ETH zürich

ETH AI CENTER

# Bandits with Preference Feedback: A Stackelberg Game Perspective

Barna Pasztor*, Parnian Kassraie*, Andreas Krause



## Dueling Bandits

At every step t

Choose actions $\boldsymbol{x}_t, \boldsymbol{x}'_t$ → Receive binary feedback $y_t$

$\mathbb{P}(y_t = 1) = s\left(f(\boldsymbol{x}_t) - f(\boldsymbol{x}'_t)\right)$  $\boldsymbol{x}_t \succ \boldsymbol{x}'_t$

Repeat

- Kernelized reward function:
- Goal: Sublinear regret $f \in \mathcal{H}_k,\ \|f\|_k \leq B$

$$R(T) = \sum_{t=1}^{T} \frac{\mathbb{P}(\boldsymbol{x}^\star \succ \boldsymbol{x}_t) + \mathbb{P}(\boldsymbol{x}^\star \succ \boldsymbol{x}'_t) - 1}{2} \qquad \boldsymbol{x}^\star = \arg\max f(\boldsymbol{x})$$

**Challenges**

- Continuous action space
- Expensive to query, qualitative preference feedback
- Complexity of exploration & exploitation

**Contributions**

- Stackelberg Game formulation
- Practical confidence bounds for kernelized utilities
- SOTA performance with no-regret guarantee

## Reward Estimation

Preference-based inference is *equivalent* to learning with direct feedback, up to choice of kernel.

$$h_t \leftarrow \arg\min \frac{\lambda}{2}\|h\|^2_{\tilde{k}} + \sum_{\tau=1}^{t} -y_\tau \log\left[s(h(\boldsymbol{x}_\tau, \boldsymbol{x}'_\tau))\right] - (1-y_\tau)\log\left[1 - s(h(\boldsymbol{x}_\tau, \boldsymbol{x}'_\tau))\right]$$

- $h_t(\boldsymbol{x}, \boldsymbol{x}')$ estimates the utility gap

$$h_t(\mathbf{x}, \mathbf{x}') = \sum_{\tau=1}^{t} \alpha_\tau \tilde{k}((\mathbf{x}, \mathbf{x}'), (\mathbf{x}_t, \mathbf{x}'_t))$$

$$\tilde{k}((\boldsymbol{x}, \boldsymbol{x}'), (\boldsymbol{x}_t, \boldsymbol{x}'_t)) = k(\boldsymbol{x}, \boldsymbol{x}_t) + k(\boldsymbol{x}', \boldsymbol{x}'_t) - k(\boldsymbol{x}, \boldsymbol{x}'_t) - k(\boldsymbol{x}', \boldsymbol{x}_t)$$

- $\sigma_t(\boldsymbol{x}, \boldsymbol{x}')$ quantifies the estimation uncertainty

$$\sigma_t^2(\mathbf{x}, \mathbf{x}') = \tilde{k}(\mathbf{x}, \mathbf{x}') - \mathbf{k}_t^T(\mathbf{x}, \mathbf{x}')(K_t + \lambda\kappa\mathbf{I}_t)^{-1}\mathbf{k}_t(\mathbf{x}, \mathbf{x}')$$

**Theorem (Anytime Preference-based Conf Seq)**

Choosing $\beta_t \asymp \gamma_t + \log(1/\delta)$ satisfies

$\forall t \geq 1, \mathbf{x}, \mathbf{x}' \in \mathcal{X} : |\mathbb{P}(\mathbf{x} \succ \mathbf{x}') - s(h_t(\mathbf{x}, \mathbf{x}'))| \leq \beta_t \sigma_t(\mathbf{x}, \mathbf{x}')$

with probability greater than $1 - \delta$.

I want to have a healthy and balanced dinner. What should I cook?

Scenario A
MaxMin
Pair

Generation 1

Generation 2

Which one do you prefer?

Scenario B
Greedy
Pair

Generation 1

Generation 2

Which one do you prefer?

Scenario C
MaxInfo
Pair

Generation 1

Generation 2

Which one do you prefer?

## Stackelberg Game Perspective

View actions as players in a Stackelberg Game

- With objective $\mathbb{P}(\boldsymbol{x} \succ \boldsymbol{x}')$, both players choose $\boldsymbol{x}^\star$ via backward induction
- True preference is unknown
- Approximate it with a lower-bound

$$\mathrm{LCB}_t(\boldsymbol{x}, \boldsymbol{x}') = s(h_t(\boldsymbol{x}, \boldsymbol{x}')) - \beta_t \sigma_t(\boldsymbol{x}, \boldsymbol{x}')$$

**MaxMinLCB Acquisition Function**

$$\mathbf{x}_t = \arg\max_{\mathbf{x}} \mathrm{LCB}_t(\mathbf{x} \succ \omega(\mathbf{x}))$$

$$\text{s.t. } \omega(\mathbf{x}) = \arg\min_{\mathbf{x}'} \mathrm{LCB}_t(\mathbf{x} \succ \mathbf{x}')$$

$$\mathbf{x}'_t = \omega(\mathbf{x}_t)$$

Organically balances exploration & exploitation

- What's the role of the Leader?
- What's the role of the Follower?

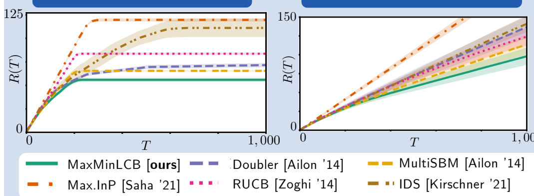
**Theorem (Regret – Informal)**

With an appropriate choice of $\beta_t$, MaxMinLCB satisfies

$$\mathbb{P}\left(\forall T \geq 1; R(T) \leq C_1(\gamma_T + \log 1/\delta)\sqrt{T}\right) \geq 1 - \delta$$

## Experiments

Ackley Function | Yelp Recommendation




Duelling Regret after 2000 steps

| $f$ | MaxMinLCB | Doubler | MultiSBM | MaxInP | RUCB | IDS |
|---|---|---|---|---|---|---|
| Branin | **104** ± 13 | 114 ± 9 | **89** ± 13 | 340 ± 2 | **101** ± 14 | 163 ± 22 |
| Matyas | 125 ± 5 | 136 ± 4 | **106** ± 7 | 136 ± 6 | **106** ± 6 | 128 ± 5 |
| Rosenbrock | **27** ± 4 | 44 ± 12 | **25** ± 5 | 109 ± 2 | 58 ± 7 | 58 ± 13 |
| Ackley | **56** ± 2 | 72 ± 2 | 65 ± 0.5 | 120 ± 1 | 84 ± 0.7 | 111 ± 9 |
| Eggholder | **113** ± 6 | 154 ± 4 | 134 ± 3 | 230 ± 34 | 213 ± 40 | 141 ± 12 |
| Hoelder | **141** ± 26 | **154** ± 3 | **136** ± 15 | 204 ± 20 | 200 ± 28 | **132** ± 15 |
| Michalewicz | **138** ± 14 | 183 ± 11 | **155** ± 10 | 260 ± 40 | 269 ± 46 | 188 ± 21 |
| Yelp | **175** ± 22 | 263 ± 28 | **199** ± 25 | 409 ± 15 | 214 ± 22 | 255 ± 22 |

**MaxMinLCB performs consistently among the top** for a variety of challenging utility functions with saddle points, local minima, and multiple global optima.