

---

# Prior-Dependent Allocations for Bayesian Fixed-Budget Best-Arm Identification in Structured Bandits

---

Nicolas Nguyen\*  
University of Tübingen

Imad Aouali\*  
Criteo AI Lab  
CREST, ENSAE, IP Paris

András György  
Google DeepMind

Claire Vernade  
University of Tübingen

## Abstract

We study the problem of Bayesian fixed-budget best-arm identification (BAI) in structured bandits. We propose an algorithm that uses fixed allocations based on the prior information and the structure of the environment. We provide theoretical bounds on its performance across diverse models, including the first prior-dependent upper bounds for linear and hierarchical BAI. Our key contribution lies in introducing novel proof techniques that yield tighter bounds for multi-armed BAI compared to existing approaches. Our work provides new insights into Bayesian fixed-budget BAI in structured bandits, and extensive experiments demonstrate the consistent and robust performance of our method in practice across various settings.

## 1 INTRODUCTION

Best arm identification (BAI) addresses the challenge of finding the optimal arm in a bandit environment (Lattimore and Szepesvári, 2020), with wide-ranging applications in online advertising, drug discovery or hyperparameter tuning. BAI is commonly approached through two primary paradigms: *fixed-confidence* and *fixed-budget*. In the fixed-confidence setting (Even-Dar et al., 2006; Kaufmann et al., 2016), the objective is to find the optimal arm with a pre-specified confidence level. Conversely, fixed-budget BAI (e.g., Audibert et al., 2010) involves identifying the optimal arm within a fixed number of observations. Within this fixed-budget context, two main metrics are used: the

---

\* Equal contribution.

*probability of error* (PoE) (Audibert et al., 2010; Karnin et al., 2013; Carpentier and Locatelli, 2016), which is the likelihood of incorrectly identifying the optimal arm, and the *simple regret* (Bubeck et al., 2009; Russo, 2016; Komiya et al., 2023), which corresponds to the expected performance disparity between the chosen and the optimal arm. We study PoE minimization in fixed-budget BAI in the Bayesian setting (Atsidakou et al., 2022), a problem closer to statistical hypothesis testing (Bernardo and Rueda, 2002) than to regret minimization (Lattimore and Szepesvári, 2020), and often viewed as a more difficult question in general (Carpentier and Locatelli, 2016; Degenne, 2023).

### The question of adaptivity in fixed-budget BAI.

Existing algorithms for PoE minimization in fixed-budget BAI are largely frequentist and often employ elimination strategies, such as Sequential Halving (SH) (Karnin et al., 2013), that can be derived for linear models using optimal designs (Azizi et al., 2021; Hoffman et al., 2014; Katz-Samuels et al., 2020; Yang and Tan, 2022). These algorithms are called *adaptive* in the sense that the sampling strategy progressively allocates more budget to higher-valued arms. That being said, BAI is a hard problem whose complexity remains hard to fully understand (Degenne, 2023), and for which simple baselines can be deceptively strong on some problem instances. Wang et al. (2024) showed that for two-armed Bernoulli bandits, there is no algorithm that is strictly better than the uniform sampling algorithm in all instances when comparing the frequentist PoE. In fact, Degenne (2023) established that there is no superior adaptive algorithms in several BAI problems, including Gaussian BAI. This questions the role of adaptivity for fixed-budget BAI problems (Qin, 2022).

In the Bayesian setting, however, when prior knowledge is available to the learner, the competitiveness of the uniform strategy must be reassessed. As a motivating example, consider a scenario with 3 arms, where prior information strongly suggests that either of the first two arms is more likely to be optimal than the third. How much budget should then be allocated to the seemingly

suboptimal third arm? Moreover, if there is greater confidence in the first arm compared to the second, what is the optimal budget distribution between them? In this paper, we address the following general question: *How to choose a prior-dependent fixed allocation of the budget to guarantee a low expected PoE?*

**Prior works on Bayesian BAI.** Bayesian approaches have focused on the minimization of the simple regret (Azizi et al., 2023; Komiya et al., 2023), or were studied under a frequentist lens (Hoffman et al., 2014; Yang et al., 2024), which do not capture the advantages of knowing informative priors. Only recently, Atsidakou et al. (2022) introduced a Bayesian version of Sequential Halving for BAI in the  $K$ -armed Gaussian multi-armed bandit (MAB) setting. They incorporate prior knowledge in two ways: the allocations in each elimination phase prioritize arms with higher observation noise variance, and the final decision is based on the posterior mean which is a function of the prior and of the observations. However, their analysis directly relies on averaging frequentist upper bounds over the prior and require assumptions on the prior distribution; unfortunately, for technical reasons,<sup>1</sup> it does not generalize to more informative priors and structured bandits. In particular, to the best of our knowledge, it cannot be extended to linear bandits through optimal design. Our work provides a novel algorithm that naturally applies to structured bandits and gives a general answer to Bayesian fixed-budget BAI with non-adaptive allocation strategies.

**Contributions.** **1)** We present and analyze a fixed-budget BAI algorithm, which we call Prior-Informed BAI (PI-BAI), that leverages prior information for efficient exploration. Our main contributions include establishing prior-dependent upper bounds on its expected PoE in multi-armed, linear, and hierarchical bandits. Specifically, in the MAB setting, our upper bound is tighter and valid under milder assumptions on the prior than that of Atsidakou et al. (2022). **2)** The proof techniques developed for PI-BAI provide a fully Bayesian perspective, significantly diverging from existing proofs that rely on frequentist techniques. This gives a more comprehensive framework for understanding and analyzing Bayesian BAI algorithms, while also enabling us to relax previously held assumptions. **3)** As a result, unlike earlier approaches, our algorithms and proofs are naturally applicable to structured problems, such as linear and hierarchical bandits, leading to the first Bayesian BAI algorithm with a prior-dependent PoE bound in these settings. **4)** We empirically evaluate PI-BAI’s variants in numerous setups. Our experiments on synthetic and real-world

data show the generality of PI-BAI and highlight its good performance.

## 2 A NEW ALGORITHM CONCEPT

**Notation.** For any positive integer  $K$ , let  $\Delta_K$  be the  $K$ -simplex and  $\Delta_K^+ = \{\omega \in \Delta_K; \omega_i > 0, \forall i \in [K]\}$ , where  $[K] = \{1, 2, \dots, K\}$ . For any positive-definite matrix  $A \in \mathbb{R}^{d \times d}$  and vector  $v \in \mathbb{R}^d$ , we define  $\|v\|_A = \sqrt{v^\top A v}$ . Also,  $\lambda_1(A)$  and  $\lambda_d(A)$  denote the maximum and minimum eigenvalues of  $A$ , respectively. We denote by  $e_i$  the  $i$ -th vector of the canonical basis, and by  $\text{KL}$  the Kullback-Leibler divergence.

**Background.** We consider scenarios involving  $K$  arms. In each round  $t \in [n]$ , the agent selects an arm  $A_t \in [K]$ , and then receives a stochastic reward  $Y_t \sim P(\cdot; \theta, A_t)$ , where  $\theta$  is an *unknown* parameter vector and  $P(\cdot; \theta, i)$  is the *known* reward distribution of arm  $i$ , given  $\theta$ . We denote by  $r(i; \theta) = \mathbb{E}_{Y \sim P(\cdot; \theta, i)}[Y]$  the mean reward of arm  $i$ , given  $\theta$ . We adopt the Bayesian view where  $\theta$  is assumed to be sampled from a *known* prior distribution  $P_0$ . Given a bandit instance characterized by  $\theta \sim P_0$ , the goal is to find the (unique) optimal arm  $i_*(\theta) = \text{argmax}_{i \in [K]} r(i; \theta)$  by interacting with the bandit instance for a fixed-budget of  $n$  rounds. These interactions are summarized by *the history*  $H_n = \{(A_t, Y_t)\}_{t=1}^n$ , and we denote  $J_n \in [K]$  the arm selected by the agent after  $n$  rounds. In this Bayesian setting, Atsidakou et al. (2022) study the Bayesian risk, which they call *expected PoE*:

$$\mathcal{P}_n = \mathbb{E} [\mathbb{P}(J_n \neq i_*(\theta) \mid \theta)] ,$$

that corresponds to the average PoE over instances sampled from the prior,  $\theta \sim P_0$ . This is different from the frequentist counterpart where the performance is assessed for a single  $\theta$ .

### 2.1 Prior-Informed BAI

We consider the following algorithm:

---

**Algorithm 1** Prior-Informed BAI: PI-BAI(Alloc)

---

**Input:** Budget  $n$ , prior  $P_0$ , function `Alloc` (e.g. Sec 3)  
 Compute allocations  $\omega = \text{Alloc}(n, P_0)$ .

**for**  $i = 1, \dots, K$  **do**

    Get  $n_i = \lfloor \omega_i n \rfloor$  samples of arm  $i$   
     Compute mean posterior reward  $\mathbb{E}[r(i; \theta) \mid H_n]$

Set  $J_n = \text{argmax}_{i \in [K]} \mathbb{E}[r(i; \theta) \mid H_n]$ .

---

PI-BAI (Algorithm 1) takes as input a prior distribution  $P_0$ , a budget  $n$  and an allocation function `Alloc`. Then, PI-BAI uses the function `Alloc` to compute a vector of prior-dependent allocation weights  $\omega = (\omega_i)_{i \in [K]} \in \Delta_K$ . This vector is central to the good performance of our method; choices how to compute it are discussed in

<sup>1</sup>We provide a detailed discussion in Appendix A and Section 5.

details in the next section. After that, PI-BAI proceeds in a straightforward manner: it collects  $n_i = \lfloor \omega_i n \rfloor$  samples for each arm  $i$ , and finally returns the arm with the highest mean posterior reward, defined as

$$J_n = \operatorname{argmax}_{i \in [K]} \mathbb{E} [r(i; \theta) \mid H_n]. \quad (1)$$

This posterior expectation can be computed in closed-form in Gaussian bandits with Gaussian priors. Note that a fully Bayesian algorithm would set  $J_n = \operatorname{argmin}_{j \in [K]} \mathbb{P}(j \neq i_*(\theta) \mid H_n)$ ; however, this cannot be computed explicitly even in the simple Gaussian setting, and we instead opt for a decision rule that is deterministic conditionally on observations, as in top-two based algorithms in the fixed-confidence setting (Jourdan et al., 2022).

Decision rule (1) naturally generalizes to structured bandit settings: the structure is a direct prior information that is taken into account in the computation of the posterior updates and the allocation vector. Similarly, and despite additional technicalities, our novel proof scheme (see Section 5) is preserved across all settings we consider.

**Analysis overview.** For any prior-dependent allocation vector  $\omega \in \Delta_K^+$  (computed by `Alloc`), we derive an upper bound on the PoE in multi-armed, linear and hierarchical bandits (see below Theorems 4.1, 4.2 and 4.4) of the form

$$\mathcal{P}_n \leq C(P_0, \omega, n), \quad (2)$$

where the bound  $C$  depends on the prior  $P_0$ , the budget  $n$  and the allocation weights  $\omega$ . The bound is valid for any  $\omega \in \Delta_K^+$ , but its guarantee, as well as the performance of the algorithm, weakens for worse choices of  $\omega$ . To alleviate this, a key aspect of our work is to derive good (prior-dependent) allocation strategies.

### 3 ALLOCATIONS STRATEGIES

In this section, we propose three approaches to derive allocation strategies to be used in PI-BAI, each leveraging the prior in different ways, and we show in the next section how they can be implemented in structured bandit models.

**Allocation by optimization.** Since the bound (2) is valid for any  $\omega \in \Delta_K^+$ , we can define the *optimized allocation weights* as

$$\operatorname{Opt}(n, P_0) = \operatorname{argmin}_{\omega \in \Delta_K^+} C(P_0, \omega, n). \quad (3)$$

We call the resulting algorithm PI-BAI(Opt), where we set the allocation function `Alloc` = `Opt`. In general, (3) is non-convex and may require advanced numerical methods (e.g., L-BFGS-B of Virtanen et al., 2020) to

find a suitable optimum. Fortunately, this optimization is performed only once (offline) before interacting with the environment.

In our experiments, as a heuristic, we define a mixture of optimized weights:  $\alpha \omega_i^{\operatorname{opt}} + (1 - \alpha) \frac{\mu_{0,i} \sigma_{0,i}}{\sum_{k \in [K]} \mu_{0,k} \sigma_{0,k}}$  where  $\alpha \in (0, 1)$  whose performance is still covered by our upper bound. We tested various  $\alpha$  values (Appendix D.7) and found the best performance at  $\alpha \approx 0.5$ . For notational simplicity, we still refer to this allocation weight (with  $\alpha = 0.5$ ) as  $\omega^{\operatorname{opt}}$ .

**Allocation by G-optimal design.** Optimal experimental design (Lattimore and Szepesvári, 2020, Chapter 21) is an optimization problem that returns an allocation over arms that balances exploration and exploitation, usually used in linear bandit algorithms with fixed-design (Abbasi-Yadkori et al., 2011). We notice that this idea can also be leveraged to obtain allocations for Bayesian BAI. Finding an (approximate) Bayesian G-optimal design (López-Fidalgo, 2023, Chapter 4) is equivalent to maximizing the log-determinant of the regularized information matrix defined as

$$\operatorname{G-opt}(n, P_0) = \operatorname{argmax}_{\omega \in \Delta_K^+} \log \det \left( \frac{n}{\sigma^2} \Sigma_\omega + \Sigma_0^{-1} \right), \quad (4)$$

where  $(x_1, \dots, x_k)$  are the action vectors (unit vectors in the MAB setting) and  $\Sigma_\omega = \sum_{k \in [K]} \omega_k x_k x_k^\top$ . This leads to budget allocations that minimize the worst-case posterior variance in all directions.

Note that both allocation weights `Opt`( $n, P_0$ ) (3) and `G-opt`( $n, P_0$ ) (4) are prior-dependent and fixed in advance before interacting with the environment (hence they are independent of the actual instantiation  $\theta$  of the environment). Therefore, both enjoy the theoretical guarantees we derive in Sections 4.1 and 4.2 that hold for any allocation weights fixed in advance.

**Allocation by warm-up.** Finally, we introduce an improper ‘adaptive’ strategy that calls a possibly prior-dependent *warm-up* policy  $\pi_w$  to interact with the bandit environment for  $n_w$  rounds (the warm-up phase). This no longer corresponds exactly to the PI-BAI algorithm as this `Alloc` strategy also interacts with the environment and outputs instance-dependent weights  $\omega_{\pi_w}$ , thus our theoretical guarantees do not apply to this framework. The warm-up policy can be any decision-making strategy that (preferably) also uses the prior. Inspired by its strong performance in BAI (Lee et al., 2024), we use Thompson sampling for  $\pi_w$ . The allocation weights are then set proportional to the number of pulls for each arm during the warm-up phase:  $\omega_i^{\operatorname{TS}} = \frac{1}{n_w} \sum_{t \in [n_w]} \mathbb{1}\{A_t = i\}$  for all  $i \in [K]$ . These strategies are further discussed in Appendix B.4.

## 4 BANDIT SETTINGS

In this section, we present how to concretely implement our algorithm PI-BAI for several major bandit settings using the three types of allocations discussed above, and prove the corresponding theoretical guarantees.

### 4.1 Multi-Armed Bandits

In this setting, each component of  $\theta = (\theta_i)_{i \in [K]}$  is sampled independently from the prior distribution. We focus on the Gaussian case where  $\theta_i \sim \mathcal{N}(\mu_{0,i}, \sigma_{0,i}^2)$ , with  $\mu_{0,i}$  and  $\sigma_{0,i}^2$  being the *known* prior reward mean and variance for arm  $i$ . Then, given  $\theta$ , the reward distribution of arm  $i$  is  $\mathcal{N}(\theta_i, \sigma^2)$  where  $\sigma^2$  is the (known) observation noise variance,<sup>2</sup>

$$\begin{aligned} \theta_i &\sim \mathcal{N}(\mu_{0,i}, \sigma_{0,i}^2) & \forall i \in [K] \\ Y_t \mid \theta, A_t &\sim \mathcal{N}(\theta_{A_t}, \sigma^2) & \forall t \in [n]. \end{aligned} \quad (5)$$

Under (5), the posterior distribution of  $\theta_i$  given  $H_n$  is a Gaussian  $\mathcal{N}(\hat{\mu}_{n,i}, \hat{\sigma}_{n,i}^2)$  (Bishop, 2006) with

$$\hat{\sigma}_{n,i}^{-2} = \sigma_{0,i}^{-2} + n_i \sigma^{-2}, \quad \hat{\mu}_{n,i} = \hat{\sigma}_{n,i}^2 \left( \frac{\mu_{0,i}}{\sigma_{0,i}^2} + \frac{B_{n,i}}{\sigma^2} \right), \quad (6)$$

where, defining  $\mathcal{T}_i = \{t \in [n], A_t = i\}$  to be the set of rounds when arm  $i$  is chosen,  $n_i = |\mathcal{T}_i|$  is the number of times arm  $i$  is chosen and  $B_{n,i} = \sum_{t \in \mathcal{T}_i} Y_t$  is the sum of rewards of arm  $i$ . The *mean posterior reward* in this case is  $\mathbb{E}[r(i; \theta) \mid H_n] = \hat{\mu}_{n,i}$ .

**PoE bound for MAB.** Theorem 4.1 presents an upper bound on the expected PoE of PI-BAI for MAB (5) for a *fixed and possibly prior-dependent* choice of positive allocation weights. The general proof scheme is presented in Section 5, and the full proof is provided in Appendix C.2.

**Theorem 4.1** (Upper bound for MAB). *For all  $\omega \in \Delta_K^+$ , the expected PoE of PI-BAI that uses allocation  $\omega$  under the MAB problem (5) is upper bounded as*

$$\mathcal{P}_n \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{1}{\sqrt{1+n\phi_{i,j}(\omega)}} \exp\left(-\frac{(\mu_{0,i}-\mu_{0,j})^2}{2(\sigma_{0,i}^2+\sigma_{0,j}^2)}\right),$$

where

$$\phi_{i,j}(\omega) = \frac{\sigma_{0,i}^4 \omega_i \left( \frac{\sigma^2}{n} + \omega_j \sigma_{0,j}^2 \right) + \sigma_{0,j}^4 \omega_j \left( \frac{\sigma^2}{n} + \omega_i \sigma_{0,i}^2 \right)}{\sigma^2 \sigma_{0,i}^2 \left( \frac{\sigma^2}{n} + \omega_j \sigma_{0,j}^2 \right) + \sigma^2 \sigma_{0,j}^2 \left( \frac{\sigma^2}{n} + \omega_i \sigma_{0,i}^2 \right)}.$$

In particular,  $\phi_{i,j}(\omega) = \Omega(1)$  depends on the prior parameters and allocation weights, and  $\mathcal{P}_n = \mathcal{O}(1/\sqrt{n})$ .

As a sanity check, note that if the prior is informative, either with small prior variances  $\sigma_{0,i}^2 \rightarrow 0$  or large prior

<sup>2</sup>The noise variance  $\sigma^2$  could be arm-dependent and we can provide a similar analysis, but we choose to keep notation simple.

gaps  $|\mu_{0,i} - \mu_{0,j}| \rightarrow \infty$ , then  $\mathcal{P}_n \rightarrow 0$  for any fixed allocation weights  $\omega \in \Delta_K^+$ .

**Tightness of our bound.** First we compare our upper bound to the lower bound of Atsidakou et al. (2022) for the 2-armed Gaussian setting with  $\sigma_{0,1}^2 = \sigma_{0,2}^2 = \sigma_0^2$ :

$$\frac{e^{-\frac{(\mu_{0,1}-\mu_{0,2})^2}{4\sigma_0^2}}}{\sqrt{1+\frac{n\sigma_0^2}{2\sigma^2}}} \stackrel{\text{Th. 4.1}}{\geq} \mathcal{P}_n \stackrel{\text{LB}}{\geq} \frac{e^{-\frac{(\mu_{0,1}-\mu_{0,2})^2}{4\sigma_0^2}}}{2e\sqrt{1+\frac{2n(8\log(2n)+1)\sigma_0^2}{\sigma^2}}}$$

Comparing the formulas above, we can see that in 2-arm problems we achieve optimal guarantees up to log factors and multiplicative universal constants.

To further test the tightness of our bound, we numerically compare it with that of BayesElim (Atsidakou et al., 2022) using the 3-armed bandit example in Section 1, where one arm is a priori suboptimal, while one of the other two is a priori optimal, reflected by prior means  $\mu_0 = (1, 1.9, 2)$ . We consider two scenarios: one with homogeneous variances  $\sigma_{0,1} = \sigma_{0,2} = \sigma_{0,3} = 0.3$  and another with heterogeneous variances  $(\sigma_{0,1}, \sigma_{0,2}, \sigma_{0,3}) = (0.1, 0.5, 0.5)$ . Since BayesElim's bound does not handle heterogeneous prior variances, we use an average prior variance  $\sigma_0^2 = \frac{1}{K} \sum_{i \in [K]} \sigma_{0,i}^2$  for comparison. PI-BAI is instantiated with the following allocation strategies: Opt, using (3) with Theorem 4.1; Uniform, with  $\omega_i = 1/K$  for all  $i$ ; Random, where  $\omega_i$  are generated uniformly at random from  $[0, 1]$  and then normalized to sum up to 1; and Heuristic, which gives more weight to arms with higher prior means and variances as  $\omega_i \propto \mu_{0,i} \sigma_{0,i}$ .

As shown in Figure 1, our bound is tighter than the upper bound of Atsidakou et al. (2022) for any allocations and in both variance settings. This gain is due to two main reasons: our bound is tighter by a factor  $\log_2^{3/2}(K)$  because our algorithm does not rely on elimination phases. Moreover, our proof technique yields terms that explicitly depend on the structure of the prior, which provides deeper insights beyond the  $\mathcal{O}(1/\sqrt{n})$  rate typically obtained by averaging the frequentist bound of  $\mathcal{O}(e^{-n/f(\theta)})$  (Audibert et al., 2010; Carpentier and Locatelli, 2016) over a Gaussian prior, though it does not necessarily result in faster rates.

**Allocation by optimization: a simple example in the MAB setting.** To provide intuition, we present the explicit solution of (3) with Theorem 4.1 for the MAB setting with  $K = 2$ : denoting  $\omega^{\text{opt}} = \text{Opt}(n, P_0)$ ,

$$\omega_1^{\text{opt}} = \Pi_{[0,1]} \left( \frac{1}{2} - \frac{(\sigma_{0,2}^2 - \sigma_{0,1}^2)\sigma^2}{2n\sigma_{0,1}^2\sigma_{0,2}^2} \right), \quad \text{and} \quad \omega_2^{\text{opt}} = 1 - \omega_1^{\text{opt}},$$

where  $\Pi_{[a,b]}(\cdot)$  denotes truncation into  $[a, b]$ . First, if the prior variances are equal ( $\sigma_{0,1}^2 = \sigma_{0,2}^2$ ), allocating an equal number of samples to each arm is optimal. Conversely, if  $\sigma_{0,1}^2 \ll \sigma_{0,2}^2$  and the budget  $n$  is small,

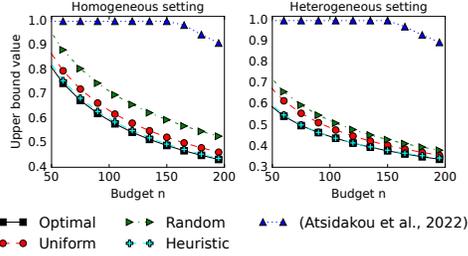


Figure 1: Bound of PI-BAI with different weights compared to BayesElim (Atsidakou et al., 2022).

we get  $\omega_2^{\text{opt}} \gg \omega_1^{\text{opt}}$ , allocating most of the budget to the arm with higher prior variance (lower initial confidence). This is only relevant for small budgets since when  $n \rightarrow +\infty$ , the optimal strategy converges to uniform allocation as the influence of the prior vanishes asymptotically. On the other hand, a saturation phenomenon happens for ‘too small’ budgets: if  $n < 2\sigma^2 \left| \frac{\sigma_{0,2}^2}{\sigma_{0,1}^2} - \frac{\sigma_{0,1}^2}{\sigma_{0,2}^2} \right|$ , the weight of the arm with larger prior variance is equal to 1 due to the truncation. Finally, this allocation strategy does not depend on the prior gap  $\Delta_0$ , which is expected as identifying the optimal arm is equivalent to identifying the worst arm in the case of two arms. This equivalence does not hold when  $K > 2$ .

**Allocation by optimal design.** G-optimal design can be applied for MAB, by using  $x_i = e_i$  for any  $i \in [K]$  and  $\Sigma_0 = \text{diag}(\sigma_{0,i}^2)_{i \in [K]}$ .

**Allocation by warm-up.** Thompson Sampling is well-understood for MAB (Russo et al., 2018) even beyond Gaussian priors so this allocation strategy is fairly versatile, though not covered by our theory since the resulting weights are instance-dependent.

## 4.2 Linear Bandits

In linear bandits (Abbasi-Yadkori et al., 2011), arms share a common low-dimensional representation through the parameter  $\theta \in \mathbb{R}^d$ . We denote  $\mathcal{X} = \{x_1, \dots, x_K\}$  the set of arms where each  $x_i \in \mathbb{R}^d$ . In the Gaussian case, the reward distribution is parametrized by  $\theta$ ,

$$\begin{aligned} \theta &\sim \mathcal{N}(\mu_0, \Sigma_0) \\ Y_t | \theta, A_t &\sim \mathcal{N}(\theta^\top x_{A_t}, \sigma^2) \quad \forall t \in [n], \end{aligned} \quad (7)$$

for some prior mean  $\mu_0 \in \mathbb{R}^d$  and variance  $\Sigma_0 \in \mathbb{R}^{d \times d}$ . Similarly to (5), this model offers closed-form formulas (Agrawal and Goyal, 2013), where the posterior of  $\theta$  given the history  $H_n$  containing  $n_i$  samples from arm

$i$  is a Gaussian  $\mathcal{N}(\hat{\mu}_n, \hat{\Sigma}_n)$ :

$$\begin{aligned} \hat{\Sigma}_n^{-1} &= \Sigma_0^{-1} + \sigma^{-2} \sum_{i \in [K]} n_i x_i x_i^\top \\ \hat{\mu}_n &= \hat{\Sigma}_n (\Sigma_0^{-1} \mu_0 + \sigma^{-2} B_n), \end{aligned} \quad (8)$$

where  $B_n = \sum_{t \in [n]} Y_t x_{A_t}$ . The mean posterior reward of arm  $i$  is given by  $\mathbb{E}[r(i; \theta) | H_n] = \hat{\mu}_n^\top x_i$ . Note that the MAB (5) can be recovered from (7) when  $x_i = e_i \in \mathbb{R}^K$  and  $\Sigma_0 = \text{diag}(\sigma_{0,i}^2)_{i \in [K]}$ .

**PoE bound for linear bandits.** Importantly, our error analysis extends to structured bandit problems. We begin with linear bandits in the next theorem and discuss the specific case of hierarchical models further below.

**Theorem 4.2** (Upper bound for linear bandits). *Assume that  $x_i \neq x_j$  for any  $i \neq j$ , and that  $\mathcal{X}$  spans  $\mathbb{R}^d$ . Then, for all  $\omega \in \Delta_K^+$ , the expected PoE of PI-BAI using allocation  $\omega$  under the linear bandit problem (7) is upper bounded as*

$$\mathcal{P}_n \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{1}{\sqrt{1 + \frac{nc_{i,j}(\omega)}{\|x_i - x_j\|_{\Sigma_0}^2}}} \exp\left(-\frac{(\mu_0^\top x_i - \mu_0^\top x_j)^2}{2\|x_i - x_j\|_{\Sigma_0}^2}\right),$$

where  $c_{i,j}(\omega)$  is given in (26), and  $\tilde{\Sigma}_n = (\frac{1}{n}\Sigma_0^{-1} + \sigma^{-2} \sum_{i \in [K]} w_i x_i x_i^\top)^{-1}$  is deterministic.

In particular,  $\mathcal{P}_n = \mathcal{O}(1/\sqrt{n})$  and we recover Theorem 4.1 when  $x_i = e_i \in \mathbb{R}^K$  and  $\Sigma_0 = \text{diag}(\sigma_{0,i}^2)_{i \in [K]}$ . Full expressions and proofs are given in Appendix C.3. Also, this bound captures the benefit of using informative priors since  $\mathcal{P}_n \rightarrow 0$  when the prior variances are small, that is,  $\Sigma_0 \rightarrow 0_{L \times L}$ , or when the prior gaps are large, in the sense that  $|\mu_0^\top x_i - \mu_0^\top x_j| \rightarrow \infty$ .

**Allocation by optimization.** Similarly to the MAB setting, the bound in Theorem 4.2 can be numerically optimized. This is a non-convex problem so there is no guarantee to reach a global optimum,<sup>3</sup> and, in practice, initialization plays an important role.

**Allocation by warm-up.** We use Linear Thompson Sampling (e.g., Abeille and Lazaric, 2017) with the given informed prior  $\mathcal{N}(\mu_0, \Sigma_0)$ .

**G-optimal design allocations.** In this setting, the most appropriate allocation strategy is by Bayesian G-optimal design (4). We specifically compute our upper bound with this choice below.

**Corollary 4.3** (Upper bound of PI-BAI(G-opt)). *Under the assumptions of Theorem 4.2, the expected PoE of PI-BAI(G-opt) in problem (7) with a diagonal covariance matrix  $\Sigma_0$  is upper bounded as*

$$\mathcal{P}_n \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{1}{\sqrt{1 + \frac{n}{2d\sigma^2} c_{i,j}}} \exp\left(-\frac{(\mu_0^\top x_i - \mu_0^\top x_j)^2}{2\|x_i - x_j\|_{\Sigma_0}^2}\right),$$

<sup>3</sup>Our code is provided for reproducibility purposes.

where  $c_{i,j} = \Omega(1)$ . The full expressions and proofs are given in Appendix C.3.

### 4.3 Hierarchical Bandits

Hierarchical (or *mixed-effect*) models are rich models commonly used in practice where actions are typically structured into clusters, or *effects* (Aouali et al., 2023; Bishop, 2006; Hong et al., 2022b; Wainwright et al., 2008). The correlations are then captured as

$$\begin{aligned} \mu &\sim \mathcal{N}(\nu, \Sigma) \\ \theta_i &\sim \mathcal{N}(b_i^\top \mu, \sigma_{0,i}^2) \quad \forall i \in [K] \\ Y_t | \mu, \theta, A_t &\sim \mathcal{N}(\theta_{A_t}, \sigma^2) \quad \forall t \in [n] \end{aligned} \quad (9)$$

where  $\mu = (\mu_\ell)_{\ell \in [L]}$  is an unknown latent parameter composed of  $L$  effects  $\mu_\ell$  and it is sampled from a multivariate Gaussian with mean  $\nu \in \mathbb{R}^L$  and covariance  $\Sigma \in \mathbb{R}^{L \times L}$ . Then, given  $\mu$ , the mean rewards  $\theta_i$  are independently sampled as  $\theta_i \sim \mathcal{N}(b_i^\top \mu, \sigma_{0,i}^2)$ , where  $(b_i)_{i \in [K]}$  represent *known mixing weights*. In particular,  $b_i^\top \mu$  creates a linear mixture of the  $L$  effects, with  $b_{i,\ell}$  indicating a known score that quantifies the association between arm  $i$  and the effect  $\ell$ . Concrete examples of  $b_{i,\ell}$  are provided in Appendix B.1. Note that arm correlations arise because  $\theta_i$  are derived from the same effect parameter  $\mu$ . Finally, given  $\mu$  and  $\theta$ , the reward distribution of arm  $i$  is similar to the MAB (5) and only depends on  $\theta_i$  as  $\mathcal{N}(\theta_i, \sigma^2)$ .

Similarly to the linear case, an explicit marginal posterior distribution  $\mathbb{P}(\theta_i | H_n) = \mathcal{N}(\hat{\mu}_{n,i}, \hat{\sigma}_{n,i}^2)$  can be computed for closed-form posterior mean and variance parameters (see Appendix C.1).

**PoE bound for hierarchical bandits.** We derive a bound on the PoE that takes advantage of the hierarchical structure (9). To the best of our knowledge, these are the first *prior-dependent* bounds for fixed-budget Bayesian BAI in these settings.

**Theorem 4.4** (Upper bound for hierarchical bandits). *For all  $\omega \in \Delta_K^+$ , the expected PoE of PI-BAI using allocation  $\omega$  under the hierarchical bandit problem (9) is upper bounded as*

$$\mathcal{P}_n \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{1}{\sqrt{1 + \frac{c_{i,j}(\omega)}{\hat{\sigma}_{n,i} + \hat{\sigma}_{n,j}}}} \exp\left(\frac{-(\nu^\top b_i - \nu^\top b_j)^2}{2(\|b_i - b_j\|_\Sigma^2 + \sigma_{0,i}^2 + \sigma_{0,j}^2)}\right),$$

where  $c_{i,j}(\omega)$  is given in (30) and  $\hat{\sigma}_{n,i}$  in (17).  $c_{i,j} = \Omega(1)$  and  $\hat{\sigma}_{n,i}^2 = \Omega(1/n)$ , and they depend on both the prior parameters and allocation weights. In particular,  $\mathcal{P}_n = \mathcal{O}(1/\sqrt{n})$ . Full expressions and proofs are given in Appendix C.4.

The term  $\|b_i - b_j\|_\Sigma^2 + \sigma_{0,i}^2 + \sigma_{0,j}^2$  accounts for the prior uncertainty of both arms and effects. If the effects are

deterministic ( $\Sigma = 0_{L \times L}$ ) then our bound recovers the upper bound of MAB with prior mean  $\mu_{0,i} = \nu^\top b_i$ . On the other hand, if the arms are deterministic given the effects ( $\sigma_{0,i}^2 = 0$ ), the bound only depends on the effect covariance. Finally, if the prior is informative by its gap ( $|\nu^\top b_i - \nu^\top b_j| \rightarrow \infty$ ) or by its variance ( $\Sigma \rightarrow 0_{L \times L}$  and  $\sigma_{0,i}^2 \rightarrow 0$ ), then  $\mathcal{P}_n \rightarrow 0$ .

**Allocation by optimization.** Again, numerical optimization can be leveraged for the upper bound in Theorem 4.4, but in our experiments (see Section 6) we see that this method reaches its limits due to the complexity of the bound to be optimized.

**Allocation by optimal design.** (9) is a special case of a linear bandit (7), by realizing that  $\theta_i = b_i^\top \mu + \eta_i$  where  $\eta_i \sim \mathcal{N}(0, \sigma_{0,i}^2)$  and the  $\eta_i$  are independent of  $\mu$ . Hence, (9) can be rewritten by replacing  $\nu$  with  $\bar{\nu} \in \mathbb{R}^{L+K}$  defined as  $\bar{\nu}^\top = (\nu^\top, 0, \dots, 0)$  and  $\Sigma$  with a block-diagonal matrix  $\bar{\Sigma} = \langle \Sigma, \sigma_{0,1}^2, \dots, \sigma_{0,K}^2 \rangle$  with actions  $\bar{b}_i \in \mathbb{R}^{L+K}$  defined as  $\bar{b}_i^\top = (b_i^\top, e_i^\top)$ , leading to a linear bandit

$$\begin{aligned} \bar{\mu} &\sim \mathcal{N}(\bar{\nu}, \bar{\Sigma}) \\ Y_t | \bar{\mu}, A_t &\sim \mathcal{N}(\bar{b}_{A_t}^\top \bar{\mu}, \sigma^2) \quad \forall t \in [n]. \end{aligned} \quad (10)$$

The only practical benefit of this point of view is to allow us to compute G-optimal design allocations (4) and inherit the theoretical guaranties from the linear case. In general, adhering to the initial formulation in (9) and the subsequent derivations in (17) is more computationally efficient and this motivates the concept of hierarchical bandits in the first place. Additional discussions can be found in Appendix B.2.

**Allocation by warm-up.** A mixed-effect Thompson Sampling algorithm has been proposed recently by Aouali et al. (2023), so we can directly use this algorithm as warm up with our given prior parameters.

### 4.4 Robustness to Prior Misspecification

While the assumption of knowing the prior distribution is common in the Bayesian bandit literature (Russo et al., 2018; Kveton et al., 2021), we acknowledge that it may not always be the case in practice. Lemma 4.5 quantifies the effect of prior misspecification when using the uniform allocation strategy yielding  $\omega = (\frac{1}{K})_{i \in [K]}$ , and when the prior variances are homogeneous. Its proof is provided in Appendix C.5.

**Lemma 4.5** (Upper bound for PI-BAI with misspecified prior parameters). *The expected PoE of PI-BAI(Uni) under the MAB problem (5) with mis-*

specified priors  $\mathcal{N}(\tilde{\mu}_{0,i}, \tilde{\sigma}_0^2)$ <sup>4</sup> is upper bounded as

$$\mathcal{P}_n \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{1}{\sqrt{1 + \frac{n\sigma_0^2}{K\sigma^2}}} \exp\left(-\frac{(\mu_{0,i} - \mu_{0,j})^2}{4\sigma_0^2}\right) d_n(i, j),$$

where

$$d_n(i, j) = \exp\left(\frac{\text{KL}(P_{ij}, \tilde{P}_{ij})}{1 + \frac{n\sigma_0^2}{K\sigma^2}}\right),$$

with  $P_{ij} = \mathcal{N}(\mu_{0,i} - \mu_{0,j}, 2\sigma_0^2)$  and  $\tilde{P}_{ij} = \mathcal{N}(\frac{\tilde{\sigma}_0^2}{\sigma_0^2}(\tilde{\mu}_{0,i} - \tilde{\mu}_{0,j}), 2\sigma_0^2)$  are respectively the true and the misspecified prior gap distributions. Note that  $\text{KL}(P_{ij}, \tilde{P}_{ij}) = \frac{\sigma_0^2}{\tilde{\sigma}_0^4} (\frac{\tilde{\sigma}_0^2}{\sigma_0^2}(\mu_{0,i} - \mu_{0,j}) - (\tilde{\mu}_{0,i} - \tilde{\mu}_{0,j}))^2$ .

In case of well specified variances, the misspecification factor  $d_n(i, j)$  depends only on the true and assumed prior gaps  $(\mu_{0,i} - \mu_{0,j})$  and  $(\tilde{\mu}_{0,i} - \tilde{\mu}_{0,j})$ . If the assumed prior preserves the prior gaps, then  $d_n(i, j) = 1$ , and Theorem 4.1 is recovered. However, if the prior gaps are misspecified, then  $(d_n(i, j) > 1)$ , which increases the bound obtained in Theorem 4.1. This multiplicative factor does not affect the overall convergence rate of  $\mathcal{O}(1/\sqrt{n})$ , and  $d_n = e^{\mathcal{O}(1/n)}$ , indicating that the impact of prior misspecification becomes less significant as the budget increases. This is also observed empirically (see Appendix D.1). That said, we give two solutions to circumvent the assumption of prior knowledge.

**Offline learning of priors.** Fixed-budget BAI can be significantly enhanced by incorporating readily available offline data as follows. We first estimate arm means using offline data (potentially noisy). Then, we model the uncertainty in these estimates with Gaussian covariance, creating a Gaussian prior for multi-armed and linear bandits. For hierarchical bandits, we fit a Gaussian Mixture Model (GMM) to the offline estimates, resulting in the desired hierarchical prior. For instance, in our MovieLens experiments (Section 6), we derived the offline estimates through low-rank matrix factorization of the user-item ratings.

**Online learning of priors using hierarchical models.** In the absence of offline data, one approach to address prior misspecification online is to use hierarchical models. For example, consider the MAB problem (5), where the prior means  $\mu_0$  are now unknown. Then we adopt the hierarchical model (9), assuming that  $\mu_0$  is sampled from a known hyper-prior as  $\mu_0 \sim \mathcal{N}(\nu, \Sigma)$ , with  $L = K$  and canonical mixing weights  $b_i = e_i \in \mathbb{R}^K$ . This approach to learning the prior online has good empirical performance (Section 6) and it can be extended to linear and hierarchical bandits.

<sup>4</sup>Formally, this means that the mean posterior reward for arm  $i$  is calculated using the misspecified prior  $\mathcal{N}(\tilde{\mu}_{0,i}, \tilde{\sigma}_0^2)$  instead of the true prior  $\mathcal{N}(\mu_{0,i}, \sigma_0^2)$ .

## 5 GENERAL PROOF SCHEME

We outline the key technical insights to derive our Bayesian proofs. The idea is general and can be applied to all our settings (MAB, linear and hierarchical bandits). Specific proofs for these settings are provided in Appendix C.

**From Frequentist to Bayesian proof.** To analyze their algorithm in the MAB setting, [Atsidakou et al. \(2022\)](#) rely on the strong restriction that  $\sigma_{0,i} = \sigma_0$  for all arms  $i \in [K]$  and tune their allocations as a function of the noise variance  $\sigma^2$  such that in the Gaussian setting, the posterior variances  $\hat{\sigma}_{n,i}^2$  in (6) are equal for all arms  $i \in [K]$ . This assumption is needed to directly leverage results from [Karnin et al. \(2013\)](#), thus allowing them to bound the frequentist PoE as  $\mathbb{P}(J_n \neq i_*(\theta) \mid \theta) \leq B(\theta)$  for a fixed instance  $\theta$ . Then, the expected PoE,  $\mathcal{P}_n = \mathbb{E}[\mathbb{P}(J_n \neq i_*(\theta) \mid \theta)]$ , is bounded by directly computing  $\mathbb{E}_{\theta \sim P_0}[B(\theta)]$ . We believe it is not possible to extend such technique for general choices of allocations  $n_i$  and prior variances  $\sigma_{0,i}^2$ . Thus, we pursue an alternative approach, establishing the result in a fully Bayesian fashion. We start with a key observation.

**Key reformulation of the expected PoE.** We observe that the expected PoE can be reformulated as follows

$$\mathcal{P}_n = \mathbb{E}[\mathbb{P}(J_n \neq i_*(\theta) \mid \theta)] = \mathbb{E}[\mathbb{P}(J_n \neq i_*(\theta) \mid H_n)].$$

This swap of measures means that to bound  $\mathcal{P}_n$ , we no longer bound the probability of

$$J_n = \underset{i \in [K]}{\text{argmax}} \mathbb{E}[r(i; \theta) \mid H_n] \neq \underset{i \in [K]}{\text{argmax}} r(i; \theta) = i_*(\theta)$$

for any fixed  $\theta$ . Instead, we only need to bound that probability when  $\theta$  is drawn from the posterior distribution. Precisely, we bound the probability that the arm  $i$  maximizing the posterior mean  $\mathbb{E}[r(i; \theta) \mid H_n]$  differs from the arm  $i$  maximizing the posterior sample  $r(i; \theta) \mid H_n$ . This is achieved by first noticing that  $\mathcal{P}_n$  can be rewritten as (see proof in Appendix C.2)

$$\mathcal{P}_n = \sum_{\substack{i,j \in [K] \\ i \neq j}} \mathbb{E}[\mathbb{P}(i_*(\theta) = i \mid J_n = j, H_n) \mathbb{1}\{J_n = j\}].$$

The rest of the proof consists in bounding the above conditional probabilities for distinct  $i$  and  $j$ , and this depends on the setting (MAB, linear or hierarchical). Roughly speaking, this is achieved as follows.  $\mathbb{P}(i_*(\theta) = i \mid H_n, J_n = j)$  is the probability that arm  $i$  maximizes the posterior sample  $r(\cdot; \theta)$ , given that arm  $j$  maximizes the posterior mean  $\mathbb{E}[r(\cdot; \theta) \mid H_n]$ . We show that this probability decays exponentially with the squared difference  $(\mathbb{E}[r(i; \theta) \mid H_n] - \mathbb{E}[r(j; \theta) \mid H_n])^2$ .

Taking the expectation of this term under the history  $H_n$  gives the desired  $\mathcal{O}(1/\sqrt{n})$  rate. This proof introduces a novel perspective for Bayesian BAI, distinguished by its tighter prior-dependent bounds on the expected PoE and ease of extension to more complex settings like linear and hierarchical bandits. However, its application to adaptive algorithms could be challenging, particularly due to the complexity of taking expectations under the history  $H_n$  in that case. Also, extending this proof to non-Gaussian distributions is an interesting direction for future work, since in general the posterior variances depend on sample observations.

**Connections with Bayesian Decision Theory.** Note that the ‘key reformulation’ above is inspired from a classical result in Bayesian Decision Theory (Robert et al., 2007) that we exploit in a different manner to analyse the Bayes risk of our (non-optimal but computable) decision rule.

## 6 EXPERIMENTS

In all experiments, we set the observation noise to  $\sigma = 1$  and run algorithms  $10^4$  times. We display the expected PoE along with its (narrow) standard error. Additional experiments and details are provided in Appendix D. We consider four allocations for PI-BAI (discussed in Section 3): PI-BAI(Uni) (uniform weights), PI-BAI(Opt) (optimized weights), PI-BAI(G-opt) (G-optimal design) and PI-BAI(TS) (warmed-up with Thompson Sampling) with warm-up length  $n_w = K$ . The tuning of  $n_w$  and the choice of the warm-up policy are discussed in Appendix D.7. The code is available in this [Github repository](#).

**MAB.** We set  $K = 10$ ,  $\mu_{0,i} \sim \mathcal{U}([0, 1])$  and  $\sigma_{0,i}$  are evenly spaced between 0.1 and 0.5. We compare PI-BAI to Bayesian methods: top-two Thompson sampling (TTTS<sup>5</sup>) (Jourdan et al., 2022; Russo, 2016) and BayesElim (Atsidakou et al., 2022), and to frequentist ones: Successive Rejects (SR) (Audibert et al., 2010) and Sequential Halving (SH) (Karnin et al., 2013).

**Linear bandit.** We set  $d = 4$ ,  $K = 30$ ,  $\mu_{0,i} \sim \mathcal{U}([0, 1])$  and  $\Sigma_0 = \text{diag}(\sigma_{0,i}^2)_{i \in [K]}$ , with  $\sigma_{0,i}$  are evenly spaced between 0.1 and 0.5. We compare our methods with two algorithms designed for linear bandits: BayesGap (Hoffman et al., 2014) and GSE (Azizi et al., 2021). These represent the current (tractable<sup>6</sup>) state-of-the-art approaches, leveraging G-optimal design for successive elimination. We did not include other methods that use the same elimination approach but have lower performance in these settings (Alieva et al., 2021; Yang

<sup>5</sup>TTTS does not come with theoretical guarantees in the fixed-budget setting.

<sup>6</sup>Katz-Samuels et al. (2020) has an algorithm with tighter bounds but it is not tractable.

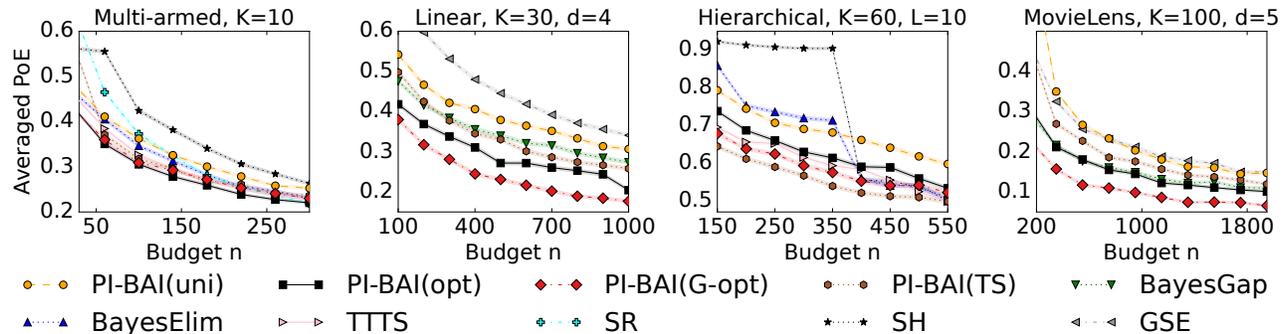
and Tan, 2022).

**Hierarchical bandit.** We set  $K = 60$  and  $L = 10$ , with mixing weights  $b_i$  sampled uniformly in  $[0, 1]$  and normalized to sum to 1. We set  $\nu_i \sim \mathcal{U}([-1, 1])$ , and  $\Sigma$  and  $\Sigma_0$  are diagonal with entries evenly spaced in  $[0.1^2, 0.5^2]$ . The MAB frequentist baselines do not require priors, while the Bayesian ones need a Gaussian MAB prior (5) since they are not suitable for a hierarchical prior. To obtain such a Gaussian MAB prior, we marginalize the effect parameters in (9). Also, although there is a connection to linear bandits, we do not include those baselines as they do not perform well due to their inefficient representation of the structure, resulting in a high-dimensional  $(K + L)$  linear model.

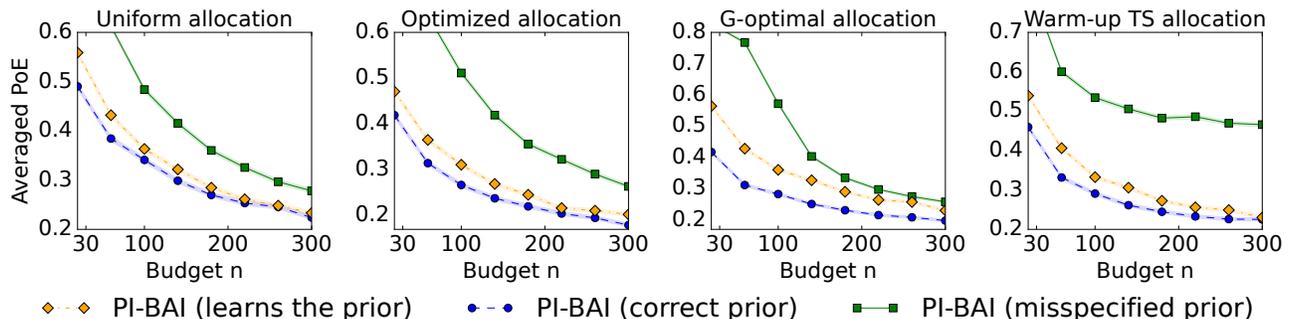
**Results on synthetic data (first three columns of Figure 2a).** Overall, variants of PI-BAI outperform the baselines. The best variant depends on the specific setting, but PI-BAI(G-opt) is the best overall, closely followed by PI-BAI(Opt). In hierarchical experiments, PI-BAI(TS) exceeds all baselines, highlighting its effectiveness in capturing the underlying structure. We emphasize that all Bayesian baselines in our experiments (*e.g.*, BayesElim and TTTS) are implemented using the same informative prior. Overall, these results reaffirm that leveraging prior information is a powerful and practical tool to scale BAI to cases with a large number of arms and limited data. Additional experiments in various settings are in Appendix D.2.

**MovieLens (fourth column of Figure 2a).** We also consider the real-world MovieLens dataset (Lam and Herlocker, 2016), which contains 1 million ratings from 6,040 users for 3,883 movies, forming a sparse rating matrix  $M$  of size  $6,040 \times 3,883$ . To learn a suitable prior for our algorithm, we complete  $M$  using alternating least squares with rank  $d = 5$  (low-rank factorization), resulting in the decomposition  $M = U^\top V$ , where rows  $U_i$  and  $V_j$  represent user  $i$  and movie  $j$ , respectively. We then construct a linear Gaussian prior (Section 4.2) by setting the mean  $\mu_0 = \frac{1}{3883} \sum_{j \in [3883]} V_j$  and the covariance  $\Sigma_0 = \text{diag}(v)$ , where  $v \in \mathbb{R}^d$  is the empirical variance of  $V_j$  for  $j \in [3883]$  along each dimension. This prior is fixed across all MovieLens experiments and does not depend on the rewards generated by the bandit instance, resulting in a prior that is not data-dependent. We run  $10^4$  simulations with  $K = 100$  randomly subsampled movies. In this experiment, the bandit instances are not sampled from a Gaussian prior, yet our algorithm employs a learned Gaussian prior, highlighting the robustness of our method to prior misspecification. Despite this mismatch, PI-BAI with this prior performs very well, as seen in the fourth column of Figure 2a.

**Online learning of priors (Figure 2b).** We con-



(a) Comparison of PI-BAI instantiated with different allocation weights to the baselines BayesElim, TTTS, SR, SH in multi-armed and hierarchical bandits, and to BayesGap and GSE in linear bandits and MovieLens.



(b) For a given allocation strategy, we compare PI-BAI with learned vs. correct vs. misspecified prior.

Figure 2: Average PoE with varying  $n$ : comparison to baselines and impact of prior misspecification.

sider the same MAB setting as before with  $K = 10$ ,  $\mu_{0,i} \sim \mathcal{U}([0, 1])$ , and  $\sigma_{0,i}$  evenly spaced in  $[0.1, 0.5]$ . We compare three variants of PI-BAI: First, PI-BAI using the correct prior mean  $\mu_{0,i}$ . Second, PI-BAI with a misspecified prior  $\tilde{\mu}_0 = \mu_0 + 0.5\epsilon$  where  $\epsilon \sim \mathcal{N}(0, 1)$ . Third, PI-BAI that learns the prior online through the hierarchical model with  $b_i = e_i \in \mathbb{R}^K$ ,  $\nu_i \sim \mathcal{U}([0, 1])$ , and  $\Sigma = I_K$ , as described in the last paragraph of Section 4.4. This comparison is conducted when PI-BAI variants are instantiated with either Uniform, Optimized, G-optimal or TS allocation weights. As expected, prior misspecification affects PI-BAI’s performance, but this effect diminishes as the budget  $n$  grows, validating our theory. Additionally, PI-BAI in the hierarchical model effectively learns the correct prior  $\mu_0$ , with performance converging to that of PI-BAI with the correct prior  $\mu_0$ . Finally, PI-BAI with theoretically grounded allocations (Uniform, Optimized, G-optimal) is more robust to prior misspecification than PI-BAI with TS warm-up, likely because the misspecified prior is used twice: by the warm-up policy TS and by PI-BAI.

## 7 CONCLUSION

We revisit the Bayesian fixed-budget BAI for PoE minimization and propose a simple yet effective algorithm for multi-armed, linear, and hierarchical bandits.

Our novel proof technique provides a generic, prior-dependent upper bound on the expected PoE that is tighter than existing bounds for MAB. It presents the first bound on the expected PoE for linear and hierarchical bandits. Finally, we address prior misspecification through both theoretical and empirical analysis. Our work illuminates the trade-off between adaptivity and generality in BAI algorithm design and opens several avenues for future research; the competitiveness of this non-adaptive algorithm with elimination schemes is of broader interest, especially as it matches known lower bounds. Theoretically, it would be valuable to explore prior-dependent lower bounds beyond the simple  $K = 2$  case.

## ACKNOWLEDGEMENTS

N. Nguyen and C. Vernade are funded by the Deutsche Forschungsgemeinschaft (DFG) under both the project 468806714 of the Emmy Noether Programme and under Germany’s Excellence Strategy – EXC number 2064/1 – Project number 390727645. Both also thank the international Max Planck Research School for Intelligent Systems (IMPRS-IS).

## References

- Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320, 2011.
- Marc Abeille and Alessandro Lazaric. Linear Thompson sampling revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.
- Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pages 127–135, 2013.
- Ayya Alieva, Ashok Cutkosky, and Abhimanyu Das. Robust pure exploration in linear bandits with limited budget. In *International Conference on Machine Learning*, pages 187–195. PMLR, 2021.
- Imad Aouali. Diffusion models meet contextual bandits with large action spaces. *arXiv preprint arXiv:2402.10028*, 2024.
- Imad Aouali, Branislav Kveton, and Sumeet Katariya. Mixed-effect thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 2087–2115. PMLR, 2023.
- Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Bayesian off-policy evaluation and learning for large action spaces. *arXiv preprint arXiv:2402.14664*, 2024.
- Alexia Atsidakou, Sumeet Katariya, Sujay Sanghavi, and Branislav Kveton. Bayesian fixed-budget best-arm identification. *arXiv preprint arXiv:2211.08572*, 2022.
- Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.
- Javad Azizi, Branislav Kveton, Mohammad Ghavamzadeh, and Sumeet Katariya. Meta-learning for simple regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6709–6717, 2023.
- Mohammad Javad Azizi, Branislav Kveton, and Mohammad Ghavamzadeh. Fixed-budget best-arm identification in structured bandits. *arXiv preprint arXiv:2106.04763*, 2021.
- Soumya Basu, Branislav Kveton, Manzil Zaheer, and Csaba Szepesvari. No regrets for learning the prior in bandits. In *Advances in Neural Information Processing Systems 34*, 2021.
- José M Bernardo and Raúl Rueda. Bayesian hypothesis testing: A reference approach. *International Statistical Review*, 70(3):351–372, 2002.
- Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4 of *Information science and statistics*. Springer, 2006.
- Sébastien Bubeck, Rémi Munos, and Gilles 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.
- Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR, 2016.
- Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*, pages 2249–2257, 2012.
- Pierre Clavier, Tom Huix, and Alain Durmus. Vits: Variational inference thomson sampling for contextual bandits. *arXiv preprint arXiv:2307.10167*, 2023.
- Rémy Degenne. On the existence of a complexity in fixed budget bandit identification. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1131–1154. PMLR, 2023.
- Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.
- Sarah Filippi, Olivier Cappé, Aurelien Garivier, and Csaba Szepesvari. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pages 586–594, 2010.
- Matthew Hoffman, Bobak Shahriari, and Nando Freitas. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*, pages 365–374. PMLR, 2014.
- Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, and Craig Boutilier. Latent bandits revisited. In *Advances in Neural Information Processing Systems 33*, 2020.
- Joey Hong, Branislav Kveton, Sumeet Katariya, Manzil Zaheer, and Mohammad Ghavamzadeh. Deep hierarchy in bandits. *arXiv preprint arXiv:2202.01454*, 2022a.
- Joey Hong, Branislav Kveton, Manzil Zaheer, and Mohammad Ghavamzadeh. Hierarchical Bayesian bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 2022b.
- Joey Hong, Branislav Kveton, Manzil Zaheer, Sumeet Katariya, and Mohammad Ghavamzadeh. Multi-task

- off-policy learning from bandit feedback. In *International Conference on Machine Learning*, pages 13157–13173. PMLR, 2023.
- Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne de Heide, and Emilie Kaufmann. Top two algorithms revisited. *Advances in Neural Information Processing Systems*, 35:26791–26803, 2022.
- Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International conference on machine learning*, pages 1238–1246. PMLR, 2013.
- Julian Katz-Samuels, Lalit Jain, Kevin G Jamieson, et al. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *Advances in Neural Information Processing Systems*, 33:10371–10382, 2020.
- Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016.
- Junpei Komiyama, Kaito Ariu, Masahiro Kato, and Chao Qin. Rate-optimal bayesian simple regret in best arm identification. *Mathematics of Operations Research*, 2023.
- Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020.
- Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-Wei Hsu, Martin Mladenov, Craig Boutilier, and Csaba Szepesvari. Meta-Thompson sampling. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.
- Shyong Lam and Jon Herlocker. MovieLens Dataset. <http://grouplens.org/datasets/movielens/>, 2016.
- Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- Jongyeong Lee, Junya Honda, and Masashi Sugiyama. Thompson exploration with best challenger rule in best arm identification. In *Asian Conference on Machine Learning*, pages 646–661. PMLR, 2024.
- Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The journal of machine learning research*, 18(1):6765–6816, 2017.
- Jesús López-Fidalgo. *Optimal Experimental Design: A Concise Introduction for Researchers*, volume 226. Springer Nature, 2023.
- Nicolas Nguyen and Claire Vernade. Lifelong best-arm identification with misspecified priors. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.
- Amit Peleg, Naama Pearl, and Ron Meir. Metalearning linear bandits by prior update. In *International Conference on Artificial Intelligence and Statistics*, pages 2885–2926. PMLR, 2022.
- My Phan, Yasin Abbasi Yadkori, and Justin Domke. Thompson sampling and approximate inference. *Advances in Neural Information Processing Systems*, 32, 2019.
- Chao Qin. Open problem: Optimal best arm identification with fixed-budget. In *Conference on Learning Theory*, pages 5650–5654. PMLR, 2022.
- Christian P Robert et al. *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer, 2007.
- Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418. PMLR, 2016.
- Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.
- David J Spiegelhalter and Steffen L Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20(5):579–605, 1990.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.
- Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- Po-An Wang, Kaito Ariu, and Alexandre Proutiere. On universally optimal algorithms for a/b testing. In *International Conference on Machine Learning, 21-27 July 2024, Vienna, Austria*, volume 235, pages 50065–50091. International Conference on Machine Learning, 2024.

- R. Wolke and H. Schwetlick. Iteratively reweighted least squares: Algorithms, convergence analysis, and numerical comparisons. *SIAM Journal on Scientific and Statistical Computing*, 9(5):907–921, 1988.
- Junwen Yang and Vincent Tan. Minimax optimal fixed-budget best arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 35:12253–12266, 2022.
- Le Yang, Siyang Gao, and Chin Pang Ho. Improving the knowledge gradient algorithm. *Advances in Neural Information Processing Systems*, 36, 2024.

## Checklist

1. For all models and algorithms presented, check if you include:
  - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
  - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
  - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes
2. For any theoretical claim, check if you include:
  - (a) Statements of the full set of assumptions of all theoretical results. Yes
  - (b) Complete proofs of all theoretical results. Yes
  - (c) Clear explanations of any assumptions. Yes
3. For all figures and tables that present empirical results, check if you include:
  - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
  - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
  - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
  - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
  - (a) Citations of the creator If your work uses existing assets. Yes
  - (b) The license information of the assets, if applicable. Not Applicable
  - (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
  - (d) Information about consent from data providers/curators. Not Applicable
  - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
  - (a) The full text of instructions given to participants and screenshots. Not Applicable
  - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
  - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

## ORGANIZATION OF THE APPENDIX

The supplementary material is organized as follows. In Appendix A, we mention additional existing works relevant to our work, and discuss in depth the differences between our work and them. In Appendix B, we provide additional general additional remarks. In Appendix C, we give complete statements and proofs of our theoretical results. In Appendix D, we supply additional numerical experiments.

## A EXTENDED RELATED WORK

**Hierarchical Bayesian bandits.** Bayesian bandits algorithms under hierarchical models have been heavily studied in recent years (Aouali, 2024; Aouali et al., 2023; Basu et al., 2021; Hong et al., 2020, 2022a,b; Kveton et al., 2021; Nguyen and Vernade, 2023; Peleg et al., 2022). All the aforementioned papers propose methods to explore efficiently in hierarchical bandit environments to minimize the (Bayesian) regret. Roughly speaking, the idea of all hierarchical Bayesian bandits is to take advantage of the correlations between arms (learned through the posterior covariance) to have more information of the arms that are not pulled at each round. Beyond regret minimization, Bayesian structured models also found success in simple regret minimization (Azizi et al., 2023) and off-policy learning in contextual bandits (Aouali et al., 2024; Hong et al., 2023).

**Bayesian simple regret minimization and differences with Komiyama et al. (2023).** Bayesian simple regret (SR) minimization have been studied in Komiyama et al. (2023) for Bernoulli bandits. However, we would like to emphasize that our work is very different from theirs. First, SR and PoE may yield to different rate in the Bayesian setting (Komiyama et al., 2023). While this observation is not true in the frequentist setting (Audibert et al., 2010), Komiyama et al. (2023) proved an asymptotic  $\mathcal{O}(1/n)$  lower bound on the Bayesian SR, while Atsidakou et al. (2022) proved a  $\mathcal{O}(1/\sqrt{n})$  lower bound on the PoE. This is because the relationship between SR and PoE is not clear as in the frequentist setting. In fact, we have for a fixed environment  $\theta \in \mathbb{R}^K$ ,

$$\text{SR}(\theta) \leq \Delta_{\max}(\theta) \mathbb{P}(J_n \neq i_*(\theta) \mid \theta),$$

where  $\Delta_{\max}(\theta) = \max_{i \in [K]} \theta_i - \min_{i \in [K]} \theta_i$ . Integrating with respect to the prior distribution, we bound the expected simple regret as

$$\mathbb{E}[\text{SR}(\theta)] \leq \mathbb{E}[\Delta_{\max}(\theta) \mathbb{P}(J_n \neq i_*(\theta) \mid \theta)],$$

where the right hand side term involves two non-independent quantities. Note that upper bounding  $\Delta_{\max}$  by a pure constant (*e.g.* with high probability for unbounded support reward distributions) would lead to a suboptimal rate of  $\mathcal{O}(1/\sqrt{n})$  for the simple regret.

The second difference between these two works is that ours consider the finite-time regime, while theirs derive asymptotic upper and lower bounds. We strongly believe that finite-time guarantees are more interesting in our setting because in practical scenarios, having access to an informative prior is beneficial in the data-poor regime in order to identify the best arm. Moreover, despite deriving Bayesian guarantees, their algorithm is frequentist in the sense that it does not use the prior distribution.

Finally, their work focus on Bernoulli distributions, while our work focus on Gaussian distributions. We believe that it is not straightforward to extend their method and analysis to distributions with unbounded support as Gaussians.

**Differences on the prior assumption with Atsidakou et al. (2022).** This latter work considers the  $K$ -arms Gaussian setting (5) and restricts the prior to have equal prior variances  $\sigma_{0,i}^2 = \sigma_0^2$  for all  $i \in [K]$ . Since their algorithm (BayesElim) is elimination-based and operates in  $R$  rounds, at each elimination round  $r \in [R]$ , they perform uniform allocation,  $n_i = \frac{n}{KR}$ . Coupling this with the assumption of prior variance yields to equal posterior covariance,  $\sigma_{n,i}^2 = \sigma_n^2$  for all  $i \in [K]$ , *i.e.* arms have the same posterior uncertainty at the end of each round. This observation allows them to use frequentist-based analysis as in Karnin et al. (2013). Note that similarly as in frequentist analysis of Karnin et al. (2013), their analysis imposes to discard all previously collected observations at the end of each round, impacting the empirical performances of their method. We believe that it is not possible to adapt their analysis to heterogeneous prior variances, where each arm would have a different posterior uncertainty at the end of each round.

We now compare the upper bound of BayesElim in the setting where their results are valid, that is, when the prior variances are homogeneous,  $\sigma_{0,i}^2 = \sigma_0^2$ . Their bound and ours read

$$\mathcal{P}_n^{\text{BayesElim}} \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \log_2(K) \frac{e^{-\frac{(\mu_{0,i} - \mu_{0,j})^2}{4\sigma_0^2}}}{\sqrt{1 + \frac{n\sigma_0^2}{K \log_2(K)\sigma^2}}}, \quad \text{and} \quad \mathcal{P}_n^{\text{PI-BAI}} \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{\exp\left(-\frac{(\mu_{0,i} - \mu_{0,j})^2}{4\sigma_0^2}\right)}{\sqrt{1 + \frac{n\sigma_0^2}{K\sigma^2}}}.$$

By omitting elimination phases, we gain roughly a factor  $\log_2^{3/2}(K)$  over the bound of [Atsidakou et al. \(2022\)](#) (highlighted in blue). This additional factor can be significant when bounding a probability ( $\log_2^{3/2}(K) \approx 6$  for  $K = 10$ ). This difference makes our bound smaller even in their setting with homogeneous prior variances and choosing uniform allocation weights for PI-BAI.

## B ADDITIONAL DISCUSSIONS

### B.1 Motivating Practical Examples for Hierarchical Bandits

We discuss motivating examples for using hierarchical models in pure exploration settings.

**Hyper-parameter tuning.** Here, the goal is to find the best configuration for a neural network using  $n$  epochs ([Li et al., 2017](#)). A configuration  $i$  is represented by a scalar  $\theta_i \in \mathbb{R}$  which quantifies the expected performance of a neural network with such configuration. Again, it is intuitive to learn all  $\theta_i$  individually. Roughly speaking, this means running each configuration for  $\lfloor \frac{n}{K} \rfloor$  epochs and selecting the one with the highest performance. This is statistically inefficient since the number of configurations can be combinatorially large. Fortunately, we can leverage the fact that configurations often share the values of many hyper-parameters. Precisely, a configuration is a combination of multiple hyper-parameters, each set to a specific value. Then we represent each hyper-parameter  $\ell \in [L]$  by a scalar parameter  $\mu_\ell \in \mathbb{R}$ , and the configuration parameter  $\theta_i$  is a mixture of its hyper-parameters  $\mu_\ell$  weighted by their values. That is,  $\theta_i = \sum_{\ell \in [L]} b_{i,\ell} \mu_\ell + \epsilon_i$ , where  $b_{i,\ell}$  is the value of hyper-parameter  $\ell$  in configuration  $i$  and  $\epsilon_i$  is a random noise to incorporate uncertainty due to model misspecification.

**Drug design.** In clinical trials,  $K$  drugs are administrated to  $n$  subjects, with the goal of finding the optimal drug design. Each drug is parameterized by its expected efficiency  $\theta_i \in \mathbb{R}$ . As in the previous example, it is intuitive to learn each  $\theta_i$  individually by assigning a drug to  $\lfloor \frac{n}{K} \rfloor$  subjects. However, this is inefficient when  $K$  is large. We leverage the idea that drugs often share the same components: each drug parameter  $\theta_i$  is a combination of component parameters  $\mu_\ell$ , each accounting for a specific dosage. More precisely, the parameter of drug  $i$  can be modeled as  $\theta_i = \sum_{\ell \in [L]} b_{i,\ell} \mu_\ell + \epsilon_i$  where  $\epsilon_i$  accounts for uncertainty due to model misspecification. Similarly to the hyper-parameter tuning example, this models correlations between drugs and it can be leveraged for more efficient use of the whole budget of  $n$  epochs.

### B.2 Additional Remarks on Hierarchical Models

The two-level prior we consider has a shared latent parameter  $\mu = (\mu_\ell)_{\ell \in [L]} \in \mathbb{R}^L$  representing  $L$  effects impacting each of the  $K$  arm means:

$$\begin{aligned} \mu &\sim Q \\ \theta_i &\sim P_{0,i}(\cdot; \mu) && \forall i \in [K] \\ Y_t \mid \mu, \theta, A_t &\sim P(\cdot; \theta_{A_t}) && \forall t \in [n], \end{aligned}$$

where  $Q$  is the latent prior on  $\mu \in \mathbb{R}^L$ .

In the Gaussian setting (9), the maximum likelihood estimate of the reward mean of action  $i$ ,  $B_{n,i}/n_i$ , contributes to (15) proportionally to its precision  $n_i/(n_i\sigma_{0,i}^2 + \sigma^2)$  and is weighted by its mixing weight  $b_i$ . (16) is a standard Gaussian posterior, and (17) takes into account the information of the conditional posterior. Finally, (17) takes into account the arm correlation through its dependence on  $\hat{\sigma}_{n,i}$  and  $\hat{\mu}_{n,i}$ . While the properties of conjugate priors are useful for inference, other models could be considered with approximate inference techniques ([Clavier et al., 2023](#); [Phan et al., 2019](#)).

**Link with linear bandit (cont.).** The slightly unusual characteristic of (10) is that the prior distribution has correlated components. This can be addressed by the whitening trick ([Bishop, 2006](#)), defining  $\tilde{\mu} = \bar{\Sigma}^{-1/2} \mu$  and

$\tilde{b}_i = \bar{\Sigma}^{1/2} \bar{b}_i$ , giving

$$\begin{aligned} \tilde{\mu} &\sim \mathcal{N}(\bar{\Sigma}^{-1/2} \bar{\nu}, I_{L+K}) \\ Y_t | \tilde{\mu}, A_t &\sim \mathcal{N}(\tilde{b}_{A_t}^\top \tilde{\mu}, \sigma^2) \quad \forall t \in [n], \end{aligned} \quad (11)$$

where  $I_{L+K}$  is the  $(L+K)$ -dimensional identity matrix. Then, (11) corresponds to a linear bandit model with  $K$  arms and  $d = K + L$  features. However, this model comes with some limitations. First, when computing posteriors under (11), the time and space complexities are  $\mathcal{O}((K+L)^3)$  and  $\mathcal{O}((K+L)^2)$  respectively, compared to the  $\mathcal{O}(K+L^3)$  and  $\mathcal{O}(K+L^2)$  for our model (17). The feature dimension  $d = K + L$  can be reduced to  $d = K$  through the following QR decomposition:  $\tilde{B} = (\tilde{b}_1, \dots, \tilde{b}_K) \in \mathbb{R}^{(K+L) \times K}$  can be expressed as  $\tilde{B} = VR$ , where  $V \in \mathbb{R}^{(K+L) \times K}$  is an orthogonal matrix and  $R \in \mathbb{R}^{K \times K}$ . This leads to the following model  $\tilde{\mu} \sim \mathcal{N}(V^\top \bar{\Sigma}^{-1/2} \bar{\nu}, V^\top V)$  and  $Y_t | \tilde{\mu}, A_t \sim \mathcal{N}(R_{A_t}^\top \tilde{\mu}, \sigma^2)$ , yet the feature dimension  $d$  remains at the order of  $K$ , and computational efficiency is not improved with respect to  $K$ .

**From hierarchical bandit to MAB.** Marginalizing the hyper-prior in (9) leads to a MAB model,

$$\begin{aligned} \theta_i &\sim \mathcal{N}(b_i^\top \nu, \sigma_{0,i}^2 + b_i^\top \Sigma b_i) \quad \forall i \in [K] \\ Y_t | \mu, \theta, A_t &\sim \mathcal{N}(\theta_{A_t}, \sigma^2) \quad \forall t \in [n]. \end{aligned} \quad (12)$$

In this marginalized model, the agent does not know  $\mu$  and he doesn't want to model it. Therefore, only  $\theta$  is learned. The marginalized prior variance  $\sigma_{0,i}^2 + b_i^\top \Sigma b_i$  accounts for the uncertainty of the not-modeled effects. From (6), the corresponding posterior covariance of an arm  $i \in [K]$  is

$$\hat{\sigma}_{n,i}^{-2} = \frac{1}{\sigma_{0,i}^2 + b_i^\top \Sigma b_i} + \frac{\omega_i n}{\sigma^2}.$$

We illustrate the benefit of using hierarchical models over MAB models with a toy experiment. In the first setting, we uniformly draw a vector  $u \in [0, 1]$  and set  $\sigma_0 = 0.1u$  and  $\Sigma = 2I_L$ . In the second setting, we set  $\sigma_0 = u$  and  $\Sigma = 10^{-3}I_L$ . In both settings, we consider  $K = 50$  arms, and  $L = 10$  effects. Each  $\nu_i$  and  $b_i$  are sampled from  $[-1, 1]$ , and the allocation vector is set to uniform allocation,  $\omega_i^{\text{uni}} = \frac{1}{K}$  for any  $i \in [K]$ . Figure 3 shows the average posterior covariance  $\frac{1}{K} \sum_{i \in [K]} \sigma_{n,i}^2$  across all arms for both the (marginalized) standard MAB model (12) in blue and the hierarchical model (9) in red. The results show that the benefits of using hierarchical models are more pronounced when the initial uncertainty of the effects  $\Sigma$  is greater than the initial uncertainty of the mean rewards  $(\sigma_{0,i}^2)_{i \in [K]}$ .

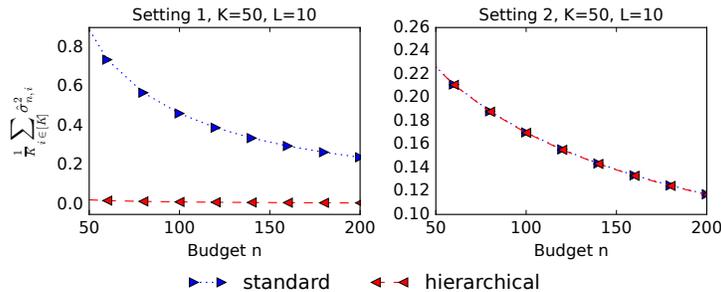


Figure 3: Average posterior covariance across all arms for standard MAB and hierarchical model for two settings.

### B.3 Beyond Gaussian Distributions

**Beyond linear models.** The standard linear model (7) can be generalized beyond linear mean rewards. The *Generalized Linear Bandit* (GLB) model with prior  $P_0$  writes (Filippi et al., 2010; Kveton et al., 2020)

$$\begin{aligned} \theta &\sim P_0 \\ Y_t | \theta, A_t &\sim P(\cdot; \theta, A_t) \quad \forall t \in [n], \end{aligned} \quad (13)$$

where the reward distribution  $P(\cdot; \theta, A_t)$  belongs to some exponential family with mean reward  $r(A_t; \theta) = \phi(\theta^\top x_{A_t})$ .  $\phi$  is called the *link function*. The *log-likelihood* of such reward distribution can be written as

$$\mathcal{L}_n(\theta) = \sum_{t=1}^n \log P(Y_t; \theta, A_t) = \sum_{t=1}^n Y_t \theta^\top x_{A_t} - A(\theta^\top x_{A_t}) + h(Y_t),$$

where  $A$  is a *log-partition* function and  $h$  another function. Importantly, (13) encompasses the *logistic bandit* model with the particular link function  $\phi(z) = \frac{1}{1+e^{-z}}$ .

The main challenge of (13) is that closed-form posterior generally does not exist. One method is to approximate the posterior distribution of  $\theta$  given  $H_n$  with Laplace approximation, that is,  $\mathbb{P}(\theta | H_n)$  is approximated with a multivariate Gaussian distribution with mean  $\hat{\theta}_{\text{MAP}}$  and covariance  $\hat{\Sigma}_{\text{Lap}}$ ,

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} \mathcal{L}_n(\theta) P_0(\theta), \quad \hat{\Sigma}_{\text{Lap}}^{-1} = \sum_{t=1}^n \dot{\phi}(\hat{\theta}_{\text{MAP}}^\top x_{A_t}) x_{A_t} x_{A_t}^\top,$$

where  $\phi$  is assumed continuously differentiable and increasing. Note that  $\hat{\theta}_{\text{MAP}}$  can be computed efficiently by iteratively reweighted least squares (Wolke and Schwetlick, 1988).

**Logistic Bandit.** In the particular case where the reward distribution is Bernoulli, the model writes

$$\begin{aligned} \theta &\sim \mathcal{N}(\mu_0, \Sigma_0) \\ Y_t | \theta, A_t &\sim \mathcal{B}(\phi(\theta^\top x_{A_t})) \quad \forall t \in [n], \end{aligned} \quad (14)$$

where  $\phi$  is the logistic function. Then, using a result from Spiegelhalter and Lauritzen (1990), the mean posterior reward can be approximated as

$$\mathbb{E}_{\theta \sim \mathcal{N}(\hat{\theta}_{\text{MAP}}, \hat{\Sigma}_{\text{Lap}})} [\phi(\theta^\top x_i)] \approx \frac{\phi(\hat{\theta}_{\text{MAP}}^\top x_i)}{\sqrt{1 + \frac{\pi}{8} \|x_i\|_{\hat{\Sigma}_{\text{Lap}}}}},$$

and we set the decision after  $n$  rounds as  $J_n = \operatorname{argmax}_{i \in [K]} \frac{\phi(\hat{\theta}_{\text{MAP}}^\top x_i)}{\sqrt{1 + \frac{\pi}{8} \|x_i\|_{\hat{\Sigma}_{\text{Lap}}}}$ .

Proving an upper bound on the expected PoE of this algorithm is challenging. Particularly, upper bounding the expectation with respect to  $H_n$  is hard because one needs to show that  $\hat{\theta}_{\text{MAP}}$  concentrates in norm towards its expectation  $\mathbb{E}_{H_n} [\hat{\theta}_{\text{MAP}}]$ . We leave this study for future work. However, we provide numerical experiments for this setting in Appendix D.4.

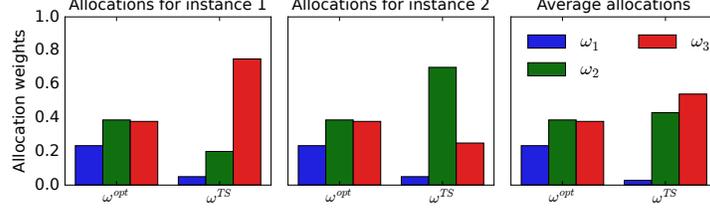
## B.4 Differences Between Optimized and Warmed-up Weights.

To illustrate the differences between optimized weights (PI-BAI( $\omega^{\text{opt}}$ )) and learned weights with Thompson sampling as a warm-up policy (PI-BAI( $\omega^{\text{TS}}$ )), we return to our motivating example in Section 4.1, where  $K = 3$ ,  $\mu_0 = (1, 1.9, 2)$  and  $\sigma_{0,i} = 0.3$  for all  $i \in [3]$ . We set the budget as  $n = 100$ . We repeat  $10^4$  times the following experiment: we sample a bandit instance from the prior and run Thompson sampling for  $n_w = 20$  rounds, then construct the allocation weights  $\omega^{\text{TS}}$ . Computing the weights  $\omega^{\text{opt}}$  by numerical optimization of (3) is done once at the beginning of these experiments.

Figure 4 shows an empirical comparison of the weights on 2 problem instances and on average over  $10^4$  runs. We see that, in this example, both allocation strategies assign high weights to arms 2 and 3, while allocating a small weight to arm 1. This is because, based on the prior information, arm 1 is highly unlikely to be the optimal arm. Then, the primary objective revolves around selecting the optimal arm among arms 2 and 3. Also, while the allocation weights  $\omega_i^{\text{TS}}$  vary with each bandit instance, their average values in all instances are similar to those of  $\omega_i^{\text{opt}}$ . Thus PI-BAI( $\omega^{\text{TS}}$ ) is more adaptive than PI-BAI( $\omega^{\text{opt}}$ ), while both have similar average behavior.

## C MISSING PROOFS

In this section, we give complete proof of our theoretical results. In Appendix C.1, we give proofs for the Bayesian posterior derivations and we provide technical results. Then we provide the complete proofs of our results for MAB (Appendix C.2), linear bandits (Appendix C.3) and hierarchical bandits (Appendix C.4).


 Figure 4: Allocations weights  $\omega_i^{\text{opt}}$  and  $\omega_i^{\text{TS}}$ .

### C.1 Technical Proofs and Posteriors Derivations

**Bayesian computations for hierarchical model.** We detail the computations of posterior distribution computations of the hierarchical Gaussian model,

$$\begin{aligned}
 \mu &\sim \mathcal{N}(\nu, \Sigma) \\
 \theta_i &\sim \mathcal{N}(b_i^\top \mu, \sigma_{0,i}^2) && \forall i \in [K] \\
 Y_t \mid \mu, \theta, A_t &\sim \mathcal{N}(\theta_{A_t}, \sigma^2) && \forall t \in [n],
 \end{aligned}$$

where we recall that  $B_{n,i} = \sum_{t \in \mathcal{T}_i} Y_t$  and  $\mathcal{T}_i = \{t \in [n], A_t = i\}$ .

**Lemma C.1** (Gaussian posterior update). *For any  $\rho \in \mathbb{R}$ ,  $\mu \in \mathbb{R}^L$ ,  $b \in \mathbb{R}^L$  and  $\sigma, \sigma_0 > 0$ ,  $m \in \mathbb{N}$ , we have*

$$\int_{\rho} \prod_{t \in [m]} \mathcal{N}(Y_t; \rho, \sigma^2) \mathcal{N}(\rho; b^\top \mu, \sigma_0^2) d\rho \propto \mathcal{N}(\mu; \mu_m, \Sigma_m),$$

with

$$\Sigma_m = \frac{m}{m\sigma_0^2 + \sigma^2} bb^\top; \quad \mu_m = \Sigma_m^{-1} \frac{\sum_{t \in [m]} Y_t}{m\sigma_0^2 + \sigma^2} b.$$

*Proof of Lemma C.1.* By keeping only terms that depend on  $\mu$ ,

$$\begin{aligned}
 f(\mu) &= \int_{\rho} \prod_{t \in [m]} \mathcal{N}(Y_t; \rho, \sigma^2) \mathcal{N}(\rho; b^\top \mu, \sigma_0^2) d\rho \\
 &\propto \int_{\rho} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{t \in [m]} (Y_t - \rho)^2 - \frac{1}{2\sigma_0^2} (\rho - b^\top \mu)^2 \right\} d\rho \\
 &\propto \int_{\rho} \exp \left\{ -\frac{1}{2}\rho^2 \left( \frac{1}{\sigma_0^2} + \frac{m}{\sigma^2} \right) - 2\rho \left( \frac{\sum_{t \in [m]} Y_t}{\sigma^2} + \frac{b^\top \mu}{\sigma_0^2} \right) \right\} d\rho \exp \left\{ -\frac{1}{2\sigma_0^2} \mu^\top bb^\top \mu \right\} \\
 &\propto \exp \left\{ \frac{1}{2} \left( \frac{1}{\sigma^2} \sum_{t \in [m]} Y_t + \frac{b^\top \mu}{\sigma_0^2} \right)^2 \frac{\sigma_0^2 \sigma^2}{\sigma^2 + m\sigma_0^2} - \frac{1}{2\sigma_0^2} \mu^\top bb^\top \mu \right\} \\
 &\propto \exp \left\{ \frac{\sum_{t \in [m]} Y_t b^\top \mu}{\sigma^2 + m\sigma_0^2} + \frac{\sigma^2}{2(\sigma^2 + m\sigma_0^2)} \mu^\top bb^\top \mu - \frac{1}{\sigma_0^2} bb^\top \mu \right\} \\
 &\propto \exp \left\{ -\frac{1}{2} \left( \mu^\top \frac{m}{\sigma^2 + m\sigma_0^2} bb^\top \mu - 2\mu^\top \frac{\sum_{t \in [m]} Y_t}{\sigma^2 + m\sigma_0^2} b \right) \right\} \\
 &\propto \mathcal{N}(\mu; \mu_m, \Sigma_m).
 \end{aligned}$$

□

**Lemma C.2** (Joint effect posterior). *For any  $n \in [N]$ , the joint effect posterior is a multivariate Gaussian  $Q_n(\mu) = \mathcal{N}(\check{\mu}, \check{\Sigma}_n)$ , where*

$$\check{\Sigma}_n^{-1} = \Sigma^{-1} + \sum_{i \in [K]} \frac{n_i}{n_i \sigma_{0,i}^2 + \sigma^2} b_i b_i^\top, \quad \check{\mu}_n = \check{\Sigma}_n \left( \Sigma^{-1} \nu + \sum_{i \in [K]} \frac{B_{n,i}}{n_i \sigma_{0,i}^2 + \sigma^2} b_i \right). \quad (15)$$

*Proof of Lemma C.2.* The joint effect posterior can be written as

$$\begin{aligned} Q_n(\mu) &\propto \int_{\theta} \mathcal{L}_{\theta}(Y_{A_1}, \dots, Y_{A_n}) P_0(\theta | \mu) d\theta Q(\mu) \\ &= \prod_{i \in [K]} \int_{\theta_i} \prod_{t \in \mathcal{T}_i} \mathcal{N}(Y_t; \theta_i, \sigma^2) \mathcal{N}(\theta_i; b_i^\top \mu, \sigma_{0,i}^2) d\theta_i \mathcal{N}(\mu; \nu, \Sigma). \end{aligned}$$

Applying Lemma C.1 gives

$$\int_{\theta_i} \prod_{t \in \mathcal{T}_i} \mathcal{N}(Y_t; \theta_i, \sigma^2) \mathcal{N}(\theta_i; b_i^\top \mu, \sigma_{0,i}^2) d\theta_i \propto \mathcal{N}(\mu; \check{\mu}_{n,i}, \check{\Sigma}_{n,i}),$$

with

$$\check{\Sigma}_{n,i}^{-1} = \frac{n_i}{\sigma_{0,i}^2 n_i + \sigma^2} b_i b_i^\top, \quad \check{\mu}_{n,i} = \check{\Sigma}_{n,i} b_i \frac{n_i}{\sigma_{0,i}^2 n_i + \sigma^2} \frac{B_{n,i}}{n_i}.$$

Therefore, the joint effect posterior is a product of Gaussian distributions,

$$Q_n(\mu) \propto \prod_{i \in [K]} \mathcal{N}(\mu; \check{\mu}_{n,i}, \check{\Sigma}_{n,i}) \mathcal{N}(\mu; \nu, \Sigma) \propto \mathcal{N}(\mu; \check{\mu}_n, \check{\Sigma}_n),$$

where

$$\begin{aligned} \check{\Sigma}_n &= \Sigma^{-1} + \sum_{i \in [K]} \check{\Sigma}_{n,i} = \Sigma^{-1} + \sum_{i \in [K]} \frac{n_i}{\sigma_{0,i}^2 n_i + \sigma^2} b_i b_i^\top \\ \check{\mu}_n &= \check{\Sigma}_n^{-1} \left( \Sigma^{-1} \nu + \sum_{i \in [K]} \check{\Sigma}_{n,i}^{-1} \check{\mu}_{n,i} \right) = \check{\Sigma}_n^{-1} \left( \Sigma^{-1} \nu + \sum_{i \in [K]} \frac{B_{n,i}}{n_i \sigma_{0,i}^2 + \sigma^2} b_i \right). \end{aligned}$$

□

**Lemma C.3** (Conditional arm posteriors). *For any  $n \in [n]$  and any arm  $i \in [K]$ , the conditional posterior distribution of arm  $i$  is a Gaussian distribution  $P_{n,i}(\theta_i | \mu) = \mathcal{N}(\tilde{\mu}_{n,i}, \tilde{\sigma}_{n,i}^2)$ , where*

$$\tilde{\sigma}_{n,i}^{-2} = \frac{1}{\sigma_{0,i}^2} + \frac{n_i}{\sigma^2}, \quad \tilde{\mu}_{n,i} = \tilde{\sigma}_{n,i}^2 \left( \frac{\mu^\top b_i}{\sigma_{0,i}^2} + \frac{B_{n,i}}{\sigma^2} \right). \quad (16)$$

*Proof of Lemma C.3.* The conditional posterior of arm  $i$  can be written as

$$\begin{aligned} P_{n,i}(\theta_i | \mu) &\propto \mathcal{L}_{\theta_i}(Y_{A_1}, \dots, Y_{A_n}) P_{0,i}(\theta_i | \mu) \\ &\propto \prod_{t \in \mathcal{T}_i} \mathcal{N}(Y_t; \theta_i, \sigma^2) \mathcal{N}(\theta_i; b_i^\top \mu, \sigma_{0,i}^2) \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{t \in \mathcal{T}_i} (Y_t - \theta_i)^2 - \frac{1}{2\sigma_{0,i}^2} (\theta_i - b_i^\top \mu)^2 \right\} \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{t \in \mathcal{T}_i} (-2Y_t \theta_i + \theta_i^2) - \frac{1}{2\sigma_{0,i}^2} (\theta_i^2 - 2\theta_i b_i^\top \mu) \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left( \theta_i^2 \left( \frac{n_i}{\sigma^2} + \frac{1}{\sigma_{0,i}^2} \right) - 2\theta_i \left( \frac{1}{\sigma^2} \sum_{t \in \mathcal{T}_i} Y_t + \frac{1}{\sigma_{0,i}^2} b_i^\top \mu \right) \right) \right\} \\ &\propto \mathcal{N}(\theta_i; \tilde{\mu}_{n,i}, \tilde{\sigma}_{n,i}^2) \end{aligned}$$

□

**Lemma C.4** (Marginal arm posterior). *For any  $n \in [n]$  and any arm  $i \in [K]$ , the marginal posterior distribution of arm  $i$  is a Gaussian distribution  $\mathbb{P}(\theta_i | H_n) = \mathcal{N}(\hat{\mu}_{n,i}, \hat{\sigma}_{n,i}^2)$ , where*

$$\hat{\sigma}_{n,i}^2 = \tilde{\sigma}_{n,i}^2 + \frac{\tilde{\sigma}_{n,i}^4}{\sigma_{0,i}^4} b_i^\top \check{\Sigma}_n b_i, \quad \hat{\mu}_{n,i} = \tilde{\sigma}_{n,i}^2 \left( \frac{\check{\mu}_n^\top b_i}{\sigma_{0,i}^2} + \frac{B_{n,i}}{\sigma_i^2} \right). \quad (17)$$

*Proof of Lemma C.4.* The marginal distribution of arm  $i$  can be written as

$$\begin{aligned} \int_{\mu} P_{n,i}(\theta_i | \mu) Q_n(\mu) d\mu &= \int_{\mu} \mathcal{N}(\theta_i; \check{\mu}_{n,i}, \tilde{\sigma}_{n,i}^2) \mathcal{N}(\mu; \check{\mu}_n, \check{\Sigma}_n) d\mu \\ &\propto \int_{\mu} \mathcal{N}\left(\theta_i; \tilde{\sigma}_{n,i}^2 \left( \frac{\mu^\top b_i}{\sigma_{0,i}^2} + \frac{B_{n,i}}{\sigma_i^2} \right), \tilde{\sigma}_{n,i}^2\right) \mathcal{N}(\mu; \check{\mu}_n, \check{\Sigma}_n) d\mu. \end{aligned}$$

The line above is a convolution of Gaussian measures, and can be written as (Bishop, 2006),

$$\begin{aligned} \int_{\mu} P_{n,i}(\theta_i | \mu) Q_n(\mu) d\mu &\propto \mathcal{N}\left(\theta_i; \tilde{\sigma}_{n,i}^2 \left( \frac{\check{\mu}_n^\top b_i}{\sigma_{0,i}^2} + \frac{B_{n,i}}{\sigma_i^2} \right), \tilde{\sigma}_{n,i}^2 + \frac{\tilde{\sigma}_{n,i}^2}{\sigma_{0,i}^2} b_i^\top \check{\Sigma}_n \frac{\tilde{\sigma}_{n,i}^2}{\sigma_{0,i}^2} b_i\right) d\mu \\ &= \mathcal{N}(\theta_i; \hat{\mu}_{n,i}, \hat{\sigma}_{n,i}^2). \end{aligned}$$

□

**Lemma C.5** (Technical lemma). *Let  $a > 0$  and  $X \sim \mathcal{N}(\mu, \sigma^2)$ . Then  $\mathbb{E}_X \left[ e^{-\frac{X^2}{2a^2}} \right] = \frac{1}{\sqrt{1 + \frac{2\sigma^2}{a^2}}} e^{-\frac{\mu^2}{2(\sigma^2 + a^2)}}$ .*

## C.2 Proofs for Multi-armed Bandits

From now, we consider that  $[\omega_k n] = \omega_k n \in \mathbb{N}$  for sake of simplicity.

**Theorem C.6** (Complete statement of Theorem 4.1). *For all  $\omega \in \Delta_K^+$ , the expected PoE of PI-BAI( $\omega$ ) under the MAB problem (5) is upper bounded as*

$$\mathcal{P}_n \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{e^{-\frac{(\mu_{0,i} - \mu_{0,j})^2}{2(\sigma_{0,i}^2 + \sigma_{0,j}^2)}}}{\sqrt{1 + n\phi_{i,j}(\omega)}}, \quad \text{where } \phi_{i,j}(\omega) = \frac{\sigma_{0,i}^4 \omega_i \left( \frac{\sigma^2}{n} + \omega_j \sigma_{0,j}^2 \right) + \sigma_{0,j}^4 \omega_j \left( \frac{\sigma^2}{n} + \omega_i \sigma_{0,i}^2 \right)}{\sigma^2 \sigma_{0,i}^2 \left( \frac{\sigma^2}{n} + \omega_j \sigma_{0,j}^2 \right) + \sigma^2 \sigma_{0,j}^2 \left( \frac{\sigma^2}{n} + \omega_i \sigma_{0,i}^2 \right)}. \quad (18)$$

*Remark C.7.* When  $\sigma_{0,i}^2 = \sigma_{0,j}^2$ ,  $\lim_{n \rightarrow +\infty} \phi_{i,j} = \lim_{n \rightarrow +\infty} \frac{\sigma_{0,i}^4 \omega_i \left( \frac{\sigma^2}{n} + \omega_j \sigma_{0,j}^2 \right) + \sigma_{0,j}^4 \omega_j \left( \frac{\sigma^2}{n} + \omega_i \sigma_{0,i}^2 \right)}{\sigma^2 \sigma_{0,i}^2 \left( \frac{\sigma^2}{n} + \omega_j \sigma_{0,j}^2 \right) + \sigma^2 \sigma_{0,j}^2 \left( \frac{\sigma^2}{n} + \omega_i \sigma_{0,i}^2 \right)} = \frac{2\sigma_0^2}{\sigma^2} \frac{\omega_i \omega_j}{\omega_i + \omega_j} = \mathcal{O}(1)$ .

**Proof of Theorem 4.1** We first write  $\mathcal{P}_n$  as a double sum over all possible distinct arms,

$$\begin{aligned} \mathbb{E} [\mathbb{P}(J_n \neq i_*(\theta) | H_n)] &= \mathbb{E} [\mathbb{1}\{J_n \neq i_*(\theta)\}] \\ &= \mathbb{E} \left[ \mathbb{E} \left[ \sum_{i=1}^K \sum_{j=1}^K \mathbb{1}\{i \neq j\} \mathbb{1}\{i_*(\theta) = i\} \mathbb{1}\{J_n = j\} | H_n \right] \right] \\ &= \sum_{\substack{i,j \in [K] \\ i \neq j}} \mathbb{E} [\mathbb{E} [\mathbb{1}\{i_*(\theta) = i\} \mathbb{1}\{J_n = j\} | H_n]] \\ &= \sum_{\substack{i,j \in [K] \\ i \neq j}} \mathbb{E} [\mathbb{P}(i_*(\theta) = i \cap J_n = j | H_n)] \end{aligned}$$

Since  $J_n : H_n \rightarrow [K]$ ,  $\mathbb{P}(J_n = j | H_n) = \mathbb{1}\{J_n = j\}$ . For the conditional probability of  $i_*(\theta) = i$  given the event  $J_n = j$  to be well defined, we want to make sure to condition on an event with non-zero probability. Considering

both events  $\{J_n = j\}$  or  $\{J_n \neq j\}$  under  $H_n$ ,

$$\begin{aligned}
 \sum_{\substack{i,j \in [K] \\ i \neq j}} \mathbb{E} [\mathbb{P}(i_*(\theta) = i \cap J_n = j \mid H_n)] &= \sum_{\substack{i,j \in [K] \\ i \neq j}} \mathbb{E} [\mathbb{P}(i_*(\theta) = i \cap J_n = j \mid H_n) (\mathbb{1}\{J_n = j\} + \mathbb{1}\{J_n \neq j\})] \\
 &= \sum_{\substack{i,j \in [K] \\ i \neq j}} \mathbb{E} [\mathbb{P}(i_*(\theta) = i \mid J_n = j, H_n) \mid J_n = j] \mathbb{P}(J_n = j). \\
 &= \sum_{\substack{i,j \in [K] \\ i \neq j}} \mathbb{E} [\mathbb{P}(i_*(\theta) = i \mid J_n = j, H_n) \mathbb{1}\{J_n = j\}],
 \end{aligned}$$

where the second equality holds by tower rule conditionally on  $H_n$ , and the last equation holds because  $J_n$  is deterministic conditionally on  $H_n$ . Overall, this gives

$$\mathcal{P}_n = \sum_{\substack{i,j \in [K] \\ i \neq j}} \mathbb{E} [\mathbb{P}(i_*(\theta) = i \mid J_n = j, H_n) \mathbb{1}\{J_n = j\}].$$

By definition of  $i_*(\theta)$  in the MAB setting and applying Hoeffding inequality for sub-Gaussian random variables,

$$\begin{aligned}
 \mathbb{P} \left( \operatorname{argmax}_{k \in [K]} \theta_k = i \mid H_n, J_n = j \right) &\leq \mathbb{P}(\theta_i \geq \theta_j \mid H_n, J_n = j) \\
 &= \mathbb{P}((\theta_i - \theta_j) - (\hat{\mu}_{n,i} - \hat{\mu}_{n,j}) \geq -(\hat{\mu}_{n,i} - \hat{\mu}_{n,j}) \mid H_n, J_n = j) \\
 &\leq \exp \left( -\frac{(\hat{\mu}_{n,i} - \hat{\mu}_{n,j})^2}{2(\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2)} \right). \tag{19}
 \end{aligned}$$

Therefore,

$$\mathbb{E} [\mathbb{P}(i_*(\theta) = i \mid J_n = j, H_n) \mathbb{1}\{J_n = j\}] \leq \mathbb{E} \left[ \exp \left( -\frac{(\hat{\mu}_{n,i} - \hat{\mu}_{n,j})^2}{2(\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2)} \right) \right] \tag{20}$$

We now want to compute this above expectation with respect to  $H_n$ .

First, we remark that because the scheduling of arms  $(A_1, \dots, A_n)$  is deterministic, the law of  $H_n = (A_1, Y_{A_1}, \dots, A_n, Y_{A_n})$  is the law of  $(Y_{A_1}, \dots, Y_{A_n})$ . Denoting  $\pi_{H_n}$  the marginal distribution of  $H_n$ ,

$$\pi_{H_n}(H_n) = \pi_{H_n}(Y_{A_1}, \dots, Y_{A_n}) = \int_{\theta} \mathcal{L}_{\theta}(Y_{A_1}, \dots, Y_{A_n}) P_0(\theta) d\theta,$$

where  $\mathcal{L}_{\theta}(Y_{A_1}, \dots, Y_{A_n})$  denotes the likelihood of  $(Y_{A_1}, \dots, Y_{A_n})$  given parameter  $\theta$  and  $P_0(\theta) = \prod_{i \in [K]} P_{0,i}(\theta_i)$  since each mean reward  $\theta_i$  is drawn independently from  $P_{0,i}$  in the MAB setting. Since rewards given parameter  $\theta$  are independent and identically distributed,

$$\begin{aligned}
 \pi_{H_n}(H_n) &= \int_{\theta} \prod_{i \in [K]} \mathcal{L}_{\theta_i}((Y_t)_{t \in \mathcal{T}_i}) P_{0,i}(\theta_i) d\theta_i \\
 &= \int_{\theta} \prod_{i \in [K]} \mathcal{N}((Y_t)_{t \in \mathcal{T}_i}; \theta_i \mathbf{1}_{\omega_i n}, \sigma^2 I_{\omega_i n}) \mathcal{N}(\theta_i; \mu_{0,i}, \sigma_{0,i}^2) d\theta_i, \tag{21}
 \end{aligned}$$

where  $\mathbf{1}_q$  denotes the vector of size  $q$  whose all components are 1s.

(21) is a convolution of Gaussians and can be computed easily (Bishop, 2006),

$$\mathcal{N}((Y_t)_{t \in \mathcal{T}_i}; \theta_i \mathbf{1}_{\omega_i n}, \sigma^2 I_{\omega_i n}) \mathcal{N}(\theta_i; \mu_{0,i}, \sigma_{0,i}^2) = \mathcal{N}((Y_t)_{t \in \mathcal{A}_i}; \mu_{0,i} \mathbf{1}_{\omega_i n}, \sigma^2 I_{\omega_i n} + \sigma_{0,i}^2 \mathbf{1}_{\omega_i n} \mathbf{1}_{\omega_i n}^{\top}).$$

The above covariance matrix exhibits  $\sigma^2 + \sigma_{0,i}^2$  on the diagonal and  $\sigma_{0,i}^2$  out of diagonal.

We are now ready to compute some useful statistics : for any  $i \in [K]$ ,

$$\mathbb{E} [\hat{\mu}_{n,i}] = \mathbb{E} \left[ \frac{\sigma^2}{\sigma^2 + \sigma_{0,i}^2 \omega_i n} \mu_{0,i} + \frac{\sigma_{0,i}^2}{\sigma^2 + \sigma_{0,i}^2 \omega_i n} \sum_{t \in \mathcal{T}_i} Y_t \right] = \mu_{0,i} \quad (22)$$

$$\mathbb{V} (\hat{\mu}_{n,i}) = \frac{\sigma_{0,i}^4}{(\sigma^2 + \sigma_{0,i}^2 \omega_i n)^2} \mathbb{V} \left( \sum_{t \in \mathcal{T}_i} Y_t \right) = \frac{\sigma_{0,i}^4}{(\sigma^2 + \sigma_{0,i}^2 \omega_i n)} \omega_i n \quad (23)$$

$$\mathbb{E} \left[ \frac{1}{\omega_i n} \sum_{t \in \mathcal{T}_i} Y_t \right] = \mu_{0,i} \quad (24)$$

$$\mathbb{V} \left( \frac{1}{\omega_i n} \sum_{t \in \mathcal{T}_i} Y_t \right) = \frac{1}{\omega_i^2 n^2} (\omega_i n (\sigma^2 + \sigma_{0,i}^2) + (\omega_i^2 n^2 - \omega_i n) \sigma_{0,i}^2) = \frac{\sigma^2}{\omega_i n} + \sigma_{0,i}^2 \quad (25)$$

Applying Lemma C.5 on (20) and simplifying terms gives

$$\mathbb{E} [\mathbb{P} (i_*(\theta) = i \mid J_n = j, H_n) \mathbb{1}\{J_n = j\}] \leq \mathbb{E} \left[ \exp \left( - \frac{(\hat{\mu}_{n,i} - \hat{\mu}_{n,j})^2}{2(\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2)} \right) \right] = \frac{e^{-\frac{(\mu_{0,i} - \mu_{0,j})^2}{2(\sigma_{0,i}^2 + \sigma_{0,j}^2)}}}{\sqrt{1 + n\phi_{i,j}}}.$$

### C.3 Proofs for Linear Bandits

**Theorem C.8** (Complete statement of Theorem 4.2). *Assume that  $x_i \neq x_j$  for any  $i \neq j$ . Then, for all  $\omega \in \Delta_K^+$ , the expected PoE of PI-BAI using allocation  $\omega$  under linear bandit problem (7) is upper bounded as*

$$\mathcal{P}_n \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{1}{\sqrt{1 + \frac{c_{i,j}(\omega)}{\|x_i - x_j\|_{\Sigma_n}^2}}} e^{-\frac{(\mu_0^\top x_i - \mu_0^\top x_j)^2}{2\|x_i - x_j\|_{\Sigma_0}^2}},$$

where:

$$\text{Cov}(\hat{\mu}_n) = \frac{1}{\sigma^4} \hat{\Sigma}_n \left( \underbrace{\sum_{i \in [K]} (\omega_i n (\sigma^2 + x_i^\top \Sigma_0 x_i)) x_i x_i^\top}_{\text{variance terms}} + \underbrace{\sum_{i \in [K]} \sum_{j \in [K] \setminus \{i\}} x_i^\top \Sigma_0 x_j \omega_i \omega_j n^2 x_i x_j^\top}_{\text{covariance between arms}} \right) \hat{\Sigma}_n$$

$$c_{i,j} = \|x_i - x_j\|_{\text{Cov}(\hat{\mu}_n)}^2. \quad (26)$$

#### Proof of Theorem 4.2.

The proof for the linear model follows the same steps as the MAB model by rewriting  $\mathcal{P}_n$  as

$$\mathcal{P}_n = \sum_{\substack{i,j \in [K] \\ i \neq j}} \mathbb{E} [\mathbb{P} (i_*(\theta) = i \mid J_n = j, H_n) \mathbb{1}\{J_n = j\}].$$

By definition of  $i_*(\theta)$  and  $J_n$  in the linear bandit setting,

$$\begin{aligned} \mathbb{P} (i_*(\theta) = i \mid H_n, J_n = j) &= \mathbb{P} (\forall k \in [K], \theta^\top x_i \geq \theta^\top x_k \mid H_n, J_n = j) \\ &\leq \mathbb{P} (\theta^\top x_i \geq \theta^\top x_j \mid J_n = j, H_n) \\ &\leq \exp \left( - \frac{\|\hat{\mu}_n\|_{(x_i - x_j)(x_i - x_j)^\top}^2}{2\|x_i - x_j\|_{\Sigma_n}^2} \right), \end{aligned}$$

where the last inequality follows from Hoeffding inequality for sub-Gaussian random variables. Taking the expectation with respect to  $H_n$ ,

$$\mathbb{E} [\mathbb{P} (i_*(\theta) = i \mid H_n, J_n = j) \mathbb{1}\{J_n = j\}] \leq \mathbb{E} \left[ \exp \left( - \frac{\|\hat{\mu}_n\|_{(x_i - x_j)(x_i - x_j)^\top}^2}{2\|x_i - x_j\|_{\Sigma_n}^2} \right) \right]. \quad (27)$$

Then we remark that the expectation of  $\hat{\mu}_n$  with respect to  $H_n$  is

$$\mathbb{E} [\hat{\mu}_n] = \mathbb{E} \left[ \hat{\Sigma}_n \left( \Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} \sum_{t \in [n]} Y_t x_{A_t} \right) \right] = \hat{\Sigma}_n \left( \Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} \mathbb{E} \left[ \sum_{t \in [n]} Y_t x_{A_t} \right] \right),$$

since the scheduling  $(A_1, \dots, A_n)$  is known beforehand. Now,

$$\mathbb{E} \left[ \sum_{t \in [n]} Y_t x_{A_t} \right] = \sum_{t \in [n]} \mathbb{E} [Y_t] x_{A_t} = \sum_{t \in [n]} \mu_0^\top x_{A_t} x_{A_t},$$

where  $\mathbb{E} [Y_t]$  was obtained by marginalizing the likelihood over the prior distribution as in (21).

Rearranging the terms permits to conclude that  $\mathbb{E} [\hat{\mu}_n] = \mu_0$ . Then we can compute the expectation in (27) by applying Lemma C.5, Sylvester identity, and some simplifications:

$$\mathbb{E} \left[ \exp \left( - \frac{\|\hat{\mu}_n\|_{(x_i - x_j)(x_i - x_j)^\top}^2}{2\|x_i - x_j\|_{\hat{\Sigma}_n}^2} \right) \right] = \frac{1}{\sqrt{1 + \frac{\|x_i - x_j\|_{\text{Cov}(\hat{\mu}_n)}^2}{\|x_i - x_j\|_{\hat{\Sigma}_n}^2}}} e^{-\frac{1}{2} \|\mu_0\|_{\Lambda_{i,j}}^2}. \quad (28)$$

where from Lemma C.5,

$$\begin{aligned} \Lambda_{i,j} &= \text{Cov}(\hat{\mu}_n)^{-1} - \text{Cov}(\hat{\mu}_n)^{-1} \left( \text{Cov}(\hat{\mu}_n)^{-1} + \frac{(A_i - A_j)(A_i - A_j)}{\|A_i - A_j\|_{\hat{\Sigma}_n}^2} \right)^{-1} \text{Cov}(\hat{\mu}_n)^{-1} \\ &= \text{Cov}(\hat{\mu}_n)^{-1} - \text{Cov}(\hat{\mu}_n)^{-1} \left( I_d + \text{Cov}(\hat{\mu}_n) \frac{(x_i - x_j)(x_i - x_j)}{\|x_i - x_j\|_{\hat{\Sigma}_n}^2} \right)^{-1} \\ &= \frac{(x_i - x_j)(x_i - x_j)^\top}{\|x_i - x_j\|_{\hat{\Sigma}_n}^2 + \|x_i - x_j\|_{\text{Cov}(\hat{\mu}_n)}^2}. \end{aligned}$$

The last equality follows from an application of Sherman-Morrison identity. Applying the law of total expectation,

$$\text{Cov}(\theta) = \mathbb{E} [\text{Cov}(\theta | H_n)] + \text{Cov}(\mathbb{E}[\theta | H_n]) = \hat{\Sigma}_n + \text{Cov}(\hat{\mu}_n).$$

Therefore,

$$\|x_i - x_j\|_{\hat{\Sigma}_n}^2 + \|x_i - x_j\|_{\text{Cov}(\hat{\mu}_n)}^2 = \|x_i - x_j\|_{\hat{\Sigma}_n + \text{Cov}(\hat{\mu}_n)}^2 = \|x_i - x_j\|_{\Sigma_0}^2.$$

Plugging these into (28), we obtain

$$\mathbb{E} [\mathbb{P}(i_*(\theta) = i | H_n, J_n = j)] \leq \frac{e^{-\frac{\|\mu_0\|_{(x_i - x_j)(x_i - x_j)^\top}^2}{2\|x_i - x_j\|_{\hat{\Sigma}_0}^2}}}{\sqrt{1 + \frac{\|x_i - x_j\|_{\text{Cov}(\hat{\mu}_n)}^2}{\|x_i - x_j\|_{\hat{\Sigma}_n}^2}}} = \frac{e^{-\frac{(\mu_0^\top x_i - \mu_0^\top x_j)^2}{2\|x_i - x_j\|_{\hat{\Sigma}_0}^2}}}{\sqrt{1 + \frac{\|x_i - x_j\|_{\text{Cov}(\hat{\mu}_n)}^2}{\|x_i - x_j\|_{\hat{\Sigma}_n}^2}}}.$$

### Computation of $\text{Cov}(\hat{\mu}_n)$ .

By definition of Gaussian posteriors in linear bandit in (8),

$$\text{Cov}(\hat{\mu}_n) = \text{Cov} \left( \hat{\Sigma}_n \left( \Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} B_n \right) \right) = \hat{\Sigma}_n \text{Cov} \left( \Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} B_n \right) \hat{\Sigma}_n = \frac{1}{\sigma^4} \hat{\Sigma}_n \text{Cov}(B_n) \hat{\Sigma}_n,$$

and

$$\begin{aligned}
 \text{Cov}(B_n) &= \sum_{t \in [n]} \mathbb{V}(Y_t x_{A_t}) + \sum_{t \in [n]} \sum_{t' \in [n], t \neq t'} \text{Cov}(Y_t x_{A_t}, Y_{t'} x_{A_{t'}}) \\
 &= \sum_{t \in [n]} \mathbb{V}(Y_t) x_{A_t} x_{A_t}^\top + \sum_{t \in [n]} \sum_{t' \in [n], t \neq t'} \text{Cov}(Y_t, Y_{t'}) x_{A_t} x_{A_{t'}}^\top \\
 &= \sum_{k \in [K]} \omega_k n \mathbb{V}(Y_{x_k}) x_k x_k^\top + \sum_{\substack{i, j \in [K] \\ i \neq j}} \omega_i \omega_j n^2 \text{Cov}(Y_{x_i}, Y_{x_j}) x_i x_j^\top \\
 &= \sum_{k \in [K]} \omega_k n (\sigma^2 + x_k^\top \Sigma_0 x_k) x_k x_k^\top + \sum_{\substack{i, j \in [K] \\ i \neq j}} \omega_i \omega_j n^2 (x_i^\top \Sigma_0 x_j) x_i x_j^\top.
 \end{aligned}$$

### Proof of Corollary 4.3.

We first prove a useful lemma that holds for Bayesian G-optimal design.

**Lemma C.9.** *Let  $\mathcal{X}$  a finite set such that  $|\mathcal{X}| = K$ ,  $\xi : \mathcal{X} \rightarrow [0, 1]$  a distribution on  $\mathcal{X}$  so that  $\sum_{x \in \mathcal{X}} \xi(x) = 1$ ,  $V_n(\xi) = \sum_{x \in \mathcal{X}} \xi(x) x x^\top + \frac{\sigma^2}{n} \Sigma_0^{-1}$ ,  $\Sigma_0 \in \mathbb{R}^{d \times d}$  a diagonal matrix, and  $f(\xi) = \log \det(V_n(\xi))$ . If  $\xi^* = \text{argmin}_{\xi \in \Delta_K} f(\xi)$ , then  $\max_{x \in \mathcal{X}} \|x\|_{V_n(\xi^*)^{-1}}^2 \leq d$ .*

*Proof of Lemma C.9.* By concavity of  $\xi \mapsto f(\xi)$ , we have for any  $\xi$  that

$$0 \geq \langle \nabla f(\xi^*), \xi - \xi^* \rangle = \sum_{x \in \mathcal{X}} \xi(x) [\nabla f(\xi^*)]_x - \langle \xi^*, \nabla f(\xi^*) \rangle,$$

and since this holds for any pdf  $\xi$ , choosing  $\xi = \delta_{x'}$  for an arbitrary action  $x'$  yields

$$[\nabla f(\xi^*)]_{x'} \leq \langle \xi^*, \nabla f(\xi^*) \rangle \quad \text{for any } x' \in \mathcal{X}.$$

Since r.h.s. does not depend on  $x'$ ,

$$\max_{x \in \mathcal{X}} [\nabla f(\xi^*)]_x \leq \langle \xi^*, \nabla f(\xi^*) \rangle. \quad (29)$$

By the property of the gradient of log-determinant,  $[\nabla f(\xi^*)]_x = \|x\|_{V_n(\xi^*)^{-1}}$ . Therefore, for any  $\xi$ ,

$$\begin{aligned}
 \langle \xi, \nabla f(\xi) \rangle &= \sum_{x \in \mathcal{X}} \xi(x) \|x\|_{V^{-1}(\xi)}^2 \\
 &= \sum_{x \in \mathcal{X}} \xi(x) x^\top V_n(\xi)^{-1} x \\
 &= \text{Tr} \left( \sum_{x \in \mathcal{X}} \xi(x) x^\top V_n(\xi)^{-1} x \right) \\
 &= \text{Tr} \left( \sum_{x \in \mathcal{X}} \xi(x) x x^\top \left( \sum_{x' \in \mathcal{X}} \xi(x') x' x'^\top + \frac{\sigma^2}{n} \Sigma_0^{-1} \right)^{-1} \right) \\
 &= \text{Tr} \left( E \left( E + \frac{\sigma^2}{n} \Sigma_0^{-1} \right)^{-1} \right) \quad \text{where } E = \sum_{x \in \mathcal{X}} \xi(x) x x^\top \\
 &= \text{Tr} \left( \left( I_d + \frac{\sigma^2}{n} \Sigma_0^{-1} E^{-1} \right)^{-1} \right) \\
 &= \text{Tr} \left( I_d - \frac{\sigma^2}{n} \Sigma_0^{-1} \left( I_d + \frac{\sigma^2}{n} E^{-1} \Sigma_0^{-1} \right)^{-1} E^{-1} \right) \quad (\text{Woodbury identity}) \\
 &= \text{Tr}(I_d) - \frac{\sigma^2}{n} \text{Tr} \left( \left( E \Sigma_0 + \frac{\sigma^2}{n} I_d \right)^{-1} \right) \\
 &\leq \text{Tr}(I_d) = d.
 \end{aligned}$$

All putting together in (29) with  $\xi^* = \operatorname{argmin}_{\xi \in \Delta_K} f(\xi)$  implies  $\max_{x \in \mathcal{X}} \|x\|_{V(\xi^*)^{-1}} \leq d$ .  $\square$

A direct implication of Lemma C.9 is that  $\max_{x \in \mathcal{X}} \|x\|_{\Sigma_n} \leq \frac{d\sigma^2}{n}$ . Therefore,

$$\mathcal{P}_n \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{e^{-\frac{(\mu_0^\top x_i - \mu_0^\top x_j)^2}{2\|x_i - x_j\|_{\Sigma_0}^2}}}{\sqrt{1 + \frac{c_{i,j}}{\|x_i - x_j\|_{\Sigma_n}^2}}} \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{e^{-\frac{(\mu_0^\top x_i - \mu_0^\top x_j)^2}{2\|x_i - x_j\|_{\Sigma_0}^2}}}{\sqrt{1 + \frac{c_{i,j}}{2 \max_{x \in \mathcal{X}} \|x\|_{\Sigma_n}^2}}} \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{e^{-\frac{(\mu_0^\top x_i - \mu_0^\top x_j)^2}{2\|x_i - x_j\|_{\Sigma_0}^2}}}{\sqrt{1 + n \frac{c_{i,j}}{2d\sigma^2}}}.$$

#### C.4 Proofs for Hierarchical Bandits

We begin by stating the complete proof.

**Theorem C.10** (Complete statement of Theorem 4.4). *For all  $\omega \in \Delta_K^+$ , the expected PoE of PI-BAI using allocation  $\omega$  under the hierarchical bandit problem (9) is upper bounded as*

$$\mathcal{P}_n \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{1}{\sqrt{1 + \frac{c_{i,j}}{\bar{\sigma}_{n,i} + \bar{\sigma}_{n,j}}}} e^{-\frac{(\nu^\top b_i - \nu^\top b_j)^2}{2(\|b_i - b_j\|_{\Sigma}^2 + \sigma_{0,i}^2 + \sigma_{0,j}^2)}},$$

where

$$\begin{aligned} c_{i,j} = \mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j}) &= \frac{\sigma_{0,i}^4}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)^2} \mathbb{V}(B_{n,i}) + \frac{\sigma_{0,j}^4}{(\sigma_{0,j}^2 \omega_j n + \sigma^2)^2} \mathbb{V}(B_{n,j}) \\ &+ \frac{\sigma^4}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)^2} \sum_{k \in [K]} \frac{(b_k^\top \check{\Sigma}_n b_i)^2}{(\sigma^2 + \omega_k n \sigma_{0,k}^2)^2} \mathbb{V}(B_{n,k}) + \frac{\sigma^4}{(\sigma_{0,j}^2 \omega_j n + \sigma^2)^2} \sum_{k \in [K]} \frac{(b_k^\top \check{\Sigma}_n b_j)^2}{(\sigma^2 + \omega_k n \sigma_{0,k}^2)^2} \mathbb{V}(B_{n,k}) \\ &+ \frac{\sigma^4}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)^2} \sum_{k \in [K]} \sum_{k' \in [K] \setminus \{k\}} \frac{(b_k^\top \check{\Sigma}_n b_i) \cdot (b_{k'}^\top \check{\Sigma}_n b_i)}{(\sigma_{0,k}^2 \omega_k n + \sigma^2) \cdot (\sigma_{0,k'}^2 \omega_{k'} n + \sigma^2)} \operatorname{cov}(B_{n,k}, B_{n,k'}) \\ &+ \frac{\sigma^4}{(\sigma_{0,j}^2 \omega_j n + \sigma^2)^2} \sum_{k \in [K]} \sum_{k' \in [K] \setminus \{k\}} \frac{(b_k^\top \check{\Sigma}_n b_j) \cdot (b_{k'}^\top \check{\Sigma}_n b_j)}{(\sigma_{0,k}^2 \omega_k n + \sigma^2) \cdot (\sigma_{0,k'}^2 \omega_{k'} n + \sigma^2)} \operatorname{cov}(B_{n,k}, B_{n,k'}) \\ &- \frac{2\sigma^4}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)(\sigma_{0,j}^2 \omega_j n + \sigma^2)} \left[ \sum_{k \in [K]} \frac{(b_k^\top \check{\Sigma}_n b_i)(b_{k'}^\top \check{\Sigma}_n b_j)}{(\sigma^2 + \omega_k n \sigma_{0,k}^2)^2} \mathbb{V}(B_{n,k}) + \sum_{k \in [K]} \sum_{k' \in [K] \setminus \{k\}} \frac{(b_k^\top \check{\Sigma}_n b_i) \cdot (b_{k'}^\top \check{\Sigma}_n b_j)}{(\sigma_{0,k}^2 \omega_k n + \sigma^2) \cdot (\sigma_{0,k'}^2 \omega_{k'} n + \sigma^2)} \operatorname{cov}(B_{n,k}, B_{n,k'}) \right] \\ &- \frac{2\sigma_{0,i}^2 \sigma_{0,j}^2}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)(\sigma_{0,j}^2 \omega_j n + \sigma^2)} \operatorname{cov}(B_{n,i}, B_{n,j}) \\ &+ \frac{2\sigma^2 \sigma_{0,i}^2}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)^2} \left[ \sum_{k \in [K] \setminus \{i\}} \frac{b_k^\top \check{\Sigma}_n b_i}{(\sigma_{0,k}^2 \omega_k n + \sigma^2)} \operatorname{cov}(B_{n,k}, B_{n,i}) + \frac{b_i^\top \check{\Sigma}_n b_i}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)} \mathbb{V}(B_{n,i}) \right] \\ &+ \frac{2\sigma^2 \sigma_{0,j}^2}{(\sigma_{0,j}^2 \omega_j n + \sigma^2)^2} \left[ \sum_{k \in [K] \setminus \{j\}} \frac{b_k^\top \check{\Sigma}_n b_j}{(\sigma_{0,k}^2 \omega_k n + \sigma^2)} \operatorname{cov}(B_{n,k}, B_{n,j}) + \frac{b_j^\top \check{\Sigma}_n b_j}{(\sigma_{0,j}^2 \omega_j n + \sigma^2)} \mathbb{V}(B_{n,j}) \right] \\ &- \frac{2\sigma^2 \sigma_{0,j}^2}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)(\sigma_{0,j}^2 \omega_j n + \sigma^2)} \left[ \sum_{k \in [K] \setminus \{j\}} \frac{b_k^\top \check{\Sigma}_n b_i}{\sigma^2 + \omega_k n \sigma_{0,k}^2} \operatorname{cov}(B_{n,k}, B_{n,i}) + \frac{b_j^\top \check{\Sigma}_n b_i}{\sigma^2 + \omega_j n \sigma_{0,j}^2} \mathbb{V}(B_{n,j}) \right] \\ &- \frac{2\sigma^2 \sigma_{0,i}^2}{(\sigma_{0,j}^2 \omega_j n + \sigma^2)(\sigma_{0,i}^2 \omega_i n + \sigma^2)} \left[ \sum_{k \in [K] \setminus \{i\}} \frac{b_k^\top \check{\Sigma}_n b_j}{\sigma^2 + \omega_k n \sigma_{0,k}^2} \operatorname{cov}(B_{n,k}, B_{n,j}) + \frac{b_i^\top \check{\Sigma}_n b_j}{\sigma^2 + \omega_i n \sigma_{0,i}^2} \mathbb{V}(B_{n,i}) \right], \end{aligned} \tag{30}$$

where we defined  $\check{\Sigma}_n$  from (15),

$$\begin{aligned} \check{\Sigma}_n^{-1} &= \Sigma^{-1} + \sum_{k \in [K]} \frac{\omega_k n}{\sigma^2 + \omega_k n \sigma_{0,k}^2} b_k b_k^\top, & \mathbb{V}(B_{n,k}) &= \omega_k n \sigma^2 + \omega_k^2 n^2 (\sigma_{0,k}^2 + b_k^\top \Sigma b_k) \\ \operatorname{cov}(B_{n,k}, B_{n,k'}) &= \omega_k \omega_{k'} n^2 b_k^\top \Sigma b_{k'}. \end{aligned}$$

*Proof of Theorem C.10.* This proof follows the same idea of the proof of Theorem 4.1. We first write  $\mathcal{P}_n$  as

$$\mathcal{P}_n = \sum_{\substack{i,j \in [K] \\ i \neq j}} \mathbb{E} [\mathbb{P}(i_*(\theta) = i \mid J_n = j, H_n) \mathbf{1}\{J_n = j\}].$$

Following (19), by applying Hoeffding inequality for sub-Gaussian random variables,

$$\begin{aligned} \mathbb{P}\left(\operatorname{argmax}_{k \in [K]} \theta_k = i \mid H_n, J_n = j\right) &\leq \mathbb{P}(\theta_i \geq \theta_j \mid H_n, J_n = j) \\ &= \mathbb{P}((\theta_i - \theta_j) - (\hat{\mu}_{n,i} - \hat{\mu}_{n,j}) \geq -(\hat{\mu}_{n,i} - \hat{\mu}_{n,j}) \mid H_n, J_n = j) \\ &\leq \exp\left(-\frac{(\hat{\mu}_{n,i} - \hat{\mu}_{n,j})^2}{2(\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2)}\right), \end{aligned}$$

where  $\hat{\mu}_{n,i}$  and  $\hat{\sigma}_{n,i}^2$  are given by (17). Taking the expectation with respect to the history  $H_n$ ,

$$\mathbb{E}[\mathbb{P}(i_*(\theta) = i \mid J_n = j, H_n) \mathbf{1}\{J_n = j\}] \leq \mathbb{E}\left[e^{-\frac{(\hat{\mu}_{n,j} - \hat{\mu}_{n,i})^2}{2(\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2)}}\right].$$

Since  $\hat{\mu}_{n,i} - \hat{\mu}_{n,j} \sim \mathcal{N}(\mathbb{E}[\hat{\mu}_{n,i}] - \mathbb{E}[\hat{\mu}_{n,j}], \mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j}))$ , applying Lemma C.5 gives

$$\mathbb{E}\left[e^{-\frac{(\hat{\mu}_{n,j} - \hat{\mu}_{n,i})^2}{2(\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2)}}\right] = \frac{1}{\sqrt{1 + \frac{\mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j})}{\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2}}} \exp\left(-\frac{(\mathbb{E}[\hat{\mu}_{n,i}] - \mathbb{E}[\hat{\mu}_{n,j}])^2}{2(\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2)} \frac{1}{1 + \frac{\mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j})}{\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2}}\right)$$

Therefore,

$$\mathbb{E}[\mathbb{P}(i_*(\theta) = i \mid J_n = j, H_n)] \leq \frac{1}{\sqrt{1 + \frac{\mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j})}{\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2}}} e^{-\frac{(\mathbb{E}[\hat{\mu}_{n,i}] - \mathbb{E}[\hat{\mu}_{n,j}])^2}{2(\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2 + \mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j}))}}.$$

Now we want to simplify  $\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2 + \mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j})$ . on one hand, by the law of total variance,

$$\mathbb{V}(\theta_i - \theta_j) = \mathbb{E}[\mathbb{V}(\theta_i - \theta_j \mid H_n)] + \mathbb{V}(\mathbb{E}[\theta_i - \theta_j \mid H_n]) = \hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2 + \mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j}),$$

On the other hand,

$$\begin{aligned} \mathbb{V}(\theta_i - \theta_j) &= \mathbb{E}[\mathbb{V}(\theta_i - \theta_j \mid \mu)] + \mathbb{V}(\mathbb{E}[\theta_i - \theta_j \mid \mu]) = \sigma_{0,i}^2 + \sigma_{0,j}^2 + \mathbb{V}((b_i - b_j)^\top \mu) \\ &= \sigma_{0,i}^2 + \sigma_{0,j}^2 + \|b_i - b_j\|_\Sigma^2. \end{aligned}$$

Combining these two last equations gives  $\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2 + \mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j}) = \sigma_{0,i}^2 + \sigma_{0,j}^2 + \|b_i - b_j\|_\Sigma^2$ .

Therefore,

$$\mathbb{E}[\mathbb{P}(i_*(\theta) = i \mid J_n = j, H_n)] \leq \frac{1}{\sqrt{1 + \frac{\mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j})}{\hat{\sigma}_{n,i}^2 + \hat{\sigma}_{n,j}^2}}} e^{-\frac{(\mathbb{E}[\hat{\mu}_{n,i}] - \mathbb{E}[\hat{\mu}_{n,j}])^2}{2(\sigma_{0,i}^2 + \sigma_{0,j}^2 + \|b_i - b_j\|_\Sigma^2)}}. \quad (31)$$

**Computing  $\mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j})$ .**

The rest of the proof consists to compute  $\mathbb{E}[\hat{\mu}_{n,i}]$  and  $\mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j})$  for  $(i, j)$ . Denoting  $Q$  the latent prior distribution  $\mu \sim Q$  and  $\pi_{H_n}$  the law of  $H_n$ ,

$$\begin{aligned} \pi_{H_n}(H_n) &= \pi_{H_n}(Y_{A_1}, \dots, Y_{A_n}) \\ &= \iint_{(\theta, \mu)} \mathcal{L}_\theta(Y_{A_1}, \dots, Y_{A_n}) P_0(\theta \mid \mu) Q(\mu) d\theta d\mu \\ &= \iint_{(\theta, \mu)} \prod_{i \in [K]} \mathcal{L}_{\theta_i}((Y_t)_{t \in \mathcal{T}_i}) P_{0,i}(\theta_i \mid \mu) Q(\mu) d\theta_i d\mu \\ &= \int_\mu \left[ \prod_{i \in [K]} \int_{\theta_i} \mathcal{N}((Y_t)_{t \in \mathcal{T}_i}; \theta_i \mathbf{1}_{\omega_i n}, \sigma^2 I_{\omega_i n}) \mathcal{N}(\theta_i; b_i^\top \mu, \sigma_{0,i}^2) d\theta_i \right] Q(\mu) d\mu. \end{aligned}$$

From properties of Gaussian convolutions (Bishop, 2006),

$$\int_{\theta_i} \mathcal{N}((Y_t)_{t \in \mathcal{T}_i}; \theta_i \mathbf{1}_{\omega_i n}, \sigma^2 I_{\omega_i n}) \mathcal{N}(\theta_i; b_i^\top \mu, \sigma_{0,i}^2) d\theta_i = \mathcal{N}((Y_t)_{t \in \mathcal{T}_i}; (b_i^\top \mu) \mathbf{1}_{\omega_i n}, \sigma^2 I_{\omega_i n} + \mathbf{1}_{\omega_i n} \mathbf{1}_{\omega_i n}^\top \sigma_{0,i}^2).$$

Therefore,

$$\begin{aligned} & \prod_{i \in [K]} \int_{\theta_i} \mathcal{N}((Y_t)_{t \in \mathcal{T}_i}; \theta_i \mathbf{1}_{\omega_i n}, \sigma^2 I_{\omega_i n}) \mathcal{N}(\theta_i; b_i^\top \mu, \sigma_{0,i}^2) d\theta_i \\ &= \mathcal{N}(H_n; \sum_{i \in [K]} e_i(\mathbb{R}^K) \otimes \left( \sum_{t \in [\omega_i n]} e_t(\mathbb{R}^{\omega_i n}) \otimes b_i^\top \right) \otimes \mu, I_K \otimes (\sigma^2 I_{\omega_i n} + \mathbf{1}_{\omega_i n} \mathbf{1}_{\omega_i n}^\top \sigma_{0,i}^2)), \end{aligned}$$

where we define explicitly  $e_i(\mathbb{R}^K)$  as the  $i^{\text{th}}$  base vector of  $\mathbb{R}^K$ .

Therefore,

$$\begin{aligned} \pi(H_n) &= \int_{\mu} \mathcal{N}(H_n; \sum_{i \in [K]} e_i(\mathbb{R}^K) \otimes \left( \sum_{t \in [\omega_i n]} e_t(\mathbb{R}^{\omega_i n}) \otimes b_i^\top \right) \otimes \mu, I_K \otimes (\sigma^2 I_{\omega_i n} + \mathbf{1}_{\omega_i n} \mathbf{1}_{\omega_i n}^\top \sigma_{0,i}^2)) \mathcal{N}(\mu; \nu, \Sigma) d\mu \\ &= \mathcal{N}(H_n; \hat{\mu}, \hat{\Sigma}), \end{aligned}$$

where  $\hat{\mu} \in \mathbb{R}^n$ ,  $\hat{\Sigma} \in \mathbb{R}^{n \times n}$  with

$$\begin{aligned} \hat{\mu} &= \sum_{i \in [K]} e_i(\mathbb{R}^K) \otimes \left( \sum_{t \in [\omega_i n]} e_t(\mathbb{R}^{\omega_i n}) \otimes b_i^\top \right) \otimes \nu \\ \hat{\Sigma} &= I_K \otimes (\sigma^2 I_{\omega_i n} + \mathbf{1}_{\omega_i n} \mathbf{1}_{\omega_i n}^\top \sigma_{0,i}^2) \\ &+ \left[ \sum_{i \in [K]} e_i(\mathbb{R}^K) \otimes \left( \sum_{t \in [\omega_i n]} e_t(\mathbb{R}^{\omega_i n}) \otimes b_i^\top \right) \right] \Sigma \left[ \sum_{i \in [K]} e_i(\mathbb{R}^K) \otimes \left( \sum_{t \in [\omega_i n]} e_t(\mathbb{R}^{\omega_i n}) \otimes b_i^\top \right) \right]^\top. \end{aligned} \quad (32)$$

The covariance matrix  $\hat{\Sigma}$  seems complex but has a simple structure. The first term  $I_K \otimes (\sigma^2 I_{\omega_i n} + \mathbf{1}_{\omega_i n} \mathbf{1}_{\omega_i n}^\top \sigma_{0,i}^2)$  is the same as in the standard model. The remaining term accounts for the correlation between distinct arms  $(i, j)$ , and this correlation is of the form  $b_i^\top \Sigma b_j$ .

Now we are ready to compute  $\mathbb{E}[\hat{\mu}_{n,k}]$  for any arm  $k \in [K]$ : from (16) and (17),

$$\begin{aligned} \mathbb{E}[\hat{\mu}_{n,k}] &= \mathbb{E} \left[ \tilde{\sigma}_{n,k}^2 \left( \frac{\check{\mu}_n^\top b_k}{\sigma_{0,k}^2} + \frac{B_{n,k}}{\sigma^2} \right) \right] = \mathbb{E} \left[ \frac{\sigma^2 \sigma_{0,i}^2}{\sigma_{0,k}^2 \omega_k n + \sigma^2} \left( \frac{\check{\mu}_n^\top b_k}{\sigma_{0,k}^2} + \frac{B_{n,k}}{\sigma^2} \right) \right] \\ &= \frac{\sigma^2}{\sigma_{0,k}^2 \omega_k n + \sigma^2} \mathbb{E}[\check{\mu}_n^\top b_k] + \frac{\sigma_{0,k}^2}{\sigma_{0,k}^2 \omega_k n + \sigma^2} \mathbb{E}[B_{n,k}]. \end{aligned}$$

From (15),

$$\check{\mu}_n^\top b_k = \left( \nu^\top \Sigma^{-1} + \sum_{i \in [K]} \frac{B_{n,i}}{\sigma^2 + \omega_i n \sigma_{0,i}^2} b_i^\top \right) \check{\Sigma}_n b_k = \nu^\top \Sigma^{-1} \check{\Sigma}_n b_k + \sum_{i \in [K]} \frac{B_{n,i}}{\sigma^2 + \omega_i n \sigma_{0,i}^2} b_i^\top \check{\Sigma}_n b_k.$$

By linearity,

$$\mathbb{E}[\hat{\mu}_{n,k}] = \frac{\sigma^2}{\sigma_{0,k}^2 \omega_k n + \sigma^2} \left( \nu^\top \Sigma^{-1} \check{\Sigma}_n b_k + \sum_{i \in [K]} \frac{\mathbb{E}[B_{n,i}]}{\sigma^2 + \omega_i n \sigma_{0,i}^2} b_i^\top \check{\Sigma}_n b_k \right) + \frac{\sigma_{0,k}^2}{\sigma_{0,k}^2 \omega_k n + \sigma^2} \mathbb{E}[B_{n,k}].$$

From Equation (32),  $\mathbb{E}[B_{n,i}] = \omega_i n \nu^\top b_i$ . Therefore,

$$\mathbb{E}[\hat{\mu}_{n,k}] = \frac{\sigma^2}{\sigma_{0,k}^2 \omega_k n + \sigma^2} \left( \nu^\top \Sigma^{-1} \check{\Sigma}_n b_k + \sum_{i \in [K]} \frac{\omega_i n \nu^\top b_i}{\sigma^2 + \omega_i n \sigma_{0,i}^2} b_i^\top \check{\Sigma}_n b_k \right) + \frac{\sigma_{0,k}^2}{\sigma_{0,k}^2 \omega_k n + \sigma^2} \omega_k n \nu^\top b_k = \nu^\top b_k \quad (33)$$

Now we are ready to compute  $\mathbb{V}(\hat{\mu}_{n,i} - \hat{\mu}_{n,j})$  for any  $(i, j)$ . From (33),

$$\begin{aligned}
 \hat{\mu}_{n,i} - \hat{\mu}_{n,j} &= \frac{\sigma^2}{\sigma_{0,i}^2 \omega_i n + \sigma^2} \check{\mu}_n^\top b_i + \frac{\sigma_{0,i}^2}{\sigma_{0,i}^2 \omega_i n + \sigma^2} B_{n,i} - \frac{\sigma^2}{\sigma_{0,j}^2 \omega_j n + \sigma^2} \check{\mu}_n^\top b_j - \frac{\sigma_{0,j}^2}{\sigma_{0,i}^2 \omega_j n + \sigma^2} B_{n,j} \\
 &= \underbrace{\frac{\sigma^2}{\sigma_{0,i}^2 \omega_i n + \sigma^2} \nu^\top \Sigma^{-1} \check{\Sigma}_n b_i - \frac{\sigma^2}{\sigma_{0,j}^2 \omega_j n + \sigma^2} \nu^\top \Sigma^{-1} \check{\Sigma}_n b_j}_{\text{does not depend on observations}} \\
 &\quad + \underbrace{\frac{\sigma^2}{\sigma_{0,i}^2 \omega_i n + \sigma^2} \cdot \sum_{k \in [K]} \frac{B_{n,k}}{\omega_k n \sigma_{0,k}^2 + \sigma^2} b_k^\top \check{\Sigma}_n b_i}_{(1)} + \underbrace{\frac{-\sigma^2}{\sigma_{0,j}^2 \omega_j n + \sigma^2} \cdot \sum_{k \in [K]} \frac{B_{n,k}}{\omega_k n \sigma_{0,k}^2 + \sigma^2} b_k^\top \check{\Sigma}_n b_j}_{(2)} \\
 &\quad + \underbrace{\frac{\sigma_{0,j}^2}{\sigma_{0,i}^2 \omega_i n + \sigma^2} B_{n,i}}_{(3)} + \underbrace{\frac{-\sigma_{0,j}^2}{\sigma_{0,j}^2 \omega_j n + \sigma^2} B_{n,j}}_{(4)}.
 \end{aligned}$$

Since (1), (2), (3) and (4) are correlated,

$$\mathbb{V}(\hat{\mu}_{n,i} + \hat{\mu}_{n,j}) = \mathbb{V}((1) + (2) + (3) + (4)) = \sum_{i=1}^4 \mathbb{V}((i)) + \sum_{i=1}^4 \sum_{j=1, j \neq i}^4 2\text{cov}((i), (j)). \quad (34)$$

We now compute each term of (34):

$$\begin{aligned}
 \mathbb{V}((1)) &= \frac{\sigma^4}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)^2} \mathbb{V} \left( \sum_{k \in [K]} \frac{B_{n,k}}{\omega_k n \sigma_{0,k}^2 + \sigma^2} b_k^\top \check{\Sigma}_n b_i \right) \\
 &= \frac{\sigma^4}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)^2} \left( \sum_{k \in [K]} \frac{(b_k^\top \check{\Sigma}_n b_i)^2}{(\sigma^2 + \omega_k n \sigma_{0,k}^2)^2} \mathbb{V}(B_{n,k}) + \sum_{(k,k'), k \neq k'} \text{cov} \left( \frac{B_{n,k}}{\sigma^2 + \omega_k n \sigma_{0,k}^2} b_k^\top \check{\Sigma}_n b_i, \frac{B_{n,k'}}{\sigma^2 + \omega_{k'} n \sigma_{0,k'}^2} b_{k'}^\top \check{\Sigma}_n b_i \right) \right) \\
 &= \frac{\sigma^4}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)^2} \left( \sum_{k \in [K]} \frac{(b_k^\top \check{\Sigma}_n b_i)^2}{(\sigma^2 + \omega_k n \sigma_{0,k}^2)^2} \mathbb{V}(B_{n,k}) + \sum_{(k,k'), k \neq k'} \frac{(b_k^\top \check{\Sigma}_n b_i)(b_{k'}^\top \check{\Sigma}_n b_i)}{(\sigma^2 + \omega_k n \sigma_{0,k}^2)(\sigma^2 + \omega_{k'} n \sigma_{0,k'}^2)} \text{cov}(B_{n,k}, B_{n,k'}) \right) \\
 \mathbb{V}((2)) &= \frac{\sigma^4}{(\sigma_{0,j}^2 \omega_j n + \sigma^2)^2} \left( \sum_{k \in [K]} \frac{(b_k^\top \check{\Sigma}_n b_j)^2}{(\sigma^2 + \omega_k n \sigma_{0,k}^2)^2} \mathbb{V}(B_{n,k}) + \sum_{(k,k'), k \neq k'} \frac{(b_k^\top \check{\Sigma}_n b_i)(b_{k'}^\top \check{\Sigma}_n b_j)}{(\sigma^2 + \omega_k n \sigma_{0,k}^2)(\sigma^2 + \omega_{k'} n \sigma_{0,k'}^2)} \text{cov}(B_{n,k}, B_{n,k'}) \right), \\
 \mathbb{V}((3)) &= \frac{\sigma_{0,i}^4}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)^2} \mathbb{V}(B_{n,i}) \\
 \mathbb{V}((4)) &= \frac{\sigma_{0,j}^4}{(\sigma_{0,j}^2 \omega_j n + \sigma^2)^2} \mathbb{V}(B_{n,j}), \\
 \text{cov}((1), (2)) &= -\text{cov} \left( \frac{\sigma^2}{\sigma_{0,i}^2 \omega_i n + \sigma^2} \cdot \sum_{k \in [K]} \frac{B_{n,k}}{\omega_k n \sigma_{0,k}^2 + \sigma^2} b_k^\top \check{\Sigma}_n b_i, \frac{\sigma^2}{\sigma_{0,j}^2 \omega_j n + \sigma^2} \cdot \sum_{k \in [K]} \frac{B_{n,k}}{\omega_k n \sigma_{0,k}^2 + \sigma^2} b_k^\top \check{\Sigma}_n b_j \right) \\
 &= -\frac{\sigma^4}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)(\sigma_{0,j}^2 \omega_j n + \sigma^2)} \text{cov} \left( \sum_{k \in [K]} \frac{B_{n,k}}{\omega_k n \sigma_{0,k}^2 + \sigma^2} b_k^\top \check{\Sigma}_n b_i, \sum_{k \in [K]} \frac{B_{n,k}}{\omega_k n \sigma_{0,k}^2 + \sigma^2} b_k^\top \check{\Sigma}_n b_j \right) \\
 &= -\frac{\sigma^4}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)(\sigma_{0,j}^2 \omega_j n + \sigma^2)} \left[ \sum_{k \in [K]} \frac{(b_k^\top \check{\Sigma}_n b_i)(b_k^\top \check{\Sigma}_n b_j)}{(\sigma^2 + \omega_k n \sigma_{0,k}^2)} \mathbb{V}(B_{n,k}) \right. \\
 &\quad \left. + \sum_{(k,k'), k \neq k'} \frac{(b_k^\top \check{\Sigma}_n b_i)(b_{k'}^\top \check{\Sigma}_n b_j)}{(\sigma^2 + \omega_k n \sigma_{0,k}^2)(\sigma^2 + \omega_{k'} n \sigma_{0,k'}^2)} \text{cov}(B_{n,k}, B_{n,k'}) \right] \\
 \text{cov}((3), (4)) &= -\text{cov} \left( \frac{\sigma_{0,i}^2}{\sigma_{0,i}^2 \omega_i n + \sigma^2} B_{n,i}, \frac{\sigma_{0,j}^2}{\sigma_{0,j}^2 \omega_j n + \sigma^2} B_{n,j} \right) \\
 &= -\frac{\sigma_{0,i}^2 \sigma_{0,j}^2}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)(\sigma_{0,j}^2 \omega_j n + \sigma^2)} \text{cov}(B_{n,i}, B_{n,j}) \\
 \text{cov}((1), (4)) &= -\text{cov} \left( \frac{\sigma^2}{\sigma_{0,i}^2 \omega_i n + \sigma^2} \cdot \sum_{k \in [K]} \frac{B_{n,k}}{\omega_k n \sigma_{0,k}^2 + \sigma^2} b_k^\top \check{\Sigma}_n b_i, \frac{\sigma_{0,j}^2}{\sigma_{0,j}^2 \omega_j n + \sigma^2} B_{n,j} \right) \\
 &= -\frac{\sigma^2 \sigma_{0,j}^2}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)(\sigma_{0,j}^2 \omega_j n + \sigma^2)} \text{cov} \left( \sum_{k \in [K]} \frac{B_{n,k}}{\sigma^2 + \omega_k n \sigma_{0,k}^2} b_k^\top \check{\Sigma}_n b_i, B_{n,j} \right) \\
 &= -\frac{\sigma^2 \sigma_{0,j}^2}{(\sigma_{0,i}^2 \omega_i n + \sigma^2)(\sigma_{0,j}^2 \omega_j n + \sigma^2)} \left[ \sum_{k \in [K] \setminus \{j\}} \frac{b_k^\top \check{\Sigma}_n b_i}{\sigma_{0,k}^2 \omega_k n + \sigma^2} + \frac{b_j^\top \check{\Sigma}_n b_i}{\sigma_{0,j}^2 \omega_j n + \sigma^2} \mathbb{V}(B_{n,j}) \right].
 \end{aligned}$$

The remaining terms are obtained by symmetry. Finally, for any  $(i, j)$  :

$$\begin{aligned}\mathbb{V}(B_{n,i}) &= \mathbb{V}\left(\sum_{t \in \mathcal{A}_i} Y_t\right) = \omega_i n \sigma^2 + \omega_i^2 n^2 (\sigma_{0,i}^2 + b_i^\top \Sigma b_i) \\ \text{cov}(B_{n,i}, B_{n,j}) &= \text{cov}\left(\sum_{t \in \mathcal{A}_i} Y_t, \sum_{t \in \mathcal{A}_j} Y_t\right) = \sum_{t \in \mathcal{A}_i} \sum_{t \in \mathcal{A}_j} \text{cov}(Y_t, Y_t) = \omega_i \omega_j n^2 b_i^\top \Sigma b_j.\end{aligned}$$

*Remark C.11* (Computing the upper bound for hierarchical bandit with Theorem 4.2). The reader can wonder why transforming the hierarchical model into a linear model thanks to (10), and plug directly the transformed prior and actions to the linear upper bound (Theorem 4.2). While this is what we do to optimize numerically the bound, it is challenging to give explicit terms with this method. In fact, it would yield to the following upper bound,

$$\mathcal{P}_n \leq \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{1}{\sqrt{1 + \frac{\|\bar{b}_i - \bar{b}_j\|_{\text{Cov}_n}^2}{\|b_i - b_j\|_{\bar{\Sigma}_n}^2}}} e^{-\frac{(\bar{\nu}^\top \bar{b}_i - \bar{\nu}^\top \bar{b}_j)^2}{2(\|b_i - b_j\|_{\bar{\Sigma}_n}^2)}} = \sum_{\substack{i,j \in [K] \\ i \neq j}} \frac{1}{\sqrt{1 + \frac{\|\bar{b}_i - \bar{b}_j\|_{\text{Cov}_n}^2}{\|b_i - b_j\|_{\bar{\Sigma}_n}^2}}} e^{-\frac{(\nu^\top b_i - \nu^\top b_j)^2}{2(\|b_i - b_j\|_{\bar{\Sigma} + \sigma_{0,i}^2 + \sigma_{0,j}^2}^2)}},$$

where

$$\bar{\Sigma}_n = \left( \bar{\Sigma} + \frac{1}{\sigma^2} \sum_{i \in [K]} \omega_i n \bar{b}_i \bar{b}_i^\top \right)^{-1}, \quad \text{Cov}_n = \frac{1}{\sigma^4} \bar{\Sigma}_n \left( \sum_{i \in [K]} \omega_i n (\sigma^2 + \bar{b}_i \bar{\Sigma} \bar{b}_i) \bar{b}_i \bar{b}_i^\top + \sum_{\substack{i,j \in [K] \\ i \neq j}} \bar{b}_i \bar{\Sigma} \bar{b}_j \omega_i \omega_j n^2 \bar{b}_i \bar{b}_j^\top \right) \bar{\Sigma}_n.$$

However, computing  $\|\bar{b}_i - \bar{b}_j\|_{\text{Cov}_n}^2$  and  $\|b_i - b_j\|_{\bar{\Sigma}_n}^2$  is computationally challenging because it requires first to compute  $\bar{\Sigma}$  with block-matrix inversion, then to recover the marginal and posterior covariances  $\tilde{\sigma}_{n,i}^2$ ,  $\tilde{\sigma}_{n,i}^2$  and  $\hat{\sigma}_{n,i}^2$  from (15), (16) and (17). □

## C.5 Proof for Prior Misspecification

*Proof of Lemma 4.5.* Let's denote as  $(\tilde{\mu}_{n,i}, \tilde{\sigma}_{n,i}^2)$  the posterior mean and variance of arm  $i$  when using the misspecified priors  $\mathcal{N}(\tilde{\mu}_{0,i}, \tilde{\sigma}_0^2)$ , that is,

$$\tilde{\sigma}_{n,i}^{-2} = \tilde{\sigma}_0^{-2} + n_i \sigma^{-2}, \quad \tilde{\mu}_{n,i} = \tilde{\sigma}_{n,i}^2 (\tilde{\sigma}_0^{-2} \mu_{0,i} + \sigma^{-2} B_{n,i}),$$

where we recall  $B_{n,i} = \sum_{t \in \mathcal{T}_i} Y_t$ , where  $\mathcal{T}_i = \{t \in [n] : A_t = i\}$  and  $|\mathcal{T}_i| = n_i$ . The user's decision  $J_n$  is defined as  $J_n = \arg \max_{i \in [K]} \tilde{\mu}_{n,i}$ . Then, for a fixed bandit  $\theta \in \mathbb{R}^K$ ,

$$\begin{aligned}\mathbb{P}(J_n \neq i_*(\theta) \mid \theta) &\leq \mathbb{P}(\exists j \neq i_*(\theta), \tilde{\mu}_{n,j} > \tilde{\mu}_{n,i_*(\theta)} \mid \theta) \\ &\leq \sum_{j \in [K]} \mathbb{P}(\tilde{\mu}_{n,j} > \tilde{\mu}_{n,i_*(\theta)} \mid \theta).\end{aligned}$$

Then, using the misspecified posterior means formulas and the fact that we perform uniform allocations,

$$\mathbb{P}(\tilde{\mu}_{n,j} > \tilde{\mu}_{n,i_*(\theta)} \mid \theta) = \mathbb{P}\left(\sum_{t \in \mathcal{T}_i} Y_t - \sum_{t \in \mathcal{T}_{i_*(\theta)}} Y_t > \frac{\tilde{\sigma}_{0,i}^2}{\sigma^2} \mu_{0,i_*(\theta)} - \frac{\tilde{\sigma}_{0,i}^2}{\sigma^2} \mu_{0,i} \mid \theta\right).$$

Using Lemma 3 from [Atsidakou et al. \(2022\)](#), we can upper bound this latter probability as

$$\mathbb{P}(\tilde{\mu}_{n,j} > \tilde{\mu}_{n,i_*(\theta)} \mid \theta) \leq \exp\left(-\frac{1}{4\sigma_0^2} \left(\frac{\tilde{\sigma}_{0,i}^2}{\sigma^2} \mu_{0,i} - \frac{\tilde{\sigma}_{0,i_*(\theta)}^2}{\sigma^2} \mu_{0,i_*(\theta)}\right)^2 - \frac{1}{2\sigma_0^2} \left(\frac{\tilde{\sigma}_{0,i}^2}{\sigma^2} \mu_{0,i} - \frac{\tilde{\sigma}_{0,i_*(\theta)}^2}{\sigma^2} \mu_{0,i_*(\theta)}\right) (\theta_i - \theta_{i_*(\theta)})\right).$$

Considering all the possible  $i_*(\theta) \in [K]$  and integrating with respect to bandit instances  $\theta_i \sim \mathcal{N}(\mu_{0,i}, \sigma_0^2)$  with Lemma C.5 gives the result. □

## D ADDITIONAL EXPERIMENTS

In this section, we consider the two following settings (we set  $\sigma = 1$  for both settings):

**Random setting.** For MAB, each  $\mu_{0,i}$  is sampled from  $\mathcal{U}([0, 1])$  and  $\sigma_{0,i}$ 's are evenly spaced between 0.1 and 0.5. For linear bandits, each  $\mu_{0,i}$  is sampled from  $\mathcal{U}([0, 1])$  and  $\Sigma_0$  is a diagonal matrix whose entries are evenly spaced between 0.1 and 0.5. For hierarchical bandits,  $\nu_i \sim \mathcal{U}([-1, 1])$  and  $\Sigma$  and  $\Sigma_0$  are diagonal with entries evenly spaced in  $[0.1, 0.5]$ .

**Fixed setting.** For MAB, each  $\mu_{0,i}$ 's are evenly spaced between 0 and 1, and  $\sigma_{0,i}$ 's are evenly spaced between 0.1 and 0.5. For linear bandits, each  $\mu_0$  is set flat,  $\mu_0 = (1, \dots, 1)$  and  $\Sigma_0$  is a diagonal matrix whose entries are evenly spaced between 0.1 and 0.5. For hierarchical bandits,  $\nu_i$ 's are evenly spaced between  $-1$  and  $1$ .  $\Sigma$  and  $\Sigma_0$  are diagonal with entries evenly spaced in  $[0.1, 0.5]$ .

Note that the random setting corresponds to the setting used for Figure 2a.

### D.1 Prior Misspecification

We conduct additional synthetic experiments in the standard Gaussian model (5), where we instantiate PI-BAI with different allocations (uniform, optimized, G-optimal and warm-up allocations with TS) with misspecified prior. We set all the parameters as in Section 6, that is,  $\sigma = 1$ ,  $K = 10$ ,  $\sigma_{0,i}$  evenly spaced between 0.1 and 0.5 and  $\mu_{0,i} \sim \mathcal{U}([0, 1])$ . Then construct the misspecified prior mean as  $\tilde{\mu}_0 = \mu_0 + \alpha u$ , where  $u \sim \mathcal{N}(0, I_K)$ . For the variance misspecification, we set the true prior variance to  $\sigma_{0,i} = 0.3$  for all arms, then construct the misspecified prior as  $\tilde{\sigma}_{0,i} = \sigma_{0,i} + \alpha$ . Results are shown in Figure 5 (first row for prior misspecification, second row for variance misspecification), where we let  $\alpha$  vary from 0 (no misspecification) to 0.5. Overall, it shows that the effect of prior misspecification vanishes as  $n$  increases, as expected (see Lemma 4.5), except for the TS warm-up. This is because TS is more aggressive, that is, it does not explore suboptimal arms based on his prior belief, yielding to long-term consequences when sticking to the same allocation rule.

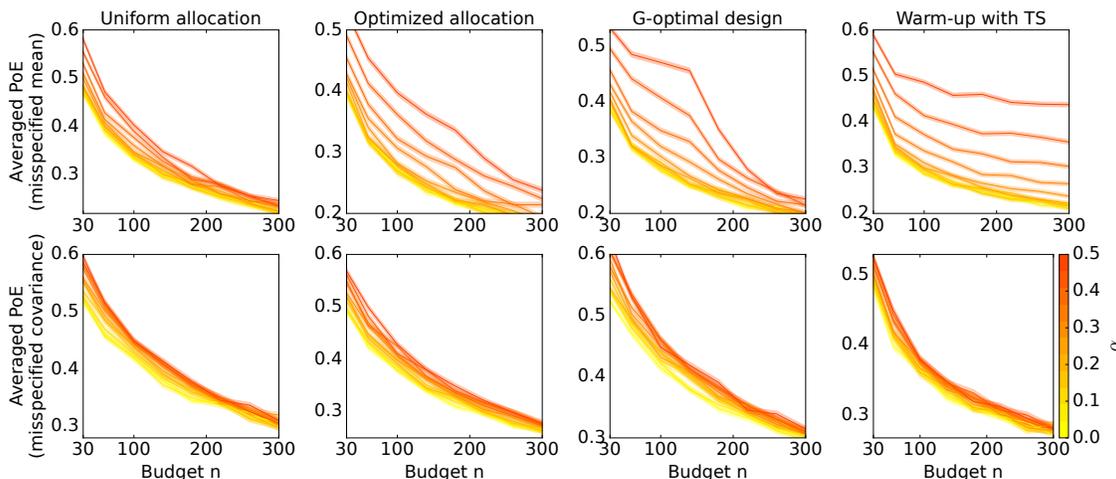


Figure 5: Average PoE of PI-BAI with varying budget and allocations (uniform, optimized, G-optimal and TS warmed-up weights) for different level of mean misspecification (first row) and variance misspecification (second row).

### D.2 Additional Setting

We provide empirical results for the Fixed setting. All remaining parameters  $K$ ,  $d$  and  $L$  are the same as in Figure 2a of Section 6. Figure 6 shows that our methods outperform existing baselines in this setting.

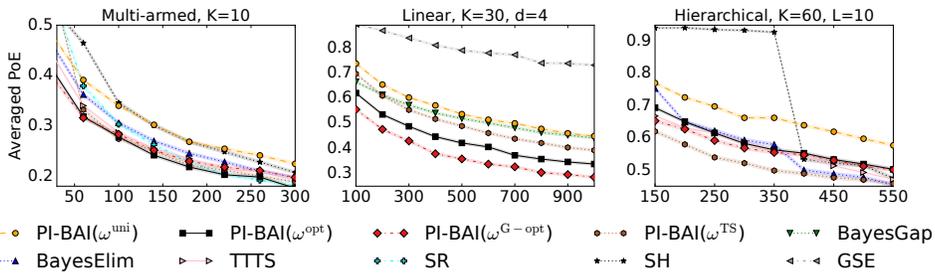


Figure 6: Average PoE with varying budgets for the Fixed setting: for MAB the  $\mu_{0,i}$ 's are evenly spaced between 0 and 1 and  $\sigma_{0,i}$ 's are evenly spaced between 0.1 and 0.5. For linear bandits, each  $\mu_{0,i}$  is sampled from  $\mathcal{U}([0, 1])$  and  $\Sigma_0$  is a diagonal matrix whose entries are evenly spaced between 0.1 and 0.5. For hierarchical bandits,  $\nu_i \sim \mathcal{U}([-1, 1])$  and  $\Sigma$  and  $\Sigma_0$  are diagonal with entries evenly spaced in  $[0.1^2, 0.5^2]$ .

### D.3 MovieLens Experiments

We provide more information on our MovieLens experiments in Figure 2a. The MovieLens dataset provides 1 million ratings ratings given by 6040 users to 3952 movies, forming a sparse rating matrix  $M$  of size  $6,040 \times 3,952$ . To learn a prior suitable for our algorithm, we complete  $M$  using alternating least squares with rank  $d = 5$  (low-rank factorization), resulting in the decomposition  $M = U^T V$ , where rows  $U_i$  and  $V_j$  represent user  $i$  and movie  $j$ , respectively. We then construct a linear Gaussian prior (see Section 4.2) by setting the mean  $\mu_0 = \frac{1}{3952} \sum_{i \in [3952]} V_j$  and the covariance  $\Sigma_0 = \text{diag}(v)$  where  $v \in \mathbb{R}^d$  is the empirical variance of  $V_j$  for  $j \in [3952]$  along each dimension. We use a subset of  $K = 100$  randomly picked movies. In our experiments, the bandit instances aren't sampled from a Gaussian prior, but we learned a Gaussian prior that we employ in our algorithm. Despite this mismatch, PI-BAI with such prior performs very well. Of course, the prior is very informative as we used the whole dataset to learn it. But still, the bandit instances are not sampled from it. All results are averaged over  $10^4$  rounds.

### D.4 Logistic Bandits

We consider the Random and Fixed setting for  $K = 30$  arms and  $d \in \{3, 4\}$ . As explained in Appendix B, We use a Laplace approximation to estimate the posterior means. For the warmed-up allocations, we use TS with Bernoulli-logistic likelihood (Chapelle and Li, 2012). The G-optimal design weights are obtained as if we face a linear Gaussian bandit.

Figure 7 shows that the generalization of PI-BAI with G-optimal design allocations on has good performances beyond linear settings.

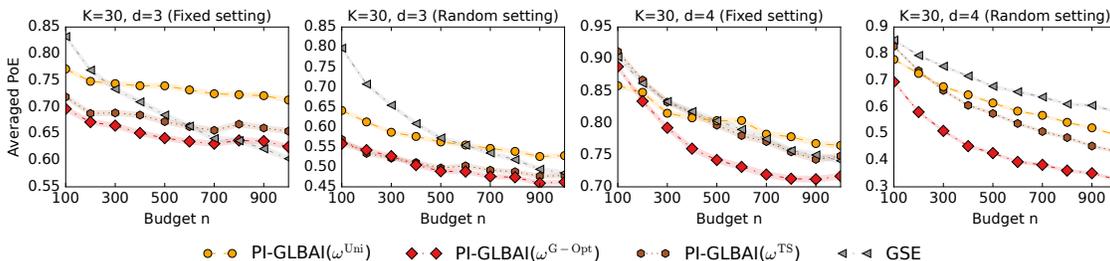


Figure 7: Average PoE with varying budgets for the Fixed setting and the Random setting in the GLB framework.

### D.5 Choice of Baselines

**A remark on TTS.** Top two sampling algorithms is a family of algorithms that is known to have good performances in BAI. In Section 6, we used TS-TCI with  $\beta = 0.5$  from Jourdan et al. (2022) and denoted it as TTS for sake of notation simplicity.

**Influence of elimination.** We empirically compare the influence of using elimination on top of our methods. The elimination procedure is the same as the one used in [Atsidakou et al. \(2022\)](#). There are  $\lfloor \log_2(K) \rfloor$  rounds, and each lasts  $\lfloor \frac{n}{R} \rfloor$  steps. At each round, we pull each remaining arm  $i$   $\lfloor \frac{\omega_i n}{R} \rfloor$  times. At the end of the round, half of arms are eliminated. These correspond to the arms that have the least posterior mean reward (so  $\hat{\mu}_{n,i}$  in the MAB setting). The allocation  $\omega$  is then normalized to allocate more budget to remaining arms. Note that we draft all observations at the end of each round, as it is the case in all methods that rely on successive halvings ([Atsidakou et al., 2022](#); [Azizi et al., 2021](#); [Karnin et al., 2013](#)). Figure 8 shows that using elimination does not give better performances, and hence we chose to not add these baselines in Section 6.

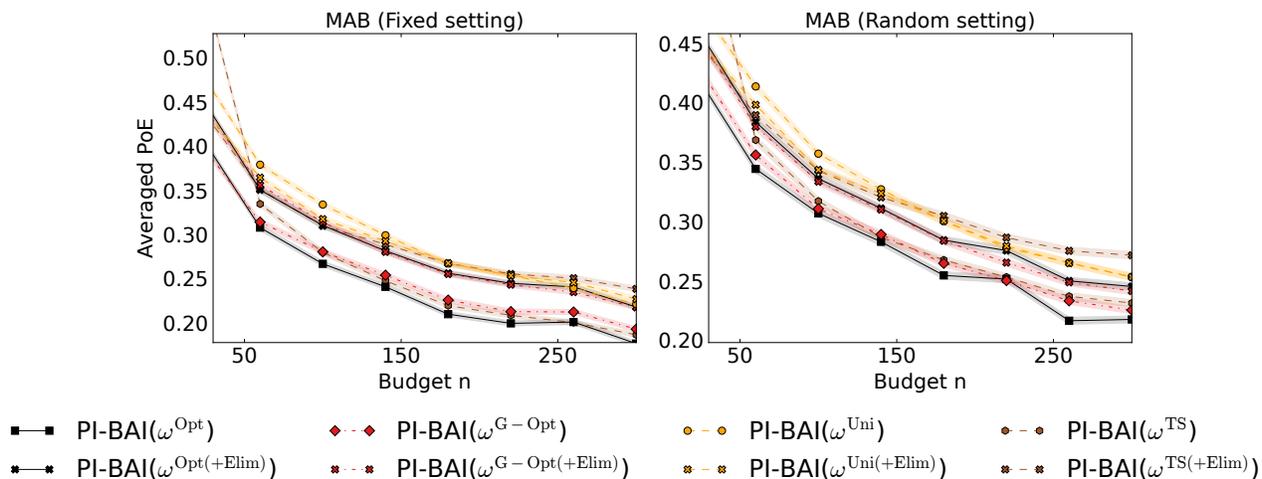


Figure 8: Average PoE of PI-BAI instantiated with different weights with or without elimination in the MAB setting.

## D.6 Simple Regret

Figure 9 compares the performances of our methods based on the Bayesian simple regret (see discussion in Appendix A). Overall, it shows that our method outperform existing baselines in terms of simple regret in the Random and Fixed settings. We do not compare our method to [Komiyama et al. \(2023\)](#), since it is designed for Bernoulli bandits.

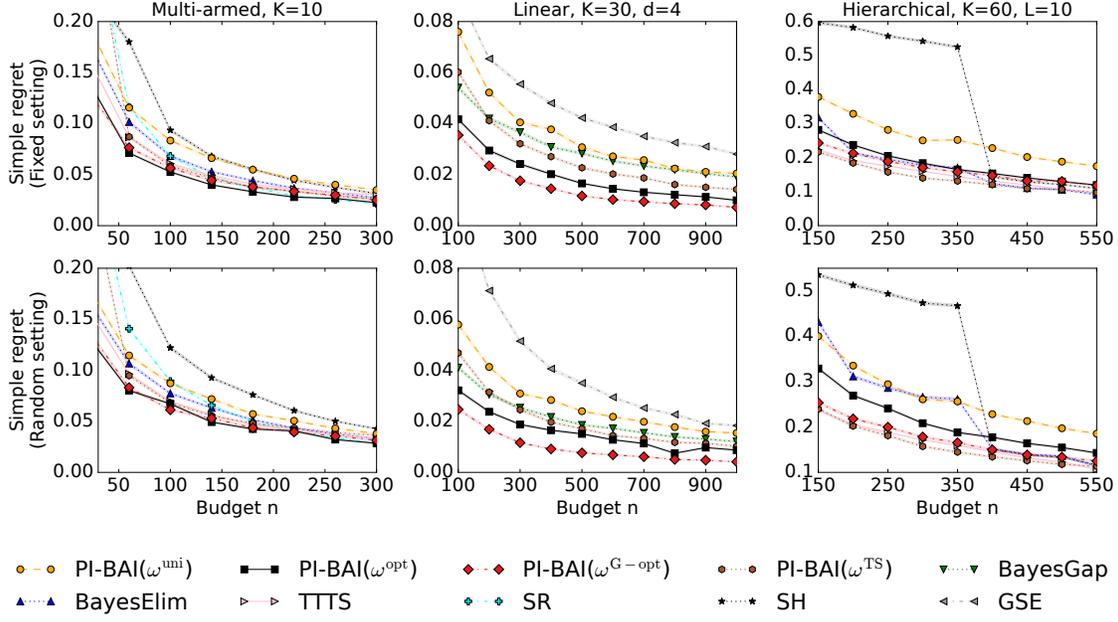
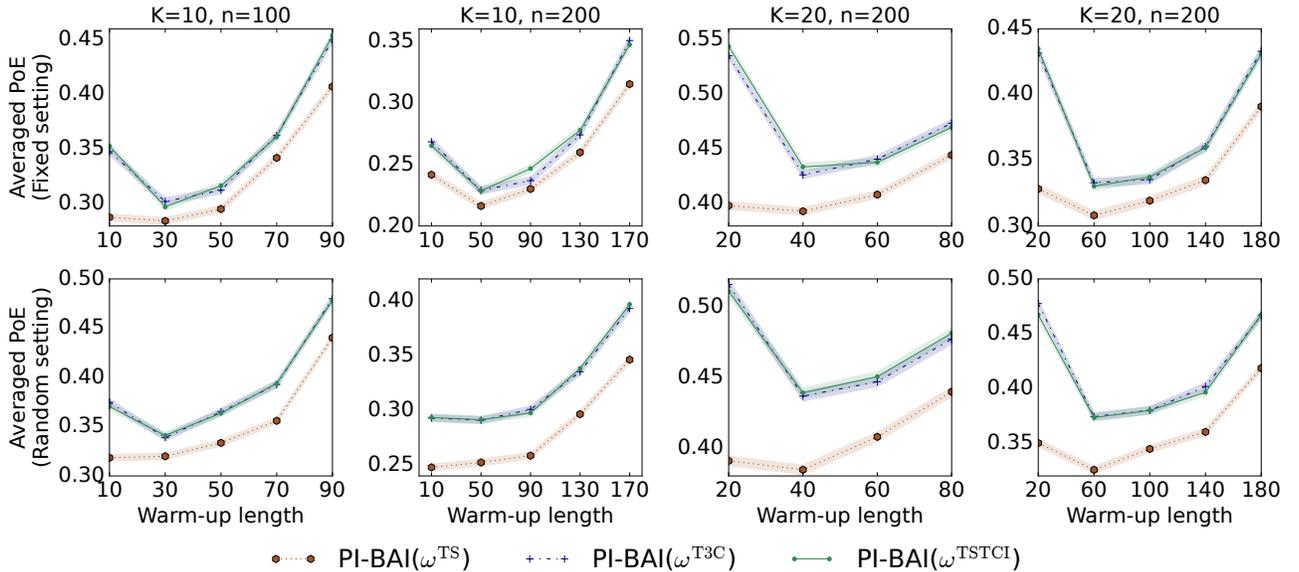


Figure 9: Average simple regret with varying budgets for fixed and randomized settings.

## D.7 Hyperparameters

**Warm-up length  $n_w$ .** We try different values of warm-up length  $n_w$  for our warm-up policies. We emphasize that methods based on TTS require  $n_w > K$  because each arm has to be pulled at the beginning. Figure 10 suggests picking  $n_w = 2K$  for the warm-up with T3C and TSTCI, and  $n_w = K$  for the warm-up with TS.


 Figure 10: Average PoE of PI-BAI instantiated with different warm-up policies for different warm-up lengths  $n_w$ .

**Choice of warm-up policy.** We evaluate different warm-up policies, *TS* and two Top-Two algorithms, TSTCI and T3C from Jourdan et al. (2022). The experiments shown in Figure 11 are run in the same setting as in Section 6, with  $K = 10$  arms in the MAB setting, and with  $K = 60$  and  $d = 4$  in the hierarchical setting. Figure 11 suggests to pick TS as a warm-up policy for the MAB setting and meTS of Aouali et al. (2023) for the hierarchical setting.

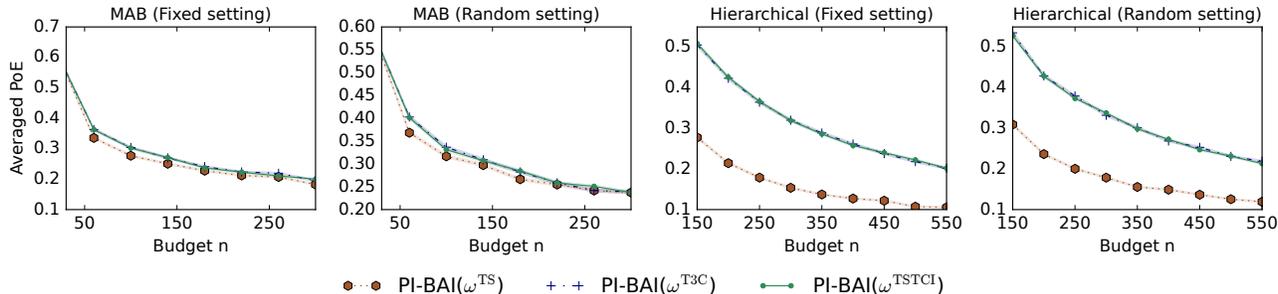


Figure 11: Average PoE of PI-BAI instantiated with different warm-up policies.

**Mixture parameter  $\alpha$ .** We discuss the choice of the mixture parameter  $\alpha$  in the settings presented in Section 6 and Appendix D.2. We recall that in the "Random setting", each  $\mu_{0,i}$  is sampled from  $\mathcal{U}([0, 1])$ , whereas in the "fixed setting",  $\mu_{0,i}$  are evenly spaced between 0 and 1. In both settings,  $\sigma_{0,i}$  are evenly spaced from 0.1 to 0.5. We recall that we use the heuristic  $\alpha\omega_i^{\text{opt}} + (1 - \alpha)\frac{\mu_{0,i}\sigma_{0,i}}{\sum_{k \in [K]} \mu_{0,k}\sigma_{0,k}}$  in our experiments. Figure 12 shows that for the fixed setting, adding the vector  $\frac{\mu_{0,i}\sigma_{0,i}}{\sum_{k \in [K]} \mu_{0,k}\sigma_{0,k}}$  helps improve the performances. This is not necessarily the case in the random setting. We find that the best performance is reached at  $\alpha \approx 0.5$ .

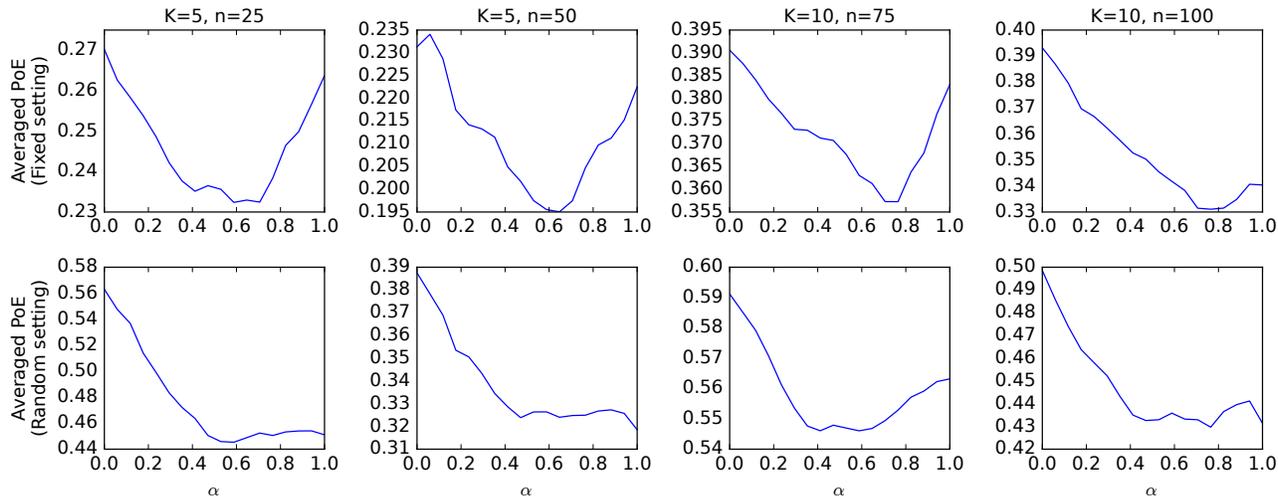


Figure 12: Average PoE of PI-BAI( $\omega^{\text{opt}}$ ) for different mixture parameter  $\alpha$ .

### D.8 Confidence Intervals on Sampled Instances

We show different type of confidence intervals in the Random and Fixed settings. In the first row of Figure 13, for each instance, we repeat the experiments 100 times on a same bandit environment to get a probability of error. Then, we sample 1000 different bandit instances and show one standard deviation of the resulting probability of error. In the second row of Figure 13, we plot the PoE of each method subtracted by the PoE of the method having the least PoE in each setting (PI-BAI( $\omega^{\text{opt}}$ ) in MAB, PI-BAI( $\omega^{\text{G-opt}}$ ) in linear bandits and PI-BAI( $\omega^{\text{TS}}$ ) in hierarchical bandits).

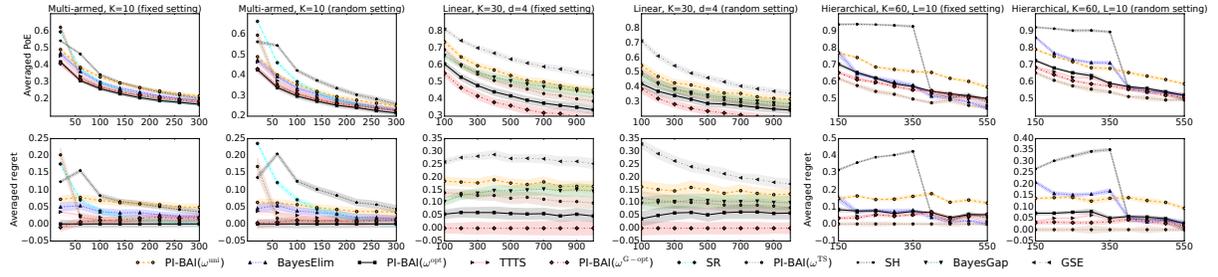


Figure 13: PoE of several methods (first row) and PoE of each method subtracted by the PoE of the most performing method in each setting (second row). For each instance, we repeat the experiments 100 times and we average the results over 1000 instances. The confidence intervals show one standard deviation.

## E COMPUTE RESSOURCES

All experiments have been run locally on a 2021 Mac Book Pro with 8 CPUs.