Greedy Algorithm for Structured Bandits: A Sharp Characterization of Asymptotic Success/Failure

Aleksandrs Slivkins (Microsoft Research), Yunzong Xu, Shiliang Zuo (U. of Illinois Urbana-Champaign)

1. Introduction: Scope and Results

- **Explore** (to collect info) vs **exploit** (this info to make decisions)
  - Exploration adds complexity, is costly/unfair for the current user
- Alternative: **greedy** algorithm (exploitation only)
  - easier to deploy & adopt, aligns with user incentives.
  - natural dynamics in online platforms (due to myopic user behavior)
  - widely believed to perform poorly
- How well does Greedy perform, really?
  - Not well-understood! Even a basic question: success vs failure – linear vs sublinear regret
  - failure is the common case for **unstructured bandits** [1]
- Our scope: **success vs failure of Greedy in structured bandits**
  - known reward/feedback structure
  - a few (very) specific success & failure examples known, e.g., [2, 3, 4, 5]
- Our results: **sharp characterization of success vs failure**
  - applies to bandits with arbitrary *finite* structures
  - extends to contextual bandits and arbitrary auxiliary feedback
  - extends (with some caveats) to bandits with *infinite* reward structures
- Success vs failure **for each problem instance**, not in the worst case

2. Setting: Structured Contextual Bandits

- **Protocol** in each round $t = 1, 2, \ldots$
  - context $x_t \in \mathcal{X}$ arrives, algorithm selects arm $a_t \in \mathcal{A}$.
  - reward r_t: unit-variance Gaussian with mean $f^*(x_t, a_t)$.
- **Problem structure:** unknown true reward function $f^* \in \mathcal{F}$.
  - $\mathcal{F}$: known function class (e.g., linear, Lipschitz, polynomial).
- **Finiteness**: $\mathcal{X}$, $\mathcal{A}$, $\mathcal{F}$ are all finite (unless specified otherwise).
- Goal: minimize cumulative regret

$$R(T) = \sum_{t=1}^{T} \left[r^*(x_t) - r_t \right], \quad r^*(x) = \max_{a \in \mathcal{A}} f^*(x, a).$$

- **Greedy** Algorithm (Exploitation-Only): in each round t
  - Warm-up phase: collect T_0 samples for some context-arm pairs.
  - In each round $t > T_0$: predict a model via a **regression oracle**:

$$f_t = \arg\min_{f \in \mathcal{F}} \sum_{s \le t} (f(x_s, a_s) - r_s)^2.$$

  Choose the best arm for this model: $a_t = \arg\max_{a \in \mathcal{A}} f_t(x_t, a)$.
- **Structured Bandits**: special case with no contexts
- **Generalization**: Decision-Making with Structured Feedback (**DMSO**) [6]
  - arbitrary auxiliary feedback, incl. episodic reinforcement learning

3. Sharp Dichotomy for Structured Bandits

- **Key concepts:**
  - a problem instance $(f^*, \mathcal{F})$ is **self-identifiable** if revealing the expected reward $f^*(a)$ of any suboptimal arm a identifies this arm as suboptimal for any $f \in \mathcal{F}$.
  - $f_{\text{dec}} \in \mathcal{F}$ is a **decoy** if $f_{\text{dec}}(a_{\text{dec}}) = f^*(a_{\text{dec}}) < f^*(a^*)$ for some "decoy arm" a_{dec}.
  - Lemma: self-identifiability $\Leftrightarrow$ no decoys
- **Examples** (each reward function $f \in \mathcal{F}$ is a vector of expected rewards)
  1. Suppose $\mathcal{F} = \{(2,1), (2,3)\}$, and $f^* = (2,1)$; then f^* is not self-identifiable.
  2. Suppose $\mathcal{F} = \{(2,1,3), (2,3,1), (7,6,5)\}$, and $f^* = (2,1,3)$; then f^* is self-identifiable.

> **Theorem (Finite Structured Bandits):** Fix instance $(f^*, \mathcal{F})$.
> - "Success" if self-identifiable (for any warmup data): $\mathbb{E}[R(t)] \le T_0 + (K/\Gamma)^2\, O(\log t)$, where $\Gamma = \Gamma(f^*, \mathcal{F})$ is the "smallest gap" between f^* and any other function in $\mathcal{F}$.
> - "Failure" if a decoy exists (and warmup data consists of one sample for each arm):
>
> $$\Pr[\text{Greedy gets stuck on a decoy arm}] = p_{\text{dec}} > 0 \quad \Rightarrow \quad \mathbb{E}[R(t)] = \Omega(t).$$

- **Sharp Dichotomy:**

  Greedy succeeds for any warm-up data $\Leftrightarrow$ the problem instance is self-identifiable.
- **Significance:** "$\Rightarrow$" substantiates the common belief that Greedy performs poorly, while "$\Leftarrow$" makes the characterization precise and suggests when Greedy may suffice
- Similar results for **Structured Contextual Bandits** and **DMSO**:
  - under suitable generalizations of "self-identifiability" and "decoys"
  - DMSO requires a non-standard (MLE-based) version of Greedy & more involved analysis

4. Examples: Some Well-Studied Structures

Structure	Self-Identifiable?	Greedy Outcome
Linear bandits	✗	Fails
Linear contextual (diverse contexts)	✓	Succeeds
Linear contextual (degenerate contexts)	✗	Fails
Lipschitz bandits (contextual or not)	✗	Fails
Quadratic & Polynomial bandits	✗	Fails

All examples are **discretized** (in a consistent way), to satisfy the finiteness assumption.

Informal Takeaways:

- Greedy fails as a common case for most/all bandit structures of interest
- For *contextual* bandits it can go either way, depending on the structure.
- The success of Greedy appears to require context diversity and a parametric reward structure.

5. Structured Bandits with Infinite $\mathcal{F}$

- **Key Idea:** stronger, parameterized notions of self-identifiability & decoys
- **Definitions:** for some "margin" $\varepsilon > 0$,
  - An instance $(f^*, \mathcal{F})$ is **ε-self-identifiable** if every suboptimal arm a remains suboptimal for all $f \in \mathcal{F}$ satisfying $|f(a) - f^*(a)| \le \varepsilon$.
  - **ε-interior:** $\text{int}(\mathcal{F}, \varepsilon)$ is the subset of $\mathcal{F}$ whose nearby perturbations (within ℓ_2-distance ε) are still contained in $\mathcal{F}$.

> **Theorem (Infinite Function Class):** Fix instance $(f^*, \mathcal{F})$.
> - If $(f^*, \mathcal{F})$ is ε-self-identifiable, then
>
> $$\mathbb{E}[R(t)] \le T_0 + (K/\varepsilon)^2\, O(\log t).$$
>
> - If a decoy $f_{\text{dec}} \in \text{int}(\mathcal{F}, \varepsilon)$ exists, then Greedy gets stuck on a decoy arm with "constant" probability: at least $\exp(-O(K^2/\varepsilon^2))$.

- "margin" ε separating instances for which the positive result applies from instances for which the negative result applies

6. Conclusions

- **Main result:** sharp characterization via self-identifiability and "decoys", extends to contextual bandits and DMSO.
- **Elaborations:**
  - Greedy fails in most/all common bandit structures, unless context diversity and reward structure gives self-identifiability.
  - Self-identifiability makes the problem instance *intrinsically easy*: in some sense, any "reasonable" algorithm achieves sublinear regret.
- **Caveats:** the sharp characterization only applies to finite structures and comes with (possibly) very weak constants.
  - Partial fix: the margin-based characterization for infinite $\mathcal{F}$.
- **Future directions:**
  - Extend to infinite action sets and "approximate" greedy behaviors.
  - Better constants / regret rates for particular structures

References and acknowledgements

[1] K. Banihashem, M. Hajiaghayi, S. Shin, and A. Slivkins. Bandit social learning under myopic behavior. In *NeurIPS*, 2023.

[2] H. Bastani, M. Bayati, and K. Khosravi. Mostly exploration-free algorithms for contextual bandits. *Manag. Sci.*, 2021.

[3] S. Kannan, J. Morgenstern, A. Roth, B. Waggoner, and Z.S. Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *NeurIPS*, 2018.

[4] J.M. Harrison, N.B. Keskin, and A. Zeevi. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Manag. Sci.*, 2012.

[5] A.V. den Boer and B. Zwart. Simultaneously learning and optimizing using controlled variance pricing. *Manag. Sci.*, 2014.

[6] D.J. Foster, S.M. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making. *arXiv*, 2021.

LaTeX TikZposter