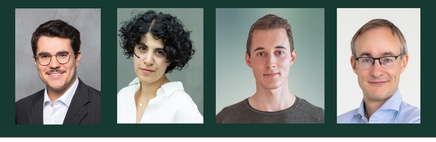
Lifelong Bandit Optimization: No Prior and No Regret

ETH zürich — Learning & Adaptive Systems

Felix Schur*, Parnian Kassraie*, Jonas Rothfuss, Andreas Krause (UAI '23)



TL;DR

Meta-learning for lifelong kernelized bandit optimization with oracle optimal performance guarantees

Overview

- We solve a sequence of kernelized bandit optimization tasks, where we assume the kernel to be unknown, but shared across all tasks.
- We develop LiBO, an algorithm that sequentially meta-learns an approximate kernel and solves the incoming tasks with the latest kernel estimate.
- Our method pairs with any kernelized bandit algorithm, ensuring oracle optimal performance, meaning that the LIBO's task-specific regret approaches the regret of an algorithm with oracle knowledge of the true kernel over time.
- We also propose F-LiBO, which solves the lifelong problem in a federated manner.

Problem Setting

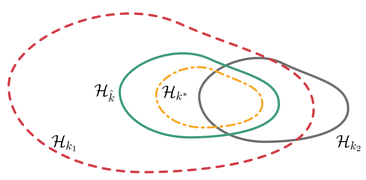
- Sequentially interacting with the environments

$$y_{s,i} = f_s(\boldsymbol{x}_{s,i}) + \varepsilon_{s,i} \qquad 1 \le i \le n \text{ and } 1 \le s \le m$$

$\varepsilon_{s,i}$: also σ^2 sub-Gussian, i.i.d.

$f_s : \mathcal{X} \to R,\ f_s \in \mathcal{H}_{k^*},\ \|f_s\|_{k^*} \le B$

- Approximating the kernel



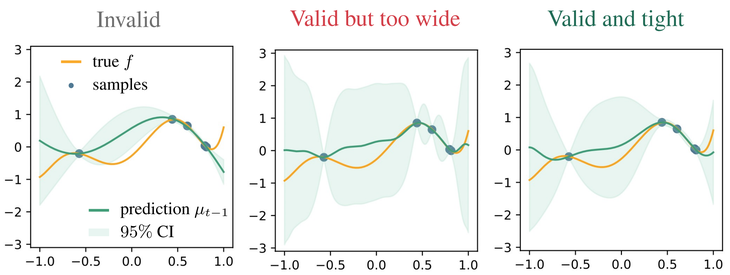
Invalid

Valid but too wide

Valid and tight

True sets (valid)

- We want valid and tight confidence bands for the reward functions such that the bandit solvers converge properly
- We achieve this with a good kernel approximation



- Applications

Bandits — Safe BO

Bayesian Optimization — Model-Based RL

Meta-Learning the Kernel

- Assume the kernel is a linear combination of known base kernels

$$k^*(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=1}^{p} \eta_j^* k_j(\boldsymbol{x}, \boldsymbol{x}') \qquad p < \infty$$

η_j^*: unknown, non-negative $\qquad$ k_j: known, finite-dimensional

- We use group Lasso to find the sparsity pattern with offline data

$$\hat{\beta} \implies \hat{J} \implies \hat{k} \implies \mathcal{H}_{\hat{k}}$$

$$\hat{\beta} \coloneqq \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{dm}} \left(\frac{1}{mn} \|y - \Phi\beta\|_2^2 + \lambda \sum_{j=1}^{p} \|\beta^{(j)}\|_2 \right)$$

$$\hat{k} = \frac{1}{|\hat{J}|} \sum_{j \in \hat{J}} k_j \qquad \hat{J} \coloneqq \left\{ j \in \{1, \dots, p\} \;\middle|\; \|\hat{\beta}^{(j)}\|_2 / \sqrt{m} > \omega \right\}$$

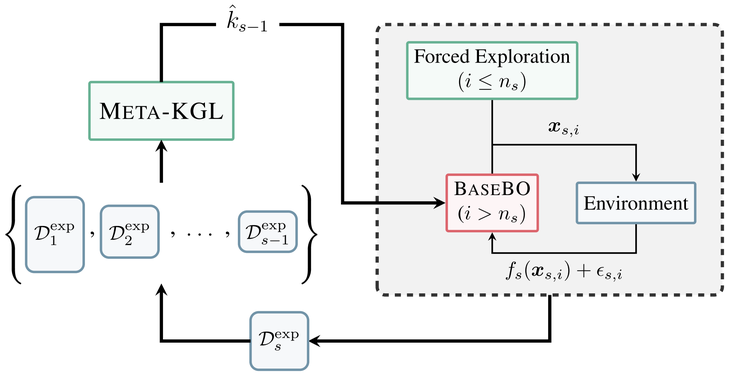
- Estimator is consistent

Theorem (Informal)

Under mild assumptions on the data model, $\mathcal{H}_{\hat{k}}$ is with high probability **a consistent estimator** *in the number of samples n and the number of tasks m. That is,*

$$\lim_{n \to \infty} \mathbb{P}\left[\mathcal{H}_{\hat{k}} = \mathcal{H}_{k^\star}\right] = 1, \text{ and } \lim_{m \to \infty} \mathbb{P}\left[\mathcal{H}_{\hat{k}} = \mathcal{H}_{k^\star}\right] = 1.$$

Lifelong Bandit Optimization



- Key features
  - Solves a sequence of kernelized bandit tasks in lifelong setting
  - Pairs with any kernelized bandit algorithm
  - Uses forced exploration to improve convergence

- Guarantee

Theorem (Informal)

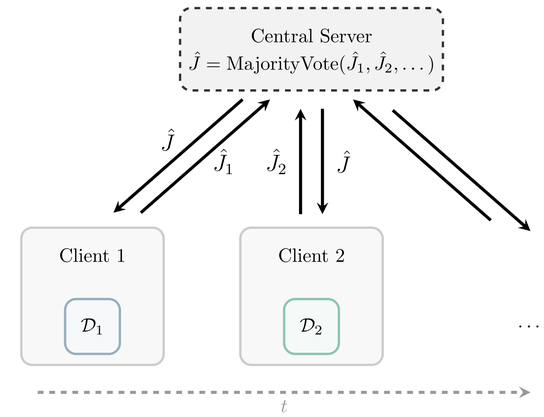
Under mild assumptions, LiBO *paired with* BaseBO *achieves oracle optimal performance with high probability, i.e.*

$$R(n, m) = \mathcal{O}\left(B\sqrt{n}m + R^\star(n, m)\right) = \mathcal{O}\left(R^\star(n, m)\right)$$

where $R^\star(n, m)$ is the lifelong regret of BaseBO*, if given knowledge of the kernel.*

Federated Setting

- We develop F-LiBO for the federated setting:

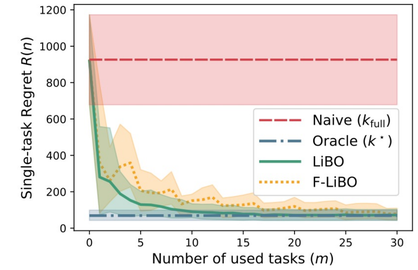


- Under similar assumptions we can prove the same convergence guarantees as for the non-federated setting

Experiments

- Offline kernel convergence experiment

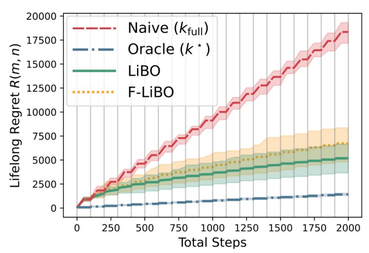
Sythetic 2D consine features



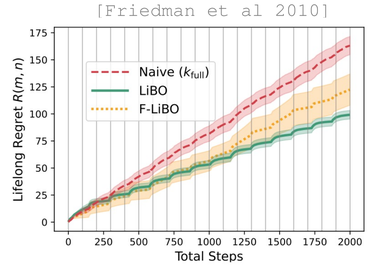
Regret of UCB using the estimated kernel quickly converges to the regret achieved with oracle knowledge of the kernel

- Lifelong bandit optimization

Sythetic 2D consine features



GLMNET data [Friedman et al 2010]



- Meta-learning with LiBO improves the performance on both synthetic and real world data
- In the beginning, when kernel estimate is invalid, the regret behaves the same as the naive bandit.
- Eventually, the regret becomes oracle optimal.