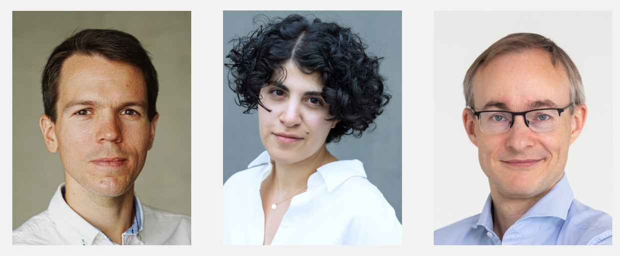
Bandits with Preference Feedback: A Stackelberg Game Perspective

Barna Pasztor*, Parnian Kassraie*, Andreas Krause

ETH AI CENTER ETH zürich



Challenges

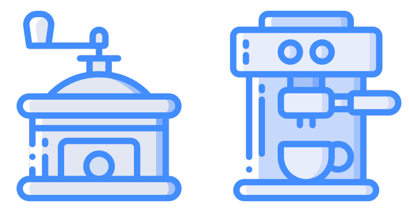
- Continuous action space
- Qualitative preference feedback
- Costly sampling
- Complexity of exploration & exploitation

Contributions

- Stackelberg Game formulation
- Practical confidence bounds for kernelized utilities
- No-regret guarantee
- Very promising performance

At every step t

Choose actions $\boldsymbol{x}_t, \boldsymbol{x}'_t$

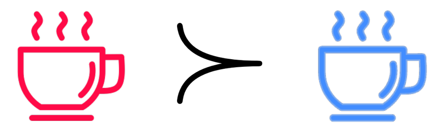


unknown utility

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

bernoulli sigmoid function

Receive binary feedback y_t



Repeat

Setting: continuous & kernelized — compact domain in $\mathbb{R}^d$, $f \in \mathcal{H}_k$, $\|f\|_k \le B$

Goal: sublinear regret

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

$\boldsymbol{x}^\star = \arg\max f(\boldsymbol{x})$

Stackelberg Perspective

View actions as players in a Stackelberg Game

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

MaxMinLCB Acquisition Function

$$\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}} \mathrm{LCB}_t(\boldsymbol{x} \succ \omega(\boldsymbol{x})) \qquad \text{Leader}$$

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

$$\boldsymbol{x}'_t = \omega(\boldsymbol{x}_t) \qquad \text{Follower}$$

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 β_t, MaxMinLCB satisfies

$$\mathbb{P}\left(\forall T \ge 1:\ R(T) \le C_1 (\gamma_T + \log 1/\delta)\sqrt{T}\right) \ge 1 - \delta$$

check the paper for formal statement.

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

$$\tilde{k}((\boldsymbol{x}_1, \boldsymbol{x}'_1), (\boldsymbol{x}_2, \boldsymbol{x}'_2)) = k(\boldsymbol{x}_1, \boldsymbol{x}_2) + k(\boldsymbol{x}'_1, \boldsymbol{x}'_2) - k(\boldsymbol{x}_1, \boldsymbol{x}'_2) - k(\boldsymbol{x}'_1, \boldsymbol{x}_2)$$

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

To construct a lower-bound for

- $h_t(\boldsymbol{x}, \boldsymbol{x}')$ estimates the utility gap $f(\boldsymbol{x}) - f(\boldsymbol{x}')$
- $\sigma_t(\boldsymbol{x}, \boldsymbol{x}')$ quantifies the estimation uncertainty

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

Theorem (Anytime Preference-based Conf Seq)

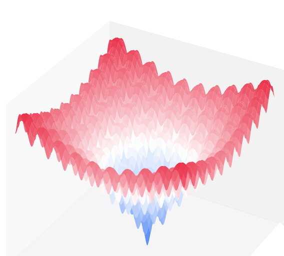
Choosing $\beta_t \asymp \gamma_t + \log(1/\delta)$ satisfies

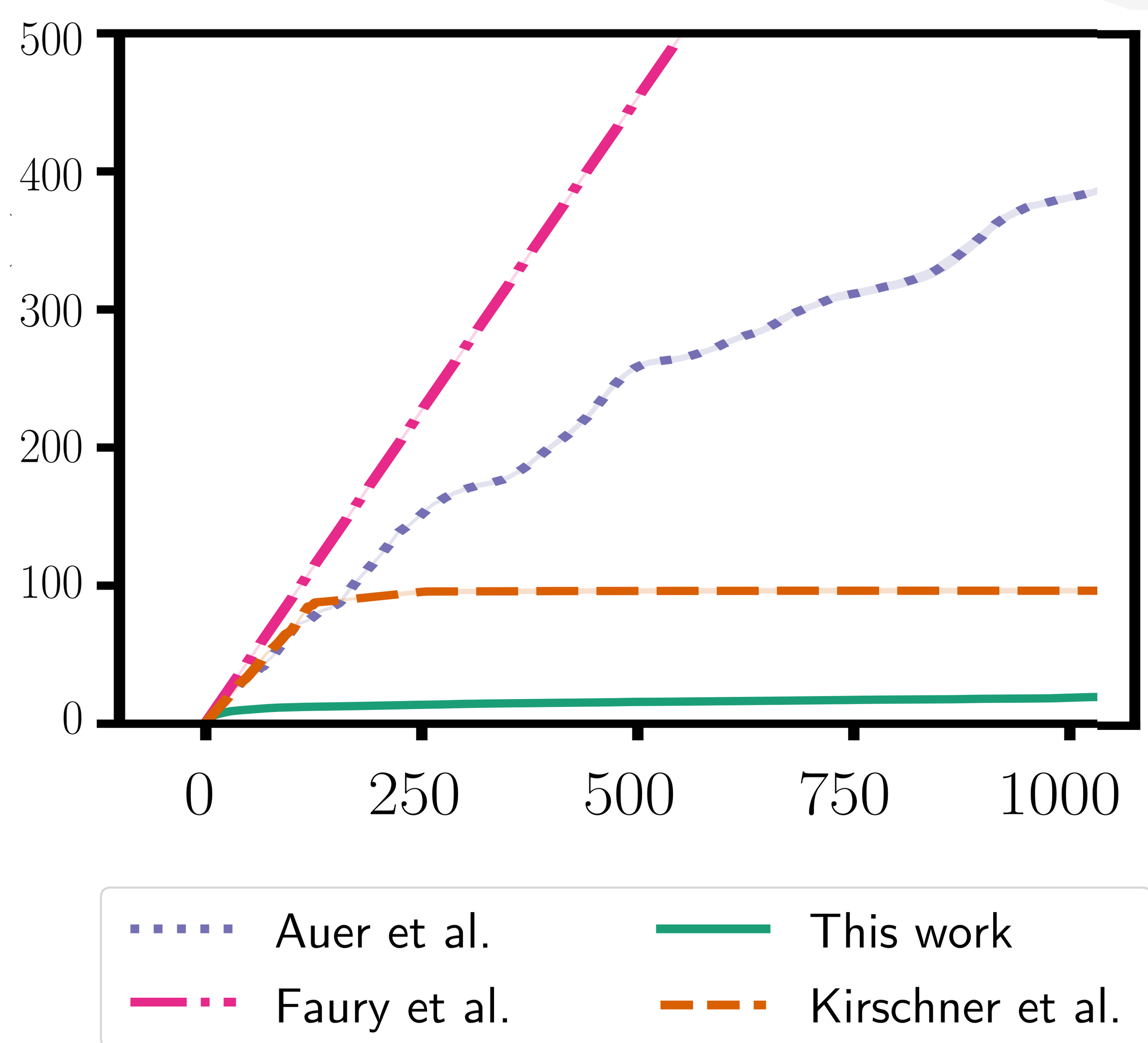
$$\forall t \ge 1, \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}:\ \left|\mathbb{P}(\boldsymbol{x} \succ \boldsymbol{x}') - s(h_t(\boldsymbol{x}, \boldsymbol{x}'))\right| \le \beta_t \sigma_t(\boldsymbol{x}, \boldsymbol{x}')$$

with probability greater than $1 - \delta$.

Result

regret of logistic bandit on Ackley reward using different conf. seqs.

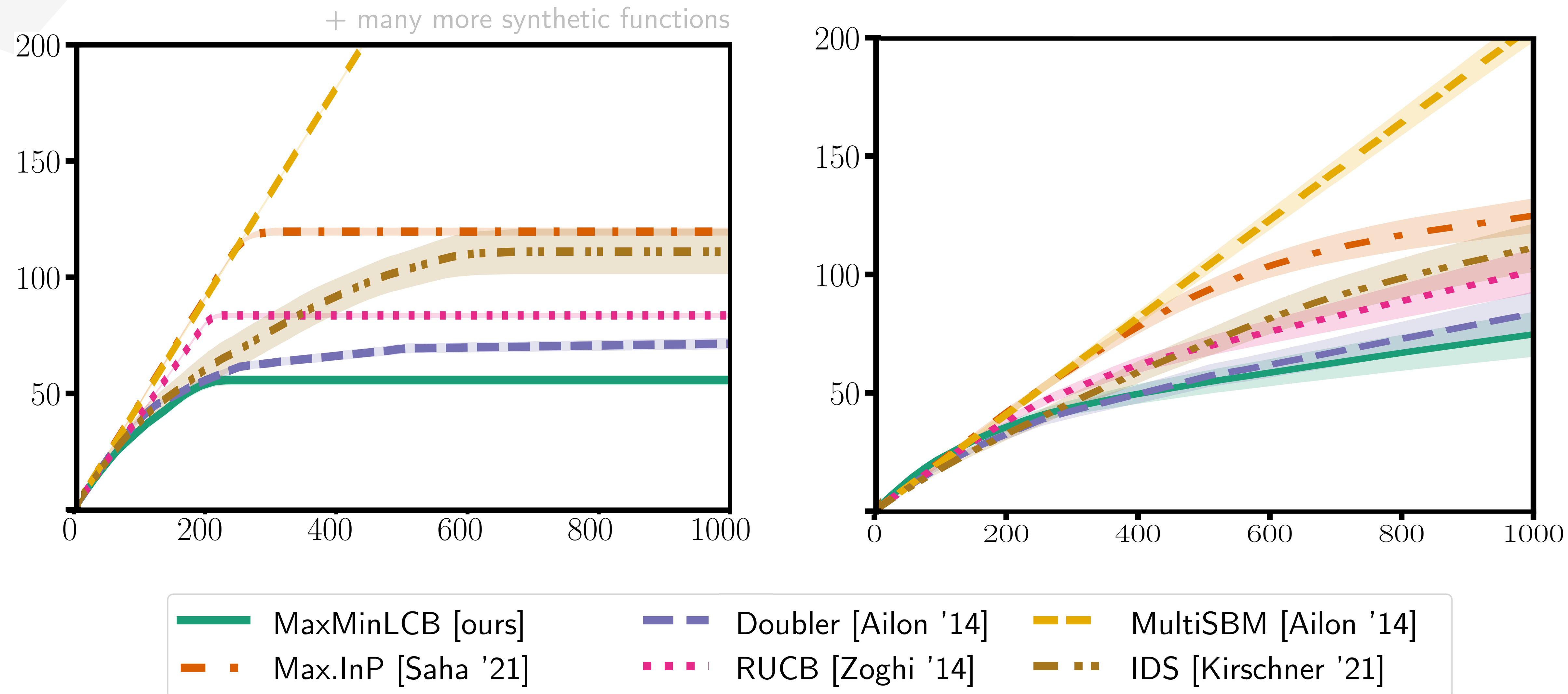




pref-based bandit benchmark on the 2D Ackley function

+ many more synthetic functions

pref-based bandit benchmark on text embeddings of Yelp reviews



Up Next

Applications in RLHF
adaptive fine-tuning of LLMs to niche domains, personalized & pluralist usage

Learning w Finite Recall
choosing an action from recent history to improve costs & feedback quality

Welfare Maximization
accepting feedback from multiple sources and aggregating the preference