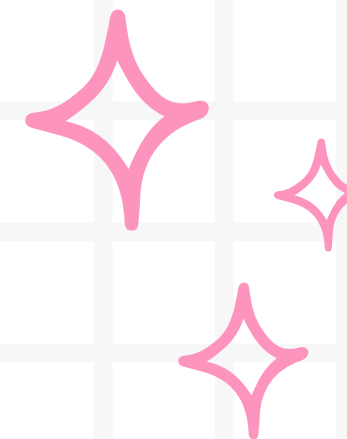


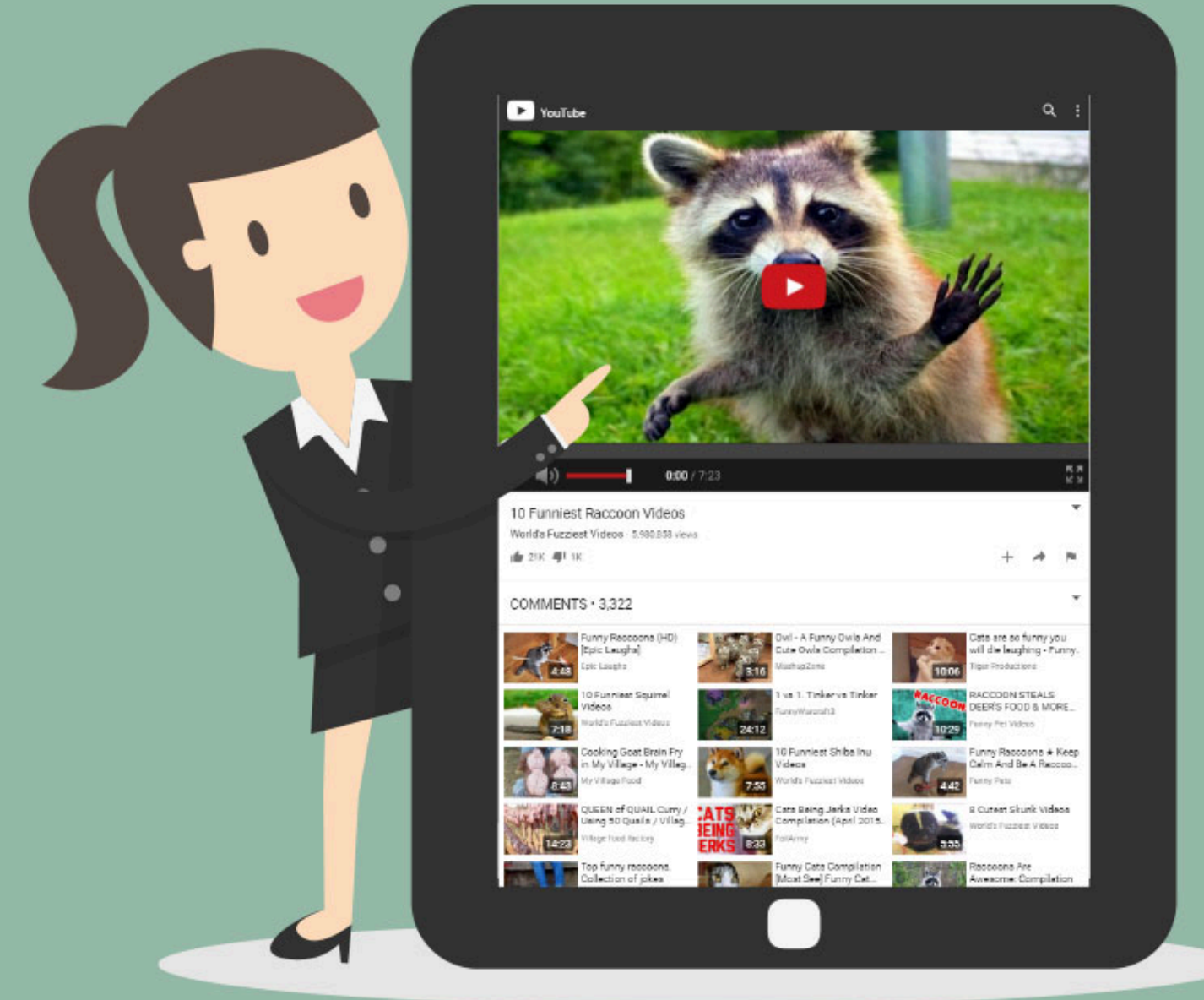
BANDITS MEET MECHANISM DESIGN TO COMBAT CLICKBAIT IN ONLINE RECOMMENDATION



Benjamín Blancaire, Jean Fuentes, Marcos Santelices

PLATAFORMAS DE RECOMENDACIÓN

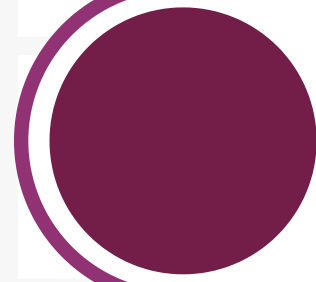
NETFLIX



CLiCKBAiT



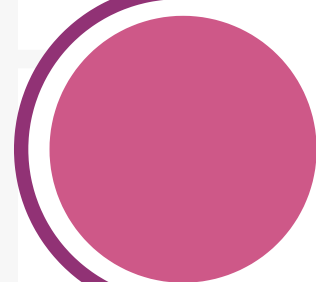
CLiCKBAiT



EXPERiENCiA DEL USUARIO DEGRADADA



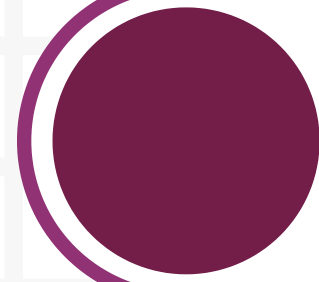
DAÑANDO LA REPUTACiÓN DE LA PLATAFORMA



DESAFÍOS EN EL APRENDiZAJE DE LOS ALGORiTMOS

ESTADO DEL ARTE

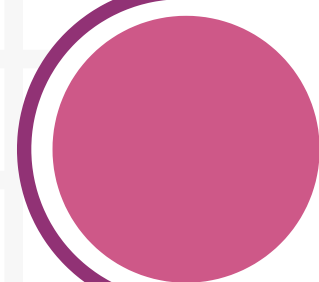
Multi-Armed Bandits (MAB)



BRAZOS O "ARMS"



RECOMPENSAS

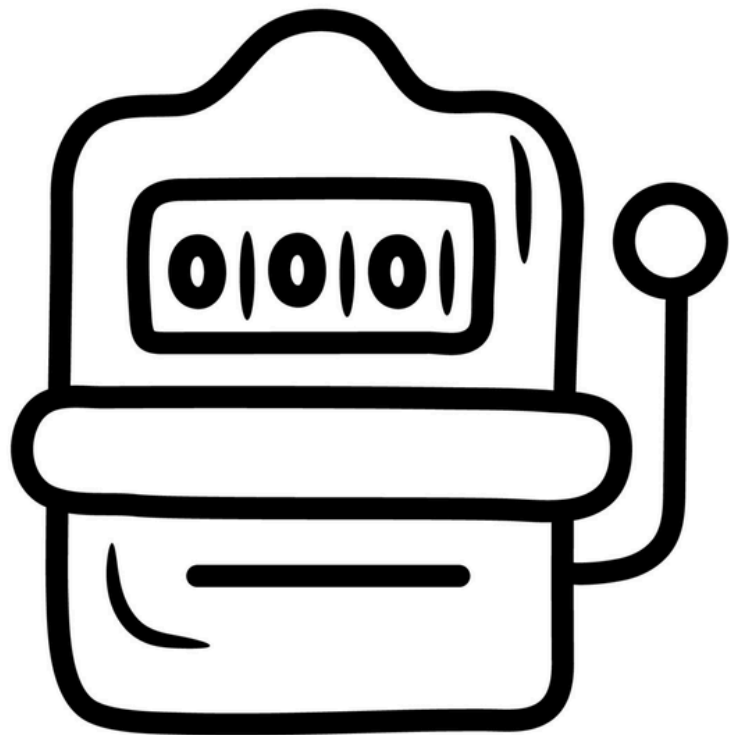


POST CLICK REWARDS



DECISIONES DEL ALGORITMO

Desafío



Arm Bandit

1

MANIPULACIÓN POR PARTE
DE LOS VENDEDORES

2

DIMENSIÓN **ESTRATÉGICA**
ADICIONAL

Trabajos relacionados



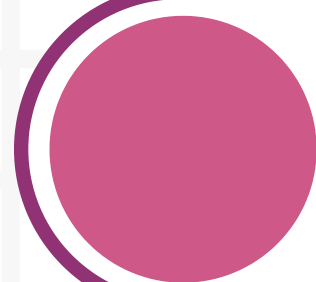
MODELO ESTRATÉGICO DE BANDITS

BRAVERMAN ET AL. (2019)



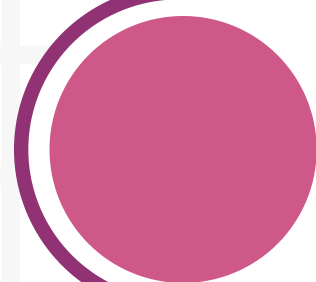
BANDITS ROBUSTOS A MANIPULACIONES

FENG ET AL. (2020) Y DONG ET AL. (2022)



DISEÑO DE SUBASTAS EN BANDITS

BABAI OFF ET AL. (2009, 2015)



DESAFÍOS DE INCENTIVOS

NISAN Y RONEN (1999), FREEMAN ET AL. (2020) Y ZHANG Y CONITZER (2021)

Contribución

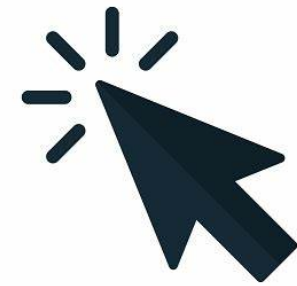
**STRATEGIC
CLICK BANDIT**

**EQUILIBRIO
DE NASH**

UCB-S

Propuesta

Strategic Click Bandit



The Strategic Click-Bandit Problem

Model 1: The Strategic Click-Bandit Problem

```
1 Learner commits to algorithm  $M$ , which is shared with all arms
2 Arms choose strategies  $(s_1, \dots, s_K) \in [0, 1]^K$  (unknown to  $M$ )
3 for  $t = 1, \dots, T$  do
4   | Algorithm  $M$  selects arm  $i_t \in [K]$ 
5   | Arm  $i_t$  is clicked with probability  $s_{i_t}$ , i.e.,  $c_{t,i_t} \sim \text{Bern}(s_{i_t})$ 
6   | if  $i_t$  was clicked ( $c_{t,i_t} = 1$ ) then
7   |   | Arm  $i_t$  receives utility 1 from the click
8   |   |  $M$  observes post-click reward  $r_{t,i_t}$  drawn from a distribution with mean  $\mu_{i_t}$ 
```

Función de utilidad

Política

(A1) $u: [0, 1] \times [0, 1] \rightarrow \mathbb{R}$ is L -Lipschitz w.r.t. the ℓ_1 -norm.

(A2) $u^*(\mu) := \max_{s \in [0, 1]} u(s, \mu)$ is monotonically increasing.

(A3) $s^*(\mu) := \operatorname{argmax}_{s \in [0, 1]} u(s, \mu)$ is H -Lipschitz and is bounded away from zero.

Brazo

$$v_i(M, s_i, s_{-i}) := \mathbb{E}_M \left[\sum_{t=1}^T \mathbb{1}_{\{i_t=i\}} c_{t,i} \right] \quad v_i(M, \sigma_i, \sigma_{-i}) := \mathbb{E}_{s \sim \sigma} [v_i(M, s_i, s_{-i})]$$

Equilibrio de Nash (NE)

“Conjunto de estrategias en el que ningún brazo puede mejorar unilateralmente su utilidad”

Lemma 5.1. *For any post-click rewards μ_1, \dots, μ_K , there always exists a (possibly mixed) Nash equilibrium for the arms under the UCB-S mechanism.*

Strategic Regret

- Regret es el costo de de tomar desiciones sub óptimas, en este caso se evalúe el desempeño de acuerdo a las estrategias de los "agentes" brazos
- Se define como la suma en el total de rondas de las diferencias entre el valor de utilidad del mejor brazo vs la utilidad del brazo seleccionado en la ronda

$$R_T(M, \mathbf{s}) := \mathbb{E} \left[\sum_{t=1}^T u(s^*, \mu^*) - u(s_{i_t}, \mu_{i_t}) \right]$$

Strategic Regret

Regret fuerte (RT+)

Máximo regret entre todos los posibles equilibrios de Nash.

$$R_T^+(M) := \max_{\sigma \in \text{NE}(M)} R_T(M, \sigma),$$

Regret débil (RT-)

Mínimo regret posible en cualquier equilibrio de Nash.

$$R_T^-(M) := \min_{\sigma \in \text{NE}(M)} R_T(M, \sigma),$$

Limitaciones de algoritmos Incentive-Unware

Proposition 4.1. *Let μ -Oracle be the algorithm with oracle knowledge of μ_1, \dots, μ_K that plays $i_t = \operatorname{argmax}_{i \in [K]} \mu_i$ in every round t , whereas (s, μ) -Oracle is the algorithm with oracle knowledge of μ_1, \dots, μ_K and s_1, \dots, s_K that always plays $i_t = \operatorname{argmax}_{i \in [K]} u(s_i, \mu_i)$ with ties broken in favor of the larger μ . We then have*

(i) *Under every equilibrium $\sigma \in \text{NE}(\mu\text{-Oracle})$, the μ -Oracle suffers regret $\Omega(\beta T)$, i.e.,*

$$R_T^-(\mu\text{-Oracle}) = \Omega(\beta T).$$

(ii) *Under every $\sigma \in \text{NE}((s, \mu)\text{-Oracle})$, the (s, μ) -Oracle suffers regret $\Omega(\min\{\beta, \eta\}T)$, i.e.,*

$$R_T^-((s, \mu)\text{-Oracle}) = \Omega(\min\{\beta, \eta\}T).$$

Los algoritmos que escogen los brazos con recompensa post-click más alta y con mayor utilidad, su regret crece como $\Omega(T)$ (salvo una constante)

UCB-S (Upper Confidence Bound con Screening)

Aprendizaje en línea

Estimar las recompensas post-clic (μ_i) a través de las tasas de clics estratégicamente manipuladas (S_i).

Diseño de incentivos

Alinear las estrategias de los brazos con los objetivos del sistema de recomendación.

Características UCB-S

Screening

if $\bar{s}_{i_t}^t < \min_{\mu \in [\underline{\mu}_{i_t}^t, \bar{\mu}_{i_t}^t]} s^*(\mu)$ **or** $\underline{s}_{i_t}^t > \max_{\mu \in [\underline{\mu}_{i_t}^t, \bar{\mu}_{i_t}^t]} s^*(\mu)$ **then**
| Ignore arm i_t in future rounds: $A_t \leftarrow A_{t-1} \setminus \{i_t\}$

Parámetros

$$\begin{aligned}\underline{s}_i^t &= \hat{s}_i^t - \sqrt{2 \log(T) / n_t(i)}, & \bar{s}_i^t &= \hat{s}_i^t + \sqrt{2 \log(T) / n_t(i)}, \\ \underline{\mu}_i^t &= \hat{\mu}_i^t - \sqrt{2 \log(T) / m_t(i)}, & \bar{\mu}_i^t &= \hat{\mu}_i^t + \sqrt{2 \log(T) / m_t(i)}.\end{aligned}$$

Características UCB-S

Caracterización Equilibrio de Nash

Theorem 5.2. *For all $s \in \text{supp}(\sigma)$ with $\sigma \in \text{NE}(\text{UCB-S})$ and all $i \in [K]$:*

$$s_i = s^*(\mu_i) + \mathcal{O} \left(H \cdot \max \left\{ \Delta_i, \sqrt{\frac{K \log(T)}{T}} \right\} \right).$$

In particular, for all arms $i^ \in [K]$ with $\Delta_{i^*} = 0$, i.e., maximal post-click rewards:*

$$s_{i^*} = s^*(\mu_{i^*}) + \mathcal{O} \left(H \sqrt{\frac{K \log(T)}{T}} \right).$$

Bajo el mecanismo UCB-S, todo equilibrio de Nash es tal que las estrategias de los brazos con recompensas post-click maximales difieren con la estrategia deseada en a lo más $\tilde{\mathcal{O}}(\sqrt{K/T})$

Características UCB-S

Cota superior Regret Fuerte

Theorem 5.3. *Let $\Delta_i := \mu^* - \mu_i$ and let L and H denote the Lipschitz constants of $u(s, \mu)$ and $s^*(\mu)$, respectively. The strong strategic regret of UCB-S is bounded as*

$$R_T^+(\text{UCB-S}) = LH \cdot \mathcal{O} \left(\sqrt{KT \log(T)} + \sum_{i: \Delta_i > 0} \frac{\log(T)}{\Delta_i} \right). \quad (3)$$

In other words, the above regret bound is achieved under any equilibrium $\sigma \in \text{NE}(\text{UCB-S})$.

Bajo UCB-S, el regret fuerte está acotado superiormente por $\tilde{\mathcal{O}}(\sqrt{KT})$

Características UCB-S

Cota inferior Regret Débil

Theorem 5.5. *Let M be any mechanism with $\text{NE}(M) \neq \emptyset$. There exists a utility function u satisfying (A1)-(A3) and post-click rewards μ_1, \dots, μ_K such that for all Nash equilibria $\sigma \in \text{NE}(M)$:*

$$R_T(M, \sigma) = \Omega(\sqrt{KT}).$$

In other words, $R_T^-(M) = \Omega(\sqrt{KT})$.

Bajo UCB-S, el regret débil está acotado inferiormente por $\Omega(\sqrt{KT})$

Métricas utilizadas para evaluar el algoritmo

Utilidad del recomendador
($u(s, \mu)$)

Comportamiento
estratégico de los brazos

Regret estratégico
(RT)

Resultados

Rendimiento teórico del algoritmo

- UCB-S garantiza un regret estratégico sublineal $O(K/T)$ en cualquier equilibrio de Nash.
- Se minimiza el impacto de brazos subóptimos al limitar la frecuencia con la que son seleccionados.

Resultados

Validación experimental

- Brazos óptimos (μ) convergen rápidamente a $s \star (\mu_i)$, mientras que los subóptimos presentan desviaciones proporcionales a Δ_i .
- Algoritmos sin incentivos (como UCB estándar) sufren un regret significativamente mayor, ya que no desincentivan estrategias desalineadas.

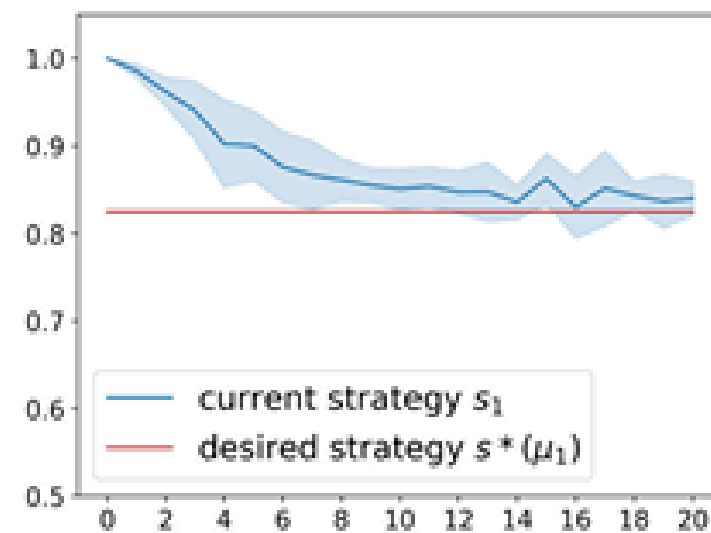
Resultados

Validación experimental

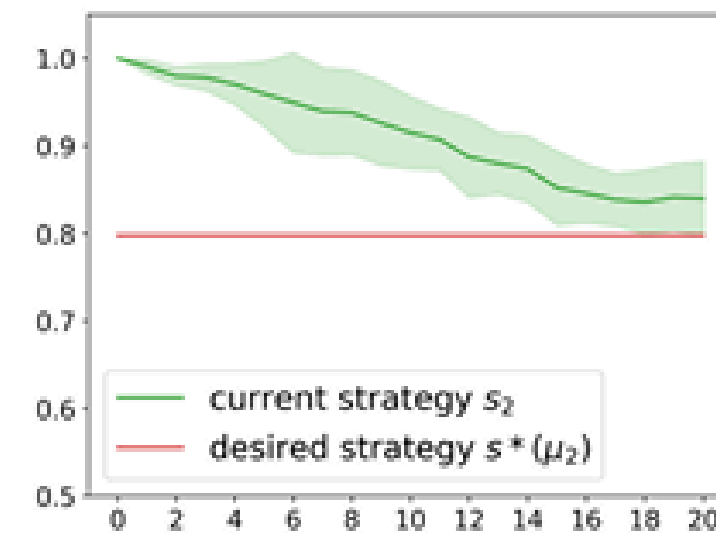
- Incluso cuando los brazos usan estrategias adaptativas simples (como gradient ascent), UCB-S logra mantener bajas pérdidas al ajustarse a las dinámicas del entorno.

Resultados

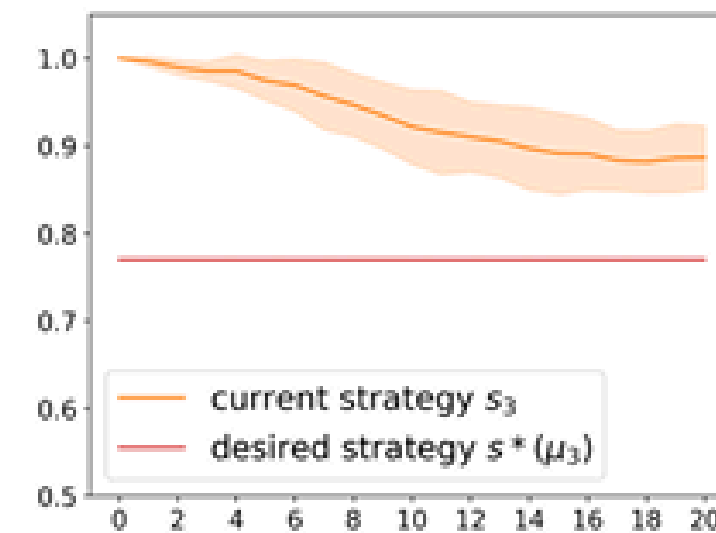
Validación experimental



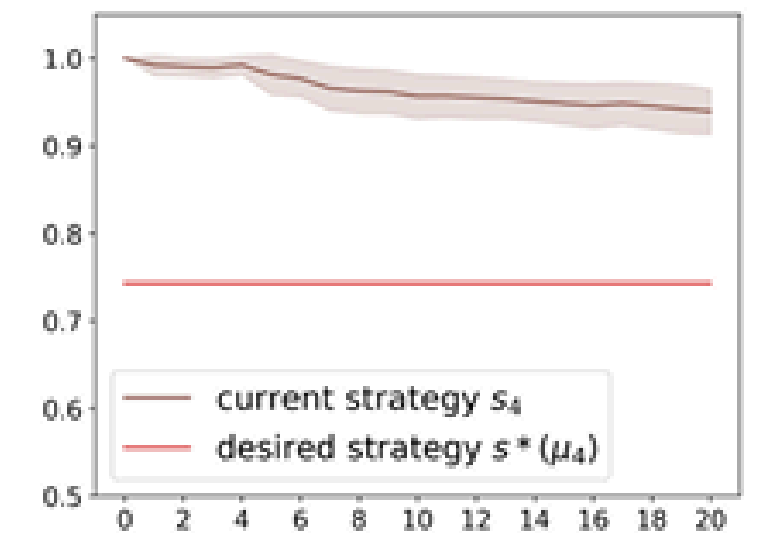
(a) Optimal arm with mean $\mu_1 = 0.75$.



(b) Suboptimal arm with mean $\mu_2 = 0.725$.



(c) Suboptimal arm with mean $\mu_3 = 0.7$.



(d) Suboptimal arm with mean $\mu_4 = 0.675$.

Resultados

Validación experimental

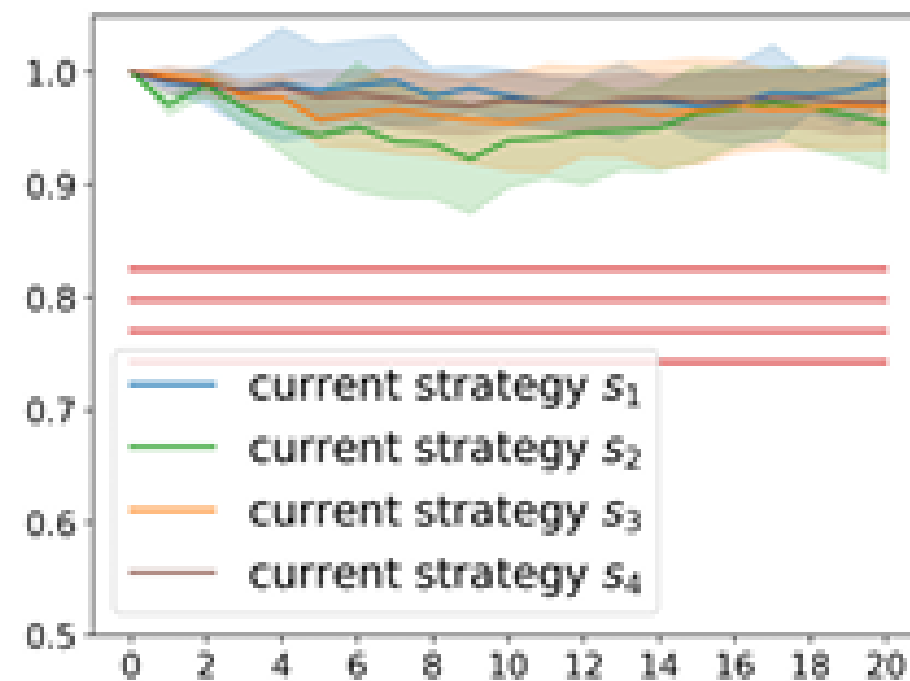


Figure 3: Strategic arm behavior when interacting with incentive-*unaware* standard UCB.

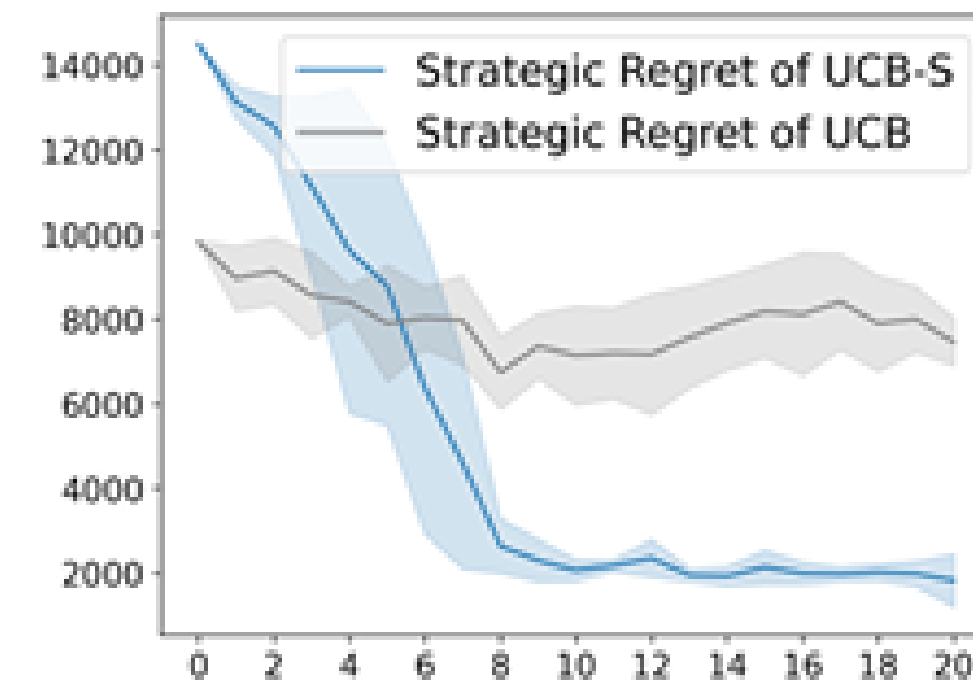
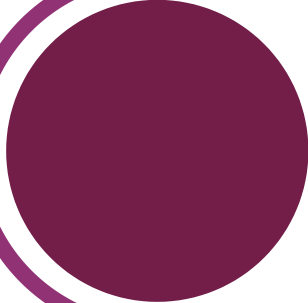
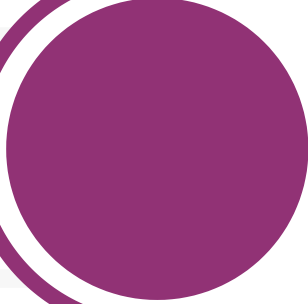
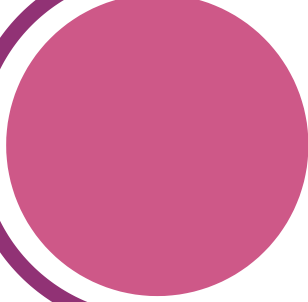
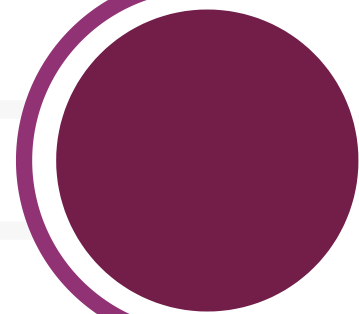


Figure 4: Strategic regret of UCB-S and standard UCB as arms adapt their strategies.

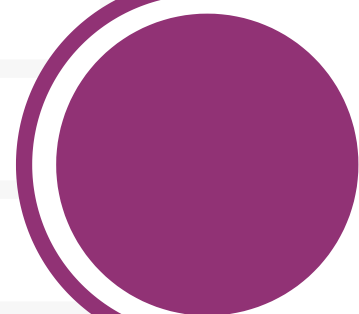
Trabajo a futuro

-  ESTUDIAR SI LOS INCENTIVOS SIGUEN SIENDO EFICACES CON ESTRATEGIAS DE BRAZOS ADAPTATIVAS.
-  ESTUDIAR SI ES POSIBLE GENERAR UN MECANISMO EN QUE EXISTAN DISTINTAS ESTRATEGIAS DESEABLES DE EQUILIBRIO DE NASH DOMINANTES.
-  ESTUDIAR SI CTR DEPENDE MÁS DE PARTE DEL USUARIO O DE INFORMACIÓN CONTEXTUAL.

Trabajo a futuro

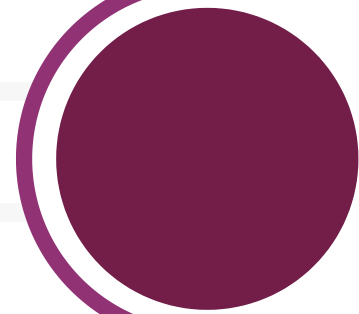


CONSIDERAR RECOMENDACIONES MULTI-SLOT, SE PUEDE ESCOGER DISTINTAS ACCIONES A LA VEZ.

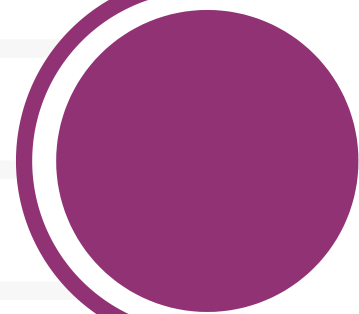


POTENCIALMENTE, EXTENDER EL FILTRO DE CONFIANZA DE CADA ÁRBOL A OTROS MÉTODOS DE APRENDIZAJE.

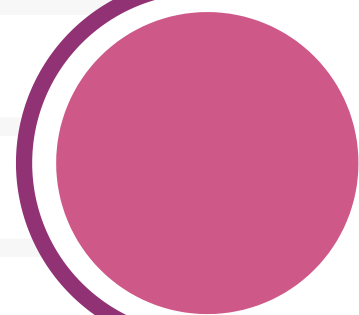
Nuestras observaciones



EL PAPER NO POSEE UNA FUENTE DATOS REALES



NO SE PRESENTA EL CÓDIGO REAL PARA REALIZAR LAS PRUEBAS. SOLO SE COMENTA SOBRE LOS RESULTADOS



NO HAY ANÁLISIS DE SENSIBILIDAD, EN EL SENTIDO DE QUE SE PRUEBA UN CASO EN PARTICULAR SIN CAMBIAR LA FUNCIÓN DE UTILIDAD

BIBLIOGRAFÍA

- Kleine Büning, T., Saha, A., Dimitrakakis, C., & Xu, H. (2024). Bandits Meet Mechanism Design to Combat Clickbait in Online Recommendation. Proceedings of the International Conference on Learning Representations (ICLR). Recuperado de <https://openreview.net/forum?id=lsxeNvYqCja>

Mark Braverman, Jieming Mao, Jon Schneider, and S Matthew Weinberg. Multi-armed bandit problems with strategic arms. In *Conference on Learning Theory*, pages 383–416. PMLR, 2019.

Zhe Feng, David Parkes, and Haifeng Xu. The intrinsic robustness of stochastic bandits to strategic manipulation. In *International Conference on Machine Learning*, pages 3092–3101. PMLR, 2020.

Jing Dong, Ke Li, Shuai Li, and Baoxiang Wang. Combinatorial bandits under strategic manipulations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 219–229, 2022.

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 79–88, 2009.

Moshe Babaioff, Robert D Kleinberg, and Aleksandrs Slivkins. Truthful mechanisms with implicit payment computation. *Journal of the ACM (JACM)*, 62(2):1–37, 2015.

Noam Nisan and Amir Ronen. Algorithmic mechanism design. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 129–140, 1999.

Rupert Freeman, David M Pennock, Chara Podimata, and Jennifer Wortman Vaughan. No-regret and incentive-compatible prediction with expert advice. *arXiv preprint arXiv:2002.08837*, 2020.

Hanrui Zhang and Vincent Conitzer. Incentive-aware pac learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5797–5804, 2021.