Lifelong Best-Arm Identification with Misspecified Priors

Anonymous Author(s)
Affiliation
Address
email

Abstract

We address the problem of lifelong fixed-budget best-arm identification (BAI), which arises in realistic sequential A/B testing scenarios where the value of each arm is correlated across test phases. We propose a hierarchical Gaussian model and develop a Bayesian fixed-budget BAI algorithm. Our main contribution is to investigate the impact of prior misspecification on the missidentification probability along the learning trajectory through an upper bound on a novel risk metric. We conduct extensive empirical evaluations of our algorithm against state-of-the-art methods on various types of martingales with different dependency structures. Our results show that our approach outperforms other algorithms across a wide range of settings.

1 Introduction

Practical A/B testing often involves sequentially testing changing options with limited trial budget for each test phase: a new treatment (or version) ($B^{(1)}$) is proposed to replace and improve a legacy one (A) and is tested on a fixed and finite proportion of the traffic. If the test fails ('$A \gtrsim B^{(1)}$'), A remains in place and a new treatment ($B^{(2)}$) must be designed to be tested again against A.

Each of theses phases is a statistical test that can be efficiently performed with a *fixed-budget Best Arm Identification* (BAI) algorithm [White, 2013]. However, for small sample sizes, or when alternatives are hard to distinguish, the error probability may remain high.

Oftentimes though, treatments are tuned models or small modifications of the legacy version, and have close performance to previously tested ones. We leverage this observation to propose an incremental method that adaptively uses all the previous test phases to improve the estimation during the future ones. We call this new problem *Lifelong Best-Arm Identification*.

We build on and extend the *Bayesian Fixed-Budget BAI* framework introduced by Atsidakou et al. [2022], in which the learner can encode side information in a *prior* over bandit instances. The metric studied therein is an expected probability of missidentification (or error) with respect to a prior over bandit instances, but that prior is assumed to be known. In practice, this assumption is often unrealistic and the learner may have to use a misspecified prior. Our contribution is to address the following main questions:

1. What is the cost of using a misspecified prior for *Bayesian Fixed-Budget BAI*?
2. Can we sequentially improve the expected probability of error?

We answer (1) with a general technical result on prior misspecification (Lemma 1). Then we introduce the notion of Lifelong Error for algorithms that sequentially learn a sequence of priors, and we prove a bound on the performance of our algorithm called META-BAYESELIM (Theorem 1).

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

Notation For an integer L, we denote $[L] = \{1, ..., L\}$. We denote by $i_*(\theta)$ the optimal arm of the bandit instance $\theta \in \mathbb{R}^K$, and just i_* when there is no ambiguity. We denote $X_n \to^{\mathcal{L}} X$ when the random variable X_n converges to X *in law*. Throughout, we use the standard notation as in Lattimore and Szepesvári [2020] for ease of readability but the paper is self-contained.

2 Problem setting

We consider a sequence of m fixed-budget BAI problems with n rounds, also called *tasks*. The s-th task is parameterised by a K-tuple $\theta_{s,*} = (\theta^{(1)}_{s,*}, ..., \theta^{(K)}_{s,*}) \in \mathbb{R}^K$. We assume that each instance is sampled *i.i.d.* from an (unknown) distribution: $\theta_{s,*} \sim P_* = \mathcal{N}(\mu_*, \sigma_*^2 I_K)$. The best arm of task $\theta_{s,*}$ is denoted as $i_*(\theta_{s,*})$.

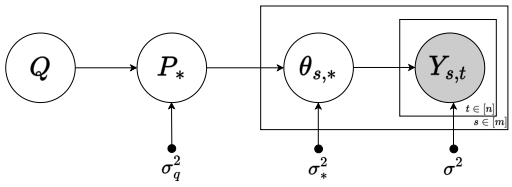


Figure 1: Hierarchical Gaussian Graphical model.

Bayesian Fixed-Budget BAI. Each task consists of n rounds and a policy consists of a *selection rule* that chooses which arm to pull for each round, and a final *decision rule* that decides which arm is the best at the end. More specifically, at each round $t \in [n]$ of a task $s \in [m]$, the agent chooses an arm $A_{s,t} \in [K]$ and observes a (stochastic) reward $Y_{s,t} \sim \mathcal{N}\left(\theta^{(A_{s,t})}_{s,*}, \sigma^2\right)$. By this selection process, the learner collects a history $H_s = (A_{s,1}, Y_{s,1}, ..., A_{s,n}, Y_{s,n})$ and uses it to make a decision $J_s \in [K]$ at the end of task s.

In order to incorporate uncertainty on the choice of the prior in a Bayesian way , we introduce a prior distribution Q over priors on bandit instances $P_* \sim Q$. For computational reasons, we set Q to be a conjugate prior, hence here a Gaussian distribution.

During each test phase $s \in [m]$, the policy π_s must first select actions sequentially to collect H_s, and after round n, the decision follows a given rule. We assume that each π_s is instantiated with a prior over the arms distributions P_s. This sequence of priors is chosen by a meta-policy $H_1, ..., H_{s-1} \mapsto P_s$ which combines evidence from previous phases to progressively (better) set π_s. Thus, in all generality, π_s's decision J_s depends on all the history $H_1, ...H_s$ through its prior.

Lifelong Error Probability We introduce a novel metric to evaluate policies for our problem that accounts for each error rate along the sequence. For a given $m < \infty$, the *Lifelong Error* is the averaged expected probability of missidentification (also refereed as *probability of error*) of $\pi = (\pi_1, ...\pi_m)$ characterized by its sequence of misspecified priors:

$$\mathcal{LE}(\pi, m; P_*) = \frac{1}{m} \sum_{s=1}^{m} \mathbb{E}_{\theta_* \sim P_*} \left[\mathbb{P}_\pi \left(J_s \neq i_* \mid \theta_*, P_s\right)\right] \tag{1}$$

where $\mathbb{P}\left(J \neq i_*\left(\theta_{s,*}\right) \mid \theta_{s,*}, P_s\right)$ is the probability of selecting a suboptimal arm in instance $\theta_{s,*}$ when using a prior $P_s \neq P_*$ almost-surely. This metric extends and generalizes the recently studied *expected probability of missidentification* [Atsidakou et al., 2022], whereby the learner is assumed to know P_* and is evaluated on $\mathcal{E} = \mathbb{E}_{\theta_* \sim P_*}\left[\mathbb{P}\left(J \neq i_* \mid \theta_*, P_*\right)\right]$. Note that indeed, when our learner is given P_*, each term in Eq. (1) is equal and our metric is consistent with theirs. We formally analyse the convergence and consistency properties of our metric in Section 4.

Informally, if the meta-learning process is consistent, $P_s \to^{\mathcal{L}}_{s\to\infty} P_*$, and then $\mathcal{LE}(\pi, m; P_*) \to_{m\to\infty} \mathcal{E}$ almost surely. Otherwise, if $P_s \to^{\mathcal{L}}_{s\to\infty} \tilde{P}$ with $\tilde{P} \neq P_*$ almost everywhere, then $\mathcal{LE}(\pi, m; P_*)$ converges almost surely to the expected probability of error for a fixed misspecified prior, which we bound in a technical result of independent interest in Lemma 1.

Hierarchical Gaussian case In this paper we focus on the hierarchical Gaussian generative model described in Figure 1. More precisely, the (unknown) prior distribution over instances is $P_*(\theta) = \mathcal{N}(\theta|\mu_*, \sigma_*^2)$. We assume σ_*^2 known. We consider μ_* as a random variable whose (known) prior is $Q(\mu) = \mathcal{N}\left(\mu|\mu_q, \sigma_q^2\right)$. Similarly, each prior distribution used at task s are assumed Gaussian: $P_s\left(\mu\right) = \mathcal{N}\left(\mu|\mu_s, \sigma_0^2 I_K\right)$. Since σ_0^2 is assumed known, these prior distributions are fully characterized by μ_s which is considered as a random variable that is sampled from the meta-posterior.

3 Meta-*BayesElim*

Algorithm 1 META-BAYESELIM

Input: Meta prior Q, fixed-budget n, parameters $(\mu_q, \sigma_q^2, \sigma_0^2)$
  $Q_1 \leftarrow Q$
  for $s = 1, \ldots$ **do**
    Sample a prior from the current meta-posterior: $P_s \sim Q_s$
    Run BE for n rounds with prior P_s
    Output $J_s = \text{BE}(P_s, n)$ ▷ highest posterior mean of remaining arms
    Update meta-posterior Q_{s+1} with collected observations (Eq 2)
  end for

We build META-BAYESELIM by adding an outer-loop to BAYESELIM (BE) [Atsidakou et al., 2022] that sequentially updates the choice of prior P_s given as input to the selecting policy BE. The main ideas of BE (see below and Appendix B for details), are to perform successive rejects [Audibert et al., 2010, Karnin et al., 2013] based on the posterior means of the arms, given a budget and a prior. The general pseudo-code of META-BAYESELIM is given in Algorithm 1 and we detail the updates in the Gaussian setting below.

META-BAYESELIM starts with a prior P_1 sampled from the Meta-prior $Q_1 = (\mu_q, \sigma_q^2)$ at the beginning of the first task. The choice of these hyperparameters is left to the practitioner: it is recommended to choose a large enough variance σ_q^2 as one wants maximal adaptivity to the data, *i.e.* low-confidence on the (arbitrary) choice of meta-prior μ_q (see examples in Section 5). Then, at each task $s \in [m]$, a distribution P_s is sampled from the current meta-posterior Q_s and an instance of $\text{BE}(P_s, n)$ is launched with a (potentially misspecified) prior P_s to select actions for n rounds.

For each task $s \in [m]$, and each arm $i \in [K]$, we denote $T_s^{(i)} = \sum_{t \in [n]} \mathbf{1}\{A_{s,t} = i\}$ the number of pulls of arm i, and $\sum_t Y_{s,t}^{(i)}$ the according sum of rewards.

Using the collected data at task $s \in [m]$, the parameters of the meta-posterior Q_{s+1} are updated as:

$$\hat{\mu}_{s+1}^{(i)} = \hat{\sigma}_{s+1,i}^2 \left(\frac{\hat{\mu}_s^{(i)}}{\hat{\sigma}_{s,i}^2} + \frac{T_s^{(i)}}{T_s^{(i)}\sigma_0^2 + \sigma^2} \frac{\sum_t Y_{s,t}^{(i)}}{T_s^{(i)}} \right), \quad \hat{\sigma}_{s+1,i}^2 = \left(\frac{1}{\hat{\sigma}_{s,i}^2} + \frac{T_s^{(i)}}{T_s^{(i)}\sigma_0^2 + \sigma^2} \right)^{-1} \tag{2}$$

The recursive updates of Eq. (2) reflect the transfer of information from task s to task $s+1$. The meta-posterior variance $\hat{\sigma}_{s+1,i}^2$ is strictly decreasing (concentration of the posterior) as we factor in the newly collected information. Similarly, one can see that the posterior mean is easily rewritten as a convex combination of the previous value $\hat{\mu}_s$ and the new empirical average $\sum_t Y_{s,t}/T_s$. We discuss and quantify in depth the impact the hyperparameters in Section 5.

Inner-loop updates and decisions: The full pseudo-code of BAYESELIM is recalled in Alg. 2 (Appendix B) for completeness. It builds on the classic *successive-rejects* algorithm but uses the posterior means of the arms for both the sequential eliminations and the final decision.

The algorithm is a phase-based algorithm. At the end of each phase, it eliminates half of the arms according to their empirical posterior means. Importantly, the prior $\tilde{P}$ used for the inference is given as input arbitrarily. The schedule is determined before the start of the exploration: during each phase, each arm is pulled [1] $n_{\texttt{phase}} = \lfloor n/(K\lceil \log_2(K) \rceil) \rfloor$ times and then half of the arms with the lowest posterior means are eliminated.

The latter are computed using classical Gaussian identities as above. Given a Gaussian prior $\tilde{P}(\mu) = \mathcal{N}\left(\mu | \tilde{\mu}, \sigma_0^2 I_K\right)$, the posterior mean of an arm i at the end of round r is:

$$\bar{\mu}_r^{(i)} = \bar{\sigma}^2 \left(\frac{\tilde{\mu}^{(i)}}{\sigma_0^2} + \frac{\sum_{t \in [n_r]} Y_{i,t}}{\sigma^2} \right), \quad \bar{\sigma}^2 = \left(\frac{1}{\sigma_0^2} + \frac{n_{\texttt{phase}}}{\sigma^2} \right)^{-1} \tag{3}$$

[1] In general, $n_{r,i} = \lfloor \frac{n}{\lceil \log_2(K) \rceil} \frac{\sigma_i^2}{\sum_k \sigma_k^2} \rfloor$ but we assume $\sigma_i^2 = \sigma^2$ here.

Remark 1 (Beyond Gaussian case) *A similar algorithm can be generalized beyond the Gaussian hierarchical model. However, it requires conjugancy properties at both the meta level and the intra-task level. Bernoulli distributions for arms with beta-priors on instances and categorical meta-prior satisfies these properties (See Section 3.2 of Kveton et al. [2021] for more details).*

Remark 2 (Frequentist algorithms) *It would be possible to design a frequentist algorithm based on the successive elimination procedure of [Karnin et al., 2013] together with the usual biased regularization ideas in the meta-learning literature [Denevi et al., 2020]. Though out of the scope of this paper, we discuss initial ideas in Appendix A.*

4 Analysis

We quantify the cost of learning an informative prior and provide an upper bound on the Lifelong Error. We make two major contributions. First, we quantify the cost of running BAYESELIM with a prior $\tilde{P} \neq P_*$ almost everywhere. This extends and gives a practical point of view on the metric introduced by Atsidakou et al. [2022]. Then we bound the Lifelong Errorfor META-BAYESELIM and show that for a correct choice of prior variance $\sigma_0^2 = \sigma_*^2$, META-BAYESELIM 's performance converge to the optimal error probability (Theorem 1).

4.1 Main results

Our first result bounds the expected probability of error of BAYESELIM when given a prior $\tilde{P} \neq P_*$. This technical result is of independent interest and we provide details on the proof further below.

Lemma 1 *Let denote $\tilde{P}(\mu) = \mathcal{N}(\mu|\tilde{\mu}, \sigma_0^2)$ and $P_*(\theta) = \mathcal{N}(\theta|\mu_*, \sigma_*^2)$. The expected probability of error in a $n-$fixed-budget BAI using* BAYESELIM *with a given prior $\tilde{P}$ is upper bounded as follows:*

$$\mathbb{E}_{\theta_* \sim P_*}\left[\mathbb{P}\left(J \neq i_* \mid \theta_*, \tilde{P}\right)\right] \leq 2\log(K) C_{env}^n(\sigma_*^2) \sum_{i\in[K]} \sum_{j\in[K]} e^{-\frac{1}{4\sigma_*^2}(\mu_i^* - \mu_j^*)^2} \phi\left(P_*^{ij}, \tilde{P}^{ij}\right)$$

where C_{env}^n is a constant that depends on the parameters of the environment, the budget n and σ_*^2, and ϕ measures the distance between the unknown distribution P_* and the prior $\tilde{P}$:

$$C_{env}^n(\sigma_*^2) := \sqrt{\frac{\log_2(K)K\sigma^2}{n\sigma_*^2 + \log_2(K)K\sigma^2}}, \quad \phi\left(P_*^{ij}, \tilde{P}^{ij}\right) = e^{C_{env}^n(\sigma_*^2)^2 \cdot \frac{\sigma_*^2}{\sigma_0^4}\left[\frac{\sigma_0^2}{\sigma_*^2}(\mu_i^* - \mu_j^*) - (\tilde{\mu}_i - \tilde{\mu}_j)\right]^2} \tag{4}$$

Remark 3 *For any couple (i, j), $\phi\left(P_*^{ij}, \tilde{P}^{ij}\right) > 1$ for any misspecified prior $\tilde{P}$ and the equality $\phi\left(P_*^{ij}, \tilde{P}^{ij}\right) = 1$ holds if and only if $\tilde{P} = P_*$. In this precise case, we recover exactly the bound stated in Atsidakou et al. [2022].*

The proof is postponed to the next subsection and the necessary technical lemmas are in Appendix C.

We are now ready to prove our a Lifelong Error bound for META-BAYESELIM .

Theorem 1 *Consider the Lifelong fixed-budget BAI setting as described in Section 2 where the prior distributions $(P_1, ..., P_m)$ are generated according to* META-BAYESELIM *. Then with probability at least $1 - m\delta$,*

$$\mathcal{LE}(\pi^{\text{MBE}}, m; P_*) \leq \underbrace{O\left(C_{env}^n(\sigma_*^2) \sum_{i\in[K]} \sum_{j\in[K]} e^{-\frac{1}{4\sigma_*^2}\left(\mu_i^* - \mu_j^*\right)^2}\right)}_{:=\mathcal{E}^*:\ \text{Upper bound of Atsidakou et al. [2022]}} \times \underbrace{\frac{1}{m}\sum_{s=1}^{m} e^{\frac{C_1^\delta}{s} + \frac{C_2^\delta}{\sqrt{s}}|\kappa-1| + (\kappa-1)^2 C_3^\delta}}_{\text{Cost of learning an informative prior}}$$

where we introduced the non-negative problem-dependent constants $C_1^\delta, C_2^\delta, C_3^\delta$ to exhibit dependence on $(\kappa-1)$ and s, and where $\kappa = \frac{\sigma_0^2}{\sigma_*^2}$ represents the prior variance misspecification. The $O(\cdot)$ notation only hides a $\log_2(K)$ dependency and universal constants.

Before discussing the proof of this result, we make a few observations. First, we can distinguish 2 cases depending on the value of κ. In the case where $\kappa = 1$ *i.e.* $\sigma_0^2 = \sigma_*^2$, the Lifelong Error is upper bounded with high probability by:

$$\mathcal{LE}(\pi^{\text{MBE}}, m; P_*) \leq \mathcal{E}^* \times \frac{1}{m} \sum_{s=1}^{m} e^{\mathcal{O}\left(\frac{1}{s}\right)} \tag{5}$$

So, in the long run (as $m \to +\infty$), the cost of learning an informative prior vanishes, and we recover $\mathcal{E}^*$, *i.e.* the case where the learner knows P_*. This is because the multiplicative term in Eq. (5) converges to 1 (Césaro summation). The astute reader will notice that as the bound holds with probability $1 - m\delta$, asymptotic reasoning may be problematic and would require to use accordingly smaller confidence values $\delta_m \to 0$. But we hope that the discussion is useful to understand the metric and the empirical results we present in the next section.

In the case where $\kappa \neq 1$ *i.e.* $\sigma_0^2 \neq \sigma_*^2$, Theorem 1 shows that the learner will suffer a non-vanishing cost of learning P_*, *even in the long run*. This cost is explicitly characterized by the term $e^{(\kappa-1)^2 C_3^\delta}$, which is the asymptotic Césaro limit of this rightmost term.

Remark 4 *Theorem 1 exhibits a $\delta-$PAC, or* Probably approximately correct *bound on the random quantity $\mathcal{LE}$, where the randomness comes from the sequential sampling of the tasks, which can be controlled with high probability. Another choice could be to bound the (deterministic) quantity $\mathbb{E}_{\mu_s \sim Q}[\mathcal{LE}(\pi^{\text{MBE}}, m; P_*)]$. However, this metric choice would only move the problem of prior misspecification one layer up as we would then only get guarantees for an algorithm that knows the meta-prior Q. Moreover, this type of high-probability bound is strictly stronger than bounds in expectation in general. Thus, we believe that a PAC-style bound makes more sense in our setting.*

The proof of this result is a nearly direct consequence of a more general upper bound for any fixed sequence of priors $(P_1, ..., P_m)$ that we prove in Appendix C. We need to handle additionally the concentration of the posterior parameters $(\hat{\mu}_s)_{s\geq 1}$; more details are provided in Appendix C.

4.2 Proof of Lemma 1

Proof: Denote $R = \log_2(K)$. Our analysis is a generalisation of the work of Atsidakou et al. [2022]. We start by bounding the expected probability of error when using a prior $\tilde{P}$ for a *fixed* task θ_*: we first take a union bound over the rounds where the bad event of eliminating the best arm i^* happened,

$$\begin{aligned}
\mathbb{P}\left(J_s \neq i_* \mid \theta_*, \tilde{P}\right) &\leq \mathbb{P}\left(\bigcup_{r\in[R-1]} \left\{ i_*(\theta_*) \notin \mathcal{S}_{r+1} \mid \theta_*, \{i_*(\theta_*) \in \mathcal{S}_r\} \right\}\right) \\
&\leq \sum_{r\in[R-1]} \mathbb{P}\left(i_*(\theta_*) \notin \mathcal{S}_{r+1} \mid \theta_*, \{i_*(\theta_*) \in \mathcal{S}_r\}\right)
\end{aligned}$$

We then apply our technical lemmas (see Appendix C): Lemma 3 states that under the bad event, at least one remaining arm must be badly estimated,

$$\begin{aligned}
\mathbb{P}\left(J_s \neq i_* \mid \theta_*, \tilde{P}\right) &\underset{\text{Lem. 3}}{\leq} 2 \sum_{r\in[R-1]} e^{-\frac{n}{4RK\sigma^2}\left(\theta^*_{j_{r,\theta_*}} - \theta^*_{i_*}\right)^2 - \frac{\left(\tilde{\mu}_{i_*} - \tilde{\mu}_{j_{r,\theta_*}}\right)\left(\theta^*_{j_{r,\theta_*}} - \theta^*_{i_*}\right)}{2\sigma_0^2}} \\
&\leq 2R \max_r \left\{ e^{-\frac{n}{4RK\sigma^2}\left(\theta^*_{j_{r,\theta_*}} - \theta^*_{i_*}\right)^2 - \frac{\left(\tilde{\mu}_{i_*} - \tilde{\mu}_{j_{r,\theta_*}}\right)\left(\theta^*_{j_{r,\theta_*}} - \theta^*_{i_*}\right)}{2\sigma_0^2}} \right\}
\end{aligned}$$

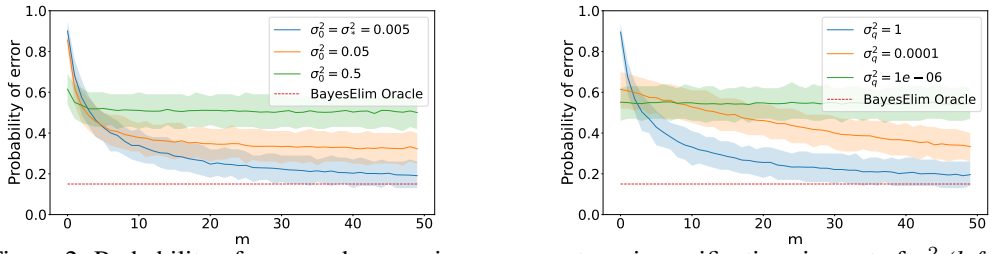


Figure 2: Probability of error under covariance parameter misspecification: impact of σ_0^2 *(left)* and σ_q^2 *(right)*. We set $K = 10$, $\mu_* = (0, 0.1, ..., 0.9)$, $\sigma_*^2 = 5.10^{-3}$, observation noise $\sigma^2 = 10^{-1}$, and budget $n = 30$ for each task.

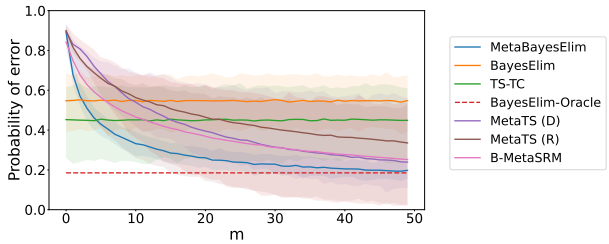


Figure 3: Comparison of the probability of expected missidentification of META-BAYESELIM compared to other fixed-budget BAI algorithms (described in Sec. 5.2. Some baselines also progressively learn the right prior: META-TS(D) , META-TS(R) , B-METASRM . TSTC and BAYESELIM use the same non-informative prior for each task.

Finally, we bound the maximum by summing over possible (i, j):

$$\mathbb{P}\left(J_s \neq i_* \mid \theta_*, \tilde{P}\right) \leq 2R \sum_{i\in[K]} \sum_{j\in[K]} e^{-\frac{n}{4RK\sigma^2}\left(\theta_j^* - \theta_i^*\right)^2 - \frac{\left(\tilde{\mu}_i - \tilde{\mu}_j\right)\left(\theta_j^* - \theta_i^*\right)}{2\sigma_0^2}}$$

To obtain the result claimed in Lemma 1 we integrate the last quantity with respect to P_*:

$$\begin{aligned}\mathbb{E}_{\theta_* \sim P_*}\left[\mathbb{P}\left(J \neq i_* \mid \theta_*, \tilde{P}\right)\right] &\leq 2R \sum_{i\in[K]} \sum_{j\in[K]} \iint_{(\theta_i^*, \theta_j^*)} e^{-\frac{n}{4RK\sigma^2}\left(\theta_j^* - \theta_i^*\right)^2 - \frac{\left(\tilde{\mu}_i - \tilde{\mu}_j\right)\left(\theta_j^* - \theta_i^*\right)}{2\sigma_0^2}} P_*\left(d(\theta_i^*, \theta_j^*)\right) \\ &\underset{\text{Lem. 4}}{\leq} 2RC_{env}^n(\sigma_*^2) \sum_{i\in[K]} \sum_{j\in[K]} e^{-\frac{1}{4\sigma_*^2}(\mu_i^* - \mu_j^*)^2} \phi\left(P_*^{ij}, \tilde{P}^{ij}\right)\end{aligned}$$

□

The upper bound for $\mathcal{LE}(\pi^{\text{MBE}}, m; P_*)$ in Theorem 1 follows from an application of the previous Lemma to a particular sequence of (random) priors. We then use the concentration of this sequence of priors with high probability (details are given in Appendix C.)

5 Experiments

We simulate Lifelong n-fixed budget BAI problems with $K = 10$. The (unknown) bandit instances are generated around linearly separated means $\mu_* = (0, 0.1, ..., 0.9)$ with variance $\sigma_*^2 = 5.10^{-3}$ (this noise level is chosen so that the best arm changes approximately 15% of the time). We set the noise of the observations $\sigma^2 = 10^{-1}$ for a per-task budget of $n = 30$ such that each task remain hard for a naive (non-meta-learning) algorithm.

In each experiment, we start from a non-informative prior $\mu_q = \left(\frac{1}{2}, ..., \frac{1}{2}\right)$ to verify if the involved algorithms adapt to the bandit tasks. The choice of σ_q is studied in a separate experiment as it influences the inference. For each task, the results are averaged over 100 runs on the same environment to obtain an *empirical* probability of error. Then we repeat 100 times this m-task procedure in order

to have confidence intervals over probabilities. We denote BAYESELIM-ORACLE the algorithm BAYESELIM that knows and uses the prior P_*.

We plot the probability of error at task $s \in [m]$ as a function of s, rather than the actual Lifelong Error, though it would converge to the same limit. The intention is to show the progress of the meta-learners more directly.

5.1 Misspecification on the covariance parameters

The variance parameters σ_0 and σ_q are constant and assumed to be known in our theoretical results, and play a crucial role on the convergence of the Lifelong Error (Section 4.1). We study numerically the consequences of misspecifying these parameters.

Influence of σ_0^2 The covariance parameter σ_0^2 reflects the confidence the user has on the prior that is learnt over tasks. As seen in Section 4.1, it should ideally match σ_*^2. Fig. 2 shows that as the value of σ_0^2 decreases towards σ_*^2, the probability of error converges to the performances of BAYESELIM-ORACLE . This result is consistent with Theorem 1. Intuitively, the more σ_0^2 decreases, the more the learner exploits the information collected by the observations during the last tasks. However, for the first tasks, the probability of error suffers from choosing a too small σ_0^2: the learner's confidence in its prior is too strong while too little experience have been collected so far to have an informative one. As we saw in Theorem 1, the long-term cost of learning the prior does not vanish when $\sigma_0^2 \neq \sigma_*^2$.

Influence of σ_q^2 Before the beginning of the first task, the algorithm samples a distribution from the meta-prior distribution with covariance σ_q^2. Having a too strong confidence on this meta-prior (*i.e.* $\sigma_q^2 \to 0$) makes it difficult for the algorithm to update its belief with the observations collected trough interactions with bandit instances. On Figure 2, we show that σ_q directly influences the rate of convergence of the probability of error.

5.2 Comparison to other fixed-budget BAI algorithms

For this benchmark, we set $\sigma_0^2 = \sigma_*^2$ and $\sigma_q^2 = 1$. We compare the performances of META-BAYESELIM with the six following baselines:

- BAYESELIM [Atsidakou et al., 2022] uses the *same* non-informative prior $\tilde{P}$ with mean $\tilde{\mu} = \left(\frac{1}{2}, ..., \frac{1}{2}\right)$ and covariance $\tilde{\sigma}^2 = 1$.
- TSTC [Jourdan et al., 2022] is a variant of Top-Two Thomson Sampling [Russo, 2016] that is the state-of-the-art for the fixed-confidence setting (see benchmarks in [Jourdan et al., 2022]). We adapt it to our setting easily because this strategy is anytime, *i.e.* it does not depend on the confidence level δ. At each round t of each task, the *leader arm* is sampled with probability β. Otherwise, it samples the *challenger arm*. In TSTC , the leader is sampled according to *Thompson-sampling* with a prior $\tilde{\mu}$. The challenger is chosen such that a *Transportation cost w.r.t.* the leader is minimised. The Gaussian setting leads to closed-form solutions. We use the *same* non-informative prior every task, $\tilde{P}$, with mean $\tilde{\mu} = \left(\frac{1}{2}, ..., \frac{1}{2}\right)$ and covariance $\tilde{\sigma}^2 = 1$, and set $\beta = 0.5$ as in prior work.
- META-TS(R) [Kveton et al., 2021] (Meta-Thompson sampling - Random version) is a Meta-learning algorithm that learns an informative prior while minimizing the per-task cumulative (Bayesian) regret. At the end of each task, the best arm is sampled proportionally to the number of pulls.
- META-TS(D) (Meta-Thompson sampling - Deterministic version) similar to META-TS(D) but at the end of each task, arm with the highest posterior mean is returned deterministically.
- B-METASRM [Azizi et al., 2022b] consists in applying ADATS [Basu et al., 2021] to the *simple regret minimisation* setting. Its *simple regret* is bounded as $\tilde{\mathcal{O}}(\frac{m}{\sqrt{n}})$. The prior is updated as the number of tasks increases. Contrarily to META-TS(D) , the prior P_s is not sampled from the meta-posterior Q_s but is directly computed (this is possible in the Gaussian case), leading to a variance reduction and better theoretical performances.

- BAYESELIM-ORACLE corresponds to BAYESELIM using P_* as a prior.

Figure 3 shows that META-BAYESELIM adapts sequentially in the sense that it converges to the probability of error of BAYESELIM-ORACLE that knows P_*. This observation is consistent with the bounds stated in Section 4.1. We remark that B-METASRM (based on ADATS) has better performances than META-TS(R) in this BAI setting, which confirms the observations in Kveton et al. [2021] that marginalizing when possible improves performance.

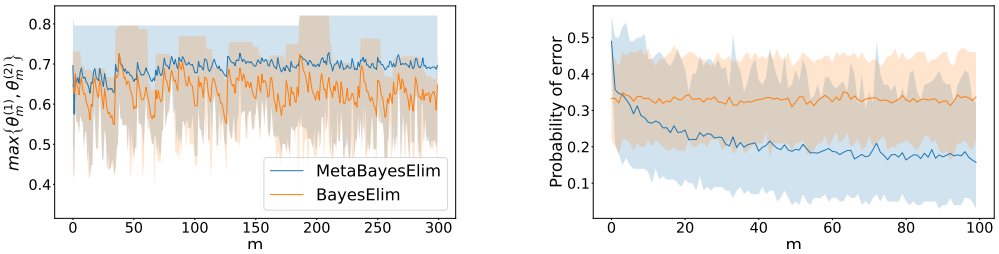
5.3 Example of a sequential A/B testing problem

We consider a more realistic lifelong testing task which violates the Gaussian assumptions we've made so far. We simulate the sequential A/B testing task that we stated with in introduction: we start with $K = 2$ treatments sampled from P_*, (1) and (2), and when a decision is made by the fixed-budget BAI algorithm, the chosen option is kept for the next round, and a new treatment is sampled from the distribution of the eliminated arm (new challenger). This process creates a martingale as follows:

$$\underset{s=1}{\begin{cases} \theta_1^{(1)} \sim \mathcal{N}\left(\mu_*^{(1)}, \sigma_*^2\right) \\ \theta_1^{(2)} \sim \mathcal{N}\left(\mu_*^{(2)}, \sigma_*^2\right) \end{cases}} \xrightarrow{Alg} \underset{s=2}{\begin{cases} \hat{\max}\left\{\theta_1^{(1)}, \theta_1^{(2)}\right\} := \theta_2^{(1)} \\ \theta_2^{(2)} \sim \mathcal{N}\left(\mu_*^{\hat{\min}\left\{\theta_1^{(1)}, \theta_1^{(2)}\right\}}, \sigma_*^2\right) \end{cases}} \xrightarrow{Alg} \cdots \xrightarrow{Alg} \underset{s=m}{\hat{\max}\left\{\theta_m^{(1)}, \theta_m^{(2)}\right\}}$$

We run BAYESELIM and META-BAYESELIM on an instance of this problem with $\mu_* = (0.5, 0.56)$, $\sigma_*^2 = 10^{-2}$, $\sigma^2 = 10^{-1}$, and a small budget $n = 10$ (the task is hard for a non-meta learning algorithm). We initialize both algorithms with a non-informative[2] prior $\sigma_q^2 = 1$, $\mu_q = (0.53, 0.53)$. We show two indicators of performance: as before, the probability of error (returning the action with current lower mean) over tasks is shown in Figure 4b and the point-wise value of the current maximum option is shown in Figure 4a. The latter shows two interesting effects: the martingale generated by MBE converges to a higher value, and its overall variance is much lower, two desirable properties in practice. The probabilities of error in Figure 4b show similar trends as observed in previous experiments where our assumptions on the environment were not violated.

Though this experiment remains a toy problem, we believe it is an interesting and promising result that could open novel areas of research in sequential testing. We discuss these ideas further in Conclusion.



(a) value of the maximal arm for each task.

(b) probability of error.

Figure 4: Sequential A/B testing with $n = 10, K = 2, \mu_* = (0.5, 0.56), \sigma^2 = 10^{-1}, \sigma_*^2 = 10^{-2}$. META-BAYESELIM starts with a non-informative prior $\mu_q = (0.53, 0.53), \sigma_q^2 = 1$ and BAYESELIM uses it for each task.

6 Related Work

Learning the prior for Bayesian inference is a successful idea in Machine Learning. Though it has been explored in classical single-task supervised learning [Rivasplata et al., 2018, Dziugaite et al., 2021] where the dataset can be split before training to 'learn' a data-dependent prior, it fits

[2] the choice of the prior mean μ_q has no effect with this variance, we also tried $\mu_q = (0, 0)$ with no noticeable impact.

more closely the Meta-Learning setting [Thrun and Pratt, 1998] where multiple tasks can be used to meta-learn the task distribution. More specifically, Bayesian and PAC Bayes frameworks [Amit and Meir, 2018, Rothfuss et al., 2021] have exploited the idea of learning a meta-prior.

Online Meta-Learning and Lifelong Learning both refer to the idea of performing meta-learning in an online learning scenario where tasks, and sometimes even data, are streamed with either full-information [Alquier et al., 2017, Khodak et al., 2019] or bandit feedback [Jedor et al., 2020, Cella et al., 2020, Kveton et al., 2021, Basu et al., 2021, Simchowitz et al., 2021]. The latter setting, also called *Lifelong bandits*, has recently received attention, especially around 'Meta-regret' minimization problems (sum of the per-task regrets) where a structure of the task space can be learnt [Azizi et al., 2022a, Schur et al., 2022].

Pure Exploration and Best-Arm identification are comparatively less thoroughly studied in the Lifelong setting and we are only aware of [Azizi et al., 2022b] tackling the expected simple-regret in the fixed-confidence setting and optimizing. For single-task problems, pure exploration in bandits is now well-understood both in the simple regret minimization setting [Bubeck et al., 2009, Audibert et al., 2010] and in BAI with fixed-confidence [Gabillon et al., 2012, Kaufmann et al., 2016] and fixed-budget [Degenne, 2023, Karnin et al., 2013, Even-Dar et al., 2006] constraints.

7 Conclusion

Existing works on Bayesian Fixed-budget BAI assumed the prior to be known [Atsidakou et al., 2022], which can be impractical. We relax this assumption and study a PAC-style new metric for Lifelong Best-Arm Identification. Based on related works [Kveton et al., 2021, Azizi et al., 2022b, Basu et al., 2021], we propose META-BAYESELIM , the first algorithm to sequentially adapt to an unknown environment in the fixed-budget BAI setting. Our theoretical and numerical analysis of META-BAYESELIM show that it converges towards the same error probability as algorithms that assume the prior known.

Our work also opens new directions of research around sequential testing in structured and/or changing environments.

Martingales of testing problems. Our sequential A/B testing problem is the first of a kind, to the best of our knowledge. Many practical applications of machine learning algorithms feature this *performative* aspect [Perdomo et al., 2020]: the learner's action impact the next learning problems. We believe there is a lack of understanding of this key issue in sequential testing problems, both in fixed-confidence and fixed-budget settings. Another key practical issue is that of making 'safe' decisions such that the chain remains as stable as possible by only changing treatment when the new one is significantly better than the legacy one. One could for instance introduce meta-switching costs [Dekel et al., 2014], or conservative constraints [Wu et al., 2016].

On stopping easy problems. Consider the case $K = 2$ and a particular environment where $\sigma_*^2 \ll \Delta_*$ (the meta-gap). In such setting, when a good prior is known, testing is no longer needed as with high-probability the best arm is always the same. An interesting metric is the *stopping time* for which the lifelong algorithm outputs the best arm indicated by the prior information. We believe that our results pave the way to give a problem-dependent bound on the expectation of this stopping time.

META-BAYESELIM in linear bandit setting. As mentioned in Atsidakou et al. [2022], an interesting direction could be to extend the Bayesian FB-BAI setting to linear bandit models so far only studied with frequentist metrics [Soare et al., 2014, Azizi et al., 2021]. We could extend our lifelong-learning setting to adapt to the case where the prior is unknown to the learner. For the Gaussian hierarchical model, the computations in the contextual setting are tractable [Basu et al., 2021].

References

P. Alquier, M. Pontil, et al. Regret bounds for lifelong learning. In *Artificial Intelligence and Statistics*, pages 261–269. PMLR, 2017.

R. Amit and R. Meir. Meta-learning by adjusting priors based on extended pac-bayes theory. In *International Conference on Machine Learning*, pages 205–214. PMLR, 2018.

A. Atsidakou, S. Katariya, S. Sanghavi, and B. Kveton. Bayesian fixed-budget best-arm identification. *arXiv preprint arXiv:2211.08572*, 2022.

J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.

M. Azizi, T. Duong, Y. Abbasi-Yadkori, A. György, C. Vernade, and M. Ghavamzadeh. Non-stationary bandits and meta-learning with a small set of optimal arms. *arXiv preprint arXiv:2202.13001*, 2022a.

M. Azizi, B. Kveton, M. Ghavamzadeh, and S. Katariya. Meta-learning for simple regret minimization. *arXiv preprint arXiv:2202.12888*, 2022b.

M. J. Azizi, B. Kveton, and M. Ghavamzadeh. Fixed-budget best-arm identification in structured bandits. *arXiv preprint arXiv:2106.04763*, 2021.

S. Basu, B. Kveton, M. Zaheer, and C. Szepesvári. No regrets for learning the prior in bandits. *Advances in neural information processing systems*, 34:28029–28041, 2021.

J. Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, pages 23–37. Springer, 2009.

L. Cella, A. Lazaric, and M. Pontil. Meta-learning with stochastic linear bandits. In *International Conference on Machine Learning*, pages 1360–1370. PMLR, 2020.

R. Degenne. On the existence of a complexity in fixed budget bandit identification. *arXiv preprint arXiv:2303.09468*, 2023.

O. Dekel, J. Ding, T. Koren, and Y. Peres. Bandits with switching costs: T 2/3 regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 459–467, 2014.

G. Denevi, D. Stamos, C. Ciliberto, and M. Pontil. Online-within-online meta-learning. *Advances in Neural Information Processing Systems*, 32, 2019.

G. Denevi, M. Pontil, and C. Ciliberto. The advantage of conditional meta-learning for biased regularization and fine tuning. *Advances in Neural Information Processing Systems*, 33:964–974, 2020.

G. K. Dziugaite, K. Hsu, W. Gharbieh, G. Arpino, and D. Roy. On the role of data in pac-bayes bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 604–612. PMLR, 2021.

E. Even-Dar, S. Mannor, Y. Mansour, and S. Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25, 2012.

M. Jedor, J. Louëdec, and V. Perchet. Lifelong learning in multi-armed bandits. *arXiv preprint arXiv:2012.14264*, 2020.

M. Jourdan, R. Degenne, D. Baudry, R. de Heide, and E. Kaufmann. Top two algorithms revisited. In *NeurIPS 2022-36th Conference on Neural Information Processing System*, 2022.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016.

K. Khetarpal, C. Vernade, B. O'Donoghue, S. Singh, and T. Zahavy. Pomrl: No-regret learning-to-plan with increasing horizons. *arXiv preprint arXiv:2212.14530*, 2022.

M. Khodak, M.-F. F. Balcan, and A. S. Talwalkar. Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32, 2019.

I. Kuzborskij and F. Orabona. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950. PMLR, 2013.

B. Kveton, M. Konobeev, M. Zaheer, C.-w. Hsu, M. Mladenov, C. Boutilier, and C. Szepesvari. Meta-thompson sampling. In *International Conference on Machine Learning*, pages 5884–5893. PMLR, 2021.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

A. Maurer and T. Jaakkola. Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6(6), 2005.

J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

O. Rivasplata, E. Parrado-Hernández, J. S. Shawe-Taylor, S. Sun, and C. Szepesvári. Pac-bayes bounds for stable algorithms with instance-dependent priors. *Advances in Neural Information Processing Systems*, 31, 2018.

J. Rothfuss, V. Fortuin, M. Josifoski, and A. Krause. Pacoh: Bayes-optimal meta-learning with pac-guarantees. In *International Conference on Machine Learning*, pages 9116–9126. PMLR, 2021.

D. Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418. PMLR, 2016.

F. Schur, P. Kassraie, J. Rothfuss, and A. Krause. Lifelong bandit optimization: No prior and no regret. *arXiv preprint arXiv:2210.15513*, 2022.

M. Simchowitz, C. Tosh, A. Krishnamurthy, D. J. Hsu, T. Lykouris, M. Dudik, and R. E. Schapire. Bayesian decision-making under misspecified priors with applications to meta-learning. *Advances in Neural Information Processing Systems*, 34:26382–26394, 2021.

M. Soare, A. Lazaric, and R. Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

S. Thrun and L. Pratt. Learning to learn: Introduction and overview. *Learning to learn*, pages 3–17, 1998.

J. White. *Bandit algorithms for website optimization*. " O'Reilly Media, Inc.", 2013.

Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262. PMLR, 2016.

A Frequentist and Bayesian approach in the Gaussian case

The frequentist equivalent of the Bayesian inference problem we addressed in this paper (Gaussian case) is a least squares regression with biased regularization. These ideas were proposed early in the meta-learning literature [Baxter, 2000] and have been studied extensively since then [Denevi et al., 2020, Kuzborskij and Orabona, 2013, Maurer and Jaakkola, 2005].

Under a linear model, the biased regularization algorithm is

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^K} \sum_{s=1}^{t} \|Y_s - X_s^\top \theta\|^2 + \lambda \|\theta - \theta_0\|^2$$

In our case, $X_s = (\mathbf{1}\{A_s = i\})_{i=1..K}$ is the indicator vector of the chosen arms and the closed-form solution is

$$\hat{\theta}_i = \frac{\sum_{s=1}^{t} Y_s^{(i)} - T_t^{(i)} \theta_0^{(i)}}{T_t^{(i)} + t\lambda}$$

The parameter θ_0 is the regularization bias and λ balances out the confidence the learner has in this bias, similarly to the effect of the prior variance choice in our model. As usual in learning theory, the parameter λ can be optimized to minimize a bias-variance trade-off of the resulting estimator. Similarly to Cella et al. [2020], Denevi et al. [2019], Khetarpal et al. [2022], one could progressively improve the choice of bias θ_0 over tasks using the collected data from previous tasks. This would result in a similar estimator to ours, with the additional issue of having to tune the parameter $\lambda_{(m,n)}$ (now potentially depending on the current task, the budget and other problem-dependent quantities).

To obtain a frequentist equivalent of META-BAYESELIM , one would need to prove high-probability concentration bounds for the estimator discussed above (for well-chosen values of λ), and then use the successive-rejects algorithm. This alone is beyond the scope of this paper. Bayesian regret bounds on this frequentist algorithm could be obtain using the ideas in Atsidakou et al. [2022].

B BAYESELIM

In this subsection we recall the psuedo-code of BAYESELIM . We recall $R = \lceil \log_2(K) \rceil$, and the posterior means are computed as follows :

$$\bar{\mu}_r^{(i)} = \bar{\sigma}^2 \left(\frac{\tilde{\mu}^{(i)}}{\sigma_0^2} + \frac{\sum_{t \in [n_r]} Y_{i,t}}{\sigma^2} \right), \quad \bar{\sigma}^2 = \left(\frac{1}{\sigma_0^2} + \frac{n_{\texttt{phase}}}{\sigma^2} \right)^{-1} \tag{6}$$

Algorithm 2 BAYESELIM with prior $\tilde{P}$

Input: Fixed-budget n
  $S_1 \leftarrow [K]$
  for $r = 1, ...R$ **do**
    for $i \in \mathcal{S}_r$ **do**
      Get $n_{r,i}$ samples of arm i
      Compute posterior means $(\bar{\mu}_i)_i$
    end for
    Update $\mathcal{S}_{r+1}$.
  end for

C Detailed proofs

We first recall the table of notations.

Notation	Signification
$\mathcal{S}_r$	set of active arms at round r when playing BAYESELIM
$R = \lceil \log_2(K) \rceil$	total amount of rounds
σ^2	variance of the observations of arm
n	budget per task
$Y_{i,t}$	observation of arm i at round t
P_*	prior distributions on bandit instances
$(\mu_*, \sigma_*^2 I_K)$	parameters of P_*
P_s	prior used by the algorithm at task s
$(\mu_s, \sigma_0^2 I_K)$	parameters of P_s
Q	meta-prior
$(\mu_q, \sigma_q^2 I_K)$	parameters of the meta-prior
Q_s	meta-posterior at task s
$\left(\hat{\mu}_s, \hat{\Sigma}_s\right)$	parameters of the meta-posterior Q_s at task s

Table 1: Notations used in this paper

We include the main technical lemmas in the main paper. Lemma 2 and Lemma 3 correspond to Lemma 4.3 and Lemma 4.4 in Atsidakou et al. [2022] respectively ; we state these Lemmas in Appendix C.1 for completeness, but we do not provide the proofs. Lemma 4 is mainly a technical lemma. Lemma 5 corresponds to Lemma 6 of Kveton et al. [2021]. The only difference is that we bound $|\mu_i^*|$ for any coordinate i with probability $1 - \frac{\delta}{K}$ instead of bounding the norm of the whole vector $\|\mu^*\|_\infty$ with probability $1 - \delta$. Since the arguments of the proofs are not different from theirs, we do not provide the proof.

C.1 Main Lemmas

Lemma 2 *For any instance θ_* and round $r \in [R]$, suppose we use the prior $\tilde{P}(\mu) = \mathcal{N}(\mu|\tilde{\mu}, \sigma_0^2 I_K)$. then for any arm $i \in S_r$, the posterior means are correctly ordered with high probability:*

$$\mathbb{P}\left(\bar{\mu}_{i,n_{r,i}} > \bar{\mu}_{i_*,n_{r,i_*}}|\theta_*\right) \leq 2\exp\left(-\frac{n}{4R|\mathcal{S}_r|\sigma^2}\left(\theta_i^* - \theta_{i_*}^*\right)^2 - \frac{\left(\tilde{\mu}_{i_*} - \tilde{\mu}_i\right)\left(\theta_i^* - \theta_{i_*}^*\right)}{2\sigma_0^2}\right)$$

Lemma 3 *With the same notations as in Lemma 2, there exists an arm $j_{r,\theta_*} \in S_r \setminus \{i_*\}$ such that the probability of wrongly eliminating i^* in round r is bounded as:*

$$\mathbb{P}\left(i_* \notin S_{r+1}|\left\{i_* \in S_r\right\}, \theta_*\right) \leq 2\exp\left(-\frac{n}{4R|\mathcal{S}_r|\sigma^2}\left(\theta_{j_{r,\theta_*}}^* - \theta_{i_*}^*\right)^2 - \frac{\left(\tilde{\mu}_{i_*} - \tilde{\mu}_{j_{r,\theta_*}}\right)\left(\theta_{j_{r,\theta_*}}^* - \theta_{i_*}^*\right)}{2\sigma_0^2}\right)$$

Lemma 4 *Let $c_1, c_2 > 0$. Assume the case where P_* is a product of 2 Gaussian distributions: $P_*\left(d(\theta_i, \theta_j)\right) = \mathcal{N}\left(d\theta_i|\mu_i^*, \sigma_*^2\right)\mathcal{N}\left(d\theta_j|\mu_j^*, \sigma_*^2\right)$. Then for any positive constant $c_1, c_2 > 0$ we have the following identity:*

$$\int_{(\theta_i,\theta_j)} e^{-c_1(\theta_i-\theta_j)^2 - c_2\frac{(\theta_i-\theta_j)(\tilde{\mu}_i-\tilde{\mu}_j)}{\sigma_0^2}} P_*\left(d(\theta_i, \theta_j)\right) = \frac{1}{\sqrt{1+4c_1\sigma_*^2}} e^{-\frac{c_1\sigma_*^2+c_2-c_2^2}{\sigma_*^2(4c_1\sigma_*^2+1)}\left(\mu_i^*-\mu_j^*\right)^2}$$
$$.e^{\frac{\sigma_*^2 c_2^2}{(4c_1\sigma_*^2+1)\sigma_0^4}\left(\left(\frac{\sigma_0}{\sigma_*}\right)^2\left(\mu_i^*-\mu_j^*\right)-(\tilde{\mu}_i-\tilde{\mu}_j)\right)^2}$$

Lemma 5 *Let $\mu_* \sim \mathcal{N}\left(\mu_q, \sigma_q^2 I_K\right)$ and the prior parameter in task s sampled such that $\mu_s | H_{1:s-1} \sim \mathcal{N}\left(\hat{\mu}_s, \hat{\Sigma}_s\right)$. Then for each arm $i \in [K]$, with probability at least $1 - \frac{m\delta}{K}$,*

$$|\mu_s^i - \mu_*^i| \leq 2\sqrt{2\frac{\sigma_0^2 + \sigma^2}{(\sigma_0^2 + \sigma^2)\,\sigma_q^{-2} + s - 1}\log\left(\frac{4K}{\delta}\right)}$$

C.2 Proof of technical Lemmas

Proof: [Lemma 4] This is mainly a technical proof which combines classical results in Bayesian statistics. Since we integrate *w.r.t.* a joint measure that is the product of 2 measures, we can integrate one at a time the term with respect to $P(d\theta_i)$ then $P(d\theta_j)$ (*Fubini* theorem). Since the proof is mostly computational, we only give to the reader the big lines of the proof.

Let denote $\mathcal{I} := \iint_{(\theta_i,\theta_j)} e^{-c_1(\theta_i-\theta_j)^2 - c_2\frac{(\theta_i-\theta_j)(\tilde{\mu}_i-\tilde{\mu}_j)}{\sigma_0^2}} P_*\left(d(\theta_i,\theta_j)\right)$

Integrate *w.r.t.* $\mathcal{N}(d\theta_i|\mu_i^*, \sigma_*^2)$:

$$\int_{\theta_i} e^{-c_1(\theta_i^2 - 2\theta_i\theta_j) - \frac{c_2}{\sigma_0^2}\theta_i(\tilde{\mu}_i - \tilde{\mu}_j)} e^{-\frac{1}{2\sigma_*^2}(\theta_i - \tilde{\mu}_i)^2} \mathcal{N}(d\theta_i|\mu_i^*, \sigma_*^2) = \sqrt{2\pi\sigma_a^2} e^{\frac{1}{2\sigma_a^2}m_a^2} e^{-\frac{1}{2\sigma_*^2}\tilde{\mu}_i^2}$$

where

$$m_a = \frac{\sigma_*^2}{2c_1\sigma_*^2 + 1}\left(2\theta_j c_1 - \frac{c_2(\tilde{\mu}_i - \tilde{\mu}_j)}{\sigma_0^2} + \frac{\tilde{\mu}_i}{\sigma_*^2}\right), \quad \sigma_a^2 = \frac{\sigma_*^2}{2c_1\sigma_*^2 + 1}$$

Integrate *w.r.t.* $\mathcal{N}(d\theta_j|\mu_j^*, \sigma_*^2)$: the last yields to :

$$\begin{aligned}\mathcal{I} &= \int_{\theta_j} e^{-c_1\theta_j^2 + \frac{c_2}{\sigma_0^2}\theta_j(\tilde{\mu}_i - \tilde{\mu}_j)} e^{-\frac{1}{2\sigma_*^2}(\theta_j - \tilde{\mu}_j)^2} \frac{\sqrt{2\pi\sigma_a^2}}{2\pi\sigma_*^2} e^{\frac{1}{2\sigma_a^2}m_a^2} e^{-\frac{\tilde{\mu}_i^2}{2\sigma_*^2}} \mathcal{N}(d\theta_j|\mu_j^*, \sigma_*^2) \\ &= \frac{\sigma_a\sigma_b}{\sigma_*^2}.\exp\left(\frac{1}{2}\left(\frac{\sigma_*^2}{2c_1\sigma_*^2 + 1}\left(\frac{\tilde{\mu}_i\sigma_0^2 - c_2\sigma_*^2(\tilde{\mu}_i - \tilde{\mu}_j)}{\sigma_0^2\sigma_*^2}\right)^2 + \frac{m_b^2}{\sigma_b^2} - \frac{1}{\sigma_*^2}\left((\tilde{\mu}_i^2 + \tilde{\mu}_j^2)\right)\right)\right)\end{aligned}$$

where

$$\begin{cases} m_b = \dfrac{1}{4c_1\sigma_*^2 + 1}.\dfrac{c_2\left(\tilde{\mu}_i - \tilde{\mu}_j\right)\sigma_*^2\left(2c_1\sigma_*^2 + 1\right) + \tilde{\mu}_j\sigma_0^2\left(2c_1\sigma_*^2 + 1\right) + 2c_1\left(\frac{\tilde{\mu}_i}{\sigma_*^2} - \frac{c_2(\tilde{\mu}_i - \tilde{\mu}_j)}{\sigma_0^2}\right)\sigma_*^4\sigma_0^2}{\sigma_0^2} \\ \sigma_b^2 = \dfrac{\sigma_*^2\left(2c_1\sigma_*^2 + 1\right)}{4c_1\sigma_*^2 + 1} \end{cases}$$

Rearranging the terms yields the desired identity.

□

C.3 Proof of Theorem 1

We recall

$$C_{env}^n(\sigma_*^2) := \sqrt{\frac{\log_2(K)K\sigma^2}{n\sigma_*^2 + \log_2(K)K\sigma^2}} \tag{7}$$

A direct application of Lemma 1 gives :

$$\mathcal{LE}(\pi^{\text{MBE}}, m; P_*) \leq 2\log(K) C^n_{env}(\sigma_*^2)^2 \sum_i \sum_j e^{-\frac{1}{4\sigma_*^2}(\mu_i^* - \mu_j^*)^2} . \frac{1}{m} \sum_{s=1}^{m} e^{C^n_{env}(\sigma_*^2)^2 . \frac{\sigma_*^2}{\sigma_0^4}\left[\frac{\sigma_0^2}{\sigma_*^2}(\mu_i^* - \mu_j^*) - (\mu_i^s - \mu_j^s)\right]^2} \tag{8}$$

This bound, however, is not fully explicit. The last term on the right-hand side depends on quantities $\mu_s^{(i,j)}$ which will concentrate with $s > 0$. To study this phenomenon, it remains to bound the term $e^{C^n_{env}(\sigma_*^2)^2 . \frac{\sigma_*^2}{\sigma_0^4}\left[\frac{\sigma_0^2}{\sigma_*^2}(\mu_i^* - \mu_j^*) - (\mu_i^s - \mu_j^s)\right]^2}$ at each task s.

Let $\kappa = \frac{\sigma_0^2}{\sigma_*^2}$:

$$\begin{aligned}
\left[\frac{\sigma_0^2}{\sigma_*^2}(\mu_i^* - \mu_j^*) - (\mu_i^s - \mu_j^s)\right]^2 &= \left[(\mu_i^* - \mu_i^s) - \left(\mu_j^* - \mu_j^s\right) - (\kappa - 1)\left(\mu_j^* - \mu_i^*\right)\right]^2 \\
&\leq (\mu_i^* - \mu_i^s)^2 + \left(\mu_j^* - \mu_j^s\right)^2 - 2\left(\mu_i^* - \mu_i^s\right)\left(\mu_j^* - \mu_j^s\right) + (\kappa - 1)^2\left(\mu_j^* - \mu_i^*\right)^2 \\
&- 2(\kappa - 1)\left(\mu_j^* - \mu_i^*\right)\left[(\mu_i^* - \mu_i^s) - \left(\mu_j^* - \mu_j^s\right)\right] \\
&\leq (\mu_i^* - \mu_i^s)^2 + \left(\mu_j^* - \mu_j^s\right)^2 + (\kappa - 1)^2\left(\mu_j^* - \mu_i^*\right)^2 \\
&+ 2|\kappa - 1|\left(\mu_j^* - \mu_i^*\right)\left[(\mu_i^* - \mu_i^s) - \left(\mu_j^* - \mu_j^s\right)\right]
\end{aligned} \tag{8}$$

We use the Lemma 5 to exploit the concentration of posterior distributions. Let's call ψ_s the upper bound stated in this lemma so that for *any* arm $i \in [K]$, with probability at least $1 - \frac{m\delta}{K}$,

$$|\mu_s^i - \mu_*^i| \leq \psi_s(\delta) := 2\sqrt{2\frac{\sigma_0^2 + \sigma^2}{(\sigma_0^2 + \sigma^2)\,\sigma_q^{-2} + s - 1}\log\left(\frac{4K}{\delta}\right)}$$

Thus, this bound holds simultaneously for *all* arms with probability at least $1 - m\delta$ (union bound on all arms).

From properties of Gaussian distributions, there exits a positive constant $B^* < +\infty$ such that for any arm i, with probability at least $1 - \frac{m\delta}{K}$,

$$|\mu_i^*| \leq B_*(\delta)$$

so that this bound holds simultaneously for *all* arms with probability $1 - m\delta$.

We are now ready bound each term of Eq. (8) with probability at least $1 - 2m\delta$: since $(\mu_i^* - \mu_i^s)^2 \leq \psi_s^2(\delta)$ and $\left(\mu_j^* - \mu_i^*\right) \leq |\mu_j^*| + |\mu_i^*| \leq 2B_*(\delta)$,

$$\begin{aligned}
\left[\frac{\sigma_0^2}{\sigma_*^2}(\mu_i^* - \mu_j^*) - (\mu_i^s - \mu_j^s)\right]^2 &\leq 2\psi_s^2(\delta) + (\kappa - 1)^2 4B_*^2(\delta) + 2|\kappa - 1| 2B_*(\delta) 2\psi_s(\delta) \\
&\leq \underbrace{2\psi_s^2(\delta)}_{\tilde{\mathcal{O}}(\frac{1}{s})} + |\kappa - 1| . \underbrace{4B_*(\delta)\psi_s(\delta)}_{\tilde{\mathcal{O}}(\frac{1}{\sqrt{s}})} + (\kappa - 1)^2 . 4B_*^2(\delta)
\end{aligned}$$

Injecting this last inequality to Eq. (8) gives the desired identity.

Finally, in order to have a bound which holds with probability at least $1 - m\delta$, we just have to set δ to $\frac{\delta}{2}$.

D A general Lifelong Error bound for any arbitrary sequence of priors

Theorem 2 *Assume m prior distributions $(P_1, ..., P_m)$ are given to the learner, where for any s, $P_s(\mu) = \mathcal{N}\left(\mu|\mu_s, \sigma_0^2 I_K\right)$. At each task s, the learner runs* BAYESELIM *with n rounds using prior P_s. In this setting, the associated Lifelong Error is upper bounded as follows:*

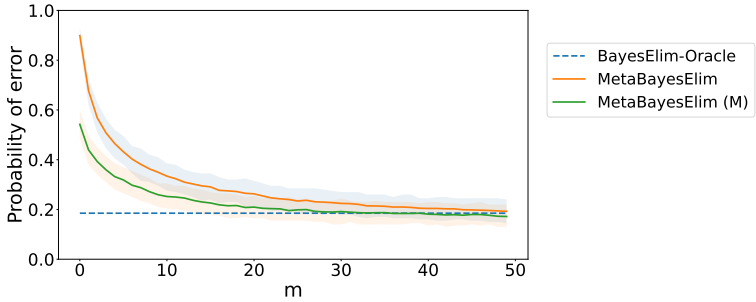


Figure 5: Probability of error when sampling prior μ_s from meta-posterior (META-BAYESELIM) or computing from marginalization (META-BAYESELIM (M)) at each task s.

$$\mathcal{LE}(\pi^{BE}, m; P_*) \leq \frac{2\log_2(K)\, C_{env}^n(\sigma_*^2)}{m} \sum_{i\in[K]} \sum_{j\in[K]} e^{-\frac{1}{4\sigma_*^2}\left(\mu_i^* - \mu_j^*\right)^2} \sum_{s=1}^{m} \phi\left(P_*^{ij}, P_s^{ij}\right)$$

The term $\sum_{s=1}^{m} \phi\left(P_*^{ij}, P_s^{ij}\right)$ reflects the cost of running BE with m possibly misspecified priors. An example of an explicit bound is given in Theorem 1 for our Gaussian setting.

The proof is a direct application of Lemma 1 and a sum over tasks of the errors. For a given choice of model and prior, the divergences ϕ's can be derived and bounded explicitly.

E Additional numerical experiments

E.1 META-BAYESELIM without sampling

In this section, we introduce META-BAYESELIM (M) which applies the same idea of ADATS from Basu et al. [2021] to META-BAYESELIM . More precisely, at each task s, the prior P_s is not sampled from the meta-posterior anymore but directly computed from a marginalization *w.r.t.* the meta-posterior distribution. In the Gaussian case,

$$P_s(\mu) = \mathcal{N}\left(\mu | \hat{\mu}_s, \hat{\Sigma}_s + \sigma_0^2 I_K\right)$$

where $\hat{\Sigma}_s$ is the diagonal covariance matrix whose entries are the $\hat{\sigma}_s^2$ defined in Eq.(2). Figure 5 shows that META-BAYESELIM (M) performs slightly better than META-BAYESELIM for a small amount of instances m, then converges asymptotically to the same value. This is not surprising since the prior is computed *via* marginalizing the meta-prior, which yields to a reduction of variance. However, this marginalization is only possible in very particular cases due to computational tractability. The sampling scheme $\mu_s \sim Q_s$ applies to a broader class of distributions with the use of sampling methods (*e.g.* MCMC sampling).

E.2 Case of strongly changing environments

We test the robustness of our meta-algorithm in cases where there is little or no structure. In these situation it should not be relevant to learn a specific prior and the question is whether doing so would significantly hurt performance or not.

We study the case of strongly changing environment ; we set $\sigma_*^2 = 0.5$, such that the best arm changes 70% of the time. The data observed in previous epochs might not be useful for the current task. Figure 6 shows that setting $\sigma_0^2 \ll \sigma_*^2$ is detrimental in this situation: there is almost no structure in the bandit instances $\theta_{*,s}$, so, setting a too low value of σ_0^2 biases the outcomes too much and not appropriately.

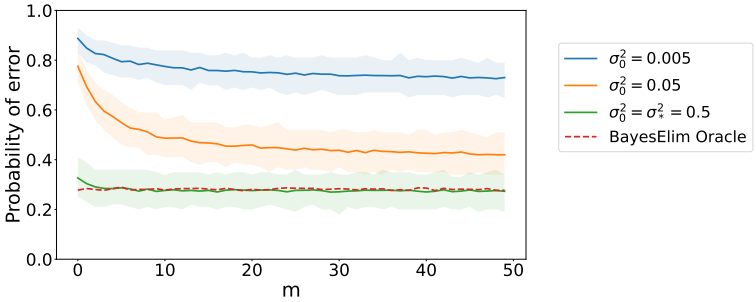


Figure 6: Probability of error under misspecification of σ_0^2. We set $K = 10$, $\mu_* = (0, 0.1, ..., 0.9)$, $\sigma_*^2 = 0.5$, observation noise $\sigma^2 = 10^{-1}$, and budget $n = 30$ for each task.

With the exact choice $\sigma_0^2 = \sigma_*^2$, META-BAYESELIM stagnates around the performance of the Oracle (green curve), which itself is not very good (around 30% error rate). This means that, at least in this case, learning the correct prior does not hurt but it also does not help.