Electronic Companion of "Thompson Sampling with Information Relaxation Penalties"

Seungki Min
KAIST
skmin@kaist.ac.kr

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

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

Current Revision: February 2022

A. An Illustrative Example

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

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

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

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

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

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

*The authors wish to thank Daniel Russo, Martin Haugh, David Brown, Jim Smith, and anonymous reviewers for helpful comments. A preliminary version of this paper appeared in the conference proceedings Advances in Neural Information Processing Systems 32 (NeurIPS 2019) (Min et al., 2019).

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

A.1. Inner Problems Induced by Different Penalty Functions

No penalty. To clarify the role of penalties, we first consider the case of zero penalty, i.e., $z_t \equiv 0$, which was not discussed in §3. With zero penalty, the DM at any time earns the current realized reward without adjustment. The clairvoyant DM, who is informed of the outcome ω, can find the best action sequence for this particular outcome ω. Recall that $R_{a,n}$ is defined to be the reward from the 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 μ_a whenever playing arm a. The hindsight optimal action sequence is to pull arm 3 (the arm with the largest mean reward μ_a) eight times in a row and the DM can earn a total reward of $T \times \mu_3 = 6.296$ at most.

	Payoffs under z_t^{TS}								**Maximal payoff**
	$n=1$	2	3	4	5	6	7	8	
Arm 1	.235	.235	.235	.235	.235	.235	.235	.235	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 arm a. Recall that $\hat{\mu}_{a,T-1}(\omega)$ is the Bayesian estimate on mean reward of arm a after observing reward realizations $R_{a,1}, \cdots, R_{a,T-1}$. In this particular example, we have $(\hat{\mu}_{1,T-1}, \hat{\mu}_{2,T-1}, \hat{\mu}_{3,T-1}) = \left(\frac{6}{11}, \frac{6}{9}, \frac{6}{11}\right)$ and the maximal payoff is $T \times \hat{\mu}_{2,T-1} = 5.333$, which can be obtained by playing arm 2 throughout the entire time horizon.

IRS.V-Zero penalty. Finally, let us focus on $z_t^{\text{IRS.V-ZERO}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \hat{\mu}_{a_t, n_{t-1}(\mathbf{a}_{1:t-1}, a_t)}$ under which the DM earns $\hat{\mu}_{a,n-1}(\omega)$ from the 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 $z_t^{\text{IRS.FH}}$								Maximal payoff
	$n=1$	2	3	4	5	6	7	8	
Arm 1	6/11	6/11	6/11	6/11	6/11	6/11	6/11	6/11	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	

	Payoffs under $z_t^{\text{IRS.V-ZERO}}$								Maximal payoff
	$n=1$	2	3	4	5	6	7	8	
Arm 1	3/4	3/5	4/6	5/7	6/8	6/9	6/10	6/11	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[1] ω.

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

A.2. IRS Performance Bounds

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

W^0	W^{TS}	$W^{\text{IRS.FH}}$	$W^{\text{IRS.V-ZERO}}$	$W^{\text{IRS.V-EMAX}}$	$W^{\text{ideal}} = V^*$
6.805	6.429	6.279	6.111	6.075	6.063

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

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

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

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

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

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

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

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

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

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

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

B. Algorithms in Detail

B.1. Implementation of IRS.V-Zero

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

Algorithm 5: Arm selection rule of $\pi^{\text{IRS.V-ZERO}}$ when remaining time is T and current belief is $\mathbf{y}$

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

$\tilde{\theta}_a \sim \mathcal{P}_a(y_a), \tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}), \quad \forall n \in \{1, \dots, T\}, \forall a \in \{1, \dots, K\}$
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
$\tilde{M}_{0,0} \leftarrow 0, \tilde{M}_{0,n} \leftarrow -\infty, \forall n \in \{1, \dots, T\}$
for $a = 1, \cdots, K$ **do**
  for $n = 0, \cdots, T$ **do**
    $\tilde{M}_{a,n} \leftarrow \max_{0 \le m \le n}\{\tilde{M}_{a-1,n-m} + \tilde{S}_{a,m}\}$
    $\tilde{L}_{a,n} \leftarrow \operatorname{argmax}_{0 \le m \le n}\{\tilde{M}_{a-1,n-m} + \tilde{S}_{a,m}\}$
  end
end
$\tau \leftarrow T$
for $a = K, \cdots, 1$ **do**
  $\tilde{n}^*_a \leftarrow \tilde{L}_{a,\tau}$
  $\tau \leftarrow \tau - \tilde{n}^*_a$
end
return $\operatorname{argmax}_a \tilde{n}^*_a$

B.2. Implementation of IRS.V-EMax

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

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

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 $\mathbb{E}\left[W^{\mathrm{TS}}\left(T-t, \mathbf{y}_t\right) \middle| H_{t-1}\right] = W^{\mathrm{TS}}\left(T-t, \mathbf{y}_{t-1}\right)$.

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

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

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

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

i.e., when letting a^* be the maximizer of (72), it is optimal to play arm a^* after making the best effort within the allocation $\mathbf{n}_{1:K} - \mathbf{e}_a$. For all feasible counts $\mathbf{n}_{1:K}$'s such that $\sum_{a=1}^K n_a \leq T$, we can compute $M(\mathbf{n}_{1:K}, \omega)$'s by sequentially solving (72) in an appropriate order. By doing so, we can obtain the maximal value of the original inner problem (∗) by evaluating

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

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

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

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

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

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

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

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

end

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

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

end

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

for *each* $\mathbf{n}_{1:K} \in N_{\leq T} \setminus \{\mathbf{0}\}$ *in order with increasing* $\sum_{a=1}^K n_a$ **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 $W^{\text{TS}}(T, \mathbf{y}) = T \times \mathbb{E}_{\mathbf{y}}\left[\max_a \mu_a(\theta_a)\right]$ 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

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

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

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

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

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

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

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

$$\sum_{n=1}^{m} (T-t_n) \times \left(\Gamma^\lambda_n - \Gamma^\lambda_{n-1}\right) \tag{76}$$

$$= \sum_{n=1}^{m} (T-t_n) \times \Gamma^\lambda_n - \sum_{n=1}^{m} (T-t_n) \times \Gamma^\lambda_{n-1} \tag{77}$$

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

$$= \sum_{n=0}^{m} (T-t_n) \times \Gamma^\lambda_n - (T-t_0) \times \Gamma^\lambda_0 - \sum_{n=0}^{m} (T-t_{n+1}) \times \Gamma^\lambda_n + (T-t_{m+1}) \times \Gamma^\lambda_m \tag{79}$$

$$= -\Gamma^\lambda_m - T \times \Gamma^\lambda_0 + \sum_{n=0}^{m} (t_{n+1} - t_n) \times \Gamma^\lambda_n. \tag{80}$$

Consider the minimum value among $\Gamma^\lambda_0, \ldots, \Gamma^\lambda_m$ and let $n^* \triangleq \operatorname{argmin}_{0 \leq n \leq m} \Gamma^\lambda_n$. In order to minimize (80), it should satisfy that $t_{n+1} - t_n = T - m + 1$ for $n = n^*$ and $t_{n+1} - t_n = 1$ for $n \neq n^*$. For such t_n's, (75) reduces to

$$\sum_{n=1}^{m} \hat{\mu}_{a,n-1} + (T-m) \times \lambda - \left(-\Gamma^\lambda_m - T \times \Gamma^\lambda_0 + \sum_{n=0}^{m} \Gamma^\lambda_n + (T-m) \times \min_{0 \leq n \leq m} \Gamma^\lambda_m\right) \tag{81}$$

$$= \sum_{n=1}^{m} \hat{\mu}_{a,n-1} + (T-m) \times \left(\lambda - \min_{0 \leq n \leq m} \Gamma^\lambda_m\right) + T \times \Gamma^\lambda_0 - \sum_{n=0}^{m-1} \Gamma^\lambda_n. \tag{82}$$

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

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

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

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

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

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

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

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

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

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

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

end

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

C. Proofs for §3

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

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

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

C.1. Proof of Theorem 1

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

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

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

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

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

where the inequality holds since $\Pi_{\mathbb{F}} \subseteq \Pi_{\mathbb{G}}$. ■

Strong duality. Fix T and $\mathbf{y}$. Let $V_t^{\text{in}}(\mathbf{a}_{1:t-1}, \omega)$ and $Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega)$ be, respectively, the value function and the state-action value (Q-value) function that are associated with the inner problem (∗) given a particular outcome ω under the ideal penalty (22). With $V_{T+1}^{\text{in}} \equiv 0$, we have the following Bellman equation for the inner problem:

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

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

We argue by induction to show that

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

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

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

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

$t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega))$ for all $\mathbf{a}_{1:t} \in \mathcal{A}^t$. For any $\mathbf{a}_{1:t-1} \in \mathcal{A}^{t-1}$ and $a \in \mathcal{A}$, then,

$$
\begin{aligned}
Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) &= r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) - z_t^{\text{ideal}}(\mathbf{a}_{1:t-1} \oplus a, \omega) + V_{t+1}^{\text{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega) && (91)\\
&= \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) + V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega)\right) \middle| H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right] && (92)\\
&\quad \underbrace{-V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega)\right) + V_{t+1}^{\text{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega)}_{=0} && (93)\\
&= \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) + V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega)\right) \middle| H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right] && (94)\\
&= \mathbb{E}_{\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)}\left[R_a + V^*\left(T - t, \mathcal{U}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a, R_a)\right) \right] && (95)\\
&= Q^*(T - t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a), && (96)
\end{aligned}
$$

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

$$
\begin{aligned}
V_t^{\text{in}}(\mathbf{a}_{1:t-1}, \omega) &= \max_{a \in \mathcal{A}} \left\{ Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) \right\} && (97)\\
&= \max_{a \in \mathcal{A}} \left\{ Q^*(T - t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a) \right\} && (98)\\
&= V^*(T - t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)). && (99)
\end{aligned}
$$

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

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

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

C.2. Proof of Remark 2

We proceed on the proof of strong duality. The policy π^{ideal} solves the same inner problem with respect to a randomly sampled outcome $\tilde{\omega}$. When the remaining time is T and the current belief is $\mathbf{y}$, it takes an action with the largest Q-value: together with (100), it yields

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

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

C.3. Proof of Remark 3

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

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

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

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

and thus z_t^{TS} is dual feasible.

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

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

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

D. Proofs for §4

D.1. Notes on Regularity

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

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

Proof. By Jensen's inequality,

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

Consequently,

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

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

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

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

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

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 $\pi^{\text{IRS.FH}}$. Let $\tilde{\omega}$ be the sampled outcome used by $\pi^{\text{IRS.FH}}$. By Proposition 5, we have $\lim_{n\to\infty} \hat{\mu}_{a,n}(\tilde{\omega}) = \mu_a(\tilde{\theta}_a)$ for almost all $\tilde{\omega}$. This, together with the assumption that $\mu_i(\theta_i) \neq \mu_j(\theta_j)$ for $i \neq j$, since $\operatorname{argmax}_a \mu_a(\tilde{\theta}_a)$ is uniquely defined for almost all $\tilde{\omega}$, yields

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

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

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

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

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

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

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

Define

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

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

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

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

For such N, we have

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

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

To argue by contradiction, suppose that $A^{\circ}_T \neq A^{\mathrm{TS}}$ 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:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

D.3. Proof of Theorem 2

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

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

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

■

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

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

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

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

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

By setting $t=x$, we get

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

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

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

Combining this with (127), we get

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

For random variable X, by taking expectation, we get

$$\begin{aligned}\mathbb{E}\left[\max\{X+\varphi(X),0\}\right]&\geq\mathbb{E}\left[\max\{\mu+\varphi(X),0\}+(X-\mu)\cdot\mathbf{1}\{\mu+\varphi(\mu)\geq 0\}\right] &(130)\\ &\geq\mathbb{E}\left[\max\{\mu+\varphi(X),0\}\right]+\mathbb{E}(X-\mu)\cdot\mathbf{1}\{\mu+\varphi(\mu)\geq 0\} &(131)\\ &=\mathbb{E}\left[\max\{\mu+\varphi(X),0\}\right]. &(132)\end{aligned}$$

■

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

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

Proof. Define

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

By (129), we have

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

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

$$\begin{aligned}\mathbb{E}\left[\max\{X(\omega)+\varphi(X(\omega),\omega),0\}\right]&\geq\mathbb{E}\left[\max\{\mu(\omega)+\varphi(X(\omega),\omega),0\}+(X(\omega)-\mu(\omega))\cdot I(\omega)\right] &(136)\\ &=\mathbb{E}\left[\mathbb{E}\left(\max\{\mu(\omega)+\varphi(X(\omega),\omega),0\}+(X(\omega)-\mu(\omega))\cdot I(\omega)|\,\mathcal{H}\right)\right] &(137)\\ &=\mathbb{E}\left[\max\{\mu(\omega)+\varphi(X(\omega),\omega),0\}\right]+\mathbb{E}\left[\left((X(\omega)-\mu(\omega))\cdot I(\omega)|\,\mathcal{H}\right)\right] &(138)\\ &=\mathbb{E}\left[\max\{\mathbb{E}\left(X|\mathcal{H}\right)(\omega)+\varphi(X(\omega),\omega),0\}\right] &(139)\\ &\quad+\mathbb{E}\left[\underbrace{\left(\mathbb{E}(X|\mathcal{H})(\omega)-\mu(\omega)\right)}_{=0}\cdot I(\omega)\right] &(140)\\ &=\mathbb{E}\left[\max\{\mathbb{E}\left(X|\mathcal{H}\right)(\omega)+\varphi(X(\omega),\omega),0\}\right]. &(141)\end{aligned}$$

■

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

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

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

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

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

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

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

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

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

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

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

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

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

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

■

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

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

What we want to show is

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

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

Further define

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

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

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

Fix k and n. Define a sub-σ-field

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

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

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

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

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

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

$$\begin{aligned}
U_{k,n} &= \mathbb{E}\left[\max_{0\leq i\leq T}\left\{(i-n)^+ \times X + C_i\right\}\right] && (159)\\
&\overset{\text{Corollary 2}}{\geq} \mathbb{E}\left[\max_{0\leq i\leq T}\left\{\mathbb{E}(X|\mathcal{H})\cdot\mathbf{1}\{i\geq n+1\} + (i-n-1)^+ \times X + C_i\right\}\right] && (160)\\
&\overset{\text{(a)}}{=} \mathbb{E}\left[\max_{0\leq i\leq T}\left\{\hat{\mu}_{k,n}\cdot\mathbf{1}\{i\geq n+1\} + (i-n-1)^+ \times \hat{\mu}_{a,T-1} + C_i\right\}\right] && (161)\\
&\overset{\text{(b)}}{=} U_{k,n+1}. && (162)
\end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

for brevity, we get

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

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

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

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

■

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

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

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

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

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

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

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

D.4. Proof of Theorem 3

D.4.1. Suboptimality Decomposition

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

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

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

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

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

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

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

Proposition 6 (Suboptimality decomposition)**.** *Given a non-anticipating policy $\pi \in \Pi_{\mathbb{F}}$ and a dual-feasible penalty function z_t, the suboptimality gap is the sum of the instantaneous suboptimalities of individual actions taken by π along the sample path: i.e.,*

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

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

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

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

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

By taking expectation, we obtain the second equality. ■

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

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

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

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

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

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

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

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

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

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

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

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

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

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

D.4.2. Recursive Structure of IRS Penalty Functions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
& z_s(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s}, \omega; T, \mathbf{y}) && (199)\\
&= \varphi^z(\mathcal{M}_{s-1}(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s-1}, \omega), T-s+1, \mathbf{y}_{s-1}(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s-1}, \omega; \mathbf{y})) && (200)\\
&= \varphi^z \begin{pmatrix} \mathcal{M}_{s-t}(\mathbf{a}'_{t:s-1}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)), \\ (T-t+1)-(s-t), \\ \mathbf{y}_{s-t}(\mathbf{a}'_{t:s-1}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega); \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y}) \end{pmatrix} && (201)\\
&= z_{s-t+1}(\mathbf{a}'_{t:s}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega); T-t+1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})). && (202)
\end{aligned}$$

For rewards, similarly, we have $r_s(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s}, \omega) = r_{s-t+1}(\mathbf{a}'_{t:s}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega))$. Therefore, the

sub-inner problem (198) is reformulated as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

for some $\sigma > 0$.

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

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

Then, the following holds:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

D.4.4. Proof of Theorem 3

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

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

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

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

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

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

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

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

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

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\mu_t^U(A_t^*) - \mu_t^L(A_t^\pi)\right] && (242)\\
&= \mathbb{E}_t\left[\mu_t^U(A_t^*) - \mu_t^L(A_t^\pi)\right] + \mathbb{E}_t\left[U_t(A_t^\pi) - U_t(A_t^*)\right] + \mathbb{E}_t\left[L_t(A_t^\pi) - L_t(A_t^\pi)\right] && (243)\\
&= \mathbb{E}_t\left[\mu_t^U(A_t^*) - U_t(A_t^*)\right] + \mathbb{E}_t\left[L_t(A_t^\pi) - \mu_t^L(A_t^\pi)\right] + \mathbb{E}_t\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right] && (244)\\
&\leq \mathbb{E}_t\left[\left(\mu_t^U(A_t^*) - U_t(A_t^*)\right)^+\right] + \mathbb{E}_t\left[\left(L_t(A_t^\pi) - \mu_t^L(A_t^\pi)\right)^+\right] + \mathbb{E}_t\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right]. && (245)
\end{aligned}$$

We further observe that

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

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

$$\begin{aligned}
W(T,\mathbf{y}) - V(\pi,T,\mathbf{y}) &\overset{\text{Prop 6}}{=} \mathbb{E}\left[\sum_{t=1}^T Q_t^{z,\text{in}}(A_t^*) - Q_t^{z,\text{in}}(A_t^\pi)\right] && (247)\\
&\le \mathbb{E}\left[\sum_{t=1}^T \mu_t^U(A_t^*) - \mu_t^L(A_t^\pi)\right] && (248)\\
&= \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_t\left[\mu_t^U(A_t^*) - \mu_t^L(A_t^\pi)\right]\right] && (249)\\
&\le \mathbb{E}\left[\sum_{t=1}^T \left(\frac{C_U}{T} + \frac{C_L}{T} + \mathbb{E}_t\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right]\right)\right] && (250)\\
&\le C_U + C_L + \sum_{t=1}^T \mathbb{E}\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right]. && (251)
\end{aligned}$$

■

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

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

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

Further define

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

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

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

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

We construct the confidence intervals as follows:

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

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

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

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

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

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

Further observe that

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

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

By utilizing Cauchy–Schwartz inequality, we deduce that

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

Combining all these results, we conclude that

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

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

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

We construct the confidence intervals as follows:

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

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

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

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

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

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

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

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

Consequently,

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

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

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

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

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

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

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

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

Combining all these results, we conclude that

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

$$\le 2\sqrt{L}\left[\frac{1}{\sqrt{\nu}} + \sqrt{2\log T}\left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right)\right]. \tag{277}$$

(3) Suboptimality analysis for Irs.V-Zero (62)**.** As discussed in Remark 11, for Irs.FH, we have

$$Q_t^{z,\text{in}}(a_t^{z,*}) - Q_t^{z,\text{in}}(a) = \max_{0\le n\le T-t}\{\hat{\mu}_t^{\pi}(a_t^{z,*},n)\} - \hat{\mu}_t^{\pi}(a,0). \tag{278}$$

We construct the confidence intervals as follows:

$$U_t(a) \triangleq \hat{\mu}_t^{\pi}(a,0) + \sqrt{2\log T}\times\Delta_t^{\pi}(a,T-t), \quad L_t(a) \triangleq \hat{\mu}_t^{\pi}(a,0). \tag{279}$$

By Lemma 6, we have

$$\mathbb{E}\left[\left(\max_{0\le n\le T-t}\hat{\mu}_t^{\pi}(a,n) - U_t(a)\right)^{+}\middle|\,\mathcal{F}_{t-1}\right] \le \frac{\Delta_t^{\pi}(a,T-t)}{\sqrt{2\log T}}e^{-\frac{2\log T}{2}} \le \frac{\sqrt{L/\nu}}{T}, \tag{280}$$

where

$$\mathbb{E}\left[\hat{\mu}_t^{\pi}(a,0) - L_t(a)\middle|\,\mathcal{F}_{t-1}\right] = 0. \tag{281}$$

The rest of the proof is almost identical to the case of Irs.FH:

$$W^{\textsc{Irs.V-Zero}}(T,\mathbf{y}) - V(\pi^{\textsc{Irs.V-Zero}},T,\mathbf{y}) \le \sqrt{\frac{L}{\nu}} + \sum_{t=1}^{T}\mathbb{E}\left[U_t(A_t^{\pi}) - L_t(A_t^{\pi})\right]. \tag{282}$$

$$= \sqrt{\frac{L}{\nu}} + \sqrt{2\log T}\sum_{t=1}^{T}\Delta_t^{\pi}(A_t^{\pi},T-t) \tag{283}$$

$$\le \sqrt{L}\left[\frac{1}{\sqrt{\nu}} + \sqrt{2\log T}\left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right)\right]. \tag{284}$$