Thompson Sampling with Information Relaxation Penalties*

Seungki Min Costis Maglaras Ciamac C. Moallemi

Initial Version: November 2018
Current Revision: February 12, 2019

Abstract

We consider a finite time horizon multi-armed bandit (MAB) problem in a Bayesian framework, for which we develop a general set of control policies that leverage ideas from information relaxations of stochastic dynamic optimization problems. In crude terms, an information relaxation allows the decision maker (DM) to have access to the future (unknown) rewards and incorporate them in her optimization problem to pick an action at time t, but penalizes the decision maker for using this information. In our setting, the future rewards allow the DM to better estimate the unknown mean reward parameters of the multiple arms, and optimize her sequence of actions. By picking different information penalties, the DM can construct a family of policies of increasing complexity that, for example, include Thompson Sampling and the true optimal (but intractable) policy as special cases.

We systematically develop this framework of *information relaxation sampling*, propose an intuitive family of control policies for our motivating finite time horizon Bayesian MAB problem, and prove associated structural results and performance bounds. Numerical experiments suggest that this new class of policies performs well, in particular in settings where the finite time horizon introduces significant tension in the problem. Finally, inspired by the finite time horizon Gittins index, we propose an index policy that builds on our framework that particularly outperforms to the state-of-the-art algorithms in our numerical experiments.

1. Introduction

Dating back to the earliest work (Bradt et al., 1956; Gittins, 1979), multi-armed bandit (MAB) problems have been considered within a Bayesian framework, in which the unknown parameters are modeled as random variables drawn from a known prior distribution. In this setting, the problem can be viewed as a Markov decision process (MDP) with state that is an information state describing the beliefs of unknown parameters that evolves stochastically upon each play of an arm according to Bayes' rule.

Under the objective of expected performance, where the expectation is taken with re-

*Graduate School of Business, Columbia University; email: {smin20, c.maglaras, ciamac}@gsb.columbia.edu

spect to the prior distribution over unknown parameters, the (Bayesian) optimal policy is characterized by Bellman equations immediately following from the MDP formulation. In the discounted infinite-horizon setting, the celebrated Gittins index (Gittins, 1979) determines an optimal policy, despite the fact that its computation is still challenging. In the non-discounted finite-horizon setting, which we consider, the problem becomes more difficult (Berry and Fristedt, 1985), and except for some special cases, the Bellman equations are neither analytically nor numerically tractable, due to the curse of dimensionality. In this paper, we focus on the determination of the optimal policy (OPT) as an ideal goal that can be tackled by dynamic programming (DP).

We introduce the *information relaxation* framework (Brown et al., 2010), a recently developed technique that provides a systemic way of obtaining the performance bounds on the optimal policy. It is common in multi-period stochastic DP problems to consider admissible policies that are required to make decisions based only on the previously revealed information. In our framework, we consider the non-anticipativity as a constraint imposed on the policy space that can be relaxed, as in a usual Lagrangian relaxation. Under such a relaxation, the decision maker (DM) is allowed to access to the future information and is asked to solve an optimization problem so as to maximize her total reward, in the presence penalties that punish the violation of the non-anticipativity. When the penalties satisfy a condition (dual feasibility, formally defined in §3), the expected value of maximal reward adjusted by the penalties provides an upper bound of the expected performance of the (non-anticipating) optimal policy.

The idea of relaxing the non-anticipativity constraint has been studied over time in the different contexts (Rockafellar and Wets, 1991; Davis and Karatzas, 1994; Rogers, 2002; Haugh and Kogan, 2004), and later formulated as a formal framework by Brown et al. (2010), upon which our methodology is developed. This framework has been applied across a variety of applications including optimal stopping problems (Desai et al., 2012), linear-quadratic control (Haugh and Lim, 2012), dynamic portfolio execution (Haugh and Wang, 2014) and others (see Brown and Haugh (2017)).

Our contribution is to apply the information relaxation techniques to the finite-horizon stochastic MAB problem exploiting the structures of Bayesian learning process. In particular:

1. we propose a series of information relaxations and penalties with increasing complexity;
2. we systematically obtain the upper bounds on the best achievable expected performance that are in trade-off between tightness and computational complexity;
3. and, we obtain the associated (randomized) policies that generalize Thompson Sampling (TS) in the finite-horizon setting.

In our framework, which we call *information relaxation sampling*, each of penalty functions (and information relaxations) determines one policy and one performance bound given a particular problem instance specified by the time horizon and prior belief. As a base case for our algorithms, we have TS (Thompson, 1933) and the conventional regret benchmark that has been popularized for Bayesian regret analysis since Lai and Robbins (1985). On the other extreme, the optimal policy OPT and its expected performance follow from the "ideal" penalty which is intractable to specify. By picking increasingly strict information penalties, we can improve the policy and the associated bound between the two extremes of TS and Opt.

As an illustrating example, one of our algorithms, Irs.FH, provides a very simple modification of TS that takes into account the length of the time horizon T. Recalling that TS makes a decision based on sampled parameters from the posterior distribution in each epoch, we focus on the fact that knowing the parameters is as informative as having an infinite number of future reward observations in terms of the best arm identification. We let the policy, say $\pi^{\text{Irs.FH}}$, to make a decision based on the future Bayesian estimates, updated with only $T-1$ future reward realizations for each arm, where the rewards are randomly generated based on the posterior belief at the moment. When $T = 1$ (equivalently, at the last decision epoch), such a policy takes a myopically best action based only on the current estimates, which is indeed an optimal decision, whereas TS would still explore unnecessarily. While keeping the recursive structure in the sequential decision making process of TS, it naturally performs less exploration than TS as the remaining time horizon diminishes.

Beyond this, we propose other algorithms that more explicitly quantify the benefit of exploration and more explicitly trade-off exploration versus exploitation, at the cost of additional computational complexity. As we increase complexity, we achieve policies that improve performance, and separately provide tighter tractable computational upper bounds on the expected performance of any policy for a particular problem instance.

2. Notation and Preliminaries

Problem. We consider a classical stochastic MAB problem with K *independent arms* and *finite-horizon* T. At each decision epoch $t = 1, \cdots, T$, the decision maker (DM) pulls an arm $a_t \in \mathcal{A} \triangleq \{1, \cdots, K\}$ and earns a *stochastic reward* associated with arm a_t. More formally, the reward from n^{th} pull of arm a is denoted by $R_{a,n}$ which is independently drawn from unknown distribution $\mathcal{R}_a(\theta_a)$, where $\theta_a \in \Theta_a$ is the *parameter* associated with arm a. We also have a prior distribution $\mathcal{P}_a(y_a)$ over unknown parameter θ_a, where $y_a \in \mathcal{Y}_a$, which we call *belief*, is a hyperparameter describing the prior distribution:

$$\theta_a \sim \mathcal{P}_a(y_a), \quad R_{a,n}|\theta_a \sim \mathcal{R}_a(\theta_a), \quad \forall n \in [T], \quad \forall a \in \mathcal{A}. \tag{1}$$

We define two mean reward functions $\mu_a(\theta_a) \triangleq \mathbb{E}[R_{a,n}|\theta_a]$ and $\bar{\mu}_a(y_a) \triangleq \mathbb{E}_{\theta_a \sim \mathcal{P}_a(y_a)}[\mu_a(\theta_a)]$ as a function of unknown parameter θ_a and prior belief y_a respectively. Through out the paper, we assume that the rewards are absolutely integrable over the prior distribution: i.e., $\mathbb{E}[|R_{a,n}|] < \infty$ or more explicitly, $\mathbb{E}_{\theta_a \sim \mathcal{P}_a(y_a), r \sim \mathcal{R}_a(\theta_a)}[|r|] < \infty$ for all $a \in \mathcal{A}$.

For brevity, we denote $\boldsymbol{\theta} \triangleq (\theta_1, \cdots, \theta_K) \in \Theta$ and $\mathbf{y} \triangleq (y_1, \cdots, y_K) \in \mathcal{Y}$ be the vector of parameters and beliefs across arms, respectively. We additionally define an *outcome* ω as a combination of the parameters and all future reward realizations that incorporates all uncertainties in the environment that the DM encounters:

$$\omega \triangleq \Big(\boldsymbol{\theta}, (R_{a,n})_{a \in \mathcal{A}, n \in [T]}\Big) \ \sim \ \mathcal{I}(T, \mathbf{y}) \tag{2}$$

where $\mathcal{I}(T, \mathbf{y})$ represents the distribution of outcome.

Policy. Given an action sequence up to time t, $\mathbf{a}_{1:t} \triangleq (a_1, \cdots, a_t) \in \mathcal{A}^t$, define the number of pulls $n_t(\mathbf{a}_{1:t}, a) \triangleq \sum_{s=1}^{t} \mathbf{1}\{a_s = a\}$ for each arm a, and the corresponding reward realization $r_t(\mathbf{a}_{1:t}; \omega) \triangleq R_{a_t, n_t(\mathbf{a}_{1:t}, a_t)}$. The *natural filtration* $\mathcal{F}_t(\mathbf{a}_{1:t}; \omega) \triangleq \sigma\left(T, \mathbf{y}, (a_s, r_s(\mathbf{a}_{1:s}; \omega))_{s\in[t]}\right)$ encodes the observations revealed up to time t (inclusive).

Let $\mathbf{a}^{\pi}_{1:t}$ be the action sequence taken by a policy π. The (Bayesian) *performance* of a policy π is defined as the expected total reward over the randomness associated with the outcome, i.e.,

$$V(\pi, T, \mathbf{y}) \triangleq \mathbb{E}_{\omega\sim\mathcal{I}(T,\mathbf{y})}\left[\sum_{t=1}^{T} r_t(\mathbf{a}^{\pi}_{1:t}; \omega)\right]. \tag{3}$$

A policy π is called *non-anticipating* if its every action a^{π}_t is $\mathcal{F}_{t-1}$-measurable, and we define $\Pi_{\mathbb{F}}$ be a set of all non-anticipating policies, including randomized ones.

MDP formulation. We assume that we are equipped with a *Bayesian update function* $\mathcal{U}_a : \mathcal{Y}_a \times \mathbb{R} \mapsto \mathcal{Y}_a$ so that after observing $R_{a,1} = r$ from an arm a, the belief is updated from y_a to $\mathcal{U}_a(y_a, r)$ according to Bayes' rule. We will often use $\mathcal{U} : \mathcal{Y} \times \mathcal{A} \times \mathbb{R} \mapsto \mathcal{Y}$ to describe the updating of the entire belief vector $\mathbf{y}$; i.e., after observing $R_{a,1} = r$ from some arm a, the belief vector is updated from $\mathbf{y}$ to $\mathcal{U}(\mathbf{y}, a, r)$ where only the a^{th} component is updated in this step.

In a Bayesian framework, the MAB problem has a recursive structure. Given a time horizon T and prior belief $\mathbf{y}$, suppose the DM had just earned r by pulling an arm a at time $t = 1$. The remaining problem for the DM is equivalent to a problem with time horizon $T - 1$ and prior belief $\mathcal{U}(\mathbf{y}, a, r)$. We further know the (unconditional) distribution of what the DM will observe when pulling an arm a, a doubly stochastic random variable, and we denote it by $\mathcal{R}_a(\mathcal{P}_a(y_a))$. Following from this Markovian structure, we obtain the Bellman equations for the MAB problem:

$$Q^*(T, \mathbf{y}, a) \triangleq \mathbb{E}_{r\sim\mathcal{R}_a(\mathcal{P}_a(y_a))}\left[r + V^*(T - 1, \mathcal{U}(\mathbf{y}, a, r))\right] \tag{4}$$

$$V^*(T, \mathbf{y}) \triangleq \max_{a\in\mathcal{A}} Q^*(T, \mathbf{y}, a), \tag{5}$$

with $V^*(0, \mathbf{y}) \triangleq 0$ for all $\mathbf{y} \in \mathcal{Y}$. While the Bellman equation is intractable to analyze, it offers a characterization of the Bayesian optimal policy (Opt) and the best achievable performance V^*: i.e., $V^*(T, \mathbf{y}) = V(\text{OPT}, T, \mathbf{y}) = \sup_{\pi\in\Pi_{\mathbb{F}}} V(\pi, T, \mathbf{y})$.

3. Information Relaxation Sampling

We propose a general framework, which we refer to as *information relaxation sampling* (IRS), that takes as an input a 'penalty function', and produces as outputs a policy and an associated performance bound.

Information relaxation penalties and inner problem. If we relax the nonanticipativity constraint imposed on policy space $\Pi_{\mathbb{F}}$ (i.e., a_t^π is $\mathcal{F}_{t-1}$-measurable), the DM will be allowed to first observe all future outcomes in advance, and then pick an action (i.e., a_t^π is $\sigma(\omega)$-measurable). To compensate for this relaxation, we impose a penalty on the DM for violating the nonanticipativity constraint.

We introduce a *penalty function* $z_t(\mathbf{a}_{1:t};\omega,T,\mathbf{y})$ to denote the penalty that the DM incurs at time t, when taking an action sequence $\mathbf{a}_{1:t}$ given a particular instance specified by ω, T and $\mathbf{y}$. The clairvoyant DM can find the best action sequence that is optimal for a particular outcome ω in the presence of penalties z_t, by solving the following (deterministic) optimization problem, referred as the *inner problem*:

$$\text{maximize}_{\mathbf{a}_{1:T}\in\mathcal{A}^T} \quad \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t};\omega) - z_t(\mathbf{a}_{1:t};\omega,T,\mathbf{y}). \tag{$*$}$$

Definition 1 (Dual feasibility). *A penalty function z_t is dual feasible if it is ex-ante zero-mean, i.e.,*

$$\mathbb{E}\left[z_t(\mathbf{a}_{1:t};\omega,T,\mathbf{y})\,|\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1};\omega)\right]=0, \quad \forall \mathbf{a}_{1:t}\in\mathcal{A}^t, \quad \forall t\in[T] \tag{6}$$

To clarify the notion of conditional expectation, we remark that the penalty function $z_t(\cdot;\omega,T,\mathbf{y})$ is a stochastic function of the action sequence $\mathbf{a}_{1:t}$ since the outcome ω is random.[1] The dual feasibility condition requires that the DM who makes decisions on the natural filtration will receive zero penalties in expectation.

IRS performance bound. Let $W^z(T,\mathbf{y})$ be the expected maximal value of the inner problem ($*$), when the outcome ω is randomly drawn from its prior distribution $\mathcal{I}(T,\mathbf{y})$, i.e., the expected total payoff that a clairvoyant DM can achieve in the presence of penalties:

$$W^z(T,\mathbf{y}) \triangleq \mathbb{E}_{\omega\sim\mathcal{I}(T,\mathbf{y})}\left[\max_{\mathbf{a}_{1:T}\in\mathcal{A}^T}\left\{\sum_{t=1}^{T} r_t(\mathbf{a}_{1:t};\omega) - z_t(\mathbf{a}_{1:t};\omega,T,\mathbf{y})\right\}\right]. \tag{7}$$

We can obtain this value numerically via simulation: draw outcomes $\omega^{(1)},\omega^{(2)},\cdots,\omega^{(S)}$ independently from $\mathcal{I}(T,\mathbf{y})$, solve the inner problem for each outcome separately, and then take the average of the maximal value over samples. The following theorem shows that W^z is indeed a valid performance bound of the stochastic MAB problem.

Theorem 1 (Weak duality and strong duality). *If the penalty function z_t is dual feasible, W^z is an upper bound on the optimal value V^*: for any T and $\mathbf{y}$,*

$$\textit{(Weak duality)} \qquad W^z(T,\mathbf{y}) \ge V^*(T,\mathbf{y}). \tag{8}$$

There exists a dual feasible penalty function, referred as the ideal penalty z_t^{ideal}, such that

$$\textit{(Strong duality)} \qquad W^{\text{ideal}}(T,\mathbf{y}) = V^*(T,\mathbf{y}). \tag{9}$$

[1] As in usual probability theory, $Z(\omega)\triangleq\mathbb{E}[X(\omega)|Y(\omega)]$ represents the expected value of a random variable $X(\omega)$ given the information $Y(\omega)$, and $Z(\omega)$ is itself a random variable that has a dependency on ω.

The ideal penalty function z_t^{ideal} has a following functional form:

$$z_t^{\text{ideal}}(\mathbf{a}_{1:t};\omega,T,\mathbf{y}) \triangleq r_t(\mathbf{a}_{1:t};\omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t};\omega)\,|\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1};\omega)\right] + V^*\left(T-t,\mathbf{y}_t(\mathbf{a}_{1:t};\omega)\right) - \mathbb{E}\left[V^*\left(T-t,\mathbf{y}_t(\mathbf{a}_{1:t};\omega)\right)\middle|\,\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1};\omega)\right]. \tag{10}$$

Recall that a dual feasible penalty function does not penalize (in expectation) non-anticipating policies, which include Opt. Even when the future information is available, the DM can earn V^* under the penalties by implementing Opt without taking advantage of future information. When she makes use of future information, she can always outperform Opt, which leads to the weak duality result. The ideal penalty z_t^{ideal} precisely penalizes for the additional profit extracted from using the future information, therefore removing any incentive to deviate from Opt and resulting in the strong duality.

The ideal penalty is, of course, intractable, but its structure tells us what a good penalty may look like. It implies that there are two sources of additional profit: in DP terminology, one from knowing future immediate rewards and one from knowing future state transitions, each of which will be taken into account later in this paper. As another implication, it shows that relaxing more the available information can always be compensated by adding associated terms in the penalty function. Hence, it is sufficient to consider the full-information relaxation (i.e., a_t^π is $\sigma(\omega)$-measurable), as we do in this paper, in a sense that a partial-information relaxation (e.g., a_t^π is measurable w.r.t. $\mathcal{G}_{t-1}$ such that $\mathcal{F}_{t-1} \subseteq \mathcal{G}_{t-1} \subseteq \sigma(\omega)$) is equivalent to the setting with the full relaxation and a more complicated penalty function. Under the full-information relaxation, the actual amount of information available for the DM can be equivalently controlled by adjusting the penalty function.

IRS policy. Given a penalty function z_t, we characterize an IRS, possibly randomized, policy $\pi^z \in \Pi_{\mathbb{F}}$ as follows. The policy π^z specifies 'which arm to pull when the remaining time is T and current belief is $\mathbf{y}$'. Given T and $\mathbf{y}$, (i) it first samples the outcome $\tilde{\omega}$ from $\mathcal{I}(T,\mathbf{y})$ randomly, (ii) solves the inner problem to find a best action sequence $\tilde{\mathbf{a}}_{1:T}^*$ with respect to $\tilde{\omega}$ in the presence of penalties z_t, and (iii) takes the first action $\tilde{a}_1^*$ that the clairvoyant optimal solution $\tilde{\mathbf{a}}_{1:T}^*$ suggests. Analogous to Thompson sampling, the procedure (i)-(iii) is repeated at every decision epoch, while updating the remaining time T and belief $\mathbf{y}$.

Algorithm 1: Information Relaxation Sampling (IRS) policy

Function $\texttt{IRS}(T, \mathbf{y}; z)$

  Sample an outcome $\tilde{\omega} \sim \mathcal{I}(T, \mathbf{y})$

  Find the best action sequence with respect to $\tilde{\omega}$ under penalties z_t:

  $\tilde{\mathbf{a}}^*_{1:T} \leftarrow \operatorname{argmax}_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}; \tilde{\omega}) - z_t(\mathbf{a}_{1:t}; \tilde{\omega}, T, \mathbf{y}) \right\}$

  return $\tilde{a}^*_1$

Procedure $\texttt{IRS-Outer}(T, \mathbf{y}; z)$

  $\mathbf{y}_0 \leftarrow \mathbf{y}$

  for $t = 1, 2, \cdots, T$ **do**

    Pull $a_t \leftarrow \texttt{IRS}(T - t + 1, \mathbf{y}_{t-1}; z)$

    Earn and observe a reward r_t and update belief $\mathbf{y}_t \leftarrow \mathcal{U}(\mathbf{y}_{t-1}, a_t, r_t)$

  end

In step (i), sampling $\tilde{\omega} \sim \mathcal{I}(T, \mathbf{y})$ means sampling the parameters $\tilde{\theta}_a \sim \mathcal{P}_a(y_a)$ and then generating the future rewards $\tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}_a)$ for all $n \in [T]$ and all $a \in \mathcal{A}$. It is equivalent to simulating a plausible future scenario based on the current belief $\mathbf{y}$, and π^z takes the best action optimized to this synthesized future. Note that only the first action $\tilde{a}^*_1$ out of the optimal solution $\tilde{\mathbf{a}}^*_{1:T}$ is utilized, and at the following decision epoch a new outcome is sampled from the updated posterior. For a MAB problem with time horizon T, in total, it solves T different instances of the inner problem throughout the entire decision process.

Remark 1. *The ideal penalty yields the Bayesian optimal policy: i.e.,* $V(\pi^{\text{ideal}}, T, \mathbf{y}) = V^*(T, \mathbf{y})$.

Choice of penalty functions. IRS policies include Thompson Sampling and the Bayesian optimal policy as two extremal cases. We propose a set of penalty functions spanning these two. While deferring the detailed explanations in §3.1 - §3.3, we briefly list the penalty functions:

$$z_t^{\text{TS}}(\mathbf{a}_{1:t}; \omega) \triangleq r_t(\mathbf{a}_{1:t}; \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}; \omega) \,|\boldsymbol{\theta} \right] \tag{11}$$

$$z_t^{\text{IRS.FH}}(\mathbf{a}_{1:t}; \omega) \triangleq r_t(\mathbf{a}_{1:t}; \omega) - \mathbb{E}\left[\mu_{a_t}(\theta_{a_t}) \,| R_{a_t,1}, \cdots, R_{a_t,T-1} \right] \tag{12}$$

$$z_t^{\text{IRS.V-Zero}}(\mathbf{a}_{1:t}; \omega) \triangleq r_t(\mathbf{a}_{1:t}; \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}; \omega) \,| \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}; \omega) \right] \tag{13}$$

$$\begin{aligned} z_t^{\text{IRS.V-EMax}}(\mathbf{a}_{1:t}; \omega) \triangleq{}& r_t(\mathbf{a}_{1:t}; \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}; \omega) \,| \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}; \omega) \right] \\ &+ W^{\text{TS}}\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}; \omega)\right) - \mathbb{E}\left[W^{\text{TS}}\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}; \omega)\right) \middle| \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}; \omega) \right] \end{aligned} \tag{14}$$

where $\mathbf{y}_t(\mathbf{a}_{1:t}; \omega)$ represents the posterior belief at time t as used in IRS-Outer in Algorithm 1. To help understanding, we provide an identity as an example: $\mathbb{E}\left[r_t(\mathbf{a}_{1:t}; \omega) \,| \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}; \omega) \right] = \mathbb{E}\left[\mu_{a_t}(\theta_{a_t}) \,| \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}; \omega) \right] = \mathbb{E}\left[\mu_{a_t}(\theta_{a_t}) \middle| R_{a_t,1}, \cdots, R_{a_t, n_{t-1}(\mathbf{a}_{1:t-1}, a_t)} \right] = \bar{\mu}_{a_t}\left([\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}; \omega)]_{a_t} \right)$ – they all represent the mean reward that the DM expects to get from arm a_t before making a decision at time t.

Remark 2. *All penalty functions* (10)-(14) *are dual feasible.*

Penalty function	Policy	Performance bound	Inner problem	Run time
z_t^{TS}	TS	W^{TS}	Find a best arm given parameters.	$O(K)$
$z_t^{\text{IRS.FH}}$	$\pi^{\text{IRS.FH}}$	$W^{\text{IRS.FH}}$	Find a best arm given finite observations.	$O(K)$ or $O(KT)$
$z_t^{\text{IRS.V-ZERO}}$	$\pi^{\text{IRS.V-ZERO}}$	$W^{\text{IRS.V-ZERO}}$	Find an optimal allocation of T pulls.	$O(KT^2)$
$z_t^{\text{IRS.V-EMAX}}$	$\pi^{\text{IRS.V-EMAX}}$	$W^{\text{IRS.V-EMAX}}$	Find an optimal action sequence.	$O(KT^K)$
z_t^{ideal}	OPT	V^*	Solve Bellman equations.	-

Table 1: List of algorithms following from the penalty functions (10)-(14). TS refers to Thompson sampling and OPT refers to the Bayesian optimal policy. Run time represents the time complexity of solving one instance of inner problem, representing the time required to obtain one sample of performance bound or to make a single decision in policy.

Table 1 summarizes our algorithms investigated in this paper. We derive a variety of penalty designs by exploiting the structures in the causal process of a Bayesian learner. As we sequentially increase its complexity, from z^{TS} to z^{ideal}, the penalty function more accurately penalizes the benefit of knowing the future outcomes, more explicitly preventing the DM from exploiting the future information. It makes the inner problem closer to the original stochastic optimization problem that results in a better performing policy and a tighter performance bound. As a result, we achieve a family of algorithms that are intuitive and tractable, exhibiting a trade-off between quality and computational efficiency.

3.1. Thompson Sampling

With the penalty function $z_t^{\text{TS}}(\mathbf{a}_{1:t};\omega) \triangleq r_t(\mathbf{a}_{1:t};\omega) - \mu_{a_t}(\theta_{a_t})$, the inner problem $(*)$ reduces to

$$\max_{\mathbf{a}_{1:T}\in\mathcal{A}^T} \left\{ \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t};\omega) - z_t(\mathbf{a}_{1:t};\omega) \right\} = \max_{\mathbf{a}_{1:T}\in\mathcal{A}^T} \left\{ \sum_{t=1}^{T} \mu_{a_t}(\theta_{a_t}) \right\} = T \times \max_{a\in\mathcal{A}} \mu_a(\theta_a). \tag{15}$$

Given an outcome ω, in the presence of penalties, a hindsight optimal action sequence is to keep pulling one arm $a_1^* = \operatorname{argmax}_a \mu_a(\theta_a)$, T times in a row. The resulting performance bound is equivalent to the conventional regret benchmark, i.e.,

$$W^{\text{TS}}(T,\mathbf{y}) = \mathbb{E}\left[T \times \max_{a\in\mathcal{A}} \mu_a(\theta_a) \right] \tag{16}$$

which measures how much the DM could have achieved if the parameters are revealed in advance. The corresponding IRS policy π^{TS} is equivalent to Thompson Sampling: when the sampled outcome $\tilde{\omega}$ is used instead, it pulls the arm $\operatorname{argmax}_a \mu_a(\tilde{\theta}_a)$ where each $\tilde{\theta}_a \sim \mathcal{P}_a(y_a)$, and this sampling-based decision making is repeated at each epoch, while updating the belief sequentially, as described in IRS-OUTER in Algorithm 1.

Note that the optimal solution is determined by the parameters $\boldsymbol{\theta}$ only – it does not need to consider the future rewards, and thus it takes $O(K)$ computations to make a single decision in policy or to obtain a single sample of performance bound.

3.2. IRS.FH

Let $\mu_{a,T-1}(\omega)$ be the expected mean reward of an arm a inferred from $T-1$ reward realizations $R_{a,1}, \cdots, R_{a,T-1}$. Given (12), the optimal solution to the inner problem $(*)$ is to pull an arm with the highest $\mu_{a,T-1}(\omega)$ from beginning to the end:

$$\mu_{a,T-1}(\omega) \triangleq \mathbb{E}\left[\mu_a(\theta_a)\,|R_{a,1}, \cdots, R_{a,T-1}\right], \quad W^{\text{IRS.FH}}(T, \mathbf{y}) = \mathbb{E}\left[T \times \max_{a\in\mathcal{A}} \mu_{a,T-1}(\omega)\right]. \quad (17)$$

IRS.FH is almost identical to TS except that $\mu_a(\theta_a)$ is replaced with $\mu_{a,T-1}(\omega)$. Note that $\mu_{a,T-1}(\omega)$ is less informative than $\mu_a(\theta_a)$ from the DM's perspective, since she will never be able to learn $\mu_a(\theta_a)$ perfectly within a finite horizon. In terms of mean reward estimation, knowing the parameters is equivalent to having the infinite number of observations. The inner problem of TS asks the DM to "identify the best arm based on the infinite number of samples" whereas that of IRS.FH asks her to "identify the best arm based on the finite number of samples", which takes into account the length of time horizon explicitly.

Focusing on the randomness of $\mu_a(\theta_a)$ and $\mu_{a,T-1}(\omega)$, we observe that the distribution of $\mu_{a,T-1}(\omega)$ will be more concentrated around its mean $\bar{\mu}_a(y_a)$. Following from Jensen's inequality, we have $W^{\text{IRS.FH}} \leq W^{\text{TS}}$ for any problem instance, saying that IRS.FH yields a performance bound tighter than the conventional benchmark. In terms of policy, the variance of $\mu_{a,T-1}(\tilde{\omega})$ (and $\mu_a(\theta_a)$) also governs the degree of random exploration, deviating from the myopic decision of pulling an arm with the largest $\bar{\mu}_a(y_a)$. When it approaches the end of the horizon ($T \searrow 1$), $\pi^{\text{IRS.FH}}$ naturally explores less than TS.

Sampling $\mu_{a,T-1}(\tilde{\omega})$ at once. In order to obtain $\mu_{a,T-1}(\tilde{\omega})$ for a synthesized outcome $\tilde{\omega}$, one may apply Bayes' rule sequentially for each reward realization, which will take $O(KT)$ computations in total. It can be done in $O(K)$ if the prior distribution $\mathcal{P}_a$ is a conjugate prior of the reward distribution $\mathcal{R}_a$, in which the belief can be updated in a batch by the use of sufficient statistics of observations. In the case of the Beta-Bernoulli MAB or the Gaussian MAB, for example, $\mu_{a,T-1}(\tilde{\omega})$ can be represented as a convex combination of the current estimate $\bar{\mu}_a(y_a)$ and the sample mean $\frac{1}{T-1}\sum_{n=1}^{T-1}\tilde{R}_{a,n-1}$. We further know that the distribution of $\sum_{n=1}^{T-1}\tilde{R}_{a,n-1}$ is Binomial$(T-1, \tilde{\theta}_a)$ for the Beta-Bernoulli case, and $\mathcal{N}((T-1)\cdot\tilde{\theta}_a, (T-1)\cdot\sigma_a^2)$ for the Gaussian case, where σ_a^2 represents the noise variance. After sampling the parameter $\tilde{\theta}_a$, we can sample $\sum_{n=1}^{T-1}\tilde{R}_{a,n-1}$ directly from the known distribution, and use it to compute $\mu_{a,T-1}(\tilde{\omega})$ without sequentially updating the belief. In such cases, a single decision of $\pi^{\text{IRS.FH}}$ can be made within $O(K)$ operations, similar in complexity to TS.

3.3. IRS.V-Zero and IRS.V-EMax

IRS.V-Zero. Let $\mu_{a,n}(\omega)$ be the expected mean reward of arm a inferred from the first n reward realizations:

$$\mu_{a,n}(\omega) \triangleq \mathbb{E}\left[\mu_a(\theta_a)\,|R_{a,1}, \cdots, R_{a,n}\right]. \quad (18)$$

Under this penalty, the DM earns $\mu_{a,n-1}(\omega)$ from the n^{th} pull of an arm a: for example, if $\mathbf{a}_{1:T} = (1, 2, 2, 2, 2, 1)$, the total payoff is $\mu_{1,0} + \mu_{2,0} + \mu_{2,1} + \mu_{2,2} + \mu_{2,3} + \mu_{1,1}$.

Given an outcome ω, the total payoff is determined only by the total number of pulls of each arm, and not the sequence in which the arms had been pulled. Therefore, solving the inner problem $(*)$ is equivalent to "finding the optimal allocation $(n_1^*, n_2^*, \cdots, n_K^*)$ among T remaining opportunities": omitting ω for brevity, the inner problem reduces to

$$\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} \mu_{a_t, n_{t-1}(\mathbf{a}_{1:t-1}, a_t)} \right\} = \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{a=1}^{K} \sum_{n=1}^{n_T(\mathbf{a}_{1:T}, a)} \mu_{a,n-1} \right\} = \max_{\mathbf{n}_{1:K} \in N_T} \left\{ \sum_{a=1}^{K} S_{a,n_a} \right\} \tag{19}$$

where $S_{a,n} \triangleq \sum_{m=1}^{n} \mu_{a,m-1}$ is the cumulative payoff from the first n pulls of an arm a, and $N_T \triangleq \{(n_1, \cdots, n_K) \in \mathbb{Z}_+^K : \sum_{a=1}^{N} n_a = T\}$ is the set of all feasible allocations. Once the $S_{a,n}$'s are computed, this inner problem can be solved within $O(KT^2)$ operations by sequentially applying sup convolution K times. The detailed implementation is provided in §A.1.

Given an optimal allocation $\tilde{\mathbf{n}}^*$, the policy $\pi^{\text{IRS.V-ZERO}}$ needs to select which arm to pull next. In principle, any arm a that was included in the solution of the inner problem, $\tilde{n}_a^* > 0$, would be fine, but we suggest a selection rule in which the arm that needs most pulls is chosen, i.e., $\text{argmax}_a \tilde{n}_a^*$. It guarantees $\pi^{\text{IRS.V-ZERO}}$ to behave like TS when T is large, as formally stated in Proposition 1.

IRS.V-EMax. IRS.V-EMAX includes an additional cost for using the information of future belief transitions. Compared to the ideal penalty z_t^{ideal} (10), $z_t^{\text{IRS.V-EMAX}}$ (14) is obtained by replacing the true value function $V^*(T, \mathbf{y})$ with $W^{\text{TS}}(T, \mathbf{y})$ (16), as a tractable approximation. The use of W^{TS} leads to a simple expression for the conditional expectation with respect to the natural filtration. Since $\boldsymbol{\theta} | \mathcal{F}_{t-1}$ is distributed with $\mathcal{P}(\mathbf{y}_{t-1})$, we have $\mathbb{E}\left[W^{\text{TS}}(T-t, \mathbf{y}_t) \middle| \mathcal{F}_{t-1} \right] = (T-t) \times \mathbb{E}\left[\max_a \mu_a(\theta_a) | \mathcal{F}_{t-1} \right] = (T-t) \times \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{P}(\mathbf{y}_{t-1})}\left[\max_a \mu_a(\theta_a) \right] = W^{\text{TS}}(T-t, \mathbf{y}_{t-1})$.

We observe that, given ω, the future belief $\mathbf{y}_t(\mathbf{a}_{1:t}; \omega)$ is completely determined by how many times each arm had been pulled, irrespective of the sequence of the pulls. For example, consider two action sequences $\mathbf{a}_{1:t}^A = (1, 1, 2, 1, 2)$ and $\mathbf{a}_{1:t}^B = (2, 1, 1, 2, 1)$. Even though the order of observations would differ, the agent will observe $(R_{1,1}, R_{1,2}, R_{1,3})$ from arm 1 and $(R_{2,1}, R_{2,2})$ from arm 2 in both cases that end up with the same belief $\mathbf{y}_t(\mathbf{a}_{1:t}^A; \omega) = \mathbf{y}_t(\mathbf{a}_{1:t}^B; \omega)$.

Following from the observation above, the state (belief) space can be efficiently parameterized with the pull counts $\mathbf{n}_{1:K} = (n_1, \cdots, n_K)$ instead of action sequence $\mathbf{a}_{1:t}$. Since the total number of possible future beliefs is $O(T^K)$, not $O(K^T)$, the inner problem $(*)$ can be solved by dynamic programming in $O(c_W T^K + KT^K)$ operations, where c_W is the cost of numerically calculating $W^{\text{TS}}(T, \mathbf{y})$ (see §A.2 for the detail).

IRS.Index policy. Finally, we propose IRS.INDEX, which does not strictly belong to the IRS framework, and does not produce a performance bound, but it exhibits strong empirical performance.

Roughly speaking, IRS.INDEX approximates the finite-horizon Gittins index (Kaufmann et al., 2012) using IRS.V-EMAX. For each arm in isolation, it internally solves the single-armed bandit problem in which there is a competing outside option that yields a deterministic (known) reward. Applying Irs.V-EMax to a single-armed bandit problem, we can find if the stochastic arm is worth trying against a particular value of outside option in $O(KT)$. The threshold value that makes the arm barely worth trying can be obtained by binary search, repeatedly solving the singe-armed bandit problems while varying the value of outside option. The policy $\pi^{\text{Irs.Index}}$ plays an arm with the largest threshold value. See §A.3.

4. Analysis

Remark 3 (Optimality at the end). *When $T = 1$, all $\pi^{\text{Irs.FH}}$, $\pi^{\text{Irs.V-Zero}}$, $\pi^{\text{Irs.V-EMax}}$, and $\pi^{\text{Irs.Index}}$ take the optimal action that is pulling the myopically best arm $a^* = \text{argmax}_a \bar{\mu}_a(y_a)$.*

Proposition 1 (Asymptotic behavior). *Assume $\mu_i(\theta_i) \neq \mu_j(\theta_j)$ almost surely for any two distinct arms $i \neq j$. As $T \nearrow \infty$, the distribution of Irs.FH's action converges to that of Thompson Sampling:*

$$\lim_{T\to\infty} \mathbb{P}\left[\text{Irs.FH}(T, \mathbf{y}) = a\right] = \mathbb{P}\left[\text{TS}(\mathbf{y}) = a\right], \quad \forall a \in \mathcal{A}. \tag{20}$$

Similarly, so does Irs.V-Zero[2]:

$$\lim_{T\to\infty} \mathbb{P}\left[\text{Irs.V-Zero}(T, \mathbf{y}) = a\right] = \mathbb{P}\left[\text{TS}(\mathbf{y}) = a\right], \quad \forall a \in \mathcal{A}. \tag{21}$$

TS, Irs.FH$(T, \mathbf{y})$ and Irs.V-Zero$(T, \mathbf{y})$ denote the action taken by policies π^{TS}, $\pi^{\text{Irs.FH}}$ and $\pi^{\text{Irs.V-Zero}}$, repsectively, when the remaining time is T and the prior belief is $\mathbf{y}$. These are random variables, since each of these policies uses a randomly sampled outcome $\tilde{\omega}$ on its own.

Remark 3 and Proposition 1 state that Irs.FH and Irs.V-Zero behave like TS during the initial decision epochs, gradually shift toward the myopic scheme and end up with optimal decision; in contrast, TS will continue to explore throughout. The transition from exploration to exploitation under these IRS policies occurs smoothly, without relying on an auxiliary control parameter. While maintaining their recursive structure, IRS policies take into account the horizon T, and naturally balance exploitation and exploration.

Theorem 2 (Monotonicity in performance bounds). *Irs.FH and Irs.V-Zero monotonically improve the performance bound:*

$$W^{\text{TS}}(T, \mathbf{y}) \geq W^{\text{Irs.FH}}(T, \mathbf{y}) \geq W^{\text{Irs.V-Zero}}(T, \mathbf{y}). \tag{22}$$

Note that $W^{\text{TS}}(T, \mathbf{y}) = \mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{P}(\mathbf{y})}\left[T \times \max_a \mu_a(\theta_a)\right]$ is the conventional regret benchmark.

In addition, we have $W^{\text{Irs.V-EMax}} \geq W^{\text{ideal}}$ since W^{ideal} is the lowest attainable upper bound (Theorem 1). Empirically (§D), we also observe $W^{\text{Irs.V-Zero}} \geq W^{\text{Irs.V-EMax}}$.

We interpret that the tightness of performance bound $W^z - V^*$ reflects the degree of optimism that each algorithm would possess. Recall that W^z is the expected value of the

[2] We assume a particular selection rule such that $a^{\text{Irs.V-Zero}} = \text{argmax}_a \tilde{n}^*_a$ as discussed in §3.3.

best possible payoff when the agent is informed with some future outcomes in advance. The weak duality $W^z \geq V^*$ implies that IRS algorithms are basically optimistic in a sense that the agent would believe that she can earn more than the optimal policy in a hope that the additional information is true. Even with the same outcome ω, depending on the penalties z_t, the agent would have different anticipation about the future payoff. As we incorporate the actual learning process, the agent's anticipation becomes less optimistic and the performance bound gets tighter.

We define the 'suboptimality gap' of an IRS policy π^z to be $W^z(T, \mathbf{y}) - V(\pi^z, T, \mathbf{y})$, and analyze it instead of the conventional (Bayesian) regret, $W^{\text{TS}}(T, \mathbf{y}) - V(\pi^z, T, \mathbf{y})$. While its non-negativity is guaranteed from weak duality (Theorem 1), more desirably, the optimal policy yields a zero suboptimality gap (Theorem 1 & Remark 1). It coincides with the conventional regret measure only for TS.

Theorem 3 (Suboptimality gap). *For the Beta-Bernoulli MAB, for any T and $\mathbf{y}$,*

$$W^{\text{TS}}(T, \mathbf{y}) - V(\pi^{\text{TS}}, T, \mathbf{y}) \leq 3K + 2\sqrt{\log T} \times 2\sqrt{KT} \tag{23}$$

$$W^{\text{IRS.FH}}(T, \mathbf{y}) - V(\pi^{\text{IRS.FH}}, T, \mathbf{y}) \leq 3K + 2\sqrt{\log T} \times \left(2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right) \tag{24}$$

$$W^{\text{IRS.V-ZERO}}(T, \mathbf{y}) - V(\pi^{\text{IRS.V-ZERO}}, T, \mathbf{y}) \leq 2K + \sqrt{\log T} \times \left(2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right) \tag{25}$$

We do not have a theoretical guarantee for monotonicity in the actual performance $V(\pi^z, T, \mathbf{y})$ among IRS policies. Instead, Theorem 3 indirectly shows the improvements in suboptmality: although all the bounds have the same asymptotic order of $O(\sqrt{KT \log T})$,[3] the IRS policies improve the leading coefficient or the additional term.

The proof of Theorem 3, provided in C.4, relies on an interesting property of IRS policies, which is a generalization of TS. Russo and Van Roy (2014) observed that TS is randomized in a way that, conditional on the past observations, the probability of choosing an action a equals to the probability that the action a is chosen by someone who knows the parameters. Analogously, the IRS policy π^z is randomized in a way that, conditional on the past observations and the past actions, the probability of choosing an action a matches the probability that the action a is chosen by someone who knows the entire future but penalized (see Proposition 7). Recall that the penalties are designed to penalize the gain of having additional future information. A better choice of penalty function prevents the policy π^z from picking up an action that is overly optimized to a randomly sampled future realization, which in turn improves the quality of the decision making.

[3] Bubeck and Liu (2013) had shown that the Bayesian regret of TS is bounded by $14\sqrt{KT}$ when the rewards have a bounded support in $[0, 1]$ including Beta-Bernoulli MAB. Despite of its lower asymptotic order, however, the actual number given in (23) is tighter than $14\sqrt{KT}$ for small T. As a side note, Lai (1987) showed that the Bayesian regret of the optimal policy has an asymptotic lower bound of $O(\log^2 T)$.

5. Numerical Experiments

We visualize the effectiveness of IRS policies and performance bounds in case of Gaussian MAB with five arms ($K = 5$) with different noise variances. More specifically, each arm $a \in \mathcal{A}$ has the unknown mean reward $\theta_a \sim \mathcal{N}(0, 1^2)$ and yields the stochastic rewards $R_{a,n} \sim \mathcal{N}(\theta_a, \sigma_a^2)$ where $\sigma_1 = 0.1$, $\sigma_2 = 0.4$, $\sigma_3 = 1.0$, $\sigma_4 = 4.0$ and $\sigma_5 = 10.0$. Our experiment includes the state-of-the-art algorithms that are particularly suitable in a Bayesian framework: Bayesian Upper Confidence Bound (Bayes-UCB, Kaufmann et al. (2012), with a quantile of $1 - \frac{1}{t}$), Information Directed Sampling (IDS, Russo and Van Roy (2017)), and Optimistic Gittins Index (OGI, Farias and Gutin (2016), one-step look ahead approximation with a discount factor of $\gamma_t = 1 - \frac{1}{t}$). Irs.V-EMax algorithm is omitted here because of its time complexity. In §D, we provide the detailed simulation procedures and the results for the other settings including Irs.V-EMax.

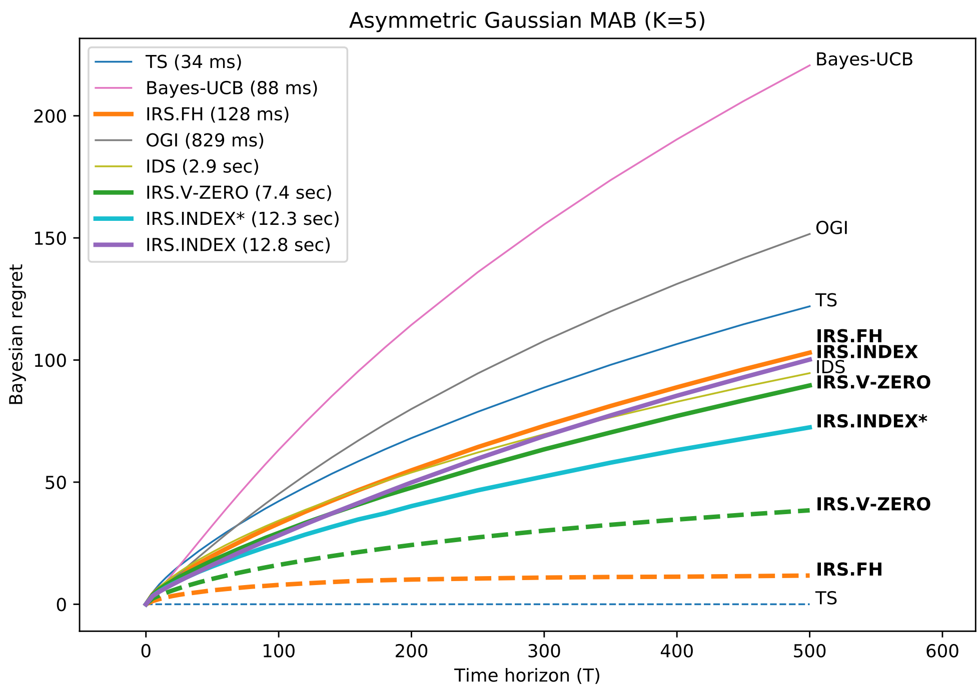

Figure 1: Regret plot for Gaussian MAB with five arms whose noise variances are different. The solid lines represent the (Bayesian) regret of algorithms, $W^{\text{TS}}(T, \mathbf{y}) - V(\pi, T, \mathbf{y})$, and the dashed lines represent the regret bounds that IRS algorithms produce, $W^{\text{TS}}(T, \mathbf{y}) - W^z(T, \mathbf{y})$. The times on the top left corner represent the average length of time required to simulate each policy for a single problem instance with $T = 500$.

Figure 1 shows the Bayesian regrets (solid lines, $W^{\text{TS}}(T, \mathbf{y}) - V(\pi, T, \mathbf{y})$) and the regret bounds (dashed lines, $W^{\text{TS}}(T, \mathbf{y}) - W^z(T, \mathbf{y})$) that are measured at the different values of $T = 5, 10, \cdots, 500$. Note that lower regret curves are better, and higher bound curves are better. Also, the regret bound produced by TS is zero, since $W^{\text{TS}}(T, \mathbf{y})$ is the benchmark (16) used in this regret plot.

We first observe a clear improvement in both performances and bounds as we incorporate more complicated penalty functions from TS to IRS.V-ZERO. As stated in Theorem 2, the monotonicity in the bound curves can be observed. The suboptimality gap (the gap between a regret curve and its corresponding bound curve) gets tightened, which is consistent with the implication of Theorem 3. As a trade-off, however, it requires a longer running time.

In this particular example, it is crucial to incorporate how much we can learn about each of the arms during the remaining time periods, which heavily depends on the noise level σ_a.[4] Comparing IRS.FH with TS, as a simple modification for finite-horizon setting, the performance has improved significantly without degrading its computational efficiency. We also observe that IRS policies and IDS outperform to BAYES-UCB, OGI and TS algorithms, since they explicitly incorporate the value of exploration – how quickly the posterior distribution will be concentrated upon each observation.

This example also illustrates us the significance of having a tighter performance bound. Benchmarking to $W^{\text{IRS.V-ZERO}}$, when $T = 500$, IRS.INDEX* policy achieves 94% $\left(= \frac{V(\pi^{\text{IRS.INDEX*}}, T, \mathbf{y})}{W^{\text{IRS.V-ZERO}}(T, \mathbf{y})}\right)$ of it. If the conventional benchmark W^{TS} is used instead, as in a usual regret analysis, we might have concluded that IRS.INDEX* only achieves 88% $\left(= \frac{V(\pi^{\text{IRS.INDEX*}}, T, \mathbf{y})}{W^{\text{TS}}(T, \mathbf{y})}\right)$ of that (looser) bound, which may suggest a larger margin of possible improvement.

[4] In order for the posterior distribution to be concentrated so as to have the standard deviation of 0.1, for example, one observation is enough for arm 1 whereas 100 and 10,000 observations are required for arm 3 and arm 5, respectively.

References

D. A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Champman and Hall, 1985.

R. N. Bradt, S. M. Johnson, and S. Karlin. On sequential designs for maximizing the sum of n observations. *The Annals of Mathematical Statistics*, pages 1060–1074, 1956.

David B. Brown and Martin B. Haugh. Information relaxation bounds for infinite horizon markov decision processes. *Operations Research*, 65(5):1355–1379, 2017.

David B. Brown, James E. Smith, and Peng Sun. Information relaxations and duality in stochastic dynamic programs. *Operations Research*, 58(4):785–801, 2010.

Sebastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for Thompson Sampling. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 1(638-646), 2013.

M. H. A. Davis and I. Karatzas. *A Deterministic Approach to Optimal Stopping.* Wiley, 1994.

Vijay V. Desai, Vivek F. Farias, and Ciamac C. Moallemi. Pathwise optimization for optimal stopping problems. *Management Science*, 58(12):2292–2308, 2012.

Vivek F. Farias and Eli Gutin. Optimistic Gittins indices. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, (3161-3169), 2016.

J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2):148–177, 1979.

Martin B. Haugh and Leonid Kogan. Pricing American options: A duality approach. *Operations Research*, 52(2):258–270, 2004.

Martin B. Haugh and Andrew E.B. Lim. Linear-quadratic control and information relaxations. *Operations Research Letters*, 40:521–528, 2012.

Martin B. Haugh and Chun Wang. Dynamic portfolio execution and information relaxations. *SIAM Journal of Financial Math*, 5:316–359, 2014.

Emilie Kaufmann, Olivier Cappe, and Aurelien Garivier. On Bayesian upper confidence bounds for bandit problems. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics,*, 22:592–600, 2012.

T. L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 15(3):1091–1114, 1987.

T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Olivier Marchal and Julyan Arbel. On the sub-Gaussianity of the Beta and Dirichlet distributions. 2017.

Rockafellar and Wets. Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16(1):119–147, 1991.

L. C. G. Rogers. Monte carlo valuation of American options. *Mathematical Finance*, 12(3): 271–286, 2002.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Daniel Russo and Benjamin Van Roy. Learning to optimize via Information-Directed Sampling. *Operations Research*, 66(1):230–252, 2017.

W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

A. Algorithms in Detail

A.1. Implementation of IRS.V-Zero

We provide a pseudo-code of $\pi^{\text{Irs.V-Zero}}$ introduced in §3.3. The same logic can be directly used to compute the performance bound $W^{\text{Irs.V-Zero}}$ if the sampled outcome $\tilde{\omega}$ is replaced with the true outcome ω.

Algorithm 2: Irs.V-Zero policy

Function `Irs.V-Zero`($T, \mathbf{y}$)

- $\tilde{\theta}_a \sim \mathcal{P}_a(y_a), \tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}), \forall n \in [T], \forall a \in [K]$
- **for** $a = 1, \cdots, K$ **do**
  - $\tilde{y}_{a,0} \leftarrow y_a, \tilde{S}_{a,0} \leftarrow 0$
  - **for** $n = 1, \cdots, T$ **do**
    - $\tilde{S}_{a,n} \leftarrow \tilde{S}_{a,n-1} + \bar{\mu}_a(\tilde{y}_{a,n-1})$
    - $\tilde{y}_{a,n} \leftarrow \mathcal{U}_a(\tilde{y}_{a,n-1}, \tilde{R}_{a,n})$
  - **end**
- **end**
- $\tilde{M}_{0,0} \leftarrow 0, \tilde{M}_{0,n} \leftarrow -\infty, \forall n = 1, \cdots, T$
- **for** $a = 1, \cdots, K$ **do**
  - **for** $n = 0, \cdots, T$ **do**
    - $\tilde{M}_{a,n} \leftarrow \max_{0 \le m \le n}\{\tilde{M}_{a-1,n-m} + \tilde{S}_{a,m}\}$
    - $\tilde{A}_{a,n} \leftarrow \text{argmax}_{0 \le m \le n}\{\tilde{M}_{a-1,n-m} + \tilde{S}_{a,m}\}$
  - **end**
- **end**
- $m \leftarrow T$
- **for** $a = K, \cdots, 1$ **do**
  - $\tilde{n}_a^* \leftarrow \tilde{A}_{a,m}$
  - $m \leftarrow m - \tilde{n}_a^*$
- **end**
- **return** $\text{argmax}_a \, \tilde{n}_a^*$

A.2. Implementation of IRS.V-EMax

Given the penalty function $z_t^{\text{IRS.V-EMAX}}$ defined in (14), we define the payoff of pulling an arm a one more time after pulling each arm a', $n_{a'}$ times: with $\mathbf{n}_{1:K} \triangleq (n_1, \cdots, n_K) \in \mathbb{Z}^K$,

$$\begin{aligned} r^z(\mathbf{n}_{1:K}, a; \omega) \quad \triangleq \quad & \bar{\mu}_a([\mathbf{y}_t(\mathbf{n}_{1:K}; \omega)]_a) - W^{\text{TS}}\left(T-t-1, \mathbf{y}_{t+1}(\mathbf{n}_{1:K} + \mathbf{e}_a; \omega)\right) \\ & + W^{\text{TS}}\left(T-t-1, \mathbf{y}_t(\mathbf{n}_{1:K}; \omega)\right) \end{aligned} \tag{26}$$

where $\mathbf{e}_a \in \mathbb{Z}^K$ is a basis vector such that a^{th} component is one and the others are zero. Note that we used the fact that $\mathbb{E}\left[W^{\text{TS}}\left(T-t, \mathbf{y}_t\right) \middle| \mathcal{F}_{t-1}\right] = W^{\text{TS}}\left(T-t, \mathbf{y}_{t-1}\right)$. We also use the notation of $\mathbf{y}_t(\mathbf{n}_{1:K}; \omega)$ to denote the belief as a function of pull counts $\mathbf{n}_{1:K}$, based on the observation that the belief is completely determined by how many times each arm was pulled, $\mathbf{n}_{1:K}$, no matter in what order they were pulled.

Consider a subproblem of $(*)$ such that maximizes the total payoff given the number of pulls $\mathbf{n}_{1:K}$ across arm: with $t = \sum_{a=1}^K n_a$,

$$M(\mathbf{n}_{1:K}; \omega) \triangleq \max_{\mathbf{a}_{1:t} \in \mathcal{A}^t} \left\{ \sum_{s=1}^t r_s(\mathbf{a}_{1:s}; \omega) - z_s^{\text{IRS.V-EMAX}}(\mathbf{a}_{1:s}; \omega); \ \sum_{s=1}^t \mathbf{1}\{a_s = a\} = n_a, \forall a \right\}. \tag{27}$$

Then, it should satisfy

$$M(\mathbf{n}_{1:K}; \omega) = \max_{a \in \mathcal{A}} \left\{ M(\mathbf{n}_{1:K} - \mathbf{e}_a; \omega) + r^z(\mathbf{n}_{1:K} - \mathbf{e}_a, a; \omega); \ n_a \geq 1 \right\}. \tag{28}$$

For all feasible counts $\mathbf{n}_{1:K}$'s such that $\sum_{a=1}^K n_a \leq T$, we can compute $M(\mathbf{n}_{1:K}; \omega)$'s by sequentially solving (28) in an appropriate order. After all, we can obtain the maximal value to original inner problem $(*)$ by evaluating

$$\max_{\mathbf{n}_{1:K}} \left\{ M(\mathbf{n}_{1:K}; \omega); \quad \sum_{a=1}^K n_a = T \right\}. \tag{29}$$

The optimal action sequence $\mathbf{a}_{1:T}^*$ can be elicited by tracking $M(\mathbf{n}_{1:K}, \omega)$'s backward.

Algorithm 3: IRS.V-EMAX policy

Function `Irs.V-EMax`$(T, \mathbf{y})$

1 Sample an outcome $\tilde{\omega} \sim \mathcal{I}(T, \mathbf{y})$

2 $\tilde{y}_{a,0} \leftarrow y_a, \quad \tilde{y}_{a,n} \leftarrow \mathcal{U}_a(\tilde{y}_{a,n-1}, \tilde{R}_{a,n}), \quad \forall n \in [T], \quad \forall a \in [K]$

3 **for** *each* $\mathbf{n}_{1:K} \in N_{\leq T}$ **do**

4 $\quad \tilde{\Gamma}[\mathbf{n}_{1:K}] \leftarrow \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{P}(\tilde{\mathbf{y}}(\mathbf{n}_{1:K}))} [\max_a \mu_a(\theta_a)]$

end

5 **for** *each* $\mathbf{n}_{1:K} \in N_{<T}$ **do**

6 $\quad \tilde{r}^z[\mathbf{n}_{1:K}, a] \leftarrow \bar{\mu}_a(\tilde{y}_{a,n_a-1}) + \left(T - \sum_{a=1}^K n_a - 1\right) \times \left(\tilde{\Gamma}[\mathbf{n}_{1:K}] - \tilde{\Gamma}[\mathbf{n}_{1:K} + \mathbf{e}_a]\right), \quad \forall a$

end

7 $\tilde{M}[\mathbf{0}] \leftarrow 0$

8 **for** *each* $\mathbf{n}_{1:K} \in N_{\leq T} \setminus \{\mathbf{0}\}$ *in order* **do**

9 $\quad \tilde{M}[\mathbf{n}_{1:K}] \leftarrow \max_a \left\{\tilde{M}[\mathbf{n}_{1:K} - \mathbf{e}_a] + \tilde{r}^z[\mathbf{n}_{1:K} - \mathbf{e}_a, a]\right\}$

10 $\quad \tilde{A}[\mathbf{n}_{1:K}] \leftarrow \operatorname{argmax}_a \left\{\tilde{M}[\mathbf{n}_{1:K} - \mathbf{e}_a] + \tilde{r}^z[\mathbf{n}_{1:K} - \mathbf{e}_a, a]\right\}$

end

11 $\mathbf{m}_{1:K} \leftarrow \operatorname{argmax} \left\{\tilde{M}[\mathbf{n}_{1:K}]; \sum_a n_a = T\right\}$

12 **for** $t = T, \cdots, 1$ **do**

13 $\quad \tilde{a}_t^* \leftarrow \tilde{A}[\mathbf{m}_{1:K}]$

14 $\quad m_{\tilde{a}_t^*} \leftarrow m_{\tilde{a}_t^*} - 1$

end

15 **return** $\tilde{a}_1^*$

Here, $\tilde{\mathbf{y}}(\mathbf{n}_{1:K}) \triangleq (\tilde{y}_{1,n_1}, \cdots, \tilde{y}_{K,n_K})$, $N_{\leq T} \triangleq \{\mathbf{n}_{1:K}; \sum_a n_a \leq T\}$, $N_{<T} \triangleq \{\mathbf{n}_{1:K}; \sum_a n_a < T\}$, and in line 8, $\mathbf{n}_{1:K}$ iterates over $N_{\leq T} \setminus \{\mathbf{0}\}$ in an order that $\sum_{a=1}^K n_a$ is non-decreasing.

Since $|N_{\leq T}| = O(T^K)$, it requires $O(KT^K)$ operations to compute all $M(\mathbf{n}_{1:K}, \omega)$'s. However, another practical issue is the cost of computing $W^{\mathrm{TS}}(T, \mathbf{y}) = T \times \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{P}(\mathbf{y})} [\max_a \mu_a(\theta_a)]$ which has to be evaluated $O(T^K)$ times in total. In general, there is no simple closed form expression in general, and it should be evaluated with numerical integration or sampling.

A.3. IRS.Index Policy

Single-armed bandit problem. Consider a problem with a single arm a that yields stochastic rewards $R_{a,n} \sim \mathcal{R}_a(\theta_a)$ and with an outside option that yields a deterministic reward λ. We have a prior distribution $\mathcal{P}_a(y_a)$ over unknown parameter θ_a whereas the deterministic reward λ is known apriori.

Given an outcome ω, we have the future belief trajectory $(y_{a,n})_{n \in \{0, \cdots, T\}}$ where $y_{a,n}$ is the belief after observing first n rewards.

$$y_{a,0} \triangleq y_a, \quad y_{a,n} \triangleq \mathcal{U}_a(y_{a,n-1}, R_{a,n}), \quad \forall n \in [T] \tag{30}$$

We adopt the penalty function $z_t^{\text{IRS.V-EMAX}}$ in which the true value function $V^*(T, y_a, \lambda)$ is approximated by $W^{\mathrm{TS}}(T, y_a, \lambda) = \mathbb{E}_{\theta_a \sim \mathcal{P}_a(y_a)} [T \times \max\{\mu_a(\theta_a), \lambda\}]$. We define $\mathcal{A} \triangleq \{0, 1\}$

such that $a_t = 1$ if pulling the stochastic arm and $a_t = 0$ if choosing the outside option at time t. The associated inner problem is

$$\text{maximize} \quad \sum_{t=1}^{T} \mu_{a,n_t-1} \cdot \mathbf{1}\{a_t = 1\} + \lambda \cdot \mathbf{1}\{a_t = 0\} - (T-t) \times \left(\Gamma^\lambda_{n_t} - \Gamma^\lambda_{n_{t-1}}\right) \tag{31}$$

$$\text{subject to} \quad n_t = \sum_{s=1}^{t} \mathbf{1}\{a_t = 1\}, \quad a_t \in \{0,1\}, \quad \forall t = 1, \cdots, T \tag{32}$$

where $\mu_{a,n} \triangleq \bar{\mu}_a(y_{a,n})$ and

$$\Gamma^\lambda_n \triangleq \mathbb{E}_{\theta_a \sim \mathcal{P}_a(y_{a,n})} \left[\max(\mu_a(\theta_a), \lambda)\right]. \tag{33}$$

Proposition 2. *The optimization problem* (31) *can be reformulated as*

$$\max_{0 \le n \le T} \left\{ T \times \Gamma^\lambda_0 + (T-n) \times \left(\lambda - \min_{0 \le s \le n} \Gamma^\lambda_s\right) + \sum_{s=1}^{n} \left(\mu_{a,s-1} - \Gamma^\lambda_{s-1}\right) \right\}. \tag{34}$$

Here, the decision variable n is the total number of pulls of stochastic arm.

Proof. Fix $m \triangleq n_T$. Note that if $a_t = 0$, $(T-t) \times (\Gamma^\lambda_{n_t} - \Gamma^\lambda_{n_{t-1}}) = 0$ since $n_t = n_{t-1}$. The objective function can be represented as

$$\sum_{n=1}^{m} \mu_{a,n-1} + (T-m) \times \lambda - \sum_{n=1}^{m} (T-t_n) \times \left(\Gamma^\lambda_n - \Gamma^\lambda_{n-1}\right) \tag{35}$$

where $t_n \triangleq \inf\{t; n_t \ge n\}$. It suffices to find $(t_1, \cdots, t_m)$ with $1 \le t_1 < t_2 < \cdots < t_m \le T$ that minimize $\sum_{n=1}^{m} (T-t_n) \times \left(\Gamma^\lambda_n - \Gamma^\lambda_{n-1}\right)$. With $t_0 \triangleq 0$ and $t_{m+1} \triangleq T+1$, note that

$$\sum_{n=1}^{m} (T-t_n) \times \left(\Gamma^\lambda_n - \Gamma^\lambda_{n-1}\right) = \sum_{n=1}^{m} (T-t_n) \times \Gamma^\lambda_n - \sum_{n=1}^{m} (T-t_n) \times \Gamma^\lambda_{n-1} \tag{36}$$

$$= \sum_{n=1}^{m} (T-t_n) \times \Gamma^\lambda_n - \sum_{n=0}^{m-1} (T-t_{n+1}) \times \Gamma^\lambda_n \tag{37}$$

$$= \sum_{n=0}^{m} (T-t_n) \times \Gamma^\lambda_n - (T-t_0) \times \Gamma^\lambda_0 - \sum_{n=0}^{m} (T-t_{n+1}) \times \Gamma^\lambda_n + (T-t_{m+1}) \times \Gamma^\lambda_m \tag{38}$$

$$= -\Gamma^\lambda_m - T \times \Gamma^\lambda_0 + \sum_{n=0}^{m} (t_{n+1} - t_n) \times \Gamma^\lambda_n \tag{39}$$

In order to minimize (39), we need to set $t_{n^*+1} - t_{n^*} = T - m + 1$ for $n^* \triangleq \operatorname{argmin}_{0 \le n \le m} \Gamma^\lambda_n$ and $t_{n+1} - t_n = 1$ for $n \ne n^*$. For such t_n's, (35) reduces to

$$\sum_{n=1}^{m} \mu_{a,n-1} + (T-m) \times \lambda - \left(-\Gamma^\lambda_m - T \times \Gamma^\lambda_0 + \sum_{n=0}^{m} \Gamma^\lambda_n + (T-m) \times \min_{0 \le n \le m} \Gamma^\lambda_m\right) \tag{40}$$

$$= \sum_{n=1}^{m} \mu_{a,n-1} + (T-m) \times \left(\lambda - \min_{0 \le n \le m} \Gamma^\lambda_m\right) + T \times \Gamma^\lambda_0 - \sum_{n=0}^{m-1} \Gamma^\lambda_n. \tag{41}$$

By taking its maximum value over $m = 0, \cdots, T$, we obtain (34). ■

Let $\varphi_a(\lambda)$ be the (maximal) relative benefit of pulling the stochastic arm against not pulling.

$$\varphi_a(\lambda) \triangleq \max_{1\leq n\leq T} \left\{ T \times \Gamma_0^\lambda + (T-n) \times \left(\lambda - \min_{0\leq s\leq n} \Gamma_s^\lambda\right) + \sum_{s=1}^{n} \left(\mu_{a,s-1} - \Gamma_{s-1}^\lambda\right) \right\} - T \times \lambda \quad (42)$$

Note that max was taken over $n \geq 1$. We interpret that, given the future belief trajectory $(y_{a,n})_{n\in\{0,\cdots,T\}}$, the stochastic arm is worth trying against the deterministic reward λ if $\varphi_a(\lambda) \geq 0$ and not worth trying if $\varphi_a(\lambda) < 0$.

The value of $\varphi_a(\lambda)$ can be computed in $O(T)$ operations by precalculating $\sum_{s=1}^{n} \mu_{a,s-1}$, $\min_{0\leq s\leq n} \Gamma_s^\lambda$ and $\sum_{s=1}^{n} \Gamma_{s-1}^\lambda$ over $n = 1, \cdots, T$ sequentially. The single-armed bandit problem has an additional advantage in terms of computational efficiency: W^{TS} often admits a closed form expression. In cases of the Beta-Bernoulli MAB and the Gaussian MAB,

$$\mathbb{E}_{\theta\sim\mathrm{Beta}(\alpha,\beta)}\left[\max(\theta,\lambda)\right] = \lambda \times F_{\alpha,\beta}^{\mathrm{beta}}(\lambda) + \frac{\alpha}{\alpha+\beta} \times \left(1 - F_{\alpha+1,\beta}^{\mathrm{beta}}(\lambda)\right) \quad (43)$$

$$\mathbb{E}_{\theta\sim\mathcal{N}(m,\nu^{-2})}\left[\max(\theta,\lambda)\right] = m + (\lambda - m) \times \Phi(\nu(\lambda - m)) + \nu^{-1} \times \phi(\nu(\lambda - m)) \quad (44)$$

where $F_{\alpha,\beta}^{\mathrm{beta}}(\cdot)$ represents c.d.f. of Beta(α,β) distribution, $\Phi(\cdot)$ and $\phi(\cdot)$ represent c.d.f. and p.d.f. of standard normal distribution. With these expressions, Γ_n^λ's can be computed very efficiently without using numerical integration or sampling in contrast to the multi-armed case of IRS.V-EMAX.

Algorithm 4: IRS.INDEX Policy

Function `Irs.Single.Worth-Trying`$(a, T, \lambda, (\tilde{y}_{a,n})_{n\in\{0,\cdots,T\}})$

$\tilde{\Gamma}_n^\lambda \leftarrow \mathbb{E}_{\theta_a \sim \mathcal{P}_a(\tilde{y}_{a,n})}\left[\max(\mu_a(\theta_a), \lambda)\right], \forall n = 0, \cdots, T$

$\tilde{S}_{a,0}^\mu \leftarrow 0,\ \tilde{S}_0^\Gamma \leftarrow 0,\ \tilde{m}_0^\Gamma \leftarrow \tilde{\Gamma}_0^\lambda$

for $n = 1, \cdots, T$ **do**

  $\tilde{S}_{a,n}^\mu \leftarrow \tilde{S}_{a,n-1}^\mu + \bar{\mu}_a(\tilde{y}_{a,n-1})$

  $\tilde{S}_n^\Gamma \leftarrow \tilde{S}_{a,n-1}^\Gamma + \tilde{\Gamma}_n^\lambda$

  $\tilde{m}_n^\Gamma \leftarrow \min\left(\tilde{m}_{n-1}^\Gamma, \tilde{\Gamma}_{n-1}^\lambda\right)$

end

$\tilde{\varphi}_a \leftarrow \max_{1\le n\le T}\left\{\tilde{S}_{a,n}^\mu + T \times \tilde{\Gamma}_0^\lambda + (T-n) \times \left(\lambda - \tilde{m}_n^\Gamma\right) - \tilde{S}_n^\Gamma\right\} - T \times \lambda$

if $\tilde{\varphi}_a \ge 0$ **then**

  return true

else

  return false

end

Function `Irs.Index`$(T, \mathbf{y})$

Sample an outcome $\tilde{\omega} \sim \mathcal{I}(T, \mathbf{y})$

$\tilde{y}_{a,0} \leftarrow y_a, \quad \tilde{y}_{a,n} \leftarrow \mathcal{U}_a(\tilde{y}_{a,n-1}, \tilde{R}_{a,n}), \quad \forall n \in [T], \quad \forall a \in [K]$

for $a = 1, \cdots, K$ **do**

  $\tilde{\lambda}_a^* \leftarrow \inf\left\{\lambda; \texttt{Irs.Single.Worth-Trying}(a, T, \lambda, (\tilde{y}_{a,n})_{n\in\{0,\cdots,T\}}) = \text{true}\right\}$

end

return $\text{argmax}_a\, \tilde{\lambda}_a^*$

Index policy. We now return to the original MAB problem with K arms. Recall that the single-arm bandit algorithm tells us whether an arm (given a sampled future) is worth trying against the deterministic reward λ. We use the algorithm as a module to compute the index of each arm.

More specifically, for each arm $a = 1, \cdots, K$ seperately, IRS.INDEX policy samples the future belief trajectory $\{\tilde{y}_{a,n}\}_{n\in\{0,\cdots,T\}}$ and finds a threshold deterministic reward that makes the arm barely worth trying:

$$\tilde{\lambda}_a^* \triangleq \sup\left\{\lambda \in \mathbb{R}\ ;\ \tilde{\varphi}_a(\lambda) \ge 0\right\}. \tag{45}$$

Although the monotonicity of $\tilde{\varphi}_a(\cdot)$ is not theoretically proven, we observe that the binary search works well in our numerical experiments.

The value $\tilde{\lambda}_a^*$ is used as the index of an arm a. IRS.INDEX policy chooses the arm $\text{argmax}_a\, \tilde{\lambda}_a^*$ with the largest index. The entire procedure for single decision making requires $O(c_b \times KT)$ operations where c_b represents the number of iterations in binary search.

Some numerical experiment include a heuristic variation of index policy, IRS.INDEX*, that is obtained by using

$$\varphi_a(\lambda) \triangleq \max_{1\le n\le T} \left\{ \sum_{s=1}^{n} \left(\mu_{a,s-1} - \lambda - \left(\Gamma_s^\lambda - \Gamma_0^\lambda \right) \right) \right\} \tag{46}$$

instead of (42).

B. Proofs for §3

Proposition 3 (Mean equivalence). *If the penalty function z_t is dual feasible, it does not penalize any non-anticipating policy $\pi \in \Pi_{\mathbb{F}}$ in expectation, i.e.,*

$$\mathbb{E}^{\pi}_{\omega\sim\mathcal{I}(T,\mathbf{y})}\left[\sum_{t=1}^{T} r_t(\mathbf{a}^\pi_{1:t};\omega) - z_t(\mathbf{a}^\pi_{1:t};\omega)\right] = \mathbb{E}^{\pi}_{\omega\sim\mathcal{I}(T,\mathbf{y})}\left[\sum_{t=1}^{T} r_t(\mathbf{a}^\pi_{1:t};\omega)\right] \equiv V(\pi,T,\mathbf{y}). \tag{47}$$

Proof. We define an appending operator $\oplus$ that concatenates an element into a vector such that $\mathbf{a}_{1:t} \equiv \mathbf{a}_{1:t-1} \oplus a_t$. When $\pi \in \Pi_{\mathbb{F}}$ and z_t is dual feasible, omitting ω for brevity,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}^\pi_{1:t}) - z_t(\mathbf{a}^\pi_{1:t})\right] &= \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}^\pi_{1:t}) - \mathbb{E}\left[z_t(\mathbf{a}^\pi_{1:t}) \middle| \mathcal{F}_{t-1}\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}^\pi_{1:t}) - \mathbb{E}\left[\sum_{a=1}^{K} z_t(\mathbf{a}^\pi_{1:t-1}\oplus a)\cdot \mathbf{1}\{a^\pi_t = a\} \middle| \mathcal{F}_{t-1}\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(r_t(\mathbf{a}^\pi_{1:t}) - \sum_{a=1}^{K} \underbrace{\mathbb{E}\left[z_t(\mathbf{a}^\pi_{1:t-1}\oplus a) \middle| \mathcal{F}_{t-1}\right]}_{=0} \cdot \mathbf{1}\{a^\pi_t = a\} \right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}^\pi_{1:t})\right].
\end{aligned}$$

■

B.1. Proof of Theorem 1

Week duality. Define $\mathcal{G}_t \triangleq \mathcal{F}_t \cup \sigma(\omega)$ and consider a relaxed policy space $\Pi_{\mathbb{G}} \triangleq \{\pi : a^\pi_t$ is $\mathcal{G}_{t-1}$-measurable, Then, we have

$$\begin{aligned}
V^*(T,\mathbf{y}) &\triangleq \sup_{\pi\in\Pi_{\mathbb{F}}} \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}^\pi_{1:t})\right] \overset{\text{Prop 3}}{=} \sup_{\pi\in\Pi_{\mathbb{F}}} \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}^\pi_{1:t}) - z_t(\mathbf{a}^\pi_{1:t};)\right] \\
&\le \sup_{\pi\in\Pi_{\mathbb{G}}} \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}^\pi_{1:t}) - z_t(\mathbf{a}^\pi_{1:t})\right] = \mathbb{E}\left[\max_{\mathbf{a}_{1:T}\in\mathcal{A}^T} \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}) - z_t(\mathbf{a}_{1:t})\right] \\
&= W^z(T,\mathbf{y})
\end{aligned}$$

since $\Pi_{\mathbb{F}} \subseteq \Pi_{\mathbb{G}}$. ■

Strong duality. Fix T and $\mathbf{y}$. Let $V_t^{\text{in}}(\mathbf{a}_{1:t-1};\omega)$ and $Q_t^{\text{in}}(\mathbf{a}_{1:t-1},a;\omega)$ be the value function and Q-value function associated with the inner problem $(*)$ given a particular outcome ω under the ideal penalty (10). With $V_{T+1}^{\text{in}} \equiv 0$, we have the Bellman equations for the inner problem:

$$Q_t^{\text{in}}(\mathbf{a}_{1:t-1},a;\omega) \triangleq r_t(\mathbf{a}_{1:t-1}\oplus a;\omega) - z_t^{\text{ideal}}(\mathbf{a}_{1:t-1}\oplus a;\omega) + V_{t+1}^{\text{in}}(\mathbf{a}_{1:t-1}\oplus a;\omega) \quad (48)$$
$$V_t^{\text{in}}(\mathbf{a}_{1:t-1};\omega) = \max_{a\in\mathcal{A}}\left\{Q_t^{\text{in}}(\mathbf{a}_{1:t-1},a;\omega)\right\} \quad (49)$$

We argue by induction to show that

$$V_t^{\text{in}}(\mathbf{a}_{1:t-1};\omega) = V^*(T-t+1,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega)) \quad (50)$$
$$Q_t^{\text{in}}(\mathbf{a}_{1:t-1},a;\omega) = Q^*(T-t+1,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega),a) \quad (51)$$

for all $\mathbf{a}_{1:t-1}\in\mathcal{A}^{t-1}$, $a\in\mathcal{A}$ and $t\in[T]$.

As a terminal case, when $t=T+1$, the claim holds trivially, since $V_t^{\text{in}}(\mathbf{a}_{1:T};\omega)=0=V^*(0,\mathbf{y}_T(\mathbf{a}_{1:T};\omega))$. Now assume that the claim holds for $t+1$: $V_{t+1}^{\text{in}}(\mathbf{a}_{1:t};\omega)=V^*(T-t,\mathbf{y}_t(\mathbf{a}_{1:t};\omega))$ for all $\mathbf{a}_{1:t}\in\mathcal{A}^t$. For any $\mathbf{a}_{1:t-1}\in\mathcal{A}^{t-1}$ and $a\in\mathcal{A}$, then,

$$Q_t^{\text{in}}(\mathbf{a}_{1:t-1},a;\omega) = r_t(\mathbf{a}_{1:t-1}\oplus a;\omega) - z_t^{\text{ideal}}(\mathbf{a}_{1:t-1}\oplus a;\omega) + V_{t+1}^{\text{in}}(\mathbf{a}_{1:t-1}\oplus a;\omega) \quad (52)$$
$$= \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1}\oplus a;\omega) + V^*\left(T-t,\mathbf{y}_t(\mathbf{a}_{1:t-1}\oplus a;\omega)\right)\middle|\,\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1};\omega)\right] \quad (53)$$
$$\underbrace{-V^*\left(T-t,\mathbf{y}_t(\mathbf{a}_{1:t-1}\oplus a;\omega)\right) + V_{t+1}^{\text{in}}(\mathbf{a}_{1:t-1}\oplus a;\omega)}_{=0} \quad (54)$$
$$= \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1}\oplus a;\omega) + V^*\left(T-t,\mathbf{y}_t(\mathbf{a}_{1:t-1}\oplus a;\omega)\right)\middle|\,\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1};\omega)\right] \quad (55)$$
$$= \mathbb{E}_{r\sim\mathcal{R}_a(\mathcal{P}_a([\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega)]_a))}\left[r + V^*\left(T-t,\mathcal{U}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega),a,r)\right)\right] \quad (56)$$
$$= Q^*(T-t,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega),a) \quad (57)$$

where the last equality follows from the original Bellman equation (4). Consequently,

$$V_t^{\text{in}}(\mathbf{a}_{1:t-1};\omega) = \max_{a\in\mathcal{A}}\left\{Q_t^{\text{in}}(\mathbf{a}_{1:t-1},a;\omega)\right\} \quad (58)$$
$$= \max_{a\in\mathcal{A}}\left\{Q^*(T-t,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega),a)\right\} \quad (59)$$
$$= V^*(T-t,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega)) \quad (60)$$

Therefore the claim holds for all $t=1,\cdots,T$. In particular for $t=1$, we have

$$V_1^{\text{in}}(\emptyset;\omega) = V^*(T,\mathbf{y}), \quad Q_1^{\text{in}}(\emptyset,a;\omega) = Q^*(T,\mathbf{y},a), \quad \forall\omega. \quad (61)$$

Note that the maximal value of inner problem does not depend on ω – deterministic with respect to the randomness of ω. As its expected value, $W^{\text{ideal}}(T,\mathbf{y}) = V^*(T,\mathbf{y})$. ■

B.2. Proof of Remark 1

We continue on the proof of strong duality. π^{ideal} solves the same inner problem with respect to a randomly sampled outcome $\tilde{\omega}$. When the remaining time is T and the current belief is $\mathbf{y}$, it takes an action with the largest Q-value: together with (61),

$$a^{\pi^{\text{ideal}}} = \operatorname*{argmax}_a Q_1^{\text{in}}(\emptyset,a;\tilde{\omega}) = \operatorname*{argmax}_a Q^*(T,\mathbf{y},a). \quad (62)$$

Therefore, at each moment, no matter what $\tilde{\omega}$ is chosen, the policy π^{ideal} always takes the same action that Bayesian optimal policy would choose. Although there might be some ambiguity regarding tie-breaking in argmax, it does not affect the expected performance. Therefore, $V(\pi^{\text{ideal}}, T, \mathbf{y}) = V^*(T, \mathbf{y})$. ■

B.3. Proof of Remark 2

Except $z_t^{\text{IRS.FH}}$, all the other penalty functions have a form of $z_t = X_t - \mathbb{E}[X_t|\mathcal{G}_{t-1}]$ such that $\mathcal{G}_{t-1} \supseteq \mathcal{F}_{t-1}$. Since $\mathbb{E}[\mathbb{E}(X_t|\mathcal{G}_{t-1})|\mathcal{F}_{t-1}] = \mathbb{E}[X_t|\mathcal{F}_{t-1}]$ by Tower property, $\mathbb{E}[z_t|\mathcal{F}_{t-1}] = \mathbb{E}[X_t - \mathbb{E}[X_t|\mathcal{G}_{t-1}]|\mathcal{F}_{t-1}] = 0$.

For $z_t^{\text{IRS.FH}}$ (12), observe that

$$\begin{aligned}\mathbb{E}[\mathbb{E}(\mu_{a_t}(\theta_{a_t})|R_{a_t,1},\cdots,R_{a_t,T-1})|\mathcal{F}_{t-1}] &= \mathbb{E}\left[\mathbb{E}\left(\mu_{a_t}(\theta_{a_t})\middle|\{R_{a,n}\}_{a\in\mathcal{A},n\in[T-1]}\right)\middle|\mathcal{F}_{t-1}\right] \\ &= \mathbb{E}[\mu_{a_t}(\theta_{a_t})|\mathcal{F}_{t-1}] = \mathbb{E}[r_t(\mathbf{a}_{1:t};\omega)|\mathcal{F}_{t-1}]\end{aligned}$$

and thus $\mathbb{E}\left[z_t^{\text{IRS.FH}}|\mathcal{F}_{t-1}\right] = 0$. ■

C. Proofs for §4

Within this section, without loss of generality, we redefine the outcome ω to include an infinite number of future reward realizations:

$$\omega \triangleq (R_{a,n})_{a\in\mathcal{A},n\in\mathbb{N}}. \tag{63}$$

The original definition of outcome (2) can be thought as a truncated version of (63), in which only the first T reward realizations are adopted for each arm. Although ω does not contain $\boldsymbol{\theta}$, having an infinite number of rewards is sufficient since $\mu_a(\theta_a) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n R_{a,i}$ by Proposition 5.

With this new definition, we can describe the distribution of outcome irrespectively of T and we denote it with $\mathcal{I}(\mathbf{y})$ (one can imagine $\mathcal{I}(\infty, \mathbf{y})$ with respect to the original definition).

C.1. Notes on Regularity

Proposition 4. *If* $\mathbb{E}|R_{a,n}| < \infty$ *for all* a,

$$\mathbb{E}|\mu_a(\theta_a)| < \infty \quad \textit{and} \quad W^{\text{TS}}(T, \mathbf{y}) < \infty \tag{64}$$

Proof. By Jensen's inequality,

$$\mathbb{E}|\mu_a(\theta_a)| = \mathbb{E}[|\mathbb{E}(R_{a,n}|\theta_a)|] \le \mathbb{E}[\mathbb{E}(|R_{a,n}||\theta_a)] \le \mathbb{E}|R_{a,n}| < \infty \tag{65}$$

Consequently,

$$\mathbb{E}\left[\max_a \mu_a(\theta_a)\right] \le \mathbb{E}\left[\sum_{a=1}^K |\mu_a(\theta_a)|\right] = \sum_{a=1}^K \mathbb{E}|\mu_a(\theta_a)| < \infty \tag{66}$$

■

Proposition 5. *If* $\mathbb{E}|R_{a,n}| < \infty$*,*

$$\lim_{n\to\infty} \mu_{a,n}(\omega) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} R_{a,i} = \mu_a(\theta_a) \quad \textit{almost surely} \tag{67}$$

where $\mu_{a,n}(\omega) \triangleq \mathbb{E}\left[\mu_a(\theta_a)|R_{a,1},\cdots,R_{a,n}\right]$ *defined in* (18).

Proof. Fix a and let $\mathcal{H}_n \triangleq \sigma\left(R_{a,1},\cdots,R_{a,n}\right)$. First note that, by SLLN, $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} R_{a,i} = \mu_a(\theta_a)$ almost surely. Therefore, $\mu_a(\theta_a)$ is $\mathcal{H}_\infty$-measurable. Also note that $\mu_{a,n} = \mathbb{E}\left(\mu_a(\theta_a)|\mathcal{H}_n\right)$ is a Doob's Martingale adapted to $\mathcal{H}_n$. By Levy's upward theorem, since $\mu_a(\theta_a) \in \mathcal{L}^1$ by Proposition 4, $\mu_{a,n}$ converges to $\mathbb{E}\left(\mu_a(\theta_a)|\mathcal{H}_\infty\right) = \mu_a(\theta_a)$ almost surely as $n \to \infty$. ■

C.2. Proof of Proposition 1

Asymptotic behavior of Irs.FH. Let $\tilde{\omega}$ be the sampled outcome used inside of $\textsc{Irs}.\text{FH}(T,\mathbf{y})$. Following from Proposition 5, we have $\lim_{n\to\infty}\mu_{a,n}(\tilde{\omega}) = \mu_a(\tilde{\theta}_a)$ for almost all $\tilde{\omega}$. Together with the assumption that $\mu_i(\theta_i) \neq \mu_j(\theta_j)$ for $i \neq j$, since $\text{argmax}_a\, \mu_a(\tilde{\theta}_a)$ is uniquely defined for almost all $\tilde{\omega}$, we have

$$\lim_{n\to\infty} \underset{a}{\text{argmax}}\, \mu_{a,n}(\tilde{\omega}) = \underset{a}{\text{argmax}}\, \mu_a(\tilde{\theta}_a) \quad \text{a.s.} \tag{68}$$

Since the almost sure convergence guarantees the convergence in distribution, for any a,

$$\begin{aligned}
\lim_{T\to\infty} \mathbb{P}\left[\textsc{Irs}.\text{FH}(T,\mathbf{y}) = a\right] &= \lim_{T\to\infty} \mathbb{P}\left[\underset{a'}{\text{argmax}}\, \mu_{a',T-1}(\tilde{\omega}) = a\right] && (69)\\
&= \mathbb{P}\left[\underset{a'}{\text{argmax}}\, \mu_{a'}(\tilde{\theta}_{a'}) = a\right] && (70)\\
&= \mathbb{P}\left[\text{TS}(\mathbf{y}) = a\right] && (71)
\end{aligned}$$

Note that we are not particularly assuming that $\textsc{Irs}.\text{FH}(T,\mathbf{y})$ and $\text{TS}(\mathbf{y})$ share the randomness. The sampled parameters used in $\text{TS}(\mathbf{y})$ may not be necessarily same to that of $\textsc{Irs}.\text{FH}(T,\mathbf{y})$, but their distribution is identical since they are given the same prior. ■

Asymptotic behavior of Irs.V-Zero. We drop $\tilde{\omega}$ for brevity in what follows, but we use the notation of $\tilde{x}$ to indicate the dependency of x on $\tilde{\omega}$. Let $\tilde{a}_T^\circ = \textsc{Irs}.\text{V-}\textsc{Zero}(T,\mathbf{y})$ in which $\tilde{\omega}$ is used, and let $\tilde{a}^{\text{TS}} \triangleq \text{argmax}_a\, \mu_a(\tilde{\theta}_a)$. Similar to above, it suffices to show that $\lim_{T\to\infty}\tilde{a}_T^\circ = \tilde{a}^{\text{TS}}$ for almost all $\tilde{\omega}$.

Define

$$\tilde{\Delta} \triangleq \min_{a\neq\tilde{a}^{\text{TS}}} \left|\mu_{\tilde{a}^{\text{TS}}}(\tilde{\theta}_{\tilde{a}^{\text{TS}}}) - \mu_a(\tilde{\theta}_a)\right| \quad \text{and} \quad \tilde{M} \triangleq \sup_{a\in\mathcal{A},n\geq 0} |\tilde{\mu}_{a,n}|\,. \tag{72}$$

We have $0 < \tilde{\Delta} < 2\tilde{M} < \infty$ almost surely since $\mu_i(\tilde{\theta}_i) \neq \mu_j(\tilde{\theta}_j)$ for $i \neq j$ and $\lim_{n\to\infty}\tilde{\mu}_{a,n} = \mu_a(\tilde{\theta}_a) < \infty$ almost surely for all a. In addition, there exists $\tilde{N} \in \mathbb{N}$ such that

$$\left|\tilde{\mu}_{a,n} - \mu_a(\tilde{\theta}_a)\right| < \frac{\tilde{\Delta}}{4}, \quad \forall n \geq \tilde{N}, \quad \forall a \in \mathcal{A} \tag{73}$$

For such $\tilde{N}$, we have

$$\inf_{n\geq\tilde{N}} \tilde{\mu}_{\tilde{a}^{\mathrm{TS}},n} \geq \sup_{n\geq\tilde{N}} \tilde{\mu}_{a,n} + \frac{\tilde{\Delta}}{2}, \quad \forall a \neq \tilde{a}^{\mathrm{TS}} \tag{74}$$

Note that $\tilde{a}^{\mathrm{TS}}$, $\tilde{\Delta}$, $\tilde{M}$ and $\tilde{N}$ are determined only by $\tilde{\omega}$, regardless of T.

To argue by contradiction, suppose $\tilde{a}^{\circ}_T \neq \tilde{a}^{\mathrm{TS}}$ for some large T such that $T \geq 2\tilde{N} + \frac{8\tilde{M}\tilde{N}}{\tilde{\Delta}} + 2$. Define the optimal solution to the inner problem of IRS.V-ZERO:

$$\tilde{\mathbf{n}}^{\circ}_{1:K} \triangleq \operatorname*{argmax}_{\mathbf{n}_{1:K}} \left\{ \sum_{a=1}^{K}\sum_{s=1}^{n_a} \tilde{\mu}_{a,s-1} \ ; \ \sum_{a=1}^{K} n(a) = T \right\} \tag{75}$$

where the ties are broken arbitrarily in argmax{}. Given IRS.V-ZERO's selection rule $\tilde{a}^{\circ}_T = \operatorname{argmax}_a \tilde{n}^{\circ}(a)$, the assumption $\tilde{a}^{\circ}_T \neq \tilde{a}^{\mathrm{TS}}$ implies that $\tilde{n}^{\circ}(\tilde{a}^{\circ}_T) \geq \lfloor \frac{T}{2} \rfloor$ $(> \tilde{N})$.

Case 1: If $\tilde{n}^{\circ}(\tilde{a}^{\mathrm{TS}}) \geq \tilde{N}$, consider a deviation of pulling $\tilde{a}^{\mathrm{TS}}$ one more time but pulling $\tilde{a}^{\circ}_T$ one less time: define $\tilde{\mathbf{n}}^{\dagger}_{1:K}$ such that $\tilde{n}^{\dagger}(\tilde{a}^{\mathrm{TS}}) = \tilde{n}^{\circ}(\tilde{a}^{\mathrm{TS}}) + 1$, $\tilde{n}^{\dagger}(\tilde{a}^{\circ}_T) = \tilde{n}^{\circ}(\tilde{a}^{\circ}_T) - 1$ and $\tilde{n}^{\dagger}(a) = \tilde{n}^{\circ}(a)$ for $a \notin \{\tilde{a}^{\mathrm{TS}}, \tilde{a}^{\circ}_T\}$. Then, since $\tilde{n}^{\circ}(\tilde{a}^{\mathrm{TS}}) \geq \tilde{N}$ and $\tilde{n}^{\circ}(\tilde{a}^{\circ}_T) \geq \tilde{N}$, by (74),

$$\sum_{a=1}^{K}\sum_{s=1}^{\tilde{n}^{\dagger}(a)} \tilde{\mu}_{a,s-1} - \sum_{a=1}^{K}\sum_{s=1}^{\tilde{n}^{\circ}(a)} \tilde{\mu}_{a,s-1} = \tilde{\mu}_{\tilde{a}^{\mathrm{TS}},\tilde{n}^{\circ}(\tilde{a}^{\mathrm{TS}})} - \tilde{\mu}_{\tilde{a}^{\circ}_T,\tilde{n}^{\circ}(\tilde{a}^{\circ}_T)-1} \geq \frac{\tilde{\Delta}}{2} > 0 \tag{76}$$

The allocation $\tilde{\mathbf{n}}^{\dagger}_{1:K}$ achieves a strictly better payoff than $\tilde{\mathbf{n}}^{\circ}_{1:K}$, which contradicts to the assumption that $\tilde{\mathbf{n}}^{\circ}_{1:K}$ is an optimal allocation.

Case 2: If $\tilde{n}^{\circ}(\tilde{a}^{\mathrm{TS}}) < \tilde{N}$, consider a deviation $\tilde{\mathbf{n}}^{\dagger}_{1:K}$ such that

$$\tilde{n}^{\dagger}(a) \triangleq \begin{cases} \tilde{n}^{\circ}(\tilde{a}^{\mathrm{TS}}) + (\tilde{n}^{\circ}(\tilde{a}^{\circ}_T) - \tilde{N}) & \text{if } a = \tilde{a}^{\mathrm{TS}} \\ \tilde{N} & \text{if } a = \tilde{a}^{\circ}_T \\ \tilde{n}^{\circ}(a) & \text{otherwise} \end{cases} \tag{77}$$

By making this change,

$$\begin{aligned}
& \sum_{a=1}^{K}\sum_{s=1}^{\tilde{n}^{\dagger}(a)} \tilde{\mu}_{a,s-1} - \sum_{a=1}^{K}\sum_{s=1}^{\tilde{n}^{\circ}(a)} \tilde{\mu}_{a,s-1} && (78)\\
= \quad & \sum_{s=\tilde{n}^{\circ}(\tilde{a}^{\mathrm{TS}})+1}^{\tilde{n}^{\circ}(\tilde{a}^{\mathrm{TS}})+(\tilde{n}^{\circ}(\tilde{a}^{\circ}_T)-\tilde{N})} \tilde{\mu}_{\tilde{a}^{\mathrm{TS}},s-1} - \sum_{s=\tilde{N}+1}^{\tilde{n}^{\circ}(\tilde{a}^{\circ}_T)} \tilde{\mu}_{\tilde{a}^{\circ}_T,s-1} && (79)\\
\geq \quad & -(\tilde{N} - \tilde{n}^{\circ}(\tilde{a}^{\mathrm{TS}})) \cdot 2\tilde{M} + \sum_{s=\tilde{N}+1}^{\tilde{n}^{\circ}(\tilde{a}^{\circ}_T)} \tilde{\mu}_{\tilde{a}^{\mathrm{TS}},s-1} - \sum_{s=\tilde{N}+1}^{\tilde{n}^{\circ}(\tilde{a}^{\circ}_T)} \tilde{\mu}_{\tilde{a}^{\circ}_T,s-1} && (80)\\
\geq \quad & -(\tilde{N} - \tilde{n}^{\circ}(\tilde{a}^{\mathrm{TS}})) \cdot 2\tilde{M} + (\tilde{n}^{\circ}(\tilde{a}^{\circ}_T) - \tilde{N}) \cdot \frac{\Delta}{2} && (81)\\
\geq \quad & (\tilde{n}^{\circ}(\tilde{a}^{\circ}_T) - \tilde{N}) \cdot \frac{\Delta}{2} - 2\tilde{N}\tilde{M} && (82)
\end{aligned}$$

Since $T \geq 2\tilde{N} + \frac{8\tilde{M}\tilde{N}}{\tilde{\Delta}} + 2$ and $\tilde{n}^\circ(\tilde{a}_T^\circ) \geq \lfloor \frac{T}{2} \rfloor$, the last term is strictly positive, which means that $\tilde{\mathbf{n}}_{1:K}^\dagger$ is strictly better than $\tilde{\mathbf{n}}_{1:K}^\circ$. We got a contradiction.

We've shown that for almost all $\tilde{\omega}$, when T is large enough, the optimal solution $\tilde{\mathbf{n}}_{1:K}^\circ$ must impose more than a half of the pulls on the arm $\tilde{a}^{\mathrm{TS}} = \mathrm{argmax}_a\, \mu_a(\tilde{\theta}_a)$. Therefore, $\lim_{T\to\infty} \tilde{a}_T^\circ = \tilde{a}^{\mathrm{TS}}$ for almost all $\tilde{\omega}$, which completes the proof.

C.3. Proof of Theorem 2

The first inequality $W^{\mathrm{TS}}(T, \mathbf{y}) \geq W^{\textsc{Irs.FH}}(T, \mathbf{y})$ immediately follows from Jensen's inequality: since $\max(\cdots)$ is a convex function,

$$\mathbb{E}\left[\max_a \mu_a(\theta_a)\right] \geq \mathbb{E}\left[\max_a \mathbb{E}\left(\mu_a(\theta_a) \middle| R_{a,1}, \cdots, R_{a,T-1}\right)\right]. \tag{83}$$

To show $W^{\textsc{Irs.FH}}(T, \mathbf{y}) \geq W^{\textsc{Irs.V-Zero}}(T, \mathbf{y})$, we need to modify Jensen's inequality a bit.

Lemma 1 (Variant of Jensen's inequality). *Suppose $\varphi : \mathbb{R} \mapsto \mathbb{R}$ is a* ***non-decreasing*** *(deterministic) function. Then, for any real-valued random variable X such that $\mathbb{E}|X| < \infty$,*

$$\mathbb{E}\left[\max\left\{X + \varphi(X), 0\right\}\right] \geq \mathbb{E}\left[\max\left\{\mathbb{E}(X) + \varphi(X), 0\right\}\right]. \tag{84}$$

Proof. Define $\mu \triangleq \mathbb{E}(X)$ and $f_x(t) \triangleq \max\{t + \varphi(x), 0\}$. Since $f_x(\cdot)$ is a convex function for each $x \in \mathbb{R}$,

$$f_x(t) \geq f_x(\mu) + (t-\mu)\cdot f_x'(\mu) = \max\{\mu + \varphi(x), 0\} + (t-\mu)\cdot \mathbf{1}\{\mu + \varphi(x) \geq 0\}, \quad \forall t, \quad \forall x \tag{85}$$

By setting $t = x$,

$$\max\{x + \varphi(x), 0\} = f_x(x) \geq \max\{\mu + \varphi(x), 0\} + (x-\mu)\cdot \mathbf{1}\{\mu + \varphi(x) \geq 0\}, \quad \forall x \tag{86}$$

Note that, since $\mathbf{1}\{\mu+\varphi(x) \geq 0\}$ is increasing in x, (i) for any $x \geq \mu$, $(x-\mu) \geq 0$ and $\mathbf{1}\{\mu+\varphi(x)\} \geq \mathbf{1}\{\mu + \varphi(\mu)\}$, and (ii) for any $x < \mu$, $(x-\mu) < 0$ and $\mathbf{1}\{\mu + \varphi(x)\} \leq \mathbf{1}\{\mu + \varphi(\mu)\}$. Therefore,

$$(x-\mu)\cdot \mathbf{1}\{\mu + \varphi(x) \geq 0\} \geq (x-\mu)\cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\}, \quad \forall x \in \mathbb{R} \tag{87}$$

Combining with (86),

$$\max\{x + \varphi(x), 0\} \geq \max\{\mu + \varphi(x), 0\} + (x-\mu)\cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\}, \quad \forall x \in \mathbb{R} \tag{88}$$

For random variable X, by taking expectation,

$$\begin{aligned} \mathbb{E}\left[\max\{X + \varphi(X), 0\}\right] &\geq \mathbb{E}\left[\max\{\mu + \varphi(X), 0\} + (X-\mu)\cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\}\right] && (89)\\ &\geq \mathbb{E}\left[\max\{\mu + \varphi(X), 0\}\right] + \mathbb{E}(X-\mu)\cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\} && (90)\\ &= \mathbb{E}\left[\max\{\mu + \varphi(X), 0\}\right] && (91) \end{aligned}$$

■

Corollary 1. *On a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, suppose* $\varphi(x,\omega) : \mathbb{R} \times \Omega \mapsto \mathbb{R}$ *be a function such that (i) the mapping* $x \mapsto \varphi(x,\omega)$ *is* ***non-decreasing*** *for each* $\omega \in \Omega$ *and (ii) for some sub-*σ*-field* $\mathcal{H} \subseteq \mathcal{F}$*, the mapping* $\omega \mapsto \varphi(x,\omega)$ *is* $\mathcal{H}$***-measurable*** *for each* $x \in \mathbb{R}$ *(i.e.,* $\varphi(\cdot,\omega)$ *is a deterministic function conditioned on* $\mathcal{H}$*). Then*

$$\mathbb{E}\left[\max\left\{X(\omega) + \varphi(X(\omega),\omega), 0\right\}\right] \geq \mathbb{E}\left[\max\left\{\mathbb{E}(X|\mathcal{H})(\omega) + \varphi(X(\omega),\omega), 0\right\}\right] \tag{92}$$

Proof. Define

$$\mu(\omega) \triangleq \mathbb{E}(X|\mathcal{H})(\omega), \quad I(\omega) \triangleq \mathbf{1}\{\mu(\omega) + \varphi(\mu(\omega),\omega) \geq 0\} \tag{93}$$

Following from (88), we have

$$\max\{x + \varphi(x,\omega), 0\} \geq \max\{\mu(\omega) + \varphi(x,\omega), 0\} + (x - \mu(\omega)) \cdot I(\omega), \quad \forall x \in \mathbb{R}, \quad \text{for each } \omega \in \Omega \tag{94}$$

Since $\mu(\omega)$ and $I(\omega)$ are $\mathcal{H}$-measurable,

$$\begin{aligned}
\mathbb{E}\left[\max\{X(\omega) + \varphi(X(\omega),\omega), 0\}\right] &\geq \mathbb{E}\left[\max\{\mu(\omega) + \varphi(X(\omega),\omega), 0\} + (X(\omega) - \mu(\omega)) \cdot I(\omega)\right] &(95)\\
&= \mathbb{E}\left[\mathbb{E}\left(\max\{\mu(\omega) + \varphi(X(\omega),\omega), 0\} + (X(\omega) - \mu(\omega)) \cdot I(\omega) \middle| \mathcal{H}\right)\right] &(96)\\
&= \mathbb{E}\left[\max\{\mu(\omega) + \varphi(X(\omega),\omega), 0\}\right] + \mathbb{E}\left[\mathbb{E}\left(\left(X(\omega) - \mu(\omega)\right) \cdot I(\omega) \middle| \mathcal{H}\right)\right] &(97)\\
&= \mathbb{E}\left[\max\{\mathbb{E}\left(X|\mathcal{H}\right)(\omega) + \varphi(X(\omega),\omega), 0\}\right] &(98)\\
&\quad + \mathbb{E}\left[\underbrace{\left(\mathbb{E}(X|\mathcal{H})(\omega) - \mu(\omega)\right)}_{=0} \cdot I(\omega)\right] &(99)\\
&= \mathbb{E}\left[\max\{\mathbb{E}\left(X|\mathcal{H}\right)(\omega) + \varphi(X(\omega),\omega), 0\}\right] &(100)
\end{aligned}$$

■

Corollary 2. *On a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, let* $(C_0, \cdots, C_T)$ *be* $\mathcal{H}$***-measurable real-valued random variables*** *for some sub-*σ*-field* $\mathcal{H} \subseteq \mathcal{F}$ *(i.e.,* C_i*'s are constants conditioned on* $\mathcal{H}$*). Then*

$$\mathbb{E}\left[\max_{0 \leq i \leq T}\left\{(i-n)^+ \times X + C_i\right\}\right] \geq \mathbb{E}\left[\max_{0 \leq i \leq T}\left\{\mathbb{E}\left(X \middle| \mathcal{H}\right) \cdot \mathbf{1}\{i \geq n+1\} + (i-n-1)^+ \times X + C_i\right\}\right] \tag{101}$$

for any $n = 0, 1, \cdots, T$*.*

Proof. When $n = T$, both sides become $\mathbb{E}\left[\max_{0 \leq i \leq T}\left\{C_i\right\}\right]$ which makes the claim true. Fix $n < T$ and define

$$\varphi(x,\omega) \triangleq \max_{n+1 \leq i \leq T}\left\{(i-n-1) \times x + C_i(\omega)\right\} - \max_{0 \leq i \leq n}\left\{C_i(\omega)\right\}. \tag{102}$$

Note that $\varphi(x,\omega)$ satisfies the conditions in Corollary 1. Therefore,

$$\mathbb{E}\left[\max_{0\le i\le T}\left\{(i-n)^+\times X+C_i\right\}\right] \tag{103}$$

$$=\mathbb{E}\left[\max\left\{\max_{n+1\le i\le T}\{(i-n)\times X+C_i\},\quad \max_{0\le i\le n}C_i\right\}\right] \tag{104}$$

$$=\mathbb{E}\left[\max\left\{X+\max_{n+1\le i\le T}\{(i-n-1)\times X+C_i\},\quad \max_{0\le i\le n}C_i\right\}\right] \tag{105}$$

$$=\mathbb{E}\left[\max\left\{\underbrace{X(\omega)+\max_{n+1\le i\le T}\{(i-n-1)\times X(\omega)+C_i(\omega)\}-\max_{0\le i\le n}C_i(\omega)}_{=\varphi(X(\omega),\omega)},\quad 0\right\}+\max_{0\le i\le n}C_i(\omega)\right] \tag{106}$$

$$\overset{\text{Cor 1}}{\ge}\mathbb{E}\left[\max\left\{\mathbb{E}\left(X|\mathcal{H}\right)(\omega)+\max_{n+1\le i\le T}\{(i-n-1)\times X(\omega)+C_i(\omega)\}-\max_{0\le i\le n}C_i(\omega),\quad 0\right\}+\max_{0\le i\le n}C_i(\omega)\right] \tag{107}$$

$$=\mathbb{E}\left[\max\left\{\max_{n+1\le i\le T}\{\mathbb{E}\left(X|\mathcal{H}\right)+(i-n-1)\times X+C_i\},\quad \max_{0\le i\le n}C_i\right\}\right] \tag{108}$$

$$=\mathbb{E}\left[\max_{0\le i\le T}\left\{\mathbb{E}\left(X|\,\mathcal{H}\right)\cdot\mathbf{1}\{i\ge n+1\}+(i-n-1)^+\times X+C_i\right\}\right] \tag{109}$$

■

Proof of $W^{\textbf{Irs.FH}}(T,\mathbf{y})\ge W^{\textbf{Irs.V-Zero}}(T,\mathbf{y})$. Define

$$N_T\triangleq\left\{\mathbf{n}_{1:K}\in\mathbb{Z}_+^K:\sum_{a=1}^K n_a=T\right\}\quad\text{and}\quad S_a(n_a)\triangleq\sum_{s=1}^{n_a}\mu_{a,s-1} \tag{110}$$

What we want to show is

$$W^{\text{Irs.FH}}\equiv\mathbb{E}\left[T\times\max_a\{\mu_{a,T-1}\}\right]=\mathbb{E}\left[\max_{\mathbf{n}_{1:K}\in N_T}\left\{\sum_{a=1}^K n_a\times\mu_{a,T-1}\right\}\right] \tag{111}$$

$$\ge\mathbb{E}\left[\max_{\mathbf{n}_{1:K}\in N_T}\left\{\sum_{a=1}^K S_a(n_a)\right\}\right]\equiv W^{\text{Irs.V-Zero}}. \tag{112}$$

Further define

$$U_{k,n}\triangleq\mathbb{E}\left[\max_{\mathbf{n}_{1:K}\in N_T}\left\{\left(\sum_{a=1}^{k-1}S_a(n_a)\right)+\Big(S_k(n_k\wedge n)+(n_k-n)^+\times\mu_{k,T-1}\Big)+\left(\sum_{a=k+1}^K n_a\times\mu_{a,T-1}\right)\right\}\right] \tag{113}$$

where $a\wedge b\triangleq\min(a,b)$. Observe that $W^{\text{Irs.FH}}=U_{1,0}$, $W^{\text{Irs.V-Zero}}=U_{K,T}$ and $U_{k+1,0}=U_{k,T}$. Therefore, it suffices to show that

$$U_{k,n}\ge U_{k,n+1},\quad\forall k=1,\cdots,K,\quad\forall n=0,\cdots,T-1. \tag{114}$$

Fix k and n. Define a sub-σ-field

$$\mathcal{H}\triangleq\sigma\left(\{R_{a,s}\}_{a=k,s\le n}\cup\{R_{a,s}\}_{a\ne k,s\le T-1}\right). \tag{115}$$

For each $i = 0, \cdots, T$, define

$$C_i \triangleq \max_{\mathbf{n}_{1:K} \in N_T} \left\{ \left(\sum_{a=1}^{k-1} S_a(n_a) \right) + S_k(i \wedge n) + \left(\sum_{a=k+1}^{K} n_a \times \mu_{a,T-1} \right) \; ; \; n_k = i \right\}. \tag{116}$$

Note that C_i's are $\mathcal{H}$-measurable and

$$U_{k,n} = \mathbb{E}\left[\max_{0 \le i \le T} \left\{ (i-n)^+ \times \mu_{k,T-1} + C_i \right\} \right] \tag{117}$$

With $X \triangleq \mu_{k,T-1}$, by Corollary 2,

$$\begin{aligned} U_{k,n} &= \mathbb{E}\left[\max_{0 \le i \le T} \left\{ (i-n)^+ \times X + C_i \right\} \right] && (118) \\ &\overset{\text{Cor 2}}{\ge} \mathbb{E}\left[\max_{0 \le i \le T} \left\{ \mathbb{E}\left(X \middle| \mathcal{H} \right) \cdot \mathbf{1}\{i \ge n+1\} + (i-n-1)^+ \times X + C_i \right\} \right] && (119) \\ &\overset{\text{(a)}}{=} \mathbb{E}\left[\max_{0 \le i \le T} \left\{ \mu_{k,n} \cdot \mathbf{1}\{i \ge n+1\} + (i-n-1)^+ \times \mu_{k,T-1} + C_i \right\} \right] && (120) \\ &\overset{\text{(b)}}{=} U_{k,n+1} && (121) \end{aligned}$$

(a) holds since $\mathbb{E}(X|\mathcal{H}) = \mathbb{E}(\mu_{k,T-1}|\mathcal{H}) = \mathbb{E}(\mu_{k,T-1}|R_{k,1}, \cdots, R_{k,n}) = \mu_{k,n}$, and (b) holds since $S_k(i \wedge n) + \mu_{k,n} \cdot \mathbf{1}\{i \ge n+1\} = \sum_{s=1}^{n} \mu_{k,s-1} \cdot \mathbf{1}\{i \ge s\} + \mu_{k,n} \cdot \mathbf{1}\{i \ge n+1\} = \sum_{s=1}^{n+1} \mu_{k,s-1} \cdot \mathbf{1}\{i \ge s\} = S_k(i \wedge (n+1))$. ■

C.4. Proof of Theorem 3

As in §B.1, we define Q-values of inner problem with respect to a particular outcome ω, a penalty function $z_t(\cdot)$, a time horizon T and a prior belief $\mathbf{y}$.

$$\begin{aligned} Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a; \omega, T, \mathbf{y}) &= r_t(\mathbf{a}_{1:t-1} \oplus a; \omega) - z_t(\mathbf{a}_{1:t-1} \oplus a; \omega, T, \mathbf{y}) && (122) \\ &\quad + V_{t+1}^{z,\text{in}}(\mathbf{a}_{1:t-1} \oplus a; \omega, T, \mathbf{y}) \\ V_t^{z,\text{in}}(\mathbf{a}_{1:t-1}; \omega, T, \mathbf{y}) &= \max_{a \in \mathcal{A}} \left\{ Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a; \omega, T, \mathbf{y}) \right\} && (123) \end{aligned}$$

with $V_{T+1}^{z,\text{in}}(\cdot; \omega, T, \mathbf{y}) \equiv 0$. Additionally define

$$\begin{aligned} \mathcal{S}^z(\mathbf{a}_{1:T}; \omega, T, \mathbf{y}) &\triangleq \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}; \omega) - z_t(\mathbf{a}_{1:t}; \omega, T, \mathbf{y}) && (124) \\ a_t^{z,*}(\mathbf{a}_{1:t-1}; \omega, T, \mathbf{y}) &\triangleq \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a; \omega, T, \mathbf{y}) \right\} && (125) \end{aligned}$$

We have $V_1^{z,\text{in}}(\emptyset; \omega, T, \mathbf{y}) = \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \mathcal{S}^z(\mathbf{a}_{1:T}; \omega, T, \mathbf{y})$.

Proposition 6 (Suboptimality decomposition). *Given a policy $\pi \in \Pi_{\mathbb{F}}$ and a dual feasible penalty function z_t,*

$$\begin{aligned} W^z(T, \mathbf{y}) - V(\pi, T, \mathbf{y}) &= \mathbb{E}\left[\max_{\mathbf{a}_{1:T}} \{ \mathcal{S}^z(\mathbf{a}_{1:T}; \omega, T, \mathbf{y}) \} - \mathcal{S}^z(\mathbf{a}_{1:T}^{\pi}; \omega, T, \mathbf{y}) \right] && (126) \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \max_{a'} \left\{ Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}^{\pi}, a'; \omega, T, \mathbf{y}) \right\} - Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}^{\pi}, a_t^{\pi}; \omega, T, \mathbf{y}) \right] && (127) \end{aligned}$$

where the expectation is taken with respect to the randomness of outcome ω and the randomness of policy π.

Proof. The first equality immediately follows from the definition of W^z and the mean-equivalence (Proposition 3). Now fix ω, T and $\mathbf{y}$. Consider the (pathwise) suboptimality of a particular action sequence $\mathbf{a}_{1:T}$ compared to the clairvoyant optimal solution. It can decomposed into the suboptimalty gap of individual actions over time.

$$\max_{\mathbf{a}'_{1:T}} \{\mathcal{S}^z(\mathbf{a}'_{1:T})\} - \mathcal{S}^z(\mathbf{a}_{1:T}) = \sum_{t=1}^{T} \max_{a'} \left\{ Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a') \right\} - Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a_t) \tag{128}$$

By taking expectation, we obtain the second equality. ■

We continue with the extended definition of outcome (63). Define a shift operator $\mathcal{M}_t : \mathcal{A}^t \times \Omega \mapsto \Omega$.

$$\mathcal{M}_t(\mathbf{a}_{1:t}, \omega) \triangleq (R_{a,n_a}; \forall n_a > n_t(\mathbf{a}_{1:t}, a), \forall a \in \mathcal{A}) \tag{129}$$

$\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)$ encodes the remaining reward realizations after taking $\mathbf{a}_{1:t-1}$.

Remark 4 (Recursive structure of remaining uncertainties). *Conditioned on $\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}; \omega)$, the remaining uncertainties are sufficiently described by $\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}; \omega)$.*

$$\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega) | \, \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}; \omega) \quad \sim \quad \mathcal{I}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}; \omega)) \tag{130}$$

Remark 5 (Recursive structure in IRS penalties). *Each of the penalty functions* (10)-(14) *has the following form*

$$z_t(\mathbf{a}_{1:t}; \omega, T, \mathbf{y}) = \varphi^z(\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}; \omega), T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}; \omega)) \tag{131}$$

for some function $\varphi^z : \Omega \times \mathbb{N} \times \mathcal{Y} \mapsto \mathbb{R}$: the penalty is completely determined by the remaining rewards $\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}; \omega)$, the remaining time horizon $T - t + 1$ and the belief $\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}; \omega)$ at that moment.

Remark 4 immediately follows from Bayes' rule, and Remark 5 can be easily verified. We observe a recursive structure in the sequential inner problems that the DM solves throughout the decision making process, characterized by the following property.

Proposition 7 (Generalized posterior sampling). *For any of penalty functions* (10)-(14)*, the IRS policy π^z is randomized in a way that it chooses an action a with the probability that the action a is indeed the best action $a_t^{z,*}$ at that moment: i.e.,*

$$\mathbb{P}\left[a_t^{\pi^z} = a \middle| \, \mathcal{F}_{t-1} \right] = \mathbb{P}\left[a_t^{z,*} = a \middle| \, \mathcal{F}_{t-1} \right], \quad \forall a, \quad \forall t. \tag{132}$$

The source of uncertainty in LHS is the randomness of decision maker (embedded in $\tilde{\omega}$) and that of RHS is the randomness of nature (embedded in ω). $a_t^{z,*}$ abbreviates $a_t^{z,*}(\mathbf{a}^{\pi^z}_{1:t-1}; \omega, T, \mathbf{y})$ defined in (125) *and $\mathcal{F}_{t-1}$ abbreviates $\mathcal{F}_{t-1}(\mathbf{a}^{\pi^z}_{1:t-1}; \omega)$. Here we assume that the tie-breaking rule in* argmax *of* (125) *is identical to the one used in π^z when solving the inner problem.*

Proof. Fix t, $\mathbf{a}_{1:t-1}$ and ω. First, $a_t^{z,*}$ is the best action that maximizes payoff in the remaining periods:

$$a_t^{z,*}(\mathbf{a}_{1:t-1};\omega,T,\mathbf{y}) = \underset{a'_t}{\operatorname{argmax}}\left\{\max_{\mathbf{a}'_{t+1:T}}\sum_{s=t}^{T} r_s(\mathbf{a}_{1:t-1}\oplus\mathbf{a}'_{t:s};\omega) - z_s(\mathbf{a}_{1:t-1}\oplus\mathbf{a}'_{t:s};\omega,T,\mathbf{y})\right\} \tag{133}$$

Following from Remark 5, for any $s\in[t,T]$,

$$z_s(\mathbf{a}_{1:t-1}\oplus\mathbf{a}'_{t:s};\omega,T,\mathbf{y}) \tag{134}$$
$$= \varphi^z(\mathcal{M}_{s-1}(\mathbf{a}_{1:t-1}\oplus\mathbf{a}'_{t:s-1};\omega),T-s+1,\mathbf{y}_{s-1}(\mathbf{a}_{1:t-1}\oplus\mathbf{a}'_{t:s-1};\omega,\mathbf{y})) \tag{135}$$
$$= \varphi^z(\mathcal{M}_{s-t}(\mathbf{a}'_{t:s-1};\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1},\omega)),(T-t+1)+(s-t),\mathbf{y}_{s-t}(\mathbf{a}'_{t:s-1};\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1},\omega),\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega)) \tag{136}$$
$$= z_{s-t+1}(\mathbf{a}'_{t:s};\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1},\omega),T-t+1,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega)) \tag{137}$$

Similarly for rewards, we have $r_s(\mathbf{a}_{1:t-1}\oplus\mathbf{a}'_{t:s};\omega) = r_{s-t+1}(\mathbf{a}'_{t:s};\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1},\omega))$. Therefore, (134) is reformulated as

$$a_t^{z,*} = \underset{a'_t}{\operatorname{argmax}}\left\{\max_{\mathbf{a}'_{t+1:T}}\sum_{s=t}^{T} r_{s-t+1}(\mathbf{a}'_{t:s};\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1},\omega)) - z_{s-t+1}(\mathbf{a}'_{t:s};\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1},\omega),T-t+1,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega))\right\} \tag{138}$$

Next, consider the IRS policy's action $a_t^{\pi^z}$. It internally solves an instance of inner problem with sampled outcome $\tilde\omega\sim\mathcal{I}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega))$, the remaining horizon $T-t+1$ and the prior belief $\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega)$:

$$a_t^{\pi^z} = \underset{a'_1}{\operatorname{argmax}}\left\{\max_{\mathbf{a}'_{2:T-t+1}}\sum_{s=1}^{T-t+1} r_s(\mathbf{a}'_{1:s};\tilde\omega) - z_s(\mathbf{a}'_{1:s};\tilde\omega,T-t+1,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega))\right\}. \tag{139}$$

Comparing (138) and (139), we observe that they have the identical functional forms except $\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1},\omega)$ is replaced with $\tilde\omega$. Since $\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1},\omega)|\mathcal{F}_{t-1}\sim\mathcal{I}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega))$ (Remark 4) and $\tilde\omega\sim\mathcal{I}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1};\omega))$,

$$\mathbb{P}\left[a_t^{z,*}(\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1},\omega)) = a\middle|\,\mathcal{F}_{t-1}\right] = \mathbb{P}\left[a_t^{\pi^z}(\tilde\omega) = a\middle|\,\mathcal{F}_{t-1}\right]. \tag{140}$$

■

Proof outline of Theorem 3. From now on, we restrict our attention to the Beta-Bernoulli MAB. We basically mirror the proof of Russo and Van Roy (2014) while generalizing the meaning of posterior sampling.

For each of penalty functions z_t^{TS}, $z_t^{\text{IRS.FH}}$ and $z_t^{\text{IRS.V-ZERO}}$, we will construct a confidence interval process $\{(L_{a,t},U_{a,t})\}_{a\in\mathcal{A},t\in[T]}$ such that each of $(L_{a,t},U_{a,t})$'s is (i) $\mathcal{F}_{t-1}$-measurable and (ii) regulates the suboptimality of action a at time t: more specifically, (ii) means that

$$Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1},a_t^{z,*}) - Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1},a_t) \le U_{a_t^{z,*},t} - L_{a_t,t},\quad \forall a_t\in\mathcal{A} \tag{**}$$

holds with a high probability $1-\delta$. Following from Proposition 6,

$$W^z(T,\mathbf{y}) - V(\pi^z, T, \mathbf{y}) \tag{141}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}^{\pi}, a_t^{z,*}) - Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}^{\pi}, a_t^{\pi})\right] \tag{142}$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{P}_{t-1}[(**)\text{ fails}] + \mathbb{E}_{t-1}\left[Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}^{\pi}, a_t^{z,*}) - Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}^{\pi}, a_t^{\pi})\middle| (**)\text{ holds}\right]\right] \tag{143}$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \delta + \mathbb{E}_{t-1}\left[U_{a_t^{z,*},t} - L_{a_t^{\pi},t}\right]\right] \tag{144}$$

$$= T\delta + \mathbb{E}\left[\sum_{t=1}^{T} U_{a_t^{\pi},t} - L_{a_t^{\pi},t}\right] \tag{145}$$

where $\mathbb{P}_{t-1}[\cdot] \triangleq \mathbb{P}[\cdot|\mathcal{F}_{t-1}]$ and $\mathbb{E}_{t-1}[\cdot] \triangleq \mathbb{E}[\cdot|\mathcal{F}_{t-1}]$. The last equality follows from

$$\mathbb{E}_{t-1}\left[U_{a_t^{z,*},t}\right] = \sum_{a=1}^{K} U_{a,t} \times \mathbb{P}_{t-1}\left[a_t^{z,*} = a\right] = \sum_{a=1}^{K} U_{a,t} \times \mathbb{P}_{t-1}\left[a_t^{\pi} = a\right] = \mathbb{E}_{t-1}\left[U_{a_t^{\pi},t}\right] \tag{146}$$

by the predictability of $U_{a,t}$ with respect to $\mathbb{F}$ and Proposition 7. Note that (145) accumulates $U_{a_t^{\pi},t} - L_{a_t^{\pi},t}$, which is the length of confidence interval of the action a_t^{π} taken by the policy at each time. We will show that, whenever the policy pulls an arm a, the confidence interval of that arm shrinks, and therefore the entire suboptimality cannot grow too fast.

Some facts in the Beta-Bernoulli MAB. Before proving Theorem 3, we characterize the Bayesian estimate $\mu_{a,n}$ in the Beta-Bernoulli MAB problem in which $\theta_a \sim \text{Beta}(\alpha_a, \beta_a)$ and $R_{a,n} \sim \text{Bernoulli}(\theta_a)$. After observing the first n reward realizations, recall that the Bayesian update results in

$$\theta_a|(R_{a,1},\cdots,R_{a,n}) \sim \text{Beta}\left(\alpha_a + \sum_{s=1}^{n} R_{a,n}, \beta_a + n - \sum_{s=1}^{n} R_{a,n}\right), \quad \mu_{a,n} = \frac{\alpha_a + \sum_{s=1}^{n} R_{a,s}}{\alpha_a + \beta_a + n}. \tag{147}$$

Note that $\{\mu_{a,n}\}_{n\geq 0}$ is a Martingale such that starts from $\mu_{a,0} = \frac{\alpha_a}{\alpha_a+\beta_a}$ and converges to $\lim_{n\to\infty} \mu_{a,n} = \theta_a$. Roughly speaking, the (unconditional) distribution of $\mu_{a,n}$, starting from a point mass $\frac{\alpha_a}{\alpha_a+\beta_a}$, diffuses toward $\text{Beta}(\alpha_a, \beta_a)$ which is the prior distribution of θ_a.[5] In the following lemma, we further characterize the distribution of $\mu_{a,n}$ more formally.

Lemma 2. *$\mu_{a,n}$ is $\frac{n}{4(\alpha_a+\beta_a)(\alpha_a+\beta_a+n)}$-sub-Gaussian, i.e.,*

$$\mathbb{E}\left[\exp\left(\lambda(\mu_{a,n} - \mathbb{E}[\mu_{a,n}])\right)\right] \leq \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4(\alpha_a+\beta_a)(\alpha_a+\beta_a+n)}\right), \quad \forall \lambda \in \mathbb{R} \tag{148}$$

Proof. Since (i) $\mathbb{E}[\mu_{a,n}] = \mu_{a,0} = \frac{\alpha_a}{\alpha_a+\beta_a}$, (ii) $R_{a,n}$'s are i.i.d. conditioned on θ_a, (iii) $\text{Bernoulli}(\theta_a)$ is $\frac{1}{4}$-sub-Gaussian (for any θ_a) and (iv) $\text{Beta}(\alpha,\beta)$ is $\frac{1}{4(\alpha+\beta+1)}$-sub-Gaussian (Marchal and Arbel,

[5] Conditioned on θ_a, $\{\mu_{a,n}\}_{n\geq 0}$ is no longer a Martingale and the distribution of $\mu_{a,n}$ starts from a point mass $\frac{\alpha_a}{\alpha_a+\beta_a}$, diffuses for a while, and ends up at a point mass θ_a. With the randomness of θ_a, $\{\mu_{a,n}\}_{n\geq 0}$ is a Martingale and the distribution of $\mu_{a,n}$ gets widened as n increases.

2017). For any $\lambda \in \mathbb{R}$,

$$
\begin{aligned}
& \mathbb{E}\left[\exp\left(\lambda(\mu_{a,n} - \mu_{a,0})\right)\right] && (149)\\
= \;& \mathbb{E}\left[\exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times \left((\alpha_a + \sum_{s=1}^{n} R_{a,s}) - (\alpha_a + \beta_a + n)\mu_{a,0}\right)\right)\right] && (150)\\
\overset{\text{(i)}}{=} \;& \mathbb{E}\left[\exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times \left(\sum_{s=1}^{n}(R_{a,s} - \theta_a) + n \cdot (\theta_a - \mu_{a,0})\right)\right)\right] && (151)\\
= \;& \mathbb{E}\left[\mathbb{E}\left\{\exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times \sum_{s=1}^{n}(R_{a,s} - \theta_a)\right)\middle|\theta_a\right\} \times \exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times n \cdot (\theta_a - \mu_{a,0})\right)\right] && (152)\\
\overset{\text{(ii)}}{=} \;& \mathbb{E}\left[\mathbb{E}\left\{\exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times (R_{a,1} - \theta_a)\right)\middle|\theta_a\right\}^n \times \exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times n \cdot (\theta_a - \mu_{a,0})\right)\right] && (153)\\
\overset{\text{(iii)}}{\leq} \;& \mathbb{E}\left[\left\{\exp\left(\frac{\lambda^2}{2(\alpha_a + \beta_a + n)^2} \times \frac{1}{4}\right)\right\}^n \times \exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times n \cdot (\theta_a - \mu_{a,0})\right)\right] && (154)\\
= \;& \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4(\alpha_a + \beta_a + n)^2}\right) \times \mathbb{E}\left[\exp\left(\frac{\lambda n}{\alpha_a + \beta_a + n} \times (\theta_a - \mu_{a,0})\right)\right] && (155)\\
\overset{\text{(iv)}}{\leq} \;& \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4(\alpha_a + \beta_a + n)^2}\right) \times \exp\left(\frac{\lambda^2 n^2}{2(\alpha_a + \beta_a + n)^2} \times \frac{1}{4(\alpha_a + \beta_a + 1)}\right) && (156)\\
\leq \;& \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4(\alpha_a + \beta_a + n)^2}\right) \times \exp\left(\frac{\lambda^2 n^2}{2(\alpha_a + \beta_a + n)^2} \times \frac{1}{4(\alpha_a + \beta_a)}\right) && (157)\\
= \;& \exp\left(\frac{\lambda^2}{2} \times \frac{n(\alpha_a + \beta_a) + n^2}{4(\alpha_a + \beta_a + n)^2(\alpha_a + \beta_a)}\right) = \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4(\alpha_a + \beta_a + n)(\alpha_a + \beta_a)}\right) && (158)
\end{aligned}
$$

■

(1) Suboptimality analysis of TS (23)**.** Define

$$\Delta_{a,t} \triangleq \sqrt{\frac{\log T}{n^{\pi}_{t-1}(a)}}, \quad U_{a,t} \triangleq \min\left\{\mu_{a,n^{\pi}_{t-1}(a)} + \Delta_{a,t}, 1\right\}, \quad L_{a,t} \triangleq \max\left\{\mu_{a,n^{\pi}_{t-1}(a)} - \Delta_{a,t}, 0\right\} \tag{159}$$

where $n^{\pi}_{t-1}(a) \triangleq n_{t-1}(\mathbf{a}^{\pi}_{1:t-1}, a)$ represents how many times the policy π had pulled an arm a before time t. $(L_{a,t}, U_{a,t})$ constructs the confidence interval on $\mu_a(\theta_a)$ $(= \theta_a)$ at time t and it is $\mathcal{F}_{t-1}$-measurable. Conditioned on $\mathcal{F}_{t-1}$, $\mu_a(\theta_a)$ is distributed with Beta$(\alpha_a + \sum_{s=1}^{n^{\pi}_{t-1}(a)} R_{a,s}, \beta_a + n^{\pi}_{t-1}(a) - \sum_{s=1}^{n^{\pi}_{t-1}(a)} R_{a,s})$ which is $\frac{1}{4(\alpha_a+\beta_a+n^{\pi}_{t-1}(a)+1)}$-sub-Gaussian. By Chernoff inequality,

$$
\begin{aligned}
\mathbb{P}_{t-1}\left[\mu_a(\theta_a) \geq U_{a,t}\right] &= \mathbb{P}_{t-1}\left[\mu_a(\theta_a) - \mu_{a,n^{\pi}_{t-1}(a)} \geq \Delta_{a,t}\right] \leq \exp\left(-\frac{\Delta^2_{a,t}}{2 \times \left(4(\alpha_a + \beta_a + n^{\pi}_{t-1}(a) + 1)\right)^{-1}}\right) && (160)\\
&\leq \exp\left(-2n^{\pi}_{t-1}(a) \times \frac{\log T}{n^{\pi}_{t-1}(a)}\right) = \frac{1}{T^2} && (161)
\end{aligned}
$$

Similarly, we have $\mathbb{P}_{t-1}\left[\mu_a(\theta_a) \leq L_{a,t}\right] \leq \frac{1}{T^2}$. We define an event $\mathcal{E}$ in which $(L_{a,t}, U_{a,t})$ is indeed a valid confidence interval for every arm a at every time t:

$$\mathcal{E} \triangleq \left\{\mu_a(\theta_a) \in (L_{a,t}, U_{a,t}), \quad \forall a, \quad \forall t\right\} \tag{162}$$

Following from the above concentration inequalities, the sequence of confidence intervals fails to contain the true mean $\mu_a(\theta_a)$ with a very low probability:

$$\mathbb{P}\left[\mathcal{E}^c\right] \leq \mathbb{E}\left[\sum_{a=1}^{K}\sum_{t=1}^{T}\mathbb{P}_{t-1}\left[\mu_a(\theta_a) \geq U_{a,t}\right] + \mathbb{P}_{t-1}\left[\mu_a(\theta_a) \leq L_{a,t}\right]\right] \leq \frac{2K}{T}. \tag{163}$$

With z_t^{TS}, the Q-value of the inner problem is

$$Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}, a_t) = \mu_{a_t}(\theta_{a_t}) + (T-t) \times \mu_{a_t^*}(\theta_{a_t^*}) \tag{164}$$

On the event $\mathcal{E}$, in which $\mu_a(\theta_a) \in (L_{a,t}, U_{a,t})$ for all a, we have

$$Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}, a_t^*) - Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}, a_t) = \mu_{a_t^*}(\theta_{a_t^*}) - \mu_{a_t}(\theta_{a_t}) \leq U_{a_t^*,t} - L_{a_t,t}. \tag{165}$$

As outlined earlier, the total suboptimality of π^{TS} is limited by

$$W^{\mathrm{TS}}(T,\mathbf{y}) - V(\pi^{\mathrm{TS}}, T, \mathbf{y}) \leq T \times \mathbb{P}\left[\mathcal{E}^c\right] + \mathbb{E}\left[\sum_{t=1}^{T} U_{a_t^\pi,t} - L_{a_t^\pi,t}\right] \leq 2K + \mathbb{E}\left[\sum_{a=1}^{K}\sum_{t=1}^{T}\min(1, 2\Delta_{a,t}) \cdot \mathbf{1}\{a_t^\pi = a\}\right]. \tag{166}$$

For each arm $a = 1, \cdots, K$,

$$\sum_{t=1}^{T}\min(1, 2\Delta_{a,t}) \cdot \mathbf{1}\{a_t^\pi = a\} \leq 1 + \sum_{n=2}^{n_T^\pi(a)} 2\sqrt{\frac{\log T}{n-1}} \leq 1 + 2\sqrt{\log T} \times \int_{x=0}^{n_T^\pi(a)} \frac{dx}{\sqrt{x}} \leq 1 + 4\sqrt{\log T} \times \sqrt{n_T^\pi(a)} \tag{167}$$

By Cauchy-Schwartz inequality and since $\sum_{a=1}^{K} n_T^\pi(a) = T$,

$$\sum_{a=1}^{K}\left(1 + 4\sqrt{\log T} \times \sqrt{n_T^\pi(a)}\right) \leq K + 4\sqrt{\log T} \times \sqrt{K\sum_{a=1}^{K} n_T^\pi(a)} = K + 4\sqrt{\log T} \times \sqrt{KT}. \tag{168}$$

Combining all results,

$$W^{\mathrm{TS}}(T,\mathbf{y}) - V(\pi^{\mathrm{TS}}, T, \mathbf{y}) \leq 3K + 4\sqrt{\log T} \times \sqrt{KT}. \tag{169}$$

■

(2) Suboptimality analysis of Irs.FH (24)**.** Note that $z_t^{\mathrm{IRS.FH}}$ yields

$$Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}, a_t^*) - Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}, a_t) = \mu_{a_t^*, n_{t-1}^\pi(a_t^*)+T-t} - \mu_{a_t, n_{t-1}^\pi(a_t)+T-t} \tag{170}$$

When $t=1$, $\mu_{a,n_{t-1}^\pi(a)+T-t}$ coincides with $\mu_{a,T-1}$. We need to bound $\mu_{a_t^*,n_{t-1}^\pi(a_t^*)+T-t}$ instead of $\mu_a(\theta_a)$. Note that, conditioned on $\mathcal{F}_{t-1}$, $\{\mu_{a,n_{t-1}^\pi(a)+n}\}_{n\geq 0}$ is a Martingale whose distribution starts from a point mass $\mu_{a,n_{t-1}^\pi(a)}$ and diffuses toward the prior distribution $\mathrm{Beta}(\alpha_a + \sum_{s=1}^{n_{t-1}^\pi(a)} R_{a,s}, \beta_a + n_{t-1}^\pi(a) - \sum_{s=1}^{n_{t-1}^\pi(a)} R_{a,s})$. For any a and $n \geq 0$, by Lemma 2, we have

$$\begin{aligned}\mathbb{E}_{t-1}\left[\exp(\lambda(\mu_{a,n_{t-1}^\pi(a)+n} - \mu_{a,n_{t-1}^\pi(a)}))\right] &\leq \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4(\alpha_a+\beta_a+n_{t-1}^\pi(a))(\alpha_a+\beta_a+n_{t-1}^\pi(a)+n)}\right) && (171)\\ &\leq \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4n_{t-1}^\pi(a)(n_{t-1}^\pi(a)+n)}\right) && (172)\end{aligned}$$

With $n = T - t$, we can conclude that $\mu_{a_t^*, n_{t-1}^\pi(a_t^*)+T-t}$ is $\frac{T-t}{4n_{t-1}^\pi(a)(T-t+n_{t-1}^\pi(a))}$-sub-Gaussian.

Define

$$\Delta_{a,t} \triangleq \sqrt{\frac{T-t}{n_{t-1}^\pi(a)+T-t} \times \frac{\log T}{n_{t-1}^\pi(a)}}, \quad U_{a,t} \triangleq \min\left\{\mu_{a,n_{t-1}^\pi(a)} + \Delta_{a,t}, 1\right\}, \quad L_{a,t} \triangleq \max\left\{\mu_{a,n_{t-1}^\pi(a)} - \Delta_{a,t}, 0\right\} \tag{173}$$

By Chernoff inequality,

$$\mathbb{P}_{t-1}\left[\mu_{a_t^*, n_{t-1}^\pi(a_t^*)+T-t} \geq U_{a,t}\right] = \mathbb{P}_{t-1}\left[\mu_{a,T-1} - \mu_{a,n_{t-1}^\pi(a)} \geq \Delta_{a,t}\right] \tag{174}$$

$$\leq \exp\left(-\frac{\Delta_{a,t}^2}{2 \times \frac{T-t}{4n_{t-1}^\pi(a)(T-t+n_{t-1}^\pi(a))}}\right) = \exp\left(-2\log T\right) = \frac{1}{T^2} \tag{175}$$

Similarly, we can show $\mathbb{P}_{t-1}\left[\mu_{a_t^*, n_{t-1}^\pi(a_t^*)+T-t} \leq L_{a,t}\right] \leq \frac{1}{T^2}$. Analogous to the proof of TS,

$$W^{\mathrm{TS}}(T, \mathbf{y}) - V(\pi^{\mathrm{TS}}, T, \mathbf{y}) \leq 2K + \mathbb{E}\left[\sum_{a=1}^{K}\sum_{t=1}^{T} \min(1, 2\Delta_{a,t}) \cdot \mathbf{1}\{a_t^\pi = a\}\right]. \tag{176}$$

Since $n_{t-1}^\pi(a) \leq t$, we have $\frac{T-t}{n_{t-1}^\pi(a)+T-t} = \left(1 + \frac{n_{t-1}^\pi(a)}{T-t}\right)^{-1} \leq \left(1 + \frac{n_{t-1}^\pi(a)}{T-n_{t-1}^\pi(a)}\right)^{-1} = 1 - \frac{n_{t-1}^\pi(a)}{T}$ and

$$\Delta_{a,t} \leq \sqrt{\left(1 - \frac{n_{t-1}^\pi(a)}{T}\right) \times \frac{\log T}{n_{t-1}^\pi(a)}} = \sqrt{\log T} \times \sqrt{\frac{1}{n_{t-1}^\pi(a)} - \frac{1}{T}} \leq \sqrt{\log T} \times \left(\frac{1}{\sqrt{n_{t-1}^\pi(a)}} - \frac{\sqrt{n_{t-1}^\pi(a)}}{2T}\right). \tag{177}$$

Consequently, for each a,

$$\sum_{t=1}^{T} \min(1, 2\Delta_{a,t}) \cdot \mathbf{1}\{a_t^\pi = a\} \leq 1 + 2\sqrt{\log T} \times \sum_{n=2}^{n_T^\pi(a)} \left(\frac{1}{\sqrt{n-1}} - \frac{\sqrt{n-1}}{2T}\right) \tag{178}$$

$$\leq 1 + 2\sqrt{\log T} \times \int_{x=0}^{n_T^\pi(a)} \left(\frac{1}{\sqrt{x}} - \frac{\sqrt{x}}{2T}\right) \tag{179}$$

$$= 1 + 2\sqrt{\log T} \times \left(2\sqrt{n_T^\pi(a)} - \frac{(n_T^\pi(a))^{3/2}}{3T}\right) \tag{180}$$

Note that, since $x^{3/2}$ is a convex function and $\sum_{a=1}^{K} n_T^\pi(a) = T$,

$$\sum_{a=1}^{K} (n_T^\pi(a))^{3/2} \geq \sum_{a=1}^{K} \left(\frac{T}{K}\right)^{3/2} = \sqrt{T^3/K} \tag{181}$$

By Cauchy-Schwarz inequality, as in TS, we have $\sum_{a=1}^{K} \sqrt{n_T^\pi(a)} \leq \sqrt{KT}$. As a result,

$$W^{\mathrm{TS}}(T, \mathbf{y}) - V(\pi^{\mathrm{TS}}, T, \mathbf{y}) \leq 2K + \mathbb{E}\left[\sum_{a=1}^{K} 1 + 2\sqrt{\log T} \times \left(2\sqrt{n_T^\pi(a)} - \frac{(n_T^\pi(a))^{3/2}}{3T}\right)\right] \tag{182}$$

$$\leq 3K + 2\sqrt{\log T} \times \left(2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right). \tag{183}$$

■

(3) Suboptimality analysis of Irs.V-Zero (25)**.** Consider an optimal solution $\mathbf{n}^*$ of inner problem of IRS.V-ZERO when the remaining time is T. As long as the optimal solution allocates at least one pull on an arm a, i.e., $n^*(a) > 0$, pulling the arm a does not incur suboptimality (such arms are all optimal and their Q-values tie). The suboptimality is incurred only when pulling a suboptimal arm a (such that $n^*(a) = 0$), in which we lose $\min_{a':n^*(a')>0}\{\mu_{a',n^*(a')-1}\} - \mu_{a,0}$ (it loses the last pull of one of optimal arms) where the term $\min_{a':n^*(a')>0}\{\mu_{a',n^*(a')-1}\}$ is limited by $\max_{0\leq n\leq T-1}\mu_{a^*,n}$ for some a^* such that $n^*(a^*) > 0$. Extending this argument, in the midst of the process, when the remaining time is $T - t + 1$, we have

$$Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a_t^*) - Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a_t) \leq \max_{0\leq n\leq T-t}\left\{\mu_{a_t^*,n_{t-1}^\pi(a_t^*)+n}\right\} - \mu_{a_t,n_{t-1}^\pi(a_t)}. \tag{184}$$

What we need to do is the regulation of $\max_{0\leq n\leq T-t}\left\{\mu_{a_t^*,n_{t-1}^\pi(a_t^*)+n}\right\}$. As before, we define

$$\Delta_{a,t} \triangleq \sqrt{\frac{T-t}{n_{t-1}^\pi(a)+T-t}\times\frac{\log T}{n_{t-1}^\pi(a)}}, \quad U_{a,t} \triangleq \min\left\{\mu_{a,n_{t-1}^\pi(a)} + \Delta_{a,t}, 1\right\}, \quad L_{a,t} \triangleq \mu_{a,n_{t-1}^\pi(a)} \tag{185}$$

Note that we take $L_{a,t}$ different from the previous case, but still $\mathcal{F}_{t-1}$-measurable. Given that $\{\mu_{a,n_{t-1}^\pi(a)+n} - \mu_{a,n_{t-1}^\pi(a)}\}_{n\geq 0}$ is a Martingale, $\left\{\exp\left(\lambda(\mu_{a,n_{t-1}^\pi(a)+n} - \mu_{a,n_{t-1}^\pi(a)})\right)\right\}_{n\geq 0}$ is a non-negative supermartingale due to the convexity of $\exp(\cdot)$. By Doob's maximal inequality and Lemma 2, for any $\lambda \geq 0$,

$$\begin{aligned}
\mathbb{P}_{t-1}\left[\max_{0\leq n\leq T-t}\left\{\mu_{a,n_{t-1}^\pi(a)+n}\right\}\geq U_{a,t}\right] &= \mathbb{P}_{t-1}\left[\max_{0\leq n\leq T-t}\left\{\mu_{a,n_{t-1}^\pi(a)+n} - \mu_{a,n_{t-1}^\pi(a)}\right\}\geq \Delta_{a,t}\right] &(186)\\
&\leq \mathbb{P}_{t-1}\left[\max_{0\leq n\leq T-t}\left\{\exp\left(\lambda(\mu_{a,n_{t-1}^\pi(a)+n} - \mu_{a,n_{t-1}^\pi(a)})\right)\right\}\geq \Delta_{a,t}\right] &(187)\\
&\leq \frac{\mathbb{E}_{t-1}\left[\exp\left(\lambda(\mu_{a,n_{t-1}^\pi(a)+T-t} - \mu_{a,n_{t-1}^\pi(a)})\right)\right]}{\exp(\lambda\Delta_{a,t})} &(188)\\
&\leq \exp\left(\frac{\lambda^2}{2}\times\frac{T-t}{4n_{t-1}^\pi(a)(n_{t-1}^\pi(a)+T-t)} - \lambda\Delta_{a,t}\right) &(189)
\end{aligned}$$

For λ that minimizes RHS and $\Delta_{a,t}$ defined above,

$$\mathbb{P}_{t-1}\left[\max_{0\leq n\leq T-t}\left\{\mu_{a,n_{t-1}^\pi(a)+n}\right\}\geq U_{a,t}\right] \leq \exp\left(-\frac{2n_{t-1}^\pi(a)(n_{t-1}^\pi(a)+T-t)}{T-t}\times\Delta_{a,t}^2\right) = \frac{1}{T^2} \tag{190}$$

Note that $\max_{0\leq n\leq T-t}\left\{\mu_{a,n_{t-1}^\pi(a)+n}\right\} \geq L_{a,t} \equiv \mu_{a,n_{t-1}^\pi(a)}$ by its defintion. We have shown that

$$\mathbb{P}\left[\mathcal{E} \triangleq \left\{\max_{0\leq n\leq T-t}\left\{\mu_{a,n_{t-1}^\pi(a)+n}\right\} \in [L_{a,t}, U_{a,t}), \quad \forall a, \forall t\right\}\right] \geq 1 - \frac{K}{T} \tag{191}$$

Therefore, using the facts derived for TS and IRS.FH,

$$\begin{aligned}
W^{\text{TS}}(T,\mathbf{y}) - V(\pi^{\text{TS}},T,\mathbf{y}) &\leq T\mathbb{P}[\mathcal{E}^c] + \mathbb{E}\left[\sum_{t=1}^T U_{a_t^\pi,t} - L_{a_t^\pi,t}\right] &(192)\\
&\leq K + \mathbb{E}\left[\sum_{a=1}^K\sum_{t=1}^T \min(1,\Delta_{a,t})\mathbf{1}\{a_t^\pi = a\}\right] &(193)\\
&\leq 2K + \sqrt{\log T}\times\left(2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right). &(194)
\end{aligned}$$

■

D. Numerical Experiments in Detail

D.1. Setups

In this section, we illustrate the detailed simulation procedure. Given a MAB problem instance specified by the prior distribution $\mathcal{P}_a(y_a)$ and the reward distribution $\mathcal{R}_a(\theta_a)$, we simulate the policies and calculate the IRS bounds for each T separately up to $T_{\max}$.

Let S be the number of simulations we perform. For each $i \in [S]$, we first sample the parameters $\theta_a^{(i)} \sim \mathcal{P}_a(y_a)$ and the rewards $R_{a,n}^{(i)} \sim \mathcal{R}_a(\theta_a^{(i)})$ for all $n \in [T_{\max}]$ and $a \in \mathcal{A}$. And then for each $T \in \{5, 10, 15, \cdots, T_{\max}\}$, we simulate each policy π (that can utilize the time horizon T): when the policy pulls an arm a_t^π at time t, it earns a reward $R^{(i)}_{a_t^\pi, n_t(\mathbf{a}^\pi_{1:t}, a_t^\pi)}$ and this reward realization is feedbacked to the policy; and this procedure is repeated over $t = 1, \cdots, T$. After simulating one sample path $\mathbf{a}^\pi_{1:T}$, $\sum_{t=1}^T \mu_{a_t^\pi}(\theta^{(i)}_{a_t^\pi})$ is recorded as a performance of π for the i^{th} sample, and the expected performance $V(\pi, T, \mathbf{y})$ is measured by its sample average over S samples.

For each of IRS bounds, similarly, we solve the associated inner problem over the same set of samples $\tilde{\omega}^{(1)}, \cdots, \tilde{\omega}^{(S)}$ for each $T \in \{5, 10, \cdots, T_{\max}\}$. The IRS bound $W^z(T, \mathbf{y})$ is evaluated by taking average of the maximal values over S samples.

More explicitly, we use the following sample averages to calculate $V(\pi, T, \mathbf{y})$ and $W^z(T, \mathbf{y})$:

$$V(\pi, T, \mathbf{y}) \approx \frac{1}{S} \sum_{i=1}^S \left(\sum_{t=1}^T \mu_{a_t^\pi}(\theta^{(i)}_{a_t^\pi}) \right), \quad W^z(T, \mathbf{y}) \approx \frac{1}{S} \sum_{i=1}^S \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^T r_t(\mathbf{a}_{1:t}; \omega^{(i)}) - z_t(\mathbf{a}_{1:t}; \omega^{(i)}) \right\}. \tag{195}$$

Note again that the same outcome $\omega^{(i)}$ is used across the different values of time horizon T and across algorithms. Sharing the randomness enhances the consistency of the estimates across T and algorithms.

Based on $V(\pi, T, \mathbf{y})$ and $W^{\text{TS}}(T, \mathbf{y})$ measured with sample averages, we calculate the Bayesian regret of a policy π: for each T,

$$W^{\text{TS}}(T, \mathbf{y}) - V(\pi, T, \mathbf{y}) = \mathbb{E}\left[\sum_{t=1}^T \max_a \mu_a(\theta_a) - \mu_{a_t^\pi}(\theta_{a_t^\pi}) \right], \tag{196}$$

which coincides with its conventional definition. We also interpret $W^{\text{TS}}(T, \mathbf{y}) - W^z(T, \mathbf{y})$ as a regret (lower) bound following from z_t, since $W^{\text{TS}}(T, \mathbf{y}) - V(\pi, T, \mathbf{y}) \geq W^{\text{TS}}(T, \mathbf{y}) - W^z(T, \mathbf{y})$ for any $\pi \in \Pi_{\mathbb{F}}$ by weak duality. In what follow, we use 20,000 samples ($S = 20,000$).

D.2. Beta-Bernoulli MAB

Two arms ($K = 2$). We first provide the results for Beta-Bernoulli MAB with two arms in which

$$\theta_a \sim \text{Beta}(1, 1), \quad R_{a,n} \sim \text{Bernoulli}(\theta_a), \quad \forall a \in [K]. \tag{197}$$

We consider relatively short time horizons ($\leq T_{\max} = 200$) since we focus on the finite horizon regime rather than asymptotic regime.

Figure 2 shows the regrets (solid lines) and the regret bounds (dashed lines) of all algorithms being considered in our paper. In particular for this case where the state (belief) space is discrete and small in its size, $O(T^4)$, we are able to solve Bellman equations (4) numerically that is shown with a curve labeled as Opt. Compared to Opt, we observe that all policies are nearly optimal. In Table 2, even for TS that exhibits the worst performance, its regret differs only by 1.5 compared to Opt when $T = 200$, which means that it chooses the suboptimal arms only six times more than Opt does, in average.

Amongst IRS algorithms, we observe a clear improvement in both performances and bounds as we incorporate more complicated penalty functions from TS to Irs.V-EMax. As visualized in Figure 2, the regret curve approaches to Opt from above and the bound curve approaches from below, which is consistent with the implication of Theorem 3. The strong duality states that, with ideal penalty, those two curves would meet at Opt, yielding zero suboptimality gap. However, we face a trade-off between the running time and the quality of policy/bound, as shown in Table 2.

We finally remark the near-optimality of Irs.Index policy. It outperforms to all the other policies, surprisingly close to Opt. Despite that it is developed based on Irs.V-EMax, it performs better than Irs.V-EMax, which leaves a necessity of further studies.

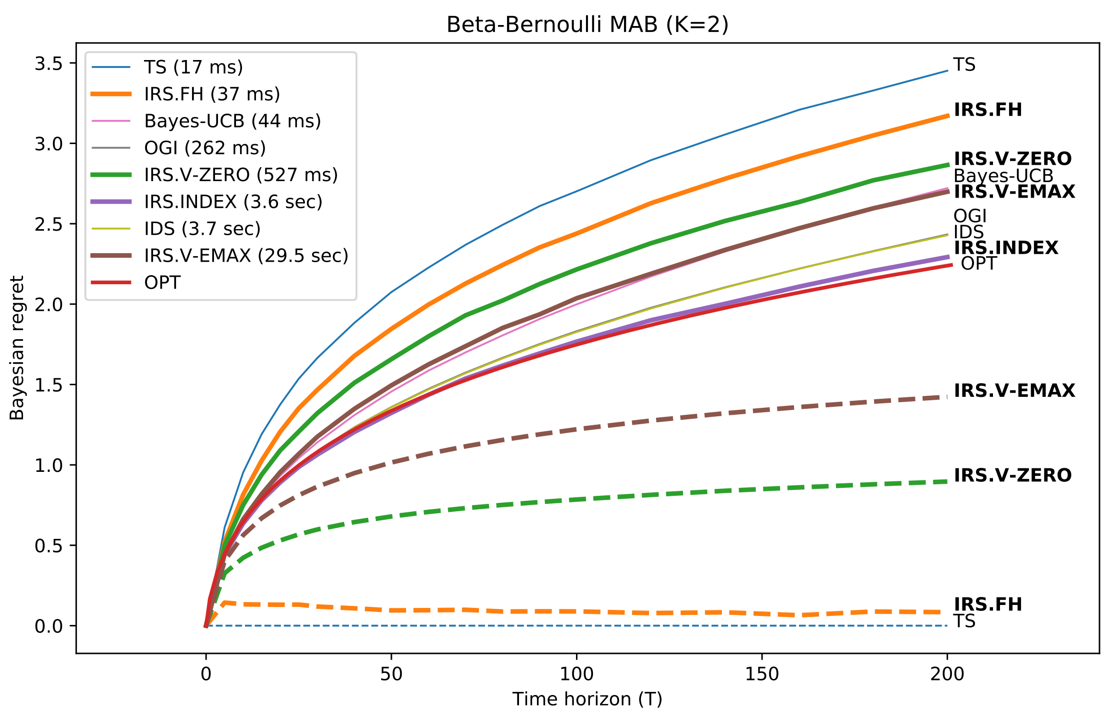

Figure 2: Regret plot for Beta-Bernoulli MAB with two arms.

Algorithm	Bayesian regret (std error)	Performance bound (std error)	Running time
TS	3.45 (0.021)	133.22 (0.332)	17 ms
IRS.FH	3.17 (0.020)	133.13 (0.331)	37 ms
IRS.V-ZERO	2.87 (0.021)	132.32 (0.318)	527 ms
IRS.V-EMAX	2.70 (0.020)	131.79 (0.008)	29.5 sec
IRS.INDEX	2.29 (0.023)	-	3.6 sec
BAYES-UCB	2.72 (0.020)	-	44 ms
IDS	2.43 (0.028)	-	3.7 sec
OGI	2.43 (0.028)	-	262 ms
OPT	2.24 (-)	131.09 (-)	-

Table 2: Summary statistics of the algorithms in Beta-Bernoulli MAB when $K = 2$ and $T = 200$. The last column shows the average time required to simulate one sample path throughout $t = 1, \cdots, T$.

Ten arms ($K = 10$). We consider Beta-Bernoulli MAB with ten arms ($K = 10$) and Beta$(1, 1)$ priors. Figure 3 and Table 3 show the results for the time horizons $T \leq T_{\max} = 500$. We no longer have IRS.V-EMAX and OPT because of the computation cost. Nevertheless, we still observe that IRS algorithms have a monotonicity in both performances and bounds, and IRS.INDEX policy performs best.

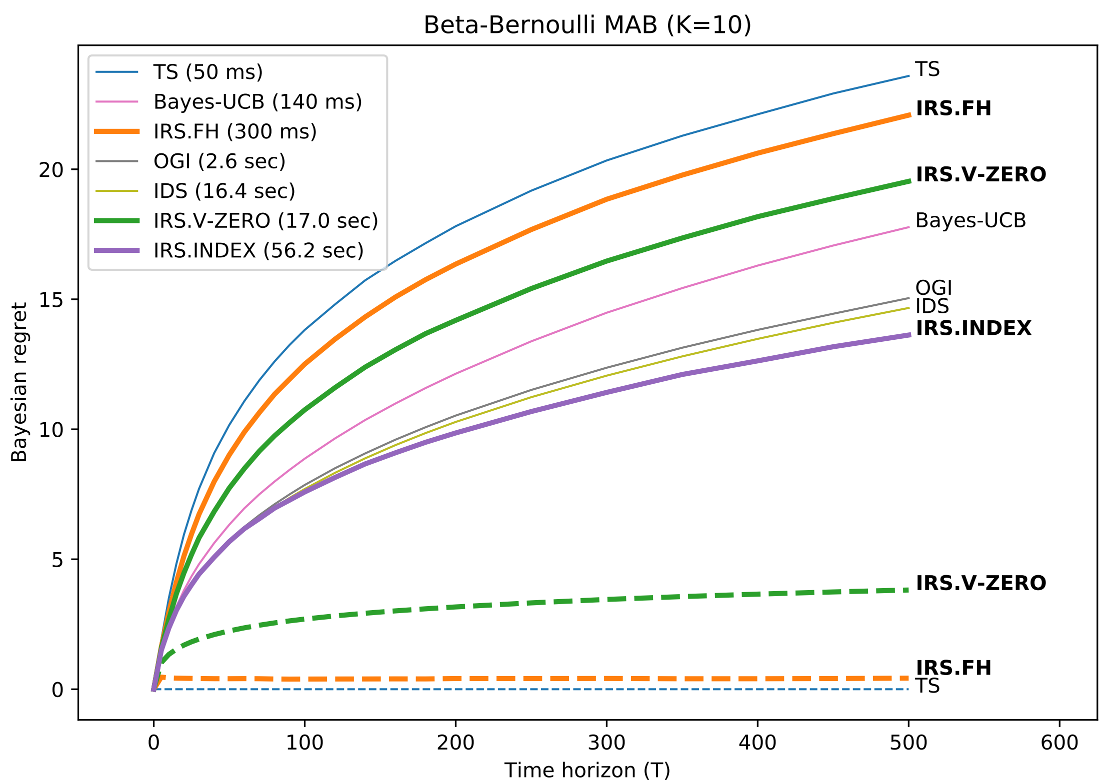

Figure 3: Regret plot for Beta-Bernoulli MAB with ten arms.

Algorithm	Bayesian regret (std error)	Performance bound (std error)	Running time
TS	23.59 (0.078)	454.01 (0.298)	50 ms
Irs.FH	22.08 (0.076)	453.59 (0.297)	300 ms
Irs.V-Zero	19.54 (0.074)	450.20 (0.290)	17.0 sec
Irs.Index	13.62 (0.080)	-	56.2 sec
Bayes-UCB	17.77 (0.077)	-	140 ms
IDS	14.67 (0.093)	-	16.4 sec
OGI	15.04 (0.092)	-	2.6 sec

Table 3: Summary statistics of algorithms in Beta-Bernoulli MAB when $K = 10$ and $T = 500$.

D.3. Gaussian MAB

Two arms ($K = 2$). We consider a case such that

$$\theta_a \sim \mathcal{N}(0, 1^2), \quad R_{a,n} \sim \mathcal{N}(\theta_a, 1^2), \quad \forall a \in [K]. \tag{198}$$

As shown in Figure 4 and Table 4, we observe the results similar to Beta-Bernoulli MABs.

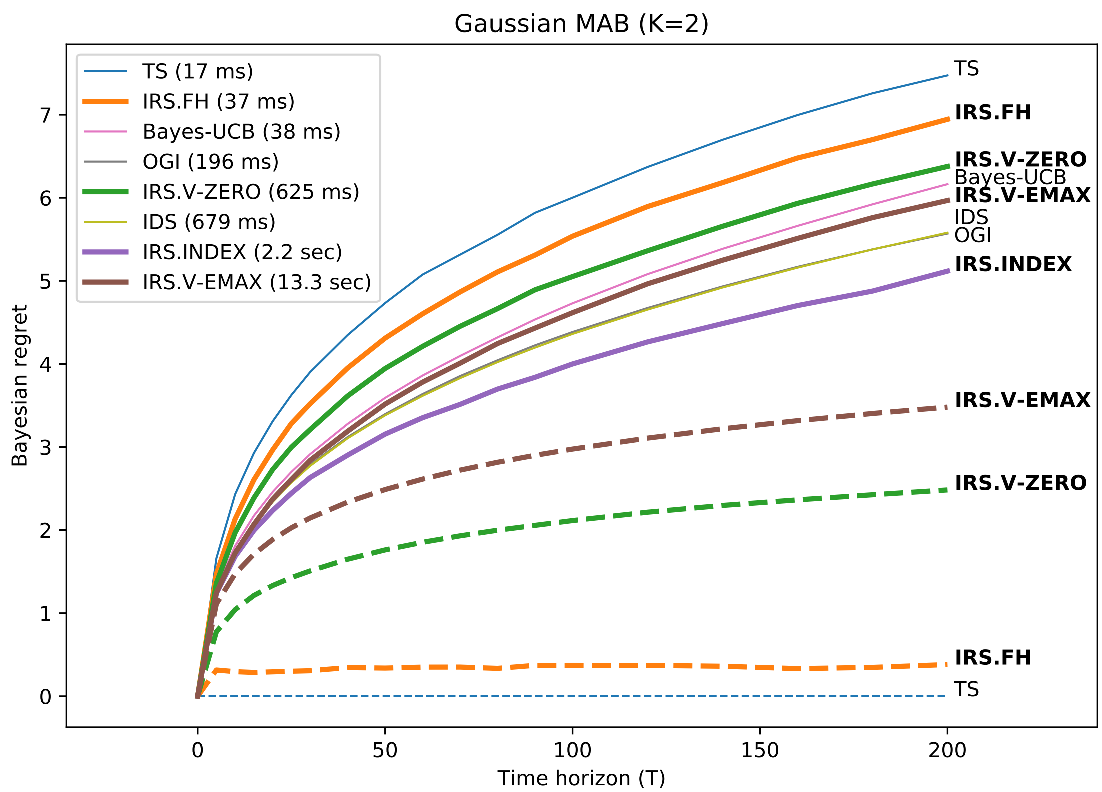

Figure 4: Regret plot for Gaussian MAB with two arms.

Algorithm	Bayesian regret (std error)	Performance bound (std error)	Running time
TS	7.47 (0.047)	112.75 (1.162)	17 ms
IRS.FH	6.94 (0.045)	112.37 (1.163)	37 ms
IRS.V-ZERO	6.38 (0.048)	110.27 (1.150)	625 ms
IRS.V-EMAX	5.97 (0.044)	109.27 (0.018)	13.3 sec
IRS.INDEX	5.12 (0.054)	-	2.2 sec
BAYES-UCB	6.16 (0.045)	-	38 ms
IDS	5.58 (0.068)	-	679 ms
OGI	5.57 (0.067)	-	196 ms

Table 4: Summary statistics of algorithms in Gaussian MAB when $K = 2$ and $T = 200$.

Ten arms ($K = 10$). The results are provided in Figure 5 and Table 5.

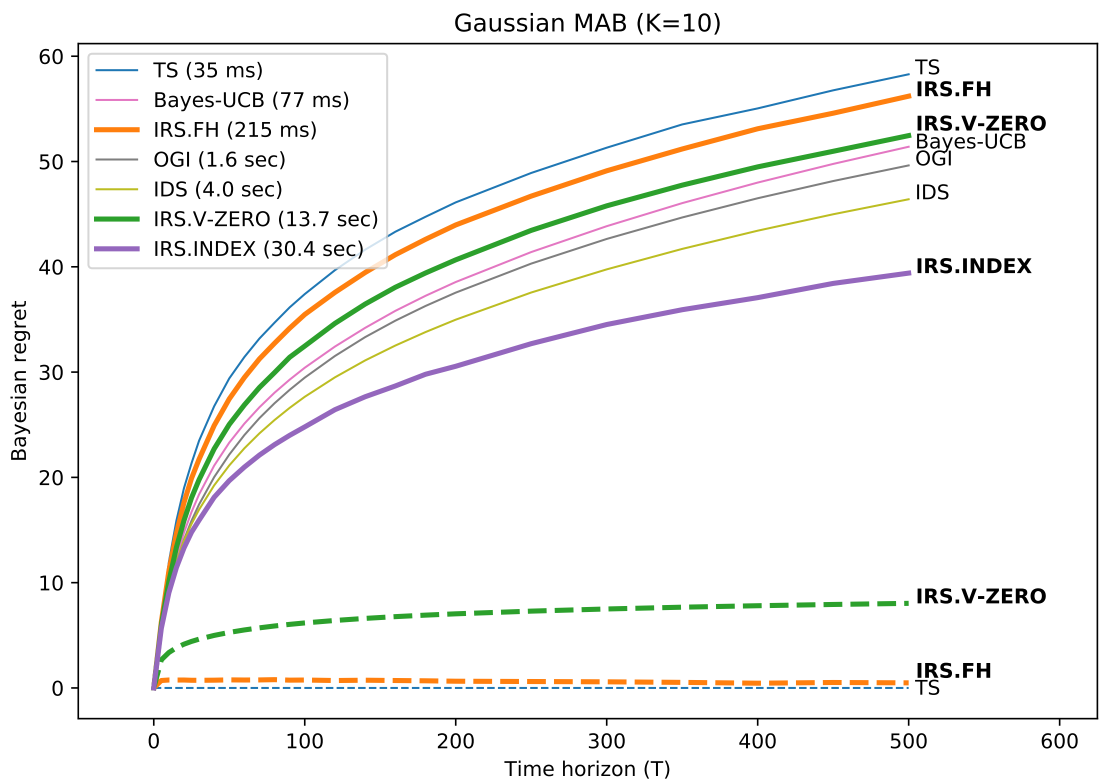

Figure 5: Regret plot for Guassian MAB with ten arms.

Algorithm	Bayesian regret (std error)	Performance bound (std error)	Running time
TS	58.28 (0.180)	766.88 (2.069)	35 ms
IRS.FH	56.20 (0.180)	766.40 (2.068)	215 ms
IRS.V-ZERO	52.46 (0.188)	758.85 (2.049)	13.7 sec
IRS.INDEX	39.40 (0.244)	-	30.4 sec
BAYES-UCB	51.40 (0.178)	-	77 ms
IDS	46.41 (0.324)	-	4.0 sec
OGI	49.63 (0.335)	-	1.6 sec

Table 5: Summary statistics of algorithms in Gaussian MAB when $K = 10$ and $T = 500$.

Five arms with different noise variances. We provide Table 6 that shows the results for the case discussed in §5.

Algorithm	Bayesian regret (std error)	Performance bound (std error)	Running time
TS	121.99 (0.615)	585.17 (2.377)	34 ms
IRS.FH	103.03 (0.628)	573.42 (2.337)	128 ms
IRS.V-ZERO	89.59 (0.690)	546.71 (2.248)	7.4 sec
IRS.INDEX	100.20 (0.657)	-	12.8 sec
IRS.INDEX*	72.43 (0.866)	-	12.3 sec
BAYES-UCB	220.66 (1.285)	-	88 ms
IDS	94.63 (0.817)	-	2.9 sec
OGI	151.61 (1.030)	-	829 ms

Table 6: Summary statistics of algorithms in Gaussian MAB when $K = 5$, $T = 500$ and $\boldsymbol{\sigma}_{1:K} = (0.1,\ 0.4,\ 1,\ 4,\ 10)$. IRS.INDEX* is a variant of IRS.INDEX introduced in §A.3.