Choosing Answers in ε-Best-Answer Identification for Linear Bandits

Marc Jourdan and Rémy Degenne

November 19, 2021

Université de Lille
CNRS
CRIStAL Centre de Recherche en Informatique, Signal et Automatique de Lille
Inria

Section 1

Motivation

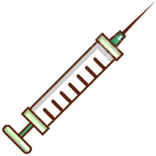
Clinical trials (phase II/III)

Treatments = Arms = Answers

$\mathcal{B}(\mu^1)$

$\mathcal{B}(\mu^2)$

$\mathcal{B}(\mu^3)$



$\mathcal{B}(\mu^4)$

For the t-th patient,

- administer a treatment a_t and
- observe a response $X_t^{a_t} \in \{0, 1\}$ such that $\mathbb{P}_\mu[X_t^{a_t} = 1] = \mu^{a_t}$.

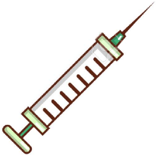
Goal: identify the best treatment (BAI), $a^*(\mu) = \arg\max_{a \in [4]} \mu^a$.

“Easy” instance

$\mathcal{B}(0.2)$

$\mathcal{B}(0.8)$

$\mathcal{B}(0.6)$



$\mathcal{B}(0.5)$

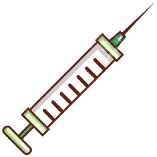
- Few samples to identify the red treatment as the best one.

“Hard” instance

$\mathcal{B}(0.2)$

$\mathcal{B}(0.8)$

$\mathcal{B}(0.799)$



$\mathcal{B}(0.5)$

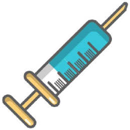
- Numerous samples to distinguish between the red and blue treatments.
- **Question:** Do we really need to identify the red treatment or would we also be satified with the blue one ?

Identifying a relatively good treatment

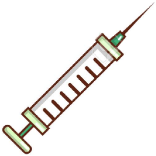
Goal: identify one treatment which is ε-close to the best treatment (ε-BAI).

$\mathcal{B}(0.2)$

$\mathcal{B}(0.8)$



$\mathcal{B}(0.799)$



$\mathcal{B}(0.5)$

- Few samples to identify the red or the blue treatments as relatively good treatments.

Question: At the end of the clinical trial, should we recommend the red treatment (BAI) or the blue one ?

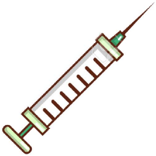
Identifying a relatively good treatment

Goal: identify one treatment which is ε-close to the best treatment (ε-BAI).

$\mathcal{B}(0.2)$

$\mathcal{B}(0.8)$

$\mathcal{B}(0.799)$



$\mathcal{B}(0.5)$

- Few samples to identify the red or the blue treatments as relatively good treatments.

Question: At the end of the clinical trial, should we recommend the red treatment (BAI) or the blue one ?

Choosing a restaurant for a special occasion

Unknown partner's taste $\mu = (\text{quantity, visual}) = (0.6, 0.5)$.

"Daily"/"Cheap" meals = Arms

$a_1 = (0.9, 0.1)$

$a_2 = (0.2, 0.7)$

$a_3 = (0.5, 0.2)$

$a_4 = (1, 0.1)$

For the t-th dinner at home,

- choose a "daily" meal a_t and
- observe a response $X_t^{a_t} \sim \mathcal{N}(\mu^{a_t}, 1)$ where $\mu^{a_t} = \langle \mu, a_t \rangle$.

Choosing a restaurant for a special occasion

"Fancy"/"Expensive" meals = Answers

$z_1 = (0.8, 0.4)$

$z_2 = (0.3, 0.9)$

$z_3 = (0.4, 0.3)$

Goal: identify one "fancy" meal which is ε-close to the favorite one of your partner whose taste is $\mu = (0.6, 0.5)$.

Question: For the special occasion, should we go eat bibimbap (BAI) or snails ?

Choosing a restaurant for a special occasion

“Fancy”/“Expensive” meals = Answers

$z_1 = (0.8, 0.4)$

$z_2 = (0.3, 0.9)$

$z_3 = (0.4, 0.3)$

Goal: identify one “fancy” meal which is ε-close to the favorite one of your partner whose taste is $\mu = (0.6, 0.5)$.

Question: For the special occasion, should we go eat bibimbap (BAI) or snails ?

Section 2

Problem Statement

Transductive linear Gaussian bandits

Transductive bandits:

- arms, $\mathcal{K} = \{a_k\}_{k \in [K]} \subseteq \mathbb{R}^d$ where Span$(\mathcal{K}) = \mathbb{R}^d$,
- answers, $\mathcal{Z} = \{z_i\}_{i \in [Z]} \subseteq \mathbb{R}^d$.

Linear Gaussian bandits:

- unknown mean parameter, $\mu \in \mathcal{M} \subseteq \mathbb{R}^d$,
- Gaussian distributions, $\nu^a = \mathcal{N}\left(\langle \mu, a \rangle, 1\right)$ for all $a \in \mathcal{K}$.

At time t, pull $a_t \in \mathcal{K}$ and observe $X_t^{a_t} \sim \nu^{a_t}$.

ε-best-answer identification (ε-BAI)

Goal: Identify one ε-optimal answer, $z \in \mathcal{Z}_\varepsilon(\mu)$ with $\varepsilon \geq 0$.

Greedy answer, $z^*(\mu) = \arg\max_{z \in \mathcal{Z}} \langle \mu, z \rangle$.
$\rightarrow$ In BAI ($\varepsilon = 0$), $z^*(\mu)$ is the unique correct answer.

ε-optimality:

- additive, $\mathcal{Z}_\varepsilon^{\mathrm{add}}(\mu) = \{z \in \mathcal{Z} : \langle \mu, z \rangle \geq \langle \mu, z^*(\mu) \rangle - \varepsilon\}$,
- multiplicative, $\mathcal{Z}_\varepsilon^{\mathrm{mul}}(\mu) = \{z \in \mathcal{Z} : \langle \mu, z \rangle \geq (1 - \varepsilon)\langle \mu, z^*(\mu) \rangle\}$.

Questions:

- How to choose among the ε-optimal answers ?
- Can we do better than the greedy answer ?

ε-best-answer identification (ε-BAI)

Goal: Identify one ε-optimal answer, $z \in \mathcal{Z}_\varepsilon(\mu)$ with $\varepsilon \geq 0$.

Greedy answer, $z^*(\mu) = \arg\max_{z \in \mathcal{Z}} \langle \mu, z \rangle$.
$\rightarrow$ In BAI ($\varepsilon = 0$), $z^*(\mu)$ is the unique correct answer.

ε-optimality:

- additive, $\mathcal{Z}_\varepsilon^{\mathrm{add}}(\mu) = \{z \in \mathcal{Z} : \langle \mu, z \rangle \geq \langle \mu, z^*(\mu) \rangle - \varepsilon\}$,
- multiplicative, $\mathcal{Z}_\varepsilon^{\mathrm{mul}}(\mu) = \{z \in \mathcal{Z} : \langle \mu, z \rangle \geq (1 - \varepsilon)\langle \mu, z^*(\mu) \rangle\}$.

Questions:

- How to choose among the ε-optimal answers ?
- Can we do better than the greedy answer ?

(ε, δ)-PAC identification strategy

Fixed-confidence setting, $\delta \in (0, 1)$

Three rules:

- *sampling* rule, $a_t \in \mathcal{K}$,
- *recommendation* rule, $z_t \in \mathcal{Z}$,
- *stopping* rule, τ_δ.

Requirement: (ε, δ)-PAC, $\mathbb{P}_\mu [z_{\tau_\delta} \notin \mathcal{Z}_\varepsilon(\mu)] \leq \delta$ and $\mathbb{P}_\mu [\tau_\delta < +\infty] = 1$.

Objective: Minimize $\mathbb{E}_\mu[\tau_\delta]$.

(ε, δ)-PAC identification strategy

Fixed-confidence setting, $\delta \in (0, 1)$

Three rules:

- *sampling* rule, $a_t \in \mathcal{K}$,
- *recommendation* rule, $z_t \in \mathcal{Z}$,
- *stopping* rule, τ_δ.

Requirement: (ε, δ)-PAC, $\mathbb{P}_\mu \left[z_{\tau_\delta} \notin \mathcal{Z}_\varepsilon(\mu)\right] \leq \delta$ and $\mathbb{P}_\mu \left[\tau_\delta < +\infty\right] = 1$.

Objective: Minimize $\mathbb{E}_\mu[\tau_\delta]$.

Contributions

1. Analyze (ε, δ)-PAC BAI for transductive linear bandits.
2. Don't choose greedily: aim at identifying the *furthest* answer !
3. LεBAI (Linear ε-BAI), asymptotically optimal and empirically competitive.

Related work

ε-BAI:

- Degenne and Koolen (2019), multiple-correct answer setting with fixed-confidence, Sticky Track-and-Stop (TaS),
- Garivier and Kaufmann (2021), (ε, δ)-PAC BAI in MAB for additive ε-optimality, ε-TaS,
- Kocák and Garivier (2021), (ε, δ)-PAC BAI in additive spectral bandits, SpectralTaS.

Fixed-confidence BAI in linear bandits (to name a few):

- Soare et al. (2014), $\mathcal{XY}$-Adaptive,
- Xu et al. (2018), LinGapE,
- Fiez et al. (2019), RAGE,
- Jedra and Proutière (2020), Lazy TaS,
- Degenne et al. (2020), LinGame.

Related work

ε-BAI:

- Degenne and Koolen (2019), multiple-correct answer setting with fixed-confidence, Sticky Track-and-Stop (TaS),
- Garivier and Kaufmann (2021), (ε, δ)-PAC BAI in MAB for additive ε-optimality, ε-TaS,
- Kocák and Garivier (2021), (ε, δ)-PAC BAI in additive spectral bandits, SpectralTaS.

Fixed-confidence BAI in linear bandits (to name a few):

- Soare et al. (2014), $\mathcal{XY}$-Adaptive,
- Xu et al. (2018), LinGapE,
- Fiez et al. (2019), RAGE,
- Jedra and Proutière (2020), Lazy TaS,
- Degenne et al. (2020), LinGame.

Section 3

Comparing ε-Optimal Answers

Notations and alternative to z

Notations:

- design matrix $V_w = \sum_{a \in \mathcal{K}} w^a a a^\intercal \in \mathbb{R}^{d \times d}$ for any $w \in (\mathbb{R}^+)^K$,
- norm $\|x\|_V = \sqrt{x^\intercal V x}$ for $x \in \mathbb{R}^d$,
- simplex of dimension $K-1$ is denoted by $\triangle_K$.

Alternative to $z \in \mathcal{Z}$: closure of the set of parameters for which z is not an ε-optimal answer, $\neg_\varepsilon z = \overline{\{\lambda \in \mathcal{M} : z \notin \mathcal{Z}_\varepsilon(\lambda)\}}$.

Identifying z as an ε-optimal answer is equivalent to rejecting the hypothesis that μ belongs to the alternative to z.

$$\forall z \in \mathcal{Z}, \quad \mathcal{H}_{0,z} : (\mu \in \neg_\varepsilon z) \quad \text{against} \quad \mathcal{H}_{1,z} : (z \in \mathcal{Z}_\varepsilon(\mu))$$

Notations and alternative to z

Notations:

- design matrix $V_w = \sum_{a \in \mathcal{K}} w^a a a^\intercal \in \mathbb{R}^{d \times d}$ for any $w \in (\mathbb{R}^+)^K$,
- norm $\|x\|_V = \sqrt{x^\intercal V x}$ for $x \in \mathbb{R}^d$,
- simplex of dimension $K - 1$ is denoted by $\triangle_K$.

Alternative to $z \in \mathcal{Z}$: closure of the set of parameters for which z is not an ε-optimal answer, $\neg_\varepsilon z = \overline{\{\lambda \in \mathcal{M} : z \notin \mathcal{Z}_\varepsilon(\lambda)\}}$.

Identifying z as an ε-optimal answer is equivalent to rejecting the hypothesis that μ belongs to the alternative to z.

$$\forall z \in \mathcal{Z}, \quad \mathcal{H}_{0,z} : (\mu \in \neg_\varepsilon z) \quad \text{against} \quad \mathcal{H}_{1,z} : (z \in \mathcal{Z}_\varepsilon(\mu))$$

Asymptotic lower bound

Theorem (Degenne and Koolen (2019))

For all (ε, δ)-PAC strategy, for all $\mu \in \mathcal{M}$,

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_\mu[\tau_\delta]}{\ln(1/\delta)} \geq T_\varepsilon(\mu)$$

where the inverse of the characteristic time is

$$T_\varepsilon(\mu)^{-1} = \max_{z \in \mathcal{Z}_\varepsilon(\mu)} \max_{w \in \triangle_K} \inf_{\lambda \in \neg_\varepsilon z} \frac{1}{2} \|\mu - \lambda\|^2_{V_w} \tag{1}$$

Asymptotic optimality: for all $\mu \in \mathcal{M}$,

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_\mu[\tau_\delta]}{\ln(1/\delta)} \leq T_\varepsilon(\mu)$$

Asymptotic lower bound

Theorem (Degenne and Koolen (2019))

For all (ε, δ)*-PAC strategy, for all* $\mu \in \mathcal{M}$*,*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_\mu[\tau_\delta]}{\ln(1/\delta)} \geq T_\varepsilon(\mu)$$

where the inverse of the characteristic time is

$$T_\varepsilon(\mu)^{-1} = \max_{z \in \mathcal{Z}_\varepsilon(\mu)} \max_{w \in \triangle_K} \inf_{\lambda \in \neg_\varepsilon z} \frac{1}{2} \|\mu - \lambda\|^2_{V_w} \tag{1}$$

Asymptotic optimality: for all $\mu \in \mathcal{M}$,

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_\mu[\tau_\delta]}{\ln(1/\delta)} \leq T_\varepsilon(\mu)$$

Furthest answer

$$T_\varepsilon(\mu)^{-1} = \max_{z \in \mathcal{Z}_\varepsilon(\mu)} \max_{w \in \triangle_K} \inf_{\lambda \in \neg_\varepsilon z} \frac{1}{2} \|\mu - \lambda\|^2_{V_w}$$

The ε-optimal answer for which its alternative is the easiest to differentiate from thanks to an optimal allocation over arms $w_F(\mu) \in \triangle_K$.

$$(z_F(\mu), w_F(\mu)) \stackrel{\text{def}}{=} \operatorname*{arg\,max}_{(z,w) \in \mathcal{Z}_\varepsilon(\mu) \times \triangle_K} \inf_{\lambda \in \neg_\varepsilon z} \frac{1}{2} \|\mu - \lambda\|^2_{V_w} \qquad (2)$$

Assumption: the furthest answer for μ is unique, $|z_F(\mu)| = 1$.

Don't choose the greedy answer: aim at identifying the *furthest* answer !

Furthest answer

$$T_\varepsilon(\mu)^{-1} = \max_{z \in \mathcal{Z}_\varepsilon(\mu)} \max_{w \in \triangle_K} \inf_{\lambda \in \neg_\varepsilon z} \frac{1}{2} \|\mu - \lambda\|^2_{V_w}$$

The ε-optimal answer for which its alternative is the easiest to differentiate from thanks to an optimal allocation over arms $w_F(\mu) \in \triangle_K$.

$$(z_F(\mu), w_F(\mu)) \overset{\text{def}}{=} \arg\max_{(z,w) \in \mathcal{Z}_\varepsilon(\mu) \times \triangle_K} \inf_{\lambda \in \neg_\varepsilon z} \frac{1}{2} \|\mu - \lambda\|^2_{V_w} \quad (2)$$

Assumption: the furthest answer for μ is unique, $|z_F(\mu)| = 1$.

Don't choose the greedy answer: aim at identifying the *furthest* answer !

- Multiplicative ε-optimality.
- $d = 2$, $\mathcal{M} = \mathbb{R}^2$, $\mathcal{Z} = \mathcal{K}$ ($K = 4$) and $\mu = (1, 0)$.
- Given ε, generate 25000 random instances: $z_1 = \mu$, $z_2 \in \mathcal{Z}_\varepsilon(\mu)$ and $(z_3, z_4) \in (\mathcal{Z} \setminus \mathcal{Z}_\varepsilon(\mu))^2$.

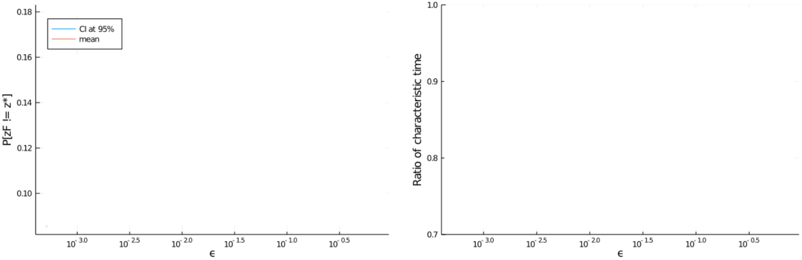


Figure: (a) Proportion of draws where $z_F(\mu) \neq z^*(\mu)$. (b) Median of the ratio between $T_\varepsilon^{\mathrm{mul}}(\mu)$ and the value at $z^*(\mu)$ (when $z_F(\mu) \neq z^*(\mu)$).

Numerical simulations

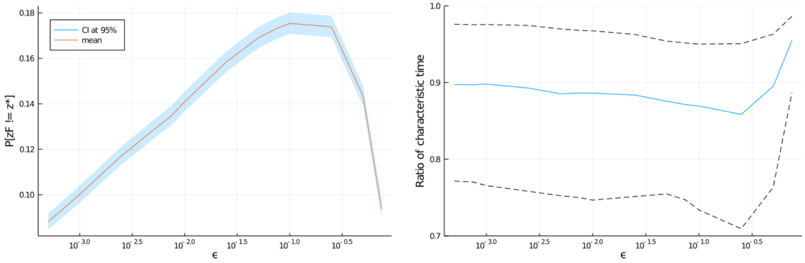


Figure: (a) Proportion of draws where $z_F(\mu) \neq z^*(\mu)$. (b) Median of the ratio between $T_\varepsilon^{\text{mul}}(\mu)$ and the value at $z^*(\mu)$ (when $z_F(\mu) \neq z^*(\mu)$).

- The furthest answer is often different from the greedy answer ($\approx 14\%$).
- The ratio of their characteristic time is on average 0.9.

Section 4

LεBAI

Notations and structure of LεBAI

- counts of pulled arms, $N^a_{t-1} = \sum_{s=1}^{t-1} \mathbf{1}_{\{a_s = a\}}$,
- OLS/ML estimator, $\mu_{t-1} = V_{N_{t-1}}^{-1} \sum_{s=1}^{t-1} X_s^{a_s} a_s$.

After pulling each arm once $(n_0 = K)$, at each round $t \geq n_0 + 1$,

- if the stopping condition for the candidate answer z_t is met, return z_t;
- else, the sampling rule returns an arm a_t to pull and the statistics are updated based on this new observation.

Assumption: set of parameter is bounded by M.

Notations and structure of LεBAI

- counts of pulled arms, $N_{t-1}^a = \sum_{s=1}^{t-1} \mathbf{1}_{\{a_s = a\}}$,
- OLS/ML estimator, $\mu_{t-1} = V_{N_{t-1}}^{-1} \sum_{s=1}^{t-1} X_s^{a_s} a_s$.

After pulling each arm once ($n_0 = K$), at each round $t \geq n_0 + 1$,

- if the stopping condition for the candidate answer z_t is met, return z_t;
- else, the sampling rule returns an arm a_t to pull and the statistics are updated based on this new observation.

Assumption: set of parameter is bounded by M.

Stopping rule

Given $z_t \in \mathcal{Z}_\varepsilon(\mu_{t-1})$, stop when the GLR exceeds $\beta(t-1, \delta)$

$$\inf_{\lambda \in \neg_\varepsilon z_t} \|\mu_{t-1} - \lambda\|^2_{V_{N_{t-1}}} > 2\beta(t-1, \delta) \tag{3}$$

Lemma

Given any sampling and recommendation rules such that $z_t \in \mathcal{Z}_\varepsilon(\mu_{t-1})$, then using (3) with the threshold

$$\beta(t, \delta) = 2K \ln\left(4 + \ln\left(\frac{t}{K}\right)\right) + K\mathcal{C}^{gG}\left(\frac{\ln\left(\frac{1}{\delta}\right)}{K}\right) \tag{4}$$

ensures that $\mathbb{P}_\mu\left[\tau_\delta < +\infty \wedge \hat{z} \notin \mathcal{Z}_\varepsilon(\mu)\right] \leq \delta$. $\mathcal{C}^{gG}(x) \approx x + \ln(x)$ is defined in Kaufmann and Koolen (2018).

Stopping rule

Given $z_t \in \mathcal{Z}_\varepsilon(\mu_{t-1})$, stop when the GLR exceeds $\beta(t-1, \delta)$

$$\inf_{\lambda \in \neg_\varepsilon z_t} \|\mu_{t-1} - \lambda\|^2_{V_{N_{t-1}}} > 2\beta(t-1, \delta) \tag{3}$$

Lemma

Given any sampling and recommendation rules such that $z_t \in \mathcal{Z}_\varepsilon(\mu_{t-1})$, then using (3) with the threshold

$$\beta(t, \delta) = 2K \ln\left(4 + \ln\left(\frac{t}{K}\right)\right) + K\mathcal{C}^{gG}\left(\frac{\ln\left(\frac{1}{\delta}\right)}{K}\right) \tag{4}$$

ensures that $\mathbb{P}_\mu\left[\tau_\delta < +\infty \wedge \hat{z} \notin \mathcal{Z}_\varepsilon(\mu)\right] \leq \delta$. $\mathcal{C}^{gG}(x) \approx x + \ln(x)$ is defined in Kaufmann and Koolen (2018).

Candidate ε-optimal answer

Question: How to choose $z_t \in \mathcal{Z}_\varepsilon(\mu_{t-1})$ to stop as early as possible ?

Natural candidates:

- greedy answer, $z_t = z^*(\mu_{t-1})$, sample inefficient,
- furthest answer, $z_t = z_F(\mu_{t-1})$, computationally inefficient.

The ε-optimal answer with highest GLR is the *instantaneous furthest* answer, $z_t = z_F(\mu_{t-1}, N_{t-1})$ where

$$z_F(\mu_{t-1}, N_{t-1}) \overset{\text{def}}{=} \arg\max_{z \in \mathcal{Z}_\varepsilon(\mu_{t-1})} \inf_{\lambda \in \neg_\varepsilon z_t} \|\mu_{t-1} - \lambda\|^2_{V_{N_{t-1}}}$$

Candidate ε-optimal answer

Question: How to choose $z_t \in \mathcal{Z}_\varepsilon(\mu_{t-1})$ to stop as early as possible ?

Natural candidates:

- greedy answer, $z_t = z^*(\mu_{t-1})$, sample inefficient,
- furthest answer, $z_t = z_F(\mu_{t-1})$, computationally inefficient.

The ε-optimal answer with highest GLR is the *instantaneous furthest* answer, $z_t = z_F(\mu_{t-1}, N_{t-1})$ where

$$z_F(\mu_{t-1}, N_{t-1}) \stackrel{\text{def}}{=} \arg\max_{z \in \mathcal{Z}_\varepsilon(\mu_{t-1})} \inf_{\lambda \in \neg_\varepsilon z_t} \|\mu_{t-1} - \lambda\|^2_{V_{N_{t-1}}}$$

Sampling rule

Maxmin saddle-point algorithm:

- the agent plays $\left(\widetilde{z}_t, w_t^{\mathcal{L}^{\mathcal{K}}}\right) \in \mathcal{Z}_\varepsilon(\mu_{t-1}) \times \triangle_K$ thanks to a $\mathcal{Z}$-oracle and a learner on $\triangle_K$ (e.g. AdaHedge), then
- the *nature* plays the closest alternative, $\lambda_t \in \arg\min_{\lambda \in \neg_\varepsilon \widetilde{z}_t} \|\mu_{t-1} - \lambda\|^2_{V_{w_t}}$ where $w_t = \frac{1}{tK}\mathbf{1}_K + \left(1 - \frac{1}{t}\right) w_t^{\mathcal{L}^{\mathcal{K}}}$ (logarithmic forced exploration).

Algorithmic ingredients:

- tracking, $a_t \in \arg\min_{a \in \mathcal{K}} N_{t-1}^a - W_t^a$ where $W_t = \sum_{s=n_0+1}^t w_s$,
- optimistic gains, $(U_t^a)_{a \in \mathcal{K}}$, used to
- update the learner with $g_t(w) = (1 - \frac{1}{t})\langle w, U_t \rangle$.

Sampling rule

Maxmin saddle-point algorithm:

- the agent plays $\left(\widetilde{z}_t, w_t^{\mathcal{L}^{\mathcal{K}}}\right) \in \mathcal{Z}_\varepsilon(\mu_{t-1}) \times \triangle_K$ thanks to a $\mathcal{Z}$-oracle and a learner on $\triangle_K$ (e.g. AdaHedge), then
- the *nature* plays the closest alternative, $\lambda_t \in \arg\min_{\lambda \in \neg_\varepsilon \widetilde{z}_t} \|\mu_{t-1} - \lambda\|^2_{V_{w_t}}$ where $w_t = \frac{1}{tK}\mathbf{1}_K + \left(1 - \frac{1}{t}\right) w_t^{\mathcal{L}^{\mathcal{K}}}$ (logarithmic forced exploration).

Algorithmic ingredients:

- tracking, $a_t \in \arg\min_{a \in \mathcal{K}} N^a_{t-1} - W^a_t$ where $W_t = \sum_{s=n_0+1}^t w_s$,
- optimistic gains, $(U^a_t)_{a \in \mathcal{K}}$, used to
- update the learner with $g_t(w) = (1 - \frac{1}{t})\langle w, U_t \rangle$.

Upper bound

Theorem

Let $\mathcal{L}^{\mathcal{K}}$ with sub-linear regret and $\mathcal{L}^{\mathcal{Z}}$ such that $\tilde{z}_s \in z_F(\mu_{s-1})$ and Assumption 1 holds true. When recommending the instantaneous furthest answer $z_t = z_F(\mu_{t-1}, N_{t-1})$ and stopping according to (3) with threshold $\beta(t, \delta)$ as in (4) for the exploration bonus $f(t) = 2\beta\left(t, t^{1/3}\right)$, L$\varepsilon$BAI yields a (ε, δ)-PAC algorithm and, for all $\mu \in \mathcal{M}$ such that $|z_F(\mu)| = 1$,

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\mu\left[\tau_\delta\right]}{\ln\left(\frac{1}{\delta}\right)} \leq T_\varepsilon(\mu)$$

Assumption 1 requires the $\mathcal{Z}$-oracle to be *not too good* with respect to a gain not optimized by $\mathcal{Z}$-oracle.

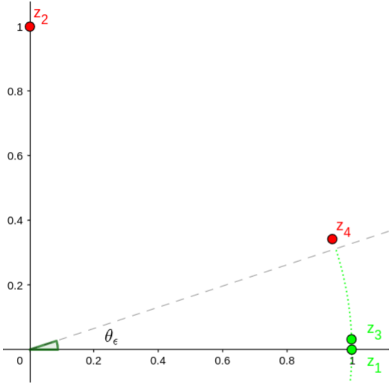
Section 5

Experiments

Hard instance

- multiplicative ε-optimality,
- $(\varepsilon, \delta) = (0.05, 0.1)$,
- 5000 runs (std of means with sub-samples of 100 runs).

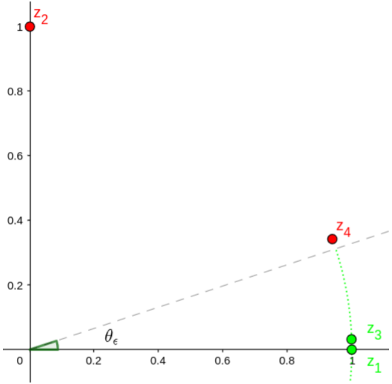
- $d = 2$, $\mathcal{M} = \mathbb{R}^2$, $Z = 4$ and $\mu = (1, 0)$,
- $z_1 = (1, 0)$, $z_2 = (0, 1)$, $z_3 = (\cos(\phi_1), \sin(\phi_1))$, $z_4 = (\cos(\phi_2), \sin(\phi_2))$ where $(\phi_1, \phi_2) = \left(\frac{1}{10}\theta_\varepsilon, \frac{11}{10}\theta_\varepsilon\right)$ and $\theta_\varepsilon = \arccos(1 - \varepsilon)$.



Hard instance

- multiplicative ε-optimality,
- $(\varepsilon, \delta) = (0.05, 0.1)$,
- 5000 runs (std of means with sub-samples of 100 runs).

- $d = 2$, $\mathcal{M} = \mathbb{R}^2$, $Z = 4$ and $\mu = (1, 0)$,
- $z_1 = (1, 0)$, $z_2 = (0, 1)$, $z_3 = (\cos(\phi_1), \sin(\phi_1))$, $z_4 = (\cos(\phi_2), \sin(\phi_2))$ where $(\phi_1, \phi_2) = \left(\frac{1}{10}\theta_\varepsilon, \frac{11}{10}\theta_\varepsilon\right)$ and $\theta_\varepsilon = \arccos(1 - \varepsilon)$.



Candidate ε-optimal answer

Table: Empirical stopping time ($\pm\ \sigma$) on the hard instance for different combinations of sampling rule and recommendation rule with $\mathcal{K} = \{e_1, e_2\}$.

	$z^*(\mu_{t-1})$	$z_F(\mu_{t-1})$	$z_F(\mu_{t-1}, N_{t-1})$
LεBAI			
ε-TaS			
Fixed			
Uniform			

Table: Empirical stopping time ($\pm\ \sigma$) on the hard instance for different combinations of sampling rule and recommendation rule with $\mathcal{K} = \{e_1, e_2\}$.

	$z^*(\mu_{t-1})$	$z_F(\mu_{t-1})$	$z_F(\mu_{t-1}, N_{t-1})$
LεBAI		$244\ (\pm 14)$	$242\ (\pm 13)$
ε-TaS		$235\ (\pm 13)$	$235\ (\pm 13)$
Fixed		$238\ (\pm 12)$	$238\ (\pm 12)$
Uniform		$284\ (\pm 16)$	$284\ (\pm 16)$

- Furthest and instantaneous furthest have almost identical performance.

Heuristic: $\widetilde{z}_t = z_t = z_F(\mu_{t-1}, N_{t-1})$.

Candidate ε-optimal answer

Table: Empirical stopping time ($\pm\ \sigma$) on the hard instance for different combinations of sampling rule and recommendation rule with $\mathcal{K} = \{e_1, e_2\}$.

	$z^*(\mu_{t-1})$	$z_F(\mu_{t-1})$	$z_F(\mu_{t-1}, N_{t-1})$
LεBAI	$264\ (\pm 11)$		$242\ (\pm 13)$
ε-TaS	$252\ (\pm 13)$		$235\ (\pm 13)$
Fixed	$256\ (\pm 12)$		$238\ (\pm 12)$
Uniform	$309\ (\pm 16)$		$284\ (\pm 16)$

- Greedy is sample-inefficient.
- LεBAI has similar performance with ε-TaS and Fixed, and outperforms Uniform.

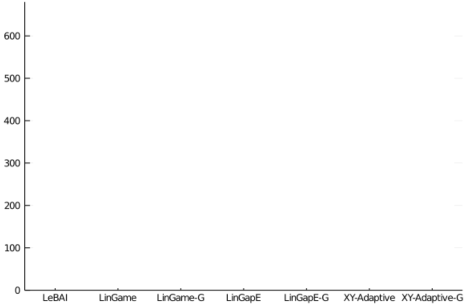


Figure: Empirical stopping time on the hard instance ($\mathcal{K} = \mathcal{Z}$).

⚠ BAI algorithms are modified to use the same stopping rule.

BAI algorithms

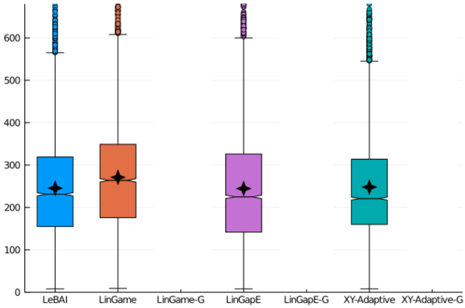


Figure: Empirical stopping time on the hard instance ($\mathcal{K} = \mathcal{Z}$).

LεBAI performs

- slightly better than LinGame and
- on par with LinGapE and $\mathcal{X}\mathcal{Y}$-Adaptive.

BAI algorithms

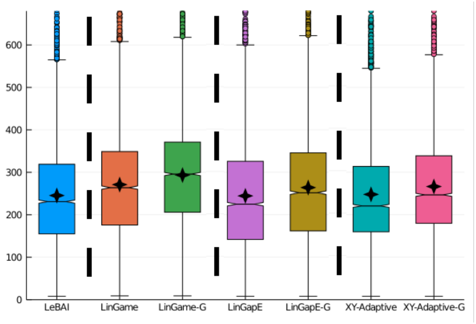


Figure: Empirical stopping time on the hard instance ($\mathcal{K} = \mathcal{Z}$).

- Sample-efficient modification of BAI algorithms for ε-BAI: use the instantaneous furthest answer instead of the greedy answer.

Conclusion

Contributions:

1. Don't choose greedily: aim at identifying the *furthest* answer !
2. LεBAI, asymptotically optimal and empirically competitive.

Open questions/problems:

- Performance of ε-BAI algorithms on BAI tasks.
- Efficient computation of the closest alternative when Z is large.
- Finite-time lower bound for multiple-correct answer.

Conclusion

Contributions:

1. Don't choose greedily: aim at identifying the *furthest* answer !
2. LεBAI, asymptotically optimal and empirically competitive.

Open questions/problems:

- Performance of ε-BAI algorithms on BAI tasks.
- Efficient computation of the closest alternative when Z is large.
- Finite-time lower bound for multiple-correct answer.

References

Degenne, R. and Koolen, W. M. (2019). Pure exploration with multiple correct answers. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14564–14573.

Degenne, R., Ménard, P., Shang, X., and Valko, M. (2020). Gamification of pure exploration for linear bandits. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2432–2442. PMLR.

Fiez, T., Jain, L., Jamieson, K. G., and Ratliff, L. J. (2019). Sequential experimental design for transductive linear bandits. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10666–10676.

Garivier, A. and Kaufmann, E. (2021). Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models. *Sequential Analysis*, 40(1):61–96.

Jedra, Y. and Proutière, A. (2020). Optimal best-arm identification in linear bandits. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kaufmann, E. and Koolen, W. (2018). Mixture martingales revisited with applications to sequential tests and confidence intervals. *arXiv preprint arXiv:1811.11419*.

Kocák, T. and Garivier, A. (2021). Epsilon best arm identification in spectral bandits. In Zhou, Z.-H., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2636–2642. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Soare, M., Lazaric, A., and Munos, R. (2014). Best-arm identification in linear bandits. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 828–836.

Xu, L., Honda, J., and Sugiyama, M. (2018). A fully adaptive algorithm for pure exploration in linear bandits. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 843–851. PMLR.

Questions ?

Appendix

Lemma

When $\overline{\mathcal{M}} = \mathbb{R}^d$ and $V_w^\dagger$ is the Moore-Penrose pseudo-inverse of V_w,

$$2T_\varepsilon^{\text{add}}(\mu)^{-1} = \max_{z \in \mathcal{Z}_\varepsilon^{\text{add}}(\mu)} \max_{w \in \triangle_K} \min_{x \in \mathcal{Z} \setminus \{z\}} \frac{(\varepsilon + \langle \mu, z - x \rangle)^2}{\|z - x\|_{V_w^\dagger}^2}$$

$$2T_\varepsilon^{\text{mul}}(\mu)^{-1} = \max_{z \in \mathcal{Z}_\varepsilon^{\text{mul}}(\mu)} \max_{w \in \triangle_K} \min_{x \in \mathcal{Z} \setminus \{z\}} \frac{\langle \mu, z - (1-\varepsilon)x \rangle^2}{\|z - (1-\varepsilon)x\|_{V_w^\dagger}^2}$$

$$T_\varepsilon^{\text{mul}}(\mu) = \min_{z \in \mathcal{Z}_\varepsilon^{\text{mul}}(\mu)} T_0(\mu, \mathcal{Z}_\varepsilon^z)$$

where $\mathcal{Z}_\varepsilon^z \overset{\text{def}}{=} \{z\} \cup \{(1-\varepsilon)x : x \in \mathcal{Z} \setminus \{z\}\}$

Furthest answer

The ε-optimal answer for which its alternative is the easiest to differentiate from thanks to an optimal allocation over arms $w_F(\mu) \in \triangle_K$.

$$(z_F(\mu), w_F(\mu)) \stackrel{\text{def}}{=} \underset{(z,w)\in\mathcal{Z}_\varepsilon(\mu)\times\triangle_K}{\arg\max} \inf_{\lambda\in\neg_\varepsilon z} \frac{1}{2}\|\mu-\lambda\|^2_{V_w} \tag{5}$$

Assumption: the furthest answer for μ is unique, $|z_F(\mu)| = 1$.

Role in asymptotic optimality:

- $z_F(\mu)$ has to be identified, e.g. $T_\varepsilon^{\mathrm{mul}}(\mu) = T_0\left(\mu, \mathcal{Z}_\varepsilon^{z_F(\mu)}\right)$ where $\mathcal{Z}_\varepsilon^z = \{z\} \cup \{(1-\varepsilon)x : x \in \mathcal{Z} \setminus \{z\}\}$
- Analysis involves a geometric quantity linked to $z_F(\mu)$

Algorithm 1: LεBAI

Input: History $\mathcal{F}_t$, $\mathcal{Z}$-oracle $\mathcal{L}^{\mathcal{Z}}$ and learner $\mathcal{L}^{\mathcal{K}}$.

Output: Candidate ε-optimal answer $\hat{z}$.

Pull once each arm $a \in \mathcal{K}$;

for $t = n_0 + 1, \cdots$ **do**

  Get $z_t =$ RECO;

  If STOP(z_t) then return z_t;

  Get $\left(\tilde{z}_t, w_t^{\mathcal{L}^{\mathcal{K}}}\right)$ from $\mathcal{L}^{\mathcal{Z}} \times \mathcal{L}^{\mathcal{K}}$;

  Let $w_t = \frac{1}{tK}\mathbf{1}_K + \left(1 - \frac{1}{t}\right) w_t^{\mathcal{L}^{\mathcal{K}}}$ and update $W_t = W_{t-1} + w_t$;

  Closest alternative: $\lambda_t \in \arg\min_{\lambda \in \neg_\varepsilon \tilde{z}_t} \|\mu_{t-1} - \lambda\|^2_{V_{w_t}}$;

  Optimistic gains: $\forall a \in \mathcal{K}$, $U_t^a = \left(\|\mu_{t-1} - \lambda_t\|_{aa^\intercal} + \sqrt{c_{t-1}^a}\right)^2$;

  Feed $\mathcal{L}^{\mathcal{K}}$ with gain $g_t(w) = (1 - \frac{1}{t})\langle w, U_t \rangle$;

  Pull $a_t \in \arg\min_{a \in \mathcal{K}} N_{t-1}^a - W_t^a$, observe $X_t^{a_t}$;

end

where $c_{t-1}^a = \min\left\{f\left(s^2\right)\|a\|^2_{V_{N_s}^{-1}}, 4M^2L_{\mathcal{K}}^2\right\}$, $L_{\mathcal{K}} = \max_{a\in\mathcal{K}}\|a\|_2$ and $f(t) = 2\beta\left(t, t^{1/3}\right)$.

Upper bound

Theorem

Let $\mathcal{L}^{\mathcal{K}}$ with sub-linear regret and $\mathcal{L}^{\mathcal{Z}}$ such that $\widetilde{z}_s \in z_F(\mu_{s-1})$ and Assumption 1 holds true. When recommending the instantaneous furthest answer $z_t = z_F(\mu_{t-1}, N_{t-1})$ and stopping according to (3) with threshold $\beta(t,\delta)$ as in (4) for the exploration bonus $f(t) = 2\beta\left(t, t^{1/3}\right)$, L$\varepsilon$BAI yields a (ε,δ)-PAC and, for all $\mu \in \mathcal{M}$ such that $|z_F(\mu)| = 1$,

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\mu}\left[\tau_{\delta}\right]}{\ln\left(\frac{1}{\delta}\right)} \leq T_{\varepsilon}(\mu)$$

Assumption

The $\mathcal{Z}$-oracle $\mathcal{L}^{\mathcal{Z}}$ with $\widetilde{z}_s \in z_F(\mu_{s-1})$ satisfies that there exist $(\alpha_0, C_0) \in [0,1) \times \mathbb{R}_+$ such that almost surely, for all $t > n_0$,
$\max_{z \in \mathcal{Z}} \sum_{s=n_0+1}^{t} \inf_{\lambda \in \neg_{\varepsilon} z} \|\mu_{s-1} - \lambda\|^2_{V_{w_s}} - \sum_{s=n_0+1}^{t} \inf_{\lambda \in \neg_{\varepsilon} \widetilde{z}_s} \|\mu_{s-1} - \lambda\|^2_{V_{w_s}} \geq -C_0 t^{\alpha_0}$.

Proof scheme

$$\mathcal{E}_t = \left\{ \forall s \leq t : \|\mu_s - \mu\|^2_{V_{N_s}} \leq f(t) \right\} \tag{6}$$

Under $\mathcal{E}_t$, if the algorithm does not stop at time $t+1$, the stopping-recommendation pair satisfies

$$2\beta(t,\delta) \geq \max_{z \in \mathcal{Z}} \inf_{\lambda \in \neg_\varepsilon z} \|\mu - \lambda\|^2_{V_{N_t}} - o\left(t + \ln\left(\frac{1}{\delta}\right)\right)$$

while the (anytime) sampling rule verifies

$$\max_{z \in \mathcal{Z}} \inf_{\lambda \in \neg_\varepsilon z} \|\mu - \lambda\|^2_{V_{N_t}} \geq \sum_{s=n_0+1}^{t} g_s\left(w_s^{\mathcal{L}^\mathcal{K}}\right) - o\left(t\right) \geq 2tT_\varepsilon(\mu)^{-1} - o\left(t\right)$$

Using $\beta(t,\delta) = \ln\left(\frac{1}{\delta}\right) + o\left(t + \ln\left(\frac{1}{\delta}\right)\right)$ (and other Lemmas) yields

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\mu\left[\tau_\delta\right]}{\ln\left(1/\delta\right)} \leq T_\varepsilon(\mu)$$

Key challenge in multiple correct answers

Difference:

- BAI: $\mu \in \neg_0 z$ for all $z \neq z^*(\mu)$, hence $\inf_{\lambda \in \neg_0 z} \|\mu - \lambda\|_w^2 = 0$ for all $w \in \mathbb{R}_+^K$.
- ε-BAI: $\mu \in \neg_\varepsilon z$ for all $z \notin \mathcal{Z}_\varepsilon(\mu)$. Need to control those strictly positive terms for ε-optimal answers that are different from the (instantaneous) furthest answer, i.e. for all $z \in \mathcal{Z}_\varepsilon(\mu) \setminus \{z_F(\mu)\}$.

Consequences:

- Assumption 1
- Forced exploration
- Requirement that $(z_t, \widetilde{z}_t) = (z_F(\mu_{t-1}, N_{t-1}), z_F(\mu_{t-1}))$

Additive furthest answer

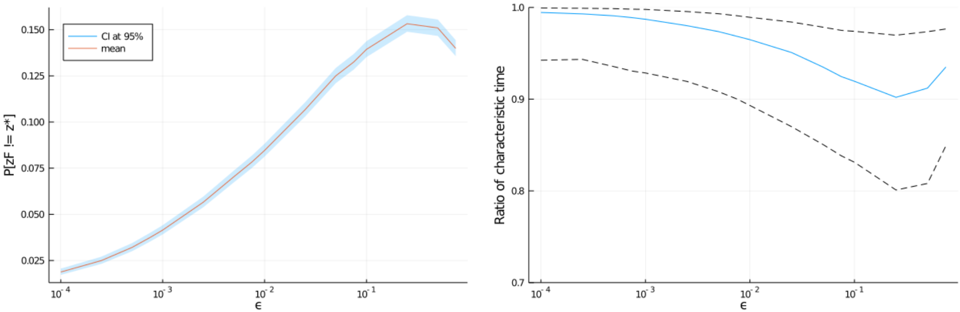


Figure: Influence of ε on (a) the proportion of draws where $z_F(\mu) \neq z^*(\mu)$, (b) the median (and first/third quartile), when $z_F(\mu) \neq z^*(\mu)$, of the ratio between $T_\varepsilon^{\text{add}}(\mu)$ and the value at $z^*(\mu)$, i.e. $\min_{w \in \triangle_K} \sup_{\lambda \in \neg_\varepsilon^{\text{add}} z^*(\mu)} \frac{1}{2}\|\mu - \lambda\|^2_{V_w}$.

BAI algorithms (bis)

Table: Average number of pulls per arm and empirical stopping time ($\pm\ \sigma$) on the hard instance ($\mathcal{K} = \mathcal{Z}$).

	a_1	a_2	a_3	a_4	**Total**
LεBAI	71	155	17	3	246 ($\pm$13)
LinGame	74	153	36	8	271 ($\pm$12)
DKM	111	141	110	110	472 ($\pm$22)
LinGapE	44	198	1	1	245 ($\pm$16)
$\mathcal{X}\mathcal{Y}$-Static	140	142	1	1	284 ($\pm$16)
$\mathcal{X}\mathcal{Y}$-Adaptive	77	169	1	1	248 ($\pm$13)
Fixed	61	173	1	1	236 ($\pm$12)
Uniform	136	136	135	135	541 ($\pm$26)

Random instances

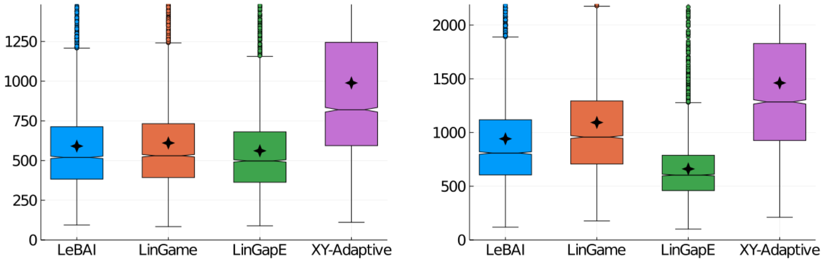


Figure: Empirical stopping time ($\mathcal{K} = \mathcal{Z}$) for $d \in \{6, 12\}$.

Original stopping rule of BAI algorithms

Table: Empirical stopping time ($\pm\ \sigma$) with their original stopping rule or with ours (3) on the hard instance ($\mathcal{K} = \mathcal{Z}$).

	LinGame	LinGapE	$\mathcal{X}\mathcal{Y}$-Adaptive
Original	$102613\ (\pm 15344)$	$146209\ (\pm 16429)$	$302417\ (\pm 29938)$
Modified	$271\ (\pm 41)$	$245\ (\pm 42)$	$248\ (\pm 37)$

Computational relaxations

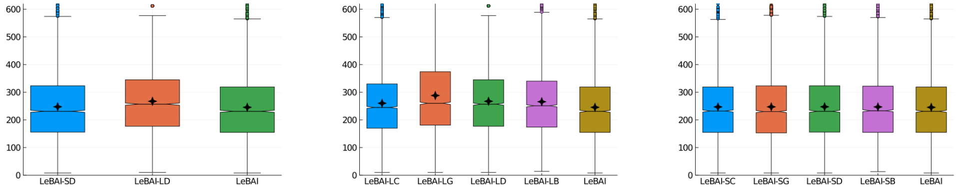


Figure: Empirical stopping time on the hard instance ($\mathcal{K} = \mathcal{Z}$) for (a) the lazy and sticky update, and different implementations of (b) the lazy scheme and (c) the sticky scheme. "-S" denotes the sticky scheme and "-L" the lazy one. The notations for implementations are: "-C" for the constant one with $T_0 = 10$, "-G" for the geometric one with $(T_0, \gamma) = (10, 0.2)$, "-D" for geometrically decreasing one with $(T_0, \gamma) = (10, 0.2)$ and "-B" for the Bernoulli one with parameter $p = 0.1$.

Tracking and forced exploration

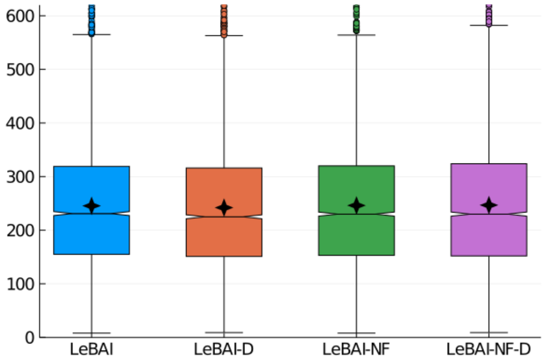


Figure: Empirical stopping time on the hard instance ($\mathcal{K} = \mathcal{Z}$). "-D" denotes when the D-Tracking is used instead of C-Tracking and "-NF" denotes the removal of forced exploration.

Drawings