Thompson Sampling with Information Relaxation Penalties*

Seungki Min
KAIST
skmin@kaist.ac.kr

Costis Maglaras
Columbia University
c.maglaras@gsb.columbia.edu

Ciamac C. Moallemi
Columbia University
ciamac@gsb.columbia.edu

Current Revision: February 2022

Abstract

We consider a finite-horizon multi-armed bandit (MAB) problem in a Bayesian setting, for which we propose an *information relaxation sampling* framework. With this framework, we define an intuitive family of control policies that include Thompson sampling (TS) and the Bayesian optimal policy as endpoints. Analogous to TS, which, at each decision epoch pulls an arm that is best with respect to the randomly sampled parameters, our algorithms sample entire future reward realizations and take the corresponding best action. However, this is done in the presence of "penalties" that seek to compensate for the availability of future information.

We develop several novel policies and performance bounds for MAB problems that vary in terms of improving performance and increasing computational complexity between the two endpoints. Our policies can be viewed as natural generalizations of TS that simultaneously incorporate knowledge of the time horizon and explicitly consider the exploration-exploitation trade-off. We prove associated structural results on performance bounds and suboptimality gaps. Numerical experiments suggest that this new class of policies perform well, in particular in settings where the finite time horizon introduces significant exploration-exploitation tension into the problem. Finally, inspired by the finite-horizon Gittins index, we propose an index policy that builds on our framework that particularly outperforms the state-of-the-art algorithms in our numerical experiments.

s

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 a state that is an information

---

state describing the beliefs of unknown parameters that evolve stochastically upon each play of an arm according to Bayes' rule.

Under the objective of expected performance, where the expectation is taken with respect to the prior distribution over unknown parameters, the (Bayesian) optimal policy (OPT) 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 Bayesian setting, and attempt to apply ideas from dynamic programming (DP) to develop tractable policies with good performance.

To this end, we apply the idea of *information relaxation* (Brown et al., 2010), a technique that provides a systematic way of obtaining the performance bounds on the optimal policy. In multi-period stochastic DP problems, admissible policies are required to make decisions based only on previously revealed information. The idea of information relaxation is to consider non-anticipativity as a constraint imposed on the policy space that can be relaxed, while simultaneously introducing a penalty for this relaxation into the objective, as in the usual Lagrangian relaxations of convex duality theory. Under such a relaxation, the decision maker (DM) is allowed to access future information and is asked to solve an optimization problem so as to maximize her total reward, in the presence of penalties that punish any violation of the non-anticipativity constraint. When the penalties satisfy a condition (dual feasibility, formally defined in §3), the expected value of the maximal reward adjusted by the penalties provides an upper bound on the expected performance of the (non-anticipating) optimal policy.

The idea of relaxing the non-anticipativity constraint has been studied in different contexts (Rockafellar and Wets, 1991; Davis and Karatzas, 1994; Rogers, 2002; Haugh and Kogan, 2004), and was later formulated as a formal framework by Brown et al. (2010), upon which our methodology is developed. This framework has been applied to a variety of applications including optimal stopping problems (Desai et al., 2012b); linear-quadratic and linear-convex control (Desai et al., 2012a; Haugh and Lim, 2012); dynamic portfolio execution (Haugh and Wang, 2014); and more (e.g., Brown and Haugh, 2017; Haugh and Lacedelli, 2019). Typically, the application of this method to a specific class of MDPs requires custom analysis. In particular, it is not always easy to determine penalty functions that (1) yield a relaxation that is tractable to solve, and (2) provide tight upper bounds on the performance of the optimal policy. Moreover, the established information relaxation theory focuses on upper bounds and provides no guidance on the development of tractable policies.

Our contribution is to apply the information relaxation techniques to the finite-horizon stochastic MAB problem, explicitly exploiting the structure of a Bayesian learning process. In particular,

1. we propose a series of information relaxations and penalties of increasing computational

2

complexity;

2. we systematically obtain the upper bounds on the best achievable expected performance that trade off between tightness and computational complexity;

3. and we develop associated (randomized) policies that generalize Thompson sampling (TS) in the finite-horizon setting.

In our framework, which we call *information relaxation sampling*, each of the penalty functions (and information relaxations) determines one policy and one performance bound given a particular problem instance specified by the time horizon and the prior beliefs. As a base case for our algorithms, we have TS (Thompson, 1933) and the conventional regret benchmark that has been used for Bayesian regret analysis since Lai and Robbins (1985). At the other extreme, the optimal policy OPT and its expected performance follow from the "ideal" penalty (which, not surprisingly, is intractable to compute). 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 example, one of our algorithms, IRS.FH, is a very simple modification of TS that naturally incorporates time horizon T. Recalling that TS makes a decision based on sampled parameters for each arm from the posterior distribution in each epoch, observe that knowledge of the parameters is essentially (assuming Bayesian consistency) as informative as having an infinite number of future reward observations from each arm. By contrast, IRS.FH makes a decision based on future Bayesian estimates, updated with only $T-1$ future reward realizations for each arm, where the rewards are sampled based on the inital posterior belief. 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 of the sequential decision-making process of TS, IRS.FH naturally performs less exploration than TS as the remaining time horizon diminishes. This mitigates a common practical criticism of TS: it explores too much.

Beyond this, we propose other algorithms that more explicitly quantify the benefit of exploration and more explicitly trade off between exploration and exploitation, at the cost of additional computational complexity. As we increase the 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. By providing natural generalizations of TS, our work provides both a deeper understanding of TS and improved policies that do not require tuning. Since TS has been shown to be asymptotically regret optimal in some settings, e.g., by the metric of growth-rate (Kaufmann et al., 2012b) or by the metric of worst-case regret (Agrawal and Goyal, 2013; Bubeck and Liu, 2013), our improvements can at best be (asymptotically) constant factor improvements by that metric. On the other hand, TS is extremely popular in practice, and we demonstrate in numerical examples that the improvements can be significant and are likely to be of practical interest.

Moreover, we develop upper bounds on performance that are useful in their own right. Suppose

that a decision maker faces a particular problem instance and is considering any particular MAB policy (be it one we suggest or otherwise). By simulating the policy, we can find a lower bound on the performance of the optimal policy. We introduce a series of upper bounds that can also be evaluated in any problem instance via simulation. Paired with the lower bound, these provide a computational, simulation-based "confidence interval" that can be helpful to the decision maker. For example, if the upper bound and lower bound are close, the suboptimality gap of the policy under consideration is guaranteed to be small, and it is not worth investing in better policies.

2. Finite-horizon Bayesian Multi-armed Bandit

2.1. Problem

We consider a Bayesian MAB problem with K *independent arms* and a *finite time horizon* T. More specifically, we define an MAB instance with a tuple $(K, T, \mathcal{R}, \Theta, \mathcal{P}, \mathcal{Y}, \mathbf{y})$ as follows. In each period $t = 1, \ldots, T$, the decision maker (DM) selects one among K arms, each of which yields a stochastic reward whenever selected. We let $\mathcal{A} \triangleq \{1, \ldots, K\}$ denote the set of arms, and let $R_{a,n}$ denote the random variable that represents the reward from the nth pull[1] of arm $a \in \mathcal{A}$. For each arm a, the rewards $\{R_{a,n}\}_{n \in \mathbb{N}}$ are independent and identically distributed according to the distribution $\mathcal{R}_a(\theta_a)$, where $\theta_a \in \Theta_a$ is the *parameter* associated with arm a:

$$R_{a,n} \sim \mathcal{R}_a(\theta_a), \quad \forall n \in \mathbb{N}, \quad \forall a \in \mathcal{A}. \tag{1}$$

The parameter θ_a is unknown to the DM, and is modeled as a random variable for which we have a family of *conjugate priors* $\{\mathcal{P}_a(y_a)\}_{y_a \in \mathcal{Y}_a}$, i.e., a space of distributions for θ_a that is closed under a Bayesian update with a reward realization $R_{a,n}$. Given a *hyperparameter* $y_a \in \mathcal{Y}_a$ (also called a *belief*), consider a probability measure $\mathbb{P}_{y_a}[\cdot]$ under which the parameter θ_a follows the *prior distribution* $\mathcal{P}_a(y_a)$:

$$\theta_a \sim \mathcal{P}_a(y_a), \quad \forall a \in \mathcal{A}. \tag{2}$$

Let $\mathbb{E}_{y_a}[\cdot]$ denote the expected value under this probability measure. For brevity, denote the vector of parameters and hyperparameters across arms by $\boldsymbol{\theta} \triangleq (\theta_1, \ldots, \theta_K)$ and $\mathbf{y} \triangleq (y_1, \ldots, y_K)$, respectively. Define $\mathcal{R}$, Θ, $\mathcal{P}$, $\mathcal{Y}$, $\mathbb{P}_{\mathbf{y}}$, and $\mathbb{E}_{\mathbf{y}}$ analogously. We will often describe an MAB instance only with a tuple $(T, \mathbf{y})$ when the other components are clear in context.

Throughout the paper, we assume that the rewards are absolutely integrable for each hyperparameter $y_a \in \mathcal{Y}_a$:

$$\mathbb{E}_{y_a}[|R_{a,1}|] < \infty, \quad \forall y_a \in \mathcal{Y}_a, \ a \in \mathcal{A}, \tag{3}$$

---

[1]One may consider an alternative stochastic model for the reward realization process in which the rewards are defined through a time index (e.g., $R_{a,t}$ denotes the reward from arm a in period t). This would be mathematically equivalent from the perspective of the DM. However, once the information set is relaxed, such a model is *not* equivalent to ours: in our model, the DM is not allowed to skip any future reward realizations, and this is crucial for some of the algorithms suggested in this paper. See the discussion in §3.3.

where the expectation is taken with respect to the random realization of the parameter θ_a and also with respect to the random realization of the reward $R_{a,1}$.

We further define the *outcome* $\omega \in \Omega$ (also referred to as the future or scenario) as a combination of the parameters and all future reward realizations, i.e.,

$$\omega \triangleq \left(\boldsymbol{\theta}, (R_{a,n})_{a \in \mathcal{A}, n \in \mathbb{N}}\right) \quad \sim \quad \mathcal{I}(\mathbf{y}), \tag{4}$$

that encodes all the uncertainties that the DM encounters in the environment and whose distribution is denoted by $\mathcal{I}(\mathbf{y})$.

Policy. Given an outcome ω, the reward at time t can be represented as a function of the DM's action sequence $\mathbf{a}_{1:t} = (a_1, \ldots, a_t) \in \mathcal{A}^t$, i.e.,

$$r_t(\mathbf{a}_{1:t}, \omega) \triangleq R_{a_t, n_t(\mathbf{a}_{1:t}, a_t)}, \tag{5}$$

where $n_t(\mathbf{a}_{1:t}, a) \triangleq \sum_{s=1}^{t} \mathbf{1}\{a_s = a\}$ counts how many times an arm a has been played up to time t (inclusive). Consequently, we define the *history* $H_t(\mathbf{a}_{1:t}, \omega)$ as the information revealed to the DM up to time t when taking an action sequence $\mathbf{a}_{1:t}$ given the outcome ω:

$$H_t(\mathbf{a}_{1:t}, \omega) \triangleq (a_1, r_1(a_1, \omega), a_2, r_2(\mathbf{a}_{1:2}, \omega), \ldots, a_t, r_t(\mathbf{a}_{1:t}, \omega)). \tag{6}$$

Let $\mathbf{A}_{1:t}^{\pi}$ be the action sequence taken under the DM's policy π. We can define the natural filtration $\mathbb{F} \triangleq (\mathcal{F}_t)_{t=0,1,\ldots,T}$ where $\mathcal{F}_t \triangleq \sigma\left(H_t(\mathbf{A}_{1:t}^{\pi}, \omega)\right)$ is the σ-field generated by the history H_t.

A policy π is called *non-anticipating* if every action A_t^{π} is measurable with respect to $\mathcal{F}_{t-1}$; i.e., each decision is made based only on the information revealed prior to that time. We denote by $\Pi_{\mathbb{F}}$ the set of all non-anticipating policies, including randomized ones. The (Bayesian) *performance* of a policy π is measured by the total reward that π earns on average, i.e.,

$$V(\pi, T, \mathbf{y}) \triangleq \mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} r_t(\mathbf{A}_{1:t}^{\pi}, \omega)\right], \tag{7}$$

where T and $\mathbf{y}$ specify, respectively, the length of the time horizon and the prior hyperparameters of given the MAB instance.

Bayesian update. Whenever the DM observes a reward realization, as a Bayesian learner, she can update her belief associated with the selected arm according to Bayes' rule. More formally, we introduce a *Bayesian update function* $\mathcal{U}_a : \mathcal{Y}_a \times \mathbb{R} \to \mathcal{Y}_a$ so that after observing a reward $r \in \mathbb{R}$ from an arm $a \in \mathcal{A}$, the hyperparameter associated with arm a is updated from y_a to $\mathcal{U}_a(y_a, r)$ (e.g., if $\theta_a \sim \mathcal{P}_a(y_a)$, then $\theta_a | R_{a,1} \sim \mathcal{P}_a(\mathcal{U}_a(y_a, R_{a,1}))$). We will often use $\mathcal{U} : \mathcal{Y} \times \mathcal{A} \times \mathbb{R} \to \mathcal{Y}$ to denote the updating of the hyperparameter vector $\mathbf{y}$; i.e., after observing a reward realization r from an arm a, the hyperparameter vector is updated from $\mathbf{y}$ to $\mathcal{U}(\mathbf{y}, a, r)$, where only the ath component is updated.

We further describe the time evolution of the DM's belief throughout the decision making process. Given an outcome ω and an action sequence $\mathbf{a}_{1:t}$, the posterior hyperparameter vector at time t can be recursively expressed as

$$\mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y}) \triangleq \mathcal{U}\left(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y}), a_t, r_t(\mathbf{a}_{1:t}, \omega)\right), \quad \forall t \geq 1, \tag{8}$$

with $\mathbf{y}_0 \triangleq \mathbf{y}$. We often write $[\mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y})]_a$ to denote the ath component of $\mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y})$. This hyperparameter vector $\mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y})$ sufficiently describes the DM's belief given the history $H_t(\mathbf{a}_{1:t}, \omega)$.

Mean reward. We introduce several notions of mean reward that play a crucial role throughout the paper. For each arm $a \in \mathcal{A}$, we let $\mu_a(\theta_a)$ denote the *conditional mean reward* given the parameter θ_a, and let $\bar{\mu}_a(y_a)$ be the *predictive mean reward* given the hyperparameter y_a:

$$\mu_a(\theta_a) \triangleq \mathbb{E}\left[R_{a,n} | \theta_a\right], \quad \bar{\mu}_a(y_a) \triangleq \mathbb{E}_{y_a}\left[\mu_a(\theta_a)\right]. \tag{9}$$

We further define the *posterior predictive mean reward process* $\{\hat{\mu}_{a,n}\}_{n \geq 0}$ by

$$\hat{\mu}_{a,n}(\omega; y_a) \triangleq \mathbb{E}_{y_a}\left[\mu_a(\theta_a) | R_{a,1}, \ldots, R_{a,n}\right], \tag{10}$$

which represents the predictive mean reward (i.e., the finite-sample Bayesian estimate of $\mu_a(\theta_a)$) after observing first n rewards associated with the arm a.

Remark 1. *Fix an arm $a \in \mathcal{A}$. The posterior predictive mean reward process $\{\hat{\mu}_{a,n}\}_{n \geq 0}$ is a martingale adapted to the filtration generated by the sequence of rewards $(R_{a,1}, R_{a,2}, R_{a,3}, \ldots)$. Furthermore, it starts at the value of the prior predictive mean reward $\bar{\mu}_a(y_a)$ and converges to the conditional mean reward $\mu_a(\theta_a)$; i.e., $\hat{\mu}_{a,0}(\omega; y_a) = \bar{\mu}_a(y_a)$ and $\lim_{n \to \infty} \hat{\mu}_{a,n}(\omega; y_a) = \mu_a(\theta_a)$ almost surely (see Proposition 5 in the Electronic Companion).*

2.2. Natural Exponential Family

We will often consider the case where the reward distribution $\mathcal{R}_a(\theta_a)$ belongs to the *natural exponential family*. In this case, the closed-form expressions are available for the aforementioned notation. For any given $\theta_a \in \Theta_a \subseteq \mathbb{R}$, the probability measure for a random reward $R_{a,n}$ is determined by

$$\mathbb{P}\left[R_{a,n} \in dr \mid \theta_a\right] = h_a(dr) \exp\left(\theta_a r - A_a(\theta_a)\right), \tag{11}$$

where $h_a(dr)$ is the *reference measure* and $A_a(\cdot)$ is the *log-partition function* that is a logarithm of the normalization factor. We then have a family of conjugate priors $\{\mathcal{P}_a(y_a)\}_{y_a \in \mathcal{Y}_a}$ where $\mathcal{Y}_a \triangleq \{y_a = (\xi_a, \nu_a) | \xi_a \in \mathbb{R}, \nu > 0\}$, so, for any given hyperparameter $y_a \in \mathcal{Y}_a$, the corresponding prior $\mathcal{P}_a(y_a)$ is also an exponential family distribution and can be described as

$$\mathbb{P}_{(\xi_a, \nu_a)}\left[\theta_a \in d\theta\right] = f_a(\xi_a, \nu_a) \exp\left(\xi_a \theta - \nu_a A_a(\theta)\right) d\theta, \tag{12}$$

where $f_a(\xi_a, \nu_a)$ is the normalization factor and ν_a represents the effective number of observations. Within this family of conjugate priors, it is well known that the posterior distribution can be expressed as

$$\mathbb{P}_{(\xi_a, \nu_a)}\left[\theta_a \in d\theta \mid R_{a,1}, \ldots, R_{a,n}\right] = \mathbb{P}_{(\xi_a + \sum_{i=1}^n R_{a,i},\ \nu_a + n)}\left[\theta_a \in d\theta\right]. \tag{13}$$

This property can also be expressed via the Bayesian update function as $\mathcal{U}_a((\xi_a, \nu_a), r) = (\xi_a + r, \nu_a + 1)$. We also have the following identities for the mean reward metrics:

$$\mu_a(\theta_a) = A_a'(\theta_a), \quad \bar{\mu}_a(\xi_a, \nu_a) = \frac{\xi_a}{\nu_a}, \quad \hat{\mu}_{a,n}(\omega; \xi_a, \nu_a) = \frac{\xi_a + \sum_{i=1}^n R_{a,n}}{\nu_a + n}, \tag{14}$$

where $A_a' \triangleq dA_a/d\theta_a$. We refer the reader to Gutiérrez-Peña and Smith (1995) for further details.

Bernoulli and Gaussian MABs. We briefly illustrate the Bernoulli MAB and Gaussian MAB as representative examples of the problem instance described by a natural exponential family. In the Bernoulli MAB, the rewards of an arm are Bernoulli random variables whose success probability is drawn from a Beta distribution. In the Gaussian MAB, the rewards of an arm are normally distributed with an unknown mean and a known noise variance where the mean is also normally distributed. Table 1 summarizes the previously defined notation.

	Bernoulli MAB	Gaussian MAB
Prior distribution	$\mu_a \sim \text{Beta}(\alpha_a, \beta_a)$	$\mu_a \sim \mathcal{N}(m_a, v_a^2)$
Reward distribution	$R_{a,n} \sim \text{Bernoulli}(\mu_a)$	$R_{a,n} \sim \mathcal{N}(\mu_a, \sigma_a^2)$
Parameter θ_a	$\theta_a = \log \frac{\mu_a}{1 - \mu_a}$	$\theta_a = \frac{\mu_a}{\sigma_a^2}$
Hyperparameters ξ_a, ν_a	$\xi_a = \alpha_a,\ \nu_a = \alpha_a + \beta_a$	$\xi_a = \frac{m_a \sigma_a^2}{v_a^2},\ \nu_a = \frac{\sigma_a^2}{v_a^2}$
Reference measure h_a	$h_a(dr) = \delta_0(dr) + \delta_1(dr)$	$h_a(dr) = \frac{1}{\sqrt{2\pi\sigma_a^2}} \exp\left(-\frac{r^2}{\sigma_a^2}\right) dr$
Log-partition function A_a	$A_a(\theta_a) = \log\left(1 + e^{\theta_a}\right)$	$A_a(\theta_a) = \frac{\sigma_a^2 \theta_a^2}{2}$
Mean reward μ_a	$\mu_a(\theta_a) = \frac{e^{\theta_a}}{1 + e^{\theta_a}}$	$\mu_a(\theta_a) = \sigma_a^2 \theta_a$
Predictive mean $\bar{\mu}_a$	$\bar{\mu}_a(\alpha_a, \beta_a) = \frac{\alpha_a}{\alpha_a + \beta_a}$	$\bar{\mu}_a(m_a, v_a^2) = m_a$

Table 1: Description of a Bernoulli MAB and a Gaussian MAB. Here, $\delta_x(dr)$ denotes a Dirac measure that has a single atom at x.

2.3. Bayesian Optimal Policy

In a Bayesian framework, the MAB problem can be viewed as a Markov decision process (MDP) in which a state corresponds to an information state (or belief state) of the DM. It has the following recursive structure that we will exploit throughout the paper. Given an MAB instance with time horizon T and prior belief $\mathbf{y}$, suppose that the DM has just earned r by pulling an arm a at time $t = 1$. Then the remaining problem for the DM is equivalent to an MAB instance with time horizon $T-1$ and prior belief $\mathcal{U}(\mathbf{y}, a, r)$. Based on this Markovian structure, we obtain the following

Bellman equations for the MAB problem: for all $T \in \mathbb{N}$ and $\mathbf{y} \in \mathcal{Y}$,

$$Q^*(T, \mathbf{y}, a) \triangleq \mathbb{E}_{\mathbf{y}} \left[R_{a,1} + V^*(T - 1, \mathcal{U}(\mathbf{y}, a, R_{a,1})) \right], \tag{15}$$

$$V^*(T, \mathbf{y}) \triangleq \max_{a \in \mathcal{A}} Q^*(T - 1, \mathbf{y}, a), \tag{16}$$

with $V^*(0, \mathbf{y}) \triangleq 0$ for all $\mathbf{y} \in \mathcal{Y}$. The value function $V^*(T, \mathbf{y})$ represents the best possible performance that a non-anticipating policy can achieve in the MAB problem specified by the time horizon T and the prior belief $\mathbf{y}$, or equivalently, the maximum expected future reward that one can earn during T remaining periods[2] when the current belief is $\mathbf{y}$.

While Bellman equations are, in general, intractable to solve and directly apply, they offer a characterization of the *Bayesian optimal policy* (OPT). At a certain moment, when the remaining time horizon is T and the belief is $\mathbf{y}$, OPT takes an action with the largest state-action value (Q-value), i.e., pulls the arm $A^* = \text{argmax}_a Q^*(T, \mathbf{y}, a)$, and this action selection procedure is repeated while updating T and $\mathbf{y}$ according to Bayes' rule as described in Algorithm 1. Such a policy achieves the best possible performance among all non-anticipating policies:

$$V^*(T, \mathbf{y}) = \sup_{\pi \in \Pi_{\mathbb{F}}} V(\pi, T, \mathbf{y}) = V(\text{OPT}, T, \mathbf{y}), \quad \forall T \in \mathbb{N}, \mathbf{y} \in \mathcal{Y}. \tag{17}$$

Algorithm 1: Bayesian optimal policy (OPT)

Function OPT$(T, \mathbf{y})$

    // T:remaining time horizon, y:current belief

1   **return** $\text{argmax}_a Q^*(T, \mathbf{y}, a)$

Procedure OPT-Outer$(T, \mathbf{y})$

    // T:time horizon, y:prior belief

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

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

3   Select $A_t \leftarrow$ OPT$(T - t + 1, \mathbf{y}_{t-1})$

4   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

2.4. Thompson Sampling

Thompson sampling (TS) is a simple heuristic that makes decisions based on random sampling. When the remaining time is T and the current belief is $\mathbf{y}$, it samples the parameters $\tilde{\boldsymbol{\theta}}$ from the

---

[2]We intentionally refrain from indexing the value function V^* by time t, since such a representation conceals the Markovian structure of the Bayesian MAB problem and leads to complicated expressions for the variables that exploit this Markovian structure. To avoid confusion, the horizon T will be written as an argument to functions whereas the time index t will be written as a subscript, throughout the paper.

prior[3] distribution at that moment, $\mathcal{P}(\mathbf{y})$, and pulls the arm that is believed to be best given the sampled parameters $\tilde{\boldsymbol{\theta}}$, i.e., takes action $A^{\mathrm{TS}} = \mathrm{argmax}_a\, \mu_a(\tilde{\theta}_a)$. Like OPT, it repeats this procedure at every decision epoch while updating the belief $\mathbf{y}$ whenever a reward realization is observed.

Algorithm 2: Arm selection rule of Thompson sampling when remaining time is T and current belief is $\mathbf{y}$

Function TS(T, $\mathbf{y}$)

    // T:remaining time horizon, y:current belief

1   Sample parameters $\tilde{\boldsymbol{\theta}} \sim \mathcal{P}(\mathbf{y})$

2   **return** $\mathrm{argmax}_a\{\mu_a(\tilde{\theta}_a)\}$

Note that TS does not take into account the time information when making a decision. It applies the identical sampling and selection rule, irrespective of the remaining time periods. This often leads to the unnecessary explorations near the end of the horizon, which motivates our framework.

3. Information Relaxation Sampling

We apply the information relaxation framework (Brown et al., 2010) to the Bayesian MAB problem and propose a general framework which we call *information relaxation sampling* (IRS). The main idea behind the information relaxation is to relax the information constraint so that the decision maker (DM) is allowed to exploit some future information that is supposed to be unknown. As in the usual Lagrangian relaxation, an upper bound on the best possible performance can be obtained by solving the relaxed problem.

To motivate in detail, let us consider a situation under which the parameters $\boldsymbol{\theta}$ are revealed to the DM when the remaining period is T and the current belief is $\mathbf{y}$. The optimal action for this DM is to keep playing the arm with the highest mean reward, i.e., $\mathrm{argmax}_a\, \mu_a(\theta_a)$, and by doing so will earn $\mathbb{E}_{\mathbf{y}}[T \times \max_a \mu_a(\theta_a)]$ on average, which is indeed an upper bound on the performance of the optimal policy, $V^*(T, \mathbf{y})$.

Let us now postulate a situation under which the same kind of DM is informed with sampled parameters $\tilde{\boldsymbol{\theta}}$ that are drawn from the distribution $\mathcal{P}(\mathbf{y})$. For this (falsely informed) DM, the optimal action is again to play the arm with the highest mean reward but now with respect to the sampled parameters, i.e., $\mathrm{argmax}_a\, \mu_a(\tilde{\theta}_a)$. This procedure effectively describes the arm selection rule of Thompson sampling in the situation specified by the remaining horizon T and the current belief $\mathbf{y}$.

Above, we motivated a performance bound, $\mathbb{E}_{\mathbf{y}}[T \times \max_a \mu_a(\theta_a)]$, and a non-anticipating policy, TS, from the relaxation of the parameter information. Analogously, we can produce another

---

[3]Conventionally, the term "posterior distribution" is used to describe the distribution that TS samples the parameters from. We explicitly use "prior distribution" instead: for example, at time $t = 1$, the parameters are apparently sampled from the prior, not the posterior, distribution. After observing a reward realization, we will have a posterior but it will become a prior at the next decision epoch.

performance bound and another policy by considering a different set of future information to relax: the performance bound is obtained by computing how much the clairvoyant DM can earn with this additional information; and the policy is obtained by speculating which action the same kind of DM will take if the additional information is replaced with sampled (simulated) instance.

We will particularly consider the relaxations of information that are less effective than the full parameter information for the DM to maximize her future payoff. This will result in tighter relaxations, in the sense of a better (tighter) performance upper bound as well as a better performing policy.

In what follows, we formalize this idea utilizing the notion of *information relaxation penalties* that allows us to describe and control the benefit from having additional information explicitly. We will first describe the general framework and then propose a specific family of penalties that are particularly suitable for Bayesian MAB problems.

Information relaxation penalties and the inner problem. Applying the information relaxation framework developed by Brown et al. (2010), we relax the non-anticipativity constraint imposed on policy space $\Pi_\mathbb{F}$ (i.e., A_t^π is $\mathcal{F}_{t-1}$-measurable). Without loss of generality,[4] we consider the perfect information relaxation under which the DM is allowed to first observe all future outcomes in advance, and then pick an action (i.e., A_t^π is $\sigma(\omega)$-measurable). As in any other Lagrangian relaxation, we impose penalties on the DM for violating the non-anticipativity 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 an outcome ω for an MAB problem with time horizon T and prior belief $\mathbf{y}$. The clairvoyant DM can find the best action sequence that is optimal for this particular outcome ω in the presence of penalties z_t, by solving the following (deterministic) optimization problem, referred to as the *inner problem*:

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

Definition 1 (Dual feasibility)**.** *Given T and $\mathbf{y}$, a penalty function z_t is dual feasible if it is a zero mean for any non-anticipating policy $\pi \in \Pi_\mathbb{F}$, i.e.,*

$$\mathbb{E}_\mathbf{y}\left[\sum_{t=1}^{T} z_t(\mathbf{A}_{1:t}^\pi, \omega; T, \mathbf{y})\right] = 0, \quad \forall \pi \in \Pi_\mathbb{F}. \tag{18}$$

We remark that the mapping $\mathbf{a}_{1:t} \mapsto z_t(\mathbf{a}_{1:t}, \omega)$ is a stochastic function of the action sequence $\mathbf{a}_{1:t}$ since the outcome ω is random. This dual feasibility condition requires that the DM who makes decisions on the natural filtration will receive zero penalties in expectation.

---

[4]Any partial information relaxation can be equivalently described within the perfect information relaxation by adding additional terms into the penalty function. See the discussion after Theorem 1.

The complexity of the inner problem depends very much on the penalty function. Assuming that the penalty function can be evaluated in $O(1)$ computation, an enumerative brute-force optimization of the inner problem may require $O(K^T)$ computations. In what follows, we will illustrate that for suitably designed penalty functions, the inner problem exhibits a recursive structure and thus can be solved effectively using dynamic programming techniques.

IRS performance bound. We 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 $\mathcal{I}(\mathbf{y})$, i.e., the expected total payoff that a clairvoyant DM can achieve in the presence of penalties.:

$$W^z(T, \mathbf{y}) \triangleq \mathbb{E}_{\mathbf{y}} \left[\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}) \right\} \right]. \tag{19}$$

Once we have an algorithm to solve the inner problem, this value can be computed numerically via simulation: let $\omega_1, \omega_2, \ldots, \omega_S$ be the samples independently drawn from $\mathcal{I}(\mathbf{y})$, and W_s be the the maximal value of the inner problem with respect to ω_s for each $s = 1, \ldots, S$ separately. The bound W^z can be computed by taking the average of these maximal values, i.e., $\frac{1}{S} \sum_{s=1}^{S} W_s$. 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^*:*

$$\text{(Weak duality)} \qquad W^z(T, \mathbf{y}) \geq V^*(T, \mathbf{y}). \tag{20}$$

There exists a dual feasible penalty function denoted by z_t^{ideal}, such that

$$\text{(Strong duality)} \qquad W^{\text{ideal}}(T, \mathbf{y}) = V^*(T, \mathbf{y}). \tag{21}$$

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

$$z_t^{\text{ideal}}(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}_{\mathbf{y}} \left[r_t(\mathbf{a}_{1:t}, \omega) \,|\, H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right] \tag{22}$$
$$+ V^* \left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y}) \right) - \mathbb{E}_{\mathbf{y}} \left[V^* \left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y}) \right) |\, H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right].$$

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 the DM 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, thereby removing any incentive to deviate from OPT and resulting in the strong duality.

The ideal penalty is, of course, intractable, but its structure highlights 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 also shows that relaxing more the available information can always be compensated by adding associated terms to the penalty function. That is, a partial information relaxation (e.g., A_t^π is measurable w.r.t. $\mathcal{G}_{t-1}$ such that $\sigma(H_{t-1}) \subseteq \mathcal{G}_{t-1} \subseteq \sigma(\omega)$) with some penalty function $z_t^{\mathbb{G}}$ is equivalent to the perfect information relaxation (i.e., A_t^π is measurable w.r.t. $\sigma(\omega)$) with a penalty function $z_t^{\mathbb{G}} + z_t^{\sigma(\omega)\backslash\mathbb{G}}$ if the additional term $z_t^{\sigma(\omega)\backslash\mathbb{G}}$ exactly penalizes the relative benefit from having more information $\sigma(\omega)$ than $\mathcal{G}_{t-1}$. Hence, it is sufficient to consider the perfect information relaxation, as we do in this paper, and the actual amount of information available for the DM can be equivalently controlled by adjusting the penalty function.

Before proceeding, we remark that the above results are already well established in Brown et al. (2010) (see Lemma 2.1 and Theorem 2.3 therein) for a general class of MDP problems, except for a subtle difference regarding the assumption on the predictability of reward realizations. In MDP problems, the reward at each state is typically assumed to be deterministic (otherwise, it is replaced with its expected value), since the stochastic evolution of the state is of a major concern. By contrast, in MAB problems it is essential to consider the randomness of rewards since learning from the noisy reward realizations is of a major concern, and therefore, we do not assume that r_t is measurable with respect to $\sigma(H_{t-1})$. As a consequence, our ideal penalty function (22) has a slightly different functional form than the one formulated in Brown et al. (2010).[5] We further exploit this fact when designing a variety of penalty functions.

IRS policy. Since the true outcome ω is not available in reality, it cannot be used in online decision making. We derive a non-anticipating policy by leveraging the idea of "posterior sampling," which utilizes the sampled outcome $\tilde{\omega}$ instead of the true outcome ω.

Given a penalty function z_t, we characterize a randomized and non-anticipating IRS policy π^z as follows. Exploiting the recursive structure of a Bayesian MAB problem, the policy π^z specifies "which arm to pull when the remaining time is T and the current belief is $\mathbf{y}$," i.e., the very first action that it would take in an MAB instance with horizon T and prior belief $\mathbf{y}$. Given T and $\mathbf{y}$, it (i) first randomly generates an outcome $\tilde{\omega}$ (i.e., sampling from $\mathcal{I}(\mathbf{y})$), (ii) solves the inner problem to find a best action sequence $\tilde{\mathbf{a}}_{1:T}^*$ with respect to this randomly generated outcome $\tilde{\omega}$ 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 TS and OPT, **it repeats steps (i)–(iii) at every decision epoch**, while updating the remaining time T and belief $\mathbf{y}$ upon each decision making and reward realization.

---

[5]Brown et al. (2010) show that $z_t^{\text{ideal}} = V^*(T-t, \mathbf{y}_t) - \mathbb{E}[V^*(T-t, \mathbf{y}_t)| H_{t-1}]$, when r_t is assumed to be measurable with respect to $\sigma(H_{t-1})$ and so $r_t - \mathbb{E}[r_t | H_{t-1}] = 0$.

Algorithm 3: Information relaxation sampling (IRS) policy

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

    // T:remaining time horizon, $\mathbf{y}$:current belief

1 Sample an outcome $\tilde{\omega} \sim \mathcal{I}(\mathbf{y})$: Equivalently, for each $a \in \mathcal{A}$,

$$\tilde{\theta}_a \sim \mathcal{P}_a(y_a), \quad \tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}_a), \quad \forall n \in \{1, \dots, T\}.$$

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

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

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

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

    // T:time horizon, $\mathbf{y}$:prior belief

1 $\mathbf{y}_0 \leftarrow \mathbf{y}$

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

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

4   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), the random generation of the outcome $\tilde{\omega}$ given the belief $\mathbf{y}$ is equivalent to, for each arm $a \in \mathcal{A}$, sampling the parameter from its posterior, $\tilde{\theta}_a \sim \mathcal{P}_a(y_a)$, and then sampling the future reward realizations, $\tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}_a)$ for $n = 1, \dots, T$. In other words, the IRS policy π^z randomly generates (simulates) a plausible future scenario within its own probability space specified by T and $\mathbf{y}$.

The optimization problem in the step (ii) is identical to the inner problem ($*$) except that the true outcome ω is replaced with the sampled one $\tilde{\omega}$. Therefore, the dynamic programming algorithm that solves the inner problem can also be utilized for this online decision-making process, not only for the computation of performance bound W^z. Note that there can be multiple solutions to this optimization problem and the tie-breaking rule may affect the performance of the policy. We do not observe that the choice of tie-breaking rule is significance in our numerical experiments. In some instances that follow, however, we will adopt a specific tie-breaking rule for the purpose of theoretical analysis.

Also note that in step (iii) only the first action $\tilde{a}^*_1$ of the optimal solution $\tilde{\mathbf{a}}^*_{1:T}$ is utilized, and at the following decision epoch a new outcome is sampled based on the updated belief. If we consider an MAB instance with time horizon T, the policy π^z solves T different instances of the inner problem throughout the entire decision-making process, with a decreasing length of time horizon, from T to 1, and with a stochastically evolving belief state. See the IRS-OUTER procedure

in Algorithm 3, which is in fact identical to that employed in OPT and TS.

Remark 2. *The ideal penalty yields the Bayesian optimal policy, i.e., $\pi^{\text{ideal}} = OPT$.*

Recall that the ideal penalty (22) yields the performance bound W^{ideal} that is equal to the best achievable performance V^*, because the DM under the ideal penalty has no incentive to utilize any future information. For the same reason, the corresponding IRS policy π^{ideal} does not utilize the (randomly generated) future information in its decision making, and tries to make the best decision based only on the information revealed so far. Therefore, its decision should always coincide with the Bayesian optimal policy's decision.

Choice of penalty functions. We have so far described the general framework that takes a penalty function z_t as input, and yields a performance bound W^z and a policy π^z as outputs. While any dual feasible penalty functions can be utilized in general, we propose the following set of penalty functions that are particularly suitable for the MAB problems:

$$z_t^{\text{TS}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\, \boldsymbol{\theta}\right], \tag{23}$$

$$z_t^{\text{IRS.FH}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}_{\mathbf{y}}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\, \hat{\boldsymbol{\mu}}_{T-1}(\omega)\right], \tag{24}$$

$$z_t^{\text{IRS.V-ZERO}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}_{\mathbf{y}}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\, H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right], \tag{25}$$

$$z_t^{\text{IRS.V-EMAX}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}_{\mathbf{y}}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\, H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] \tag{26}$$
$$+ W^{\text{TS}}\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) - \mathbb{E}_{\mathbf{y}}\left[W^{\text{TS}}\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right)\Big| H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right],$$

where $\hat{\boldsymbol{\mu}}_{T-1}(\omega; \mathbf{y}) \triangleq \left(\hat{\mu}_{a,T-1}(\omega; y_a)\right)_{a \in \mathcal{A}}$ and the dependency of some expressions on T and $\mathbf{y}$ is suppressed for clarity. Also recall that the ideal penalty is given by

$$z_t^{\text{ideal}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}_{\mathbf{y}}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\, H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] \tag{27}$$
$$+ V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) - \mathbb{E}_{\mathbf{y}}\left[V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right)| H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right].$$

We can show that these penalty functions satisfy the dual feasibility condition (Definition 1); see Electronic Companion C.3 for a formal proof.

Remark 3. *All penalty functions* (23)–(27) *are dual feasible.*

This set of penalty functions results in a set of policies that ranges from Thompson sampling (TS) to the Bayesian optimal policy (OPT) and a set of performance bounds that ranges from the conventional regret benchmark W^{TS} $(= \mathbb{E}[T \times \max_a \mu_a(\theta_a)])$ to the optimal value function W^{ideal} $(= V^*)$. More specifically, at one extreme, the simplest penalty function z_t^{TS} yields TS and W^{TS} as outputs, and at the other extreme, the ideal penalty function z_t^{ideal} yields OPT and V^* which would be optimal. The other three penalty functions ($z_t^{\text{IRS.FH}}$, $z_t^{\text{IRS.V-ZERO}}$, and $z_t^{\text{IRS.V-EMAX}}$) connect the

two extremes and are sequentially "better", where we informally say that a penalty function is better than another if it is closer to the ideal penalty function and thus yields a better performing policy and a tighter performance bound. Deferring detailed explanations to §3.1–§3.4, we briefly illustrate general principles to design "good" penalty functions and motivate these penalty functions.

In design of information relaxation penalties, we first need to determine to which information set we relax the non-anticipativity constraint, i.e., what kind of additional information will be revealed to the DM in the relaxation. Although we have described our framework based on the perfect information relaxation (i.e., the relaxation in which the DM perfectly knows the entire future outcomes ω), any imperfect information relaxation can be equivalently described within the perfect information relaxation using a properly constructed penalty function.[6] Among the suggested penalty functions,[7] z_t^{TS} is the one that corresponds to the information relaxation to the parameter information $\boldsymbol{\theta}$, $z_t^{\text{IRS.FH}}$ corresponds to the information relaxation to the posterior predictive mean rewards $\hat{\boldsymbol{\mu}}_{T-1}$ (i.e., the finite-sample mean-reward estimates), and z_t^{ideal} corresponds to no information relaxation.

One principle to motivate a better penalty function is to choose a smaller set of future information for the relaxation. When less additional information is revealed to the DM in the relaxation, the additional profit that the DM can extract from this information becomes smaller, and hence the DM has to make more realistic decisions that rely more on the currently available information rather than the future information that is supposed to be unknown. Comparing $z_t^{\text{IRS.FH}}$ with z_t^{TS}, one may notice that the finite-sample mean-reward estimates $\hat{\boldsymbol{\mu}}_{T-1}$ are less informative than the parameters $\boldsymbol{\theta}$ for the DM to exploit in her profit maximization because, in terms of mean-reward estimation, the parameters are informative as much as an infinite number of observations (i.e., $\mathbb{E}[\mu_a(\theta_a)|\boldsymbol{\theta}] = \lim_{T\to\infty} \mathbb{E}[\mu_a(\theta_a)|R_{a,1},\ldots,R_{a,T-1}] = \lim_{T\to\infty} \hat{\mu}_{a,T-1}$). In this sense, $z_t^{\text{IRS.FH}}$ is better than z_t^{TS}, and resulting policy $\pi^{\text{IRS.FH}}$ and performance bound $W^{\text{IRS.FH}}$ improve upon TS and W^{TS} toward OPT and V^*.

Another principle to motivate a better penalty function is to adopt a more precise approximation of the ideal penalty function z_t^{ideal}, particularly regarding the terms containing the optimal value function V^*. In the presence of penalties that reflect the value of the additional information more accurately, the DM has less incentive to exploit this additional information in the relaxed decision making problem, and similarly to the above argument, this leads to more realistic decisions. Among our suggestions, $z_t^{\text{IRS.V-ZERO}}$ approximates the term V^* with zero, and $z_t^{\text{IRS.V-EMAX}}$ approximates the term V^* with a tractable upper bound W^{TS}. By doing so, $z_t^{\text{IRS.V-EMAX}}$ takes into account the continuation value of each action explicitly and improves upon $z_t^{\text{IRS.V-ZERO}}$.

Consider the inner problem associated with each choice of penalty function (23)–(27). Recall

---

[6]In fact, this is the main idea underlying the existence of the ideal penalty function; see the discussion after Theorem 1.

[7]We can motivate one more penalty function that corresponds to the perfect information relaxation. Such a penalty function is simply given by $z_t \equiv 0$, which is illustrated in Electronic Companion A. However, we do not suggest its use since it is even worse than z_t^{TS}.

that each inner problem is a deterministic multi-period decision making problem that has a form of $\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}) - z_t(\mathbf{a}_{1:t})$. A penalty function z_t effectively redefines what the DM earns at each time, i.e., $r_t(\mathbf{a}_{1:t})$ is replaced with $r_t(\mathbf{a}_{1:t}) - z_t(\mathbf{a}_{1:t})$. More specifically, the penalty function z_t^{TS} effectively replaces the realized rewards associated with each arm with their expected value given parameters $\boldsymbol{\theta}$; as does $z_t^{\text{IRS.FH}}$ (with their expected value given the finite-sample mean-reward estimates $\hat{\boldsymbol{\mu}}_{T-1}$); as does $z_t^{\text{IRS.V-ZERO}}$ (with their expected value conditional on how many times the arm has previously been selected up to each point in time); as does $z_t^{\text{IRS.V-EMAX}}$ (with an approximation of the continuation value in addition to the component employed by $z_t^{\text{IRS.V-ZERO}}$).

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 2: List of algorithms following from penalty functions (23)–(27). TS refers to Thompson sampling and OPT refers to the Bayesian optimal policy. Run time represents the computational complexity of solving one instance of the inner problem ($*$), that is, the time required to obtain one sample in a computation of performance bound W^z or to decide which arm to select in each period in a run of policy π^z.

Table 2 summarizes these inner problems. As we sequentially increase the computational complexity of a penalty function, from z_t^{TS} to z_t^{ideal}, the penalty function more accurately penalizes the benefit from knowing future outcomes, i.e., more explicitly prevents the DM from exploiting future information. As a result, the inner problem becomes closer to the original stochastic optimization problem, which results in a better performing policy and a tighter performance bound. Using this approach, we achieve a family of algorithms that are intuitive and tractable, exhibiting a trade-off between quality and computational efficiency. See Electronic Companion §A for an illustrative example.

The run time in Table 2 represents the computational complexity of solving one instance of the inner problem, i.e., the time it takes to obtain one sample in a computation of performance bound W^z or to decide which arm to select in each period in a run of policy π^z. In this run-time analysis, performing the Bayesian belief updating and the sampling of a random variable is counted as a single operation.

3.1. Thompson Sampling Revisited

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

$$\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t(\mathbf{a}_{1:t}, \omega) \right\} = \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} \mu_{a_t}(\theta_{a_t}) \right\} = T \times \max_{a \in \mathcal{A}} \mu_a(\theta_a). \qquad (28)$$

Given an outcome ω, and in the presence of penalties, a hindsight optimal action sequence is to keep pulling the true best arm, i.e., $\mathrm{argmax}_a \mu_a(\theta_a)$, for T times in a row. The resulting performance bound W^{TS} reduces to the conventional performance benchmark,

$$W^{\mathrm{TS}}(T, \mathbf{y}) = \mathbb{E}_{\mathbf{y}} \left[T \times \max_{a \in \mathcal{A}} \mu_a(\theta_a) \right], \qquad (29)$$

which measures how much the DM could have achieved if the parameters had been revealed in advance.

Remark 4. *The performance bound W^{TS} is the conventional benchmark that has been widely used in the Bayesian regret analysis (Lai and Robbins, 1985; Russo and Van Roy, 2014, 2018). The Bayesian regret of a policy π is defined as*

$$BayesRegret(\pi, T, \mathbf{y}) \triangleq \mathbb{E}_{\mathbf{y}} \left[\sum_{t=1}^{T} \max_a \mu_a(\theta_a) - \mu_{A_t^\pi}(\theta_{A_t^\pi}) \right] = W^{TS}(T, \mathbf{y}) - V(\pi, T, \mathbf{y}), \qquad (30)$$

which quantifies the suboptimality of the policy π.

It is trivial to see that the corresponding policy π^{TS} is equivalent to Thompson sampling. The policy π^{TS} utilizes a sampled outcome $\tilde{\omega}$ instead of the true outcome ω; accordingly, it selects an arm $A^{\mathrm{TS}} = \mathrm{argmax}_a \mu_a(\tilde{\theta}_a)$, where $\tilde{\boldsymbol{\theta}} \sim \mathcal{P}(\mathbf{y})$, which is identical to the procedure described in Algorithm 2. In order for the policy π^{TS} to make a decision at a certain time, note that it does not need to sample future rewards, and thus it requires $O(K)$ computations only.

3.2. IRS.FH

Recall that $\hat{\mu}_{a,T-1}(\omega; y_a)$ is the posterior predictive mean reward of an arm a that the DM will have after observing $T-1$ reward realizations $R_{a,1}, \ldots, R_{a,T-1}$ given the initial belief y_a:

$$\hat{\mu}_{a,T-1}(\omega; y_a) \triangleq \mathbb{E}_{y_a} \left[\mu_a(\theta_a) | R_{a,1}, \ldots, R_{a,T-1} \right]. \qquad (31)$$

Given (24), the optimal solution to the inner problem $(*)$ is to always pull the arm with the highest posterior predictive mean reward, i.e., $\mathrm{argmax}_a \hat{\mu}_{a,T-1}(\omega; y_a)$:

$$\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t^{\mathrm{IRS.FH}}(\mathbf{a}_{1:t}, \omega) \right\} = \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} \hat{\mu}_{a_t, T-1}(\omega) \right\} = T \times \max_{a \in \mathcal{A}} \hat{\mu}_{a,T-1}(\omega). \qquad (32)$$

This inner problem yields the performance bound $W^{\text{IRS.FH}}$, such that

$$W^{\text{IRS.FH}}(T, \mathbf{y}) = \mathbb{E}_{\mathbf{y}} \left[T \times \max_{a \in \mathcal{A}} \hat{\mu}_{a,T-1}(\omega; y_a) \right], \tag{33}$$

and the policy $\pi^{\text{IRS.FH}}$ that is implemented in Algorithm 4.

Algorithm 4: Arm selection rule of $\pi^{\text{IRS.FH}}$ when remaining time is T and current belief is $\mathbf{y}$

Function `IRS.FH`$(T, \mathbf{y})$

    // T:remaining time horizon, y:current belief

1   Sample parameters $\tilde{\boldsymbol{\theta}} \sim \mathcal{P}(\mathbf{y})$ and rewards $\tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}_a)$, $\forall n \in \{1, \ldots, T\}$, $\forall a \in \mathcal{A}$.

2   **return** $\text{argmax}_a \left\{ \mathbb{E}_{y_a} \left[\mu_a(\theta_a) \,\middle|\, R_{a,1} = \tilde{R}_{a,1}, \ldots, R_{a,T-1} = \tilde{R}_{a,T-1} \right] \right\}$

IRS.FH (FH stands for finite horizon) is almost identical to TS except that the conditional mean reward $\mu_a(\theta_a)$ is replaced with the posterior predictive mean reward $\hat{\mu}_{a,T-1}(\omega)$. As a finite-sample Bayesian estimate of the conditional mean reward, $\hat{\mu}_{a,T-1}(\omega)$ is less informative than $\mu_a(\theta_a)$ from the DM's perspective. In terms of mean reward estimation, the DM will never be able to identify $\mu_a(\theta_a)$ perfectly within a finite horizon, i.e., knowing the parameters is equivalent to having an infinite number of observations. The inner problem of TS requires the DM to "identify the best arm based on an infinite number of samples," whereas that of IRS.FH requires the DM to "identify the best arm based on a finite number of samples" and takes into account the length of the time horizon explicitly. By restricting the DM's access to fewer information, IRS.FH requires the DM to be more realistic, that is, to consider the uncertainties more precisely.
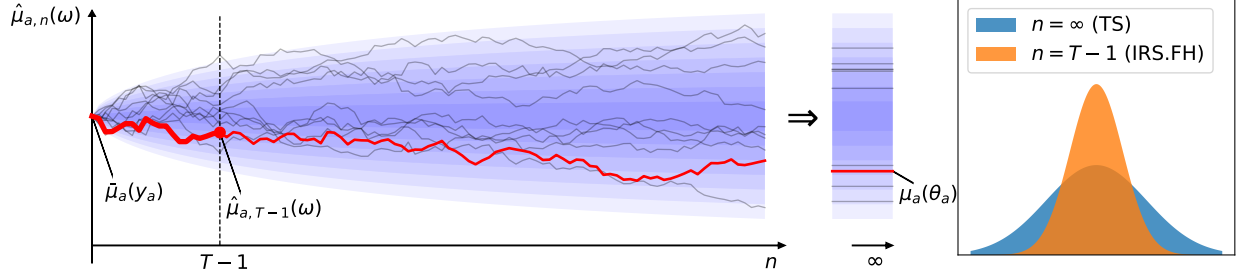

Figure 1: (Left) Sample paths of posterior predictive mean reward process of an arm a, $\left\{\hat{\mu}_{a,n}(\omega)\right\}_{n \geq 0}$. This process is a martingale that starts at (prior) predictive mean $\bar{\mu}_a$ and converges to conditional mean μ_a (Remark 1). (Right) The distributions of $\hat{\mu}_{a,T-1}$ and μ_a: $\hat{\mu}_{a,T-1}$ is more concentrated than μ_a, while all have the same mean $\bar{\mu}_a(y_a)$.

To sharpen our comparison between IRS.FH and TS, let us compare the variance of $\hat{\mu}_{a,T-1}(\omega)$ and $\mu_a(\theta_a)$ induced by the randomness of outcome ω. As depicted in Figure 1, μ_a is more widely distributed than $\hat{\mu}_{a,T-1}$ because a larger (infinite vs. $T-1$) number of samples makes it easier for the posterior to deviate from the initial prior (see also Remark 1). By Jensen's inequality, we further have $W^{\text{IRS.FH}} = \mathbb{E}[T \times \max_a \hat{\mu}_{a,T-1}(\omega)] \leq \mathbb{E}[T \times \max_a \mu_a(\theta_a)] = W^{\text{TS}}$ for any problem

instance, meaning that IRS.FH yields a performance bound that is tighter than the conventional benchmark. Also note that the same argument holds for the comparison between $\hat{\mu}_{a,T-1}(\tilde{\omega})$ and $\mu_a(\tilde{\theta}_a)$ since the synthesized outcome $\tilde{\omega}$ is identically distributed with the (true) outcome ω. The variability of $\hat{\mu}_{a,T-1}(\tilde{\omega})$ (respectively, $\mu_a(\tilde{\theta}_a)$) governs the randomness of the action taken by policy $\pi^{\text{IRS.FH}}$ (resp., π^{TS}), i.e., $A^{\text{IRS.FH}} = \text{argmax}_a \, \hat{\mu}_{a,T-1}(\tilde{\omega})$ (resp., $A^{\text{TS}} = \text{argmax}_a \, \mu_a(\tilde{\theta}_a)$). Given T and $\mathbf{y}$, the policy $\pi^{\text{IRS.FH}}$ performs fewer random explorations than TS, as it is less likely to deviate from the myopic decision to play an arm with the largest current estimate $\bar{\mu}_a(y_a)$. More desirably, the degree of exploration of $\pi^{\text{IRS.FH}}$ is controlled by the remaining time horizon as the variance of $\hat{\mu}_{a,T-1}(\omega)$ depends on T. At the last decision epoch ($T = 1$), $\pi^{\text{IRS.FH}}$ takes a myopic action that is indeed optimal.

Efficiently sampling $\hat{\mu}_{a,T-1}(\tilde{\omega})$ for natural exponential families. In order to obtain $\hat{\mu}_{a,T-1}(\tilde{\omega})$ for each arm a for a synthesized outcome $\tilde{\omega}$, one may apply Bayes' rule sequentially for each reward realization, which will take $O(T)$ computations per arm.

As discussed in §2.2, in an MAB were the reward distribution is a natural exponential family, the posterior predictive mean reward is given by

$$\hat{\mu}_{a,T-1}(\tilde{\omega}; \xi_a, \nu_a) = \frac{\xi_a + \sum_{n=1}^{T-1} \tilde{R}_{a,n}}{\nu_a + T - 1}. \tag{34}$$

Therefore, it is sufficient to sample the sum of $T - 1$ future rewards, $\sum_{n=1}^{T-1} \tilde{R}_{a,n}$, in order to sample the posterior predictive mean reward. Observe that the conditional distribution of the sum given $\tilde{\theta}_a$ also belongs to the natural exponential family, induced by a log-partition function $(T - 1)A_a(\tilde{\theta}_a)$. This distribution may be tractable to compute: for example, its distribution is Binomial$(T - 1, \mu_a(\tilde{\theta}_a))$ for the Beta-Bernoulli case, and $\mathcal{N}((T - 1) \cdot \mu_a(\tilde{\theta}_a), (T - 1) \cdot \sigma_a^2)$ for the Gaussian case. In these settings, we can sample the sum $\sum_{n=1}^{T-1} \tilde{R}_{a,n}$ directly from the tractable distribution (after sampling $\tilde{\theta}_a$) using $O(1)$ computation, and then use it to compute $\hat{\mu}_{a,T-1}(\tilde{\omega})$ without sequentially updating the belief. In such cases, a single decision of $\pi^{\text{IRS.FH}}$ can be made within $O(K)$ operations, independent of T, similar in computational complexity to TS.

3.3. IRS.V-Zero

IRS.V-ZERO introduces a further complication in that its inner problem requires the DM to consider her causal process in the course of solving the inner problem. Under the penalty $z_t^{\text{IRS.V-ZERO}}$ given in (25), the DM at time t earns $\mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\, H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right]$, the expected mean reward that she can infer from observations prior to time t. As we defined $R_{a,n}$ to be a reward from the nth pull on arm a (not the pull at time n), the posterior belief associated with each arm is determined only by the number of past pulls performed on that arm. Recall that $\hat{\mu}_{a,n}(\omega)$ is the expected mean reward of arm a that the DM can infer from the first n reward realizations:

$$\hat{\mu}_{a,n}(\omega; y_a) \triangleq \mathbb{E}_{y_a}\left[\mu_a(\theta_a) \,|\, R_{a,1}, \ldots, R_{a,n}\right]. \tag{35}$$

Therefore, the DM earns $\hat{\mu}_{a,n-1}(\omega)$ from the nth pull on arm a, irrespective of the detailed sequence of the past actions. More formally, the DM's earning at time t is

$$r_t(\mathbf{a}_{1:t}, \omega) - z_t^{\text{Irs.V-Zero}}(\mathbf{a}_{1:t}, \omega) = \mathbb{E}_{\mathbf{y}}\left[\mu_{a_t}(\theta_{a_t}) \,|\, H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] = \hat{\mu}_{a_t, n_{t-1}(\mathbf{a}_{1:t-1}, a_t)}(\omega), \qquad (36)$$

where $n_{t-1}(\mathbf{a}_{1:t-1}, a)$, defined in (5), denotes the number of pulls conducted on a particular arm a prior to time t.

Let $S_{a,n}(\omega) \triangleq \sum_{i=1}^{n} \hat{\mu}_{a,i-1}(\omega)$ be the cumulative payoff from the first n pulls of an arm a. Given an outcome ω, we observe that the total payoff is determined only by the total number of pulls on each arm, and not the sequence in which the arms have been pulled. Therefore, solving the inner problem $(*)$ is equivalent to "finding the optimal allocation $(n_1^*, n_2^*, \ldots, n_K^*)$ among T remaining opportunities": more formally,

$$\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} \hat{\mu}_{a_t, n_{t-1}(\mathbf{a}_{1:t-1}, a_t)} \right\} = \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{a=1}^{K} \sum_{n=1}^{n_T(\mathbf{a}_{1:T}, a)} \hat{\mu}_{a,n-1} \right\} = \max_{\mathbf{n}_{1:K} \in N_T} \left\{ \sum_{a=1}^{K} S_{a,n_a} \right\}, \qquad (37)$$

where $N_T \triangleq \{(n_1, \ldots, n_K) \in \mathbb{N}_0^K : \sum_{a=1}^{K} n_a = T\}$ is the set of all feasible allocations. Once the $S_{a,n}$'s are computed, we can solve this inner problem within $O(KT^2)$ operations by sequentially applying the sup convolution K times (more specifically, by recursively computing the maximal cumulative payoff that can be achieved by optimally allocating n pulls over arms $1, \ldots, a$, for $n = 1, 2, \ldots, T$ and $a = 1, 2, \ldots, K$). The detailed implementation is provided in §B.1.

The policy $\pi^{\text{Irs.V-Zero}}$ further needs to decide which arm to pull given the optimal allocation $(\tilde{n}_1^*, \tilde{n}_2^*, \ldots, \tilde{n}_K^*)$ that is obtained for the sampled outcome $\tilde{\omega}$. In principle, any arm a that was included in the solution of the inner problem, $\tilde{n}_a^* > 0$, would suffice, but we suggest a selection rule by which the arm that needs the most pulls is chosen, i.e., $A^{\text{Irs.V-Zero}} = \text{argmax}_a \tilde{n}_a^*$. This guarantees that $\pi^{\text{Irs.V-Zero}}$ behaves like TS when T is large, as formally stated in Proposition 1.

Comparison with TS **and** Irs.FH. Recall that in the inner problems of TS and Irs.FH, the DM at time t earns $\mathbb{E}[r_t|\boldsymbol{\theta}]$ and $\mathbb{E}[r_t|\hat{\boldsymbol{\mu}}_{T-1}]$, respectively, which are the mean reward estimates that rely on the information not available at the moment; e.g., $\hat{\mu}_{a,T-1}$ is revealed only after playing the arm a for $T-1$ times. Irs.V-Zero is more restrictive for the DM in the sense that she at time t earns $\mathbb{E}[r_t|H_{t-1}]$, which does not include any information that does not belong to H_{t-1}. Irs.V-Zero reflects the fact that the nth reward of an arm will not be revealed unless the arm is pulled n times, and its inner problem requires the DM to allocate a pull in order to incorporate the next reward realization into her information set; thus learning about an arm comes at the cost of sacrificing an opportunity to learn about the other arms.

More specifically, let us focus on the total payoff of a particular allocation $(n_1, \ldots, n_K)$ under each penalty function $z_t^{\text{Irs.V-Zero}}$ and $z_t^{\text{Irs.FH}}$. The allocation yields $\sum_{a=1}^{K} S_{a,n_a}(\omega)$ in the inner problem of Irs.V-Zero whereas the same allocation yields $\sum_{a=1}^{K} n_a \times \hat{\mu}_{a,T-1}(\omega)$ in the inner

problem of IRS.FH. In terms of variability originating from the randomness of ω, we observe that each summand $S_{a,n_a}(\omega) = \sum_{i=1}^{n_a} \hat{\mu}_{a,i-1}(\omega)$ is less noisy than its counterpart $n_a \times \hat{\mu}_{a,T-1}(\omega)$ since a larger number of observations makes it easier for the posterior to deviate from the initial prior and hence the variance of individual terms $\hat{\mu}_{a,0}(\omega), \ldots, \hat{\mu}_{a,n_a-1}(\omega)$ is smaller than the variance of $\hat{\mu}_{a,T-1}(\omega)$ and, therefore, $\sum_{a=1}^K S_{a,n_a}(\omega)$ is smaller than $\sum_{a=1}^K n_a \times \hat{\mu}_{a,T-1}(\omega)$. Analogous to the comparison between IRS.FH and TS, we have that IRS.V-ZERO yields a performance bound $W^{\text{IRS.V-ZERO}}$ that is tighter than $W^{\text{IRS.FH}}$ (formally stated in Theorem 2) and a policy $\pi^{\text{IRS.V-ZERO}}$ that performs fewer random explorations than $\pi^{\text{IRS.FH}}$.

3.4. IRS.V-EMax

Under perfect information relaxation, the DM perfectly knows not only (i) what she will earn at future times but also (ii) how her belief will evolve as a result of her action sequence. The previous algorithms focus on the former component by making the DM adjust the future rewards by conditioning (e.g., $\mathbb{E}[r_t(a_t)|\boldsymbol{\theta}]$, $\mathbb{E}[r_t(a_t)|\hat{\boldsymbol{\mu}}_{T-1}]$ and $\mathbb{E}[r_t(a_t)|H_{t-1}]$). IRS.V-EMAX also focuses on the latter component as well by charging the DM an additional cost for using the information on her future belief transitions.

To motivate this in detail, recall that the ideal penalty z_t^{ideal} (22) is

$$z_t^{\text{ideal}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\, H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] \tag{38}$$
$$+ V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) - \mathbb{E}\left[V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) \,|\, H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right],$$

where $V^*\left(T - t, \mathbf{y}_t\right)$ measures the value of having a belief $\mathbf{y}_t$ at a future time $t+1$. Note that, at the moment the DM takes an action a_t, the next belief state $\mathbf{y}_t = \mathcal{U}(\mathbf{y}_{t-1}, a_t, r_t)$ is not measurable with respect to the natural filtration $\sigma(H_{t-1})$ since the next observation r_t is unknown. In DP terms, the conditional expectation $\mathbb{E}\left[V^*\left(T - t, \mathbf{y}_t\right)|H_{t-1}\right]$ captures the expected value of a (random) next state given the current state. Accordingly, the gap between its realized value and its expected value, $V^*\left(T - t, \mathbf{y}_t\right) - \mathbb{E}\left[V^*\left(T - t, \mathbf{y}_t\right)|H_{t-1}\right]$, measures the additional gain from knowing the next belief state $\mathbf{y}_t$. In addition to the term $r_t - \mathbb{E}\left[r_t|H_{t-1}\right]$ $(= z_t^{\text{IRS.V-ZERO}})$, which measures the benefit from knowing which action will yield a large immediate reward, the ideal penalty also penalizes the long-term benefit from knowing which action will lead to a favorable belief state.

The penalty function $z_t^{\text{IRS.V-EMAX}}$ is obtained from z_t^{ideal} by replacing $V^*(T, \mathbf{y})$ with $W^{\text{TS}}(T, \mathbf{y})$, which is computable. The use of $W^{\text{TS}}(T, \mathbf{y}) \triangleq \mathbb{E}_{\mathbf{y}}\left[T \times \max_a \mu_a(\theta_a)\right]$, introduced in (29), leads to a simple expression for its conditional expectation: since $\boldsymbol{\theta}|H_{t-1}$ is distributed with $\mathcal{P}(\mathbf{y}_{t-1})$, we have

$$\mathbb{E}_{\mathbf{y}}\left[W^{\text{TS}}\left(T - t, \mathbf{y}_t\right)\Big| H_{t-1}\right] = (T - t) \times \mathbb{E}_{\mathbf{y}}\left[\max_a \mu_a(\theta_a)\Big| H_{t-1}\right] \tag{39}$$
$$= (T - t) \times \mathbb{E}_{\mathbf{y}_{t-1}}\left[\max_a \mu_a(\theta_a)\right] \tag{40}$$
$$= W^{\text{TS}}\left(T - t, \mathbf{y}_{t-1}\right). \tag{41}$$

In the associated inner problem, the payoff that the DM earns at time t is

$$r_t(\mathbf{a}_{1:t}, \omega) - z_t^{\text{IRS.V-EMAX}}(\mathbf{a}_{1:t}, \omega) \tag{42}$$

$$= \hat{\mu}_{a_t, n_{t-1}(\mathbf{a}_{1:t-1}, a_t)}(\omega) - W^{\text{TS}}\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) + W^{\text{TS}}\left(T - t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right) \tag{43}$$

$$= \bar{\mu}_{a_t}\left([\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)]_{a_t}\right) - W^{\text{TS}}\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) + W^{\text{TS}}\left(T - t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right), \tag{44}$$

which is completely determined by the prior belief $\mathbf{y}_{t-1}$ and the posterior belief $\mathbf{y}_t$.

We further observe that, given ω, the future belief $\mathbf{y}_t(\mathbf{a}_{1:t}, \omega)$ depends only on how many times each arm has 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, in both cases the agent would observe $(R_{1,1}, R_{1,2}, R_{1,3})$ from arm 1 and $(R_{2,1}, R_{2,2})$ from arm 2 and end up with the same belief $\mathbf{y}_t(\mathbf{a}_{1:t}^A, \omega) = \mathbf{y}_t(\mathbf{a}_{1:t}^B, \omega)$. We may conclude from this observation that a belief state can be sufficiently parameterized with the pull counts $\mathbf{n}_{1:K} = (n_1, \ldots, n_K)$ instead of action sequence $\mathbf{a}_{1:t}$.

As a result, the total number of possible future beliefs is $O(T^K)$, not $O(K^T)$, and we can come up with a dynamic programming algorithm that solves the inner problem within $O(c_W T^K + K T^K)$ computations where c_W is the cost of numerically calculating $W^{\text{TS}}(T, \mathbf{y})$. We refer the interested reader to §B.2.

3.5. IRS.Index Policy

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

Roughly speaking, the IRS.INDEX is a single-sample approximation of the finite-horizon Gittins index (Kaufmann et al., 2012a), where the approximation is motivated by IRS.V-EMAX algorithm. It first solves the single-armed bandit problem for each arm in isolation, and makes a decision based on the results of these subproblems.

Single-armed bandit problem. Consider a special case of an MAB instance in which there is a single arm a that yields stochastic rewards $R_{a,n} \sim \mathcal{R}_a(\theta_a)$ 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 a priori.

Given an outcome $\omega_a = (\theta_a, (R_{a,n})_{n \in \mathbb{N}})$, we can simulate the future belief trajectory $(y_{a,n})_{n \in \{0, \ldots, T\}}$, where $y_{a,n}$ is the belief after n reward realizations are observed:

$$y_{a,0} \triangleq y_a, \quad y_{a,n} \triangleq \mathcal{U}_a(y_{a,n-1}, R_{a,n}), \quad \forall n = 1, \ldots, T. \tag{45}$$

Let $V^*(T, y_a, \lambda)$ be the optimal value function associated with this single-armed bandit problem. We consider the penalty function $z_t^{\text{IRS.V-EMAX}}$ in which the value function $V^*(T, y_a, \lambda)$ is approxi-

mated by $W^{\text{TS}}(T, y_a, \lambda) = \mathbb{E}_{y_a}[T \times \max(\mu_a(\theta_a), \lambda)]$. We define $\mathcal{A} \triangleq \{0, 1\}$ such that $a_t = 1$ if the stochastic arm at time t is selected, and $a_t = 0$ if the outside option is selected. The associated inner problem is

$$\text{maximize} \quad \sum_{t=1}^{T} \hat{\mu}_{a,n_t-1}(\omega_a) \cdot \mathbf{1}\{a_t = 1\} + \lambda \cdot \mathbf{1}\{a_t = 0\} - (T - t) \times \left(\Gamma_{n_t}^{\lambda}(\omega_a) - \Gamma_{n_t-1}^{\lambda}(\omega_a)\right) \quad (46)$$

$$\text{subject to} \quad n_t = \sum_{s=1}^{t} \mathbf{1}\{a_t = 1\}, \quad a_t \in \{0, 1\}, \quad \forall t = 1, \dots, T, \quad (47)$$

where $\hat{\mu}_{a,n}(\omega_a) \triangleq \mathbb{E}_{y_a}[\mu_a(\theta_a)|R_{a,1}, \dots, R_{a,n}] = \bar{\mu}_a(y_{a,n})$ and

$$\Gamma_n^{\lambda}(\omega_a) \triangleq \mathbb{E}_{y_{a,n}}[\max(\mu_a(\theta_a), \lambda)]. \quad (48)$$

With some algebra (Proposition 2 in §B.3), we can reformulate the optimization problem as

$$\max_{0 \le n \le T} \left\{ T \times \Gamma_0^{\lambda}(\omega_a) + (T - n) \times \left(\lambda - \min_{0 \le i \le n} \Gamma_i^{\lambda}(\omega_a)\right) + \sum_{i=1}^{n} \left(\hat{\mu}_{a,i-1}(\omega_a) - \Gamma_{i-1}^{\lambda}(\omega_a)\right) \right\}, \quad (49)$$

where the decision variable n is the total number of pulls on the stochastic arm.

Let $\varphi_a(\lambda, \omega_a)$ be the (maximal) relative benefit from pulling the stochastic arm against not pulling at all:

$$\varphi_a(\lambda, \omega_a) \triangleq \max_{1 \le n \le T} \left\{ T \times \Gamma_0^{\lambda} + (T - n) \times \left(\lambda - \min_{0 \le i \le n} \Gamma_i^{\lambda}\right) + \sum_{i=1}^{n} \left(\hat{\mu}_{a,i-1} - \Gamma_{i-1}^{\lambda}\right) \right\} - T \times \lambda. \quad (50)$$

Note that $\max\{\cdot\}$ was taken over $n \ge 1$. We interpret the meaning of the sign of $\varphi_a(\lambda, \omega_a)$ as follows: given an outcome ω_a, the stochastic arm is worth trying against the deterministic outside option λ if $\varphi_a(\lambda, \omega_a) \ge 0$, and not worth trying if $\varphi_a(\lambda, \omega_a) < 0$.

Given ω_a and λ, the value of $\varphi_a(\lambda, \omega_a)$ can be computed in $O(T)$ operations by precalculating $\sum_{i=1}^{n} \hat{\mu}_{a,i-1}(\omega_a)$, $\min_{0 \le i \le n} \Gamma_i^{\lambda}(\omega_a)$, and $\sum_{i=1}^{n} \Gamma_{i-1}^{\lambda}(\omega_a)$ over $n = 1, \dots, T$ sequentially. The single-armed bandit problem has an additional advantage of computational efficiency: in contrast to the implementation of IRS.V-EMAX in the multi-arm setting, the approximate value function (captured by Γ_n^{λ}) often admits a closed-form expression in the single-armed setting. In the cases of the Beta-Bernoulli MAB and the Gaussian MAB, for example, we have

$$\mathbb{E}_{\mu \sim \text{Beta}(\alpha,\beta)}[\max(\mu, \lambda)] = \lambda \times F_{\alpha,\beta}^{\text{beta}}(\lambda) + \frac{\alpha}{\alpha + \beta} \times \left(1 - F_{\alpha+1,\beta}^{\text{beta}}(\lambda)\right), \quad (51)$$

$$\mathbb{E}_{\mu \sim \mathcal{N}(m,\nu^2)}[\max(\mu, \lambda)] = m + (\lambda - m) \times \Phi\left(\nu^{-1}(\lambda - m)\right) + \nu \times \phi\left(\nu^{-1}(\lambda - m)\right), \quad (52)$$

where $F_{\alpha,\beta}^{\text{beta}}(\cdot)$ represents the c.d.f. of $\text{Beta}(\alpha, \beta)$ distribution, and $\Phi(\cdot)$ and $\phi(\cdot)$ represent the c.d.f. and the p.d.f. of the standard normal distribution, respectively. With these expressions, $\Gamma_n^{\lambda}(\omega_a)$'s can be computed very efficiently without using numerical integration or Monte Carlo sampling.

23

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

More specifically, consider a certain decision epoch when the remaining time is T and the belief is $\mathbf{y}$. For each arm $a = 1, \ldots, K$ separately, the policy $\pi^{\text{IRS.INDEX}}$ samples the future outcome $\tilde{\omega}_a$ (i.e., draws $\tilde{\theta}_a \sim \mathcal{P}_a(y_a)$ and $\tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}_a)$ for $n = 1, \ldots, T$), and finds a threshold value on the deterministic outside option that makes the arm barely worth trying:

$$\lambda_a^*(\tilde{\omega}_a) \triangleq \sup \left\{ \lambda \in \mathbb{R} \; ; \; \varphi_a(\lambda, \tilde{\omega}_a) \geq 0 \right\}. \tag{53}$$

By the definition of $\varphi_a(\lambda, \omega_a)$, the threshold value $\lambda_a^*(\tilde{\omega}_a)$ measures the value of arm a as an opportunity cost of not pulling arm a at all, given a particular outcome $\tilde{\omega}_a$. We use the value $\lambda_a^*(\tilde{\omega}_a)$ as an index of arm a so that the index policy plays the arm with the largest index, i.e., $A^{\text{IRS.INDEX}} = \operatorname{argmax}_a \lambda_a^*(\tilde{\omega}_a)$.

Although the monotonicity of the mapping $\lambda \mapsto \varphi_a(\lambda, \tilde{\omega}_a)$ is not theoretically proven, we observe that the bisection search works sufficiently well in our numerical experiments. Since each instance of single-armed bandit problems requires $O(T)$ computations to solve, the entire procedure for arm selection requires a run time of $O(c_b \times KT)$, where c_b represents the number of iterations in a bisection search. See §B.3 for the implementation details.

In addition to the IRS.INDEX policy described above, some numerical experiments include a heuristic variation of it, called IRS.INDEX*, that is obtained by using

$$\varphi_a(\lambda, \omega_a) \triangleq \max_{1 \leq n \leq T} \left\{ \sum_{i=1}^{n} \left(\hat{\mu}_{a,i-1}(\omega_a) - \lambda - \left(\Gamma_i^\lambda(\omega_a) - \Gamma_0^\lambda(\omega_a) \right) \right) \right\}, \tag{54}$$

instead of (50). This alternative formulation yields indices that are relatively stable across the different samples of outcome $\tilde{\omega}_a$.

We note that our index, $\lambda_a^*(\tilde{\omega}_a)$, is a random approximation of the finite-horizon Gittins (FH-Gittins) index studied in Kaufmann et al. (2012a), Niño-Mora (2011), and Lattimore (2016). The original FH-Gittins algorithm precisely solves the single-armed bandit problem, which is shown to be an optimal stopping problem in which one must decide when to stop pulling the stochastic arm as one's belief state evolves stochastically. Applying the information relaxation framework to the single-armed bandit problem, we solve, instead, a simple deterministic problem in which one must find a deterministic schedule optimized to a particular belief trajectory associated with a randomly generated outcome $\tilde{\omega}$. As in the previous algorithms, the penalties help us to obtain a solution close to the optimal stopping policy of the original single-armed bandit problem.

4. Analysis

In this section, we provide theoretical analyses that characterize IRS policies and performance bounds in particular for TS, IRS.FH, and IRS.V-ZERO.

Remark 5 (Single-period optimality). *When $T = 1$, all of the policies $\pi^{IRS.FH}$, $\pi^{IRS.V-ZERO}$, $\pi^{IRS.V-EMAX}$, and $\pi^{IRS.INDEX}$ take the optimal action; i.e., they pull the myopically best arm $A^* = \operatorname{argmax}_a \bar{\mu}_a(y_a)$.*

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

$$\lim_{T\to\infty} \mathbb{P}\left[A^{IRS.FH}(T, \mathbf{y}) = a\right] = \mathbb{P}\left[A^{TS}(\mathbf{y}) = a\right], \quad \forall a \in \mathcal{A}. \tag{55}$$

Similarly, so does the distribution of the $\pi^{IRS.V-ZERO}$'s action:[8]

$$\lim_{T\to\infty} \mathbb{P}\left[A^{IRS.V-ZERO}(T, \mathbf{y}) = a\right] = \mathbb{P}\left[A^{TS}(\mathbf{y}) = a\right], \quad \forall a \in \mathcal{A}. \tag{56}$$

$A^{\mathrm{TS}}(\mathbf{y})$, $A^{\mathrm{IRS.FH}}(T, \mathbf{y})$ and $A^{\mathrm{IRS.V-ZERO}}(T, \mathbf{y})$ denote the action taken by policies π^{TS}, $\pi^{\mathrm{IRS.FH}}$, and $\pi^{\mathrm{IRS.V-ZERO}}$, respectively, when the remaining time is T and the current belief is $\mathbf{y}$. These actions are random variables, since each of these policies uses a randomly sampled outcome $\tilde{\omega}$ of its own. Remark 5 can be easily verified by observing that, when $T = 1$, $r_1(a, \omega) - z_1(a, \omega; T, \mathbf{y}) = \bar{\mu}_a(y_a)$ for any $a \in \mathcal{A}$ for each of the penalty functions. The results in Proposition 1 follow from Remark 1 stating that the posterior predictive mean reward process converges to the conditional mean reward, i.e., $\lim_{n\to\infty} \hat{\mu}_{a,n}(\tilde{\omega}) = \mu_a(\tilde{\theta}_a)$. The assumption $\mu_i(\theta_i) \neq \mu_j(\theta_j)$ is made to avoid the ambiguity of the tie-breaking rule that is used in TS.

Remark 5 and Proposition 1 illustrate that $\pi^{\mathrm{IRS.FH}}$ and $\pi^{\mathrm{IRS.V-ZERO}}$ behave like TS during the initial decision epochs, gradually shift toward the myopic scheme, and end up with the optimal decision; by contrast, TS continues to explore. 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 time horizon T, and naturally balance exploitation and exploration.

Theorem 2 (Monotonicity of performance bounds). *IRS.FH and IRS.V-ZERO monotonically improve the performance bound*

$$W^{TS}(T, \mathbf{y}) \geq W^{IRS.FH}(T, \mathbf{y}) \geq W^{IRS.V-ZERO}(T, \mathbf{y}), \tag{57}$$

and also

$$W^{TS}(T, \mathbf{y}) \geq W^{IRS.V-EMAX}(T, \mathbf{y}). \tag{58}$$

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

---

[8] We assume a particular selection rule such that $\tilde{a}^{\mathrm{IRS.V-ZERO}} = \operatorname{argmax}_a \tilde{n}_a^*$, as discussed in §3.3.

Empirically (§5), we observe that $W^{\text{IRS.V-ZERO}} \geq W^{\text{IRS.V-EMAX}}$. In addition, we have $W^{\text{IRS.V-EMAX}} \geq W^{\text{ideal}}$ since W^{ideal} is the lowest attainable upper bound (Theorem 1). The second inequality (58) holds in a stronger sense: for every outcome ω, the maximal value of the inner problem associated with W^{TS} is greater than that of the inner problem associated with $W^{\text{IRS.V-EMAX}}$.

While the entire proof is provided in §D.3, we highlight here the main ideas. The first result (57) follows from the monotonicity of the information structure incorporated in each penalty function: TS, IRS.FH, and IRS.V-ZERO replace the realized rewards with $\mathbb{E}(r_t|\boldsymbol{\theta})$, $\mathbb{E}(r_t|\hat{\boldsymbol{\mu}}_{T-1})$, and $\mathbb{E}(r_t|H_{t-1})$, respectively, where $\boldsymbol{\theta}$ is more informative than $\hat{\boldsymbol{\mu}}_{T-1}$, and $\hat{\boldsymbol{\mu}}_{T-1}$ is more informative than H_{t-1} for the DM to infer the value of future reward r_t. Based on this observation, we use a variant of Jensen's inequality to prove the results.[9] The second result (58) is proven based on Theorem 4 of Desai et al. (2012a), which says that if an approximate value function $\hat{V}$ is a supersolution (Definition 2) to the Bellman equation and a penalty function $\hat{z}$ approximates the ideal penalty with $\hat{V}$ in place of V^*, the resulting performance bound $W^{\hat{z}}$ is smaller than $\hat{V}$. By showing that W^{TS} is a supersolution to (15), we prove that $W^{\text{IRS.V-EMAX}} \leq W^{\text{TS}}$ since $z_t^{\text{IRS.V-EMAX}}$ is constructed upon W^{TS}.

Although Theorem 2 compares the performance bound among IRS algorithms, we interpret that its tightness, $W^z - V^*$, reflects the degree of optimism that its corresponding policy π^z possesses. Recall that W^z is the expected value of the best possible payoff when the DM is informed of some future outcomes in advance. Weak duality $W^z \geq V^*$ implies that IRS policies are basically optimistic: an IRS policy takes an action as if it can earn more than the optimal policy in the belief that the sampled outcome is the ground truth. In this sense, the gap $W^z - V^*$ captures how optimistically the policy π^z interprets the sampled outcome. When $W^z - V^*$ is relatively small for a certain penalty function z_t, we may conclude that the penalty function z_t makes the DM less optimistic and induces a policy π^z that performs fewer random explorations.

We further compare the performance of IRS policies using an alternative suboptimality measure. 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 by weak duality (Theorem 1), more desirably, the optimal policy yields a zero suboptimality gap (Theorem 1 and Remark 2). This measure coincides with the conventional regret measure only for TS.

Theorem 3 (Suboptimality gap for natural exponential families). *Consider an MAB instance with a reward distribution that is a natural exponential family distribution, as described in §2.2, in which each arm $a \in \mathcal{A}$ is described with a log-partition function $A_a(\theta_a)$ and a hyperparameter $y_a = (\xi_a, \nu_a)$.*

---

[9] We remark that $W^{\text{IRS.FH}} \geq W^{\text{IRS.V-ZERO}}$ is not an immediate consequence of the fact that $\sigma(\hat{\boldsymbol{\mu}}_{T-1})$ is a stronger filtration than $\sigma(H_{t-1})$. It further relies on a particular structure of MAB problems: the rewards of an arm are independent and identically distributed conditionally on the parameter. See §D.3.2 for a further discussion.

Suppose that all the log-partition functions are L-smooth, i.e.,

$$\frac{d^2}{d\theta_a^2} A_a(\theta_a) \leq L, \quad \forall \theta_a \in \Theta_a, \ a \in \mathcal{A}. \tag{59}$$

Further assume that $\nu_a = \nu$ for all $a \in \mathcal{A}$. Then, for any $T \geq 2$, we have

$$W^{TS}(T, \mathbf{y}) - V(\pi^{TS}, T, \mathbf{y}) \leq 2\sqrt{L} \left[\frac{1}{\sqrt{\nu}} + \sqrt{2 \log T} \times \left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT} \right) \right], \tag{60}$$

$$W^{\text{IRS.FH}}(T, \mathbf{y}) - V(\pi^{\text{IRS.FH}}, T, \mathbf{y}) \leq 2\sqrt{L} \left[\frac{1}{\sqrt{\nu}} + \sqrt{2 \log T} \times \left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT} - \frac{1}{3}\sqrt{\frac{T}{K}} \right) \right], \tag{61}$$

$$W^{\text{IRS.V-ZERO}}(T, \mathbf{y}) - V(\pi^{\text{IRS.V-ZERO}}, T, \mathbf{y}) \leq \sqrt{L} \left[\frac{1}{\sqrt{\nu}} + \sqrt{2 \log T} \times \left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT} - \frac{1}{3}\sqrt{\frac{T}{K}} \right) \right]. \tag{62}$$

Remark 6. *For a Bernoulli MAB with symmetric arms, each of which has a prior $Beta(\alpha, \beta)$ for its mean reward, we have $L = \frac{1}{2}$ and $\sqrt{\nu} = \sqrt{\alpha + \beta}$.*

Remark 7. *For a Gaussian MAB with symmetric arms, each of which has a prior $\mathcal{N}(m, v^2)$ for its mean reward and a noise variance σ^2, we have $L = \sigma$ and $\sqrt{\nu} = v/\sigma$.*

Theorem 3 indirectly shows to the improvements to the suboptimality gaps: although all the bounds have the same asymptotic order of $O(\sqrt{KT \log T})$, the IRS policies improve the leading coefficient or the additional term.[10] These results hold for a wide range of MAB problems including the Bernoulli MAB and the Gaussian MAB as stated in Remark 6 and 7. In addition, our numerical experiments in §5 suggest that such improvements in the suboptimality gaps are roughly equally attributable to both the improvements in the performance bounds W^z and those in the actual policy performance $V(\pi^z)$.

The proof of Theorem 3, provided in §D.4, relies on an essential property of IRS policies that generalizes the "probability matching" property of TS, i.e., a matching between nature's random-

---

[10]Recall that $W^{\text{TS}} - V(\pi^{\text{TS}})$ represents the Bayesian regret of TS. It will be worth mentioning some known results that may be comparable to the bound (60) established in Theorem 3.

For the cases where the reward distributions have a bounded support in $[0, 1]$, Bubeck and Liu (2013) have shown that the Bayesian regret of TS is bounded from above by $14\sqrt{KT}$; and further shown that its asymptotic order is unimprovable in the sense that for any policy there exists a prior distribution such that the policy experiences Bayesian regret no smaller than $\frac{1}{20}\sqrt{KT}$. However, this does not imply that the regret of the Bayesian optimal policy is bounded from below by $\frac{1}{20}\sqrt{KT}$ in the context of Theorem 3, since we consider a specific prior and the policy optimized to that prior.

For Gaussian MAB in the non-Bayesian setting, Agrawal and Goyal (2013) have shown that the regret of TS is $O(\sqrt{KT \log T})$; and further shown that its asymptotic order is unimprovable in the sense that for any policy there exists an instance (i.e., the set of true mean values) such that the policy's regret is at least $\Omega(\sqrt{KT \log K})$.

While there is no result in the literature that is comparable to the other bounds (61) and (62), we conjecture that they will be tight just as the bound for TS (60) is, given the fact all three policies exhibit the identical asymptotic behavior for large T (Proposition 1).

ness and the decision maker's randomness. It is well known that TS is randomized in a way that, conditional on past observations, the probability that an action a is chosen equals 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 past observations, the probability that an action a is chosen equals the probability that the action a is chosen by someone who knows the entire future but is penalized (Proposition 7). Recall that the penalties are designed to penalize the benefit from having additional future information. A better choice of penalty function would prevent the policy π^z from picking an action that is overly optimized for a randomly sampled future realization, which in turn would improve the quality of the decision making.

Given the above observation, our proof utilizes the approach taken by Russo and Van Roy (2014) that exploits the probability matching property of TS to bound its Bayesian regret. More specifically, for each penalty function, we carefully construct a sequence of confidence intervals on the mean reward such that the corresponding policy's instantaneous suboptimality at each time (loss against the hindsight solution) is bounded by the width of the confidence interval approximately. For a better penalty function, the confidence intervals can be made tighter so that the total suboptimality can also be bounded more effectively. In our analysis, the natural exponential family is assumed in order to analyze the concentration of posterior distribution in a closed form, and the smoothness condition on the log-partition function is assumed in order to guarantee that the reward distribution is sub-Gaussian, whereas Russo and Van Roy (2014) consider an arbitrary reward distribution with a bounded support.

5. Numerical Experiments

5.1. Experimental Setup

We conduct numerical simulations to evaluate the effectiveness of our framework in comparison to alternative algorithms. In addition to the IRS algorithms discussed so far, we consider other recently developed algorithms that are particularly suitable for a Bayesian setting: the Bayesian upper confidence bound (Kaufmann et al., 2012a) (BAYES-UCB, with a quantile of $1 - \frac{1}{t}$), information-directed sampling (Russo and Van Roy, 2018) (IDS), the optimistic Gittins index (Gutin and Farias, 2016) (OGI, one-step look-ahead approximation with a discount factor of $\gamma_t = 1 - \frac{1}{t}$), and the Lagrangian index policies suggested in Brown and Smith (2020) (LAGR-RT and LAGR-OT, with a random and an optimal tie-breaking rule, respectively).

Our numerical experiments include Beta-Bernoulli MABs and Gaussian MABs. Given an MAB problem instance specified by the prior distribution $\mathcal{P}(\mathbf{y})$ and the reward distribution $\mathcal{R}$, we simulate the policies and calculate the IRS bounds with respect to the different values of time horizon T.

Let S be the number of simulations we perform. For each $s \in \{1, \ldots, S\}$, we first sample the parameters $\theta_a^{(s)} \sim \mathcal{P}_a(y_a)$ and the rewards $R_{a,n}^{(s)} \sim \mathcal{R}_a(\theta_a^{(s)})$ for all $n \in \{1, \ldots, T_{\max}\}$ and $a \in \mathcal{A}$, which is equivalent to sampling an outcome $\omega^{(s)} \sim \mathcal{I}(\mathbf{y})$. Given the sth sampled outcome $\omega^{(s)}$, for

each time horizon $T \in \{5, 10, 15, \ldots, T_{\max}\}$, we simulate each policy π (that may utilize the time horizon T); i.e., at each time $t = 1, \ldots, T$, the policy makes a decision[11] on which arm to pull, A_t^π, and then the associated reward, $r_t(\mathbf{A}_{1:t}^\pi, \omega^{(s)}) = R_{A_t^\pi, n_t(\mathbf{A}_{1:t}^\pi, A_t^\pi)}^{(s)}$, is revealed accordingly. After simulating one sample path, $\sum_{t=1}^{T} \mu_{A_t^\pi}(\theta_{A_t^\pi}^{(s)})$ is recorded as the performance of π for the sth sample, and the expected performance $V(\pi, T, \mathbf{y})$ is measured by its sample average across S samples for each T.

In order to compute IRS bounds, we use the same set of samples $\omega^{(1)}, \ldots, \omega^{(S)}$. For each penalty function z and for each $T \in \{5, 10, \ldots, T_{\max}\}$, we solve the associated inner problems with respect to $\omega^{(1)}, \ldots, \omega^{(S)}$, and the IRS bound $W^z(T, \mathbf{y})$ is evaluated by taking the average of the maximal values over S instances.

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

$$V(\pi, T, \mathbf{y}) \approx \frac{1}{S} \sum_{s=1}^{S} \left(\sum_{t=1}^{T} \mu_{A_t^\pi}(\theta_{A_t^\pi}^{(s)}) \right), \quad W^z(T, \mathbf{y}) \approx \frac{1}{S} \sum_{s=1}^{S} \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega^{(s)}) - z_t(\mathbf{a}_{1:t}, \omega^{(s)}) \right\}.$$
(63)

Note again that the same outcome $\omega^{(s)}$ is used across the different values of time horizon T and across different algorithms. Sharing the randomness enhances the consistency of the estimates. In what follows, we use 20,000 samples (i.e., $S = 20,000$).

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

$$\text{BayesRegret}(\pi, T, \mathbf{y}) \triangleq \mathbb{E} \left[\sum_{t=1}^{T} \max_a \mu_a(\theta_a) - \mu_{A_t^\pi}(\theta_{A_t^\pi}) \right] = W^{\text{TS}}(T, \mathbf{y}) - V(\pi, T, \mathbf{y}),$$
(64)

which is a conventional measure in performance analysis of Bayesian algorithms as discussed in §3.1. We further calculate the regret (lower) bound obtained from a IRS penalty function z_t:

$$\text{RegretBound}(z, T, \mathbf{y}) \triangleq W^{\text{TS}}(T, \mathbf{y}) - W^z(T, \mathbf{y}).$$
(65)

By weak duality (Theorem 1), we have $\text{BayesRegret}(\pi, T, \mathbf{y}) \geq \text{RegretBound}(z, T, \mathbf{y})$ for any $\pi \in \Pi_\mathbb{F}$. By its definition, the regret bound produced by TS is zero.

5.2. Results

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

$$\mu_a \sim \text{Beta}(1, 1), \quad R_{a,n} \sim \text{Bernoulli}(\mu_a), \quad \forall a \in \{1, 2\}.$$
(66)

---

[11]Recall that IRS policies are randomized policies that perform their own simulations at each time along the sample path. This posterior sampling procedure is independent of the random generation of true outcomes.

We consider relatively short time horizons ($\leq T_{\max} = 200$) since we are focusing on a finite-horizon regime rather than an asymptotic regime. In this particular case, since the state (belief) space is discrete and small in size, $O(T^4)$, we are able to solve the Bellman equations (15) numerically, and thus we can implement the Bayesian optimal policy, which is labeled as OPT in what follows.

Figure 2 shows the regrets (solid lines) of all the policies discussed above and the regret bounds (dashed lines) produced by the IRS algorithms.[12] Table 3 provides further details including the percentage improvement in regret over TS, i.e.,

$$\text{RegretImprovement}(\pi, T, \mathbf{y}) \triangleq 1 - \frac{\text{BayesRegret}(\pi, T, \mathbf{y})}{\text{BayesRegret}(\text{TS}, T, \mathbf{y})},$$

and the improvement in regret bound over TS benchmarked to the regret of the best performing algorithm, i.e.,

$$\text{BoundImprovement}(\pi, T, \mathbf{y}) \triangleq \frac{\text{RegretBound}(z, T, \mathbf{y}) - \text{RegretBound}(z^{\text{TS}}, T, \mathbf{y})}{\min_{\pi'} \text{BayesRegret}(\pi', T, \mathbf{y})}.$$

In Figure 2, note that lower regret curves are better, and higher bound curves are better.

Comparing the IRS algorithms (TS, IRS.FH, IRS.V-ZERO, IRS.V-EMAX, and OPT), we first observe a clear improvement in both the performance of policies and the tightness of bounds, as we adopt a more complicated penalty function, albeit one that requires a longer run time: as visualized in Figure 2, the regret curve approaches the OPT curve from above and the bound curve approaches it from below, where the OPT curve represents the lowest attainable regret that is the highest attainable regret bound at the same time. The suboptimality gap (the gap between a regret curve and its corresponding bound curve) becomes smaller, which is consistent with the implication of Theorem 3.

Finally, we note that the IRS.INDEX policy is outperforming all the other policies; i.e., the regret curve of IRS.INDEX is surprisingly close to the OPT curve. Although it is developed based on IRS.V-EMAX, it performs better than IRS.V-EMAX, and the reasons for that still need to be researched.

---

[12]There also exists a performance bound induced by the Lagrangian index policies. We omit it from Figure 2, however, since that bound is not so tight and thus not informative to be displayed in the same plot; e.g., when $T = 200$, the associated regret bound is -12.54, which is far below the current x-axis.
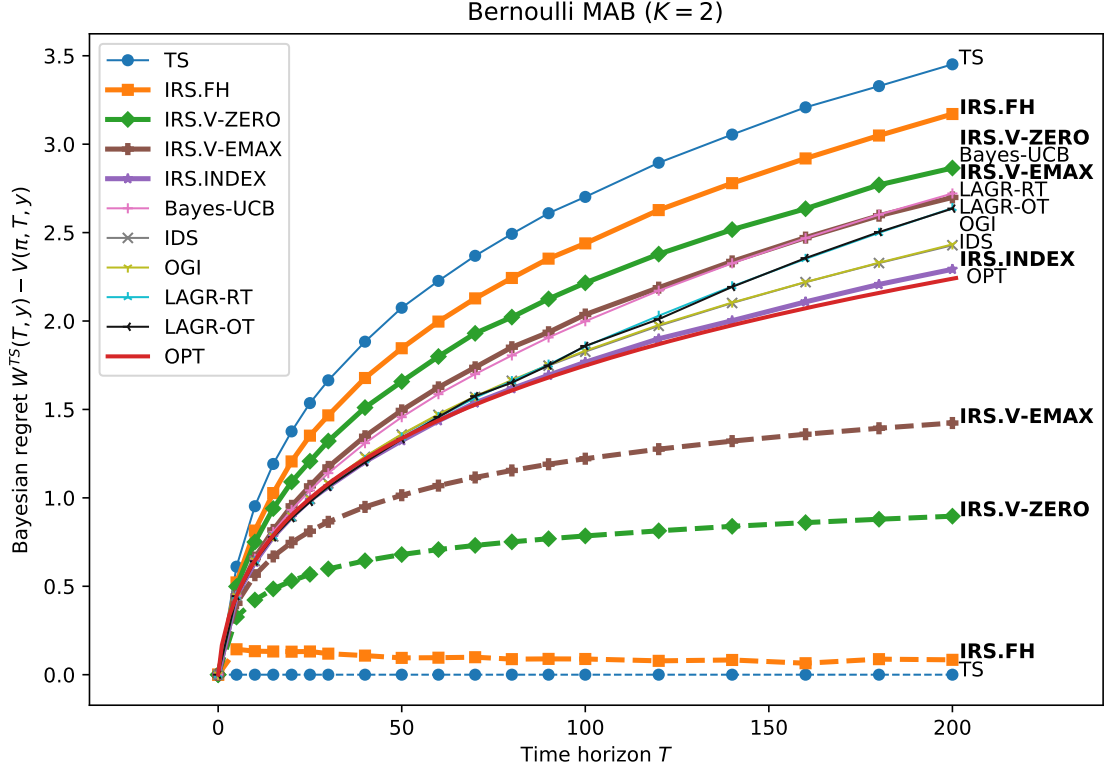
Figure 2: Regret plot for a Bernoulli MAB with two arms. The solid lines represent the (Bayesian) regret of algorithms, $W^{\mathrm{TS}}(T, \mathbf{y}) - V(\pi, T, \mathbf{y})$, and the dashed lines represent the regret bounds that IRS algorithms produce, $W^{\mathrm{TS}}(T, \mathbf{y}) - W^z(T, \mathbf{y})$. Each data point reports the average across 20,000 simulations.

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	Policy run time
TS	3.45 (0.021)	0.0%	0.00 (–)	0.0%	17 ms
Irs.FH	3.17 (0.020)	8.1%	0.08 (0.040)	3.8%	37 ms
Irs.V-Zero	2.87 (0.021)	17.0%	0.90 (0.055)	40.0%	527 ms
Irs.V-EMax	2.70 (0.020)	21.8%	1.42 (0.326)	63.6%	29.5 sec
Irs.Index	2.29 (0.023)	33.6%	–	–	3.6 sec
Bayes-UCB	2.72 (0.020)	21.2%	–	–	44 ms
IDS	2.43 (0.028)	29.6%	–	–	3.7 sec
OGI	2.43 (0.028)	29.5%	–	–	262 ms
Lagr-RT	2.64 (0.046)	23.5%	-12.54	–	19 ms*
Lagr-OT	2.64 (0.046)	23.6%	-12.54	–	14 ms*
Opt	**2.24** (–)	35.1%	**2.24** (–)	100.0%	–

Table 3: Simulation results for a Bernoulli MAB with two arms when $T = 200$. The best results are emphasized with bold letters. The third and fifth columns show the percentage improvements over TS in regret and in bound respectively; e.g., Irs.V-EMax achieves a regret that is 21.8% better than that of TS, and yields a regret bound that accounts for 63.7% of the lowest regret observed empirically. The last column shows the average time required for a policy to make decisions along one sample path including the time required posterior sampling for the case of IRS policies. *Lagr-RT and Lagr-OT require substantial offline computation prior to simulation. This takes around 20 hours in the setting of this simulation.

Bernoulli MAB with ten arms ($K = 10$). We next consider a Bernoulli MAB with ten arms and $T_{\max} = 500$. Irs.V-EMax and Opt are omitted from this simulation due to their computational cost, and so are Lagr-RT and Lagr-OT for long horizons[13] ($T > 350$). Figure 3 and Table 4 show the simulation results. We again observe a monotonic improvement in the performance of policies and the tightness of bounds among IRS algorithms, and the Irs.Index policy still performs best.

---

[13] Lagr-RT and Lagr-OT require substantial offline pre-computation. This involves a convex optimization problem with T decision variables, where a single evaluation of the objective function requires $\Theta(T^3)$ operations. As recommended by Brown and Smith (2020), we have implemented a cutting-plane method using a commercial optimization software (Gurobi), but it takes over a week to complete the pre-computation when $T = 350$.
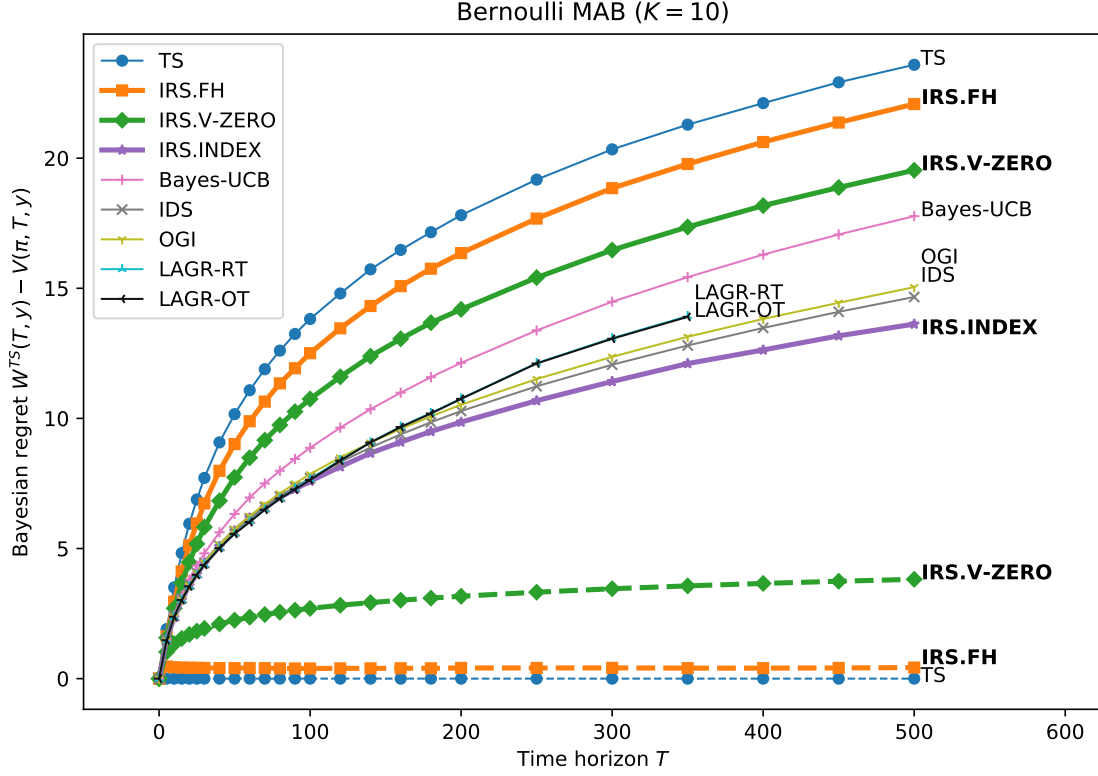
Figure 3: Regret plot for a Bernoulli MAB with ten arms. Lagr-RT and Lagr-OT are simulated only for $T \leq 350$ due to the computational cost (see Footnote 13).

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	Policy run time
TS	23.59 (0.078)	0.0%	0.00 (−)	0.0%	50 ms
Irs.FH	22.08 (0.076)	6.4%	0.43 (0.042)	4.0%	300 ms
Irs.V-Zero	19.54 (0.074)	17.2%	**3.82** (0.058)	35.6%	17.0 sec
Irs.Index	**13.62** (0.080)	42.2%	–	–	56.2 sec
Bayes-UCB	17.77 (0.077)	24.7%	–	–	140 ms
IDS	14.67 (0.093)	37.8%	–	–	16.4 sec
OGI	15.04 (0.092)	36.2%	–	–	2.6 sec

Table 4: Simulation results for a Bernoulli MAB with ten arms when $T = 500$.

Gaussian MABs ($K = 2$ or 10). We next consider Gaussian MABs in which

$$\mu_a \sim \mathcal{N}(0, 1^2), \quad R_{a,n} \sim \mathcal{N}(\mu_a, 1^2), \quad \forall a \in \{1, \dots, K\}. \tag{67}$$

Figure 4 and Table 5 show the case of two arms ($K = 2$), and Figure 5 and Table 6 show the case of ten arms ($K = 10$). The algorithms LAGR-RT and LAGR-OT are not implemented for Gaussian MABs since they require either discrete belief states or some form of state discretization. The results are similar to those of Bernoulli MABs.
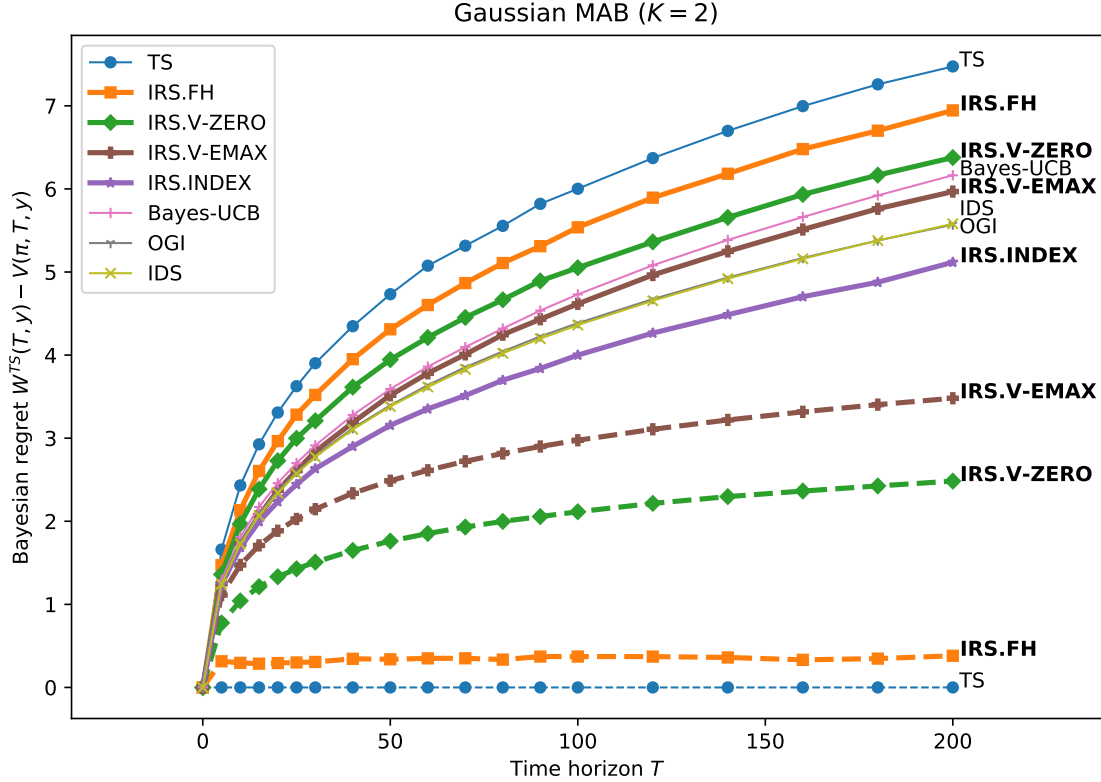

Figure 4: Regret plot for a Gaussian MAB with two arms.

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	Policy run time
TS	7.47 (0.047)	0.0%	0.00 (–)	0.0%	17 ms
IRS.FH	6.94 (0.045)	7.1%	0.38 (0.100)	7.4%	37 ms
IRS.V-ZERO	6.38 (0.048)	14.7%	2.48 (0.133)	48.5%	625 ms
IRS.V-EMAX	5.97 (0.044)	20.2%	**3.48** (1.154)	68.0%	13.3 sec
IRS.INDEX	**5.12** (0.054)	31.5%	–	–	2.2 sec
BAYES-UCB	6.16 (0.045)	17.5%	–	–	38 ms
IDS	5.58 (0.068)	25.3%	–	–	679 ms
OGI	5.57 (0.067)	25.5%	–	–	196 ms

Table 5: Simulation results for a Gaussian MAB with two arms when $T = 200$.
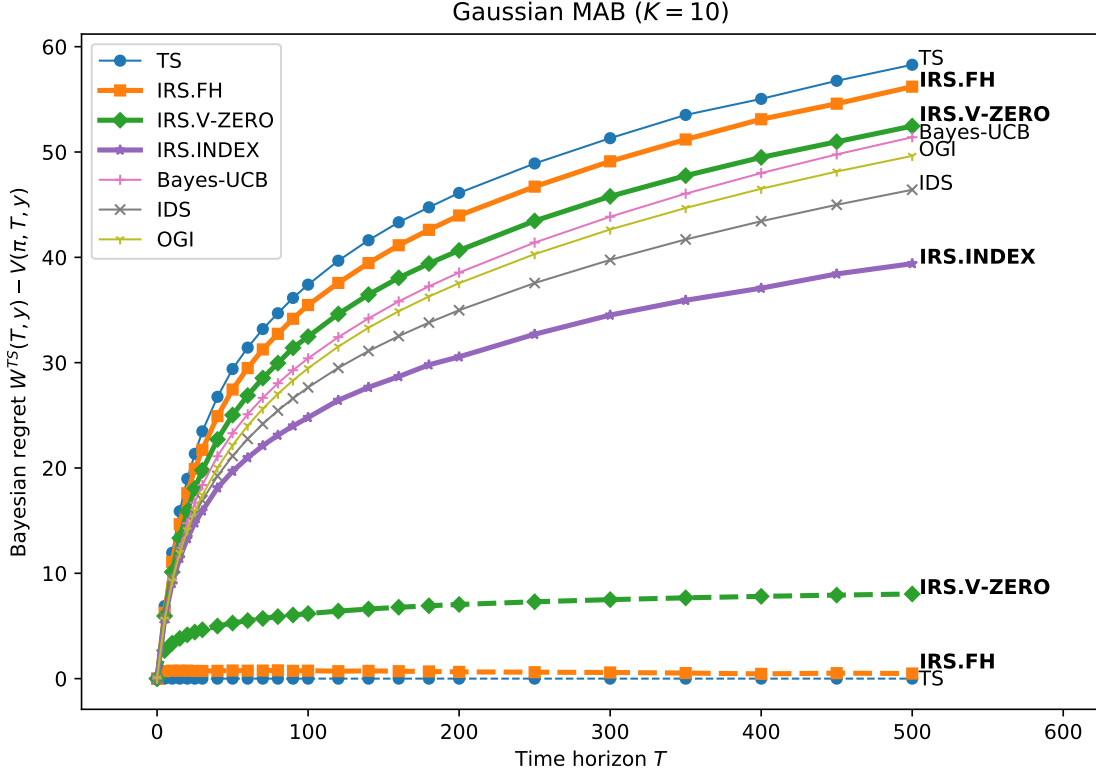
Figure 5: Regret plot for a Gaussian MAB with ten arms.

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	Policy run time
TS	58.28 (0.180)	0.0%	0.00 (–)	0.0%	35 ms
Irs.FH	56.20 (0.180)	3.6%	0.48 (0.156)	1.2%	215 ms
Irs.V-Zero	52.46 (0.188)	10.0%	**8.04** (0.216)	20.4%	13.7 sec
Irs.Index	**39.40** (0.244)	32.4%	–	–	30.4 sec
Bayes-UCB	51.40 (0.178)	11.8%	–	–	77 ms
IDS	46.41 (0.324)	20.4%	–	–	4.0 sec
OGI	49.63 (0.335)	14.8%	–	–	1.6 sec

Table 6: Simultion results for a Gaussian MAB with ten arms when $T = 500$.

Gaussian MAB with different noise variances $(K = 5)$**.** We next consider a problem where

$$\mu_a \sim \mathcal{N}(0, 1^2), \quad R_{a,n} \sim \mathcal{N}(\mu_a, \sigma_a^2), \quad \forall a \in \{1, \ldots, 5\} \tag{68}$$

and $(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) = (0.1, \ 0.4, \ 1, \ 4, \ 10)$. In this MAB instance, it is particularly crucial for the algorithms to consider how much the DM can learn about each of the arms during the remaining time periods, since the difficulty of estimating the mean reward of an arm a heavily depends on the noise level σ_a that varies across the arms.[14]

---

[14]In order for the posterior distribution to be concentrated so as to have a standard deviation of 0.1, for example, one

As shown in Figure 6, Bayes-UCB shows a particularly poor performance, as it keeps pulling arm 5 without considering the fact that arm 5 is too noisy to be learnt within such a short period of time (i.e., $T \leq 500$). By contrast, we observe that our IRS policies and IDS algorithm outperform the Bayes-UCB, OGI, and TS algorithms, since they explicitly take into account the value of exploration by quantifying the informativeness of a new observation for each arm (more specifically, by considering how the belief will change as a new reward realization is revealed). Notably, the Irs.FH policy, which is a very simple modification of TS, significantly improves the performance of TS without degrading its computational efficiency.

The example also illustrates the significance of having a tighter performance bound. If the benchmark is set to $W^{\text{Irs.V-Zero}}$, when $T = 500$, the Irs.Index* policy[15] achieves 94% $\left(= \frac{V(\pi^{\text{Irs.Index*}}, T, \mathbf{y})}{W^{\text{Irs.V-Zero}}(T, \mathbf{y})} \right)$ of the benchmark. If the benchmark is set to W^{TS} instead, as in a conventional regret analysis, we might have concluded that the Irs.Index* policy achieves only 88% $\left(= \frac{V(\pi^{\text{Irs.Index*}}, T, \mathbf{y})}{W^{\text{TS}}(T, \mathbf{y})} \right)$ of that (looser) bound, which would suggest a larger margin of possible improvement.
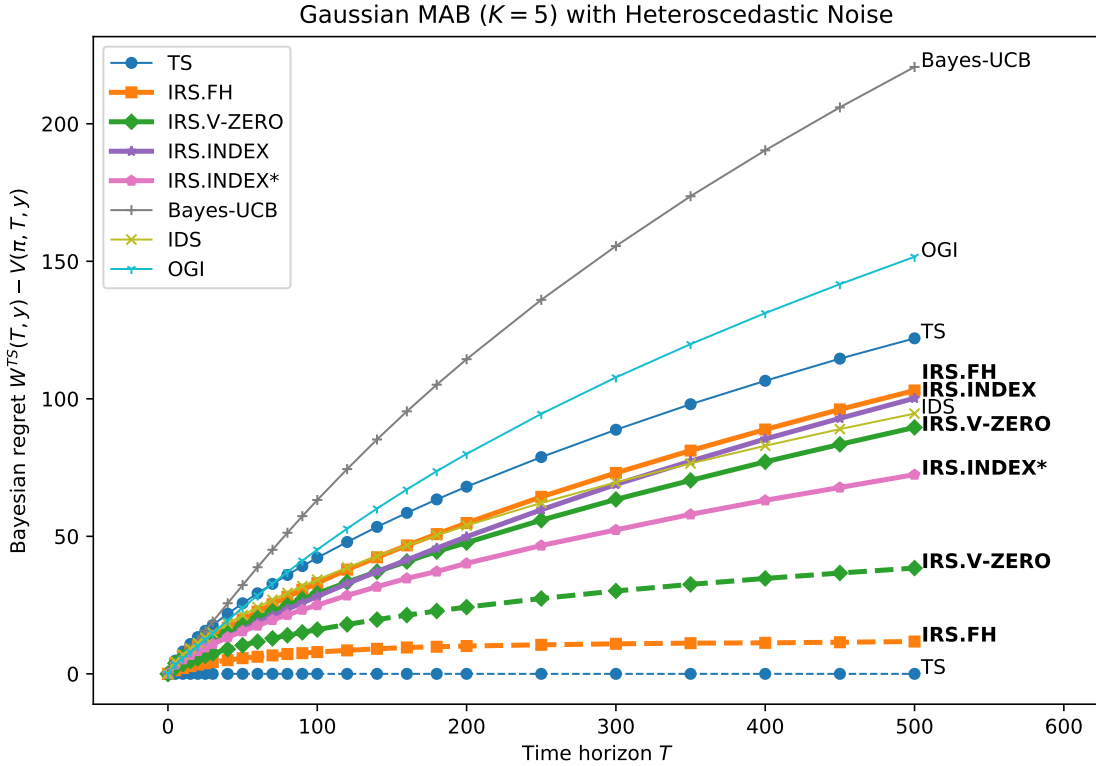

Figure 6: Regret plot for a Gaussian MAB with five arms with different noise variances.

observation is enough for arm 1 whereas 100 and 10,000 observations are required for arm 3 and arm 5, respectively.

[15]The Irs.Index* policy is a heuristic modification of the Irs.Index policy. See §B.3.

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	Policy run time
TS	121.99 (0.615)	0.0%	0.00 (−)	0.0%	34 ms
IRS.FH	103.03 (0.628)	15.5%	11.75 (0.656)	16.2%	128 ms
IRS.V-ZERO	89.59 (0.690)	26.6%	**38.47** (0.827)	53.1%	7.4 sec
IRS.INDEX	100.20 (0.657)	17.9%	–	–	12.8 sec
IRS.INDEX*	**72.43** (0.866)	40.6%	–	–	12.3 sec
BAYES-UCB	220.66 (1.285)	-80.9%	–	–	88 ms
IDS	94.63 (0.817)	22.4%	–	–	2.9 sec
OGI	151.61 (1.030)	-24.3%	–	–	829 ms

Table 7: Simulation results for a Gaussian MAB with five arms with different noise variances when $T = 500$.

6. Extensions

Below, we describe several natural generalizations of the methods developed in this paper beyond the setting of Section 2:

MAB with unknown time horizon. This paper studies finite-time horizon MABs for which we suggest algorithms that exploit the knowledge of the time horizon T and we focus on a relatively small T such that the time horizon becomes an important ingredient in optimally balancing exploration and exploitation. We briefly illustrate how to relax our framework's dependency on T, i.e., how to extend to the setting with an unknown horizon and the setting with an indefinitely long horizon.

First, our framework (penalties, policies, and upper bounds) can naturally incorporate the unknown T within the *Bayesian setting*; i.e., the horizon T is also a random variable whose prior distribution is known. As a simple case, if T is independent of the DM's actions, we can reformulate the objective function of the inner problem as $\sum_{t=1}^{\infty} \gamma_t \left(r_t(\mathbf{a}_{1:t}, \omega) - z_t(\mathbf{a}_{1:t}, \omega) \right)$ where the discount factor $\gamma_t \triangleq \mathbb{P}[T \geq t]$ is the survivor probability, and $r_t(\cdot)$ and $z_t(\cdot)$ are the reward and penalty terms used in the paper. Alternatively, we can treat the random variable T like the random reward realizations by sampling T from its prior distribution while a penalty function (additionally) penalizes for the gain from knowing T (one can imagine that the outcome ω now includes the realization of T and not only the future reward realizations). Structural results such as weak duality and strong duality will continue to hold.

Second, we can consider a practical modification of IRS policies when T is large or infinite. We can construct a dual feasible penalty function that mixes IRS.FH and IRS.V-ZERO,[16] which induces an algorithm whose complexity is $O\left(K \min\{T, T_0\}^2\right)$ for some predefined constant T_0. Alternatively, we can convert the IRS.V-EMAX or IRS.INDEX policy into an anytime policy by setting the inner

---

[16]In its inner problem, IRS.V-ZERO-like penalties are applied for the initial $\lfloor T_0/K \rfloor$ pulls and then IRS.FH-like penalties are applied for the later pulls.

problem's horizon large enough, despite that the performance bound will no longer be obtainable.

MAB in more complicated settings. Even though this paper develops a framework for the stochastic MAB with independent arms, which would be the simplest and oldest problem in the MAB literature, we believe that our framework applies to more complicated settings. Consider the following examples:

- A finite-horizon MAB with correlated arms (e.g., $R_{a,n} \sim \mathcal{N}(\mathbf{x}_a^\top \boldsymbol{\theta}, \sigma_a^2)$ where $\boldsymbol{\theta} \in \mathbb{R}^d$ is shared across the arms, and $\mathbf{x}_a \in \mathbb{R}^d$ is an arm's feature vector): IRS.V-ZERO can be immediately implemented by adopting the DP algorithm discussed in §B.2.

- MAB with the delayed reward realization: IRS.FH can be immediately implemented by simulating the DM's learning process in the presence of delay.

- MAB with a budget constraint (in which each arm consumes a certain amount of budget and the DM wants to maximize the total reward within a limited budget. See Ding et al. (2013)): all IRS algorithms can be implemented by solving a budget-constrained optimization problem instead of a horizon-constrained optimization problem.

In these extensions, we obtain not only the online decision-making policies but also their performance bounds as in this paper. Generally speaking, our framework provides a systematic way of improving TS by taking into account the exploitation-exploration trade-off more carefully, particularly in the presence of some constraint that induces incomplete learning; the main challenge would be to design a suitable penalty function that is tractable yet captures the problem-specific exploration-exploitation trade-off precisely.

7. Conclusion

Contribution to MAB literature. We first highlight that our IRS framework generalizes Thompson sampling to the finite-horizon MAB setting. As pointed out in Russo et al. (2017), TS may perform poorly in time-sensitive learning problems in which exploitation is rather more encouraged than exploration. Interpreted as a special case of IRS policies, it is clear that TS is implicitly assuming an infinite time horizon in the sense that its associated inner problem solves a best-arm identification problem with an infinite number of observations. As summarized in Table 2, IRS algorithms consider more complicated inner problems in which the benefit from exploration is limited by the time-horizon constraint. While maintaining the Bayesian recursive structure of its sequential decision-making process, we improve TS within a unified framework that also includes the Bayesian optimal policy as another special case.

Furthermore, the IRS framework provides a set of (Bayesian) performance bounds that are tighter than the conventional benchmark that has been widely used since Lai and Robbins (1985). We believe that these benchmarks would be useful, in a Bayesian setting, in measuring the optimality of an algorithm or in assessing the intrinsic difficulty of an MAB problem instance.

Contribution to information relaxation literature. The information relaxation framework is certainly a powerful tool to obtain performance bounds in a general class of decision-making problems. Although there have been several studies (Desai et al., 2012b) that elicit a decision-making policy based on this framework, they are limited to using a performance bound as a proxy for the value function. Instead of approximating the value function explicitly, the IRS framework considers simulation-based randomized policies that make each decision that is optimized to a single instance of a simulated environment, and our results show that this scheme is very powerful in online learning problems where random exploration is required.

In applying the information relaxation framework to a particular application, the most challenging task is to find a suitable penalty function that is tractable yet yields a tight performance bound. In this paper, by exploiting the recursive structures embedded in the Bayesian learning process, we derive a series of penalty functions so that users themselves can find a balance between the quality of policies/bounds and the computational cost. We also provide theoretical analyses of the tightness of performance bounds and the suboptimality of associated policies by leveraging the existing analysis developed in the MAB literature. These analytic results would be rare in the information relaxation literature due to the complex nature of the performance bound produced by the information relaxation framework.

References

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson sampling. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 99–107, 2013.

Donald A. Berry and Bert Fristedt. *Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)*. Chapman and Hall, 1985.

Russell N. Bradt, S. M. Johnson, and Samuel Karlin. On sequential designs for maximizing the sum of n observations. *The Annals of Mathematical Statistics*, 27(4):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 and James E. Smith. Index policies and performance bounds for dynamic selection problems. *Management Science*, 66(7):3029–3050, 2020.

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

Sébastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for Thompson sampling. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 2013.

Mark H. A. Davis and Ioannis Karatzas. *A deterministic approach to optimal stopping*, pages 455–466. John Wiley & Sons Ltd, 1994.

Vijay V. Desai, Vivek F. Farias, and Ciamac C. Moallemi. Bounds for Markov decision processes. In F. L. Lewis and D. Liu, editors, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, pages 452–473. John Wiley & Sons, Inc., 2012a.

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

Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, 2013.

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

E. Gutiérrez-Peña and A. F. M. Smith. Conjugate parameterizations for natural exponential families. *Journal of the American Statistical Association*, 90(432):1347–1356, 1995.

Eli Gutin and Vivek F. Farias. Optimistic Gittins indices. In *Advances in Neural Information Processing Systems 29*, pages 3153–3161, 2016.

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

Martin B. Haugh and Octavio R. Lacedelli. Information relaxation bounds for partially observed Markov decision processes. *IEEE Transactions on Automatic Control*, 65(8):3256–3271, 2019.

Martin B. Haugh and Andrew E. B. Lim. Linear-quadratic control and information relaxations. *Operations Research Letters*, 40(6):521–528, 2012.

Martin B. Haugh and Chun Wang. Dynamic portfolio execution and information relaxations. *SIAM Journal on Financial Mathematics*, 5(1):316–359, 2014.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On Bayesian upper confidence bounds for bandit problems. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22, pages 592–600, 2012a.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Proceedings of the 23th International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012b.

Tze Leueng Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tor Lattimore. Regret analysis of the finite-horizon Gittins index strategy for multi-armed bandits. In *Proceedings of the 29th Annual Conference on Learning Theory*, pages 1214–1245, 2016.

Seungki Min, Costis Maglaras, and Ciamac C Moallemi. Thompson sampling with information relaxation penalties. *Advances in Neural Information Processing Systems*, 32, 2019.

José Niño-Mora. Computing a classic index for finite-horizon bandits. *INFORMS Journal on Computing*, 23(2):254–267, 2011.

R. T. Rockafellar and Roger J.-B. 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, 2018.

Daniel Russo, David Tse, and Benjamin Van Roy. Time-sensitive bandit learning and satisficing Thompson sampling. *arXiv preprint arXiv:1704.09028*, 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.

Electronic Companion of "Thompson Sampling with Information Relaxation Penalties"

Seungki Min

KAIST

skmin@kaist.ac.kr

Costis Maglaras

Columbia University

c.maglaras@gsb.columbia.edu

Ciamac C. Moallemi

Columbia University

ciamac@gsb.columbia.edu

Current Revision: February 2022

A. An Illustrative Example

Let us consider a Bernoulli MAB with eight periods ($T = 8$) and three arms ($K = 3$) with the following priors:

$$\mu_1 \sim \text{Beta}(3,1), \quad \mu_2 \sim \text{Beta}(1,1), \quad \mu_3 \sim \text{Beta}(1,3), \tag{69}$$

where $R_{a,n} \sim \text{Bernoulli}(\mu_a)$ for each $a \in \{1,2,3\}$ and $n \in \{1,2,\cdots,8\}$. Given this prior belief, the predictive mean reward of each arm is $\bar{\mu}_1 = \mathbb{E}_{\mu_1 \sim \text{Beta}(3,1)}[\mu_1] = \frac{3}{4}$, $\bar{\mu}_2 = \frac{1}{2}$, and $\bar{\mu}_3 = \frac{1}{4}$, respectively. As an illustrative example, we examine a particular instance where the true outcome ω is given as follows:

	True means $\mu_a(\theta_a)$	**Rewards** $R_{a,n}$							
		$n=1$	2	3	4	5	6	7	8
Arm 1 ($a=1$)	0.235	0	1	1	1	0	0	0	0
Arm 2 ($a=2$)	0.443	1	0	0	1	1	1	1	0
Arm 3 ($a=3$)	0.787	1	1	1	1	0	0	1	1

Table 8: An example of the outcome in a Bernoulli MAB with $K = 3$ and $T = 8$.

If we consider only the priors, arm 1 is best since $\bar{\mu}_1$ is largest among $(\bar{\mu}_1, \bar{\mu}_2, \bar{\mu}_3)$. If, however,

we have full information about the parameter values, arm 3 is best since μ_3 is largest among (μ_1, μ_2, μ_3).

A.1. Inner Problems Induced by Different Penalty Functions

No penalty. To clarify the role of penalties, we first consider the case of zero penalty, i.e., $z_t \equiv 0$, which was not discussed in §3. With zero penalty, the DM at any time earns the current realized reward without adjustment. The clairvoyant DM, who is informed of the outcome ω, can find the best action sequence for this particular outcome ω. Recall that $R_{a,n}$ is defined to be the reward from the nth pull of arm a, not the reward from arm a at time n, and so the DM is not allowed to skip any of the reward realizations and the total reward does not depend on the order of pulls. As depicted in the table below, the optimal solution is to pull arm 1 four times, arm 2 once, and arm 3 three times, which yields a total reward of 7.

	Payoffs under zero penalty								Maximal payoff
	$n=1$	2	3	4	5	6	7	8	
Arm 1	0		1	1	1	0	0	0	0	
Arm 2	1		0	0	1	1	1	1	0	7
Arm 3	1		1	1	1	0	0	1	1	

TS penalty. Next, let us examine the penalty $z_t^{\text{TS}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mu_{a_t}(\theta_{a_t})$ under which the DM earns μ_a whenever playing arm a. The hindsight optimal action sequence is to pull arm 3 (the arm with the largest mean reward μ_a) eight times in a row and the DM can earn a total reward of $T \times \mu_3 = 6.296$ at most.

	Payoffs under z_t^{TS}								Maximal payoff
	$n=1$	2	3	4	5	6	7	8	
Arm 1	.235	.235	.235	.235	.235	.235	.235	.235	
Arm 2	.443	.443	.443	.443	.443	.443	.443	.443	6.296
Arm 3	.787	.787	.787	.787	.787	.787	.787	.787	

IRS.FH penalty. When the penalties are given by $z_t^{\text{IRS.FH}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \hat{\mu}_{a_t, T-1}(\omega)$, the DM earns $\hat{\mu}_{a, T-1}(\omega)$ whenever playing arm a. Recall that $\hat{\mu}_{a, T-1}(\omega)$ is the Bayesian estimate on mean reward of arm a after observing reward realizations $R_{a,1}, \cdots, R_{a,T-1}$. In this particular example, we have $(\hat{\mu}_{1,T-1}, \hat{\mu}_{2,T-1}, \hat{\mu}_{3,T-1}) = \left(\frac{6}{11}, \frac{6}{9}, \frac{6}{11} \right)$ and the maximal payoff is $T \times \hat{\mu}_{2,T-1} = 5.333$, which can be obtained by playing arm 2 throughout the entire time horizon.

IRS.V-Zero penalty. Finally, let us focus on $z_t^{\text{IRS.V-ZERO}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \hat{\mu}_{a_t, n_{t-1}(\mathbf{a}_{1:t-1}, a_t)}$ under which the DM earns $\hat{\mu}_{a, n-1}(\omega)$ from the nth pull of arm a. Since the payoff from an arm changes over time as the Bayesian estimate evolves, playing only one arm is no longer optimal, unlike in the previous two cases. It can be easily verified that the optimal allocation is to play arm 1 six times and arm 2 two times, as visualized in the table below.

	Payoffs under $z_t^{\text{Irs.FH}}$								Maximal payoff
	$n=1$	2	3	4	5	6	7	8	
Arm 1	6/11	6/11	6/11	6/11	6/11	6/11	6/11	6/11	
Arm 2	6/9	6/9	6/9	6/9	6/9	6/9	6/9	6/9	5.333
Arm 3	6/11	6/11	6/11	6/11	6/11	6/11	6/11	6/11	

	Payoffs under $z_t^{\text{Irs.V-Zero}}$								Maximal payoff
	$n=1$	2	3	4	5	6	7	8	
Arm 1	3/4	3/5	4/6	5/7	6/8	6/9	6/10	6/11	
Arm 2	1/2	2/3	2/4	2/5	3/6	4/7	5/8	6/9	5.314
Arm 3	1/4	2/5	3/6	4/7	5/8	5/9	5/10	6/11	

IRS.V-EMax and the ideal penalty. Regarding the penalty functions $z_t^{\text{Irs.V-EMax}}$ and z_t^{ideal}, we cannot visualize the optimal solution with a table since the total payoff depends on the detailed sequence of pulls and not only the number of pulls. While omitting the visual proof of optimality, we have that the action sequence $\mathbf{a}_{1:8}^* = (1,2,2,1,1,1,1,1)$ achieves the maximal payoff of 5.806 under $z_t^{\text{Irs.V-EMax}}$, and $\mathbf{a}_{1:8}^* = (1,1,1,1,1,1,1,1)$ achieves the maximal payoff of 6.063 under z_t^{ideal}. In particular for z_t^{ideal}, the maximal payoff depends only on the prior belief $\mathbf{y}$ and the time horizon T, irrespective of the outcome[1] ω.

We have so far illustrated how the different penalty functions induce the different inner problems and the different best actions given the same outcome ω. The readers may notice from the above examples that, as the penalty function becomes more complicated, the hindsight best action sequence becomes less dependent on a particular realization of ω. Instead, it becomes more dependent on the prior belief.

A.2. IRS Performance Bounds

The maximal payoffs above are calculated for a particular outcome given by Table 8. Recall that the IRS performance bound W^z is defined as the expected value of the maximal payoff where the expectation is taken with respect to the randomness of outcome ω over its prior distribution $\mathcal{I}(T, \mathbf{y})$. We can obtain this value by simulation, i.e., by solving a bunch of inner problems with respect to the randomly generated outcomes $\omega^{(1)}, \omega^{(2)}, \cdots, \omega^{(S)}$ and taking the average of the maximal values. For this particular Bernoulli MAB setting ($T = 8$ with given priors), we obtain the following performance bounds:

W^0	W^{TS}	$W^{\text{Irs.FH}}$	$W^{\text{Irs.V-Zero}}$	$W^{\text{Irs.V-EMax}}$	$W^{\text{ideal}} = V^*$
6.805	6.429	6.279	6.111	6.075	6.063

---

[1]For details, see the proof of the strong duality theorem in §C.1. While the maximal value does not depend on ω, the optimal action sequence still depends on ω. More specifically, it is the sequence of actions that the (non-anticipating) Bayesian optimal policy will take when ω is sequentially revealed.

We observe that the performance bounds are monotone, i.e., $W^0 > W^{\text{TS}} > W^{\text{IRS.FH}} > W^{\text{IRS.V-ZERO}} > W^{\text{IRS.V-EMAX}} > W^{\text{ideal}} = V^*$, which is consistent with Theorem 2.

A.3. Illustration of the IRS Policy (IRS.V-Zero)

We illustrate how the policy $\pi^{\text{IRS.V-ZERO}}$ makes decisions sequentially when the true outcome ω is the one specified in Table 8. At $t = 1$, it first synthesizes a future scenario based on the prior belief (i.e., sampling $\tilde{\omega}_1 \sim \mathcal{I}(\mathbf{y}_0)$) and finds the best action sequence in the presence of penalties $z_t^{\text{IRS.V-ZERO}}$ in the belief that the sampled outcome $\tilde{\omega}_1$ is the ground truth. The following table shows an example in which $\pi^{\text{IRS.V-ZERO}}$ plays arm 1.

$t = 1$	**Priors $\mathbf{y}_0$**	**Payoffs with respect to $\tilde{\omega}_1 \sim \mathcal{I}(\mathbf{y}_0)$**								**Action**
		$n = 1$	2	3	4	5	6	7	8	
Arm 1	Beta$(3, 1)$	3/4	4/5	5/6	6/7	7/8	7/9	8/10	9/11	
Arm 2	Beta$(1, 1)$	1/2	1/3	1/4	1/5	1/6	1/7	2/8	3/9	$a_1 = 1$
Arm 3	Beta$(1, 3)$	1/4	1/5	1/6	1/7	1/8	1/9	1/10	2/11	

As a result of the first action ($a_1 = 1$), we observe that $R_{1,1} = 0$ (encoded in the true outcome ω) and the associated belief is updated from Beta$(3, 1)$ to Beta$(3, 2)$ according to Bayes' rule. In order to make the next decision a_2 at time $t = 2$, $\pi^{\text{IRS.V-ZERO}}$ simulates an outcome for the remaining time horizon, i.e., $\tilde{\omega}_2 \sim \mathcal{I}(\mathbf{y}_1)$, independently of the outcome $\tilde{\omega}_1$ used at $t = 1$. Again, $\pi^{\text{IRS.V-ZERO}}$ finds the best action sequence for this new scenario and takes its first action.[2] The table below shows an instance of $\tilde{\omega}_2$ in which the policy will pull arm 2.

$t = 2$	**Priors $\mathbf{y}_1$**	**Payoffs with respect to $\tilde{\omega}_2 \sim \mathcal{I}(\mathbf{y}_1)$**							**Action**
		$n = 1$	2	3	4	5	6	7	
Arm 1	Beta$(3, 2)$	3/5	4/6	4/7	4/8	4/9	5/10	5/11	
Arm 2	Beta$(1, 1)$	1/2	2/3	3/4	3/5	4/6	4/7	5/8	$a_2 = 2$
Arm 3	Beta$(1, 3)$	1/4	1/5	1/6	1/7	1/8	1/9	1/10	

We can update the prior of arm 2 as a new reward realization $R_{2,1} = 1$ is revealed. In the following decision epochs $t = 3, 4, \cdots$, the policy repeats the same decision-making procedure – (i) samples $\tilde{\omega}_t \sim \mathcal{I}(\mathbf{y}_{t-1})$, (ii) solves the inner problem, and (iii) plays the best arm that the optimal solution suggests – while updating the priors as the true reward realizations are revealed sequentially.

The following table illustrates the last decision epoch. As there remains one time period only, the policy $\pi^{\text{IRS.V-ZERO}}$ tries to maximize $\hat{\mu}_{a,0}(\tilde{\omega}_7) = \bar{\mu}_a(\mathbf{y}_7)$, which is the expected mean reward given the prior at that moment. Such a decision is totally myopic, but it is Bayesian optimal.

---

[2]In case of IRS.V-ZERO, we select the arm with the largest pull allocation as a first action.

$t = 8$	**Priors** $\mathbf{y}_7$	**Payoffs with respect to** $\tilde{\omega}_7 \sim \mathcal{I}(\mathbf{y}_7)$	**Action**
		$n = 1$	
Arm 1	Beta$(6,3)$	6/9	
Arm 2	Beta$(2,2)$	2/4	$a_8 = 1$
Arm 3	Beta$(1,3)$	1/4	

B. Algorithms in Detail

B.1. Implementation of IRS.V-Zero

We provide a pseudo-code of the policy $\pi^{\text{IRS.V-ZERO}}$ 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 5: Arm selection rule of $\pi^{\text{IRS.V-ZERO}}$ when remaining time is T and current belief is $\mathbf{y}$

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

1    $\tilde{\theta}_a \sim \mathcal{P}_a(y_a), \tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}), \quad \forall n \in \{1, \ldots, T\}, \forall a \in \{1, \ldots, K\}$

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

3      $\tilde{y}_{a,0} \leftarrow y_a, \tilde{S}_{a,0} \leftarrow 0$

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

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

6        $\tilde{y}_{a,n} \leftarrow \mathcal{U}_a(\tilde{y}_{a,n-1}, \tilde{R}_{a,n})$

     end

   end

7    $\tilde{M}_{0,0} \leftarrow 0, \tilde{M}_{0,n} \leftarrow -\infty, \forall n \in \{1, \ldots, T\}$

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

9   **for** $n = 0, \cdots, T$ **do**

10       $\tilde{M}_{a,n} \leftarrow \max_{0 \leq m \leq n}\{\tilde{M}_{a-1,n-m} + \tilde{S}_{a,m}\}$

11       $\tilde{L}_{a,n} \leftarrow \text{argmax}_{0 \leq m \leq n}\{\tilde{M}_{a-1,n-m} + \tilde{S}_{a,m}\}$

     end

   end

12    $\tau \leftarrow T$

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

14      $\tilde{n}_a^* \leftarrow \tilde{L}_{a,\tau}$

15      $\tau \leftarrow \tau - \tilde{n}_a^*$

   end

16   **return** $\text{argmax}_a \tilde{n}_a^*$

B.2. Implementation of IRS.V-EMax

We use the notation $\mathbf{y}_t(\mathbf{n}_{1:K}, \omega)$ to denote the belief as a function of pull counts $\mathbf{n}_{1:K} \triangleq (n_1, \cdots, n_K) \in \mathbb{N}_0^K$, based on the observation that the belief is completely determined by how many times each of the arms has been pulled, $\mathbf{n}_{1:K}$, irrespective of the specific sequence in which the arms have been pulled. Given the pull counts $\mathbf{n}_{1:K}$, we define the payoff of pulling arm a one more time after pulling the individual arms $n_1, \cdots, n_K$ times respectively: with $t = \sum_{a=1}^{K} n_a$, the effective payoff associated with arm a at time t is

$$r^z(\mathbf{n}_{1:K}, a, \omega) \triangleq \hat{\mu}_{a,n_a}(\omega) - W^{\mathrm{TS}}\left(T - t - 1, \mathbf{y}_{t+1}(\mathbf{n}_{1:K} + \mathbf{e}_a, \omega)\right) + W^{\mathrm{TS}}\left(T - t - 1, \mathbf{y}_t(\mathbf{n}_{1:K}, \omega)\right),$$
(70)

where $\mathbf{e}_a \in \mathbb{N}_0^K$ is a basis vector such that the ath component is one and the others are zero. Note that we used the fact that $\mathbb{E}\left[W^{\mathrm{TS}}\left(T - t, \mathbf{y}_t\right) \big| H_{t-1}\right] = W^{\mathrm{TS}}\left(T - t, \mathbf{y}_{t-1}\right)$.

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

$$M(\mathbf{n}_{1:K}, \omega) \triangleq \max_{\mathbf{a}_{1:t} \in \mathcal{A}^t} \left\{\sum_{s=1}^{t} r_s(\mathbf{a}_{1:s}, \omega) - z_s^{\mathrm{IRS.V\text{-}EMAX}}(\mathbf{a}_{1:s}, \omega); \; \sum_{s=1}^{t} \mathbf{1}\{a_s = a\} = n_a, \forall a\right\}.$$
(71)

Consequently, the maximal value $M(\mathbf{n}_{1:K}, \omega)$ should satisfy the following Bellman equation:

$$M(\mathbf{n}_{1:K}, \omega) = \max_{a \in \mathcal{A}: n_a \geq 1} \left\{M(\mathbf{n}_{1:K} - \mathbf{e}_a, \omega) + r^z(\mathbf{n}_{1:K} - \mathbf{e}_a, a, \omega)\right\},$$
(72)

i.e., when letting a^* be the maximizer of (72), it is optimal to play arm a^* after making the best effort within the allocation $\mathbf{n}_{1:K} - \mathbf{e}_a$. 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 (72) in an appropriate order. By doing so, we can obtain the maximal value of the original inner problem $(*)$ by evaluating

$$\max_{\mathbf{n}_{1:K} \in N_T} \left\{M(\mathbf{n}_{1:K}, \omega)\right\},$$
(73)

where $N_T \triangleq \{(n_1, \cdots, n_K) \in \mathbb{N}_0^K : \sum_{a=1}^{K} n_a = T\}$, and the performance bound $W^{\mathrm{IRS.V\text{-}EMAX}}$ is the expected value of (73) with respect to the random realization of ω. The optimal action sequence $\mathbf{a}_{1:T}^*$ can be obtained by tracking $M(\mathbf{n}_{1:K}, \omega)$'s backward.

Algorithm 6: Arm selection rule of $\pi^{\text{IRS.V-Zero}}$ when remaining time is T and current belief is $\mathbf{y}$

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

1. $\tilde{\theta}_a \sim \mathcal{P}_a(y_a), \tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}), \quad \forall n \in \{1, \ldots, T\}, \forall a \in \{1, \ldots, K\}$

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

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

4. $\tilde{\Gamma}[\mathbf{n}_{1:K}] \leftarrow \mathbb{E}_{\tilde{\mathbf{y}}(\mathbf{n}_{1:K})}[\max_a \mu_a(\theta_a)]$

   end

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

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

   end

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

8.   **for** *each* $\mathbf{n}_{1:K} \in N_{\leq T} \setminus \{\mathbf{0}\}$ *in order with increasing* $\sum_{a=1}^K n_a$ **do**

9. $\tilde{M}[\mathbf{n}_{1:K}] \leftarrow \max_{a:n_a>0} \left\{\tilde{M}[\mathbf{n}_{1:K} - \mathbf{e}_a] + \tilde{r}^z[\mathbf{n}_{1:K} - \mathbf{e}_a, a]\right\}$

10. $\tilde{A}[\mathbf{n}_{1:K}] \leftarrow \text{argmax}_{a:n_a>0} \left\{\tilde{M}[\mathbf{n}_{1:K} - \mathbf{e}_a] + \tilde{r}^z[\mathbf{n}_{1:K} - \mathbf{e}_a, a]\right\}$

   end

11. $\mathbf{m}_{1:K} \leftarrow \text{argmax}_{\mathbf{n}_{1:K} \in N_T} \left\{\tilde{M}[\mathbf{n}_{1:K}]\right\}$

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

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

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

   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 in which $\sum_{a=1}^K n_a$ 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 $W^{\text{TS}}(T, \mathbf{y}) = T \times \mathbb{E}_{\mathbf{y}}[\max_a \mu_a(\theta_a)]$ which has to be evaluated $O(T^K)$ times in total. There is no simple closed-form expression in general, and it should be evaluated with numerical integration or Monte Carlo sampling.

B.3.   Implementation of IRS.Index

We first prove the identity that was utilized in §3.5, and then provide the pseudo code for IRS.INDEX policy.

Proposition 2. *The optimization problem* (46) *can be reformulated as*

$$\max_{0 \leq n \leq T} \left\{ T \times \Gamma_0^\lambda + (T - n) \times \left(\lambda - \min_{0 \leq i \leq n} \Gamma_i^\lambda\right) + \sum_{i=1}^n \left(\hat{\mu}_{a,i-1} - \Gamma_{i-1}^\lambda\right) \right\}. \tag{74}$$

48

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

Proof. Fix $m \triangleq n_T$, i.e., the total number of pulls on the stochastic arm. Note that if $a_t = 0$, then $(T-t) \times (\Gamma_{n_t}^\lambda - \Gamma_{n_{t-1}}^\lambda) = 0$ since $n_t = n_{t-1}$. The objective function can be represented as

$$\sum_{n=1}^{m} \hat{\mu}_{a,n-1} + (T-m) \times \lambda - \sum_{n=1}^{m} (T-t_n) \times \left(\Gamma_n^\lambda - \Gamma_{n-1}^\lambda\right), \tag{75}$$

where $t_n \triangleq \inf\{t; n_t \geq n\}$ represents the time at which the nth pull on the stochastic arm is made. It suffices to find the optimal pulling times $(t_1, \cdots, t_m)$ with $1 \leq t_1 < t_2 < \cdots < t_m \leq T$ by which $\sum_{n=1}^{m}(T-t_n) \times \left(\Gamma_n^\lambda - \Gamma_{n-1}^\lambda\right)$ is minimized. With $t_0 \triangleq 0$ and $t_{m+1} \triangleq T+1$, we have

$$\sum_{n=1}^{m}(T-t_n) \times \left(\Gamma_n^\lambda - \Gamma_{n-1}^\lambda\right) \tag{76}$$

$$= \sum_{n=1}^{m}(T-t_n) \times \Gamma_n^\lambda - \sum_{n=1}^{m}(T-t_n) \times \Gamma_{n-1}^\lambda \tag{77}$$

$$= \sum_{n=1}^{m}(T-t_n) \times \Gamma_n^\lambda - \sum_{n=0}^{m-1}(T-t_{n+1}) \times \Gamma_n^\lambda \tag{78}$$

$$= \sum_{n=0}^{m}(T-t_n) \times \Gamma_n^\lambda - (T-t_0) \times \Gamma_0^\lambda - \sum_{n=0}^{m}(T-t_{n+1}) \times \Gamma_n^\lambda + (T-t_{m+1}) \times \Gamma_m^\lambda \tag{79}$$

$$= -\Gamma_m^\lambda - T \times \Gamma_0^\lambda + \sum_{n=0}^{m}(t_{n+1} - t_n) \times \Gamma_n^\lambda. \tag{80}$$

Consider the minimum value among $\Gamma_0^\lambda, \ldots, \Gamma_m^\lambda$ and let $n^* \triangleq \operatorname{argmin}_{0 \leq n \leq m} \Gamma_n^\lambda$. In order to minimize (80), it should satisfy that $t_{n+1} - t_n = T - m + 1$ for $n = n^*$ and $t_{n+1} - t_n = 1$ for $n \neq n^*$. For such t_n's, (75) reduces to

$$\sum_{n=1}^{m} \hat{\mu}_{a,n-1} + (T-m) \times \lambda - \left(-\Gamma_m^\lambda - T \times \Gamma_0^\lambda + \sum_{n=0}^{m} \Gamma_n^\lambda + (T-m) \times \min_{0 \leq n \leq m} \Gamma_m^\lambda\right) \tag{81}$$

$$= \sum_{n=1}^{m} \hat{\mu}_{a,n-1} + (T-m) \times \left(\lambda - \min_{0 \leq n \leq m} \Gamma_m^\lambda\right) + T \times \Gamma_0^\lambda - \sum_{n=0}^{m-1} \Gamma_n^\lambda. \tag{82}$$

By taking its maximum value over $m = 0, \cdots, T$, we obtain (49). ∎

The following pseudo code implements the arm selection rule of the IRS.INDEX policy when remaining time is T and current belief is $\mathbf{y}$. In line 14, the infimum can be found via the bisection method, and $\tilde{\mathbf{y}}_{a,0:T} \triangleq (\tilde{y}_{a,0}, \ldots, \tilde{y}_{a,T})$ represents the sequence of beliefs under the sampled outcome.

Algorithm 7: Arm selection rule of IRS.INDEX policy when remaining time is T and current belief is $\mathbf{y}$

Function `IRS.Single.Worth-Trying`$(a, T, \lambda, \tilde{\mathbf{y}}_{a,0:T})$

1    $\tilde{\Gamma}_n^\lambda \leftarrow \mathbb{E}_{\tilde{y}_{a,n}}\left[\max(\mu_a(\theta_a), \lambda)\right], \forall n \in \{0, \ldots, T\}$

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

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

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

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

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

   end

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

8   **if** $\tilde{\varphi}_a \geq 0$ **then**

9   **return** true

   else

10   **return** false

   end

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

11    $\tilde{\theta}_a \sim \mathcal{P}_a(y_a),\ \tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}), \quad \forall n \in \{1, \ldots, T\}, \forall a \in \{1, \ldots, K\}$

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

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

14      $\tilde{\lambda}_a^* \leftarrow \inf\left\{\lambda; \text{IRS.Single.Worth-Trying}(a, T, \lambda, \tilde{\mathbf{y}}_{a,0:T}) = \text{true}\right\}$

   end

15   **return** $\operatorname{argmax}_a \tilde{\lambda}_a^*$

C. Proofs for §3

Proposition 3 (Mean equivalence). *If the penalty function z_t is dual feasible, the presence of penalties does not affect the performance of a non-anticipating policy π: i.e.,*

$$\mathbb{E}_{\mathbf{y}}^\pi\left[\sum_{t=1}^T r_t(\mathbf{A}_{1:t}^\pi, \omega) - z_t(\mathbf{A}_{1:t}^\pi, \omega)\right] = \mathbb{E}_{\mathbf{y}}^\pi\left[\sum_{t=1}^T r_t(\mathbf{A}_{1:t}^\pi, \omega)\right] =: V(\pi, T, \mathbf{y}). \tag{83}$$

Proof. The claim immediately follows from the definition of dual feasibility and the linearity of the expectation operator. ∎

C.1. Proof of Theorem 1

Despite that the results of Theorem 1 were already well established in Brown et al. (2010), we provide the detailed proof as our context is slightly different from that of Brown et al. (2010) regarding the measurability of r_t. We define an appending operator $\oplus$ that concatenates an element into a vector so that $\mathbf{a}_{1:t} = \mathbf{a}_{1:t-1} \oplus a_t$.

Weak duality. Define the filtration for the perfect information relaxation $\mathcal{G}_t \triangleq \mathcal{F}_t \cup \sigma(\omega)$ and consider a relaxed policy space $\Pi_{\mathbb{G}} \triangleq \{\pi : A_t^\pi \text{ is } \mathcal{G}_{t-1}\text{-measurable}, \forall t\}$. Then, we have

$$V^*(T, \mathbf{y}) \triangleq \sup_{\pi \in \Pi_{\mathbb{F}}} \mathbb{E}\left[\sum_{t=1}^T r_t(\mathbf{A}_{1:t}^\pi)\right] \overset{\text{Prop 3}}{=} \sup_{\pi \in \Pi_{\mathbb{F}}} \mathbb{E}\left[\sum_{t=1}^T r_t(\mathbf{A}_{1:t}^\pi) - z_t(\mathbf{A}_{1:t}^\pi)\right] \tag{84}$$

$$\leq \sup_{\pi \in \Pi_{\mathbb{G}}} \mathbb{E}\left[\sum_{t=1}^T r_t(\mathbf{A}_{1:t}^\pi) - z_t(\mathbf{A}_{1:t}^\pi)\right] = \mathbb{E}\left[\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \sum_{t=1}^T r_t(\mathbf{a}_{1:t}) - z_t(\mathbf{a}_{1:t})\right] \tag{85}$$

$$= W^z(T, \mathbf{y}), \tag{86}$$

where the inequality holds 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, respectively, the value function and the state-action value (Q-value) function that are associated with the inner problem $(*)$ given a particular outcome ω under the ideal penalty (22). With $V_{T+1}^{\text{in}} \equiv 0$, we have the following Bellman equation for the inner problem:

$$Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) \triangleq r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) - z_t^{\text{ideal}}(\mathbf{a}_{1:t-1} \oplus a, \omega) + V_{t+1}^{\text{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega), \tag{87}$$

$$V_t^{\text{in}}(\mathbf{a}_{1:t-1}, \omega) = \max_{a \in \mathcal{A}} \left\{Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega)\right\}. \tag{88}$$

We argue by induction to show that

$$V_t^{\text{in}}(\mathbf{a}_{1:t-1}, \omega) = V^*(T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)), \tag{89}$$

$$Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) = Q^*(T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a), \tag{90}$$

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

As a terminal case, when $t = T + 1$, the claim holds trivially, since $V_{T+1}^{\text{in}}(\mathbf{a}_{1:T}, \omega) = 0 = V^*(0, \mathbf{y}_T(\mathbf{a}_{1:T}, \omega))$. Now assume that the claim holds for $t + 1$: i.e., $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,

$$Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) = r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) - z_t^{\text{ideal}}(\mathbf{a}_{1:t-1} \oplus a, \omega) + V_{t+1}^{\text{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega) \tag{91}$$

$$= \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) + V^*(T - t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega)) \mid H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right] \tag{92}$$

$$\underbrace{-V^*(T - t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega)) + V_{t+1}^{\text{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega)}_{=0} \tag{93}$$

$$= \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) + V^*(T - t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega)) \mid H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right] \tag{94}$$

$$= \mathbb{E}_{\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)}\left[R_a + V^*(T - t, \mathcal{U}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a, R_a)) \right] \tag{95}$$

$$= Q^*(T - t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a), \tag{96}$$

where the last equality follows from the original Bellman equation (15). Consequently, we obtain

$$V_t^{\text{in}}(\mathbf{a}_{1:t-1}, \omega) = \max_{a \in \mathcal{A}} \left\{ Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) \right\} \tag{97}$$

$$= \max_{a \in \mathcal{A}} \left\{ Q^*(T - t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a) \right\} \tag{98}$$

$$= V^*(T - t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)). \tag{99}$$

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

$$V_1^{\text{in}}(\emptyset, \omega) = V^*(T, \mathbf{y}), \quad Q_1^{\text{in}}(\emptyset, a, \omega) = Q^*(T, \mathbf{y}, a), \quad \forall \omega. \tag{100}$$

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

C.2. Proof of Remark 2

We proceed on the proof of strong duality. The policy 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 (100), it yields

$$a^{\pi^{\text{ideal}}} = \operatorname*{argmax}_a Q_1^{\text{in}}(\emptyset, a, \tilde{\omega}) = \operatorname*{argmax}_a Q^*(T, \mathbf{y}, a). \tag{101}$$

Therefore, at each moment, irrespective of the sampled outcome $\tilde{\omega}$, the policy π^{ideal} always takes the same action that the Bayesian optimal policy would take. Although there might be some ambiguity regarding tie breaking in argmax, it does not affect the expected performance. Therefore, $V(\pi^{\text{ideal}}, T, \mathbf{y}) = V^*(T, \mathbf{y})$. ∎

C.3. Proof of Remark 3

First observe that for any non-anticipating policy $\pi \in \Pi_{\mathbb{F}}$, since A_t^{π} is $\mathcal{F}_{t-1}$-measurable, we have

$$\mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} r_t(\mathbf{A}_{1:t}^{\pi}, \omega)\right] = \mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} \mathbb{E}\left(r_t(\mathbf{A}_{1:t}^{\pi}, \omega) | \mathcal{F}_{t-1}, \boldsymbol{\theta}\right)\right] = \mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} \mu_{A_t^{\pi}}(\theta_{A_t^{\pi}})\right]. \tag{102}$$

Since $\mathbb{E}[r_t(\mathbf{a}_{1:t}, \omega) | \boldsymbol{\theta}] = \mu_{a_t}(\theta_{a_t})$ for any $\mathbf{a}_{1:t} \in \mathcal{A}^t$, we further deduce that

$$\mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} z_t^{\mathrm{TS}}(\mathbf{A}_{1:t}^{\pi}, \omega)\right] = \mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} r_t(\mathbf{A}_{1:t}^{\pi}, \omega)\right] - \mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} \mu_{A_t^{\pi}}(\theta_{A_t^{\pi}})\right] = 0, \tag{103}$$

and thus z_t^{TS} is dual feasible.

Also observe that $\mathbb{E}[r_t(\mathbf{a}_{1:t}) | \hat{\boldsymbol{\mu}}_{T-1}] = \mathbb{E}[\mu_{a_t} | \hat{\boldsymbol{\mu}}_{T-1}] = \mathbb{E}[\mu_{a_t} | \hat{\boldsymbol{\mu}}_{T-1}, H_{t-1}]$ and $\mathbb{E}[r_t(\mathbf{a}_{1:t}) | H_{t-1}] = \mathbb{E}[\mu_{a_t} | H_{t-1}]$ for any $\mathbf{a}_{1:t} \in \mathcal{A}^t$. We can easily verify that each of penalty functions (22)–(26) has a form of

$$z_t(\mathbf{a}_{1:t}, \omega) = z_t^{\mathrm{TS}}(\mathbf{a}_{1:t}, \omega) + w_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}[w_t(\mathbf{a}_{1:t}, \omega) | G_{t-1}(\mathbf{a}_{1:t-1}, \omega)], \tag{104}$$

for some deterministic function w_t and some relaxed information set $G_{t-1} \supseteq H_{t-1}$. By invoking Proposition 2.3 (iii) of Brown et al. (2010), we have that $z_t^{\mathrm{IRS.FH}} - z_t^{\mathrm{TS}}$, $z_t^{\mathrm{IRS.V\text{-}ZERO}} - z_t^{\mathrm{TS}}$, $z_t^{\mathrm{IRS.V\text{-}EMAX}} - z_t^{\mathrm{TS}}$, and $z_t^{\mathrm{ideal}} - z_t^{\mathrm{TS}}$ are dual feasible, and therefore so are $z_t^{\mathrm{IRS.FH}}$, $z_t^{\mathrm{IRS.V\text{-}ZERO}}$, $z_t^{\mathrm{IRS.V\text{-}EMAX}}$, and z_t^{ideal}. ∎

D. Proofs for §4

D.1. Notes on Regularity

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

$$\mathbb{E}_{\mathbf{y}}|\mu_a(\theta_a)| < \infty, \quad and \quad W^{TS}(T, \mathbf{y}) < \infty, \quad \forall T \in \mathbb{N}. \tag{105}$$

Proof. By Jensen's inequality,

$$\mathbb{E}_{\mathbf{y}}|\mu_a(\theta_a)| = \mathbb{E}_{\mathbf{y}}\left[|\mathbb{E}\left(R_{a,n}|\theta_a\right)|\right] \leq \mathbb{E}_{\mathbf{y}}\left[\mathbb{E}\left(|R_{a,n}||\theta_a\right)\right] = \mathbb{E}_{\mathbf{y}}|R_{a,n}| < \infty. \tag{106}$$

Consequently,

$$\mathbb{E}_{\mathbf{y}}\left[\max_a \mu_a(\theta_a)\right] \leq \mathbb{E}_{\mathbf{y}}\left[\sum_{a=1}^{K} |\mu_a(\theta_a)|\right] = \sum_{a=1}^{K} \mathbb{E}_{\mathbf{y}}|\mu_a(\theta_a)| < \infty. \tag{107}$$

The claim holds since $W^{\mathrm{TS}}(T, \mathbf{y}) = T \times \mathbb{E}_{\mathbf{y}}[\max_a \mu_a(\theta_a)]$. ∎

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

$$\lim_{n\to\infty} \hat{\mu}_{a,n}(\omega; y_a) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} R_{a,i} = \mu_a(\theta_a) \quad almost\ surely, \tag{108}$$

where $\hat{\mu}_{a,n}(\omega; y_a) \triangleq \mathbb{E}_{y_a}[\mu_a(\theta_a)|R_{a,1}, \cdots, R_{a,n}]$.

Proof. Fix a and let $\mathcal{H}_n \triangleq \sigma(R_{a,1}, \cdots, R_{a,n})$. First note that, by the strong law of large numbers, $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} R_{a,i} = \mu_a(\theta_a)$ almost surely. Therefore, $\mu_a(\theta_a)$ is measurable with respect to $\mathcal{H}_\infty \triangleq \bigcup_n \mathcal{H}_n$. Also note that $\hat{\mu}_{a,n} = \mathbb{E}(\mu_a(\theta_a)|\mathcal{H}_n)$ is a Doob martingale adapted to $\mathcal{H}_n$. By Levy's upward theorem, since $\mu_a(\theta_a) \in \mathcal{L}^1$ by Proposition 4, $\hat{\mu}_{a,n}$ converges to $\mathbb{E}(\mu_a(\theta_a)|\mathcal{H}_\infty) = \mu_a(\theta_a)$ almost surely as $n \to \infty$. ∎

D.2. Proof of Proposition 1

Asymptotic behavior of $\pi^{\text{IRS.FH}}$. Let $\tilde{\omega}$ be the sampled outcome used by $\pi^{\text{IRS.FH}}$. By Proposition 5, we have $\lim_{n\to\infty} \hat{\mu}_{a,n}(\tilde{\omega}) = \mu_a(\tilde{\theta}_a)$ for almost all $\tilde{\omega}$. This, 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 $\tilde{\omega}$, yields

$$\underset{a}{\text{argmax}}\, \mu_a(\tilde{\theta}_a) = \underset{a}{\text{argmax}} \lim_{n\to\infty} \hat{\mu}_{a,n}(\tilde{\omega}) = \lim_{n\to\infty} \underset{a}{\text{argmax}}\, \hat{\mu}_{a,n}(\tilde{\omega}) \quad \text{a.s.} \tag{109}$$

Since almost-sure convergence guarantees convergence in distribution, for any $a \in \mathcal{A}$,

$$\lim_{T\to\infty} \mathbb{P}\left[A^{\text{IRS.FH}}(T, \mathbf{y}) = a\right] = \lim_{T\to\infty} \mathbb{P}\left[\underset{a'}{\text{argmax}}\, \hat{\mu}_{a',T-1}(\tilde{\omega}) = a\right] \tag{110}$$

$$= \mathbb{P}\left[\underset{a'}{\text{argmax}}\, \mu_{a'}(\tilde{\theta}_{a'}) = a\right] \tag{111}$$

$$= \mathbb{P}\left[A^{\text{TS}}(\mathbf{y}) = a\right]. \tag{112}$$

Note that we are not assuming that $\pi^{\text{IRS.FH}}$ and π^{TS} share the randomness. The sampled parameters used in π^{TS} are not necessarily the ones used in $\pi^{\text{IRS.FH}}$, but their distributions are identical since they are drawn from the same prior. ∎

Asymptotic behavior of $\pi^{\text{IRS.V-ZERO}}$. To simplify notation, let $A_T^\circ \triangleq A^{\text{IRS.V-ZERO}}(T, \mathbf{y})$. As above, it suffices to show that $\lim_{T\to\infty} A_T^\circ = \text{argmax}_{a\in\mathcal{A}} \mu_a(\tilde{\theta}_a) := A^{\text{TS}}$ for almost all sampled outcome $\tilde{\omega}$. We hide $\tilde{\omega}$ and $\tilde{\theta}_a$ from the notation for the further simplification.

Define

$$\Delta \triangleq \min_{a \neq A^{\text{TS}}} |\mu_{A^{\text{TS}}} - \mu_a| \quad \text{and} \quad M \triangleq \sup_{a\in\mathcal{A}, n\geq 0} |\hat{\mu}_{a,n}|. \tag{113}$$

We have $0 < \Delta < 2M < \infty$ almost surely since $\mu_i(\tilde{\theta}_i) \neq \mu_j(\tilde{\theta}_j)$ for $i \neq j$ and $\lim_{n\to\infty} \hat{\mu}_{a,n} = \mu_a < \infty$

almost surely for all a. In addition, there exists $N \in \mathbb{N}$ such that

$$|\hat{\mu}_{a,n} - \mu_a| < \frac{\Delta}{4}, \quad \forall n \geq N, \quad \forall a \in \mathcal{A}. \tag{114}$$

For such N, we have

$$\inf_{n \geq N} \hat{\mu}_{a^{\mathrm{TS}},n} \geq \sup_{n \geq N} \hat{\mu}_{a,n} + \frac{\Delta}{2}, \quad \forall a \neq A^{\mathrm{TS}}. \tag{115}$$

Note that A^{TS}, Δ, M, and N do not have the dependency on T.

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

$$\mathbf{n}_{1:K}^{\circ} \triangleq \operatorname*{argmax}_{\mathbf{n}_{1:K} \in N_T} \left\{ \sum_{a=1}^{K} \sum_{s=1}^{n_a} \hat{\mu}_{a,s-1} \right\}, \tag{116}$$

where the ties are broken arbitrarily in argmax{}. We let $n^{\circ}(a)$ be the ath component of $\mathbf{n}_{1:K}^{\circ}$. According to the specified arm selection rule, we have $A_T^{\circ} = \operatorname{argmax}_a n^{\circ}(a)$ and hence $n^{\circ}(A_T^{\circ}) \geq \lfloor \frac{T}{2} \rfloor$ $(> N)$. We prove the claim for the following two cases:

Case 1: If $n^{\circ}(a^{\mathrm{TS}}) \geq N$, consider an allocation $\mathbf{n}_{1:K}^{\dagger}$ that is a deviation from the given optimal allocation $\mathbf{n}_{1:K}^{\circ}$ such that arm a^{TS} gets one pull whereas arm A_T° gets one less pull: i.e., $n^{\dagger}(A^{\mathrm{TS}}) = n^{\circ}(A^{\mathrm{TS}}) + 1$, $n^{\dagger}(A_T^{\circ}) = n^{\circ}(A_T^{\circ}) - 1$, and $n^{\dagger}(a) = n^{\circ}(a)$ for any $a \notin \{A^{\mathrm{TS}}, A_T^{\circ}\}$. The change in the total payoff from this deviation is

$$\sum_{a=1}^{K} \sum_{i=1}^{n^{\dagger}(a)} \hat{\mu}_{a,i-1} - \sum_{a=1}^{K} \sum_{i=1}^{n^{\circ}(a)} \hat{\mu}_{a,i-1} = \hat{\mu}_{A^{\mathrm{TS}},n^{\circ}(A^{\mathrm{TS}})} - \hat{\mu}_{A_T^{\circ},n^{\circ}(A_T^{\circ})-1} \geq \frac{\Delta}{2} > 0, \tag{117}$$

where the inequality follows from (115) and that $n^{\circ}(A^{\mathrm{TS}}) \geq N$ and $n^{\circ}(A_T^{\circ}) \geq N$. The allocation $\mathbf{n}_{1:K}^{\dagger}$ is strictly better than $\mathbf{n}_{1:K}^{\circ}$, which contradicts the assumption that $\mathbf{n}_{1:K}^{\circ}$ is an optimal allocation.

Case 2: If $n^{\circ}(A^{\mathrm{TS}}) < N$, consider an allocation $\mathbf{n}_{1:K}^{\dagger}$ that is a deviation from the given optimal allocation $\mathbf{n}_{1:K}^{\circ}$ such that arm A_T° gets no more than N pulls whereas arm A^{TS} gets the remains: i.e.,

$$n^{\dagger}(a) \triangleq \begin{cases} n^{\circ}(A^{\mathrm{TS}}) + (n^{\circ}(A_T^{\circ}) - N) & \text{if } a = A^{\mathrm{TS}}, \\ N & \text{if } a = A_T^{\circ}, \\ n^{\circ}(a) & \text{if } a \notin \{A^{\mathrm{TS}}, A_T^{\circ}\}. \end{cases} \tag{118}$$

By making this the deviation, the total payoff should increase by

$$\sum_{a=1}^{K}\sum_{i=1}^{n^\dagger(a)}\hat{\mu}_{a,i-1} - \sum_{a=1}^{K}\sum_{i=1}^{n^\circ(a)}\hat{\mu}_{a,i-1} \tag{119}$$

$$= \sum_{i=n^\circ(A^{\mathrm{TS}})+1}^{n^\circ(A^{\mathrm{TS}})+(n^\circ(A_T^\circ)-N)}\hat{\mu}_{A^{\mathrm{TS}},i-1} - \sum_{i=N+1}^{n^\circ(A_T^\circ)}\hat{\mu}_{A_T^\circ,i-1} \tag{120}$$

$$\geq -(N-n^\circ(A^{\mathrm{TS}}))\cdot 2M + \sum_{i=N+1}^{n^\circ(A_T^\circ)}\hat{\mu}_{A^{\mathrm{TS}},i-1} - \sum_{i=N+1}^{n^\circ(A_T^\circ)}\hat{\mu}_{A_T^\circ,i-1} \tag{121}$$

$$\geq -(N-n^\circ(A^{\mathrm{TS}}))\cdot 2M + (n^\circ(A_T^\circ)-N)\cdot\frac{\Delta}{2} \tag{122}$$

$$\geq (n^\circ(A_T^\circ)-N)\cdot\frac{\Delta}{2} - 2NM. \tag{123}$$

Since $T \geq 2N + \frac{8MN}{\Delta} + 2$ and $n^\circ(A_T^\circ) \geq \lfloor\frac{T}{2}\rfloor$, the last term is strictly positive, which is a contradiction.

We've shown that for almost all $\tilde{\omega}$, when T is large enough, the optimal allocation $\mathbf{n}_{1:K}^\circ$ must allocate more than a half of the pulls on arm $A^{\mathrm{TS}} = \mathrm{argmax}_a\,\mu_a(\tilde{\theta}_a)$. This concludes the proof.

D.3. Proof of Theorem 2

D.3.1. Proof of "$W^{\mathrm{TS}}(T,\mathbf{y}) \geq W^{\mathrm{IRS.FH}}(T,\mathbf{y})$"

Proof. It immediately follows from Jensen's inequality: since $\max(\cdots)$ is a convex function,

$$W^{\mathrm{TS}}(T,\mathbf{y}) = T\times\mathbb{E}_{\mathbf{y}}\left[\max_a\mu_a(\theta_a)\right] \geq T\times\mathbb{E}_{\mathbf{y}}\left[\max_a\mathbb{E}\left(\mu_a(\theta_a)|\,\hat{\boldsymbol{\mu}}_{T-1}\right)\right] = W^{\mathrm{IRS.FH}}(T,\mathbf{y}). \tag{124}$$

$\blacksquare$

D.3.2. Proof of "$W^{\mathrm{IRS.FH}}(T,\mathbf{y}) \geq W^{\mathrm{IRS.V\text{-}ZERO}}(T,\mathbf{y})$"

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

$$\mathbb{E}\left[\max\left\{X+\varphi(X),0\right\}\right] \geq \mathbb{E}\left[\max\left\{\mathbb{E}(X)+\varphi(X),0\right\}\right]. \tag{125}$$

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 $x\in\mathbb{R}$,

$$f_x(t) \geq f_x(\mu) + (t-\mu)\cdot f_x'(\mu) = \max\{\mu+\varphi(x),0\} + (t-\mu)\cdot\mathbf{1}\{\mu+\varphi(x)\geq 0\}, \quad \forall t,\quad\forall x. \tag{126}$$

By setting $t = x$, we get

$$\max\{x+\varphi(x),0\} = f_x(x) \geq \max\{\mu+\varphi(x),0\} + (x-\mu)\cdot\mathbf{1}\{\mu+\varphi(x)\geq 0\}, \quad\forall x. \tag{127}$$

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,

$$(x - \mu) \cdot \mathbf{1}\{\mu + \varphi(x) \geq 0\} \geq (x - \mu) \cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\}, \quad \forall x \in \mathbb{R}. \tag{128}$$

Combining this with (127), we get

$$\max\{x + \varphi(x), 0\} \geq \max\{\mu + \varphi(x), 0\} + (x - \mu) \cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\}, \quad \forall x \in \mathbb{R}. \tag{129}$$

For random variable X, by taking expectation, we get

$$\mathbb{E}\left[\max\{X + \varphi(X), 0\}\right] \geq \mathbb{E}\left[\max\{\mu + \varphi(X), 0\} + (X - \mu) \cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\}\right] \tag{130}$$

$$\geq \mathbb{E}\left[\max\{\mu + \varphi(X), 0\}\right] + \mathbb{E}(X - \mu) \cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\} \tag{131}$$

$$= \mathbb{E}\left[\max\{\mu + \varphi(X), 0\}\right]. \tag{132}$$

∎

Corollary 1. *On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $\varphi(x, \omega) : \mathbb{R} \times \Omega \to \mathbb{R}$ be a function such that (i) the mapping $x \mapsto \varphi(x, \omega)$ is **increasing** 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*

$$\mathbb{E}\left[\max\left\{X(\omega) + \varphi(X(\omega), \omega), 0\right\}\right] \geq \mathbb{E}\left[\max\left\{\mathbb{E}(X|\mathcal{H})(\omega) + \varphi(X(\omega), \omega), 0\right\}\right]. \tag{133}$$

Proof. Define
$$\mu(\omega) \triangleq \mathbb{E}(X|\mathcal{H})(\omega), \quad I(\omega) \triangleq \mathbf{1}\{\mu(\omega) + \varphi(\mu(\omega), \omega) \geq 0\}. \tag{134}$$

By (129), we have

$$\max\{x + \varphi(x, \omega), 0\} \geq \max\{\mu(\omega) + \varphi(x, \omega), 0\} + (x - \mu(\omega)) \cdot I(\omega), \quad \forall x \in \mathbb{R}, \quad \text{for each } \omega \in \Omega. \tag{135}$$

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

$$\mathbb{E}\left[\max\{X(\omega) + \varphi(X(\omega), \omega), 0\}\right] \geq \mathbb{E}\left[\max\{\mu(\omega) + \varphi(X(\omega), \omega), 0\} + (X(\omega) - \mu(\omega)) \cdot I(\omega)\right] \tag{136}$$

$$= \mathbb{E}\left[\mathbb{E}\left(\max\{\mu(\omega) + \varphi(X(\omega), \omega), 0\} + (X(\omega) - \mu(\omega)) \cdot I(\omega)|\mathcal{H}\right)\right] \tag{137}$$

$$= \mathbb{E}\left[\max\{\mu(\omega) + \varphi(X(\omega), \omega), 0\}\right] + \mathbb{E}\left[\mathbb{E}\left((X(\omega) - \mu(\omega)) \cdot I(\omega)|\mathcal{H}\right)\right] \tag{138}$$

$$= \mathbb{E}\left[\max\{\mathbb{E}(X|\mathcal{H})(\omega) + \varphi(X(\omega), \omega), 0\}\right] \tag{139}$$

$$+ \mathbb{E}\left[\underbrace{(\mathbb{E}(X|\mathcal{H})(\omega) - \mu(\omega))}_{=0} \cdot I(\omega)\right] \tag{140}$$

$$= \mathbb{E}\left[\max\{\mathbb{E}(X|\mathcal{H})(\omega) + \varphi(X(\omega), \omega), 0\}\right]. \tag{141}$$

∎

Corollary 2. *On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $(C_0, \cdots, C_T)$ be $\mathcal{H}$-**measurable real-valued ran-***

dom variables for some sub-σ-field $\mathcal{H} \subseteq \mathcal{F}$ *(i.e., C_i's are constants conditioned on $\mathcal{H}$). Then*

$$\mathbb{E}\left[\max_{0 \leq i \leq T}\left\{(i-n)^+ \times X + C_i\right\}\right] \geq \mathbb{E}\left[\max_{0 \leq i \leq T}\left\{\mathbb{E}\left(X|\mathcal{H}\right) \cdot \mathbf{1}\{i \geq n+1\} + (i-n-1)^+ \times X + C_i\right\}\right] \tag{142}$$

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

$$\varphi(x, \omega) \triangleq \max_{n+1 \leq i \leq T}\left\{(i-n-1) \times x + C_i(\omega)\right\} - \max_{0 \leq i \leq n}\left\{C_i(\omega)\right\}. \tag{143}$$

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

$$\mathbb{E}\left[\max_{0 \leq i \leq T}\left\{(i-n)^+ \times X + C_i\right\}\right] \tag{144}$$

$$= \mathbb{E}\left[\max\left\{\max_{n+1 \leq i \leq T}\left\{(i-n) \times X + C_i\right\}, \quad \max_{0 \leq i \leq n} C_i\right\}\right] \tag{145}$$

$$= \mathbb{E}\left[\max\left\{X + \max_{n+1 \leq i \leq T}\left\{(i-n-1) \times X + C_i\right\}, \quad \max_{0 \leq i \leq n} C_i\right\}\right] \tag{146}$$

$$= \mathbb{E}\left[\max\left\{X(\omega) + \underbrace{\max_{n+1 \leq i \leq T}\left\{(i-n-1) \times X(\omega) + C_i(\omega)\right\} - \max_{0 \leq i \leq n} C_i(\omega)}_{=\varphi(X(\omega),\omega)}, \quad 0\right\} + \max_{0 \leq i \leq n} C_i(\omega)\right] \tag{147}$$

$$\geq \mathbb{E}\left[\max\left\{\mathbb{E}\left(X|\mathcal{H}\right)(\omega) + \max_{n+1 \leq i \leq T}\left\{(i-n-1) \times X(\omega) + C_i(\omega)\right\} - \max_{0 \leq i \leq n} C_i(\omega), \quad 0\right\} + \max_{0 \leq i \leq n} C_i(\omega)\right] \tag{148}$$

$$= \mathbb{E}\left[\max\left\{\max_{n+1 \leq i \leq T}\left\{\mathbb{E}\left(X|\mathcal{H}\right) + (i-n-1) \times X + C_i\right\}, \quad \max_{0 \leq i \leq n} C_i\right\}\right] \tag{149}$$

$$= \mathbb{E}\left[\max_{0 \leq i \leq T}\left\{\mathbb{E}\left(X|\mathcal{H}\right) \cdot \mathbf{1}\{i \geq n+1\} + (i-n-1)^+ \times X + C_i\right\}\right]. \tag{150}$$

$\blacksquare$

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

$$N_T \triangleq \left\{\mathbf{n}_{1:K} \in \mathbb{N}_0^K : \sum_{a=1}^{K} n_a = T\right\} \quad \text{and} \quad S_a(n_a) \triangleq \sum_{i=1}^{n_a} \hat{\mu}_{a,i-1}. \tag{151}$$

What we want to show is

$$W^{\text{Irs.FH}} \equiv \mathbb{E}\left[T \times \max_a\{\hat{\mu}_{a,T-1}\}\right] = \mathbb{E}\left[\max_{\mathbf{n}_{1:K} \in N_T}\left\{\sum_{a=1}^{K} n_a \times \hat{\mu}_{a,T-1}\right\}\right] \tag{152}$$

$$\geq \mathbb{E}\left[\max_{\mathbf{n}_{1:K} \in N_T}\left\{\sum_{a=1}^{K} S_a(n_a)\right\}\right] \equiv W^{\text{Irs.V-Zero}}. \tag{153}$$

Further define

$$U_{k,n} \triangleq \mathbb{E}\left[\max_{\mathbf{n}_{1:K} \in N_T} \left\{ \left(\sum_{a=1}^{k-1} S_a(n_a)\right) + \left(S_k(n_k \wedge n) + (n_k - n)^+ \times \hat{\mu}_{a,T-1}\right) + \left(\sum_{a=k+1}^{K} n_a \times \hat{\mu}_{a,T-1}\right) \right\} \right],$$
(154)

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

$$U_{k,n} \geq U_{k,n+1}, \quad \forall k = 1, \cdots, K, \quad \forall n = 0, \cdots, T-1.$$
(155)

Fix k and n. Define a sub-σ-field

$$\mathcal{H} \triangleq \sigma\left(\{R_{a,s}\}_{a=k,1\leq s\leq n} \cup \{R_{a,s}\}_{a\neq k,1\leq s\leq T-1}\right).$$
(156)

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

$$C_i \triangleq \max \left\{ \left(\sum_{a=1}^{k-1} S_a(n_a)\right) + S_k(i \wedge n) + \left(\sum_{a=k+1}^{K} n_a \times \hat{\mu}_{a,T-1}\right) \ : \ \mathbf{n}_{1:K} \in N_T, n_k = i \right\}.$$
(157)

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

$$U_{k,n} = \mathbb{E}\left[\max_{0\leq i\leq T} \left\{(i-n)^+ \times \hat{\mu}_{k,T-1} + C_i\right\}\right].$$
(158)

With $X \triangleq \hat{\mu}_{a,T-1}$,

$$U_{k,n} = \mathbb{E}\left[\max_{0\leq i\leq T} \left\{(i-n)^+ \times X + C_i\right\}\right]$$
(159)

$$\overset{\text{Corollary 2}}{\geq} \mathbb{E}\left[\max_{0\leq i\leq T} \left\{\mathbb{E}(X|\mathcal{H}) \cdot \mathbf{1}\{i \geq n+1\} + (i-n-1)^+ \times X + C_i\right\}\right]$$
(160)

$$\overset{\text{(a)}}{=} \mathbb{E}\left[\max_{0\leq i\leq T} \left\{\hat{\mu}_{k,n} \cdot \mathbf{1}\{i \geq n+1\} + (i-n-1)^+ \times \hat{\mu}_{a,T-1} + C_i\right\}\right]$$
(161)

$$\overset{\text{(b)}}{=} U_{k,n+1}.$$
(162)

Equation (a) holds since $\mathbb{E}(X|\mathcal{H}) = \mathbb{E}(\hat{\mu}_{k,T-1}|\mathcal{H}) = \mathbb{E}(\hat{\mu}_{k,T-1}|R_{k,1}, \cdots, R_{k,n}) = \hat{\mu}_{a,n}$, and equation (b) holds since $S_k(i \wedge n) + \hat{\mu}_{k,n} \cdot \mathbf{1}\{i \geq n+1\} = \sum_{s=1}^{n} \hat{\mu}_{k,s-1} \cdot \mathbf{1}\{i \geq s\} + \hat{\mu}_{k,n} \cdot \mathbf{1}\{i \geq n+1\} = \sum_{s=1}^{n+1} \hat{\mu}_{k,s-1} \cdot \mathbf{1}\{i \geq s\} = S_k(i \wedge (n+1))$. ∎

A note on the proof. One may wonder if the above result can be derived in a simpler way by exploiting the properties of nested filtration (e.g., Proposition 2.3 of Brown et al., 2010). Unlike the proof of $W^{\text{TS}} \geq W^{\text{IRS.FH}}$, however, the proof of $W^{\text{IRS.FH}} \geq W^{\text{IRS.V-ZERO}}$ does not simply follow from the fact that $\sigma(\hat{\boldsymbol{\mu}}_{T-1})$ is larger than $\sigma(H_{t-1})$.

Consider a Bernoulli MAB with $K = 2$, $T = 2$, and a prior distribution Beta$(1, 1)$, and let us

introduce its variation whose reward function is given by $r'_t(\cdot)$ as follows:

$$r'_1(a_1) = r_1(a_1), \quad r'_2(\mathbf{a}_{1:2}) = -\kappa r_2(\mathbf{a}_{1:2}), \tag{163}$$

where $r_t(\cdot)$ is the reward function of the original Bernoulli MAB. When $\kappa > 0$, one can show that

$$W^{\text{IRS.FH}} = \mathbb{E}\left[\max_{\mathbf{a}_{1:T}}\left\{\sum_{t=1}^{T}\mathbb{E}(r'_t(\mathbf{a}_{1:t})|\hat{\boldsymbol{\mu}}_{T-1})\right\}\right] = \frac{7}{12} - \frac{5}{12}\kappa, \tag{164}$$

$$W^{\text{IRS.V-ZERO}} = \mathbb{E}\left[\max_{\mathbf{a}_{1:T}}\left\{\sum_{t=1}^{T}\mathbb{E}(r'_t(\mathbf{a}_{1:t})|H_{t-1})\right\}\right] = \frac{1}{2} - \frac{3}{8}\kappa. \tag{165}$$

If κ is large enough, we obtain $W^{\text{IRS.FH}} < W^{\text{IRS.V-ZERO}}$, which is opposite to the above result.

Recall that the additional gain from knowing the future information can be decomposed into two components; the gain from knowing the immediate reward and the gain from knowing the next belief state, where IRS.V-ZERO considers the former component only. When those two components are not aligned as in this example (i.e., a higher r'_1 leads to a worse next belief state), the DM can exploit the penalties if they penalize only for the first component (e.g., when r'_1 is smaller than expected, the DM will get compensated for this difference but she can still earn the larger reward in the next period).

This is also related to the fact that $z_t^{\text{IRS.V-ZERO}}$ does not correspond to zero penalty under the some (partial) information relaxation, but should be understood as an approximation of z_t^{ideal} under the perfect information relaxation. As opposed to TS and IRS.FH, the optimal solution to the IRS.V-ZERO's inner problem may depend on the entire outcome ω. With the terminology of Brown et al. (2010), there is a mismatch between the filtration that generates the penalties and the filtration that characterizes the relaxed policy space.

D.3.3. Proof of "$W^{\text{TS}}(T, \mathbf{y}) \geq W^{\text{IRS.V-EMAX}}(T, \mathbf{y})$"

To show that $W^{\text{TS}} \geq W^{\text{IRS.V-EMAX}}$, we take a completely different approach that utilizes Theorem 4 in Desai et al. (2012a). We here rephrase the definition and the theorem therein using our notation.

Definition 2 (Supersolution). *An approximate value function $\widehat{V} : \mathbb{N}_0 \times \mathcal{Y} \to \mathbb{R}$ is a **supersolution** to the Bellman equation* (15) *if*

$$\widehat{V}(T, \mathbf{y}) \geq \max_{a \in \mathcal{A}}\left\{\mathbb{E}_{y_a}\left[R_{a,1} + \widehat{V}(T-1, \mathcal{U}(\mathbf{y}, R_{a,1}, r))\right]\right\}, \quad \forall \mathbf{y} \in \mathcal{Y}, \quad \forall T \geq 1, \tag{166}$$

with $\widehat{V}(0, \mathbf{y}) = 0$ for all $\mathbf{y} \in \mathcal{Y}$.

Remark 8. *If $\widehat{V}(\cdot,\cdot)$ is a supersolution, then for any given ω, T, and $\mathbf{y}$,*

$$\widehat{V}(T-t+1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})) \geq \mathbb{E}_{\mathbf{y}}\left[r_t(\mathbf{a}_{1:t-1} \oplus a, \omega; \mathbf{y}) + \widehat{V}(T-t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega; \mathbf{y})) \Big| H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right], \tag{167}$$

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

Lemma 2 (Theorem 4 of Desai et al. (2012a), rephrased). *Consider a penalty function $\hat{z}_t$ generated by $\widehat{V}(\cdot,\cdot)$:*

$$\hat{z}_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}_{\mathbf{y}}\left[r_t(\mathbf{a}_{1:t}, \omega) | H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] \tag{168}$$
$$+ \widehat{V}(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y})) - \mathbb{E}_{\mathbf{y}}\left[\widehat{V}(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y})) \Big| H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right].$$

If $\widehat{V}(\cdot,\cdot)$ is a supersolution, then the performance bound induced by penalty function $\hat{z}_t$ is tighter than $\widehat{V}$: i.e.,

$$W^{\hat{z}}(T, \mathbf{y}) \leq \widehat{V}(T, \mathbf{y}). \tag{169}$$

And this holds in a stronger sense: for each outcome ω, the maximal value of the inner problem with respect to ω (denoted by $V_1^{\hat{z},\mathrm{in}}(\emptyset, \omega; T, \mathbf{y})$ in the proof) is smaller than or equal to $\widehat{V}(T, \mathbf{y})$.

Proof. Let $V_t^{\hat{z},\mathrm{in}}(\cdot)$ be the DP solution of inner problem $(*)$ for a given penalty $\hat{z}_t$ with respect to a particular outcome ω:

$$V_t^{\hat{z},\mathrm{in}}(\mathbf{a}_{1:t-1}, \omega; T, \mathbf{y}) = \max_{a \in \mathcal{A}}\left\{r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) - \hat{z}_t(\mathbf{a}_{1:t-1} \oplus a, \omega; T, \mathbf{y}) + V_{t+1}^{\hat{z},\mathrm{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega; T, \mathbf{y})\right\}, \tag{170}$$

with $V_{T+1}^{\hat{z},\mathrm{in}}(\cdot, \omega; T, \mathbf{y}) = 0$. Then, we have $W^{\hat{z}}(T, \mathbf{y}) = \mathbb{E}\left[V_1^{\hat{z},\mathrm{in}}(\emptyset, \omega; T, \mathbf{y})\right]$. To prove the claim, it suffices to show that, for any given ω,

$$V_t^{\hat{z},\mathrm{in}}(\mathbf{a}_{1:t-1}, \omega; T, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})) \leq \widehat{V}(T-t+1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})), \tag{171}$$

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

We argue by induction. As a terminal case, when $t = T+1$, the inequality (171) holds trivially since both sides are zero. Fix t and suppose that the inequality (171) holds for $t+1$. Omitting ω

for brevity, we get

$$\widehat{V}\left(T-t+1,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right) - V_t^{\hat{z},\text{in}}\left(\mathbf{a}_{1:t-1};T,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right) \tag{172}$$

$$= \widehat{V}\left(T-t+1,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right) - \max_{a\in\mathcal{A}}\left\{r_t(\mathbf{a}_{1:t-1}\oplus a) - \hat{z}_t(\mathbf{a}_{1:t-1}\oplus a;T,\mathbf{y}) + V_{t+1}^{\hat{z},\text{in}}(\mathbf{a}_{1:t-1}\oplus a;T,\mathbf{y})\right\} \tag{173}$$

$$= \min_{a\in\mathcal{A}}\left\{ \begin{array}{l} \underbrace{\widehat{V}\left(T-t,\mathbf{y}_t(\mathbf{a}_{1:t})\right) - V_{t+1}^{\hat{z},\text{in}}(\mathbf{a}_{1:t-1}\oplus a;T,\mathbf{y})}_{\geq 0 \;\; (\because \text{ induction hypothesis})} \\ + \underbrace{\widehat{V}\left(T-t+1,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1}\oplus a) + \widehat{V}\left(T-t,\mathbf{y}_t(\mathbf{a}_{1:t-1}\oplus a)\right)\Big| H_{t-1}\right]}_{\geq 0 \;\; (\because \text{ Remark 8})} \end{array}\right\} \tag{174}$$

$$\geq 0. \tag{175}$$

∎

Proof of "$W^{\text{TS}}(T,\mathbf{y}) \geq W^{\text{Irs.V-Emax}}(T,\mathbf{y})$." Recall that $z_t^{\text{Irs.V-Emax}}$ is a penalty function generated by W^{TS}. We observe that $W^{\text{TS}}(\cdot,\cdot)$ is a supersolution: for any T and $\mathbf{y}$,

$$W^{\text{TS}}(T,\mathbf{y}) = \mathbb{E}_{\mathbf{y}}\left[T \times \max_{a\in\mathcal{A}}\mu_a(\theta_a)\right] \tag{176}$$

$$= \mathbb{E}_{\mathbf{y}}\left[\max_{a\in\mathcal{A}}\mu_a(\theta_a)\right] + W^{\text{TS}}(T-1,\mathbf{y}) \tag{177}$$

$$\geq \max_{a\in\mathcal{A}}\left\{\mathbb{E}_{y_a}\left[\mu_a(\theta_a)\right] + W^{\text{TS}}(T-1,\mathbf{y})\right\} \tag{178}$$

$$= \max_{a\in\mathcal{A}}\left\{\mathbb{E}_{\mathbf{y}}\left[R_{a,1} + W^{\text{TS}}(T-1,\mathbf{y})\right]\right\} \tag{179}$$

$$= \max_{a\in\mathcal{A}}\left\{\mathbb{E}_{\mathbf{y}}\left[R_{a,1} + W^{\text{TS}}(T-1,\mathcal{U}(\mathbf{y},a,R_{a,1}))\right]\right\}. \tag{180}$$

The last equality holds since $\mathbb{E}\left[W^{\text{TS}}\left(T-1,\mathcal{U}(\mathbf{y},a_1,r_1(a_1,\omega))\right)\right] = W^{\text{TS}}(T-1,\mathbf{y})$, as argued in (39). By Lemma 2, we have $W^{\text{Irs.V-Emax}}(T,\mathbf{y}) \leq W^{\text{TS}}(T,\mathbf{y})$ which also holds in a stronger sense. ∎

D.4. Proof of Theorem 3

D.4.1. Suboptimality Decomposition

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

$$Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1},a,\omega;T,\mathbf{y}) = r_t(\mathbf{a}_{1:t-1}\oplus a,\omega) - z_t(\mathbf{a}_{1:t-1}\oplus a,\omega;T,\mathbf{y}) \tag{181}$$
$$+ V_{t+1}^{z,\text{in}}(\mathbf{a}_{1:t-1}\oplus a,\omega;T,\mathbf{y}),$$

$$V_t^{z,\text{in}}(\mathbf{a}_{1:t-1},\omega;T,\mathbf{y}) = \max_{a\in\mathcal{A}}\left\{Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1},a,\omega;T,\mathbf{y})\right\}, \tag{182}$$

with $V_{T+1}^{z,\text{in}}(\cdot, \omega; T, \mathbf{y}) \equiv 0$. Additionally define the total payoff of an action sequence and the hindsight best action under penalties:

$$\mathcal{S}^z(\mathbf{a}_{1:T}, \omega; T, \mathbf{y}) \triangleq \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}), \tag{183}$$

$$a_t^{z,*}(\mathbf{a}_{1:t-1}, \omega; T, \mathbf{y}) \triangleq \underset{a \in \mathcal{A}}{\operatorname{argmax}} \left\{ Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a, \omega; T, \mathbf{y}) \right\}. \tag{184}$$

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 non-anticipating policy $\pi \in \Pi_{\mathbb{F}}$ and a dual-feasible penalty function z_t, the suboptimality gap is the sum of the instantaneous suboptimalities of individual actions taken by π along the sample path: i.e.,*

$$W^z(T, \mathbf{y}) - V(\pi, T, \mathbf{y}) = \mathbb{E}_{\mathbf{y}} \left[\max_{\mathbf{a}_{1:T}} \left\{ \mathcal{S}^z(\mathbf{a}_{1:T}, \omega; T, \mathbf{y}) \right\} - \mathcal{S}^z(\mathbf{A}_{1:T}^{\pi}, \omega; T, \mathbf{y}) \right] \tag{185}$$

$$= \mathbb{E}_{\mathbf{y}} \left[\sum_{t=1}^{T} \max_a \left\{ Q_t^{z,\text{in}}(\mathbf{A}_{1:t-1}^{\pi}, a, \omega; T, \mathbf{y}) \right\} - Q_t^{z,\text{in}}(\mathbf{A}_{1:t-1}^{\pi}, A_t^{\pi}, \omega; T, \mathbf{y}) \right], \tag{186}$$

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 mean equivalence (Proposition 3). Now fix ω, T, and $\mathbf{y}$. Consider the (pathwise) suboptimality of the action sequence $\mathbf{A}_{1:T}^{\pi}$ compared to the clairvoyant optimal solution. It can be decomposed into the instantaneous suboptimalty incurred by the individual action at each time:

$$\max_{\mathbf{a}_{1:T}} \left\{ \mathcal{S}^z(\mathbf{a}_{1:T}) \right\} - \mathcal{S}^z(\mathbf{A}_{1:T}^{\pi}) = \sum_{t=1}^{T} \max_a \left\{ Q_t^{z,\text{in}}(\mathbf{A}_{1:t-1}^{\pi}, a) \right\} - Q_t^{z,\text{in}}(\mathbf{A}_{1:t-1}^{\pi}, A_t^{\pi}). \tag{187}$$

By taking expectation, we obtain the second equality. $\blacksquare$

The next lemma shows that the instantaneous suboptimalty of the first action can be expressed in terms of mean reward metrics for each of the IRS penalty functions.

Lemma 3. *Fix time horizon T, prior belief $\mathbf{y}$, and the true outcome ω, and hide the dependency on them in notation for $Q_1^{z,\text{in}}(\cdot)$, $a_1^{z,*}(\cdot)$, $\mu_a(\cdot)$ and $\hat{\mu}_{a,n}(\cdot)$. For each of the penalty functions z^{TS}, $z^{\text{IRS.FH}}$, and $z^{\text{IRS.V-ZERO}}$, the instantaneous suboptimalty of action $a \in \mathcal{A}$ satisfies the following:*
(1) When $z \equiv z^{TS}$,

$$Q_1^{z,\text{in}}(a_1^{z,*}) - Q_1^{z,\text{in}}(a) = \mu_{a_1^{z,*}} - \mu_a. \tag{188}$$

(2) When $z \equiv z^{\text{IRS.FH}}$,

$$Q_1^{z,\text{in}}(a_1^{z,*}) - Q_1^{z,\text{in}}(a) = \hat{\mu}_{a_1^{z,*}, T-1} - \hat{\mu}_{a, T-1}. \tag{189}$$

(3) When $z \equiv z^{\text{V-Zero}}$,

$$Q_1^{z,\text{in}}(a_1^{z,*}) - Q_1^{z,\text{in}}(a) \leq \max_{0 \leq n \leq T-1} \left\{ \hat{\mu}_{a_1^{z,*},n} \right\} - \hat{\mu}_{a,0}. \tag{190}$$

Proof. (1) When $z \equiv z^{\text{TS}}$, we have

$$Q_1^{z,\text{in}}(a) = \mu_a + (T-1) \times \max_{a'} \mu_{a'}. \tag{191}$$

Since the last term does not depend on action a, the claim follows.

(2) When $z \equiv z^{\text{IRS.FH}}$, we obtain the claim by replacing μ_a with $\hat{\mu}_{a,T-1}$ in the above proof.

(3) When $z \equiv z^{\text{IRS.V-Zero}}$, recall that the associated inner problem is to find an optimal allocation: i.e.,

$$\max_{\mathbf{n}_{1:K} \in N_T} \left\{ \sum_{a=1}^{K} \sum_{i=0}^{n_a-1} \hat{\mu}_{a,i} \right\}. \tag{192}$$

Let $\mathbf{n}_{1:K}^*$ be the optimal allocation. Observe that the suboptimality is incurred only when $n_a^* = 0$, it is no worse than $\hat{\mu}_{a^*,n_{a^*}^*} - \hat{\mu}_{a,0}$ (the loss if the payoff when pulling a one more time but pulling $a_1^{z,*}$ one less time). Since $n_{a^*}^* \leq T-1$, the claim follows. ∎

D.4.2. Recursive Structure of IRS Penalty Functions

To describe the recursive structure of Bayesian MAB problems explicitly, we define a shift operator $\mathcal{M}_t : \mathcal{A}^t \times \Omega \to \Omega$,

$$\mathcal{M}_t(\mathbf{a}_{1:t}, \omega) \triangleq (R_{a,n_a}; \forall n_a > n_t(\mathbf{a}_{1:t}, a), \forall a \in \mathcal{A}). \tag{193}$$

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

Remark 9 (Recursive structure of remaining uncertainties). *Conditioned on $\mathcal{H}_{t-1}(\mathbf{a}_{1:t-1}, \omega)$, the remaining uncertainties are sufficiently described by $\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})$, i.e.,*

$$\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega) | H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \quad \sim \quad \mathcal{I}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})). \tag{194}$$

Remark 10 (Recursive structure of IRS penalties). *Each of penalty functions (22)–(26) has the following form:*

$$z_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}) = \varphi^z(\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega), T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})), \tag{195}$$

for some function $\varphi^z : \Omega \times \mathbb{N} \times \mathcal{Y} \to \mathbb{R}$, i.e., the penalty at each time 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 prior belief $\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)$ at that moment.

Remark 9 immediately follows from Bayes' rule, and Remark 10 can be easily verified. We

observe the recursive structure of the sequential inner problems that the DM solves throughout the decision-making process, which can be characterized by the following property.

Proposition 7 (Generalized posterior sampling). *For each of penalty functions (22)–(26), the IRS policy π is randomized in such a way that it takes an action a with the probability that the action a is indeed the best action $a_t^{z,*}$ at that moment, i.e.,*

$$\mathbb{P}\left[A_t^\pi = a \,\middle|\, \mathcal{F}_{t-1}\right] = \mathbb{P}\left[a_t^{z,*}(\mathbf{A}_{1:t-1}^\pi, \omega) = a \,\middle|\, \mathcal{F}_{t-1}\right], \quad \forall a, \quad \forall t. \tag{196}$$

The source of uncertainty in the LHS is the randomness of the policy (embedded in $\tilde{\omega}$) and that in the RHS is the randomness of nature (embedded in ω). Here we assume that the tie-breaking rule in argmax *of (184) is identical to the one used when π^z solves the inner problem.*

Proof. Observe that the IRS's action A_t^π can be represented as

$$A_t^\pi = a_1^{z,*}\left(\emptyset, \tilde{\omega}; T - t + 1, \mathbf{y}_{t-1}(\mathbf{A}_{1:t-1}^\pi, \omega; \mathbf{y})\right), \tag{197}$$

where $\tilde{\omega} \sim \mathcal{I}(\mathbf{y}_{t-1}(\mathbf{A}_{1:t-1}^\pi, \omega; \mathbf{y}))$, i.e., the action that the clairvoyant DM will take in an MAB instance specified by horizon $T-t+1$, prior belief $\mathbf{y}_{t-1}(\mathbf{A}_{1:t-1}^\pi, \omega; \mathbf{y})$, and the outcome $\tilde{\omega}$. Therefore, it suffices to verify that the inner problem that π solves at time t is identically distributed with the sub-inner problem with respect to ground-truth ω (i.e., the subproblem given the past action sequence $\mathbf{A}_{1:t-1}^\pi$).

Fix time t, past actions $\mathbf{a}_{1:t-1} = \mathbf{A}_{1:t-1}^\pi$, and the true outcome ω. The sub-inner problem determining $a_t^{z,*}(\mathbf{a}_{1:t-1}, \omega)$ is

$$\max_{\mathbf{a}_{t:T}'} \left\{ \sum_{s=t}^T r_s(\mathbf{a}_{1:t-1} \oplus \mathbf{a}_{t:s}', \omega) - z_s(\mathbf{a}_{1:t-1} \oplus \mathbf{a}_{t:s}', \omega; T, \mathbf{y}) \right\}. \tag{198}$$

By Remark 10, for any $s \in \{t, \dots, T\}$, the penalty at (inner) time s is given by

$$z_s(\mathbf{a}_{1:t-1} \oplus \mathbf{a}_{t:s}', \omega; T, \mathbf{y}) \tag{199}$$

$$= \varphi^z(\mathcal{M}_{s-1}(\mathbf{a}_{1:t-1} \oplus \mathbf{a}_{t:s-1}', \omega), T - s + 1, \mathbf{y}_{s-1}(\mathbf{a}_{1:t-1} \oplus \mathbf{a}_{t:s-1}', \omega; \mathbf{y})) \tag{200}$$

$$= \varphi^z \begin{pmatrix} \mathcal{M}_{s-t}(\mathbf{a}_{t:s-1}', \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)), \\ (T - t + 1) - (s - t), \\ \mathbf{y}_{s-t}(\mathbf{a}_{t:s-1}', \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega); \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})) \end{pmatrix} \tag{201}$$

$$= z_{s-t+1}(\mathbf{a}_{t:s}', \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega); T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})). \tag{202}$$

For rewards, similarly, 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, the

sub-inner problem (198) is reformulated as

$$\max_{\mathbf{a}'_{t:T}} \left\{ \sum_{s=t}^{T} r_{s-t+1}(\mathbf{a}'_{t:s}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)) - z_{s-t+1}(\mathbf{a}'_{t:s}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega); T-t+1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})) \right\}. \tag{203}$$

Given the fact that the shifted outcome $\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)$ and the sampled outcome $\tilde{\omega}$ are identically distributed with $\mathcal{I}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y}))$ conditionally on $H_{t-1}(\mathbf{a}_{1:t-1}, \omega)$ (Remark 9), this sub-inner problem follows the same distribution with

$$\max_{\mathbf{a}'_{1:T-t+1}} \left\{ \sum_{s=1}^{T-t+1} r_s(\mathbf{a}'_{1:s}, \tilde{\omega}) - z_s(\mathbf{a}'_{1:s}, \tilde{\omega}, T-t+1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})) \right\}, \tag{204}$$

which characterizes the IRS's action A_t^π. Therefore, $a_t^{z,*}(\mathbf{A}_{1:t-1}^\pi, \omega)$ is identically distributed with A_t^π conditioned on $\mathcal{F}_{t-1}$. ∎

Remark 11. *Utilizing the recursive structure of IRS penalty functions, Lemma 3 can be extended to describe the instantaneous suboptimality of the tth action. Fix true outcome ω and past actions $\mathbf{a}_{1:t-1}$, and hide the dependency on them in notation for $Q_t^{z,\text{in}}(\cdot)$, $a_t^{z,*}(\cdot)$, $n_t(\cdot)$, $\mu_a(\cdot)$ and $\hat{\mu}_{a,n}(\cdot)$.*
(1) When $z \equiv z^{TS}$,

$$Q_t^{z,\text{in}}(a_t^{z,*}) - Q_t^{z,\text{in}}(a) = \mu_{a_t^{z,*}} - \mu_a. \tag{205}$$

(2) When $z \equiv z^{\text{IRS.FH}}$,

$$Q_t^{z,\text{in}}(a_t^{z,*}) - Q_t^{z,\text{in}}(a) = \hat{\mu}_{a_t^{z,*}, n_{t-1}(a_t^{z,*})+T-t} - \hat{\mu}_{a, n_{t-1}(a)+T-t}. \tag{206}$$

(3) When $z \equiv z^{\text{V-Zero}}$,

$$Q_t^{z,\text{in}}(a_t^{z,*}) - Q_t^{z,\text{in}}(a) \leq \max_{0 \leq n \leq T-t} \left\{ \hat{\mu}_{a_1^{z,*}, n_{t-1}(a_1^{z,*})+n} \right\} - \hat{\mu}_{a, n_{t-1}(a)}. \tag{207}$$

D.4.3. Preliminary Lemmas on MAB with Natural Exponential Family Distributions

We first describe the notion of sub-Gaussian random variable as an effective tool for bounding its tail behavior.

Definition 3 (Sub-Gaussian random variable). *A random variable X is σ-sub-Gaussian if*

$$\mathbb{E}\left[\exp\left(\lambda(X - \mathbb{E}X)\right)\right] \leq \exp\left(\frac{\sigma\lambda^2}{2}\right), \quad \forall \lambda \in \mathbb{R}, \tag{208}$$

for some $\sigma > 0$.

Lemma 4. *Given a random variable X, suppose that there exists $\sigma > 0$ such that*

$$\mathbb{P}\left[X \geq \mathbb{E}X + z\sigma\right] \leq e^{-z^2/2}, \quad \forall z \geq 0. \tag{209}$$

Then, the following holds:

$$\mathbb{E}\left[(X - (\mathbb{E}X + z\sigma))^+\right] \leq \frac{\sigma}{z} e^{-z^2/2}, \quad \forall z > 0. \tag{210}$$

Corollary 3. *If a random variable X is σ-sub-Gaussian, it satisfies the condition of Lemma 4 and hence the inequality (210) holds.*

Proof. With $\mu \triangleq \mathbb{E}X$, we have

$$\mathbb{E}\left[(X - (\mu + z\sigma))^+\right] = \int_{x=\mu+z\sigma}^{\infty} \mathbb{P}\left[X \geq x\right] dx = \int_{t=z}^{\infty} \mathbb{P}\left[X \geq \mu + t\sigma\right]\sigma dt \leq \sigma \int_{t=z}^{\infty} e^{-t^2/2} dt. \tag{211}$$

Utilizing the tail bound established for the standard normal distribution, we can show that

$$\int_{t=z}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \leq \frac{1}{z} \frac{e^{-z^2/2}}{\sqrt{2\pi}}. \tag{212}$$

By combining these two inequalities, we obtain the desired result.

The corollary simply follows from Markov inequality: for any $z \geq 0$ and $\lambda \geq 0$, we have

$$\mathbb{P}[X \geq \mu + z\sigma] = \mathbb{P}\left[e^{\lambda(X-\mu)} \geq e^{\lambda z\sigma}\right] \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda x\sigma}} \leq \exp\left(\frac{\sigma^2\lambda^2}{2} - \lambda z\sigma\right). \tag{213}$$

By taking $\lambda = \frac{z}{\sigma}$, it follows that $\mathbb{P}[X \geq \mu + z\sigma] \leq e^{-z^2/2}$. ∎

We now return to the context of MAB problems and show that the mean reward metrics are sub-Gaussian.

Lemma 5 (Sub-Gaussianity of mean reward metrics). *Consider the setting of Theorem 3, i.e., the reward distribution of arm a is described by an L-smooth log-partition function $A_a(\theta_a)$ and hyper-parameters (ξ_a, ν). Then, the conditional mean reward μ_a is $\sqrt{L/\nu}$-sub-Gaussian: i.e.,*

$$\mathbb{E}_{(\xi_a,\nu)}\left[\exp\left(\lambda(\mu_a - \bar{\mu}_a)\right)\right] \leq \exp\left(\frac{L\lambda^2}{2\nu}\right), \quad \forall \lambda \in \mathbb{R}, \tag{214}$$

where $\bar{\mu}_a = \mathbb{E}_{(\xi_a,\nu)}[\mu_a] = \frac{\xi_a}{\nu}$ is the prior predictive mean reward (i.e., the unconditional mean reward). Furthermore, the posterior predictive mean reward $\hat{\mu}_{a,n}$ is $\sqrt{\frac{Ln}{\nu(\nu+n)}}$-sub-Gaussian: i.e.,

$$\mathbb{E}_{(\xi_a,\nu)}\left[\exp\left(\lambda(\hat{\mu}_{a,n} - \bar{\mu}_a)\right)\right] \leq \exp\left(\frac{\lambda^2}{2} \times \frac{Ln}{\nu(\nu+n)}\right), \quad \forall \lambda \in \mathbb{R}. \tag{215}$$

Proof. We first prove that μ_a is $\sqrt{L/\nu}$-sub-Gaussian. Due to L-smoothness condition, $A_a(\theta_a)$ is

finite valued for all $\theta_a \in \mathbb{R}$. For any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_{(\xi_a, \nu)} \left[\exp\left(\lambda \mu_a\right) \right] \overset{(i)}{=} \mathbb{E}_{(\xi_a, \nu)} \left[\exp\left(\lambda A'_a(\theta_a)\right) \right] \tag{216}$$

$$= \int_{-\infty}^{\infty} \exp\left(\lambda A'_a(\theta_a)\right) \times f_a(\xi_a, \nu) \exp\left(\xi_a \theta_a - \nu A_a(\theta_a)\right) d\theta_a \tag{217}$$

$$= \int_{-\infty}^{\infty} f_a(\xi_a, \nu) \exp\left\{\xi_a \theta_a - \nu A_a(\theta_a) + \lambda A'_a(\theta_a)\right\} d\theta_a \tag{218}$$

$$= \int_{-\infty}^{\infty} f_a(\xi_a, \nu) \exp\left\{\xi_a \theta_a - \nu \left(A_a(\theta_a) - \lambda/\nu \cdot A'_a(\theta_a)\right)\right\} d\theta_a \tag{219}$$

$$\overset{(ii)}{\leq} \int_{-\infty}^{\infty} f_a(\xi_a, \nu) \exp\left\{\xi_a \theta_a - \nu \left(A_a(\theta_a - \lambda/\nu) - \frac{L\lambda^2}{2\nu^2}\right)\right\} d\theta_a \tag{220}$$

$$= \exp\left(\frac{L\lambda^2}{2\nu}\right) \times \int_{-\infty}^{\infty} f_a(\xi_a, \nu) \exp\left\{\xi_a \theta_a - \nu A_a(\theta_a - \lambda/\nu)\right\} d\theta_a \tag{221}$$

$$= \exp\left(\frac{\xi_a \lambda}{\nu} + \frac{L\lambda^2}{2\nu}\right) \times \int_{-\infty}^{\infty} f_a(\xi_a, \nu) \exp\left\{\xi_a (\theta_a - \lambda/\nu) - \nu A_a(\theta_a - \lambda/\nu)\right\} d\theta_a$$
$$\tag{222}$$

$$= \exp\left(\frac{\xi_a \lambda}{\nu} + \frac{L\lambda^2}{2\nu}\right) \times \int_{-\infty}^{\infty} f_a(\xi_a, \nu) \exp\left\{\xi_a \theta_a - \nu A_a(\theta_a)\right\} d\theta_a \tag{223}$$

$$= \exp\left(\frac{\xi_a \lambda}{\nu} + \frac{L\lambda^2}{2\nu}\right), \tag{224}$$

where we have utilized that (i) $\mu_a(\theta_a) = A'_a(\theta_a)$ and (ii) $A_a(\theta_a + \delta) \leq A_a(\theta_a) + \delta A'_a(\theta_a) + \frac{L}{2}\delta^2$. Since $\bar{\mu}_a = \xi_a/\nu$, we obtained the desired result.

Next we focus on the posterior predictive mean reward $\hat{\mu}_{a,n}$. Recall that we have

$$\hat{\mu}_{a,n} = \frac{\xi_a + \sum_{i=1}^{n} R_{a,i}}{\nu + n}. \tag{225}$$

For any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_{(\xi_a, \nu)} \left[\exp \left(\lambda \sum_{i=1}^{n} R_{a,i} \right) \right] = \mathbb{E}_{(\xi_a, \nu)} \left[\mathbb{E} \left\{ \exp \left(\lambda \sum_{i=1}^{n} R_{a,i} \right) \middle| \theta_a \right\} \right] \tag{226}$$

$$\stackrel{(i)}{=} \mathbb{E}_{(\xi_a, \nu)} \left[\mathbb{E} \left\{ \exp \left(\lambda R_{a,1} \right) \middle| \theta_a \right\}^n \right] \tag{227}$$

$$\stackrel{(ii)}{=} \mathbb{E}_{(\xi_a, \nu)} \left[\exp \left\{ A_a \left(\theta_a + \lambda \right) - A_a(\theta_a) \right\}^n \right] \tag{228}$$

$$\stackrel{(iii)}{\leq} \mathbb{E}_{(\xi_a, \nu)} \left[\exp \left\{ \lambda \cdot A_a'(\theta_a) + \frac{L\lambda^2}{2} \right\}^n \right] \tag{229}$$

$$\stackrel{(iv)}{=} \mathbb{E}_{(\xi_a, \nu)} \left[\exp \left\{ n\lambda \cdot \mu_a + \frac{Ln\lambda^2}{2} \right\} \right] \tag{230}$$

$$= \exp \left(n\lambda\bar{\mu}_a + \frac{Ln\lambda^2}{2} \right) \times \mathbb{E}_{(\xi_a, \nu)} \left[\exp \left(n\lambda(\mu_a - \bar{\mu}_a) \right) \right] \tag{231}$$

$$\stackrel{(v)}{\leq} \exp \left(n\lambda\bar{\mu}_a + \frac{Ln\lambda^2}{2} \right) \times \exp \left(\frac{Ln^2\lambda^2}{2\nu} \right) \tag{232}$$

$$= \exp \left(n\lambda\bar{\mu}_a \right) \times \exp \left(\frac{\lambda^2}{2} \times \frac{Ln(\nu + n)}{\nu} \right), \tag{233}$$

where we have utilized that (i) $R_{a,i}$'s are conditionally independent given θ_a, (ii) the moment-generating function of $R_{a,1}$ is given by $\mathbb{E}[\lambda R_a | \theta_a] = \exp \left(A_a(\theta_a + \lambda) - A_a(\theta_a) \right)$, (iii) $A_a(\cdot)$ is L-smooth, (iv) $A_a'(\theta_a) = \mu_a(\theta_a)$, and (v) μ_a is $\sqrt{L/\nu}$-sub-Gaussian. Given that $\mathbb{E}\left[\sum_{i=1}^{n} R_{a,i} \right] = n\bar{\mu}_a$, we just have shown that the sum $\sum_{i=1}^{n} R_{a,i}$ is $\sqrt{\frac{Ln(\nu+n)}{\nu}}$-sub-Gaussian. Therefore, its scaled version $\frac{\sum_{i=1}^{n} R_{a,i}}{\nu + n}$ is $\sqrt{\frac{Ln}{\nu(\nu+n)}}$-sub-Gaussian, and so is $\hat{\mu}_{a,n}$. ∎

Lemma 6. *Consider the setting of Theorem 3. With $\sigma_n \triangleq \sqrt{\frac{Ln}{\nu(\nu+n)}}$, the following holds:*

$$\mathbb{E} \left[\left(\max_{0 \leq i \leq n} \hat{\mu}_{a,i} - \left(\bar{\mu}_a + z\sigma_n \right) \right)^+ \right] \leq \frac{\sigma_n}{z} e^{-z^2/2}, \quad \forall z > 0. \tag{234}$$

Proof. Recall that the posterior predictive mean reward process $\{\hat{\mu}_{a,n}\}_{n \geq 0}$ is the martingale with respect to the filtration generated by reward realizations $R_{a,1}, R_{a,2}, \ldots$ and whose mean is $\bar{\mu}_a$. Therefore, $\{\exp(\lambda\hat{\mu}_{a,n})\}_{n \geq 0}$ is a positive submartingale for any given $\lambda \geq 0$. By Doob's maximal inequality, we deduce that

$$\mathbb{P} \left[\max_{0 \leq i \leq n} \hat{\mu}_{a,i} \geq \bar{\mu}_a + z\sigma_n \right] = \mathbb{P} \left[\max_{0 \leq i \leq n} \exp \left(\lambda(\hat{\mu}_{a,i} - \bar{\mu}_a) \right) \geq \exp \left(\lambda z\sigma_n \right) \right] \leq \frac{\mathbb{E} \left[\exp \left(\lambda(\hat{\mu}_{a,n} - \bar{\mu}_a) \right) \right]}{\exp \left(\lambda z\sigma_n \right)}. \tag{235}$$

By Lemma 5, since $\hat{\mu}_{a,n}$ is σ_n-sub-Gaussian, we further have

$$\frac{\mathbb{E}\left[\exp\left(\lambda(\hat{\mu}_{a,n} - \bar{\mu}_a)\right)\right]}{\exp\left(\lambda z \sigma_n\right)} \leq \frac{\exp\left(\frac{\lambda^2 \sigma_n^2}{2}\right)}{\exp\left(\lambda z \sigma_n\right)} = \exp\left(\frac{\lambda^2 \sigma_n^2}{2} - \lambda z \sigma_n\right). \tag{236}$$

Therefore, by taking $\lambda \triangleq \frac{z}{\sigma_n}$, we have $\mathbb{P}\left[\max_{0 \leq i \leq n} \hat{\mu}_{a,i} \geq \bar{\mu}_a + z\sigma_n\right] \leq e^{-z^2/2}$, and by invoking Lemma 4, we obtain the claim. ∎

D.4.4. Proof of Theorem 3

Lemma 7. *Consider one of the IRS penalty functions z^{TS}, $z^{\mathrm{IRS.FH}}$, and $z^{\mathrm{IRS.V\text{-}ZERO}}$. As discussed in Remark 11, we have*

$$Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}, a_t^{z,*}, \omega) - Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}, a, \omega) \leq \mu_t^U(\mathbf{a}_{1:t-1}, a_t^{z,*}, \omega) - \mu_t^L(\mathbf{a}_{1:t-1}, a, \omega), \tag{237}$$

for some $\mu_t^U(\mathbf{a}_{1:t-1}, a_1^{z,*}, \omega)$ and $\mu_t^L(\mathbf{a}_{1:t-1}, a, \omega)$, where $a_t^{z,*}$ abbreviates $a_t^{z,*}(\mathbf{a}_{1:t-1}, \omega)$. Suppose that there exists a sequence of confidence intervals $\{(L_t(a), U_t(a))\}_{a \in \mathcal{A}, t \in \mathbb{N}}$ such that $(L_t(\cdot), U_t(\cdot))$ is $\sigma(H_{t-1})$-measurable, and

$$\mathbb{E}_{\mathbf{y}}\left[\left(\mu_t^U(\mathbf{a}_{1:t-1}, a, \omega) - U_t(a)\right)^+ \middle| H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] \leq \frac{C_U}{T}, \quad \forall a, \forall t \tag{238}$$

$$\mathbb{E}_{\mathbf{y}}\left[\left(L_t(a) - \mu_t^L(\mathbf{a}_{1:t-1}, a, \omega)\right)^+ \middle| H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] \leq \frac{C_L}{T}, \quad \forall a, \forall t \tag{239}$$

for some constants $C_U > 0$ and $C_L > 0$. Then, for IRS policy π induced by the chosen penalty function, we have

$$W^z(T, \mathbf{y}) - V(\pi, T, \mathbf{y}) \leq C_U + C_L + \sum_{t=1}^{T} \mathbb{E}\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right]. \tag{240}$$

Proof. Let $A_t^* \triangleq a_t^{z,*}(\mathbf{A}_{1:t-1}^\pi, \omega)$, and let $\mathbb{E}_t[\cdot]$ denote $\mathbb{E}[\cdot|\mathcal{F}_{t-1}]$. By Proposition 7 we have

$$\mathbb{E}_t[U_t(A_t^\pi)] = \sum_{a \in \mathcal{A}} U_t(a) \cdot \mathbb{P}_t[A_t^\pi = a] = \sum_{a \in \mathcal{A}} L_t(a) \cdot \mathbb{P}_t[A_t^* = a] = \mathbb{E}_t[U_t(A_t^*)]. \tag{241}$$

Therefore, we have

$$\mathbb{E}_t\left[\mu_t^U(A_t^*) - \mu_t^L(A_t^\pi)\right] \tag{242}$$

$$= \mathbb{E}_t\left[\mu_t^U(A_t^*) - \mu_t^L(A_t^\pi)\right] + \mathbb{E}_t\left[U_t(A_t^\pi) - U_t(A_t^*)\right] + \mathbb{E}_t\left[L_t(A_t^\pi) - L_t(A_t^\pi)\right] \tag{243}$$

$$= \mathbb{E}_t\left[\mu_t^U(A_t^*) - U_t(A_t^*)\right] + \mathbb{E}_t\left[L_t(A_t^\pi) - \mu_t^L(A_t^\pi)\right] + \mathbb{E}_t\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right] \tag{244}$$

$$\leq \mathbb{E}_t\left[\left(\mu_t^U(A_t^*) - U_t(A_t^*)\right)^+\right] + \mathbb{E}_t\left[\left(L_t(A_t^\pi) - \mu_t^L(A_t^\pi)\right)^+\right] + \mathbb{E}_t\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right]. \tag{245}$$

We further observe that

$$\mathbb{E}_t\left[\left(\mu_t^U(A_t^*) - U_t(A_t^*)\right)^+\right] = \sum_{a\in\mathcal{A}}\mathbb{E}_t\left[\left(\mu_t^U(a) - U_t(a)\right)^+\right]\mathbb{P}_t[A_t^* = a] \le \frac{C_U}{T}\sum_{a\in\mathcal{A}}\mathbb{P}_t[A_t^* = a] = \frac{C_U}{T}.$$
(246)

Similarly, we have $\mathbb{E}_t\left[\left(L_t(A_t^\pi) - \mu_t^L(A_t^\pi)\right)^+\right] \le \frac{C_L}{T}$. Combining all these results, we have

$$W(T, \mathbf{y}) - V(\pi, T, \mathbf{y}) \overset{\text{Prop 6}}{=} \mathbb{E}\left[\sum_{t=1}^T Q_t^{z,\text{in}}(A_t^*) - Q_t^{z,\text{in}}(A_t^\pi)\right]$$
(247)

$$\le \mathbb{E}\left[\sum_{t=1}^T \mu_t^U(A_t^*) - \mu_t^L(A_t^\pi)\right]$$
(248)

$$= \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_t\left[\mu_t^U(A_t^*) - \mu_t^L(A_t^\pi)\right]\right]$$
(249)

$$\le \mathbb{E}\left[\sum_{t=1}^T \left(\frac{C_U}{T} + \frac{C_L}{T} + \mathbb{E}_t\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right]\right)\right]$$
(250)

$$\le C_U + C_L + \sum_{t=1}^T \mathbb{E}\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right].$$
(251)

∎

We are now ready to prove Theorem 3. To facilitate simpler notation, we define

$$N_{t-1}^\pi(a) \triangleq n_{t-1}(\mathbf{A}_{1:t-1}^\pi, a), \quad \hat{\mu}_t^\pi(a, n) \triangleq \hat{\mu}_{a, N_{t-1}^\pi(a)+n},$$
(252)

which represent, respectively, the number of pulls on arm a prior to time t under policy π, and the posterior predictive mean reward process given the past actions $\mathbf{A}_{1:t-1}^\pi$. Observe that for each $a \in \mathcal{A}$, the process $\{\hat{\mu}_t^\pi(a, n)\}_{n\ge 0}$ is a martingale, as discussed Remark 1.

Further define

$$\Delta_t^\pi(a, n) \triangleq \sqrt{\frac{L}{\nu + N_{t-1}^\pi(a)} \times \frac{n}{\nu + N_{t-1}^\pi(a) + n}},$$
(253)

which is measurable with respect to $\mathcal{F}_{t-1}$. In the context of Theorem 3, the prior/posterior of arm a at time t is described by the hyperparameters $\left(\xi_a + \sum_{i=1}^{N_{t-1}^\pi(a)} R_{a,i}, \nu + N_{t-1}^\pi(a)\right)$ that converges to μ_a, and therefore Lemma 5 implies that $\hat{\mu}_t^\pi(a, n)$ is $\Delta_t(a, n)$-sub-Gaussian *conditioned* on $\mathcal{F}_{t-1}$.

(1) Suboptimality analysis for TS (60)**.** As discussed in Remark 11, for TS, we have

$$Q_t^{z,\text{in}}(a_t^{z,*}) - Q_t^{z,\text{in}}(a) = \mu_{a_t^{z,*}} - \mu_a = \hat{\mu}_t^\pi(a_t^{z,*}, \infty) - \hat{\mu}_t^\pi(a, \infty).$$
(254)

We construct the confidence intervals as follows:

$$U_t(a) \triangleq \hat{\mu}_t^\pi(a, 0) + \sqrt{2\log T} \times \Delta_t^\pi(a, \infty), \quad L_t(a) \triangleq \hat{\mu}_t^\pi(a, 0) + \sqrt{2\log T} \times \Delta_t^\pi(a, \infty),$$
(255)

where $\Delta_t^\pi(a, \infty) = \lim_{n\to\infty} \Delta_t^\pi(a, n) = \sqrt{\frac{L}{\nu + N_{t-1}^\pi(a)}}$ so that μ_a is $\Delta_t^\pi(a, \infty)$-sub-Gaussian conditioned on $\mathcal{F}_{t-1}$. By Lemma 4, we have

$$\mathbb{E}\left[(\mu_a - U_t(a))^+ \,\middle|\, \mathcal{F}_{t-1}\right] \leq \frac{\Delta_t^\pi(a, \infty)}{\sqrt{2\log T}} e^{-\frac{2\log T}{2}} \leq \frac{\sqrt{L/\nu}}{T}, \tag{256}$$

where we use the fact that $2\log T \geq 1$ for any $T \geq 2$. Symmetrically, we have $\mathbb{E}\left[(L_t(a) - \mu_a)^+ \,\middle|\, \mathcal{F}_{t-1}\right] \leq \frac{\sqrt{L/\nu}}{T}$. By Lemma 7, we have

$$W^{\mathrm{TS}}(T, \mathbf{y}) - V(\pi^{\mathrm{TS}}, T, \mathbf{y}) \leq 2\sqrt{L/\nu} + \sum_{t=1}^{T} \mathbb{E}\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right] \tag{257}$$

$$= 2\sqrt{L/\nu} + 2\sqrt{2\log T} \sum_{t=1}^{T} \Delta_t^\pi(A_t^\pi, \infty). \tag{258}$$

Further observe that

$$\sum_{t=1}^{T} \Delta_t^\pi(A_t^\pi, \infty) = \sum_{t=1}^{T} \sqrt{\frac{L}{\nu + N_{t-1}^\pi(A_t^\pi)}} = \sum_{a\in\mathcal{A}} \sum_{n=0}^{N_T^\pi(a)-1} \frac{\sqrt{L}}{\sqrt{\nu + n}} = \sum_{a\in\mathcal{A}} \left(\frac{\sqrt{L}}{\sqrt{\nu}} + \sum_{n=1}^{N_T^\pi(a)-1} \frac{\sqrt{L}}{\sqrt{\nu + n}}\right) \tag{259}$$

$$\leq \sum_{a\in\mathcal{A}} \left(\frac{\sqrt{L}}{\sqrt{\nu}} + \sum_{n=1}^{N_T^\pi(a)-1} \frac{\sqrt{L}}{\sqrt{n}}\right) \leq \sum_{a\in\mathcal{A}} \left(\frac{\sqrt{L}}{\sqrt{\nu}} + \int_{x=0}^{N_T^\pi(a)} \frac{\sqrt{L}}{\sqrt{x}}dx\right) = \frac{K\sqrt{L}}{\sqrt{\nu}} + 2\sqrt{L} \sum_{a\in\mathcal{A}} \sqrt{N_T^\pi(a)}. \tag{260}$$

By utilizing Cauchy–Schwartz inequality, we deduce that

$$\sum_{a\in\mathcal{A}} \sqrt{N_T^\pi(a)} \leq \sqrt{K \sum_{a\in\mathcal{A}} N_T(a)} = \sqrt{KT}. \tag{261}$$

Combining all these results, we conclude that

$$W^{\mathrm{TS}}(T, \mathbf{y}) - V(\pi^{\mathrm{TS}}, T, \mathbf{y}) \leq 2\sqrt{L}\left[\frac{1}{\sqrt{\nu}} + \sqrt{2\log T}\left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT}\right)\right]. \tag{262}$$

(2) Suboptimality analysis for IRS.FH **(61).** As discussed in Remark 11, for IRS.FH, we have

$$Q_t^{z,\mathrm{in}}(a_t^{z,*}) - Q_t^{z,\mathrm{in}}(a) = \hat{\mu}_t^\pi(a_t^{z,*}, T - t) - \hat{\mu}_t^\pi(a, T - t). \tag{263}$$

We construct the confidence intervals as follows:

$$U_t(a) \triangleq \hat{\mu}_t^\pi(a, 0) + \sqrt{2\log T} \times \Delta_t^\pi(a, T - t), \quad L_t(a) \triangleq \hat{\mu}_t^\pi(a, 0) + \sqrt{2\log T} \times \Delta_t^\pi(a, T - t). \tag{264}$$

Given that $\hat{\mu}_t^\pi(a, T - t)$ is $\Delta_t^\pi(a, T - t)$-sub-Gaussian conditioned on $\mathcal{F}_{t-1}$, by Lemma 4, we have

$$\mathbb{E}\left[\left(\hat{\mu}_t^\pi(a, T - t) - U_t(a)\right)^+ \middle| \mathcal{F}_{t-1}\right] \leq \frac{\Delta_t^\pi(a, T - t)}{\sqrt{2\log T}} e^{-\frac{2\log T}{2}} \leq \frac{\Delta_t^\pi(a, \infty)}{\sqrt{2\log T}} e^{-\frac{2\log T}{2}} \leq \frac{\sqrt{L/\nu}}{T}. \quad (265)$$

Symmetrically, we have $\mathbb{E}\left[\left(L_t(a) - \hat{\mu}_t^\pi(a, T - t)\right)^+ \middle| \mathcal{F}_{t-1}\right] \leq \frac{\sqrt{L/\nu}}{T}$.

On the other hand, since $N_{t-1}(a) \leq t$ in any case, we have

$$\frac{1}{\nu + N_{t-1}^\pi(a)} \times \frac{T - t}{\nu + N_{t-1}^\pi(a) + T - t} = \frac{1}{\nu + N_{t-1}^\pi(a)} \times \left(1 - \frac{\nu + N_{t-1}^\pi(a)}{\nu + N_{t-1}^\pi(a) + T - t}\right) \quad (266)$$

$$= \frac{1}{\nu + N_{t-1}^\pi(a)} - \frac{1}{\nu + N_{t-1}^\pi(a) + T - t} \quad (267)$$

$$\leq \frac{1}{\nu + N_{t-1}^\pi(a)} - \frac{1}{\nu + T}. \quad (268)$$

Consequently,

$$\sum_{t=1}^T \sqrt{\frac{1}{\nu + N_{t-1}^\pi(a)} - \frac{1}{\nu + T}} = \sum_{a \in \mathcal{A}} \sum_{n=0}^{N_T^\pi(a)-1} \sqrt{\frac{1}{\nu + n} - \frac{1}{\nu + T}} \quad (269)$$

$$= \sum_{a \in \mathcal{A}} \left(\sqrt{\frac{1}{\nu} - \frac{1}{\nu + T}} + \sum_{n=1}^{N_T^\pi(a)-1} \sqrt{\frac{1}{\nu + n} - \frac{1}{\nu + T}}\right) \quad (270)$$

$$\leq \frac{K}{\sqrt{\nu}} + \sum_{a \in \mathcal{A}} \sum_{n=1}^{N_T^\pi(a)-1} \sqrt{\frac{1}{n} - \frac{1}{T}} \quad (271)$$

$$\overset{(i)}{\leq} \frac{K}{\sqrt{\nu}} + \sum_{a \in \mathcal{A}} \sum_{n=1}^{N_T^\pi(a)-1} \left(\frac{1}{\sqrt{n}} - \frac{\sqrt{n}}{2T}\right) \quad (272)$$

$$\leq \frac{K}{\sqrt{\nu}} + \sum_{a \in \mathcal{A}} \int_0^{N_T^\pi(a)} \left(\frac{1}{\sqrt{x}} - \frac{\sqrt{x}}{2T}\right) dx \quad (273)$$

$$= \frac{K}{\sqrt{\nu}} + \sum_{a \in \mathcal{A}} \left(2\sqrt{N_T^\pi(a)} - \frac{(N_T^\pi(a))^{3/2}}{2T}\right) \quad (274)$$

$$\overset{(ii)}{\leq} \frac{K}{\sqrt{\nu}} + 2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}, \quad (275)$$

where we have utilized that (i) the concavity of $\sqrt{\cdot}$, and (ii) $\min\{\sum_{a=1}^K n_a^{3/2}; \sum_{a=1}^K n_a = T\} = \sum_{a=1}^K (T/K)^{3/2} = \sqrt{T^3/K}$.

Combining all these results, we conclude that

$$W^{\text{Irs.FH}}(T, \mathbf{y}) - V(\pi^{\text{Irs.FH}}, T, \mathbf{y}) \leq 2\sqrt{\frac{L}{\nu}} + 2\sqrt{2 \log T} \sum_{t=1}^{T} \Delta_t^{\pi}(A_t^{\pi}, T - t) \tag{276}$$

$$\leq 2\sqrt{L} \left[\frac{1}{\sqrt{\nu}} + \sqrt{2 \log T} \left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT} - \frac{1}{3}\sqrt{T/K} \right) \right]. \tag{277}$$

(3) Suboptimality analysis for IRS.V-ZERO **(62).** As discussed in Remark 11, for IRS.FH, we have

$$Q_t^{z,\text{in}}(a_t^{z,*}) - Q_t^{z,\text{in}}(a) = \max_{0 \leq n \leq T-t} \left\{ \hat{\mu}_t^{\pi}(a_t^{z,*}, n) \right\} - \hat{\mu}_t^{\pi}(a, 0). \tag{278}$$

We construct the confidence intervals as follows:

$$U_t(a) \triangleq \hat{\mu}_t^{\pi}(a, 0) + \sqrt{2 \log T} \times \Delta_t^{\pi}(a, T - t), \quad L_t(a) \triangleq \hat{\mu}_t^{\pi}(a, 0). \tag{279}$$

By Lemma 6, we have

$$\mathbb{E}\left[\left(\max_{0 \leq n \leq T-t} \hat{\mu}_t^{\pi}(a, n) - U_t(a) \right)^+ \Big| \mathcal{F}_{t-1} \right] \leq \frac{\Delta_t^{\pi}(a, T - t)}{\sqrt{2 \log T}} e^{-\frac{2 \log T}{2}} \leq \frac{\sqrt{L/\nu}}{T}, \tag{280}$$

where

$$\mathbb{E}\left[\hat{\mu}_t^{\pi}(a, 0) - L_t(a) | \mathcal{F}_{t-1} \right] = 0. \tag{281}$$

The rest of the proof is almost identical to the case of IRS.FH:

$$W^{\text{Irs.V-Zero}}(T, \mathbf{y}) - V(\pi^{\text{Irs.V-Zero}}, T, \mathbf{y}) \leq \sqrt{\frac{L}{\nu}} + \sum_{t=1}^{T} \mathbb{E}\left[U_t(A_t^{\pi}) - L_t(A_t^{\pi}) \right]. \tag{282}$$

$$= \sqrt{\frac{L}{\nu}} + \sqrt{2 \log T} \sum_{t=1}^{T} \Delta_t^{\pi}(A_t^{\pi}, T - t) \tag{283}$$

$$\leq \sqrt{L} \left[\frac{1}{\sqrt{\nu}} + \sqrt{2 \log T} \left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT} - \frac{1}{3}\sqrt{T/K} \right) \right]. \tag{284}$$