - \frac{3}{2C}$, the following condition must hold:

$$N_{a,t} \leqslant \frac{(\sqrt{6} + \sqrt{C})^2 \ln t}{\Delta_a^2}. \tag{A.6}$$

As for the situation where $N_{a,t} > N_{\star,t}$, we have per Assumption 1 that

$$N_{\star,t} \geqslant \frac{\alpha_a}{1-\alpha_a} N_{a,t}$$

such that

$$N_{a,t} \leqslant \frac{(\sqrt{3/2} + \sqrt{C} + \sqrt{3/2}\sqrt{(1-\alpha_a)/\alpha_a})^2 \ln t}{\Delta_a^2}. \tag{A.7}$$

More precisely, this leads $N_{a,t} \leqslant \frac{(\sqrt{6}+\sqrt{C})^2 \ln t}{\Delta_a^2}$ for $\alpha_a \geqslant \frac{1}{2}$. By definition of *being better than the target*, we therefore have $N_{a,t} \leqslant \frac{(\sqrt{6}+\sqrt{C})^2 \ln t}{\Delta_a^2} \wedge \tilde{N}_{a,t}$ for $\tilde{N}_{a,t} \leqslant \frac{(\sqrt{6}+\sqrt{C})^2 \ln t}{\Delta_a^2}$ or $\frac{1}{2} \leqslant \alpha_a \leqslant 1$, and $N_{a,t} \leqslant \frac{(\sqrt{3/2}+\sqrt{C}+\sqrt{3/2}\sqrt{(1-\alpha_a)/\alpha_a})^2 \ln t}{\Delta_a^2} \wedge \tilde{N}_{a,t}$ otherwise.

Case B.2) When $\frac{N_{a,t}}{\tilde{N}_{a,t}} > \frac{N_{\star,t}}{\tilde{N}_{\star,t}} > 1 - \frac{3}{2C}$, then the condition given by Equation 4 simplifies to

$$\Delta_a \leqslant \sqrt{\frac{3\ln t}{2N_{a,t}}} + \sqrt{\frac{3\ln t}{2N_{a,t}}} + \sqrt{\frac{3\ln t}{2N_{\star,t}}}.$$

If $N_{a,t} \leqslant N_{\star,t}$, we have that in order for

$$\mathbb{I}\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell, N_{a,t}/\tilde{N}_{a,t} < 1\}$$

to happen, the following condition must hold:

$$N_{a,t} < \frac{14\ln t}{\Delta_a^2}.$$

By definition of *being better than the target*, we have $N_{a,t} \leqslant \left(\frac{14\ln t}{\Delta_a^2} \wedge \tilde{N}_{a,t}\right)$. As for the situation where $N_{a,t} > N_{\star,t}$, this leads to

$$\sqrt{\frac{C\ln t}{N_{a,t}}}\sqrt{1 - \left(\frac{N_{a,t}}{\tilde{N}_{a,t}} \wedge 1\right)} - \sqrt{\frac{C\ln t}{N_{\star,t}}}\sqrt{1 - \left(\frac{N_{\star,t}}{\tilde{N}_{\star,t}} \wedge 1\right)} \leqslant 0$$

such that

$$\Delta_a \leqslant \sqrt{\frac{3\ln t}{2N_{a,t}}} + \sqrt{\frac{3\ln t}{2N_{\star,t}}} \leqslant \sqrt{\frac{3\ln t}{2N_{\star,t}}} + \sqrt{\frac{3\ln t}{2N_{\star,t}}}.$$

In order for this to occur, we need

$$N_{\star,t} \leqslant \frac{6\ln t}{\Delta_a^2}$$

to hold. Thanks to Assumption 1, this is not possible for $N_{\star,t}/\tilde{N}_{\star,t} > 1 - \frac{3}{2C}$.

1.2.3 Summary

Recall that being better than the target implicitly requires that $N_{a,t} < \tilde{N}_{a,t}$. Therefore, for $\tilde{N}_{a,t} \leqslant \frac{(\sqrt{6}+\sqrt{C})^2\ln t}{\Delta_a^2}$ or $\alpha_a \geqslant 1/2$, we have $N_{a,t} < \tilde{N}_{a,t} \wedge \frac{(\sqrt{6}+\sqrt{C})^2\ln t}{\Delta_a^2}$; otherwise $N_{a,t} \leqslant \frac{(\sqrt{3/2}+\sqrt{C}+\sqrt{3/2}\sqrt{(1-\alpha_a)/\alpha_a})^2\ln t}{\Delta_a^2} \wedge \tilde{N}_{a,t}$.

1.3 Bounding the nonoccurrence of events

We can decompose

$$\sum_{t=A+1}^{T-1} \mathbb{I}\{\bar{E}_t^a \cup \bar{E}_t^\star\} \leqslant \sum_{t=1}^{T} \mathbb{I}\{\bar{E}_t^a\} + \sum_{t=1}^{T} \mathbb{I}\{\bar{E}_t^\star\}.$$

Then, using Chernoff-Hœffding and a simple union bound, we have that for $a' \in \mathcal{A}$,

$$\mathbb{P}\left[\exists s \in \{1,\dots,t\} : m_{a',t} - \mu_{a'} \geqslant \sqrt{\frac{c\ln t}{2N_{a',t}}}\right] \leqslant \frac{1}{t^{c-1}} \quad \forall c > 1.$$

Therefore, we can find that for $c = 3$,

$$\mathbb{E}\left[\sum_{t=A+1}^{T-1} \mathbb{I}\{\bar{E}_t^a \cup \bar{E}_t^\star\}\right] \leqslant \sum_{t=1}^{T} \mathbb{P}\left[\bar{E}_t^a\right] + \sum_{t=1}^{T} \mathbb{P}\left[\bar{E}_t^\star\right] \leqslant 2\sum_{t=1}^{T} t^{-2} = \frac{\pi^2}{3}. \tag{A.8}$$

1.4 Putting everything together

Taking the expectation of Equation A.1 and using Equation A.8, we have

$$\begin{aligned}\mathbb{E}[N_{a,T}] &\leqslant \ell + \mathbb{E}\Big[\sum_{t=\ell}^{T-1} \mathbb{I}\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell\}\Big] + \mathbb{E}\Big[\sum_{t=A+1}^{T-1} \mathbb{I}\{\bar{E}_t^a \cup \bar{E}_t^\star\}\Big] \\ &\leqslant \ell + \mathbb{E}\Big[\sum_{t=\ell}^{T-1} \mathbb{I}\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell\}\Big] + \frac{\pi^2}{3}.\end{aligned}$$

For $\tilde{N}_{a,t} \leqslant \frac{6\ln t}{\Delta_a^2}$ and $\ell = \frac{6\ln T}{\Delta_a^2} + 1$ or for $\tilde{N}_{a,t} > \frac{(\sqrt{6}+\sqrt{C})^2 \ln t}{\Delta_a^2}$ and $\ell = \frac{(\sqrt{6}+\sqrt{C})^2 \ln T}{\Delta_a^2} + 1$, we have

$$\sum_{t=\ell}^{T-1} \mathbb{I}\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell\} = 0.$$

For $\frac{6\ln t}{\Delta_a^2} \leqslant \tilde{N}_{a,t} \leqslant \frac{(\sqrt{6}+\sqrt{C})^2 \ln t}{\Delta_a^2}$ and $\ell = 1$ we have

$$\sum_{t=\ell}^{T-1} \mathbb{I}\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geqslant \ell\} < \tilde{N}_{a,T}.$$

Putting all this together and summing over sub-optimal actions concludes the proof of Theorem 1.

2 UCB target policy

The expected number of sub-optimal plays of action a after t episodes using target policy UCB is upper bounded by

$$\mathbb{E}[\tilde{N}_{a,t}] \leqslant \frac{8\ln t}{\Delta_a^2} + 1 + \frac{\pi^2}{3}.$$

Let us introduce the following random variable

$$X_t = \tilde{N}_{a,t} - \mathbb{E}[\tilde{N}_{a,t}].$$

By construction, $|X_t - X_{t-1}| \leqslant 1$. Thus, by an application of Azuma-Hœffding's inequality for martingales, we obtain that for all $\delta \in (0,1)$, with probability higher than $1-\delta$,

$$\tilde{N}_{a,t} - \mathbb{E}[\tilde{N}_{a,t}] < \sqrt{2t\ln(1/\delta)}.$$

Therefore we have that

$$\tilde{N}_{a,t} \leqslant \frac{8\ln t}{\Delta_a^2} + 1 + \frac{\pi^2}{3} + \sqrt{2t\ln(1/\delta)},$$

which allows to satisfy Assumption 1 for some $t \geqslant c_\Delta$ with probability higher than $1-\delta$.

Example 1. *Consider a two-actions setting with $\Delta_a = 0.3$. With probability higher than 0.9, using $C = 3$, we have $N_{\star,t} \geqslant \frac{12\ln t}{\Delta_a^2}$ for*

$$t \geqslant \frac{20\ln t}{\Delta^2} + 1 + \frac{\pi^2}{3} + \sqrt{2t\ln(10)}.$$

For example, this happens for $t \geqslant 1800$, which also leads to $N_{\star,t} \geqslant 0.58 N_{a,t}$. Note that this estimate is higly conservative and that experiments suggest that the assumption is respected when following a UCB target with $C = 2$.

2.1 α-optimal

The expected number of optimal plays after t episodes using such target policy is lower bounded by

$$\mathbb{E}[\tilde{N}_{\star,t}] \geqslant \alpha t.$$

Let us introduce the following random variable

$$X_t = \tilde{N}_{\star,t} - \mathbb{E}[\tilde{N}_{\star,t}].$$

By construction, $|X_t - X_{t-1}| \leqslant \alpha \vee (1-\alpha)$. Thus, by an application of Azuma-Hœffding's inequality for martingales, we obtain that for all $\delta \in (0,1)$, with probability higher than $1-\delta$,

$$\tilde{N}_{\star,t} - \mathbb{E}[\tilde{N}_{\star,t}] > \alpha\sqrt{2t\ln(1/\delta)} \vee (1-\alpha)\sqrt{2t\ln(1/\delta)}.$$

Therefore we have that

$$\begin{aligned} &\tilde{N}_{\star,t} > \alpha t - \alpha\sqrt{2t\ln(1/\delta)} && \text{for } 1/2 \leqslant \alpha \geqslant 1 \\ &\tilde{N}_{\star,t} > \alpha t - (1-\alpha)\sqrt{2t\ln(1/\delta)} && \text{for } 0 < \alpha < 1/2 \end{aligned}$$

which allows to satisfy Assumption 1 probability higher than $1-\delta$.

3 Additional Results

3.1 Target-UCB as a target

Figure A1 illustrates the performance of a small clique of 4 Target-UCB agents on a two-actions setting. Like for Figure 3 in the main paper, the clique size is selected according to Equation (6) for $\delta = 0.001$.

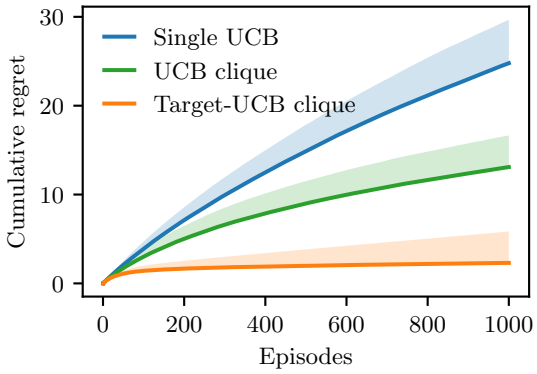

Figure A1: Single UCB and UCB clique of 4 agents vs Target-UCB clique of 4 agents (3 neighbours) on a two-actions setting ($\mu_\star = 0.9, \Delta_a = 0.1$).

3.2 Playing with humans

Figure A2 is simply a second version of Figure 5, but with no pair (Target-UCB and human target) omitted. Figure A3 shows the performance of a clique of 4 Target-UCB agents against two cliques of 4 human agents, with the first (green) one also shown in Figure 6. The second human clique shown here has much higher variance between players, but both human cliques seem to accumulate regret at a very similar rate, on average.

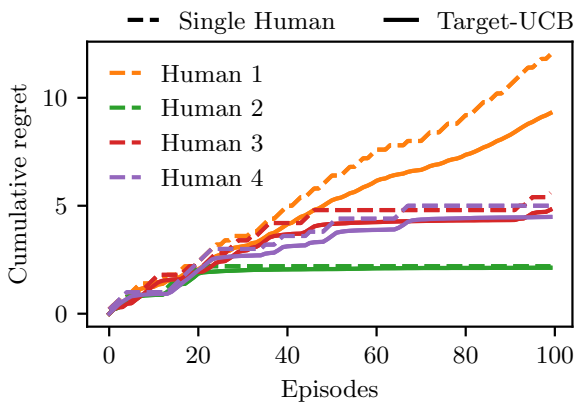

Figure A2: Target-UCB with human targets on a two-actions setting ($\mu_\star = 0.6$, $\Delta_a = 0.2$). The fourth human target and Target-UCB pair (purple) shown here was omitted in Figure 5.

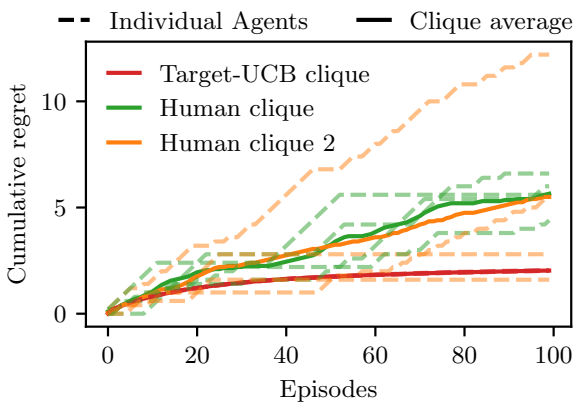

Figure A3: Cliques of humans vs Target-UCB (4 agents) on a two-actions setting ($\mu_\star = 0.6$, $\Delta_a = 0.2$). The second human clique shown here was omitted in Figure 6.

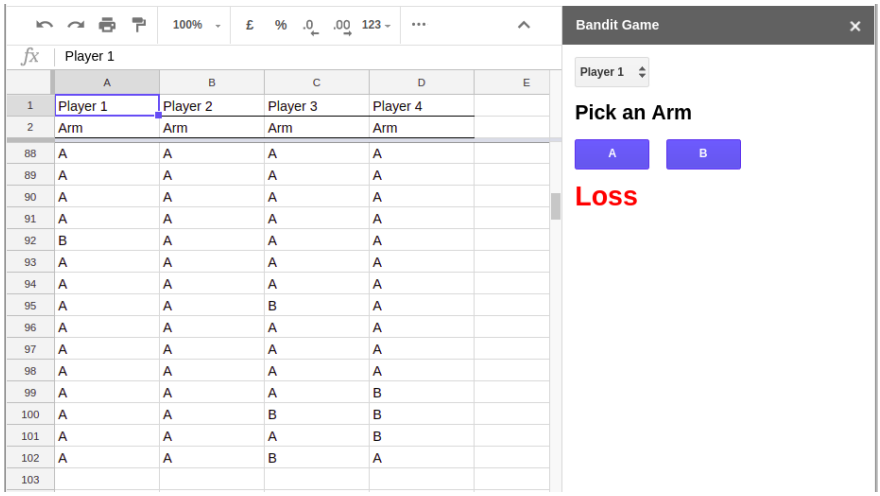

Figure A4: Screen capture of the player interface used for gathering the human clique data presented in Section 6.

4 Methodology of experiments with human subjects

In this section we describe the general process of gathering the human player data on the 2-actions setting considered in Section 6. Two versions were considered: 1) humans playing along and 2) a clique of four humans having access to each others' actions. Note that the human players used for the individual runs depicted in Figure A2 are the same as those used for the second human clique in Figure A3.

Bandit setting Bernoulli reward distributions with $\mu_\star = 0.6$ and $\mu_a = 0.4$ were used in all experiments. These distributions were identical and static for all players, but each reward was randomly and independently sampled for each action selection. This means that two players selecting the same action on the same episode could potentially obtain different results (i.e. one getting a win whereas the other gets a loss).

For each episode, each player clicked on their desired action (A or B), after which the buttons became deactivated and greyed-out. After a few seconds, the buttons were reactivated, and the reward for the selected action was displayed: either "Win!" in green or "Loss" in red. In the clique experiment, once all players had selected their action for an episode, a new row in the spreadsheet was displayed, showing the action played by each user. This marked the start of the next episode. Both the single player and the clique experiments were conducted over 100 episodes.

Game interface A player interface was created using *Google Apps Script*[1] and the *Google Sheets API*[2] in order to simplify the process of interacting with the bandit setting and accessing the actions played by other players. Figure A4 shows a screen capture of the player interface used for the human clique experiment. The interface for the single player experiment is almost identical, except there is a unique "Player" column (each player plays in its own sheet).

Single player experiment The players were gathered in a room and each was playing on a separate computer, on their dedicated interface (sheet), with no communication between them.

Clique player experiment The players were gathered in a room and each was playing on a separate computer, with no communication between them. Player IDs (i.e 1 to 4) were assigned randomly and silently, such that players did not know which person was associated to which player ID.

[1] `https://developers.google.com/apps-script/`
[2] `https://developers.google.com/sheets/api/`