On Mitigating Affinity Bias through Bandits with Evolving Biased Feedback

Matthew Faw[1], Constantine Caramanis[2], Jessica Hoffmann[3]

[1]Georgia Tech, [2]UT Austin, [3]Google DeepMind

Georgia Institute of Technology
The University of Texas at Austin
Chandra Department of Electrical and Computer Engineering
Cockrell School of Engineering
DeepMind

Paper: https://arxiv.org/abs/2503.05662
Homepage: https://matthewfaw.github.io/
Email: mfaw3@gatech.edu

Problem Setup

- Affinity bias = Unconscious tendency to favor individuals similar to us
- Studies (e.g., [UC05, OA18, PS08, RBR19]) have shown affinity bias can:
  - Arise from (often *changing*) media portrayals, cultural conditioning, and affinity biases
  - Lead to undesirable opinion formation + self-reinforcing feedback loops

Our Goal: Investigate effects of evolving biases + feedback loops in sequential decision-making problems by introducing and studying a biased multi-armed bandit model capturing key features of affinity bias

Key Features to Model

1. The system has an initial, perhaps misleading, affinity for each action.
2. Selecting an action increases the system's affinity towards that action.
3. Selecting an action (slightly) decreases the system's affinity towards other actions

Affinity Bandit Model

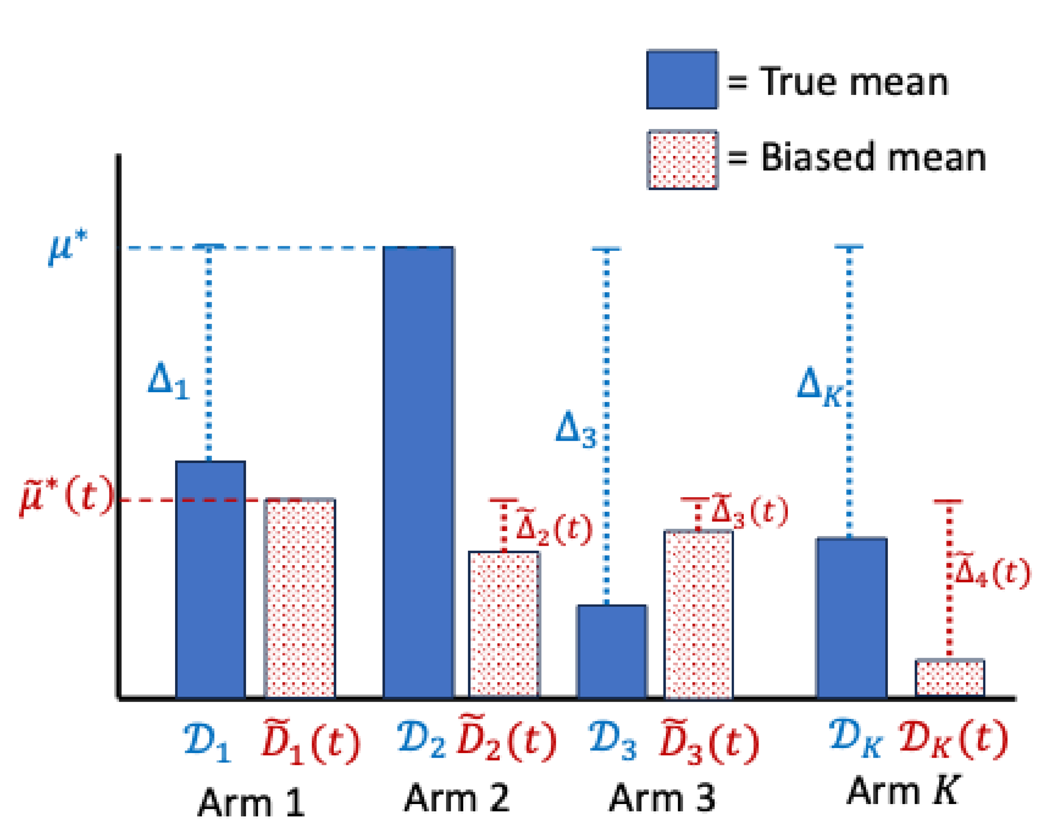
- K-armed Gaussian bandit instance $(\nu_i = N(\mu_i, 1))_{i\in[K]}$, $\Delta_i = \max_j \mu_j - \mu_i$
- Each arm $i \in [K]$ has an "initial bias" $T_i^{init} \geq 0$, $T^{init} = \sum_{i\in[K]} T_i^{init}$
- Rewards $R_{i,t} \sim \nu_i$ for each arm i are **unobserved**.
- Interaction model: For t=1,…,n
  - Select arm $A_t \in [K]$
  - Observe (biased) feedback $F_t \sim \tilde{\nu}_i(t) = N(\tilde{\mu}_i(t), 1)$
  - Update $T_{A_t}(t) \mathrel{+}= 1$
- Biased feedback model: if $A_t = i$, F_t has mean:

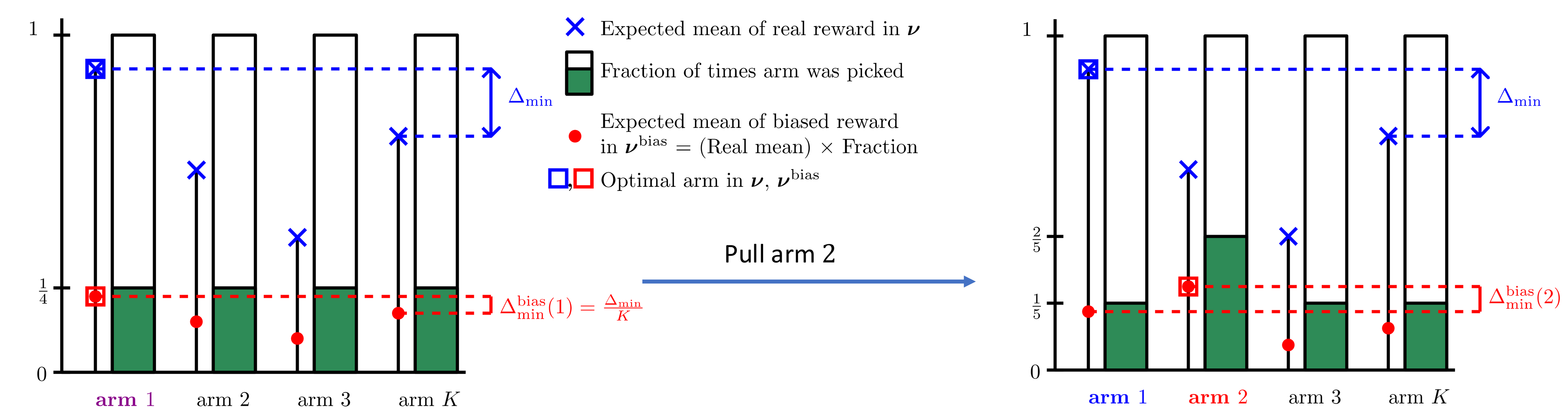
$$\tilde{\mu}_i(t) = \mu_i \cdot W_i(t) = \mu_i \cdot f\left(\frac{\underbrace{T_i^{init}}_{\text{initial bias}} + \overbrace{T_i(t-1)}^{}}{T^{init} + \underbrace{t-1}_{\text{Frac times } i \text{ played before } t}}\right) = \mu_i \cdot f\left(\frac{T_i^{bias}(t-1)}{t^{bias}-1}\right)$$

- $f(\cdot) \in [0,1]$ is *unknown*, bounded, L-Lipschitz

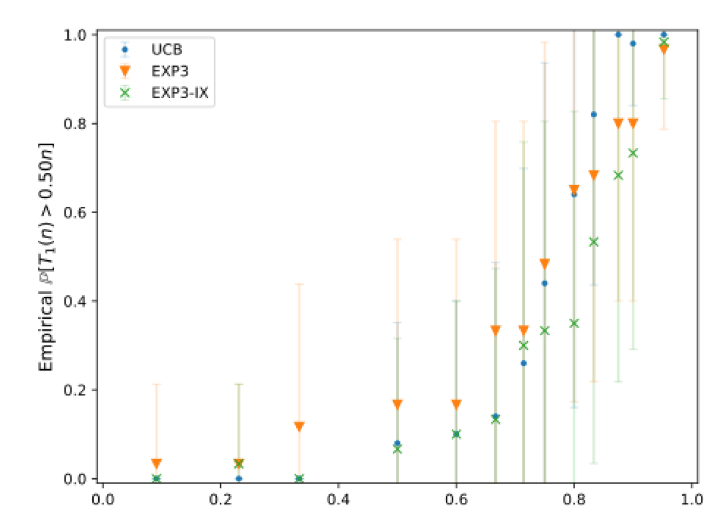
Objective: minimize regret w.r.t. **true** (unobserved) rewards:

$$R(n) = \sum_{t\in[n]} \max_{i\in[K]} \mathbb{E}[R_{i,t} - R_{A_t,t}] = \sum_{i\in[K]} \Delta_i \mathbb{E}[T_i(n)]$$



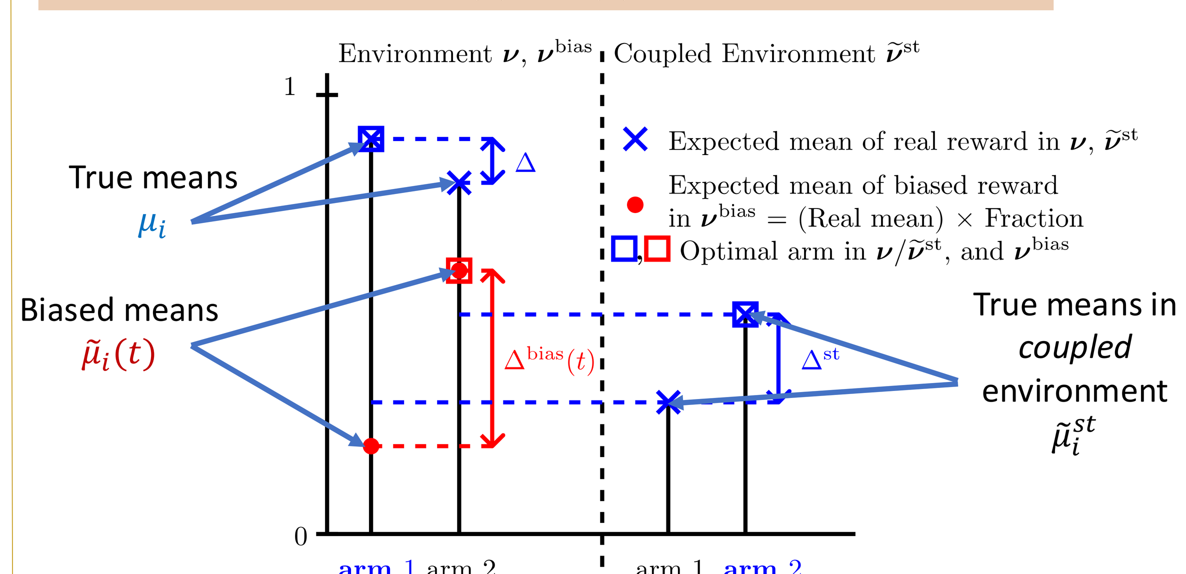


Linear Regret when Bias is Ignored



Theorem: There is a 2-armed (biased) bandit instance with $f(frac) = frac$, constant suboptimality gap and initial biases such that, for all n sufficiently large,

$$R(n) = \Omega(n).$$



Proof Outline:

- We begin by considering the following event:

$$\mathcal{B}_t = \bigcap_{s\in[t]} \{\underbrace{\tilde{\mu}_1(s) = \mu_1 W_1(s)}_{\text{Biased mean of } \textbf{optimal} \text{ arm}} \leq \tilde{\mu}_1^{st} < \tilde{\mu}_2^{st} \leq \underbrace{\mu_2 W_2(s) = \tilde{\mu}_2(s)}_{\text{Biased mean of } \textbf{suboptimal} \text{ arm}}\}$$

- Choose initial biases such that $\mathcal{B}_1$ is **deterministically** true
- As long as $\mathcal{B}_t$ is true, can couple samples between biased environment and "coupled" environment s.t.:

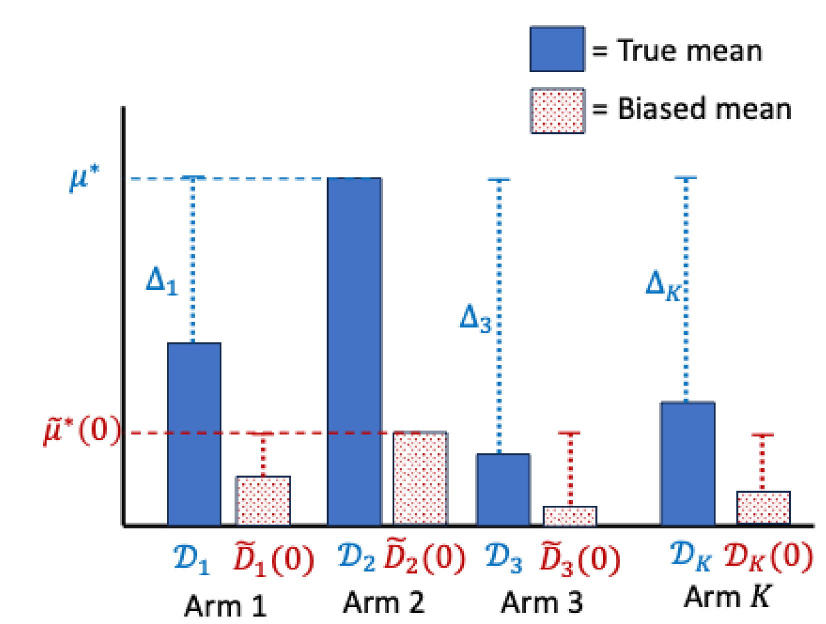
$$\tilde{T}_1^{st}(s) \geq T_1(s) \; \forall s \leq t$$

- $\mathcal{B}_t$ is challenging to reason about directly, since it depends on the dynamics in both environments. However, one can show that, for ε sufficiently small, there are times $1 < t_1 < \cdots < t_R = t$ s.t.:

$$\widetilde{\mathcal{B}}_t = \bigcap_{r\in[R]} \{\tilde{T}_1(t_r) \leq \varepsilon\, t_r\} \subseteq \mathcal{B}_t$$

- Can lower bound $\Pr[\widetilde{\mathcal{B}}_t] \geq .99$ using (anytime) high-probability guarantees for UCB
- Conclusion: since $\tilde{T}_1^{st}(n) \leq \varepsilon n$ with high probability, and $\widetilde{\mathcal{B}}_t$ implies that $T_1(n) \leq \tilde{T}_1^{st}(n)$, UCB must suffer linear regret in the biased environment.

Elimination + Round-Robin achieves Sublinear Regret



Algorithm 1 Elimination algorithm for unknown bias model

Require: Time horizon $n \in \mathbb{N}$, sampling schedule $m_r \approx \log(n)/\widetilde{\Delta}_r^2$, where $\widetilde{\Delta}_r = 2^{-r}$.
Let $\tau_0 = 0$, $t = 1$ and $\mathcal{A}_1 = [K]$
for $r = 1, 2, \ldots$ **do**
  for $\ell \in [|m_r|]$, $i \in \mathcal{A}_r$ in increasing order of index **do**
    Pull arm i, receive feedback Y_t, update $t \leftarrow t+1$.
  Compute $\widehat{\mu}_i(r)$, the empirical average of the feedback for arm i observed during round r
  Update active arms:
  $\mathcal{A}_{r+1} = \left\{i \in \mathcal{A}_r : \max_{j\in\mathcal{A}_r} \widehat{\mu}_j(r) - \widehat{\mu}_i(r) \leq \widetilde{\Delta}_r\right\}$
  Mark τ_r as the end time of round r

Theorem: Suppose Algorithm 1 is run for n time-steps in an Affinity bandit environment with reward means $\mu_i \in [0,1]$. If n is sufficiently large such that $\log(nK)/\log(\log(nK)) \gtrsim L(1 + \frac{T_{max}^{init} - T_{min}^{init}}{K})$ and $T_{max}^{init} \lesssim \log(nK)$, then the regret of Algorithm 1 is at most:

$$R(n) \lesssim f\left(\frac{1}{15\,K}\right)^{-2} \sum_{i:\Delta_i>0} \frac{\log(n)}{\Delta_i}$$

Proof Outline:

- Since arms played in round-robin manner, the biased feedback reweighting $W_i(t) \approx W_j(t)$ for all **active** i, j (assuming small initial biases)
- Thus, the ordering of the "feedback suboptimality gaps" $\widetilde{\Delta}_i(t) = \tilde{\mu}_{i^*}(t) - \tilde{\mu}_i(t)$ is (*roughly*) the same as the reward suboptimality gaps Δ_i.
- **Main challenge**: the above properties are inexact, so feedback suboptimality gap ordering may not be preserved at all time-steps. Need to ensure arms aren't mistakenly eliminated early!
- **Observation**: denote $\overline{W}_i(r)$ as the average reweighting of arm i during round r. Then, the average feedback during round r satisfies:

$$\mathbb{E}[\hat{\mu}_{i^*}(r) - \hat{\mu}_i(r)] = \mu_{i^*}\overline{W}_{i^*}(r) - \mu_i \overline{W}_i(r)$$
$$= \underbrace{(\mu_{i^*} - \mu_i)\overline{W}_{i^*}(r)}_{\text{A } \textbf{reweighting} \text{ of } \Delta_i} + \underbrace{\mu_i(\overline{W}_{i^*}(r) - \overline{W}_i(r))}_{\text{A } \textbf{bias} \text{ (can be + or −) Could change ordering of arms w.r.t. avg feedback!}}$$

- But our assumptions on bias model guarantees:

$$\mu_i|\overline{W}_{i^*}(r) - \overline{W}_i(r)| \lesssim \widetilde{\Delta}_r^2\, L\left(1 + \frac{T_{max}^{init} - T_{min}^{init}}{K}\right)\frac{\log(\log(nK))}{\log(nK)}$$

- As long as this term is $\ll \widetilde{\Delta}_r$ (the elimination criterion), bias term is negligible!

(Nearly tight) Instance-dependent Regret Lower Bound

Theorem: Fix any $K > 1$, initial biases $\left(T_i^{init}\right)_{i\in[K]}$, and time horizon n. Then, any "consistent" bandit policy in all unit-variance Gaussian environments with bounded suboptimality gaps must suffer regret at least:

$$R(n) \gtrsim \boxed{f\left(\mathcal{O}\left(\frac{\log(K)}{K}\right)\right)^{-2}} \sum_{i\in B} \frac{\log(n)}{\Delta_i} + \sum_{i\notin B:\Delta_i>0} \frac{\log(n)}{\Delta_i}$$

for some subset of arms $B \subseteq [K]$, $|B| = \Omega(K)$, as long as $\frac{\Delta_{max}}{\Delta_{min}} \leq poly(K)$.

Remarks:

- Lower bound holds against algorithms which know **exactly** (i) the bias model $f(\cdot)$, (ii) the initial biases T_i^{init}, and (iii) the time horizon n
- Our algorithm nearly achieves this bound without knowing the bias model or initial biases

Proof Outline:

- We build on the lower bound arguments of [KCG16, GMS19]
- Regret decomposition $\Longrightarrow$ suffices to lower bound a **constant fraction** of suboptimal arms $i \in [K]$ by:

$$\mathbb{E}[T_i(n)] \gtrsim \frac{K^2}{\log(K)^2} \frac{\log(n)}{\Delta_i^2}$$

Lemma 1 *(divergence decomposition + data processing)*: For any "consistent" bandit policy, any fixed time $n_0 \in [1, n]$ any suboptimal arm i, and any stopping time $\tau_i \in [n_0, n]$:

$$\mathbb{E}\left[\sum_{t=n_0+1}^{\tau_i} f\Bigg(\underbrace{\boxed{\frac{T_i^{bias}(t-1)}{t^{bias}-1}}}_{\text{Want} \lesssim \frac{1}{K}}\Bigg)^2 1\{A_t = i\}\right] \gtrsim \underbrace{\boxed{\frac{\mathbb{E}[\tau_i]}{n}}}_{\text{Want} \gtrsim const} \frac{\log(n)}{\Delta_i^2} - \underbrace{\boxed{n_0}}_{\text{Want} \lesssim \frac{\log(n)}{\Delta_i^2}}$$

- At any **fixed** time t, pigeonholing $\Longrightarrow$ there is a subset of arms $S_t \subseteq [K]$ s.t.:

$$\frac{T_i^{bias}(t-1)}{t^{bias}-1} = O\left(\frac{1}{K}\right) \text{ and } |S_t| = \Omega(K)$$

- But S_t is **random**, **time-dependent**, and may not be "stable"…
  - E.g., some bandit policy might identify the arms in S_t, pulling each of them until their fraction exceeds $1/K$.
  - However, pulling one arm decreases the fraction of **every** other arm.
- We construct the stopping times (roughly) as:

$$\tau_i = \min\left\{t \geq n_0 : \frac{T_i^{bias}(t)}{t^{bias}} \gg \frac{1}{K} \; or \; t = n\right\}$$

- Thus, we show that, for a constant fraction of arms $B' \subseteq S_{n_0}$, each arm $i \in B'$ satisfies one of the following:

  (Case 1) $\tau_i = n$, i.e., $\frac{T_i^{bias}(t)}{t^{bias}} = \tilde{O}\left(\frac{1}{K}\right)$ for every $t \in [n_0, n)$

  (Case 2) $\tau_i < n$ **and** $T_i(n_0, n) \gtrsim f\left(\frac{4}{K}\right)^{-2} \frac{\log(n)}{\Delta_i^2}$

- Pigeonholing $\Longrightarrow$ either Case 1 or Case 2 happens with constant probability for a constant fraction of arms

References

[**GMS19**]: Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. "Explore first, exploit next: The true shape of regret in bandit problems". In: Mathematics of Operations Research 44.2 (2019),

[**KCG16**]: Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. "On the complexity of best-arm identification in multi-armed bandit models". In: The Journal of Machine Learning Research 17.1 (2016)

[OA18]: Himani Oberai and Ila Mehrotra Anand. "Unconscious bias: thinking without thinking". In: Human Resource Management International Digest 26.6 (2018)

[**PS08**]: Devah Pager and Hana Shepherd. "The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets". In: Annu. Rev. Sociol 34 (2008),

[**RBR19**]: Jared A Russell, Sheri Brock, and Mary E Rudisill. "Recognizing the impact of bias in faculty recruitment, retention, and advancement processes". In: Kinesiology Review 8.4 (2019)

[**UC05**]: Eric Luis Uhlmann and Geoffrey L Cohen. "Constructed criteria: Redefining merit to justify discrimination". In: Psychological science 16.6 (2005)