Thompson Sampling with Information Relaxation Penalties*

Seungki Min
Graduate School of Business
Columbia University
smin20@gsb.columbia.edu

Costis Maglaras
Graduate School of Business
Columbia University
c.maglaras@gsb.columbia.edu

Ciamac C. Moallemi
Graduate School of Business
Columbia University
ciamac@gsb.columbia.edu

Current Revision: May 2020

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.

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.

*The authors wish to thank Daniel Russo, Martin Haugh, David Brown, and Jim Smith for helpful discussions.

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 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. In 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., Agrawal and Goyal, 2013; Kaufmann et al., 2012b; 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, a lower bound on the performance of the optimal policy can be found. 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 Multi-armed Bandit

2.1. Problem

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

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

For brevity, we let $\boldsymbol{\theta} \triangleq (\theta_1, \ldots, \theta_K) \in \Theta$ and $\mathbf{y} \triangleq (y_1, \ldots, y_K) \in \mathcal{Y}$ be the vector of parameters and beliefs across arms, respectively.

We 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,

$$\omega \triangleq (\boldsymbol{\theta}, (R_{a,n})_{a \in \mathcal{A}, n \in \mathbb{N}}) \quad \sim \quad \mathcal{I}(\mathbf{y}), \tag{2}$$

that encodes all uncertainties in the environment that the DM encounters. We denote the associated σ-field by $\sigma(\omega)$ and the prior distribution by $\mathcal{I}(\mathbf{y})$.

Mean reward. For each arm a, we define the true mean reward μ_a and its Bayesian estimate $\hat{\mu}_{a,n}$, expressed as a function of the parameter and the outcome, respectively:

$$\mu_a(\theta_a) \triangleq \mathbb{E}\left[R_{a,n} | \theta_a\right], \quad \hat{\mu}_{a,n}(\omega; y_a) \triangleq \mathbb{E}\left[\mu_a(\theta_a) | R_{a,1}, \ldots, R_{a,n}\right]. \tag{3}$$

The true mean reward μ_a is the quantity that the DM is hoping to learn from the noisy observations, and the estimate $\hat{\mu}_{a,n}$ represents the expected mean reward that a Bayesian learner can infer from the first n reward realizations of arm a. As the rewards $R_{a,1}, R_{a,2}, \ldots, R_{a,n}$ are sequentially revealed whenever the DM plays the arm a, her expectation on the mean reward evolves through $\hat{\mu}_{a,0}, \hat{\mu}_{a,1}, \ldots, \hat{\mu}_{a,n}$.

We additionally introduce the ex-ante mean reward $\bar{\mu}_a$ as a function of prior belief y_a,

$$\bar{\mu}_a(y_a) \triangleq \mathbb{E}_{\theta_a \sim \mathcal{P}_a(y_a)} \left[\mu_a(\theta_a) \right], \tag{4}$$

which represents the unconditional expected value of the true mean reward. Throughout the paper, we assume that the rewards are absolutely integrable over the prior distribution for all possible belief states, i.e., $\mathbb{E}\left[|R_{a,n}|\right] < \infty$, or, more explicitly,

$$\mathbb{E}_{r \sim \mathcal{R}_a(\mathcal{P}_a(y_a))} \left[|r| \right] < \infty, \quad \forall a \in \mathcal{A}, \tag{5}$$

where $\mathcal{R}_a(\mathcal{P}_a(y_a))$ denotes the (unconditional) distribution of reward $R_{a,n}$ as a doubly stochastic random variable.

Remark 1. *$\hat{\mu}_{a,0}(\omega; y_a) = \bar{\mu}_a(y_a)$, and $\lim_{n\to\infty} \hat{\mu}_{a,n}(\omega; y_a) = \mu_a(\theta_a)$ almost surely.*

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.,

$$r_t(\mathbf{a}_{1:t}, \omega) \triangleq R_{a_t, n_t(\mathbf{a}_{1:t}, a_t)}, \tag{6}$$

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).

While the DM is aware of the context of the game (time horizon T and prior belief $\mathbf{y}$), the rewards are sequentially revealed in response to the DM's actions. More formally, the *natural filtration* $\mathcal{F}_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y})$ incorporates the past actions and the corresponding observations revealed up to time t (inclusive),

$$\mathcal{F}_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}) \triangleq \sigma\left(T, \mathbf{y}, \{a_s, r_s(\mathbf{a}_{1:s}, \omega)\}_{s \in [t]}\right). \tag{7}$$

Let $\mathbf{a}^{\pi}_{1:t}$ be the action sequence taken by a policy π. A policy π is called *non-anticipating* if its every action a^{π}_t is measurable with respect to $\mathcal{F}_{t-1}$, and we denote by $\Pi_{\mathbb{F}}$ the set of all non-anticipating policies, including randomized ones. The (Bayesian) *performance* of a policy π is defined as the expected total reward over the randomness associated with the outcome, i.e.,

$$V(\pi, T, \mathbf{y}) \triangleq \mathbb{E}_{\omega \sim \mathcal{I}(\mathbf{y})} \left[\sum_{t=1}^{T} r_t(\mathbf{a}^{\pi}_{1:t}, \omega) \right]. \tag{8}$$

Bayesian update. In order to explicitly describe the evolution of the DM's belief, we introduce a *Bayesian update function* $\mathcal{U}_a : \mathcal{Y}_a \times \mathbb{R} \mapsto \mathcal{Y}_a$ so that after observing a reward realization r from an arm a, the belief associated with arm a is updated from y_a to $\mathcal{U}_a(y_a, r)$ according to Bayes' rule. We will often use $\mathcal{U} : \mathcal{Y} \times \mathcal{A} \times \mathbb{R} \mapsto \mathcal{Y}$ to describe the updating of the belief vector $\mathbf{y}$; i.e., after observing a reward realization r from an arm a, the belief vector is updated from $\mathbf{y}$ to $\mathcal{U}(\mathbf{y}, a, r)$

where only the a^{th} component is updated.

The future belief states can be represented using this Bayesian update function: given an outcome ω and an action sequence $\mathbf{a}_{1:t}$, the posterior belief at time t can be recursively expressed as

$$\mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y}) \triangleq \mathcal{U}\left(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y}), a_t, r_t(\mathbf{a}_{1:t}, \omega)\right), \tag{9}$$

with $\mathbf{y}_0(\emptyset, \omega; \mathbf{y}) \triangleq \mathbf{y}$.

Beta-Bernoulli and Gaussian MABs. We will consider a Beta-Bernoulli MAB and a Gaussian MAB as canonical examples for illustration of our framework as well as for numerical experiments. In the Beta-Bernoulli MAB, the rewards of an arm a are Bernoulli random variables with an unknown success probability θ_a, where the prior distribution of θ_a is $\text{Beta}(\alpha_a, \beta_a)$. In the Gaussian MAB, the rewards are normally distributed with an unknown mean θ_a and a known noise variance σ_a^2, where the prior distribution of θ_a is also a Gaussian distribution with mean m_a and variance ν_a^2. Table 1 summarizes the previously defined notations.

	Beta-Bernoulli MAB	**Gaussian MAB**
Reward distr. $\mathcal{R}_a$	$R_{a,n} \sim \text{Bernoulli}(\theta_a)$	$R_{a,n} \sim \mathcal{N}(\theta_a, \sigma_a^2)$
Prior distr. $\mathcal{P}_a$	$\theta_a \sim \text{Beta}(\alpha_a, \beta_a)$	$\theta_a \sim \mathcal{N}(m_a, \nu_a^2)$
Mean reward μ_a	$\mu_a(\theta_a) = \theta_a$	$\mu_a(\theta_a) = \theta_a$
Expected mean $\bar{\mu}_a$	$\bar{\mu}_a(\alpha_a, \beta_a) = \frac{\alpha_a}{\alpha_a + \beta_a}$	$\bar{\mu}_a(m_a, \nu_a^2) = m_a$
Bayesian update $\mathcal{U}_a$	$\mathcal{U}_a\left((\alpha_a, \beta_a), r\right) = (\alpha_a + r, \beta_a + 1 - r)$	$\mathcal{U}_a\left((m_a, \nu_a^2), r\right) = \left(\frac{m_a \cdot \nu_a^{-2} + r \cdot \sigma_a^{-2}}{\nu_a^{-2} + \sigma_a^{-2}}, \frac{1}{\nu_a^{-2} + \sigma_a^{-2}}\right)$

Table 1: Description of a Beta-Bernoulli MAB and a Gaussian MAB.

In these MAB problems, the belief state of each arm is represented by a two-dimensional vector, $y_a = (\alpha_a, \beta_a)$ and $y_a = (m_a, \nu_a^2)$, respectively, which are the sufficient statistics for Beta distribution and Gaussian distribution. More generally, when the reward distribution $\mathcal{R}_a$ is a member of an exponential family, its conjugate prior $\mathcal{P}_a$ can be represented by a low-dimensional vector y_a (the sufficient statistics for $\mathcal{P}_a$), and the Bayesian update function $\mathcal{U}_a$ admits a simple closed form. In the other cases, when the reward distributions do not belong to an exponential family, the belief y_a may be an infinite-dimensional vector that represents the entire probability density of prior/posterior distribution, and there may not be a closed-form expression for $\mathcal{U}_a$. We note that the theoretical foundations of our framework do not rely on a parsimonious representation of the belief state nor a closed-form expression for the belief update function, which may concern in practice.

2.2. Bayesian Optimal Policy

In a Bayesian framework, the MAB problem has a recursive structure. Given 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 a problem 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:

$$Q^*(T, \mathbf{y}, a) \triangleq \mathbb{E}_{r \sim \mathcal{R}_a(\mathcal{P}_a(y_a))} \left[r + V^*(T-1, \mathcal{U}(\mathbf{y}, a, r)) \right], \tag{10}$$

$$V^*(T, \mathbf{y}) \triangleq \max_{a \in \mathcal{A}} Q^*(T, \mathbf{y}, a), \tag{11}$$

with $V^*(0, \mathbf{y}) \triangleq 0$ for all $\mathbf{y} \in \mathcal{Y}$.

While Bellman equations are, in general, intractable to solve and directly apply, they offer a characterization of the *Bayesian optimal policy*[1] (OPT) and the best achievable performance V^*. 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 $a^* = \operatorname{argmax}_a Q^*(T, \mathbf{y}, a)$, 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:

$$V^*(T, \mathbf{y}) = V(\text{OPT}, T, \mathbf{y}) = \sup_{\pi \in \Pi_{\mathbb{F}}} V(\pi, T, \mathbf{y}). \tag{12}$$

Algorithm 1: Bayesian optimal policy (OPT)

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

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

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

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

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

  Pull $a_t \leftarrow$ `OPT`$(T - t + 1, \mathbf{y}_{t-1})$

  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

2.3. Thompson Sampling

Thompson sampling (TS) is a simple heuristic that makes decisions based on random sampling. When the remaining time is T and the belief is $\mathbf{y}$, it samples the parameters $\tilde{\boldsymbol{\theta}}$ from the prior[2] distribution at that moment, $\mathcal{P}(\mathbf{y})$, and pulls the arm that is believed to be best given the sampled parameters $\tilde{\boldsymbol{\theta}}$ which is $\operatorname{argmax}_a \mu_a(\tilde{\theta}_a)$. Like OPT, it repeats this procedure at every decision epoch

[1] In the frequentist setting, the optimal policy may not be well-defined since there is no performance measure consistent across different sets of parameters.

[2] Conventionally, the term "posterior distribution" is used to describe where TS samples the parameters from. We explicitly use "prior distribution" since it is the prior belief at the moment of decision making. For example, at time $t = 1$, the parameters are apparently sampled from a prior, not a posterior, distribution. After observing a reward realization, we will have a posterior but it becomes a prior at the next decision epoch.

while updating the belief $\mathbf{y}$ whenever a reward realization is observed.

Algorithm 2: Single decision making under Thompson sampling (TS)

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

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

return $\text{argmax}_a\{\mu_a(\tilde{\theta}_a)\}$

Note that TS does not take into account the time horizon T 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 propose a general framework, which we refer to as *information relaxation sampling* (IRS), that takes as an input a "penalty function" $z_t(\cdot)$, and produces as outputs a policy π^z and an associated performance bound W^z.

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). Under this relaxation, the DM will be allowed to first observe all future outcomes in advance, and then pick an action (i.e., a_t^π is $\sigma(\omega)$-measurable). To compensate for this perfect information 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 a particular instance specified by ω, T and $\mathbf{y}$. The clairvoyant DM can find the best action sequence that is optimal for a particular outcome ω in the presence of penalties z_t, by solving the following (deterministic) optimization problem, referred as the *inner problem*:

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

Definition 1 (Dual feasibility). *Given T and $\mathbf{y}$, a penalty function z_t is dual feasible if it is ex-ante zero mean, i.e.,*

$$\mathbb{E}\left[z_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}) \,|\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega; T, \mathbf{y})\right] = 0, \quad \forall \mathbf{a}_{1:t} \in \mathcal{A}^t, \quad \forall t \in [T]. \tag{13}$$

To clarify the notion of conditional expectation, 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.[3] This dual

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

feasibility condition requires that the DM who makes decisions on the natural filtration will receive zero penalties in expectation.

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

$$W^z(T, \mathbf{y}) \triangleq \mathbb{E}_{\omega \sim \mathcal{I}(\mathbf{y})} \left[\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}) \right\} \right]. \tag{14}$$

We can obtain this value 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 $\omega^{(s)}$ 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). *Given T and $\mathbf{y}$, if the penalty function z_t is dual feasible, W^z is an upper bound on the optimal value V^*:*

$$\textit{(Weak duality)} \qquad W^z(T, \mathbf{y}) \geq V^*(T, \mathbf{y}). \tag{15}$$

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

$$\textit{(Strong duality)} \qquad W^{\text{ideal}}(T, \mathbf{y}) = V^*(T, \mathbf{y}). \tag{16}$$

The ideal penalty function z_t^{ideal} has the following functional form:[4]

$$\begin{aligned} z_t^{\text{ideal}}(\mathbf{a}_{1:t}, \omega) \triangleq\ & r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right] \\ & + V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) - \mathbb{E}\left[V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) | \, \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right]. \end{aligned} \tag{17}$$

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

The ideal penalty is, of course, intractable, but its structure highlights what a good penalty may look like. It implies that there are two sources of additional profit: in DP terminology, one

[4] Throughout the paper, we will often omit T and $\mathbf{y}$ from the expressions for their brief representation. In (17), $z_t^{\text{ideal}}(\mathbf{a}_{1:t}, \omega)$, $\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega)$, and $\mathbf{y}_t(\mathbf{a}_{1:t}, \omega)$ still have a dependency on T and $\mathbf{y}$.

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 $\mathcal{F}_{t-1} \subseteq \mathcal{G}_{t-1} \subseteq \sigma(\omega)$) 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 $z_t^{\mathbb{G}} + z_t^{\sigma(\omega)\setminus\mathbb{G}}$ if the additional term $z_t^{\sigma(\omega)\setminus\mathbb{G}}$ 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 $\mathcal{F}_{t-1}$. As a consequence, our ideal penalty function (17) 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. Given a penalty function z_t, we characterize a randomized and non-anticipating IRS policy π^z as follows. The policy π^z specifies "which arm to pull when the remaining time is T and current belief is $\mathbf{y}$." Given T and $\mathbf{y}$, it (i) first samples the outcome $\tilde{\omega}$ from $\mathcal{I}(\mathbf{y})$ randomly, (ii) solves the inner problem to find a best action sequence $\tilde{\mathbf{a}}_{1:T}^*$ with respect to $\tilde{\omega}$ in the presence of penalties z_t, and (iii) takes the first action $\tilde{a}_1^*$ that the clairvoyant optimal solution $\tilde{\mathbf{a}}_{1:T}^*$ suggests. Analogous to TS and Opt, **it repeats steps (i)–(iii) at every decision epoch**, while updating

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

the remaining time T and belief $\mathbf{y}$ upon each reward realization.

Algorithm 3: Information relaxation sampling (IRS) policy

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

Sample $\tilde{\omega} \sim \mathcal{I}(\mathbf{y})$ (equivalently, $\tilde{\theta}_a \sim \mathcal{P}_a(y_a)$ and $\tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}_a)$, $\forall a \in \mathcal{A}$, $\forall n \in [T]$)

Find the best action sequence with respect to $\tilde{\omega}$ under penalties z_t:
$\tilde{\mathbf{a}}^*_{1:T} \leftarrow \operatorname{argmax}_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \tilde{\omega}) - z_t(\mathbf{a}_{1:t}, \tilde{\omega}; T, \mathbf{y}) \right\}$

return $\tilde{a}^*_1$

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

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

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

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

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

end

In step (i), sampling $\tilde{\omega} \sim \mathcal{I}(\mathbf{y})$ means sampling the parameters $\tilde{\theta}_a \sim \mathcal{P}_a(y_a)$ and then generating the future rewards $\tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}_a)$ for all $n \in [T]$ and all $a \in \mathcal{A}$. It is equivalent to simulating a plausible future scenario based on the current belief $\mathbf{y}$, with the policy π^z taking the best action optimized to this synthesized future. Note that 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 process, where the time horizon of the inner problem at each decision epoch decreases by one, from T to 1, as described in IRS-OUTER procedure in Algorithm 3.

Remark 2. *The ideal penalty yields the Bayesian optimal policy, i.e.,* $V(\pi^{\text{ideal}}, T, \mathbf{y}) = V^*(T, \mathbf{y})$.

Choice of penalty functions. IRS policies include TS and OPT as two extreme cases. We propose a set of penalty functions spanning the two. Deferring the detailed explanations to §3.1–§3.4, we briefly list the penalty functions:

$$z_t^{\text{TS}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\boldsymbol{\theta}\right], \tag{18}$$

$$z_t^{\text{IRS.FH}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\hat{\boldsymbol{\mu}}_{1:K,T-1}(\omega)\right], \tag{19}$$

$$z_t^{\text{IRS.V-ZERO}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right], \tag{20}$$

$$\begin{aligned} z_t^{\text{IRS.V-EMAX}}(\mathbf{a}_{1:t}, \omega) \triangleq\ & r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] \\ & + W^{\text{TS}}\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) - \mathbb{E}\left[W^{\text{TS}}\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) \middle| \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right], \end{aligned} \tag{21}$$

where $\hat{\boldsymbol{\mu}}_{1:K,T-1}(\omega) \triangleq (\hat{\mu}_{1,T-1}(\omega), \ldots, \hat{\mu}_{K,T-1}(\omega))$. To better the understanding of conditional expectations, we provide an identity as an example: $\mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] = \mathbb{E}\left[\mu_{a_t}(\theta_{a_t}) \,|\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] = \mathbb{E}\left[\mu_{a_t}(\theta_{a_t}) \middle| R_{a_t,1}, \ldots, R_{a_t,n_{t-1}(\mathbf{a}_{1:t-1},a_t)}\right] = \hat{\mu}_{a_t,n_{t-1}(\mathbf{a}_{1:t-1},a_t)}(\omega) = \bar{\mu}_{a_t}\left(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right)$ and all these

expressions represent the mean reward that the DM expects to get from arm a_t before making a decision at time t.

Remark 3. *All penalty functions* (17)–(21) *are dual feasible.*

Penalty function	Policy	Performance bound	Inner problem	Run time
z_t^{TS}	TS	W^{TS}	Find a best arm given parameters.	$O(K)$
$z_t^{\textsc{Irs.FH}}$	$\pi^{\textsc{Irs.FH}}$	$W^{\textsc{Irs.FH}}$	Find a best arm given finite observations.	$O(K)$ or $O(KT)$
$z_t^{\textsc{Irs.V-Zero}}$	$\pi^{\textsc{Irs.V-Zero}}$	$W^{\textsc{Irs.V-Zero}}$	Find an optimal allocation of T pulls.	$O(KT^2)$
$z_t^{\textsc{Irs.V-EMax}}$	$\pi^{\textsc{Irs.V-EMax}}$	$W^{\textsc{Irs.V-EMax}}$	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 (17)–(21). 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 of performance bound W^z or to make a single decision under policy π^z.

Table 2 summarizes the algorithms investigated in this paper. As we sequentially increase the computational complexity of a penalty function, from z^{TS} to z^{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 §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 of a performance bound W^z or to make a single decision under policy π^z. In the 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

$$\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t(\mathbf{a}_{1:t}, \omega) \right\} = \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} \mu_{a_t}(\theta_{a_t}) \right\} = T \times \max_{a \in \mathcal{A}} \mu_a(\theta_a). \tag{22}$$

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

bound W^{TS} reduces to the conventional performance benchmark,

$$W^{\mathrm{TS}}(T, \mathbf{y}) = \mathbb{E}\left[T \times \max_{a \in \mathcal{A}} \mu_a(\theta_a)\right], \tag{23}$$

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

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 $\tilde{a}^{\mathrm{TS}} = \mathrm{argmax}_a\, \mu_a(\tilde{\theta}_a)$, 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.

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, 2017). The Bayesian regret of a policy π is defined as*

$$\mathit{BayesRegret}(\pi, T, \mathbf{y}) \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \max_a \mu_a(\theta_a) - \mu_{a_t^\pi}(\theta_{a_t^\pi})\right] = W^{\mathrm{TS}}(T, \mathbf{y}) - V(\pi, T, \mathbf{y}), \tag{24}$$

which quantifies the suboptimality of the policy π.

3.2. IRS.FH

Recall that $\hat{\mu}_{a,T-1}(\omega)$ is the mean reward estimate of an arm a that the DM can infer from $T-1$ reward realizations $R_{a,1}, \ldots, R_{a,T-1}$:

$$\hat{\mu}_{a,T-1}(\omega) \triangleq \mathbb{E}\left[\mu_a(\theta_a) \,|R_{a,1}, \ldots, R_{a,T-1}\right]. \tag{25}$$

Given (19), the optimal solution to the inner problem $(*)$ is to always pull the arm $a^{\textsc{Irs.FH}} = \mathrm{argmax}_a\, \hat{\mu}_{a,T-1}(\omega)$, i.e.,

$$\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{\sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t^{\textsc{Irs.FH}}(\mathbf{a}_{1:t}, \omega)\right\} = \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{\sum_{t=1}^{T} \hat{\mu}_{a_t,T-1}(\omega)\right\} = T \times \max_{a \in \mathcal{A}} \hat{\mu}_{a,T-1}(\omega). \tag{26}$$

This inner problem yields the performance bound $W^{\textsc{Irs.FH}}$ such that

$$W^{\textsc{Irs.FH}}(T, \mathbf{y}) = \mathbb{E}\left[T \times \max_{a \in \mathcal{A}} \hat{\mu}_{a,T-1}(\omega)\right], \tag{27}$$

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

Algorithm 4: Single decision making under the IRS.FH policy

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

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

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

IRS.FH (FH stands for finite horizon) is almost identical to TS except that $\mu_a(\theta_a)$ is replaced with $\hat{\mu}_{a,T-1}(\omega)$. The main motivation is as follows: from the DM's perspective, $\hat{\mu}_{a,T-1}(\omega)$ is less informative than $\mu_a(\theta_a)$ since she will never be able to learn $\mu_a(\theta_a)$ perfectly within a finite horizon. In terms of mean reward estimation, knowing the parameters is equivalent to having the infinite number of observations. The inner problem of TS requires the DM to "identify the best arm based on an infinite number of samples," whereas that of IRS.FH requires her 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, it requires her to be more realistic, that is, to consider the uncertainties more precisely.

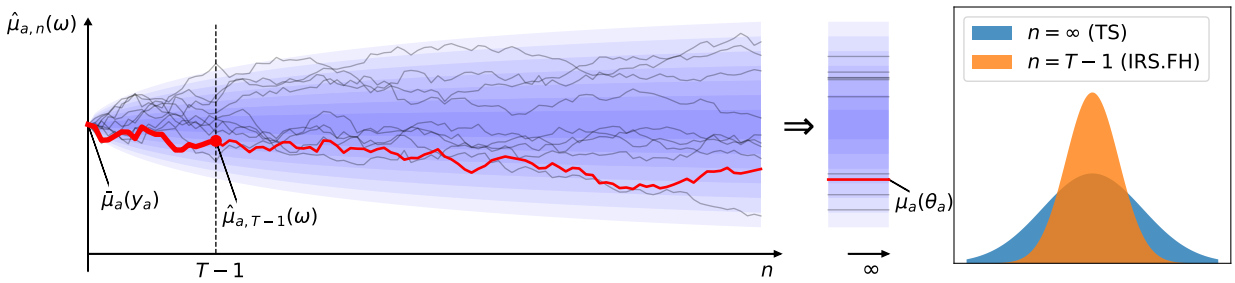

Figure 1: (Left) Sample paths of Bayesian estimate on mean reward of an arm a, $\{\hat{\mu}_{a,n}(\omega)\}_{n\geq 0}$. (Right) The distribution of the finite-horizon estimate $\hat{\mu}_{a,T-1}$ that is more concentrated than the distribution of its limit μ_n $(= \lim_{n\to\infty} \hat{\mu}_{a,n})$, while both have the same mean $\bar{\mu}_a$ $(= \hat{\mu}_{a,0} = \mathbb{E}[\mu_a])$.

To sharpen our comparison between IRS.FH and TS, let us compare the variability of $\hat{\mu}_{a,T-1}(\omega)$ and $\mu_a(\theta_a)$ focusing on the randomness of the outcome ω. As depicted in Figure 1, we observe that the distribution of $\hat{\mu}_{a,T-1}(\omega)$ is more concentrated than that of $\mu_a(\theta_a)$, while both have the same mean value $\bar{\mu}_a(y_a)$. By Jensen's inequality, we have $W^{\text{IRS.FH}} = \mathbb{E}[T \times \max_a \hat{\mu}_{a,T-1}(\omega)] \leq W^{\text{TS}} = \mathbb{E}[T \times \max_a \mu_a(\theta_a)]$ 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., $\tilde{a}^{\text{IRS.FH}} = \operatorname{argmax}_a \hat{\mu}_{a,T-1}(\tilde{\omega})$ (resp., $\tilde{a}^{\text{TS}} = \operatorname{argmax}_a \mu_a(\tilde{\theta}_a)$). 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.

Sampling $\hat{\mu}_{a,T-1}(\tilde{\omega})$ at once. In order to obtain $\hat{\mu}_{a,T-1}(\tilde{\omega})$'s for a synthesized outcome $\tilde{\omega}$, one may apply Bayes' rule sequentially for each reward realization, which will take $O(KT)$ computations in total. We illustrate that it can be done in $O(K)$ computations when the belief can be updated in a batch by the use of sufficient statistics.

To illustrate this in detail, consider the Beta-Bernoulli MAB and the Gaussian MAB introduced in Table 1. In either case, $\hat{\mu}_{a,T-1}(\tilde{\omega})$ admits the following expression:

$$\text{(Beta-Bernoulli)} \qquad \hat{\mu}_{a,T-1}(\tilde{\omega}; \alpha_a, \beta_a) = \frac{(\alpha_a + \beta_a) \times \frac{\alpha_a}{\alpha_a+\beta_a} + (T-1) \times \left(\frac{1}{T-1} \sum_{n=1}^{T-1} \tilde{R}_{a,n}\right)}{(\alpha_a + \beta_a) + (T-1)}, \tag{28}$$

$$\text{(Gaussian)} \qquad \hat{\mu}_{a,T-1}(\tilde{\omega}; m_a, \nu_a^2) = \frac{\nu_a^{-2} \times m_a + (T-1) \cdot \sigma_a^{-2} \times \left(\frac{1}{T-1} \sum_{n=1}^{T-1} \tilde{R}_{a,n}\right)}{\nu_a^{-2} + (T-1) \cdot \sigma_a^{-2}}. \tag{29}$$

Note that $\hat{\mu}_{a,T-1}(\tilde{\omega})$ can be represented as a convex combination of the current estimate $\bar{\mu}_a(y_a)$ and the sample mean $\frac{1}{T-1} \sum_{n=1}^{T-1} \tilde{R}_{a,n}$. The sufficient statistic $\sum_{n=1}^{T-1} \tilde{R}_{a,n}$ is Binomial$(T-1, \tilde{\theta}_a)$ for the Beta-Bernoulli case, and $\mathcal{N}((T-1) \cdot \tilde{\theta}_a, (T-1) \cdot \sigma_a^2)$ for the Gaussian case. After sampling the parameter $\tilde{\theta}_a$, we can sample the sufficient statistic $\sum_{n=1}^{T-1} \tilde{R}_{a,n}$ directly from the known distribution, 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, similar in computational complexity to TS.

The Beta-Bernoulli and Gaussian cases above are special instances of the more general case of a natural exponential family. In these setting, the sufficient statistic over $T-1$ future reward observations of an arm follows a distribution that is also a natural exponential family. This distribution may be tractable to compute and thus may be sampled using $O(1)$ computations, independent of T.

3.3. IRS.V-Zero

IRS.V-Zero introduces a further complication such 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 (20), the DM at time t earns $\mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right]$, 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 n^{th} 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:

$$\hat{\mu}_{a,n}(\omega) \triangleq \mathbb{E}\left[\mu_a(\theta_a) \,|R_{a,1}, \ldots, R_{a,n}\right]. \tag{30}$$

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

$$r_t(\mathbf{a}_{1:t},\omega) - z_t^{\text{Irs.V-Zero}}(\mathbf{a}_{1:t},\omega) = \mathbb{E}\left[\mu_{a_t}(\theta_{a_t})\,|\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1},\omega)\right] = \hat{\mu}_{a_t,n_{t-1}(\mathbf{a}_{1:t-1},a_t)}(\omega), \tag{31}$$

where $n_{t-1}(\mathbf{a}_{1:t-1},a)$, defined in (6), 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 of 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,

$$\max_{\mathbf{a}_{1:T}\in\mathcal{A}^T}\left\{\sum_{t=1}^{T}\hat{\mu}_{a_t,n_{t-1}(\mathbf{a}_{1:t-1},a_t)}\right\} = \max_{\mathbf{a}_{1:T}\in\mathcal{A}^T}\left\{\sum_{a=1}^{K}\sum_{n=1}^{n_T(\mathbf{a}_{1:T},a)}\hat{\mu}_{a,n-1}\right\} = \max_{\mathbf{n}_{1:K}\in N_T}\left\{\sum_{a=1}^{K} S_{a,n_a}\right\}, \tag{32}$$

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 sup convolution K times. The detailed implementation is provided in §B.1.

Given an optimal allocation $(\tilde{n}_1^*, \tilde{n}_2^*, \ldots, \tilde{n}_K^*)$, the policy $\pi^{\text{Irs.V-Zero}}$ needs to select which arm to pull next. In principle, any arm a that was included in the solution of the inner problem, $\tilde{n}_a^* > 0$, would suffice, but we suggest a selection rule by which the arm that needs the most pulls is chosen, i.e., $\tilde{a}^{\text{Irs.V-Zero}} = \text{argmax}_a\, \tilde{n}_a^*$. It 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 $\mathbb{E}[r_t(a_t)|\boldsymbol{\theta}]$ and $\mathbb{E}[r_t(a_t)|\hat{\boldsymbol{\mu}}_{1:K,T-1}]$, respectively, which are the mean reward estimates that rely on the information not available at the moment; e.g., $\mathbb{E}[r_t(a_t)|\hat{\boldsymbol{\mu}}_{1:K,T-1}] = \hat{\mu}_{a_t,T-1}$ 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 $\mathbb{E}[r_t(a_t)|\mathcal{F}_{t-1}]$ which does not include any information that does not belong to $\mathcal{F}_{t-1}$. Irs.V-Zero reflects the fact that the n^{th} 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 $S_{a,n_a}(\omega) = \sum_{m=1}^{n_a}\hat{\mu}_{a,m-1}(\omega)$ is less volatile than its counterpart $n_a \times \hat{\mu}_{a,T-1}(\omega)$ since

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 to adjust the future rewards by conditioning (e.g., $\mathbb{E}[r_t(a_t)|\boldsymbol{\theta}]$, $\mathbb{E}[r_t(a_t)|\hat{\boldsymbol{\mu}}_{1:K,T-1}]$ and $\mathbb{E}[r_t(a_t)|\mathcal{F}_{t-1}]$). IRS.V-EMAX also focuses on the second component as well by charging 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} (17) is

$$\begin{aligned} z_t^{\text{ideal}}(\mathbf{a}_{1:t}, \omega) \triangleq\ & r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega)\,|\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] \\ & + V^*\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) - \mathbb{E}\left[V^*\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right)|\,\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right], \end{aligned} \tag{33}$$

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, her 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 $\mathcal{F}_{t-1}$ since the next observation r_t is unknown. In DP terms, the conditional expectation $\mathbb{E}\left[V^*\left(T-t, \mathbf{y}_t\right)|\,\mathcal{F}_{t-1}\right]$ captures the expected value of (random) next state given the current state. Accordingly, the gap between its realized value and its expected value, $V^*\left(T-t, \mathbf{y}_t\right) - \mathbb{E}\left[V^*\left(T-t, \mathbf{y}_t\right)|\,\mathcal{F}_{t-1}\right]$, measures the additional gain from knowing the next belief state $\mathbf{y}_t$. In addition to the term $r_t - \mathbb{E}\left[r_t\,|\mathcal{F}_{t-1}\right]$ $(= z_t^{\text{IRS.V-ZERO}})$, which measures the benefit from knowing which action will incur 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 rather tractable. The use of $W^{\text{TS}}(T, \mathbf{y}) \triangleq \mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{P}(\mathbf{y})}\left[T \times \max_a \mu_a(\theta_a)\right]$, introduced in (23), leads to a simple expression for its conditional expectation: since $\boldsymbol{\theta}|\mathcal{F}_{t-1}$ is distributed with $\mathcal{P}(\mathbf{y}_{t-1})$, we have

$$\mathbb{E}\left[W^{\text{TS}}\left(T-t, \mathbf{y}_t\right)\Big|\,\mathcal{F}_{t-1}\right] = (T-t) \times \mathbb{E}\left[\max_a \mu_a(\theta_a)\Big|\,\mathcal{F}_{t-1}\right] \tag{34}$$

$$= (T-t) \times \mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{P}(\mathbf{y}_{t-1})}\left[\max_a \mu_a(\theta_a)\right] \tag{35}$$

$$= W^{\text{TS}}\left(T-t, \mathbf{y}_{t-1}\right). \tag{36}$$

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

$$
\begin{aligned}
& r_t(\mathbf{a}_{1:t}) - z_t^{\text{IRS.V-EMAX}}(\mathbf{a}_{1:t}) && (37)\\
&= \hat{\mu}_{a_t, n_{t-1}(\mathbf{a}_{1:t-1}, a_t)} - W^{\text{TS}}\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t})\right) + W^{\text{TS}}\left(T-t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right) && (38)\\
&= \bar{\mu}_{a_t}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})) - W^{\text{TS}}\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t})\right) + W^{\text{TS}}\left(T-t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right), && (39)
\end{aligned}
$$

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 $\mathbf{a}_{1:t}$, that is, with $\mathbf{y}_t(\mathbf{n}_{1:K})$ instead of $\mathbf{y}_t(\mathbf{a}_{1:t})$.

Given the above observations, we can solve the inner problem within $O(KT^K)$ computations by dynamic programming. While deferring the detailed description of procedure to §B.2, we here briefly highlight the main idea of dynamic programming.

Let us consider a subproblem of $(*)$ given a pull allocation $\mathbf{n}_{1:K}$, in which we are constrained to play each arm $n_1, \ldots, n_K$ times and we are looking for the best sequence of pulls $\mathbf{a}_{1:t}$ that maximizes the total payoff with $t = \sum_a n_a$. The maximal value of this subproblem can be computed from the result of other K subproblems parameterized with $\mathbf{n}_{1:K} - \mathbf{e}_1, \mathbf{n}_{1:K} - \mathbf{e}_2, \ldots, \mathbf{n}_{1:K} - \mathbf{e}_K$, where $\mathbf{e}_a$ is a basis vector whose a^{th} component is one; i.e., having decided to play an arm a at time t, the previous belief state should be $\mathbf{y}_{t-1}(\mathbf{n}_{1:K} - \mathbf{e}_a)$ and we can earn at most the maximal value of the subproblem with $\mathbf{n}_{1:K} - \mathbf{e}_a$ plus the payoff of transition from $\mathbf{y}_{t-1}(\mathbf{n}_{1:K} - \mathbf{e}_a)$ to $\mathbf{y}_t(\mathbf{n}_{1:K})$, as represented in (39).

Each subproblem can be solved in $O(K)$ computations if the previous subproblems and the payoffs are pre-calculated. Note that the total number of possible future beliefs is $O(T^K)$, not $O(K^T)$. Therefore, the inner problem $(*)$ can be solved by sequentially solving $O(T^K)$ subproblems, which will require $O(c_W T^K + KT^K)$ operations, where c_W is the cost of numerically calculating $W^{\text{TS}}(T, \mathbf{y})$.

3.5. IRS.Index Policy

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

Roughly speaking, IRS.INDEX is an approximated version of the finite-horizon Gittins index (Kaufmann et al., 2012a), where our indices are computed with the 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 $\omega_a = \Big(\theta_a, (R_{a,n})_{n\in[T]}\Big)$, we can simulate the future belief trajectory $(y_{a,n})_{n\in\{0,\dots,T\}}$, where $y_{a,n}$ is the belief after n reward realizations are observed:

$$y_{a,0} \triangleq y_a, \quad y_{a,n} \triangleq \mathcal{U}_a(y_{a,n-1}, R_{a,n}), \quad \forall n \in [T]. \tag{40}$$

We adopt the penalty function $z_t^{\text{IRS.V-EMAX}}$ in which the true value function $V^*(T, y_a, \lambda)$ is approximated by $W^{\text{TS}}(T, y_a, \lambda) = \mathbb{E}_{\theta_a\sim\mathcal{P}_a(y_a)}\left[T \times \max(\mu_a(\theta_a), \lambda)\right]$. We define $\mathcal{A} \triangleq \{0, 1\}$ such that $a_t = 1$ if the stochastic arm at time t is played, and $a_t = 0$ if the outside option is chosen. The associated inner problem is

$$\text{maximize} \quad \sum_{t=1}^{T} \hat{\mu}_{a,n_t-1}(\omega_a) \cdot \mathbf{1}\{a_t = 1\} + \lambda \cdot \mathbf{1}\{a_t = 0\} - (T-t) \times \Big(\Gamma^{\lambda}_{n_t}(\omega_a) - \Gamma^{\lambda}_{n_{t-1}}(\omega_a)\Big) \tag{41}$$

$$\text{subject to} \quad n_t = \sum_{s=1}^{t} \mathbf{1}\{a_t = 1\}, \quad a_t \in \{0, 1\}, \quad \forall t = 1, \dots, T, \tag{42}$$

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

$$\Gamma^{\lambda}_n(\omega_a) \triangleq \mathbb{E}_{\theta_a\sim\mathcal{P}_a(y_{a,n})}\left[\max(\mu_a(\theta_a), \lambda)\right]. \tag{43}$$

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

$$\max_{0\le n\le T} \left\{ T \times \Gamma^{\lambda}_0(\omega_a) + (T-n) \times \left(\lambda - \min_{0\le i\le n} \Gamma^{\lambda}_i(\omega_a)\right) + \sum_{i=1}^{n} \Big(\hat{\mu}_{a,i-1}(\omega_a) - \Gamma^{\lambda}_{i-1}(\omega_a)\Big) \right\}, \tag{44}$$

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:

$$\varphi_a(\lambda, \omega_a) \triangleq \max_{1\le n\le T} \left\{ T \times \Gamma^{\lambda}_0 + (T-n) \times \left(\lambda - \min_{0\le i\le n} \Gamma^{\lambda}_i\right) + \sum_{i=1}^{n} \Big(\hat{\mu}_{a,i-1} - \Gamma^{\lambda}_{i-1}\Big) \right\} - T \times \lambda. \tag{45}$$

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 $n = 1, \ldots, T$ 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,

$$\mathbb{E}_{\theta \sim \mathrm{Beta}(\alpha,\beta)} \left[\max\left(\theta, \lambda\right) \right] = \lambda \times F_{\alpha,\beta}^{\mathrm{beta}}\left(\lambda\right) + \frac{\alpha}{\alpha+\beta} \times \left(1 - F_{\alpha+1,\beta}^{\mathrm{beta}}\left(\lambda\right)\right), \tag{46}$$

$$\mathbb{E}_{\theta \sim \mathcal{N}(m,\nu^2)} \left[\max\left(\theta, \lambda\right) \right] = m + (\lambda - m) \times \Phi\left(\nu^{-1}(\lambda - m)\right) + \nu \times \phi\left(\nu^{-1}(\lambda - m)\right), \tag{47}$$

where $F_{\alpha,\beta}^{\mathrm{beta}}(\cdot)$ represents the c.d.f. of $\mathrm{Beta}(\alpha, \beta)$ 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.

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 $n \in [T]$), and finds a threshold value on the deterministic outside option that makes the arm barely worth trying:

$$\lambda_a^*(\tilde{\omega}_a) \triangleq \sup\left\{\lambda \in \mathbb{R} \;;\; \varphi_a(\lambda, \tilde{\omega}_a) \ge 0\right\}. \tag{48}$$

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 the 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, $\tilde{a}^{\text{IRS.INDEX}} = \mathrm{argmax}_a\, \lambda_a^*(\tilde{\omega}_a)$.

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 single decision making 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, IRS.INDEX*, that is obtained by using

$$\varphi_a(\lambda, \omega_a) \triangleq \max_{1 \le n \le T} \left\{ \sum_{i=1}^{n} \left(\hat{\mu}_{a,i-1}(\omega_a) - \lambda - \left(\Gamma_i^{\lambda}(\omega_a) - \Gamma_0^{\lambda}(\omega_a) \right) \right) \right\}, \tag{49}$$

instead of (45). 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 $\pi^{\text{IRS.FH}}$, $\pi^{\text{IRS.V-ZERO}}$, $\pi^{\text{IRS.V-EMAX}}$, and $\pi^{\text{IRS.INDEX}}$ 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 $\pi^{\text{IRS.FH}}$'s action converges to that of Thompson sampling:*

$$\lim_{T\to\infty} \mathbb{P}\left[\text{IRS.FH}(T, \mathbf{y}) = a\right] = \mathbb{P}\left[\text{TS}(\mathbf{y}) = a\right], \quad \forall a \in \mathcal{A}. \tag{50}$$

Similarly, so does the distribution of the $\pi^{\text{IRS.V-ZERO}}$'s action:[6]

$$\lim_{T\to\infty} \mathbb{P}\left[\text{IRS.V-ZERO}(T, \mathbf{y}) = a\right] = \mathbb{P}\left[\text{TS}(\mathbf{y}) = a\right], \quad \forall a \in \mathcal{A}. \tag{51}$$

$\text{TS}(\mathbf{y})$, $\text{IRS.FH}(T, \mathbf{y})$ and $\text{IRS.V-ZERO}(T, \mathbf{y})$ denote the action taken by policies π^{TS}, $\pi^{\text{IRS.FH}}$, and $\pi^{\text{IRS.V-ZERO}}$, respectively, when the remaining time is T and the prior 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 proof of Proposition 1 (§D.2) is based on the fact that the Bayesian estimate will eventually converge to the true mean, 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 $\pi^{\text{IRS.FH}}$ and $\pi^{\text{IRS.V-ZERO}}$ behave like TS during the initial decision epochs, gradually shift toward the myopic scheme, and end up with the optimal de-

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

cision; 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*

$$W^{\mathrm{TS}}(T,\mathbf{y}) \geq W^{\text{IRS.FH}}(T,\mathbf{y}) \geq W^{\text{IRS.V-ZERO}}(T,\mathbf{y}), \tag{52}$$

and also

$$W^{\mathrm{TS}}(T,\mathbf{y}) \geq W^{\text{IRS.V-EMAX}}(T,\mathbf{y}). \tag{53}$$

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

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).

While the entire proof is provided in §D.3, we highlight here the main ideas. The first result (52) follows from the monotonicity of the information structure incorporated in each penalty function: TS, IRS.FH and IRS.V-ZERO allow the DM to earn $\mathbb{E}(r_t|\boldsymbol{\theta})$, $\mathbb{E}(r_t|\hat{\boldsymbol{\mu}}_{1:K,T-1})$ and $\mathbb{E}(r_t|\mathcal{F}_{t-1})$ at time t, respectively, while $\boldsymbol{\theta}$ is more informative than $\hat{\boldsymbol{\mu}}_{1:K,T-1}$, and $\hat{\boldsymbol{\mu}}_{1:K,T-1}$ is more informative than $\mathcal{F}_{t-1}$, as 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.[7] The second result (53) is proven based on Theorem 4 of Desai et al. (2012a), which says that if an approximate value function $\widehat{V}$ is a supersolution (Definition 2) to the Bellman equation and a penalty function $\hat{z}$ approximates the ideal penalty with $\widehat{V}$ in place of V^*, the resulting performance bound $W^{\hat{z}}$ is smaller than $\widehat{V}$. By showing that W^{TS} is a supersolution to (10), we prove that $W^{\text{IRS.V-EMAX}} \leq W^{\mathrm{TS}}$ 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. The 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

[7] We remark that $W^{\text{IRS.FH}} \geq W^{\text{IRS.V-ZERO}}$ is not an immediate consequence of the fact that $\sigma(\hat{\boldsymbol{\mu}}_{1:K,T-1})$ is a stronger filtration than $\mathcal{F}_{t-1}$. It relies on the particular reward structure of MAB problems: a reward at a certain moment is positively correlated with the later rewards (considering the randomness of the unknown parameters). Recall that IRS.V-ZERO penalizes for the additional gain from knowing the next reward but not for the additional gain from knowing the next belief state. When these two components are not aligned (this is not the case in MAB problems), penalizing only the first component may not yield a tighter bound. See §D.3.2 for a further discussion.

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, $W^{\mathrm{TS}}(T,\mathbf{y}) - V(\pi^z,T,\mathbf{y})$. 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 the Beta-Bernoulli MAB, given any T and $\mathbf{y}$, we have*

$$W^{\mathrm{TS}}(T,\mathbf{y}) - V(\pi^{\mathrm{TS}},T,\mathbf{y}) \le 3K + 2\sqrt{\log T}\times 2\sqrt{KT}, \tag{54}$$

$$W^{\textsc{Irs.FH}}(T,\mathbf{y}) - V(\pi^{\textsc{Irs.FH}},T,\mathbf{y}) \le 3K + 2\sqrt{\log T}\times\left(2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right), \tag{55}$$

$$W^{\textsc{Irs.V-Zero}}(T,\mathbf{y}) - V(\pi^{\textsc{Irs.V-Zero}},T,\mathbf{y}) \le 2K + \sqrt{\log T}\times\left(2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right). \tag{56}$$

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.[8]

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 randomness and 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 mirrors the approach taken by Russo and Van Roy (2014), who exploit the probability matching property of TS to bound its Bayesian regret. More specifically, for each penalty function, we 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 with high probability. As we adopt a better penalty function, the confidence intervals can be made tighter so that the total suboptimality can also be bounded more effectively.

[8] Bubeck and Liu (2013) have shown that the Bayesian regret of TS is bounded by $14\sqrt{KT}$ when the rewards have a bounded support in $[0,1]$, as in the case of Beta-Bernoulli MAB, and it cannot be improved in terms of asymptotic order. Despite its lower asymptotic order, the actual number given in (54) is tighter than $14\sqrt{KT}$ for small T.

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, 2017) (IDS), and the optimistic Gittins index (Farias and Gutin, 2016) (OGI, one-step look-ahead approximation with a discount factor of $\gamma_t = 1 - \frac{1}{t}$).

Our numerical experiments include Beta-Bernoulli MABs and Gaussian MABs. Given a 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 $s \in [S]$, 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 $n \in [T_{\max}]$ and $a \in \mathcal{A}$, which is equivalent to sampling an outcome $\omega^{(s)} \sim \mathcal{I}(\mathbf{y})$. Given the s^{th} sampled outcome $\omega^{(s)}$, for each time horizon $T \in \{5, 10, 15, \ldots, T_{\max}\}$, we simulate each policy π (that can utilize the time horizon T); i.e., at each time $t = 1, \ldots, T$, the policy makes a decision which arm to pull, a_t^π, and then the reward $R^{(s)}_{a_t^\pi, n_t(\mathbf{a}_{1:t}^\pi, a_t^\pi)}$ 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 s^{th} 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 $W^z(T, \mathbf{y})$:

$$V(\pi, T, \mathbf{y}) \approx \frac{1}{S} \sum_{s=1}^{S} \left(\sum_{t=1}^{T} \mu_{a_t^\pi}(\theta_{a_t^\pi}^{(s)}) \right), \quad W^z(T, \mathbf{y}) \approx \frac{1}{S} \sum_{s=1}^{S} \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega^{(s)}) - z_t(\mathbf{a}_{1:t}, \omega^{(s)}) \right\}. \tag{57}$$

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:

$$\text{BayesRegret}(\pi, T, \mathbf{y}) \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \max_a \mu_a(\theta_a) - \mu_{a_t^\pi}(\theta_{a_t^\pi}) \right] = W^{\text{TS}}(T, \mathbf{y}) - V(\pi, T, \mathbf{y}), \tag{58}$$

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:

$$\text{RegretBound}(z, T, \mathbf{y}) \triangleq W^{TS}(T, \mathbf{y}) - W^{z}(T, \mathbf{y}). \tag{59}$$

By the 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

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

$$\theta_a \sim \text{Beta}(1, 1), \quad R_{a,n} \sim \text{Bernoulli}(\theta_a), \quad \forall a \in \{1, 2\}. \tag{60}$$

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 (10) 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. Table 3 provides further details including the percentage improvement in regret over TS, i.e.,

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

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

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

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, while it 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 which 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.

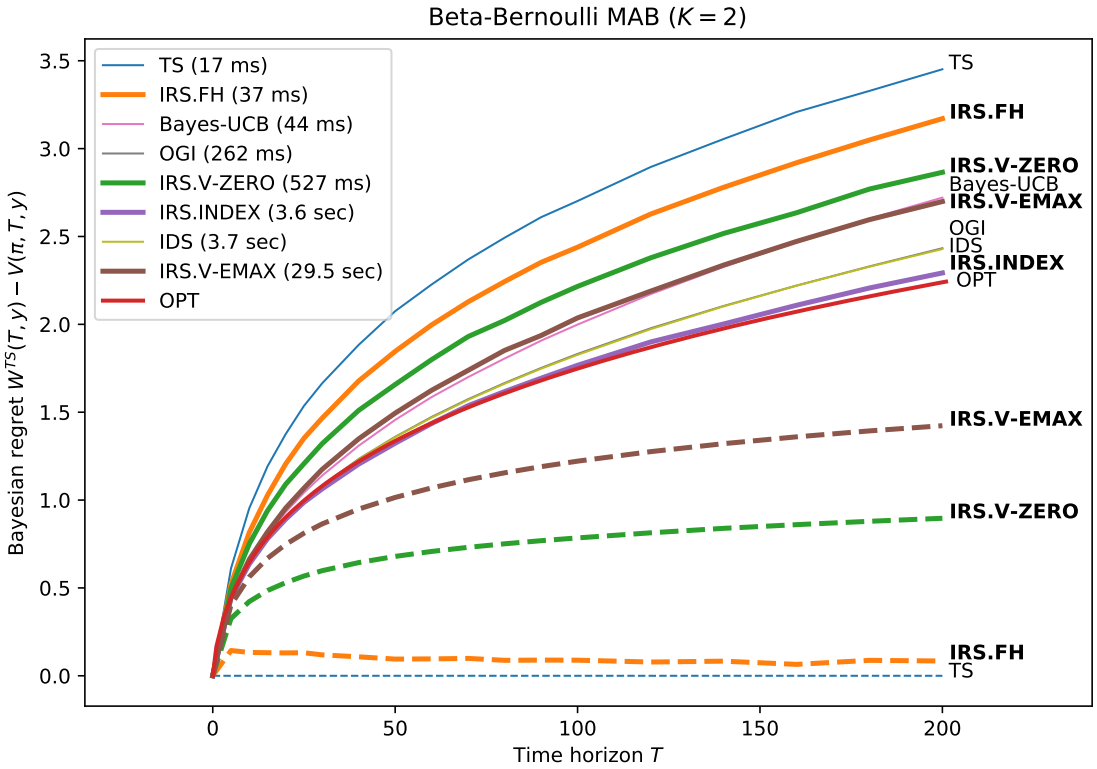

Figure 2: Regret plot for a Beta-Bernoulli MAB with two arms. The solid lines represent the (Bayesian) regret of algorithms, $W^{\mathrm{TS}}(T,\mathbf{y}) - V(\pi,T,\mathbf{y})$, and the dashed lines represent the regret bounds that IRS algorithms produce, $W^{\mathrm{TS}}(T,\mathbf{y}) - W^{z}(T,\mathbf{y})$.

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	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
OPT	**2.24** (–)	35.1%	**2.24** (–)	100.0%	–

Table 3: Regret results for the algorithms in a Beta-Bernoulli MAB when $K = 2$ and $T = 200$. 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 to simulate one sample path throughout $t = 1, \ldots, T$.

Bernoulli MAB with ten arms ($K = 10$). We next consider a Beta-Bernoulli MAB with ten arms ($K = 10$) and $T_{\max} = 500$. Figure 3 and Table 4 show the results for this case while IRS.V-EMAX and OPT are omitted due to their computational cost. 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.

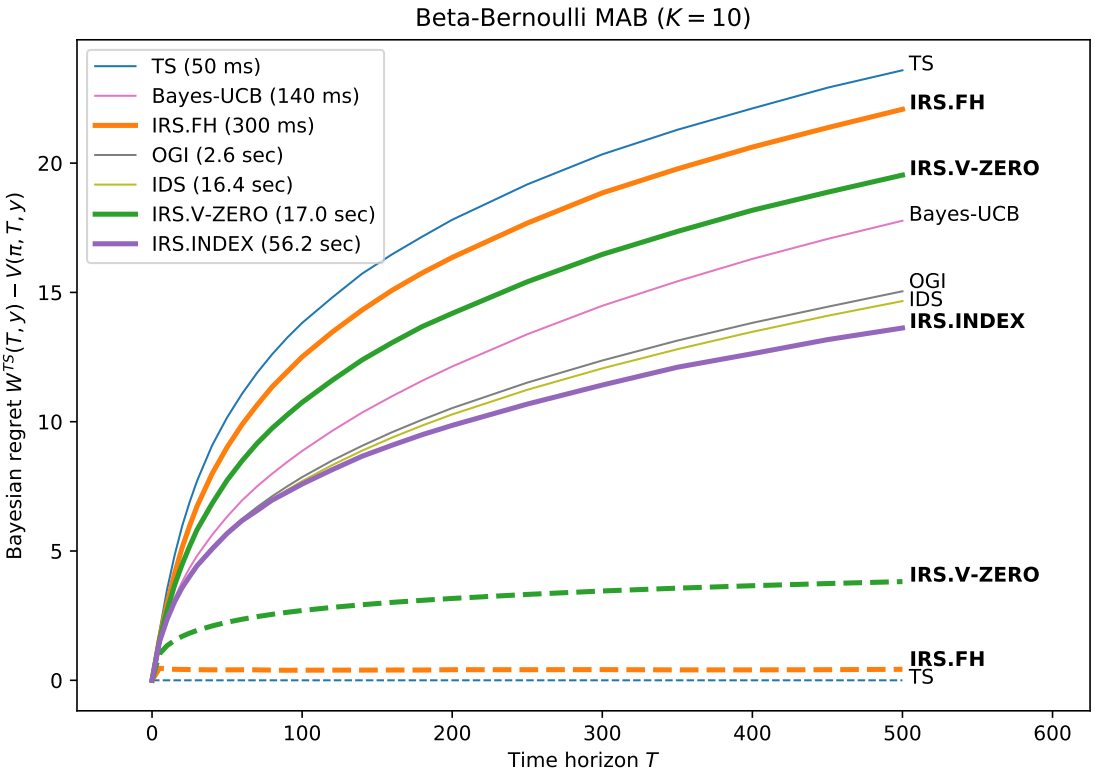

Figure 3: Regret plot for a Beta-Bernoulli MAB with ten arms.

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	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)	3.1%	300 ms
IRS.V-ZERO	19.54 (0.074)	17.2%	**3.82** (0.058)	28.0%	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: Regret results for the algorithms in a Beta-Bernoulli MAB when $K = 10$ and $T = 500$.

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

$$\theta_a \sim \mathcal{N}(0, 1^2), \quad R_{a,n} \sim \mathcal{N}(\theta_a, 1^2), \quad \forall a \in \{1, \ldots, K\}. \tag{61}$$

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 results are similar to those of Beta-Bernoulli MABs.

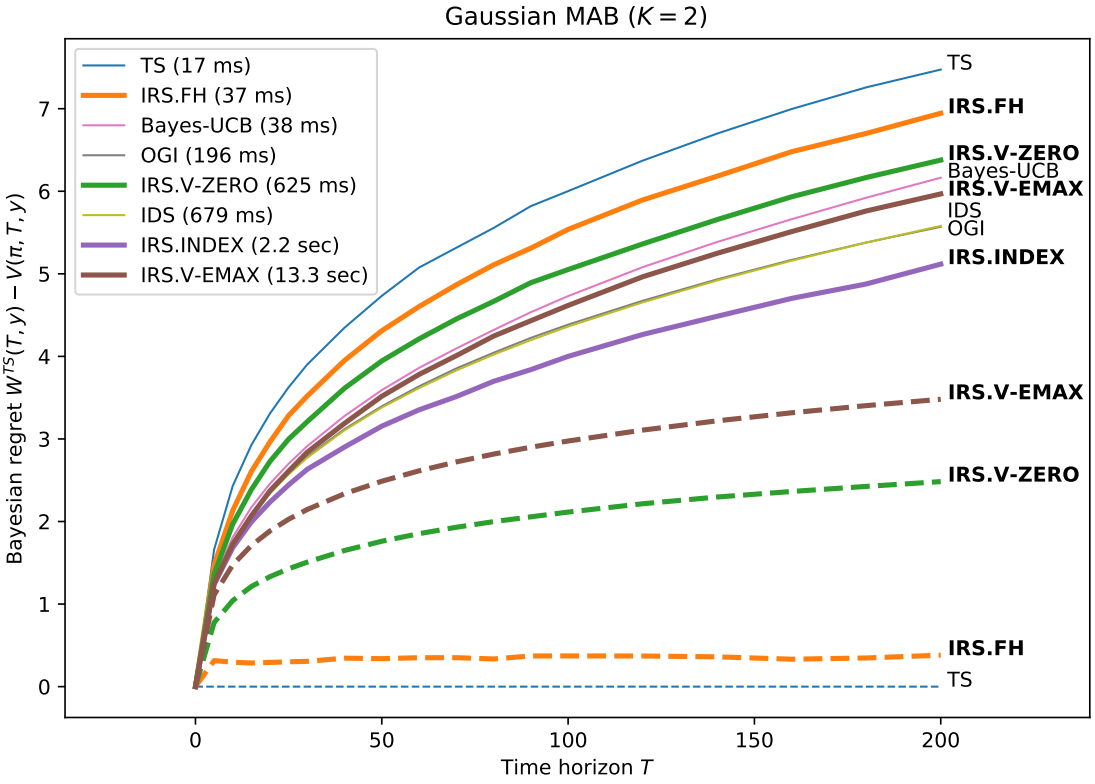

Figure 4: Regret plot for Gaussian MAB with two arms.

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	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: Regret results for the algorithms in a Gaussian MAB when $K = 2$ and $T = 200$.

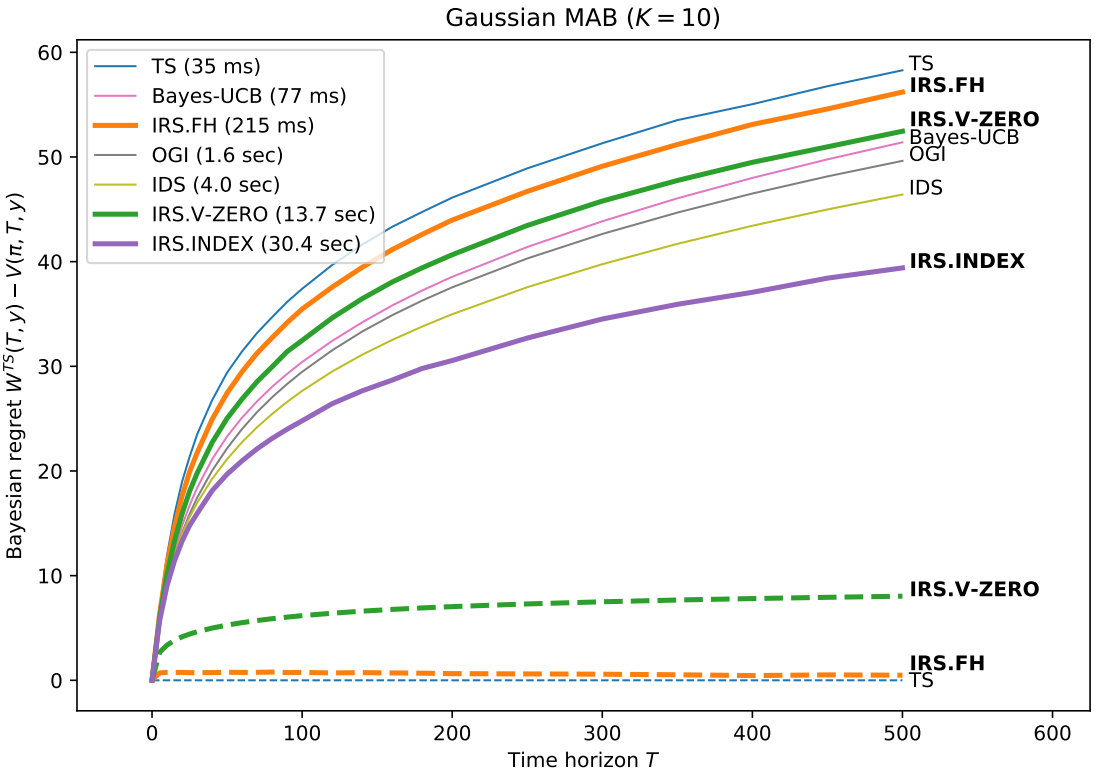

Figure 5: Regret plot for a Gaussian MAB with ten arms.

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	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: Regret results for the algorithms in a Gaussian MAB when $K = 10$ and $T = 500$.

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

$$\theta_a \sim \mathcal{N}(0, 1^2), \quad R_{a,n} \sim \mathcal{N}(\theta_a, \sigma_a^2), \quad \forall a \in \{1, \ldots, 5\} \tag{62}$$

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.[9]

[9]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 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 TS in performance 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[10] 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.

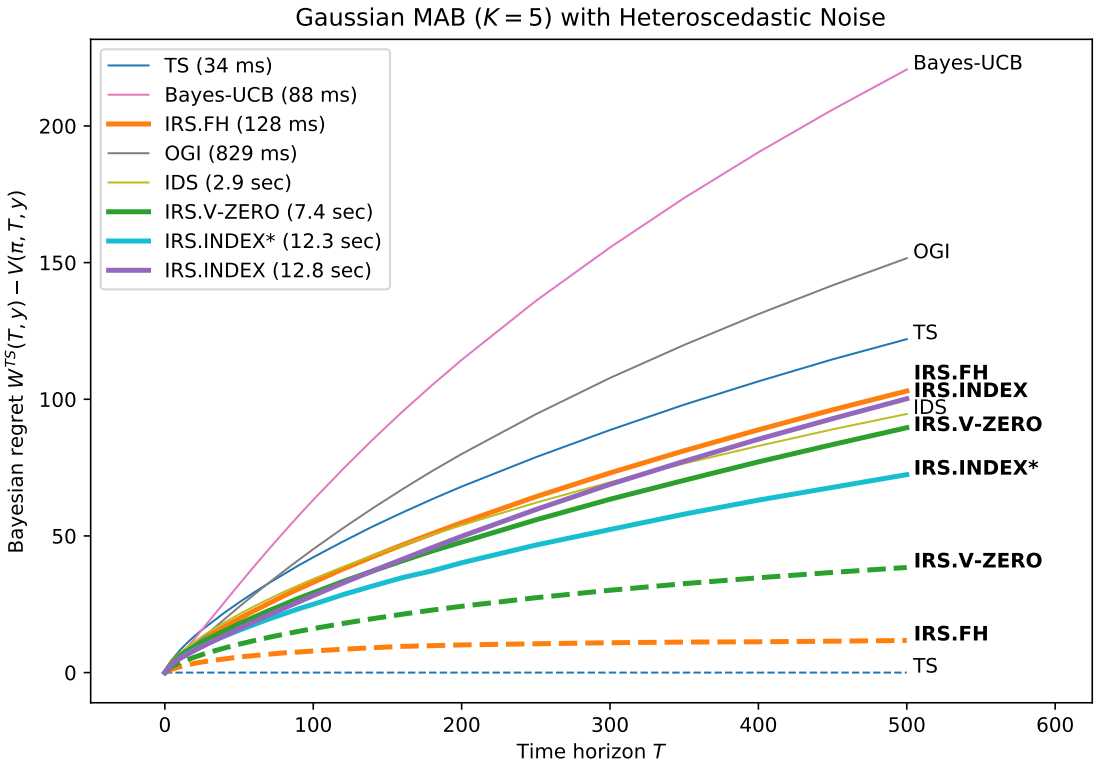

Figure 6: Regret plot for a Gaussian MAB with five arms in which noise variance varies.

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

[10]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	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: Regret results for the algorithms in a Gaussian MAB when $K = 5$, $T = 500$ and $\boldsymbol{\sigma}_{1:K} = (0.1,\ 0.4,\ 1,\ 4,\ 10)$.

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., extensions for 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 – sample 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 not only the future reward realizations). Structural results such as weak duality and strong duality will continue to hold.

Second, we can consider 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,[11] 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 IRS.V-EMAX or IRS.INDEX policy into an anytime policy by setting the inner

[11] 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 be no longer 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 MAB literature, we believe that our framework applies for 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 systemic way of improving TS by taking into account the exploitation-exploration trade-off more carefully, particularly in the presence of some constraint that incurs 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 opti-

mality of an algorithm or in assessing the intrinsic difficulty of a 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 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. *Proceeds of the 16th International Conference on Artificial Intelligence and Statistics*, pages 99–107, 2013.

Donald A. Berry and Bert Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, 1985.

Rusell N. Bradt, S. M. Johnson, and Samuel Karlin. On sequential designs for maximizing the sum of n observations. *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, James E. Smith, and Peng Sun. Information relaxations and duality in stochastic dynamic programs. *Operations Research*, 58(4):785–801, 2010.

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

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

Vijay V. Desai, Vivek F. Farias, and Ciamac C. Moallemi. 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. IEEE Press, December 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. *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, 2013.

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

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

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

Martin B. Haugh and Octavio R. Lacedelli. Information relaxation bounds for partially observed markov decision processes. *IEEE Transactions on Automatic Control*, 2019.

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

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

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

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Bshouty N.H., Stoltz G., Vayatis N., Zeugmann T. (eds) Algorithmic Learning Theory. ALT 2012. Lecture Notes in Computer Science, vol 7568*. Springer, Berlin, Heidelberg, 2012b.

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

Tor Lattimore. Regret analysis of the finite-horizon Gittins index strategy for multi-armed bandits. *29th Annual Conference on Learning Theory*, 49:1–32, 2016.

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

José Niño-Mora. Computing a classic index for finite-horzion 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, 2017.

Daniel Russo, David Tse, and Benjamin Van Roy. Time-sensitive bandit learning and satisficing Thompson sampling. 2017.

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

A. An Illustrative Example

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

$$\theta_1 \sim \text{Beta}(3,1), \quad \theta_2 \sim \text{Beta}(1,1), \quad \theta_3 \sim \text{Beta}(1,3), \tag{63}$$

where $R_{a,n} \sim \text{Bernoulli}(\theta_a)$ for each $a \in \{1,2,3\}$ and $n \in \{1,2,\cdots,8\}$. Given this prior belief, the expected mean reward of each arm is $\bar{\mu}_1 = \mathbb{E}_{\theta_1 \sim \text{Beta}(3,1)}[\theta_1] = \frac{3}{4}$, $\bar{\mu}_2 = \frac{1}{2}$, 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:

	Params θ_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 outcome in a Beta-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 θ_3 is largest among $(\theta_1, \theta_2, \theta_3)$.

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 ($z_t \equiv 0$), 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 n^{th} 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	7
Arm 2	1	0	0	1	1	1	1	0	
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 θ_a whenever playing an arm a. The hindsight optimal action sequence is to pull arm

3 (the arm with the largest mean reward θ_a) eight times in a row and the DM can earn a total reward of $T \times \theta_3 = 6.296$ at most.

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

	Payoffs under $z_t^{\mathbf{Irs.FH}}$								**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	5.333
Arm 2	6/9	6/9	6/9	6/9	6/9	6/9	6/9	6/9	
Arm 3	6/11	6/11	6/11	6/11	6/11	6/11	6/11	6/11	

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 n^{th} 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 $z_t^{\mathbf{Irs.V\text{-}Zero}}$								**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	5.314
Arm 2	1/2	2/3	2/4	2/5	3/6	4/7	5/8	6/9	
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[12] ω.

[12]For details, see the proof of the strong duality theorem in §C.1. While the maximal value does not depend

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 Beta-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

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 $\tilde{\omega}_1 \sim \mathcal{I}(8, \mathbf{y}_0)$) 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** $\tilde{\omega}_1 \sim \mathcal{I}(8, \mathbf{y}_0)$								**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

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.

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., $\tilde{\omega}_2 \sim \mathcal{I}(7, \mathbf{y}_1)$, 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.[13] 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** $\tilde{\omega}_2 \sim \mathcal{I}(7, \mathbf{y}_1)$							**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	$a_2 = 2$
Arm 2	Beta(1, 1)	1/2	2/3	3/4	3/5	4/6	4/7	5/8	
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 $\tilde{\omega}_t \sim \mathcal{I}(T - t + 1, \mathbf{y}_{t-1})$, (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.

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

[13] In case of Irs.V-Zero, we select the arm with the largest pull allocation as a first action.

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: Single decision making under the IRS.V-ZERO policy

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

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

B.2. Implementation of IRS.V-EMax

We use the notation of $\mathbf{y}_t(\mathbf{n}_{1:K}, \omega)$ to denote the belief as a function of pull counts $\mathbf{n}_{1:K} \triangleq (n_1, \cdots, n_K) \in \mathbb{N}_0^K$, based on the observation that the belief is completely by only how many times each of the arms was pulled, $\mathbf{n}_{1:K}$, irrespective of the specific sequence in which the arms were pulled. Given the pull counts $\mathbf{n}_{1:K}$, we define the payoff of pulling an arm a one more time

after pulling each arm $n_1, \cdots, n_K$ times: with $t = \sum_{a=1}^K n_a$, we get

$$r^z(\mathbf{n}_{1:K}, a, \omega) \triangleq \hat{\mu}_{a,n_a}(\omega) - W^{\mathrm{TS}}\left(T-t-1, \mathbf{y}_{t+1}(\mathbf{n}_{1:K}+\mathbf{e}_a, \omega)\right) + W^{\mathrm{TS}}\left(T-t-1, \mathbf{y}_t(\mathbf{n}_{1:K}, \omega)\right), \tag{64}$$

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

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

$$M(\mathbf{n}_{1:K}, \omega) \triangleq \max_{\mathbf{a}_{1:t} \in \mathcal{A}^t} \left\{ \sum_{s=1}^t r_s(\mathbf{a}_{1:s}, \omega) - z_s^{\text{IRS.V-EMAX}}(\mathbf{a}_{1:s}, \omega);\ \sum_{s=1}^t \mathbf{1}\{a_s = a\} = n_a, \forall a \right\}. \tag{65}$$

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

$$M(\mathbf{n}_{1:K}, \omega) = \max_{a \in \mathcal{A}: n_a \geq 1} \left\{ M(\mathbf{n}_{1:K} - \mathbf{e}_a, \omega) + r^z(\mathbf{n}_{1:K} - \mathbf{e}_a, a, \omega) \right\}. \tag{66}$$

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 (66) in an appropriate order. By doing so, we can obtain the maximal value of the original inner problem $(*)$ by evaluating

$$\max_{\mathbf{n}_{1:K} \in N_T} \left\{ M(\mathbf{n}_{1:K}, \omega) \right\}, \tag{67}$$

where $N_T \triangleq \{(n_1, \cdots, n_K) \in \mathbb{N}_0^K : \sum_{a=1}^K n_a = T\}$. The optimal action sequence $\mathbf{a}_{1:T}^*$ can be obtained by tracking $M(\mathbf{n}_{1:K}, \omega)$'s backward.

Algorithm 6: Single decision making under the IRS.V-EMAX policy

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

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

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

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

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

end

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

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

end

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

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

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

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

end

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

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

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

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

end

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 $W^{\mathrm{TS}}(T, \mathbf{y}) = T \times \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{P}(\mathbf{y})} \left[\max_a \mu_a(\theta_a)\right]$ 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

Proposition 2. *The optimization problem* (41) *can be reformulated as*

$$\max_{0 \leq n \leq T} \left\{ T \times \Gamma_0^\lambda + (T - n) \times \left(\lambda - \min_{0 \leq i \leq n} \Gamma_i^\lambda\right) + \sum_{i=1}^n \left(\hat{\mu}_{a,i-1} - \Gamma_{i-1}^\lambda\right) \right\}. \tag{68}$$

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

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

$$\sum_{n=1}^{m} \hat{\mu}_{a,n-1} + (T-m) \times \lambda - \sum_{n=1}^{m} (T-t_n) \times \left(\Gamma^\lambda_n - \Gamma^\lambda_{n-1}\right), \tag{69}$$

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

$$\sum_{n=1}^{m} (T-t_n) \times \left(\Gamma^\lambda_n - \Gamma^\lambda_{n-1}\right) = \sum_{n=1}^{m} (T-t_n) \times \Gamma^\lambda_n - \sum_{n=1}^{m} (T-t_n) \times \Gamma^\lambda_{n-1} \tag{70}$$

$$= \sum_{n=1}^{m} (T-t_n) \times \Gamma^\lambda_n - \sum_{n=0}^{m-1} (T-t_{n+1}) \times \Gamma^\lambda_n \tag{71}$$

$$= \sum_{n=0}^{m} (T-t_n) \times \Gamma^\lambda_n - (T-t_0) \times \Gamma^\lambda_0 - \sum_{n=0}^{m} (T-t_{n+1}) \times \Gamma^\lambda_n + (T-t_{m+1}) \times \Gamma^\lambda_m \tag{72}$$

$$= -\Gamma^\lambda_m - T \times \Gamma^\lambda_0 + \sum_{n=0}^{m} (t_{n+1} - t_n) \times \Gamma^\lambda_n. \tag{73}$$

In order to minimize (73), we need to set $t_{n^*+1} - t_{n^*} = T - m + 1$ for $n^* \triangleq \operatorname{argmin}_{0 \leq n \leq m} \Gamma^\lambda_n$ and $t_{n+1} - t_n = 1$ for $n \neq n^*$. For such t_n's, (69) reduces to

$$\sum_{n=1}^{m} \hat{\mu}_{a,n-1} + (T-m) \times \lambda - \left(-\Gamma^\lambda_m - T \times \Gamma^\lambda_0 + \sum_{n=0}^{m} \Gamma^\lambda_n + (T-m) \times \min_{0 \leq n \leq m} \Gamma^\lambda_m \right) \tag{74}$$

$$= \sum_{n=1}^{m} \hat{\mu}_{a,n-1} + (T-m) \times \left(\lambda - \min_{0 \leq n \leq m} \Gamma^\lambda_m \right) + T \times \Gamma^\lambda_0 - \sum_{n=0}^{m-1} \Gamma^\lambda_n. \tag{75}$$

By taking its maximum value over $m = 0, \cdots, T$, we obtain (68). ■

Algorithm 7: Single decision making under the IRS.INDEX policy

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

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

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

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

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

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

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

end

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

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

  return true

else

  return false

end

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

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

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

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

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

end

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

C. Proofs for §3

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

$$\mathbb{E}^\pi_{\omega\sim\mathcal{I}(\mathbf{y})}\left[\sum_{t=1}^T r_t(\mathbf{a}^\pi_{1:t},\omega) - z_t(\mathbf{a}^\pi_{1:t},\omega)\right] = \mathbb{E}^\pi_{\omega\sim\mathcal{I}(\mathbf{y})}\left[\sum_{t=1}^T r_t(\mathbf{a}^\pi_{1:t},\omega)\right] \equiv V(\pi,T,\mathbf{y}). \tag{76}$$

Proof. We define an appending operator $\oplus$ that concatenates an element into a vector such that

$\mathbf{a}_{1:t} \equiv \mathbf{a}_{1:t-1} \oplus a_t$. When $\pi \in \Pi_{\mathbb{F}}$ and z_t is dual feasible and ω is omitted for brevity, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}^{\pi}) - z_t(\mathbf{a}_{1:t}^{\pi})\right] = \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}^{\pi}) - \mathbb{E}\left[z_t(\mathbf{a}_{1:t}^{\pi}) \middle| \mathcal{F}_{t-1}\right]\right] \tag{77}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}^{\pi}) - \mathbb{E}\left[\sum_{a\in\mathcal{A}} z_t(\mathbf{a}_{1:t-1}^{\pi} \oplus a) \cdot \mathbf{1}\{a_t^{\pi} = a\} \middle| \mathcal{F}_{t-1}\right]\right] \tag{78}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \left(r_t(\mathbf{a}_{1:t}^{\pi}) - \sum_{a\in\mathcal{A}} \underbrace{\mathbb{E}\left[z_t(\mathbf{a}_{1:t-1}^{\pi} \oplus a) \middle| \mathcal{F}_{t-1}\right]}_{=0} \cdot \mathbf{1}\{a_t^{\pi} = a\}\right)\right] \tag{79}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}^{\pi})\right]. \tag{80}$$

■

C.1. Proof of Theorem 1

Weak duality. Define $\mathcal{G}_t \triangleq \mathcal{F}_t \cup \sigma(\omega)$ 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

$$V^*(T, \mathbf{y}) \triangleq \sup_{\pi\in\Pi_{\mathbb{F}}} \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}^{\pi})\right] \overset{\text{Prop 3}}{=} \sup_{\pi\in\Pi_{\mathbb{F}}} \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}^{\pi}) - z_t(\mathbf{a}_{1:t}^{\pi})\right] \tag{81}$$

$$\leq \sup_{\pi\in\Pi_{\mathbb{G}}} \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}^{\pi}) - z_t(\mathbf{a}_{1:t}^{\pi})\right] = \mathbb{E}\left[\max_{\mathbf{a}_{1:T}\in\mathcal{A}^T} \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}) - z_t(\mathbf{a}_{1:t})\right] \tag{82}$$

$$= W^z(T, \mathbf{y}), \tag{83}$$

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 (17). With $V_{T+1}^{\text{in}} \equiv 0$, we have the Bellman equation for the inner problem:

$$Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) \triangleq r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) - z_t^{\text{ideal}}(\mathbf{a}_{1:t-1} \oplus a, \omega) + V_{t+1}^{\text{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega), \tag{84}$$

$$V_t^{\text{in}}(\mathbf{a}_{1:t-1}, \omega) = \max_{a\in\mathcal{A}} \left\{Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega)\right\}. \tag{85}$$

We argue by induction to show that

$$V_t^{\text{in}}(\mathbf{a}_{1:t-1}, \omega) = V^*(T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)), \tag{86}$$

$$Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) = Q^*(T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a), \tag{87}$$

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

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

$$
\begin{aligned}
Q^{\text{in}}_t(\mathbf{a}_{1:t-1}, a, \omega) &= r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) - z^{\text{ideal}}_t(\mathbf{a}_{1:t-1} \oplus a, \omega) + V^{\text{in}}_{t+1}(\mathbf{a}_{1:t-1} \oplus a, \omega) && (88)\\
&= \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) + V^*\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega)\right) \middle| \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right] && (89)\\
&\quad \underbrace{-V^*\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega)\right) + V^{\text{in}}_{t+1}(\mathbf{a}_{1:t-1} \oplus a, \omega)}_{=0} && (90)\\
&= \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) + V^*\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega)\right) \middle| \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right] && (91)\\
&= \mathbb{E}_{r \sim \mathcal{R}_a(\mathcal{P}_a([\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)]_a))}\left[r + V^*\left(T-t, \mathcal{U}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a, r)\right) \right] && (92)\\
&= Q^*(T-t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a), && (93)
\end{aligned}
$$

where the last equality follows from the original Bellman equation (10). Consequently,

$$
\begin{aligned}
V^{\text{in}}_t(\mathbf{a}_{1:t-1}, \omega) &= \max_{a \in \mathcal{A}} \left\{ Q^{\text{in}}_t(\mathbf{a}_{1:t-1}, a, \omega) \right\} && (94)\\
&= \max_{a \in \mathcal{A}} \left\{ Q^*(T-t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a) \right\} && (95)\\
&= V^*(T-t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)). && (96)
\end{aligned}
$$

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

$$
V^{\text{in}}_1(\emptyset, \omega) = V^*(T, \mathbf{y}), \quad Q^{\text{in}}_1(\emptyset, a, \omega) = Q^*(T, \mathbf{y}, a), \quad \forall \omega. \tag{97}
$$

Note that the maximal value of the inner problem does not depend on ω, which 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 (97), it yields

$$
a^{\pi^{\text{ideal}}} = \operatorname*{argmax}_a Q^{\text{in}}_1(\emptyset, a, \tilde{\omega}) = \operatorname*{argmax}_a Q^*(T, \mathbf{y}, a). \tag{98}
$$

Therefore, at each moment, no matter what $\tilde{\omega}$ is chosen, the policy π^{ideal} always takes the same action that Bayesian optimal policy would 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 $\mathbb{E}\left[r_t(\mathbf{a}_{1:t},\omega)|\,\mathcal{F}_{t-1}\right]=\hat{\mu}_{a_t,n_{t-1}(\mathbf{a}_{1:t-1},a_t)}(\omega)$. Also note that

$$\mathbb{E}\left[\mathbb{E}\left(r_t(\mathbf{a}_{1:t},\omega)|\boldsymbol{\theta}\right)|\,\mathcal{F}_{t-1}\right]=\mathbb{E}\left[\mu_{a_t}(\theta_{a_t})|\,\mathcal{F}_{t-1}\right]=\hat{\mu}_{a_t,n_{t-1}(\mathbf{a}_{1:t-1},a_t)}(\omega), \tag{99}$$

and

$$\mathbb{E}\left[\mathbb{E}\left(r_t(\mathbf{a}_{1:t},\omega)|\hat{\boldsymbol{\mu}}_{1:K,T-1}(\omega)\right)|\,\mathcal{F}_{t-1}\right]=\mathbb{E}\left[\hat{\mu}_{a_t,T-1}(\omega)|\,\mathcal{F}_{t-1}\right]=\hat{\mu}_{a_t,n_{t-1}(\mathbf{a}_{1:t-1},a_t)}(\omega). \tag{100}$$

Therefore, $\mathbb{E}\left[z_t^{\mathrm{TS}}|\mathcal{F}_{t-1}\right]=\mathbb{E}\left[r_t|\,\mathcal{F}_{t-1}\right]-\mathbb{E}\left[\mathbb{E}(r_t|\boldsymbol{\theta})|\,\mathcal{F}_{t-1}\right]=0$, and $\mathbb{E}\left[z_t^{\textsc{Irs.FH}}|\mathcal{F}_{t-1}\right]=\mathbb{E}\left[r_t|\,\mathcal{F}_{t-1}\right]-\mathbb{E}\left[\mathbb{E}(r_t|\hat{\boldsymbol{\mu}}_{1:K,T-1})|\,\mathcal{F}_{t-1}\right]=0$. The other penalty functions have a form of $z_t=X_t-\mathbb{E}\left[X_t|\mathcal{F}_{t-1}\right]$ for some X_t. Therefore, $\mathbb{E}\left[z_t|\,\mathcal{F}_{t-1}\right]=\mathbb{E}\left[X_t-\mathbb{E}(X_t|\mathcal{F}_{t-1})|\,\mathcal{F}_{t-1}\right]=\mathbb{E}\left[X_t-X_t|\,\mathcal{F}_{t-1}\right]=0$. ■

D. Proofs for §4

D.1. Notes on Regularity

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

$$\mathbb{E}|\mu_a(\theta_a)|<\infty\quad\textit{and}\quad W^{\mathrm{TS}}(T,\mathbf{y})<\infty. \tag{101}$$

Proof. By Jensen's inequality,

$$\mathbb{E}|\mu_a(\theta_a)|=\mathbb{E}\left[|\mathbb{E}\left(R_{a,n}|\theta_a\right)|\right]\leq\mathbb{E}\left[\mathbb{E}\left(\,|R_{a,n}||\,\theta_a\right)\right]=\mathbb{E}|R_{a,n}|<\infty. \tag{102}$$

Consequently,

$$\mathbb{E}\left[\max_a\mu_a(\theta_a)\right]\leq\mathbb{E}\left[\sum_{a=1}^K|\mu_a(\theta_a)|\right]=\sum_{a=1}^K\mathbb{E}|\mu_a(\theta_a)|<\infty. \tag{103}$$

■

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

$$\lim_{n\to\infty}\hat{\mu}_{a,n}(\omega)=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n R_{a,i}=\mu_a(\theta_a)\quad\textit{almost surely}, \tag{104}$$

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

Proof. Fix a and let $\mathcal{H}_n\triangleq\sigma\left(R_{a,1},\cdots,R_{a,n}\right)$. 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 $\hat{\mu}_{a,n}=\mathbb{E}\left(\mu_a(\theta_a)|\mathcal{H}_n\right)$ 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 $\mathbb{E}\left(\mu_a(\theta_a)|\mathcal{H}_\infty\right) = \mu_a(\theta_a)$ almost surely as $n \to \infty$. ■

D.2. Proof of Proposition 1

Asymptotic behavior of Irs.FH. Let $\tilde{\omega}$ be the sampled outcome used by $\textsc{Irs.FH}(T, \mathbf{y})$. 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 $\operatorname{argmax}_a \mu_a(\tilde{\theta}_a)$ is uniquely defined for almost all $\tilde{\omega}$, yields

$$\operatorname*{argmax}_a \mu_a(\tilde{\theta}_a) = \operatorname*{argmax}_a \lim_{n\to\infty} \hat{\mu}_{a,n}(\tilde{\omega}) = \lim_{n\to\infty} \operatorname*{argmax}_a \hat{\mu}_{a,n}(\tilde{\omega}) \quad \text{a.s.} \tag{105}$$

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

$$\lim_{T\to\infty} \mathbb{P}\left[\textsc{Irs.FH}(T, \mathbf{y}) = a\right] = \lim_{T\to\infty} \mathbb{P}\left[\operatorname*{argmax}_{a'} \hat{\mu}_{a',T-1}(\tilde{\omega}) = a\right] \tag{106}$$

$$= \mathbb{P}\left[\operatorname*{argmax}_{a'} \mu_{a'}(\tilde{\theta}_{a'}) = a\right] \tag{107}$$

$$= \mathbb{P}\left[\mathrm{TS}(\mathbf{y}) = a\right]. \tag{108}$$

Note that we are not particularly assuming that $\textsc{Irs.FH}(T, \mathbf{y})$ and $\mathrm{TS}(\mathbf{y})$ share the randomness. The sampled parameters used in $\mathrm{TS}(\mathbf{y})$ are not necessarily the ones used in $\textsc{Irs.FH}(T, \mathbf{y})$, but their distributions are identical since they are drawn from the same prior. ■

Asymptotic behavior of Irs.V-Zero. Let $a_T^\circ(\tilde{\omega}) \triangleq \textsc{Irs.V-Zero}(T, \mathbf{y})$ in which $\tilde{\omega}$ is used, and let $a^{\mathrm{TS}}(\tilde{\omega}) \triangleq \operatorname{argmax}_a \mu_a(\tilde{\theta}_a)$. As above, it suffices to show that $\lim_{T\to\infty} a_T^\circ(\tilde{\omega}) = a^{\mathrm{TS}}(\tilde{\omega})$ for almost all $\tilde{\omega}$. We now fix $\tilde{\omega}$ and omit it from the proof for brevity.

Define

$$\Delta \triangleq \min_{a \neq a^{\mathrm{TS}}} \left|\mu_{a^{\mathrm{TS}}}(\tilde{\theta}_{a^{\mathrm{TS}}}) - \mu_a(\tilde{\theta}_a)\right| \quad \text{and} \quad M \triangleq \sup_{a\in\mathcal{A}, n\geq 0} \left|\hat{\mu}_{a,n}\right|. \tag{109}$$

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 $\lim_{n\to\infty} \hat{\mu}_{a,n} = \mu_a(\tilde{\theta}_a) < \infty$ almost surely for all a. In addition, there exists $N \in \mathbb{N}$ such that

$$\left|\hat{\mu}_{a,n} - \mu_a(\tilde{\theta}_a)\right| < \frac{\Delta}{4}, \quad \forall n \geq N, \quad \forall a \in \mathcal{A}. \tag{110}$$

For such N, we have

$$\inf_{n\geq N} \hat{\mu}_{a^{\mathrm{TS}},n} \geq \sup_{n\geq N} \hat{\mu}_{a,n} + \frac{\Delta}{2}, \quad \forall a \neq a^{\mathrm{TS}}. \tag{111}$$

Note that a^{TS}, Δ, M, and N are determined only by $\tilde{\omega}$, independently of 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:

$$\mathbf{n}^{\circ}_{1:K} \triangleq \operatorname*{argmax}_{\mathbf{n}_{1:K} \in N_T} \left\{ \sum_{a=1}^{K} \sum_{s=1}^{n_a} \hat{\mu}_{a,s-1} \right\}, \tag{112}$$

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

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

$$\sum_{a=1}^{K} \sum_{s=1}^{n^{\dagger}(a)} \hat{\mu}_{a,s-1} - \sum_{a=1}^{K} \sum_{s=1}^{n^{\circ}(a)} \hat{\mu}_{a,s-1} = \hat{\mu}_{a^{\text{TS}}, n^{\circ}(a^{\text{TS}})} - \hat{\mu}_{a^{\circ}_T, n^{\circ}(a^{\circ}_T)-1} \geq \frac{\Delta}{2} > 0, \tag{113}$$

where the inequality follows from (111). The allocation $\mathbf{n}^{\dagger}_{1:K}$ achieves a strictly better payoff than $\mathbf{n}^{\circ}_{1:K}$, which contradicts the assumption that $\mathbf{n}^{\circ}_{1:K}$ is an optimal allocation.

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

$$n^{\dagger}(a) \triangleq \begin{cases} n^{\circ}(a^{\text{TS}}) + (n^{\circ}(a^{\circ}_T) - N) & \text{if } a = a^{\text{TS}}, \\ N & \text{if } a = a^{\circ}_T, \\ n^{\circ}(a) & \text{otherwise.} \end{cases} \tag{114}$$

By making this allocation, we have

$$\sum_{a=1}^{K} \sum_{s=1}^{n^{\dagger}(a)} \hat{\mu}_{a,s-1} - \sum_{a=1}^{K} \sum_{s=1}^{n^{\circ}(a)} \hat{\mu}_{a,s-1} \tag{115}$$

$$= \sum_{s=n^{\circ}(a^{\text{TS}})+1}^{n^{\circ}(a^{\text{TS}})+(n^{\circ}(a^{\circ}_T)-N)} \hat{\mu}_{a^{\text{TS}},s-1} - \sum_{s=N+1}^{n^{\circ}(a^{\circ}_T)} \hat{\mu}_{a^{\circ}_T,s-1} \tag{116}$$

$$\geq -(N - n^{\circ}(a^{\text{TS}})) \cdot 2M + \sum_{s=N+1}^{n^{\circ}(a^{\circ}_T)} \hat{\mu}_{a^{\text{TS}},s-1} - \sum_{s=N+1}^{n^{\circ}(a^{\circ}_T)} \hat{\mu}_{a^{\circ}_T,s-1} \tag{117}$$

$$\geq -(N - n^{\circ}(a^{\text{TS}})) \cdot 2M + (n^{\circ}(a^{\circ}_T) - N) \cdot \frac{\Delta}{2} \tag{118}$$

$$\geq (n^{\circ}(a^{\circ}_T) - N) \cdot \frac{\Delta}{2} - 2NM. \tag{119}$$

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

We've shown that for almost all $\tilde{\omega}$, when T is large enough, the optimal allocation $\mathbf{n}^{\circ}_{1:K}$ must

allocate more than a half of the pulls on the arm $a^{\mathrm{TS}} = \mathrm{argmax}_a \, \mu_a(\tilde{\theta}_a)$. Therefore, $\lim_{T\to\infty} a^{\circ}_T(\tilde{\omega}) = a^{\mathrm{TS}}(\tilde{\omega})$ for almost all $\tilde{\omega}$, which completes the proof.

D.3. Proof of Theorem 2

D.3.1. "$W^{\mathsf{TS}}(T, \mathbf{y}) \geq W^{\mathsf{Irs.FH}}(T, \mathbf{y})$"

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

$$W^{\mathrm{TS}}(T, \mathbf{y}) = T \times \mathbb{E}\left[\max_a \mu_a(\theta_a)\right] \geq T \times \mathbb{E}\left[\max_a \mathbb{E}\left(\mu_a(\theta_a) \middle| \hat{\boldsymbol{\mu}}_{1:K,T-1}\right)\right] = W^{\mathrm{IRS.FH}}(T, \mathbf{y}). \quad (120)$$

■

D.3.2. "$W^{\mathsf{Irs.FH}}(T, \mathbf{y}) \geq W^{\mathsf{Irs.V\text{-}Zero}}(T, \mathbf{y})$"

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

$$\mathbb{E}\left[\max\left\{X + \varphi(X), 0\right\}\right] \geq \mathbb{E}\left[\max\left\{\mathbb{E}(X) + \varphi(X), 0\right\}\right]. \quad (121)$$

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 $x \in \mathbb{R}$,

$$f_x(t) \geq f_x(\mu) + (t - \mu) \cdot f'_x(\mu) = \max\{\mu + \varphi(x), 0\} + (t - \mu) \cdot \mathbf{1}\{\mu + \varphi(x) \geq 0\}, \quad \forall t, \quad \forall x. \quad (122)$$

By setting $t = x$, we get

$$\max\{x + \varphi(x), 0\} = f_x(x) \geq \max\{\mu + \varphi(x), 0\} + (x - \mu) \cdot \mathbf{1}\{\mu + \varphi(x) \geq 0\}, \quad \forall x. \quad (123)$$

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,

$$(x - \mu) \cdot \mathbf{1}\{\mu + \varphi(x) \geq 0\} \geq (x - \mu) \cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\}, \quad \forall x \in \mathbb{R}. \quad (124)$$

Combining this with (123), we get

$$\max\{x + \varphi(x), 0\} \geq \max\{\mu + \varphi(x), 0\} + (x - \mu) \cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\}, \quad \forall x \in \mathbb{R}. \quad (125)$$

For random variable X, by taking expectation, we get

$$\begin{aligned}
\mathbb{E}\left[\max\{X + \varphi(X), 0\}\right] &\geq \mathbb{E}\left[\max\{\mu + \varphi(X), 0\} + (X - \mu) \cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\}\right] && (126)\\
&\geq \mathbb{E}\left[\max\{\mu + \varphi(X), 0\}\right] + \mathbb{E}(X - \mu) \cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\} && (127)\\
&= \mathbb{E}\left[\max\{\mu + \varphi(X), 0\}\right]. && (128)
\end{aligned}$$

■

Corollary 1. *On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $\varphi(x,\omega) : \mathbb{R} \times \Omega \mapsto \mathbb{R}$ 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*

$$\mathbb{E}\left[\max\left\{X(\omega) + \varphi(X(\omega),\omega), 0\right\}\right] \geq \mathbb{E}\left[\max\left\{\mathbb{E}(X|\mathcal{H})(\omega) + \varphi(X(\omega),\omega), 0\right\}\right]. \tag{129}$$

Proof. Define

$$\mu(\omega) \triangleq \mathbb{E}(X|\mathcal{H})(\omega), \quad I(\omega) \triangleq \mathbf{1}\{\mu(\omega) + \varphi(\mu(\omega),\omega) \geq 0\}. \tag{130}$$

By (125), we have

$$\max\{x + \varphi(x,\omega), 0\} \geq \max\{\mu(\omega) + \varphi(x,\omega), 0\} + (x - \mu(\omega)) \cdot I(\omega), \quad \forall x \in \mathbb{R}, \quad \text{for each } \omega \in \Omega. \tag{131}$$

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

$$\begin{aligned}
\mathbb{E}\left[\max\{X(\omega) + \varphi(X(\omega),\omega), 0\}\right] &\geq \mathbb{E}\left[\max\{\mu(\omega) + \varphi(X(\omega),\omega), 0\} + (X(\omega) - \mu(\omega)) \cdot I(\omega)\right] && (132)\\
&= \mathbb{E}\left[\mathbb{E}\left(\max\{\mu(\omega) + \varphi(X(\omega),\omega), 0\} + (X(\omega) - \mu(\omega)) \cdot I(\omega) \middle| \mathcal{H}\right)\right] && (133)\\
&= \mathbb{E}\left[\max\{\mu(\omega) + \varphi(X(\omega),\omega), 0\}\right] + \mathbb{E}\left[\mathbb{E}\left(\left(X(\omega) - \mu(\omega)\right) \cdot I(\omega) \middle| \mathcal{H}\right)\right] && (134)\\
&= \mathbb{E}\left[\max\{\mathbb{E}\left(X|\mathcal{H}\right)(\omega) + \varphi(X(\omega),\omega), 0\}\right] && (135)\\
&\quad + \mathbb{E}\left[\underbrace{\left(\mathbb{E}(X|\mathcal{H})(\omega) - \mu(\omega)\right)}_{=0} \cdot I(\omega)\right] && (136)\\
&= \mathbb{E}\left[\max\{\mathbb{E}\left(X|\mathcal{H}\right)(\omega) + \varphi(X(\omega),\omega), 0\}\right]. && (137)
\end{aligned}$$

■

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

$$\mathbb{E}\left[\max_{0\leq i\leq T}\left\{(i-n)^+ \times X + C_i\right\}\right] \geq \mathbb{E}\left[\max_{0\leq i\leq T}\left\{\mathbb{E}\left(X\middle|\mathcal{H}\right) \cdot \mathbf{1}\{i \geq n+1\} + (i-n-1)^+ \times X + C_i\right\}\right] \tag{138}$$

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

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

$$\varphi(x,\omega) \triangleq \max_{n+1\leq i\leq T}\{(i-n-1) \times x + C_i(\omega)\} - \max_{0\leq i\leq n}\{C_i(\omega)\}. \tag{139}$$

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

$$\mathbb{E}\left[\max_{0\le i\le T}\left\{(i-n)^+\times X+C_i\right\}\right] \tag{140}$$

$$=\mathbb{E}\left[\max\left\{\max_{n+1\le i\le T}\left\{(i-n)\times X+C_i\right\},\quad \max_{0\le i\le n}C_i\right\}\right] \tag{141}$$

$$=\mathbb{E}\left[\max\left\{X+\max_{n+1\le i\le T}\left\{(i-n-1)\times X+C_i\right\},\quad \max_{0\le i\le n}C_i\right\}\right] \tag{142}$$

$$=\mathbb{E}\left[\max\left\{X(\omega)+\underbrace{\max_{n+1\le i\le T}\left\{(i-n-1)\times X(\omega)+C_i(\omega)\right\}-\max_{0\le i\le n}C_i(\omega)}_{=\varphi(X(\omega),\omega)},\quad 0\right\}+\max_{0\le i\le n}C_i(\omega)\right] \tag{143}$$

$$\ge\mathbb{E}\left[\max\left\{\mathbb{E}\left(X|\mathcal{H}\right)(\omega)+\max_{n+1\le i\le T}\left\{(i-n-1)\times X(\omega)+C_i(\omega)\right\}-\max_{0\le i\le n}C_i(\omega),\quad 0\right\}+\max_{0\le i\le n}C_i(\omega)\right] \tag{144}$$

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

$$=\mathbb{E}\left[\max_{0\le i\le T}\left\{\mathbb{E}\left(X|\,\mathcal{H}\right)\cdot\mathbf{1}\{i\ge n+1\}+(i-n-1)^+\times X+C_i\right\}\right]. \tag{146}$$

■

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

$$N_T\triangleq\left\{\mathbf{n}_{1:K}\in\mathbb{N}_0^K:\sum_{a=1}^K n_a=T\right\}\quad\text{and}\quad S_a(n_a)\triangleq\sum_{i=1}^{n_a}\hat{\mu}_{a,i-1}. \tag{147}$$

What we want to show is

$$W^{\textsc{Irs.FH}}\equiv\mathbb{E}\left[T\times\max_a\{\hat{\mu}_{a,T-1}\}\right]=\mathbb{E}\left[\max_{\mathbf{n}_{1:K}\in N_T}\left\{\sum_{a=1}^K n_a\times\hat{\mu}_{a,T-1}\right\}\right] \tag{148}$$

$$\ge\mathbb{E}\left[\max_{\mathbf{n}_{1:K}\in N_T}\left\{\sum_{a=1}^K S_a(n_a)\right\}\right]\equiv W^{\textsc{Irs.V-Zero}}. \tag{149}$$

Further define

$$U_{k,n}\triangleq\mathbb{E}\left[\max_{\mathbf{n}_{1:K}\in N_T}\left\{\left(\sum_{a=1}^{k-1}S_a(n_a)\right)+\left(S_k(n_k\wedge n)+(n_k-n)^+\times\hat{\mu}_{a,T-1}\right)+\left(\sum_{a=k+1}^K n_a\times\hat{\mu}_{a,T-1}\right)\right\}\right], \tag{150}$$

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

$$U_{k,n}\ge U_{k,n+1},\quad\forall k=1,\cdots,K,\quad\forall n=0,\cdots,T-1. \tag{151}$$

Fix k and n. Define a sub-σ-field

$$\mathcal{H}\triangleq\sigma\left(\{R_{a,s}\}_{a=k,1\le s\le n}\cup\{R_{a,s}\}_{a\ne k,1\le s\le T-1}\right). \tag{152}$$

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

$$C_i \triangleq \max \left\{ \left(\sum_{a=1}^{k-1} S_a(n_a) \right) + S_k(i \wedge n) + \left(\sum_{a=k+1}^{K} n_a \times \hat{\mu}_{a,T-1} \right) \ : \ \mathbf{n}_{1:K} \in N_T, n_k = i \right\}. \tag{153}$$

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

$$U_{k,n} = \mathbb{E}\left[\max_{0 \le i \le T} \left\{ (i-n)^+ \times \hat{\mu}_{k,T-1} + C_i \right\} \right]. \tag{154}$$

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

$$\begin{aligned}
U_{k,n} &= \mathbb{E}\left[\max_{0 \le i \le T} \left\{ (i-n)^+ \times X + C_i \right\} \right] && (155)\\
&\overset{\text{Corollary 2}}{\ge} \mathbb{E}\left[\max_{0 \le i \le T} \left\{ \mathbb{E}\left(X \middle| \mathcal{H} \right) \cdot \mathbf{1}\{i \ge n+1\} + (i-n-1)^+ \times X + C_i \right\} \right] && (156)\\
&\overset{\text{(a)}}{=} \mathbb{E}\left[\max_{0 \le i \le T} \left\{ \hat{\mu}_{k,n} \cdot \mathbf{1}\{i \ge n+1\} + (i-n-1)^+ \times \hat{\mu}_{a,T-1} + C_i \right\} \right] && (157)\\
&\overset{\text{(b)}}{=} U_{k,n+1}. && (158)
\end{aligned}$$

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

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 $W^{\text{TS}} \ge W^{\text{IRS.FH}}$, however, the proof of $W^{\text{IRS.FH}} \ge W^{\text{IRS.V-ZERO}}$ does not simply follow from the fact that $\sigma(\hat{\boldsymbol{\mu}}_{1:K,T-1})$ is a stronger filtration than $\mathcal{F}_t$.

Given a Beta-Bernoulli MAB with $K = 2$, $T = 2$, and a prior distribution $\text{Beta}(1,1)$, consider its variation whose reward function is given by $r'_t(\cdot)$ as follows:

$$r'_1(a_1) = r_1(a_1), \quad r'_2(\mathbf{a}_{1:2}) = -\kappa r_2(\mathbf{a}_{1:2}), \tag{159}$$

where $r_t(\cdot)$ is the reward function of the original MAB problem. When $\kappa > 0$, we have

$$W^{\text{IRS.FH}} = \mathbb{E}\left[\max_{\mathbf{a}_{1:T}} \left\{ \sum_{t=1}^{T} \mathbb{E}(r'_t(\mathbf{a}_{1:t}) | \hat{\boldsymbol{\mu}}_{1:K,T-1}) \right\} \right] = \frac{7}{12} - \frac{5}{12}\kappa, \tag{160}$$

$$W^{\text{IRS.V-ZERO}} = \mathbb{E}\left[\max_{\mathbf{a}_{1:T}} \left\{ \sum_{t=1}^{T} \mathbb{E}(r'_t(\mathbf{a}_{1:t}) | \mathcal{F}_{t-1}) \right\} \right] = \frac{1}{2} - \frac{3}{8}\kappa. \tag{161}$$

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, while IRS.V-ZERO considers the former component only. When those two gains are negatively correlated as in this example (i.e., a higher r_1' leads to a worse next belief state), penalizing only for the former component may end up with a bad outcome (i.e., not a tight upper bound).

As discussed in §3, TS and IRS.FH are obtained by relaxing some future information ($\boldsymbol{\theta}$ and $\hat{\boldsymbol{\mu}}_{1:K,T-1}$, respectively) while applying no penalty for that relaxation. In contrast, IRS.V-ZERO is not an algorithm that can be obtained by relaxing future information with no penalty. Therefore, the comparison between IRS.FH and IRS.V-ZERO is not straightforward as much as the comparison between TS and IRS.FH.

D.3.3. "$W^{\mathsf{TS}}(T,\mathbf{y}) \geq W^{\mathsf{Irs.V\text{-}EMax}}(T,\mathbf{y})$"

To show that $W^{\mathrm{TS}} \geq W^{\text{IRS.V-EMAX}}$, we take a completely different approach that mirrors the proof of Theorem 4 in Desai et al. (2012a).

Definition 2 (Supersolution). *An approximate value function $\widehat{V} : \mathbb{N}_0 \times \mathcal{Y} \mapsto \mathbb{R}$ is a **supersolution** (to the Bellman equation* (10)*) if*

$$\widehat{V}(T,\mathbf{y}) \geq \max_{a\in\mathcal{A}} \left\{ \mathbb{E}_{r\sim\mathcal{R}_a(\mathcal{P}_a(y_a))} \left[r + \widehat{V}(T-1, \mathcal{U}(\mathbf{y},a,r)) \right] \right\}, \quad \forall \mathbf{y} \in \mathcal{Y}, \quad \forall T \geq 1, \tag{162}$$

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

Remark 6. *If $\widehat{V}(\cdot,\cdot)$ is a supersolution, then for any given ω, T, and $\mathbf{y}$,*

$$\widehat{V}(T-t+1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1},\omega;\mathbf{y})) \geq \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1}\oplus a,\omega;\mathbf{y}) + \widehat{V}(T-t, \mathbf{y}_t(\mathbf{a}_{1:t-1}\oplus a,\omega;\mathbf{y})) \middle| \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1},\omega;T,\mathbf{y}) \right], \tag{163}$$

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

Lemma 2. *Consider a penalty function $\hat{z}_t$ generated by $\widehat{V}(\cdot,\cdot)$:*

$$\begin{aligned} \hat{z}_t(\mathbf{a}_{1:t},\omega;T,\mathbf{y}) \triangleq\ & r_t(\mathbf{a}_{1:t},\omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t},\omega) \,|\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1},\omega;T,\mathbf{y}) \right] \\ & + \widehat{V}\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t},\omega;\mathbf{y})\right) - \mathbb{E}\left[\widehat{V}\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t},\omega;\mathbf{y})\right) \middle| \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1},\omega;T,\mathbf{y}) \right]. \end{aligned} \tag{164}$$

If $\widehat{V}(\cdot,\cdot)$ is a supersolution,

$$W^{\hat{z}}(T,\mathbf{y}) \leq \widehat{V}(T,\mathbf{y}). \tag{165}$$

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:

$$V_t^{\hat{z},\mathrm{in}}(\mathbf{a}_{1:t-1},\omega;T,\mathbf{y}) = \max_{a\in\mathcal{A}} \left\{ r_t(\mathbf{a}_{1:t-1}\oplus a,\omega) - \hat{z}_t(\mathbf{a}_{1:t-1}\oplus a,\omega;T,\mathbf{y}) + V_{t+1}^{\hat{z},\mathrm{in}}(\mathbf{a}_{1:t-1}\oplus a,\omega;T,\mathbf{y}) \right\}, \tag{166}$$

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,

$$V_t^{\hat{z},\mathrm{in}}\left(\mathbf{a}_{1:t-1},\omega;T,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1},\omega;\mathbf{y})\right)\leq \widehat{V}\left(T-t+1,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1},\omega;\mathbf{y})\right), \tag{167}$$

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 (167) holds trivially since both sides are zero. Fix t and suppose that the inequality (167) holds for $t+1$. Omitting ω for brevity, we get

$$\widehat{V}\left(T-t+1,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right)-V_t^{\hat{z},\mathrm{in}}\left(\mathbf{a}_{1:t-1};T,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right) \tag{168}$$

$$=\widehat{V}\left(T-t+1,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right)-\max_{a\in\mathcal{A}}\left\{r_t(\mathbf{a}_{1:t-1}\oplus a)-\hat{z}_t(\mathbf{a}_{1:t-1}\oplus a;T,\mathbf{y})+V_{t+1}^{\hat{z},\mathrm{in}}(\mathbf{a}_{1:t-1}\oplus a;T,\mathbf{y})\right\} \tag{169}$$

$$=\min_{a\in\mathcal{A}}\left\{\begin{array}{l}\underbrace{\widehat{V}\left(T-t,\mathbf{y}_t(\mathbf{a}_{1:t})\right)-V_{t+1}^{\hat{z},\mathrm{in}}(\mathbf{a}_{1:t-1}\oplus a;T,\mathbf{y})}_{\geq 0\quad(\because\text{ induction hypothesis})}\\ \underbrace{+\widehat{V}\left(T-t+1,\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right)-\mathbb{E}\left[r_t(\mathbf{a}_{1:t-1}\oplus a)+\widehat{V}\left(T-t,\mathbf{y}_t(\mathbf{a}_{1:t-1}\oplus a)\right)\middle|\mathcal{F}_{t-1}\right]}_{\geq 0\quad(\because\text{ Remark 6})}\end{array}\right\} \tag{170}$$

$$\geq 0. \tag{171}$$

■

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

$$W^{\mathrm{TS}}(T,\mathbf{y})=\mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{P}(\mathbf{y})}\left[T\times\max_{a\in\mathcal{A}}\mu_a(\theta_a)\right] \tag{172}$$

$$=\mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{P}(\mathbf{y})}\left[\max_{a\in\mathcal{A}}\mu_a(\theta_a)\right]+W^{\mathrm{TS}}(T-1,\mathbf{y}) \tag{173}$$

$$\geq\max_{a\in\mathcal{A}}\left\{\mathbb{E}_{\theta_a\sim\mathcal{P}_a(y_a)}\left[\mu_a(\theta_a)\right]+W^{\mathrm{TS}}(T-1,\mathbf{y})\right\} \tag{174}$$

$$=\max_{a\in\mathcal{A}}\left\{\mathbb{E}_{r\sim\mathcal{R}_a(\mathcal{P}_a(y_a))}\left[r+W^{\mathrm{TS}}(T-1,\mathbf{y})\right]\right\} \tag{175}$$

$$=\max_{a\in\mathcal{A}}\left\{\mathbb{E}_{r\sim\mathcal{R}_a(\mathcal{P}_a(y_a))}\left[r+W^{\mathrm{TS}}(T-1,\mathcal{U}(\mathbf{y},a,r))\right]\right\}. \tag{176}$$

The last equality holds since $\mathbb{E}\left[W^{\mathrm{TS}}\left(T-1,\mathcal{U}(\mathbf{y},a_1,r_1(a_1,\omega))\right)\right]=W^{\mathrm{TS}}(T-1,\mathbf{y})$, as argued in (34). By Lemma 2, we have $W^{\text{IRS.V-EMAX}}(T,\mathbf{y})\leq W^{\mathrm{TS}}(T,\mathbf{y})$. ■

D.4. Proof of Theorem 3

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 $\mathbf{y}$.

$$Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a, \omega; T, \mathbf{y}) = r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) - z_t(\mathbf{a}_{1:t-1} \oplus a, \omega; T, \mathbf{y}) + V_{t+1}^{z,\text{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega; T, \mathbf{y}), \tag{177}$$

$$V_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, \omega; T, \mathbf{y}) = \max_{a \in \mathcal{A}} \left\{ Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a, \omega; T, \mathbf{y}) \right\}, \tag{178}$$

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:

$$\mathcal{S}^z(\mathbf{a}_{1:T}, \omega; T, \mathbf{y}) \triangleq \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}), \tag{179}$$

$$a_t^{z,*}(\mathbf{a}_{1:t-1}, \omega; T, \mathbf{y}) \triangleq \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a, \omega; T, \mathbf{y}) \right\}. \tag{180}$$

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,*

$$W^z(T, \mathbf{y}) - V(\pi, T, \mathbf{y}) = \mathbb{E}\left[\max_{\mathbf{a}_{1:T}} \{\mathcal{S}^z(\mathbf{a}_{1:T}, \omega; T, \mathbf{y})\} - \mathcal{S}^z(\mathbf{a}_{1:T}^\pi, \omega; T, \mathbf{y}) \right] \tag{181}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \max_a \left\{ Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}^\pi, a, \omega; T, \mathbf{y}) \right\} - Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}^\pi, a_t^\pi, \omega; T, \mathbf{y}) \right], \tag{182}$$

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:

$$\max_{\mathbf{a}_{1:T}} \{\mathcal{S}^z(\mathbf{a}_{1:T})\} - \mathcal{S}^z(\mathbf{a}_{1:T}^\pi) = \sum_{t=1}^{T} \max_a \left\{ Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a) \right\} - Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}^\pi, a_t^\pi). \tag{183}$$

By taking expectation, we obtain the second equality. ■

Define a shift operator $\mathcal{M}_t : \mathcal{A}^t \times \Omega \mapsto \Omega$,

$$\mathcal{M}_t(\mathbf{a}_{1:t}, \omega) \triangleq (R_{a,n_a}; \forall n_a > n_t(\mathbf{a}_{1:t}, a), \forall a \in \mathcal{A}). \tag{184}$$

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

$$\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega) | \, \mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}, \omega; T, \mathbf{y}) \quad \sim \quad \mathcal{I}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})). \tag{185}$$

Remark 8 (Recursive structure of IRS penalties)**.** *Each of penalty functions* (17)–(21) *has the following form:*

$$z_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}) = \varphi^z(\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega), T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})), \tag{186}$$

for some function $\varphi^z : \Omega \times \mathbb{N} \times \mathcal{Y} \mapsto \mathbb{R}$, 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 7 immediately follows from Bayes' rule, and Remark 8 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* (17)–(21)*, the IRS policy π^z 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.,*

$$\mathbb{P}\left[a_t^{\pi^z} = a \middle| \, \mathcal{F}_{t-1} \right] = \mathbb{P}\left[a_t^{z,*} = a \middle| \, \mathcal{F}_{t-1} \right], \quad \forall a, \quad \forall t. \tag{187}$$

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). We let $a_t^{z,*}$ abbreviate $a_t^{z,*}(\mathbf{a}_{1:t-1}^{\pi^z}, \omega; T, \mathbf{y})$ as defined in (180) *and $\mathcal{F}_{t-1}$ abbreviate $\mathcal{F}_{t-1}(\mathbf{a}_{1:t-1}^{\pi^z}, \omega; T, \mathbf{y})$. Here we assume that the tie-breaking rule in* argmax *of* (180) *is identical to the one used when π^z solves the inner problem.*

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

$$a_t^{z,*}(\mathbf{a}_{1:t-1}, \omega; T, \mathbf{y}) = \operatorname*{argmax}_{a'_t} \left\{ \max_{\mathbf{a}'_{t+1:T}} \sum_{s=t}^{T} r_s(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s}, \omega) - z_s(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s}, \omega; T, \mathbf{y}) \right\}. \tag{188}$$

By Remark 8, for any $s \in [t, T]$,

$$z_s(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s}, \omega; T, \mathbf{y}) \tag{189}$$

$$= \varphi^z(\mathcal{M}_{s-1}(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s-1}, \omega), T - s + 1, \mathbf{y}_{s-1}(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s-1}, \omega; \mathbf{y})) \tag{190}$$

$$= \varphi^z(\mathcal{M}_{s-t}(\mathbf{a}'_{t:s-1}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)), (T - t + 1) - (s - t), \mathbf{y}_{s-t}(\mathbf{a}'_{t:s-1}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega); \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})) \tag{191}$$

$$= z_{s-t+1}(\mathbf{a}'_{t:s}; \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega), T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})). \tag{192}$$

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, (189) is

reformulated as

$$a_t^{z,*} = \operatorname*{argmax}_{a'_t} \left\{ \max_{\mathbf{a}'_{t+1:T}} \sum_{s=t}^{T} r_{s-t+1}(\mathbf{a}'_{t:s}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)) - z_{s-t+1}(\mathbf{a}'_{t:s}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega), T-t+1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})) \right\}. \tag{193}$$

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

$$a_t^{\pi^z} = \operatorname*{argmax}_{a'_1} \left\{ \max_{\mathbf{a}'_{2:T-t+1}} \sum_{s=1}^{T-t+1} r_s(\mathbf{a}'_{1:s}, \tilde{\omega}) - z_s(\mathbf{a}'_{1:s}, \tilde{\omega}, T-t+1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})) \right\}. \tag{194}$$

Comparing (193) and (194), we observe that they have the identical functional forms, except that $\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)$ is replaced with $\tilde{\omega}$. Since $\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega) | \mathcal{F}_{t-1} \sim \mathcal{I}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y}))$ (Remark 7) and $\tilde{\omega} \sim \mathcal{I}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y}))$, it follows that

$$\mathbb{P}\left[a_t^{z,*}(\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)) = a \middle| \mathcal{F}_{t-1} \right] = \mathbb{P}\left[a_t^{\pi^z}(\tilde{\omega}) = a \middle| \mathcal{F}_{t-1} \right]. \tag{195}$$

■

Proof sketch of Theorem 3. For each of penalty functions z_t^{TS}, $z_t^{\textsc{Irs.FH}}$, and $z_t^{\textsc{Irs.V-Zero}}$, we construct a confidence interval process $\{(L_{a,t}, U_{a,t})\}_{a \in \mathcal{A}, t \in [T]}$ such that each of the $(L_{a,t}, U_{a,t})$'s satisfies the following conditions: (i) it is $\mathcal{F}_{t-1}$-measurable and (ii) it regulates the suboptimality of action a at time t; more specifically, (ii) means that the following holds with a high probability $1-\delta$:

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

By Proposition 6,

$$\begin{aligned}
& W^z(T, \mathbf{y}) - V(\pi^z, T, \mathbf{y}) && (196)\\
&= \mathbb{E}\left[\sum_{t=1}^{T} Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}^{\pi}, a_t^{z,*}) - Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}^{\pi}, a_t^{\pi}) \right] && (197)\\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} C \cdot \mathbb{P}_{t-1}[(**) \text{ fails}] + \mathbb{E}_{t-1}\left[Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}^{\pi}, a_t^{z,*}) - Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}^{\pi}, a_t^{\pi}) \middle| (**) \text{ holds} \right] \right] && (198)\\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} C\delta + \mathbb{E}_{t-1}\left[U_{a_t^{z,*},t} - L_{a_t^{\pi},t} \right] \right] && (199)\\
&= TC\delta + \mathbb{E}\left[\sum_{t=1}^{T} U_{a_t^{\pi},t} - L_{a_t^{\pi},t} \right], && (200)
\end{aligned}$$

where C is an almost-sure upper bound on instantaneous suboptimality, $\mathbb{P}_{t-1}[\cdot] \triangleq \mathbb{P}[\cdot | \mathcal{F}_{t-1}]$, and

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

$$\mathbb{E}_{t-1}\left[U_{a_t^{z,*},t}\right] = \sum_{a=1}^{K} U_{a,t} \times \mathbb{P}_{t-1}\left[a_t^{z,*} = a\right] = \sum_{a=1}^{K} U_{a,t} \times \mathbb{P}_{t-1}\left[a_t^{\pi} = a\right] = \mathbb{E}_{t-1}\left[U_{a_t^{\pi},t}\right], \tag{201}$$

by the predictability of $U_{a,t}$ with respect to $\mathbb{F}$ and Proposition 7. Note that (200) accumulates $U_{a_t^{\pi},t} - L_{a_t^{\pi},t}$ over $t = 1, \cdots, T$, each of which is the length of the confidence interval of the action a_t^{π} taken by the policy at each time. We will show that, whenever the policy plays an arm a, the confidence interval of that arm shrinks, and therefore the cumulative suboptimality cannot grow too fast.

Some facts about the Beta-Bernoulli MAB. From now on, we restrict our attention to a Beta-Bernoulli MAB in which $\theta_a \sim \text{Beta}(\alpha_a, \beta_a)$ and $R_{a,n} \sim \text{Bernoulli}(\theta_a)$. Recall that after the DM observes the first n reward realizations, the Bayesian updating yields

$$\theta_a|(R_{a,1}, \cdots, R_{a,n}) \sim \text{Beta}\left(\alpha_a + \sum_{s=1}^{n} R_{a,n}, \beta_a + n - \sum_{s=1}^{n} R_{a,n}\right), \quad \hat{\mu}_{a,n} = \frac{\alpha_a + \sum_{s=1}^{n} R_{a,s}}{\alpha_a + \beta_a + n}. \tag{202}$$

Note that $\{\hat{\mu}_{a,n}\}_{n\geq 0}$ is a martingale that starts from $\hat{\mu}_{a,0} = \frac{\alpha_a}{\alpha_a+\beta_a}$ and converges to $\lim_{n\to\infty} \hat{\mu}_{a,n} = \theta_a$. Roughly speaking, the (unconditional) distribution of $\hat{\mu}_{a,n}$, starting from a point mass $\frac{\alpha_a}{\alpha_a+\beta_a}$, diffuses toward $\text{Beta}(\alpha_a, \beta_a)$, which is the prior distribution[14] of θ_a. In the following lemma, we characterize the distribution of $\hat{\mu}_{a,n}$ more formally.

Lemma 3. *The future Bayesian estimate $\hat{\mu}_{a,n}$ is $\frac{n}{4(\alpha_a+\beta_a)(\alpha_a+\beta_a+n)}$-sub-Gaussian, i.e.,*

$$\mathbb{E}\left[\exp\left(\lambda(\hat{\mu}_{a,n} - \mathbb{E}[\hat{\mu}_{a,n}])\right)\right] \leq \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4(\alpha_a+\beta_a)(\alpha_a+\beta_a+n)}\right), \quad \forall \lambda \in \mathbb{R}. \tag{203}$$

Proof. Since (i) $\mathbb{E}[\hat{\mu}_{a,n}] = \hat{\mu}_{a,0} = \frac{\alpha_a}{\alpha_a+\beta_a}$, (ii) $R_{a,n}$'s are i.i.d. conditioned on θ_a, (iii) $\text{Bernoulli}(\theta_a)$ is $\frac{1}{4}$-sub-Gaussian (for any θ_a), and (iv) $\text{Beta}(\alpha, \beta)$ is $\frac{1}{4(\alpha+\beta+1)}$-sub-Gaussian (Marchal and Arbel, 2017), it follows

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

that, for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda(\hat{\mu}_{a,n} - \hat{\mu}_{a,0})\right)\right] \tag{204}$$

$$= \mathbb{E}\left[\exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times \left((\alpha_a + \sum_{s=1}^{n} R_{a,s}) - (\alpha_a + \beta_a + n)\hat{\mu}_{a,0}\right)\right)\right] \tag{205}$$

$$\overset{\text{(i)}}{=} \mathbb{E}\left[\exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times \left(\sum_{s=1}^{n}(R_{a,s} - \theta_a) + n \cdot (\theta_a - \hat{\mu}_{a,0})\right)\right)\right] \tag{206}$$

$$= \mathbb{E}\left[\mathbb{E}\left\{\exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times \sum_{s=1}^{n}(R_{a,s} - \theta_a)\right)\middle|\, \theta_a\right\} \times \exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times n \cdot (\theta_a - \hat{\mu}_{a,0})\right)\right] \tag{207}$$

$$\overset{\text{(ii)}}{=} \mathbb{E}\left[\mathbb{E}\left\{\exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times (R_{a,1} - \theta_a)\right)\middle|\, \theta_a\right\}^n \times \exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times n \cdot (\theta_a - \hat{\mu}_{a,0})\right)\right] \tag{208}$$

$$\overset{\text{(iii)}}{\leq} \mathbb{E}\left[\left\{\exp\left(\frac{\lambda^2}{2(\alpha_a + \beta_a + n)^2} \times \frac{1}{4}\right)\right\}^n \times \exp\left(\frac{\lambda}{\alpha_a + \beta_a + n} \times n \cdot (\theta_a - \hat{\mu}_{a,0})\right)\right] \tag{209}$$

$$= \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4(\alpha_a + \beta_a + n)^2}\right) \times \mathbb{E}\left[\exp\left(\frac{\lambda n}{\alpha_a + \beta_a + n} \times (\theta_a - \hat{\mu}_{a,0})\right)\right] \tag{210}$$

$$\overset{\text{(iv)}}{\leq} \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4(\alpha_a + \beta_a + n)^2}\right) \times \exp\left(\frac{\lambda^2 n^2}{2(\alpha_a + \beta_a + n)^2} \times \frac{1}{4(\alpha_a + \beta_a + 1)}\right) \tag{211}$$

$$\leq \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4(\alpha_a + \beta_a + n)^2}\right) \times \exp\left(\frac{\lambda^2 n^2}{2(\alpha_a + \beta_a + n)^2} \times \frac{1}{4(\alpha_a + \beta_a)}\right) \tag{212}$$

$$= \exp\left(\frac{\lambda^2}{2} \times \frac{n(\alpha_a + \beta_a) + n^2}{4(\alpha_a + \beta_a + n)^2(\alpha_a + \beta_a)}\right) = \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4(\alpha_a + \beta_a + n)(\alpha_a + \beta_a)}\right). \tag{213}$$

■

(1) Suboptimality analysis of TS (54)**.** Define

$$\Delta_{a,t} \triangleq \sqrt{\frac{\log T}{n^{\pi}_{t-1}(a)}}, \quad U_{a,t} \triangleq \min\left\{\hat{\mu}_{a,n^{\pi}_{t-1}(a)} + \Delta_{a,t}, 1\right\}, \quad L_{a,t} \triangleq \max\left\{\hat{\mu}_{a,n^{\pi}_{t-1}(a)} - \Delta_{a,t}, 0\right\}, \tag{214}$$

where $n^{\pi}_{t-1}(a) \triangleq n_{t-1}(\mathbf{a}^{\pi}_{1:t-1}, a)$ represents how many times the policy π has pulled an arm a before time t. The confidence interval $(L_{a,t}, U_{a,t})$ constructs the high probability lower/upper bounds on $\mu_a(\theta_a)$ $(= \theta_a)$ at time t and it is $\mathcal{F}_{t-1}$-measurable. Conditioned on $\mathcal{F}_{t-1}$, $\mu_a(\theta_a)$ is distributed with Beta$(\alpha_a + \sum_{s=1}^{n^{\pi}_{t-1}(a)} R_{a,s}, \beta_a + n^{\pi}_{t-1}(a) - \sum_{s=1}^{n^{\pi}_{t-1}(a)} R_{a,s})$, which is $\frac{1}{4(\alpha_a+\beta_a+n^{\pi}_{t-1}(a)+1)}$-sub-Gaussian. By Chernoff's inequality,

$$\mathbb{P}_{t-1}\left[\mu_a(\theta_a) \geq U_{a,t}\right] = \mathbb{P}_{t-1}\left[\mu_a(\theta_a) - \hat{\mu}_{a,n^{\pi}_{t-1}(a)} \geq \Delta_{a,t}\right] \tag{215}$$

$$\leq \exp\left(-\frac{\Delta^2_{a,t}}{2 \times \left(4(\alpha_a + \beta_a + n^{\pi}_{t-1}(a) + 1)\right)^{-1}}\right) \tag{216}$$

$$\leq \exp\left(-2n^{\pi}_{t-1}(a) \times \frac{\log T}{n^{\pi}_{t-1}(a)}\right) = \frac{1}{T^2}. \tag{217}$$

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

$$\mathcal{E} \triangleq \left\{\mu_a(\theta_a) \in (L_{a,t}, U_{a,t}), \quad \forall a, \quad \forall t\right\}. \tag{218}$$

By the above concentration inequalities, the sequence of confidence intervals contains the true mean $\mu_a(\theta_a)$ with a very high probability:

$$\mathbb{P}\left[\mathcal{E}^c\right] \leq \mathbb{E}\left[\sum_{a=1}^{K}\sum_{t=1}^{T}\mathbb{P}_{t-1}\left[\mu_a(\theta_a) \geq U_{a,t}\right] + \mathbb{P}_{t-1}\left[\mu_a(\theta_a) \leq L_{a,t}\right]\right] \leq \frac{2K}{T}. \tag{219}$$

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

$$Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}, a_t) = \mu_{a_t}(\theta_{a_t}) + (T-t) \times \mu_{a_t^*}(\theta_{a_t^*}). \tag{220}$$

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

$$Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}, a_t^*) - Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}, a_t) = \mu_{a_t^*}(\theta_{a_t^*}) - \mu_{a_t}(\theta_{a_t}) \leq U_{a_t^*,t} - L_{a_t,t}. \tag{221}$$

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

$$W^{\mathrm{TS}}(T,\mathbf{y}) - V(\pi^{\mathrm{TS}}, T, \mathbf{y}) \leq T \times \mathbb{P}\left[\mathcal{E}^c\right] + \mathbb{E}\left[\sum_{t=1}^{T} U_{a_t^\pi,t} - L_{a_t^\pi,t}\right] \leq 2K + \mathbb{E}\left[\sum_{a=1}^{K}\sum_{t=1}^{T}\min(1, 2\Delta_{a,t}) \cdot \mathbf{1}\{a_t^\pi = a\}\right]. \tag{222}$$

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

$$\sum_{t=1}^{T}\min(1, 2\Delta_{a,t}) \cdot \mathbf{1}\{a_t^\pi = a\} \leq 1 + \sum_{n=2}^{n_T^\pi(a)} 2\sqrt{\frac{\log T}{n-1}} \leq 1 + 2\sqrt{\log T} \times \int_{x=0}^{n_T^\pi(a)} \frac{dx}{\sqrt{x}} \leq 1 + 4\sqrt{\log T} \times \sqrt{n_T^\pi(a)}. \tag{223}$$

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

$$\sum_{a=1}^{K}\left(1 + 4\sqrt{\log T} \times \sqrt{n_T^\pi(a)}\right) \leq K + 4\sqrt{\log T} \times \sqrt{K\sum_{a=1}^{K} n_T^\pi(a)} = K + 4\sqrt{\log T} \times \sqrt{KT}. \tag{224}$$

Combining all the results, we obtain

$$W^{\mathrm{TS}}(T,\mathbf{y}) - V(\pi^{\mathrm{TS}}, T, \mathbf{y}) \leq 3K + 4\sqrt{\log T} \times \sqrt{KT}. \tag{225}$$

■

(2) Suboptimality analysis of Irs.FH (55)**.** Note that $z_t^{\mathrm{Irs.FH}}$ yields

$$Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}, a_t^*) - Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1}, a_t) = \hat{\mu}_{a_t^*, n_{t-1}^\pi(a_t^*)+T-t} - \hat{\mu}_{a_t, n_{t-1}^\pi(a_t)+T-t}. \tag{226}$$

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

For any a and $n \geq 0$, by Lemma 3, we have

$$\mathbb{E}_{t-1}\left[\exp(\lambda(\hat{\mu}_{a,n^\pi_{t-1}(a)+n} - \hat{\mu}_{a,n^\pi_{t-1}(a)}))\right] \leq \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4(\alpha_a+\beta_a+n^\pi_{t-1}(a))(\alpha_a+\beta_a+n^\pi_{t-1}(a)+n)}\right) \tag{227}$$

$$\leq \exp\left(\frac{\lambda^2}{2} \times \frac{n}{4n^\pi_{t-1}(a)(n^\pi_{t-1}(a)+n)}\right). \tag{228}$$

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

Define

$$\Delta_{a,t} \triangleq \sqrt{\frac{T-t}{n^\pi_{t-1}(a)+T-t} \times \frac{\log T}{n^\pi_{t-1}(a)}}, \quad U_{a,t} \triangleq \min\left\{\hat{\mu}_{a,n^\pi_{t-1}(a)} + \Delta_{a,t}, 1\right\}, \quad L_{a,t} \triangleq \max\left\{\hat{\mu}_{a,n^\pi_{t-1}(a)} - \Delta_{a,t}, 0\right\}. \tag{229}$$

By Chernoff's inequality,

$$\mathbb{P}_{t-1}\left[\hat{\mu}_{a_t^*,n^\pi_{t-1}(a_t^*)+T-t} \geq U_{a,t}\right] = \mathbb{P}_{t-1}\left[\hat{\mu}_{a,T-1} - \hat{\mu}_{a,n^\pi_{t-1}(a)} \geq \Delta_{a,t}\right] \tag{230}$$

$$\leq \exp\left(-\frac{\Delta_{a,t}^2}{2 \times \frac{T-t}{4n^\pi_{t-1}(a)(T-t+n^\pi_{t-1}(a))}}\right) = \exp\left(-2\log T\right) = \frac{1}{T^2}. \tag{231}$$

Similarly, we can show that $\mathbb{P}_{t-1}\left[\hat{\mu}_{a_t^*,n^\pi_{t-1}(a_t^*)+T-t} \leq L_{a,t}\right] \leq \frac{1}{T^2}$.

Analogously to the proof of TS, we can show that

$$W^{\mathrm{TS}}(T,\mathbf{y}) - V(\pi^{\mathrm{TS}},T,\mathbf{y}) \leq 2K + \mathbb{E}\left[\sum_{a=1}^K \sum_{t=1}^T \min(1, 2\Delta_{a,t}) \cdot \mathbf{1}\{a_t^\pi = a\}\right]. \tag{232}$$

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

$$\Delta_{a,t} \leq \sqrt{\left(1 - \frac{n^\pi_{t-1}(a)}{T}\right) \times \frac{\log T}{n^\pi_{t-1}(a)}} = \sqrt{\log T} \times \sqrt{\frac{1}{n^\pi_{t-1}(a)} - \frac{1}{T}} \leq \sqrt{\log T} \times \left(\frac{1}{\sqrt{n^\pi_{t-1}(a)}} - \frac{\sqrt{n^\pi_{t-1}(a)}}{2T}\right). \tag{233}$$

Consequently, for each a,

$$\sum_{t=1}^T \min(1, 2\Delta_{a,t}) \cdot \mathbf{1}\{a_t^\pi = a\} \leq 1 + 2\sqrt{\log T} \times \sum_{n=2}^{n^\pi_T(a)} \left(\frac{1}{\sqrt{n-1}} - \frac{\sqrt{n-1}}{2T}\right) \tag{234}$$

$$\leq 1 + 2\sqrt{\log T} \times \int_{x=0}^{n^\pi_T(a)} \left(\frac{1}{\sqrt{x}} - \frac{\sqrt{x}}{2T}\right) \tag{235}$$

$$= 1 + 2\sqrt{\log T} \times \left(2\sqrt{n^\pi_T(a)} - \frac{(n^\pi_T(a))^{3/2}}{3T}\right). \tag{236}$$

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

$$\sum_{a=1}^K (n^\pi_T(a))^{3/2} \geq \sum_{a=1}^K \left(\frac{T}{K}\right)^{3/2} = \sqrt{T^3/K}. \tag{237}$$

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

$$W^{\mathrm{TS}}(T,\mathbf{y}) - V(\pi^{\mathrm{TS}},T,\mathbf{y}) \le 2K + \mathbb{E}\left[\sum_{a=1}^{K} 1 + 2\sqrt{\log T}\times\left(2\sqrt{n_T^\pi(a)} - \frac{(n_T^\pi(a))^{3/2}}{3T}\right)\right] \tag{238}$$

$$\le 3K + 2\sqrt{\log T}\times\left(2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right). \tag{239}$$

■

(3) Suboptimality analysis of Irs.V-Zero (56)**.** Consider an optimal allocation $\mathbf{n}^*$ of the inner problem of Irs.V-Zero when the remaining time is T. For an arm a on which the optimal solution allocates at least one pull, i.e., $n^*(a) > 0$, a policy does not incur suboptimality by pulling the arm a (the arms that $n^*(a) > 0$ are all optimal and their Q-values tie). A policy incurs suboptimality only when pulling an arm a such that $n^*(a) = 0$, in which case we lose $\min_{a':n^*(a')>0}\{\hat{\mu}_{a',n^*(a')-1}\} - \hat{\mu}_{a,0}$ (we lose the last pull of one of the optimal arms) where the term $\min_{a':n^*(a')>0}\{\hat{\mu}_{a',n^*(a')-1}\}$ is limited by $\max_{0\le n\le T-1}\hat{\mu}_{a^*,n}$ for some a^* such that $n^*(a^*) > 0$. Extending this argument, at a certain time t, when the remaining time is $T-t+1$, we have

$$Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1},a_t^*) - Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1},a_t) \le \max_{0\le n\le T-t}\left\{\hat{\mu}_{a_t^*,n_{t-1}^\pi(a_t^*)+n}\right\} - \hat{\mu}_{a_t,n_{t-1}^\pi(a_t)}. \tag{240}$$

We need to regulate $\max_{0\le n\le T-t}\left\{\hat{\mu}_{a_t^*,n_{t-1}^\pi(a_t^*)+n}\right\}$. As before, we define

$$\Delta_{a,t} \triangleq \sqrt{\frac{T-t}{n_{t-1}^\pi(a)+T-t}\times\frac{\log T}{n_{t-1}^\pi(a)}}, \quad U_{a,t} \triangleq \min\left\{\hat{\mu}_{a,n_{t-1}^\pi(a)} + \Delta_{a,t}, 1\right\}, \quad L_{a,t} \triangleq \hat{\mu}_{a,n_{t-1}^\pi(a)}. \tag{241}$$

Note that we take $L_{a,t}$ that are different from those in the previous case, but still $\mathcal{F}_{t-1}$-measurable. Given that $\{\hat{\mu}_{a,n_{t-1}^\pi(a)+n} - \hat{\mu}_{a,n_{t-1}^\pi(a)}\}_{n\ge 0}$ is a martingale, $\left\{\exp\left(\lambda(\hat{\mu}_{a,n_{t-1}^\pi(a)+n} - \hat{\mu}_{a,n_{t-1}^\pi(a)})\right)\right\}_{n\ge 0}$ is a non-negative supermartingale due to the convexity of $\exp(\cdot)$. By Doob's maximal inequality and Lemma 3, for any $\lambda \ge 0$,

$$\mathbb{P}_{t-1}\left[\max_{0\le n\le T-t}\left\{\hat{\mu}_{a,n_{t-1}^\pi(a)+n}\right\} \ge U_{a,t}\right] = \mathbb{P}_{t-1}\left[\max_{0\le n\le T-t}\left\{\hat{\mu}_{a,n_{t-1}^\pi(a)+n} - \hat{\mu}_{a,n_{t-1}^\pi(a)}\right\} \ge \Delta_{a,t}\right] \tag{242}$$

$$\le \mathbb{P}_{t-1}\left[\max_{0\le n\le T-t}\left\{\exp\left(\lambda(\hat{\mu}_{a,n_{t-1}^\pi(a)+n} - \hat{\mu}_{a,n_{t-1}^\pi(a)})\right)\right\} \ge \exp\left(\lambda\Delta_{a,t}\right)\right] \tag{243}$$

$$\le \frac{\mathbb{E}_{t-1}\left[\exp\left(\lambda(\hat{\mu}_{a,n_{t-1}^\pi(a)+T-t} - \hat{\mu}_{a,n_{t-1}^\pi(a)})\right)\right]}{\exp(\lambda\Delta_{a,t})} \tag{244}$$

$$\le \exp\left(\frac{\lambda^2}{2}\times\frac{T-t}{4n_{t-1}^\pi(a)(n_{t-1}^\pi(a)+T-t)} - \lambda\Delta_{a,t}\right). \tag{245}$$

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

$$\mathbb{P}_{t-1}\left[\max_{0\le n\le T-t}\left\{\hat{\mu}_{a,n_{t-1}^\pi(a)+n}\right\} \ge U_{a,t}\right] \le \exp\left(-\frac{2n_{t-1}^\pi(a)(n_{t-1}^\pi(a)+T-t)}{T-t}\times\Delta_{a,t}^2\right) = \frac{1}{T^2}. \tag{246}$$

Note that $\max_{0\le n\le T-t}\left\{\hat{\mu}_{a,n_{t-1}^\pi(a)+n}\right\} \ge L_{a,t} \equiv \hat{\mu}_{a,n_{t-1}^\pi(a)}$ by defintion. We have shown that

$$\mathbb{P}\left[\mathcal{E} \triangleq \left\{\max_{0\le n\le T-t}\left\{\hat{\mu}_{a,n_{t-1}^\pi(a)+n}\right\} \in [L_{a,t},U_{a,t}), \quad \forall a,\forall t\right\}\right] \ge 1 - \frac{K}{T}. \tag{247}$$

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

$$W^{\mathrm{TS}}(T,\mathbf{y}) - V(\pi^{\mathrm{TS}},T,\mathbf{y}) \leq T\mathbb{P}[\mathcal{E}^c] + \mathbb{E}\left[\sum_{t=1}^{T} U_{a_t^\pi,t} - L_{a_t^\pi,t}\right] \tag{248}$$

$$\leq K + \mathbb{E}\left[\sum_{a=1}^{K}\sum_{t=1}^{T} \min(1,\Delta_{a,t})\mathbf{1}\{a_t^\pi = a\}\right] \tag{249}$$

$$\leq 2K + \sqrt{\log T} \times \left(2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right). \tag{250}$$

■