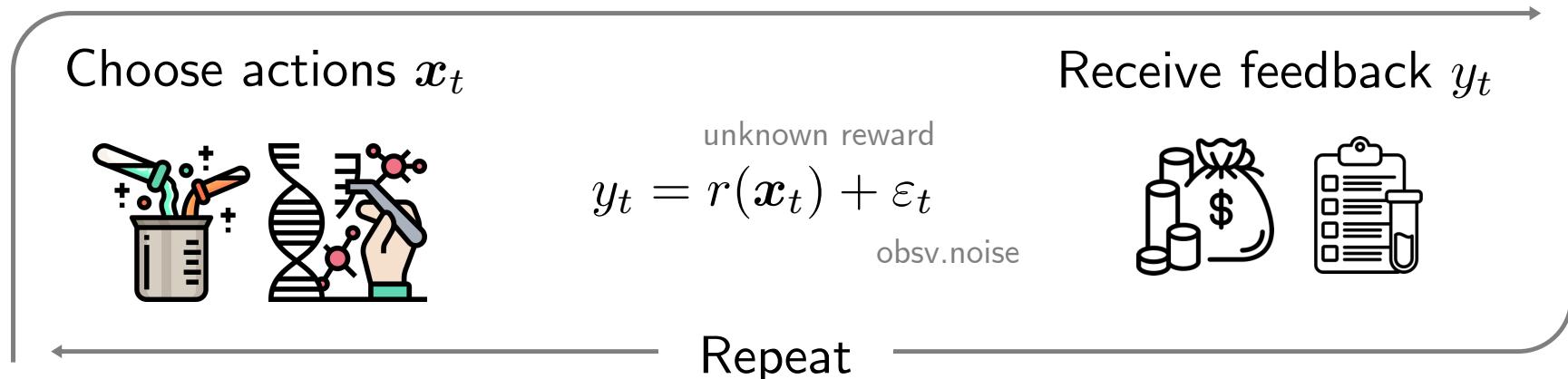


Meta Model Selection for Sequential Inference and Decision-making

Parnian Kassraie, ETH Zurich

Sequential Decision-Making & Bandits: Problem

At every step t



Goal: Choose actions that give a high reward

$$R(T) = \sum_{t=1}^T r(\mathbf{x}^*) - r(\mathbf{x}_t)$$

Motivation: maximize r given fewest samples

Sequential Decision-Making & Bandits: Solutions

Sequentially at every step:

- Estimate the reward function

based on: model for the reward e.g. r is a linear function
history $H_{t-1} = \{(x_1, y_1), \dots, (x_{t-1}, y_{t-1})\}$

- Use reward estimate to choose the next action

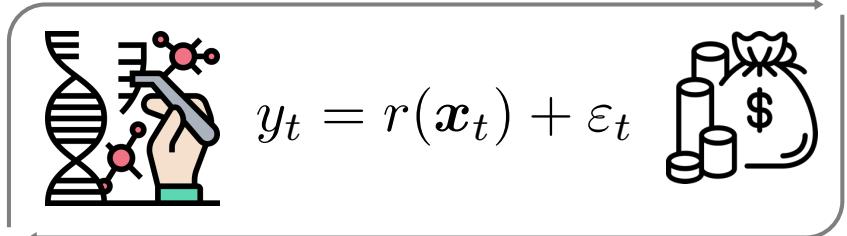
(better) estimate r
explore



maximize r
exploit

Many principles: optimism,
expected improvement,
entropy search

Heavily rely on the choice of model → Model selection is key!



Sequential Decision-Making & Bandits: Solutions

Sequentially at every step:

- Estimate the reward function

model for the reward e.g. r is a linear function
based on:

history $H_{t-1} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{t-1}, y_{t-1})\}$ samples are non-i.i.d

- Use reward estimate to choose the next action

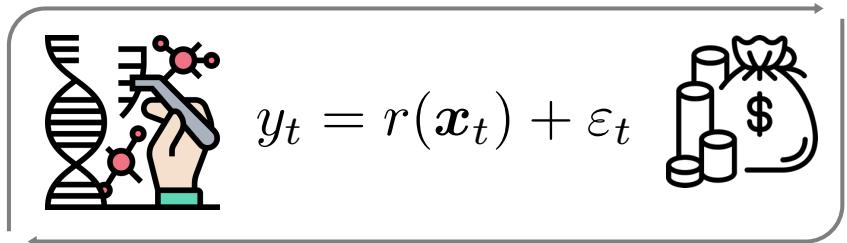
(better) estimate r
explore



maximize r
exploit

samples are
not so diverse

Model selection in this setting is not fun and games...



Problem Setting in this Talk

$$\mathbf{x}_t \in \mathcal{X} \subset \mathbb{R}^{d_0}$$

$$y_t = r(\mathbf{x}_t) + \varepsilon_t$$

i.i.d. zero-mean sub-gaussian noise

The reward is linearly parametrized by an unknown feature map

Model Class $\{\phi_j : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^d, j = 1, \dots, M\}$ $M \gg T$

$$\exists j^* \in [M] \text{ s.t. } r(\cdot) = \boldsymbol{\theta}_{j^*}^\top \boldsymbol{\phi}_{j^*}(\cdot)$$

+ typical bdd assump. $\|r\|_\infty \leq B$

Online Model Selection problem:

Find j^* while maximizing for the unknown r

$$R(T) = \sum_{t=1}^T r(\mathbf{x}^*) - r(\mathbf{x}_t) \quad \begin{aligned} & - \text{Sublinear in } T \\ & - \log M \end{aligned}$$

Warm-up Solution: Explore then Commit

For T_0 steps, take i.i.d. samples following a uniform, or “diverse” distribution

Use Group Lasso for implicit model selection

$$\hat{\boldsymbol{\theta}} = \arg \min \frac{1}{T_0} \|\mathbf{y} - \Phi \boldsymbol{\theta}\|_2^2 + \lambda \sum_{j=1}^M \|\boldsymbol{\theta}_j\|_2$$

For the remaining steps, always do

$$\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \dots, \phi_M(\mathbf{x}))$$

$$\mathbf{x}_t = \arg \max \hat{\boldsymbol{\theta}}^\top \phi(\mathbf{x})$$

Under good choice of T_0 and λ satisfies,

$$R(T) = \mathcal{O}(\sqrt[3]{T^2 \log M}) \quad \text{w.h.p.}$$

Warm-up Solution: Explore then Commit

For T_0 steps, take i.i.d. samples following a uniform, or “diverse” distribution

Incur high regret of $2BT_0$

Use the Lasso for implicit model selection

$$\hat{\boldsymbol{\theta}} = \arg \min \frac{1}{T_0} \|\mathbf{y} - \Phi \boldsymbol{\theta}\|_2^2 + \lambda \sum_{j=1}^M \|\boldsymbol{\theta}_j\|_2$$

Relies on Lasso variable selection

For the remaining steps, always do

$$\mathbf{x}_t = \arg \max \hat{\boldsymbol{\theta}}^\top \boldsymbol{\phi}(\mathbf{x})$$

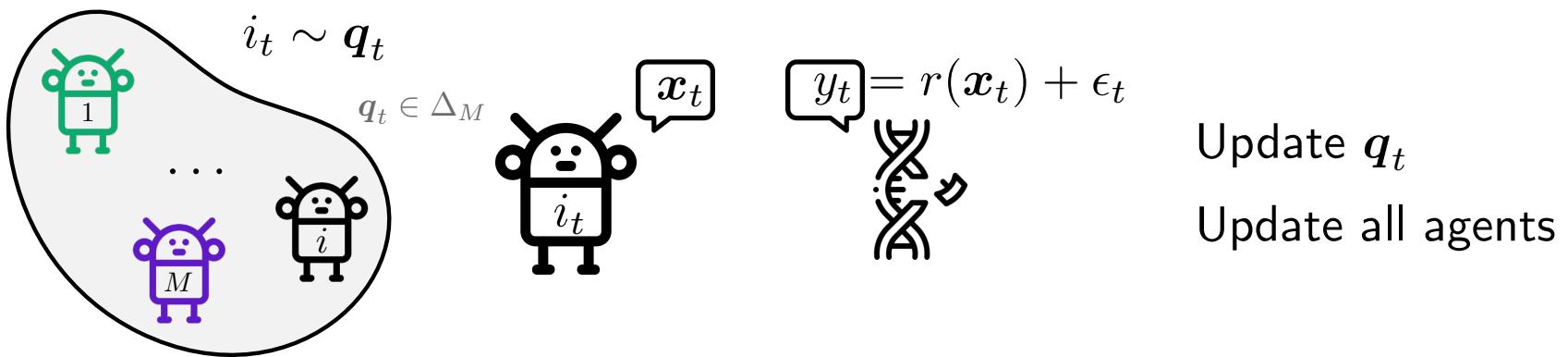
Is not any-time: only works if horizon T is known in advance

Under good choice of T_0 and λ satisfies,

$$R(T) = \mathcal{O}(\sqrt[3]{T^2 \log M}) \quad \text{w.h.p.}$$

Online Model Selection

- Instead of committing to a single model,
 - Randomly iterate over the models and at each step choose one
 - Instantiate M “agents” each using a different feature map
 - Agent j only uses ϕ_j to model the reward
 - Has action selection strategy $p_{t,j} \in \mathcal{M}(\mathcal{X})$ Which is updated at every step



Requires having observed the reward for the choice of each agent

- Reward not observed? **Hallucinate** it.

How to estimate rewards

💡 Turn group lasso into a sparse **online** regression oracle

$$\forall t \geq 1 \quad \hat{\boldsymbol{\theta}}_t = \arg \min \frac{1}{t} \|\mathbf{y}_t - \Phi_t \boldsymbol{\theta}\|_2^2 + \lambda_t \sum_{j=1}^M \|\boldsymbol{\theta}_j\|_2$$

Theorem (Anytime Conf. Seq.)

If for all $t \geq 1$

$$\lambda_t \geq \frac{c_1}{\sqrt{t}} \sqrt{\log(M/\delta) + \sqrt{d(\log(M/\delta) + (\log \log d)_+)}}$$

cost of going ‘time uniform’

then,

c_1 and c_2 made exact in the paper

$$\mathbb{P} \left(\forall t \geq 1 : \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t \right\|_2 \leq \frac{c_2 \lambda_t}{\kappa^2(\Phi_t, 2)} \right) \geq 1 - \delta$$

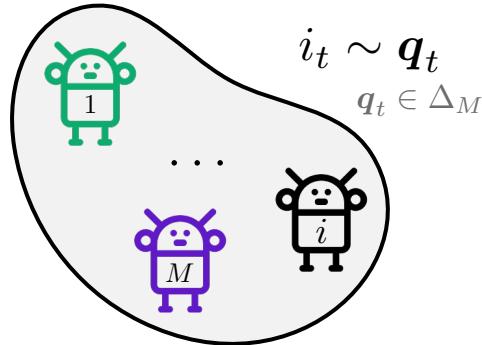
Restricted Eigenvalue property [check paper]

Hallucinate the reward of agent j as

$$\hat{r}_{t,j} = \mathbb{E}_{\mathbf{x} \sim p_{t,j}} \hat{\boldsymbol{\theta}}_t^\top \boldsymbol{\phi}(\mathbf{x})$$

$p_{t,j} \in \mathcal{M}(\mathcal{X})$ action selection strategy

How to iterate over agents



increase $q_{t,j}$ if $\hat{r}_{t,j}$ was high

💡 Exponential Weighting

Estimate of the reward obtained by agent j so far

$$q_{t,j} = \frac{\exp(\eta_t \sum_{s=1}^{t-1} \hat{r}_{s,j})}{\sum_{i=1}^M \exp(\eta_t \sum_{s=1}^{t-1} \hat{r}_{s,i})}$$

sensitivity of updates

$$\hat{r}_{t,j} = \mathbb{E}_{\mathbf{x} \sim p_{t,j}} \hat{\boldsymbol{\theta}}_t^\top \boldsymbol{\phi}(\mathbf{x})$$

Putting it all together

Find j^* while maximizing for the unknown r

Anytime Exponential weighting algorithm with Lasso reward estimates

Algorithm 1 ALEXP

Inputs: $\gamma_t, \eta_t, \lambda_t$ for $t \geq 1$

for $t \geq 1$ **do**

Draw $\mathbf{x}_t \sim (1 - \gamma_t) \sum_{j=1}^M q_{t,j} p_{t,j} + \gamma_t \text{Unif}(\mathcal{X})$

Observe $y_t = r(\mathbf{x}_t) + \epsilon_t$.

Append history $H_t = H_{t-1} \cup \{(\mathbf{x}_t, y_t)\}$.

Update agents $p_{t,j}$ for $j = 1, \dots, M$.

Calculate $\hat{\theta}_t \leftarrow \text{Lasso}(H_t, \lambda_t)$ and estimate

$$\hat{r}_{t,j} \leftarrow \mathbb{E}_{\mathbf{x} \sim p_{t+1,j}} [\hat{\theta}_t^\top \phi(\mathbf{x})]$$

Update selection distribution

$$q_{t+1,j} \leftarrow \frac{\exp(\eta_t \sum_{s=1}^t \hat{r}_{s,j})}{\sum_{i=1}^M \exp(\eta_t \sum_{s=1}^t \hat{r}_{s,i})}$$

Putting it all together: ALEXp

Find j^* while maximizing for the unknown r

Anytime Exponential weighting algorithm with Lasso reward estimates

Algorithm 1 ALEXP

Inputs: $\gamma_t, \eta_t, \lambda_t$ for $t \geq 1$

for $t \geq 1$ **do**

Draw $\mathbf{x}_t \sim (1 - \gamma_t) \sum_{j=1}^M q_{t,j} p_{t,j} + \gamma_t \text{Unif}(\mathcal{X})$

Observe $y_t = r(\mathbf{x}_t) + \epsilon_t$.

Append history $H_t = H_{t-1} \cup \{(\mathbf{x}_t, y_t)\}$.

Update agents $p_{t,j}$ for $j = 1, \dots, M$.

Calculate $\hat{\theta}_t \leftarrow \text{Lasso}(H_t, \lambda_t)$ and estimate

$$\hat{r}_{t,j} \leftarrow \mathbb{E}_{\mathbf{x} \sim p_{t+1,j}} [\hat{\theta}_t^\top \phi(\mathbf{x})]$$

Update selection distribution

$$q_{t+1,j} \leftarrow \frac{\exp(\eta_t \sum_{s=1}^t \hat{r}_{s,j})}{\sum_{i=1}^M \exp(\eta_t \sum_{s=1}^t \hat{r}_{s,i})}$$

Theorem (Online Model Selection)

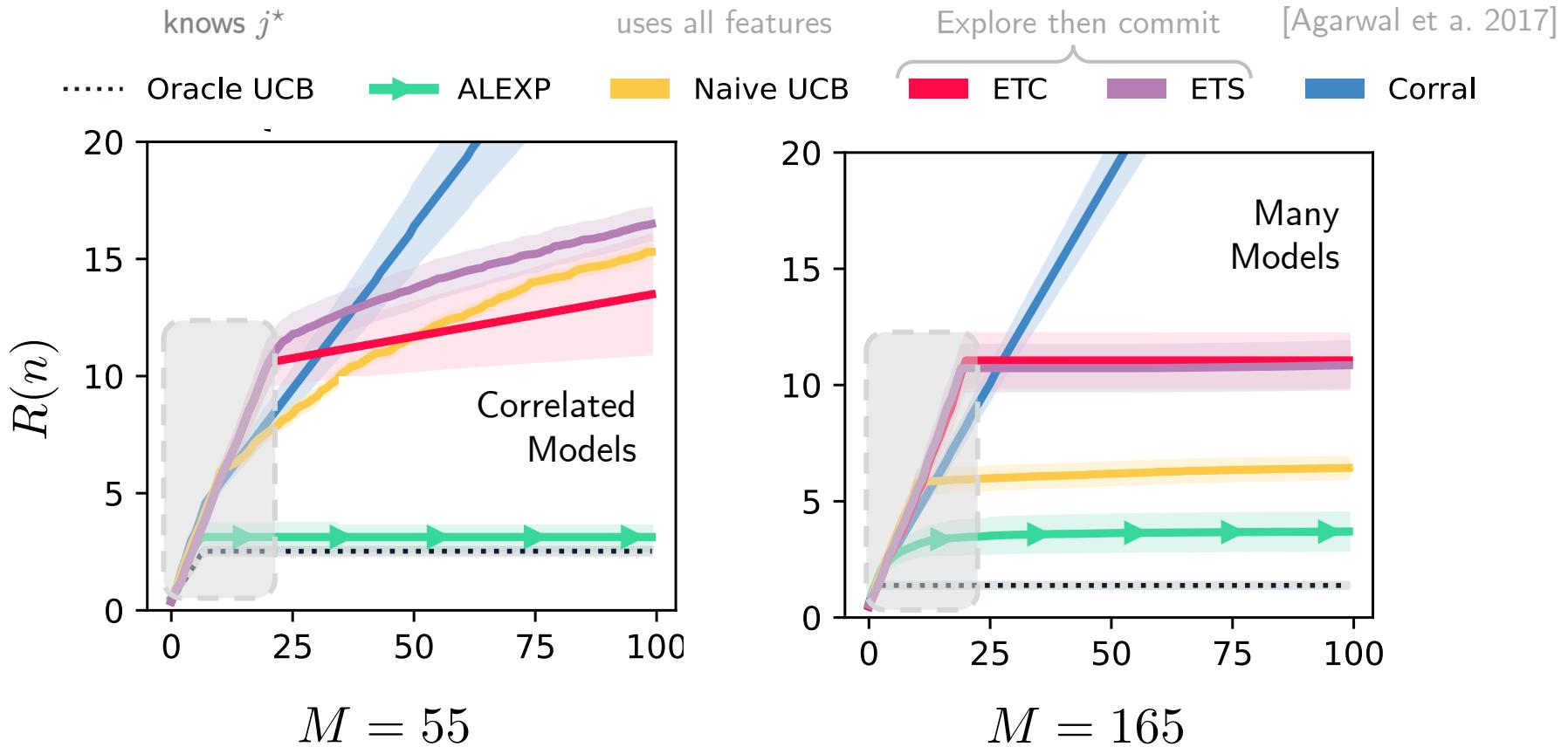
For appropriate choices of parameters, ALEXP satisfies

$$R(T) = \mathcal{O} \left(\sqrt{T \log^3 M} + T^{3/4} \sqrt{\log M} \right)$$

w.h.p. simultaneously for all $T \geq 1$.

[Addresses the open problem of Agarwal et al. 2017]

Model Selection for Optimistic algorithms



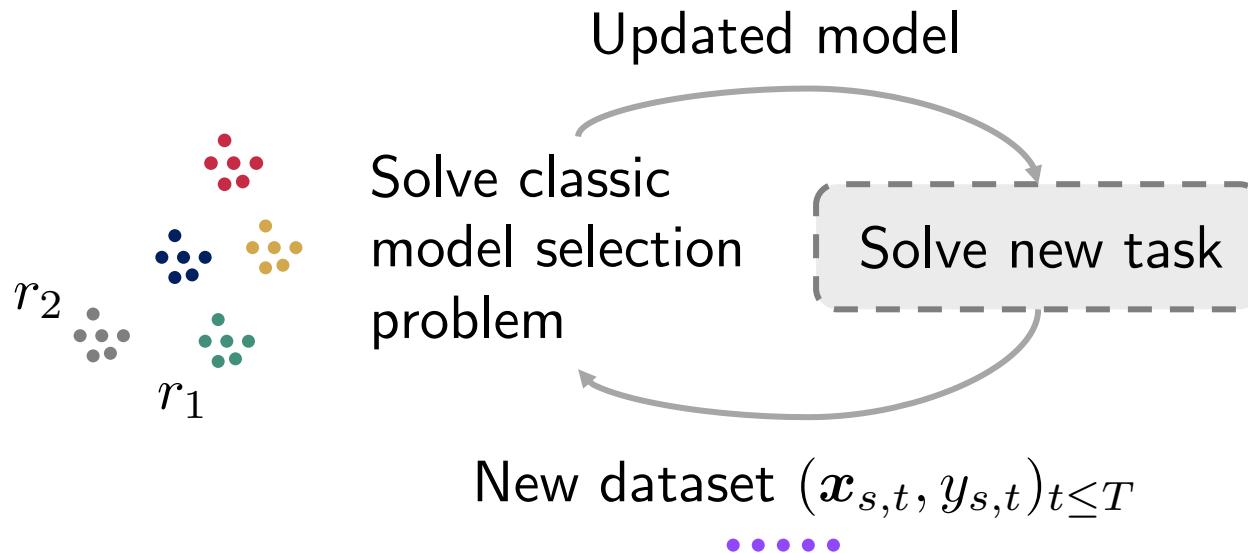
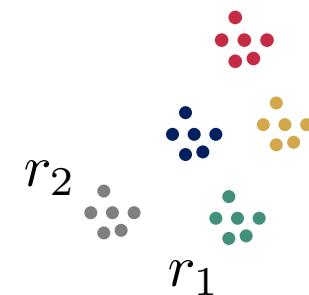
Meta Model Selection

$$\forall s \geq 1 : r_s \in \mathcal{H}$$

Suppose the bandit task is of repetitive nature,

Optimizing for different molecular properties

Recommending products to different customers



Thank you.



Kassraie, Parnian, Aldo Pacchiano, Nicolas Emmenegger, and Andreas Krause.
"Anytime Model Selection in Linear Bandits." *Under Review*, 2023.

Kassraie, Parnian, Jonas Rothfuss, and Andreas Krause. "Meta-learning hypothesis spaces for sequential decision-making." In *International Conference on Machine Learning*. PMLR, 2022.

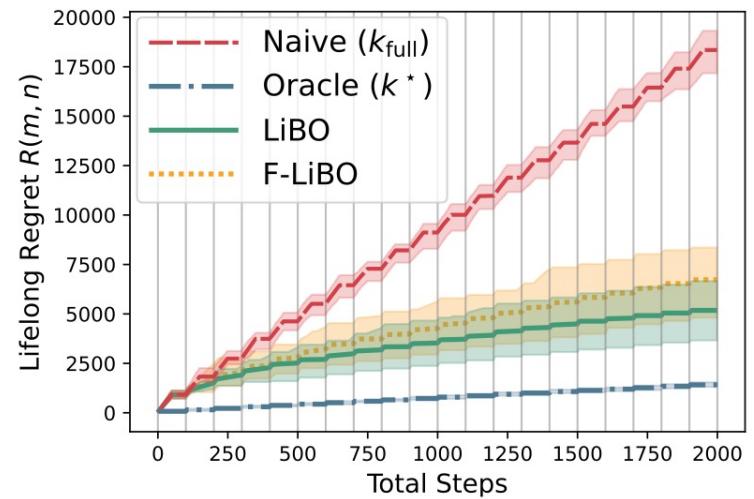
Schur, Felix, Parnian Kassraie, Jonas Rothfuss, and Andreas Krause. "Lifelong bandit optimization: no prior and no regret." In *Uncertainty in Artificial Intelligence*. PMLR, 2023.

Theorem (Lifelong Model Selection)

Under mild assumptions on the meta-data, and for an appropriate choice of λ , w.h.p.

- \hat{J} is a consistent estimator of J ,
- The optimization algorithm which uses \hat{J} achieves oracle performance $R^*(T, m)$,

as m grows.



the regret converges at a $\mathcal{O}(\log M/\sqrt{m})$ rate

$$R(T, m) = \sum_{s=1}^m \sum_{t=1}^T r_s(\mathbf{x}_s^*) - r_s(\mathbf{x}_{s,t})$$

Theorem (Regret - Informal)

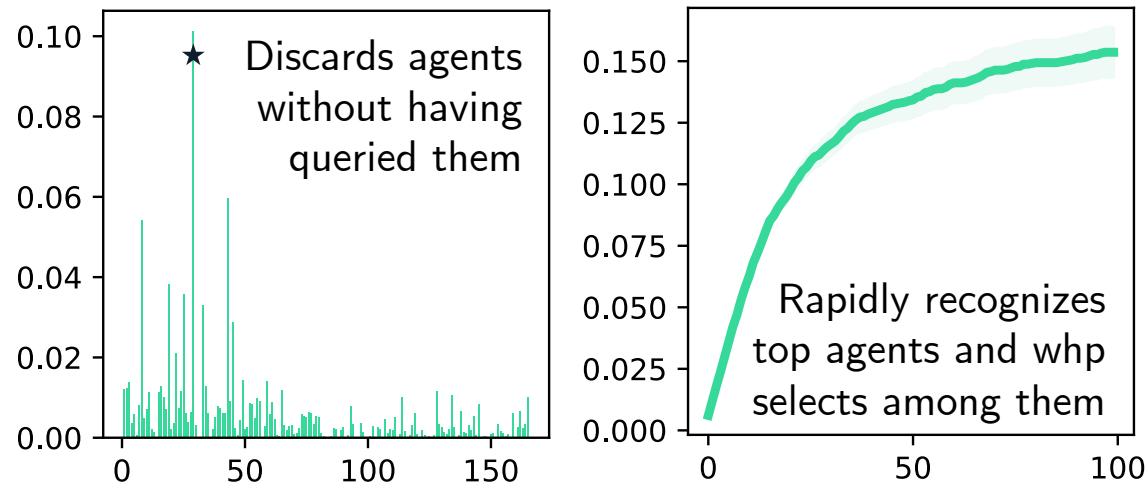
For appropriate choices of $(\gamma_t, \lambda_t, \eta_t)$,

$$R(n) = \mathcal{O} \left(C(M, \delta, d) \left(\sqrt{n} \log M + n^{3/4} \right) \right)$$

*with probability greater than $1 - \delta$,
simultaneously for all $n \geq 1$.*

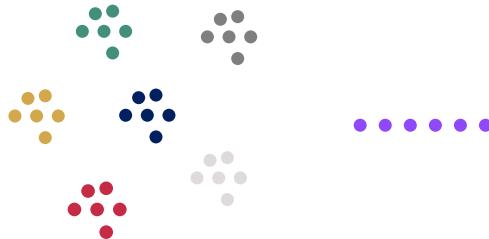
$$C(M, \delta, d) = \mathcal{O} \left(\sqrt{d \log M / \delta} + \sqrt{d \log M / \delta} \right)$$

How does ALExp work?



We consider 3 scenarios of increasing difficulty

1. Offline Data from similar tasks is available [KRK 2022]



2. Online data from similar tasks can be available [SKRK 2023]

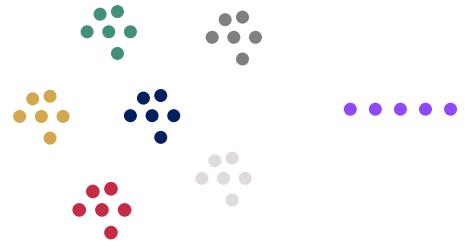


3. No data from similar tasks is available [KPEK 2023]



Meta-Model Selection: Offline

When offline data from similar tasks is available,



$$y_{s,i} = r_s(\mathbf{x}_{s,i}) + \varepsilon_{s,i} \quad i = 1, \dots, n \text{ and } s = 1, \dots, m$$

$$r_s(\cdot) = \sum_{j=1}^M \langle \boldsymbol{\theta}_s^{(j)}, \boldsymbol{\phi}_j(\cdot) \rangle \quad J \text{ is shared}$$

Classical feature selection with Lasso

$$\hat{\boldsymbol{\theta}}^{(1)}, \dots, \hat{\boldsymbol{\theta}}^{(M)} = \arg \min \frac{1}{mn} \|\mathbf{y} - \sum_{j=1}^M \Phi_j \boldsymbol{\theta}^{(j)}\|_2^2 + \lambda \sum_{j=1}^M \|\boldsymbol{\theta}^{(j)}\|_2$$

$$\hat{J} = \{j \in [M] \text{ s.t. } \hat{\boldsymbol{\theta}}^{(j)} > \omega\}$$

Solving the online optimization problem using the learned model