Continuous-in-time Limit for Bayesian Bandits

Yuhua Zhu   YUZ244@UCSD.EDU
Department of Mathematics and Halıcıoğlu Data Science Institute
University of California - San Diego
La Jolla, CA 92093, USA

Zachary Izzo   ZIZZO@STANFORD.EDU
Department of Mathematics
Stanford University
Stanford, CA 94305, USA

Lexing Ying   LEXING@STANFORD.EDU
Department of Mathematics
Stanford University
Stanford, CA 94305, USA

Editor: Qiang Liu

Abstract

This paper revisits the bandit problem in the Bayesian setting. The Bayesian approach formulates the bandit problem as an optimization problem, and the goal is to find the optimal policy which minimizes the Bayesian regret. One of the main challenges facing the Bayesian approach is that computation of the optimal policy is often intractable, especially when the length of the problem horizon or the number of arms is large. In this paper, we first show that under a suitable rescaling, the Bayesian bandit problem converges toward a continuous Hamilton-Jacobi-Bellman (HJB) equation. The optimal policy for the limiting HJB equation can be explicitly obtained for several common bandit problems, and we give numerical methods to solve the HJB equation when an explicit solution is not available. Based on these results, we propose an approximate Bayes-optimal policy for solving Bayesian bandit problems with large horizons. Our method has the added benefit that its computational cost does not increase as the horizon increases.

Keywords: Bayesian bandit, Hamilton-Jacobi-Bellman equation, optimal control, dynamic view, diffusion approximation, continuous limit

1. Introduction

Bandit problems were first introduced by Thompson (1933) with later pioneering work due to Robbins (1952) and Wald (2004). In more recent years, bandit algorithms have become widely adopted for automated decision-making tasks such as dynamic pricing (Ferreira et al., 2018), mobile health (Tewari and Murphy, 2017), Alpha Go (Silver et al., 2016), etc.

The bandit problem can be considered from one of two perspectives: Bayesian or frequentist. The Bayesian approach dominated bandit research from 1960-1980 (Bradt et al., 1956; Gittins, 1979). The objective is to minimize an average cumulative regret with respect to the Bayesian prior measure of the problem environment. It formulates the bandit problem as an optimization problem, and the goal is to find the optimal policy which minimizes

the Bayesian regret. In the frequentist setting (Lai and Robbins, 1985), the cumulative regret is viewed as an unknown deterministic quantity, and the goal is to design policies to achieve the best environment-dependent performance.

The main difficulty with Bayesian bandits is that computation of the optimal policy is often intractable, especially when the number of arms or the horizon is large. Gittin's index (Gittins, 1979) reduced the computational cost for the discounted infinite horizon setting but does not apply to undiscounted cases (Berry and Fristedt, 1985). Although computing the Bayes-optimal policy is challenging, there is a significant payoff: the performance of the policy is not only optimal in the Bayesian setting (by definition) but also has favorable frequentist regret guarantees (Lattimore, 2016). In addition, Bayesian bandits have been widely employed in economics (Bergemann and Valimaki, 2006), including in contract theory (Gur et al., 2022), dynamic pricing (Leloup and Deveaux, 2001), portfolio management (El Karoui et al., 2005), etc.

In this paper, we revisit the Bayesian perspective for the multi-armed bandit problem and analyze it using tools from PDEs. A continuous-in-time limiting HJB equation is derived as the horizon n goes to infinity for a range of bandit problems. Based on the limiting equation, a regularized Bayes-optimal policy is proposed, where regularization is employed to increase exploration and stability. Numerical schemes can be used to approximate the optimal policy, leading to improved computational efficiency when the horizon is large. In addition, the exact optimal policy for the limiting HJB equation can be obtained for certain types of bandit problems, including the classical Bernoulli and Gaussian arm reward cases, resulting in an efficient algorithm to approximate the optimal policy even if the number of arms is large. In summary, our contributions are as follows.

- We derive a continuous-in-time limit, an HJB equation, for the Bayesian multi-armed bandit problem.

- We propose a regularized version of the Bayesian multi-armed bandit problem, which encourages exploration and smooths the optimal policy.

- Based on the limiting PDE, we give an efficient algorithm for approximating the optimal policy.

Recently, the use of differential equations to analyze machine learning algorithms has received growing interest, especially in optimization algorithms Su et al. (2014); Li et al. (2017), sampling algorithms Welling and Teh (2011); Liu (2017), neural networks Mei et al. (2018); Chen et al. (2018); Han et al. (2019); Chen et al. (2020). However, fewer connections have been built between multi-armed bandits and differential equations until the recent two years. For bandit problems, Fan and Glynn (2021); Wager and Xu (2021); Kobzar and Kohn (2022) model several policies in the frequentist setting via a continuous SDE or PDE, and this continuous analysis is used to provide insights into the properties of the algorithms studied. In the Bayesian setting, Araman and Caldentey (2022); Che and Hörner (2018) give differential equation approximations, but their work can only be applied to settings with two possible environments. In this paper, we consider Bayesian bandits with general environments. There are also related works using differential equations in online learning settings, including contextual bandits (Kapralov and Panigrahy, 2011), drifting games (Wang and Kohn, 2022), etc.

2. Bayesian Bandits

Throughout the paper, we focus on K-armed stochastic bandits played over n rounds, where $n \in \mathbb{Z}_+$ is a positive integer called the horizon. At each round i, the learner chooses an arm (also called an action) A^i from the action space $\mathcal{A} = \{a_k\}_{k=1}^K$ according to a policy π^i, and the environment reveals a reward $X^i \in \mathbb{R}$. The underlying environment ν belongs to an environment class $\mathcal{E}$ and defines the arm reward distributions. More precisely, given an environment $\nu = (P_a^\nu, a \in \mathcal{A}) \in \mathcal{E}$, the reward X^i follows the distribution $P_{A^i}^\nu$.

The policy $\boldsymbol{\pi}^i$ at round i is a function which maps the history $H^i = (A^1, X^1, \cdots, A^{i-1}, X^{i-1})$ to a probability distribution over the action space $\mathcal{A}$. More precisely, we denote the set of possible histories at the beginning of round i by $\mathcal{H}^i = (\mathcal{A} \times \mathbb{R})^{i-1}$, and $\mathcal{H}^1 = \emptyset$. We denote by $\Delta(\mathcal{A})$ the set of probability measures on $\mathcal{A}$ so that the policy $\boldsymbol{\pi}^i$ is a mapping from $\mathcal{H}^i$ to $\Delta(\mathcal{A})$. We denote by Π the set of policies $\boldsymbol{\pi} = \{\boldsymbol{\pi}^i\}_{i=1}^n$, which is measurable with respect to the filtration associated with the process $\{H^i\}_{i=1}^n$. We call Π the competitor class.

Define the expectation of arm a in environment ν as $\mu_a(\nu)$, where

$$\mu_a(\nu) = \int_{\mathbb{R}} x P_a^\nu(x) dx. \tag{1}$$

The expected cumulative reward $c_n(\boldsymbol{\pi}, \nu)$ measures the performance of policy $\boldsymbol{\pi}$ in environment ν,

$$c_n(\boldsymbol{\pi}, \nu) = \mathbb{E}\left[\sum_{i=1}^n \mu_{A^i}(\nu)\right],$$

where the expectation is taken over the probability measure induced by the interaction of the policy and the environment. The goal in the K-armed bandit setting is to design a policy $\boldsymbol{\pi}^*$ that leads to the largest expected cumulative reward among all policies in the competitor class Π. The main difficulty arises because the environment is unknown, and the policy can only depend on the history sequences $\{H^i\}_{i=0}^{n-1}$.

One way to measure the performance of a policy $\boldsymbol{\pi}$ is to find functions $C : \mathcal{E} \to [0, \infty)$, $f : \mathcal{E} \to [0, \infty)$ that upper bound the regret:

$$R_n(\boldsymbol{\pi}, \nu) = n\mu^*(\nu) - c_n(\boldsymbol{\pi}, \nu) \leq C(\nu)f(n),$$

where $\mu^*(\nu) = \max_a \mu_a(\nu)$ is the expected reward of the optimal arm. This is the frequentist regret, which is environment-dependent (Lattimore and Szepesvári, 2020).

Another way to measure the performance of a policy is via the averaged cumulative reward with respect to a probability measure $\rho(\nu)$ on the environment $\mathcal{E}$,

$$c_n(\boldsymbol{\pi}, \rho) = \mathbb{E}\left[\sum_{i=1}^n \int_{\mathcal{E}} \mu_{A^i}(\nu)\rho(\nu)d\nu\right].$$

In the Bayesian setting, the environment is viewed as a random variable. According to Bayes' rule, the probability $\rho(\nu)$ will be updated conditional on the history sequence. Given a horizon n, we assume that at each round i, the environment ν^i is sampled from a prior measure $\rho^i(\nu)$ over the environment class $\mathcal{E}$. After pulling arm A^i and obtaining the reward $X^i \sim P_\nu^{A^i}$, we update $\rho^{i+1}(\nu)$ to be the posterior distribution of the environment. Given

an initial prior measure $\rho^1(\nu)$, the goal is to find the optimal policy that maximizes the averaged cumulative reward,

$$\max_{\boldsymbol{\pi} \in \Pi} c_n(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{i=1}^{n} \int_{\mathcal{E}} \mu_{A^i}(\nu)\rho^i(\nu)d\nu\right].$$

The bandit problem with the above objective function is called a *Bayesian bandit* (Chapter 35 in Lattimore and Szepesvári (2020)).

3. Continuous Limits of Bayesian Bandits

3.1 An illustrative example

Consider the "one-armed" bandit problem in which the reward of the first arm follows a Bernoulli(ν) distribution (with ν unknown) and the second arm gives a deterministic reward $\mu_2 = \frac{1}{2}$. We assume an initial prior distribution of $\nu \sim \text{Beta}(\alpha, \beta)$. Then the posterior measure of ν at round i depends on two quantities. The first quantity is q^i, which is the number of pulls of the unknown arm before round i. The second quantity is s^i, which is the cumulative reward of the unknown arm before round i. The posterior distribution of ν is $\text{Beta}(\alpha + s^i, \beta + q^i - s^i)$.

Let $w^i(s, q)$ be the optimal cumulative reward starting from round i with $s^i = s$, $q^i = q$. Similarly, let $w_k^i(s, q)$ be the optimal cumulative reward in this same setting, assuming that the k-th arm is pulled at round i. Formally, we have

$$w^i(s, q) = \max_{\boldsymbol{\pi} \in \Pi} \mathbb{E}\left[\sum_{j=i}^{n} \int_{\mathcal{E}} \mu_{A^j}(\nu)\rho^j(\nu)d\nu \;\middle|\; s^i = s, q^i = q\right],$$

$$w_k^i(s, q) = \max_{\boldsymbol{\pi} \in \Pi} \mathbb{E}\left[\sum_{j=i}^{n} \int_{\mathcal{E}} \mu_{A^j}(\nu)\rho^j(\nu)d\nu \;\middle|\; s^i = s, q^i = q, A^i = k\right].$$

As before, the dependence of the expectation on the policy $\boldsymbol{\pi}$ is through the actions A^j. If the second arm is chosen at round i, then

$$w_2^i(s, q) = \mu_2 + w^{i+1}(s, q).$$

If the first arm is chosen at round i, then

$$w_1^i(s, q) = p(s, q) + p(s, q)w^{i+1}(s + 1, q + 1) + (1 - p(s, q))w^{i+1}(s, q + 1),$$

where

$$p(s, q) = \frac{\alpha + s}{\alpha + \beta + q}.$$

The first term $p(s, q)$ represents the expectation of the reward if the first arm is pulled at round i. The second and third term hold because after pulling the first arm, one has

$$q^{i+1} = q^i + 1;$$
$$\mathbb{P}(s^{i+1} = s^i + 1 | s^i) = p(s^i, q^i), \quad \mathbb{P}(s^{i+1} = s^i | s^i) = 1 - p(s^i, q^i).$$

4

Since w^i is the optimal reward, one has

$$w^i(s,q) = \max\{w_1^i(s,q), w_2^i(s,q)\}. \tag{2}$$

Note that for horizon n, $w^{n+1}(s,q) = 0$ for all s,q by the definition of w. Therefore, one can compute $w^i(s,q)$ for all $s = \{0, \cdots, i-1\}, q = \{0, \cdots, i-1\}$ via backwards induction.

To derive a continuous limit in this setting, we rescale the reward and the number of arm pulls by $1/n$:

$$\hat{s} = \frac{1}{n}s, \quad \hat{q} = \frac{1}{n}.$$

One can define the rescaled optimal reward function

$$v^i(\hat{s}, \hat{q}) = \frac{1}{n}w^i(s,q),$$

which then satisfies the equation

$$v^i(\hat{s}, \hat{q}) = \max \left\{ \frac{1}{n}\mu_2 + v^{i+1}(\hat{s}, \hat{q}), \right.$$
$$\left. \frac{1}{n}\tilde{p}(\hat{s}, \hat{q}) + \tilde{p}(\hat{s}, \hat{q})v^{i+1}(\hat{s} + \frac{1}{n}, \hat{q} + \frac{1}{n}) + (1 - \tilde{p}(\hat{s}, \hat{q}))v^{i+1}(\hat{s}, \hat{q} + \frac{1}{n}) \right\},$$

where

$$\tilde{p}(s,q) = \frac{n^{-1}\alpha + s}{n^{-1}(\alpha + \beta) + q}.$$

By setting $\delta_t = \delta_s = \delta_q = n^{-1}$, the above equation can be equivalently written as

$$\frac{v^{i+1}(\hat{s}, \hat{q}) - v^i(\hat{s}, \hat{q})}{\delta_t} + \max \left\{ \mu_2, \ \tilde{p}(\hat{s}, \hat{q}) + \tilde{p}(\hat{s}, \hat{q})\frac{v^{i+1}(\hat{s} + \delta_s, \hat{q} + \delta_q) - v^{i+1}(\hat{s}, \hat{q} + \delta_q)}{\delta_s} \right.$$
$$\left. + \frac{v^{i+1}(\hat{s}, \hat{q} + \delta_q) - v^{i+1}(\hat{s}, \hat{q})}{\delta_q} \right\} = 0.$$

From the above equation, one sees that the rescaled value function $v^i(\hat{s}, \hat{q})$ is the numerical discretization of the following PDE:

$$\partial_t v(t, \hat{s}, \hat{q}) + \max\{\mu_2, \hat{\mu}(\hat{s}, \hat{q}) + \hat{\mu}(\hat{s}, \hat{q})\partial_{\hat{s}}v(t, \hat{s}, \hat{q}) + \partial_{\hat{q}}v(t, \hat{s}, \hat{q})\} = 0, \quad v(1, \hat{s}, \hat{q}) = 0, \tag{3}$$

where

$$\hat{\mu}(s,q) = \lim_{n\to\infty} \tilde{p}(s,q).$$

Furthermore, by moving the constant μ_2 outside of the maximum operator and introducing a control parameter $\hat{\pi} \in [0,1]$, one arrives at a Hamilton-Jacobi-Bellman equation for v:

$$\partial_t v + \max_{\hat{\pi}(t,\hat{s},\hat{q})\in[0,1]} (\hat{\mu} + \hat{\mu}\partial_{\hat{s}}v + \partial_{\hat{q}}v - \mu_2)\pi + \mu_2 = 0, \qquad v(1, \hat{s}, \hat{q}) = 0. \tag{4}$$

(3) and (4) are equivalent because when $\hat{\mu} + \hat{\mu}\partial_{\hat{s}}v + \partial_{\hat{q}}v > (<)\mu_2$, then $\hat{\pi} = 1(= 0)$, respectively. In other words, as the horizon $n \to \infty$, the rescaled value function $v^i(\hat{s}, \hat{q})$
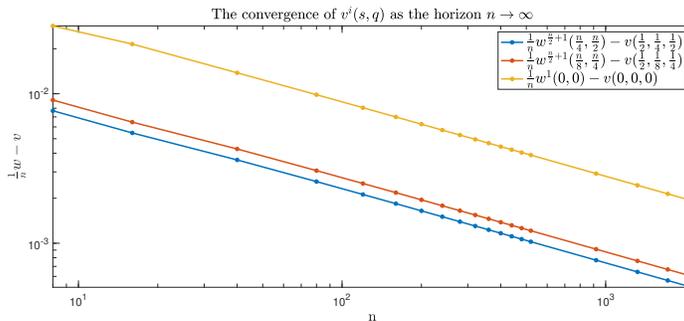
Figure 1: The above plot shows the decrease in the error $|\frac{1}{n}w^i(s,q) - v(t,\hat{s},\hat{q})|$ as the horizon $n \to \infty$ for $(i,s,q) = \{(\frac{n}{2}+1, \frac{n}{4}, \frac{n}{2}), (\frac{n}{2}+1, \frac{n}{8}, \frac{n}{4}), (1,0,0)\}$ and the corresponding $(t,s,q) = \{(\frac{1}{2}, \frac{1}{4}, \frac{1}{2})(\frac{1}{2}, \frac{1}{8}, \frac{1}{4}), (0,0,0)\}$. We set the initial hyperparameters $(\alpha, \beta) = (\frac{n}{2}, \frac{n}{2})$,.

satisfies the above HJB equation. We plot the convergence of $\frac{1}{n}w^i(\hat{s},\hat{q})$ as n increases in Figure 1.

Classical results from optimal control (see, e.g., Chapter 10.3.3 of Evans (2010)) imply that $v(t,\hat{s},\hat{q})$ in (4) solves the control problem

$$
\begin{aligned}
\max_{\hat{\pi}(\tau) \in [0,1]} \quad & \int_t^1 \hat{\mu}(\hat{s}(\tau), \hat{q}(\tau))\pi(\tau) + (1 - \pi(\tau))\mu_2 d\tau \\
\text{s.t.} \quad & d\hat{q}(\tau) = \pi(\tau)d\tau, \\
& d\hat{s}(\tau) = \hat{\mu}(\hat{s}(\tau), \hat{q}(\tau))\pi(\tau)d\tau, \\
& \hat{q}(t) = \hat{s}, \quad \hat{s}(t) = \hat{q}.
\end{aligned}
\tag{5}
$$

The preceding example illustrates that the Bayesian bandit algorithm (2) can be viewed as the discretization of an HJB equation (4), which solves the control problem (5). In other words, as the horizon $n \to \infty$, the Bayesian bandit problem will converge to a continuous control problem that can be solved via the HJB equation. In the next section, we extend this formulation to a more general setting.

3.2 Formal derivation from Bayesian bandits to the HJB equation

We return to the original K-armed bandit setting with horizon n and environments parameterized by $\boldsymbol{\nu} \in \mathcal{E} \subset \mathbb{R}^d$. The reward of the k-th arm a_k follows the distribution $P_k^{\boldsymbol{\nu}}$. We assume a prior measure $\rho(\boldsymbol{\nu})$ on the environment $\boldsymbol{\nu}$ at the beginning of the first round. We also assume that the updated measure $\rho^i(\boldsymbol{\nu})$ at the beginning of round i only depends on $(\mathbf{s}^i, \mathbf{q}^i)$. Here $\mathbf{s}^i = (s_k^i)_{k=1}^K \in \mathbb{R}^K$ is a K-dimensional vector representing the cumulative reward of each arm up to round $i-1$, and $\mathbf{q}^i = (q_k^i)_{k=1}^K \in \{0, \cdots, i-1\}^K$ is a K-dimensional vector representing the number of pulls of each arm up to round $i-1$. We note that the assumption that the state space can be reduced to $(\mathbf{s}^i, \mathbf{q}^i)$ covers many, but not all, bandit algorithms. For instance, many stochastic bandit algorithms, where the arm rewards are drawn from a stationary probability distribution, can be represented using this state space,

but the Exp3 algorithm (Auer et al., 2002) cannot. For an extensive discussion, refer to Section 2.1 of Wager and Xu (2021).

Under the above assumptions, the posterior distribution of the environment at round i can be written as a function of $\mathbf{s}^i, \mathbf{q}^i$, i.e., $\rho(\boldsymbol{\nu}|\mathbf{s}^i, \mathbf{q}^i)$. If arm k is pulled at round i, then the reward X^i follows the distribution $P_k^{\boldsymbol{\nu}}$, and

$$s_k^{i+1} = s_k^i + X^i, q_k^{i+1} = q_k^i + 1. \tag{6}$$

Let $w^i(\mathbf{s}, \mathbf{q})$ be the optimal expected cumulative reward starting from round i with $(\mathbf{s}^i, \mathbf{q}^i) = (\mathbf{s}, \mathbf{q})$. Then it satisfies the following equation:

$$w^i(\mathbf{s}, \mathbf{q}) = \max_k \left\{ \int_{\mathcal{E}} \mu_k(\boldsymbol{\nu})\rho(\boldsymbol{\nu}|\mathbf{s}, \mathbf{q})d\boldsymbol{\nu} + \int_{\mathcal{E}} \int_{\mathbb{R}} w^{i+1}(\mathbf{s} + x\mathbf{e}_k, \mathbf{q} + \mathbf{e}_k)P_k^{\boldsymbol{\nu}}(x)\rho(\boldsymbol{\nu}|\mathbf{s}, \mathbf{q}) \, dx \, d\boldsymbol{\nu} \right\}, \tag{7}$$

where $\mu_k(\boldsymbol{\nu})$ is the expected reward of the k-th arm defined in (1), and $\mathbf{e}_k$ is a K-dimensional vector with the k-th element being 1 and all other elements being 0. The first term in the max operator represents the expectation of the rewards if the k-th arm is pulled. The second term is due to the fact that s_k^{i+1} and q_k^{i+1} will follow (6) if the k-th arm is pulled at round i.

Next, we rescale the parameters to derive the continuous-in-time limit. Let

$$t = \frac{i-1}{n}, \quad \hat{\mathbf{q}} = \frac{1}{n}\mathbf{q}^i, \quad \hat{\mathbf{s}} = \frac{1}{f(n)}\mathbf{s}^i, \quad v\left(\frac{i-1}{n}, \frac{\mathbf{s}}{f(n)}, \frac{\mathbf{q}}{n}\right) = \frac{1}{f(n)}w^i(\mathbf{s}, \mathbf{q}). \tag{8}$$

We shrink the n rounds to the time interval $[0, 1]$ so that as $n \to \infty$, the discrete round $i \leq n$ will correspond to a continuous time $t = (i-1)/n \in [0, 1]$. We also rescale the number of pulls to $[0, 1]$, so that it is on the same scale as the rescaled rounds. The cumulative reward $\mathbf{s}^i$ and w^i are rescaled by $\frac{1}{f(n)}$, where $f(n)$ will be determined later. Accordingly, the rescaled expected reward $v(t, \hat{\mathbf{s}}, \hat{\mathbf{q}})$ becomes a function of the continuous time t and rescaled history $(\hat{\mathbf{s}}, \hat{\mathbf{q}})$. We refer to the function $f(n)$ as the *scaling factor*. For different scaling factors, the limiting rescaled cumulative reward $v(t, \hat{\mathbf{s}}, \hat{\mathbf{q}})$ will follow different dynamics.

Define the moments of the k-th arm w.r.t. the probability measure $P_k^{\boldsymbol{\nu}}(x)$ and Bayesian measure $\rho(\boldsymbol{\nu}|\mathbf{s}, \mathbf{q})$ by

$$\bar{\mu}_k(\mathbf{s}, \mathbf{q}) = \int_{\mathcal{E}} \int_{\mathbb{R}} x P_k^{\boldsymbol{\nu}}(x)\rho(\boldsymbol{\nu}|\mathbf{s}, \mathbf{q}) \, dx \, d\boldsymbol{\nu}, \quad \bar{\sigma}_k^2(\mathbf{s}, \mathbf{q}) = \int_{\mathcal{E}} \int_{\mathbb{R}} x^2 P_k^{\boldsymbol{\nu}}(x)\rho(\boldsymbol{\nu}|\mathbf{s}, \mathbf{q}) \, dx \, d\boldsymbol{\nu},$$
$$\bar{E}_k^p(\mathbf{s}, \mathbf{q}) = \int_{\mathcal{E}} \int_{\mathbb{R}} x^p P_k^{\boldsymbol{\nu}}(x)\rho(\boldsymbol{\nu}|\mathbf{s}, \mathbf{q}) \, dx \, d\boldsymbol{\nu}. \tag{9}$$

We assume that these moments exist and are finite. Inserting the Taylor expansion

$$w^{i+1}(\mathbf{s} + x\mathbf{e}_k, \mathbf{q} + \mathbf{e}_k) = \sum_{p=0}^{\infty} \frac{1}{p!} \partial_{s_k}^p w^{i+1}(\mathbf{s}, \mathbf{q} + \mathbf{e}_k)x^p$$

into (7) yields

$$w^i(\mathbf{s}, \mathbf{q}) = \max_k \left\{ \bar{\mu}_k(\mathbf{s}, \mathbf{q}) + w^{i+1}(\mathbf{s}, \mathbf{q} + \mathbf{e}_k) + \bar{\mu}_k(\mathbf{s}, \mathbf{q})\partial_{s_k} w^{i+1}(\mathbf{s}, \mathbf{q} + \mathbf{e}_k) + \frac{1}{2}\bar{\sigma}_k^2(\mathbf{s}, \mathbf{q})\partial_{s_k}^2 w^{i+1}(\mathbf{s}, \mathbf{q} + \mathbf{e}_k) \right.$$

$$\left. + \sum_{p=3}^{\infty} \frac{1}{p!}\bar{E}_k^p(\mathbf{s}, \mathbf{q})\partial_{s_k}^p w^{i+1}(\mathbf{s}, \mathbf{q} + \mathbf{e}_k) \right\}.$$

Therefore, the rescaled reward $v\left(\frac{i-1}{n}, \frac{\mathbf{s}}{f(n)}, \frac{\mathbf{q}}{n}\right) = \frac{1}{f(n)} w^i(\mathbf{s}, \mathbf{q})$ satisfies

$$
\begin{aligned}
v\left(\frac{i-1}{n}, \frac{\mathbf{s}}{f(n)}, \frac{\mathbf{q}}{n}\right) = \max_k \Bigg\{ &\frac{1}{f(n)}\bar{\mu}_k(\mathbf{s}, \mathbf{q}) + v\left(\frac{i}{n}, \frac{\mathbf{s}}{f(n)}, \frac{\mathbf{q}+\mathbf{e}_k}{n}\right) \\
&+ \frac{1}{f(n)}\bar{\mu}_k(\mathbf{s}, \mathbf{q})\partial_{\hat{s}_k} v\left(\frac{i}{n}, \frac{\mathbf{s}}{f(n)}, \frac{\mathbf{q}+\mathbf{e}_k}{n}\right) \\
&+ \frac{1}{2}\frac{1}{f^2(n)}\bar{\sigma}_k^2(\mathbf{s}, \mathbf{q})\partial_{\hat{s}_k}^2 v\left(\frac{i}{n}, \frac{\mathbf{s}}{f(n)}, \frac{\mathbf{q}+\mathbf{e}_k}{n}\right) \\
&+ \sum_{p=3}^{\infty} \frac{1}{p!}\frac{1}{f^p(n)}\bar{E}_k^p(\mathbf{s}, \mathbf{q})\partial_{\hat{s}_k}^p v\left(\frac{i}{n}, \frac{\mathbf{s}}{f(n)}, \frac{\mathbf{q}+\mathbf{e}_k}{n}\right) \Bigg\}.
\end{aligned}
$$

After reorganizing the terms and setting $\delta_t = \delta_q = \frac{1}{n}$, one has

$$
\begin{aligned}
\frac{v(t+\delta_t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) - v(t, \hat{\mathbf{s}}, \hat{\mathbf{q}})}{\delta_t} + \max_k \Bigg\{ &\frac{1}{\delta_t f(n)}\bar{\mu}_k(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}}) + \frac{v(t+\delta_t, \hat{\mathbf{s}}, \hat{\mathbf{q}} + \delta_q\mathbf{e}_k) - v(t+\delta_t, \hat{\mathbf{s}}, \hat{\mathbf{q}})}{\delta_q} \\
&+ \frac{1}{\delta_t f(n)}\bar{\mu}_k(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}})\partial_{\hat{s}_k} v(t+\delta_t, \hat{\mathbf{s}}, \hat{\mathbf{q}} + \delta_q\mathbf{e}_k) \\
&+ \frac{1}{2}\frac{1}{\delta_t f^2(n)}\bar{\sigma}_k^2(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}})\partial_{\hat{s}_k}^2 v(t+\delta_t, \hat{\mathbf{s}}, \hat{\mathbf{q}} + \delta_q\mathbf{e}_k) \\
&+ \sum_{p=3}^{\infty} \frac{1}{p!}\frac{1}{\delta_t f^p(n)}\bar{E}_k^p(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}})\partial_{\hat{s}_k}^p v(t+\delta_t, \hat{\mathbf{s}}, \hat{\mathbf{q}} + \delta_q\mathbf{e}_k) \Bigg\} = 0.
\end{aligned}
$$

If one assumes that there exist functions $\{\hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}})\}_{k=1}^K$, $\{\hat{\sigma}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}})\}_{k=1}^K$, such that for all $\hat{\mathbf{s}} \in \mathbb{R}^K, \hat{\mathbf{q}} \in [0,1]^K$,

$$
\begin{aligned}
&\lim_{n\to\infty} \frac{n}{f(n)}\bar{\mu}_k(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}}) = \hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}); \\
&\lim_{n\to\infty} \frac{n}{f^2(n)}\bar{\sigma}_k^2(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}}) = \hat{\sigma}_k^2(\hat{\mathbf{s}}, \hat{\mathbf{q}}); \\
&\lim_{n\to\infty} \frac{n}{f^p(n)}\bar{E}_k^p(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}}) = \hat{E}_k^p \equiv 0, \quad \text{for} \quad \forall p \geq 3,
\end{aligned}
\tag{10}
$$

then as the horizon $n \to \infty$, i.e., $\delta_t, \delta_q \to 0$, the rescaled expected cumulative reward $v(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) = \frac{1}{f(n)} w^{nt+1}(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}})$ satisfies the PDE

$$
\partial_t v + \max_k \{\hat{\mu}_k + \partial_{\hat{q}_k} v + \hat{\mu}_k \partial_{\hat{s}_k} v + \frac{1}{2}\hat{\sigma}_k^2 \partial_{\hat{s}_k}^2 v\} = 0,
$$

which can be equivalently written as the following HJB equation:

$$
\partial_t v + \max_{\hat{\boldsymbol{\pi}}(t,\hat{\mathbf{s}},\hat{\mathbf{q}}) \in \Delta^K} \sum_{k=1}^K \left(\hat{\mu}_k(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) + \partial_{\hat{q}_k} v + \hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}})\partial_{\hat{s}_k} v + \frac{1}{2}\hat{\sigma}_k^2(\hat{\mathbf{s}}, \hat{\mathbf{q}})\partial_{\hat{s}_k}^2 v \right) \hat{\pi}_k = 0. \tag{11}
$$

Here we introduce $\hat{\boldsymbol{\pi}}(t, \hat{\mathbf{s}}, \hat{\mathbf{q}})$ as the feedback control, which corresponds to the policy in the bandit problem. Since the policy is a mapping from the history H^i to probability

measures on the action space $\Delta(\mathcal{A})$, the policy at round i is described by a K-dimensional vector-valued function $\boldsymbol{\pi}^i(\mathbf{s}^i, \mathbf{q}^i)$ that satisfies $\sum_k \pi_k^i(\mathbf{s}^i, \mathbf{q}^i) = 1$. In the limit, the policy $\hat{\boldsymbol{\pi}}(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) = \lim_{n \to \infty} \boldsymbol{\pi}^{nt+1}(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}})$ is a mapping from the rescaled history $(\hat{\mathbf{s}}, \hat{\mathbf{q}})$ to the simplex Δ^K that satisfies $\sum_k \hat{\pi}_k(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) = 1$ for $\forall (t, \hat{\mathbf{s}}, \hat{\mathbf{q}})$.

We remark briefly that the selection (or even the existence) of $f(n)$ may not be obvious in the general setting described above. In general, $f(n)$ should be thought of as describing the "order" or asymptotic size of the unscaled rewards in the original bandit problem with n rounds. For instance, if the rewards are of constant size and observed with at least constant probability, then we will expect the cumulative reward of the original bandit problem to be linear in the time horizon, and we will have $f(n) = \Theta(n)$. Rather than describing necessary and sufficient technical conditions relating $f(n)$ to the arm reward distributions $P_k^{\boldsymbol{\nu}}$ and the posterior $\rho(\boldsymbol{\nu}|\mathbf{s}, \mathbf{q})$ (which may be intractable given the generality of the framework), in the remainder of the paper, we will show that in a wide range of concrete examples, our framework provides useful insights into the problem. See Remark 2 for further discussion of the scaling factor $f(n)$.

Note that the solution to (11) is not necessarily differentiable, so we are searching for a viscosity solution instead of a classical solution (Evans, 2010). One has the following guarantee on the well-posedness of the solution.

Proposition 1 *If $\{\hat{\mu}_k(\mathbf{s}, \mathbf{q})\}_k$, $\{\hat{\sigma}_k(\mathbf{s}, \mathbf{q})\}_k$ are bounded and Lipschitz continuous in $(\mathbf{s}, \mathbf{q})$, then the value function defined in (13) is the unique viscosity solution to the HJB equation (11).*

See, e.g., Nisio (2015) for the proof. Finding necessary and sufficient conditions on the problem primitives—specifically, the arm reward distributions $P_k^{\boldsymbol{\nu}}$, the prior ρ, and the scaling factor $f(n)$—under which the boundedness and Lipschitz assumptions in Proposition 1 hold is an interesting question for future work. In this paper, our goal is to demonstrate the useful insights which can be derived from our framework for specific bandit problems, once it has been determined that they meet these conditions.

Summary If the rescaled moments of all the arms satisfy (10) for some scaling factor $f(n)$, then the rescaled optimal expected cumulative reward $\frac{1}{f(n)} w^{nt+1}(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}})$ will converge to $v(t, \hat{\mathbf{s}}, \hat{\mathbf{q}})$ as $n \to \infty$, where v satisfies the HJB equation (11) with boundary condition $v(1, \hat{\mathbf{s}}, \hat{\mathbf{q}}) = 0$. In addition, the rescaled optimal policy $\boldsymbol{\pi}^{*,nt+1}(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}})$ will converge to $\hat{\boldsymbol{\pi}}^*(t, \hat{\mathbf{s}}, \hat{\mathbf{q}})$ as $n \to \infty$, where $\hat{\boldsymbol{\pi}}^*(t, \hat{\mathbf{s}}, \hat{\mathbf{q}})$ is given by

$$\pi_k^*(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) = \begin{cases} 1, & k = \underset{k}{\operatorname{argmax}} \left\{ \hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) + \partial_{\hat{q}_k} v + \hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) \partial_{\hat{s}_k} v + \frac{1}{2} \hat{\sigma}_k^2(\hat{\mathbf{s}}, \hat{\mathbf{q}}) \partial_{\hat{s}_k}^2 v \right\} \\ 0, & o.w. \end{cases} \tag{12}$$

These results form the foundation for the rest of the paper.

3.3 A formal derivation from Bayesian bandits to the optimal control problems

If one views $v(t, \mathbf{s}, \mathbf{q})$ as the optimal cumulative reward starting from time t

$$v(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) = \max_{\hat{\boldsymbol{\pi}}(\tau) \in \Delta^K} \mathbb{E} \left[\int_t^1 \hat{\boldsymbol{\mu}}(\hat{\mathbf{s}}(\tau), \hat{\mathbf{q}}(\tau)) \cdot \boldsymbol{\pi}(\tau) d\tau \right], \tag{13}$$

then $v(t, \mathbf{s}, \mathbf{q})$ is the solution to the optimal control problem (Evans, 2010)

$$
\begin{aligned}
\max_{\hat{\boldsymbol{\pi}}(\tau) \in \Delta^K} \quad & \mathbb{E} \left[\int_t^1 \hat{\boldsymbol{\mu}}(\hat{\mathbf{s}}(\tau), \hat{\mathbf{q}}(\tau)) \cdot \boldsymbol{\pi}(\tau) d\tau \right] \\
\text{s.t.} \quad & d\hat{q}_k(\tau) = \hat{\pi}_k(\tau) d\tau, \quad 1 \le k \le K; \\
& d\hat{s}_k(\tau) = \hat{\mu}_k(\hat{\mathbf{s}}(\tau), \hat{\mathbf{q}}(\tau)) \hat{\pi}_k(\tau) d\tau + \hat{\sigma}_k(\hat{\mathbf{s}}(\tau), \hat{\mathbf{q}}(\tau)) \sqrt{\hat{\pi}_k(\tau)} dB_\tau, \quad 1 \le k \le K; \\
& \hat{\mathbf{s}}(t) = \hat{\mathbf{s}}, \quad \hat{\mathbf{q}}(t) = \hat{\mathbf{q}}.
\end{aligned}
\tag{14}
$$

Therefore, as the horizon $n \to \infty$, the Bayesian bandit problem also converges to the above continuous control problem. In fact, one can derive the optimal control formulation above directly from the definition of the Bayesian bandit problem.

Given a policy $\{\boldsymbol{\pi}^i\}_i$, at each round i, the environment $\boldsymbol{\nu}$ is sampled with probability $\rho(\boldsymbol{\nu} | \mathbf{s}^i, \mathbf{q}^i)$, the k-th arm is pulled with probability π_k^i, and the reward of the k-th arm follows distribution $P_k^{\boldsymbol{\nu}}$. Assume at the beginning of round I, $(\mathbf{s}^I, \mathbf{q}^I) = (\mathbf{s}, \mathbf{q})$. The goal of Bayesian bandits is to find the optimal policy $\{\boldsymbol{\pi}^{*,i}\}_i$ that maximizes the expected cumulative reward. More precisely, our goal is to find $\{\boldsymbol{\pi}^i\}_i$ that solves the following optimization problem:

$$
\begin{aligned}
\max_{\{\boldsymbol{\pi}^i\}_i \in \Delta^K} \quad & c_n(\{\boldsymbol{\pi}^i\}) = \mathbb{E} \left[\sum_{i=I}^n \int_{\mathcal{E}} \mu_{A^i}(\boldsymbol{\nu}) \rho(\boldsymbol{\nu} | \mathbf{s}^i, \mathbf{q}^i) d\boldsymbol{\nu} \right] \\
\text{where} \quad & A^i = k \quad \text{w.p. } \pi_k^i(\mathbf{s}^i, \mathbf{q}^i), \\
& s_k^{i+1} - s_k^i = \begin{cases} 0, & \text{if } A^i \ne k \\ X^i \sim P_k^{\boldsymbol{\nu}}, \boldsymbol{\nu} \sim \rho(\boldsymbol{\nu} | \mathbf{s}^i, \mathbf{q}^i), & \text{if } A^i = k \end{cases}, \\
& q_k^{i+1} - q_k^i = \begin{cases} 0, & \text{if } A^i \ne k \\ 1, & \text{if } A^i = k \end{cases}, \\
& (\mathbf{s}^I, \mathbf{q}^I) = (\mathbf{s}, \mathbf{q}),
\end{aligned}
\tag{15}
$$

where $\mu_a(\boldsymbol{\nu})$ is the expected reward of arm a defined in (1).

Using the same rescaling as in (10) and viewing the history $(\mathbf{s}^i, \mathbf{q}^i)$ at the discrete rounds as a function $(\hat{\mathbf{s}}(t), \hat{\mathbf{q}}(t))$ over the continuous time t, the differences of the rescaled cumulative reward $\hat{\mathbf{s}}(t)$ and the rescaled number of pulls $\hat{\mathbf{q}}(t)$ after one round become

$$
\begin{aligned}
\mathbb{E}[\hat{s}_k(t + \delta_t) - \hat{s}_k(t)] &= \delta_t \frac{n}{f(n)} \pi_k^i(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}}) \bar{\mu}_k(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}}), \\
\mathbb{V}[\hat{s}_k(t + \delta_t) - \hat{s}_k(t)] &= \delta_t \frac{n}{f^2(n)} \pi_k^i(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}}) \bar{\sigma}_k^2(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}}) - (\mathbb{E}[\hat{s}_k(t + \delta_t) - \hat{s}_k(t)])^2, \\
\mathbb{E}[\hat{q}_k(t + \delta_t) - \hat{q}_k(t)] &= \delta_t \, \pi_k^i(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}}), \\
\mathbb{V}[\hat{q}_k(t + \delta_t) - \hat{q}_k(t)] &= (\delta_t)^2 \pi_k^i(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}}) - (\delta_t)^2 (\pi_k^i(f(n)\hat{\mathbf{s}}, n\hat{\mathbf{q}}))^2,
\end{aligned}
\tag{16}
$$

where $\delta_t = \frac{1}{n}$, and $\bar{\mu}_k, \bar{\sigma}_k$ are defined in (9). Accordingly, the rescaled objective function becomes

$$\hat{c}_n(\{\boldsymbol{\pi}^i\}) = \frac{1}{f(n)} c_n(\{\boldsymbol{\pi}^i\})$$

$$= \delta_t \, \mathbb{E}\left[\sum_{i=I}^{n} \sum_{k=1}^{K} \frac{n}{f(n)} \pi_k^i \left(f(n)\hat{\mathbf{s}}\left(\frac{i}{n}\right), n\hat{\mathbf{q}}\left(\frac{i}{n}\right)\right) \bar{\mu}_k \left(f(n)\hat{\mathbf{s}}\left(\frac{i}{n}\right), n\hat{\mathbf{q}}\left(\frac{i}{n}\right)\right)\right].$$
(17)

Formally, based on (16) and (17), under the assumptions (10) on the moments, the following equations hold.

$$\lim_{n\to\infty} \mathbb{E}[\hat{s}_k(t+\delta_t) - \hat{s}_k(t)]/\delta_t = \hat{\pi}_k(t,\hat{\mathbf{s}},\hat{\mathbf{q}})\hat{\mu}_k(\hat{\mathbf{s}},\hat{\mathbf{q}}),$$

$$\lim_{n\to\infty} \mathbb{V}[\hat{s}_k(t+\delta_t) - \hat{s}_k(t)]/\delta_t = \hat{\pi}_k(t,\hat{\mathbf{s}},\hat{\mathbf{q}})\hat{\sigma}_k^2(\hat{\mathbf{s}},\hat{\mathbf{q}}),$$

$$\lim_{n\to\infty} \mathbb{E}[\hat{q}_k(t+\delta_t) - \hat{q}_k(t)]/\delta_t = \hat{\pi}_k(t,\hat{\mathbf{s}},\hat{\mathbf{q}}),$$

$$\lim_{n\to\infty} \mathbb{V}[\hat{q}_k(t+\delta_t) - \hat{q}_k(t)]/\delta_t = 0,$$

$$\lim_{n\to\infty} \hat{c}_n(\{\boldsymbol{\pi}^i\}) = \mathbb{E}\left[\int_t^1 \sum_{k=1}^{K} \hat{\pi}_k(t,\hat{\mathbf{s}}(\tau),\hat{\mathbf{q}}(\tau))\hat{\mu}_k(\hat{\mathbf{s}}(\tau),\hat{\mathbf{q}}(\tau))d\tau\right].$$

This implies that the rescaled version of the Bayesian bandits (15) will converge to the continuous optimal control problem (14).

Remark 2 *Assumption (10) is critical to determine what scaling factor $f(n)$ one should use to derive a meaningful limiting HJB equation. Take the Bernoulli reward introduced in Section 3.1 for an example. Since*

$$\bar{\mu}(s,q) = \bar{\sigma}^2(s,q) = \bar{E}^p(s,q) = \frac{\alpha+s}{\alpha+\beta+q},$$

one has

$$\hat{\mu}(\hat{s},\hat{q}) = \lim_{n\to\infty} \frac{n}{f(n)} \frac{\frac{\alpha}{f(n)}+\hat{s}}{\frac{\alpha+\beta}{n}+\hat{q}}, \quad \hat{\sigma}^2(\hat{s},\hat{q}) = \lim_{n\to\infty} \frac{n}{f(n)^2} \frac{\frac{\alpha}{f(n)}+\hat{s}}{\frac{\alpha+\beta}{n}+\hat{q}},$$

$$\hat{E}^p(\hat{s},\hat{q}) = \lim_{n\to\infty} \frac{n}{f(n)^p} \frac{\frac{\alpha}{f(n)}+\hat{s}}{\frac{\alpha+\beta}{n}+\hat{q}}.$$

If the initial hyperparameters (α,β) are set such that $\lim_{n\to\infty}(\frac{\alpha}{f(n)}, \frac{\alpha+\beta}{n}) = (\hat{\alpha},\hat{\beta})$, then the only scaling factor that will induce a meaningful HJB equation is $f(n) = O(n)$. For all $f(n) = O(n^b)$ with $b < 1$, $\hat{\mu}$ diverges. For the case where $b > 1$, one has $\hat{\mu} \equiv 0$, and consequently, the limiting HJB equation does not give any useful information on the dynamics. In this particular case, one always ends up with an HJB equation without a diffusion term ($\sigma \equiv 0$), and this is consistent with the nature of the Bernoulli bandits. Note that $\hat{s}(t)$ is non-decreasing by definition, but when $\sigma > 0$, there is a chance that $\hat{s}(t)$ will decrease due to the nonzero dB_τ term. Therefore, any valid scaling factor must induce a deterministic optimal control problem. However, in the general case, depending on the choice of scaling factor $f(n)$, the limiting optimal control problem can be stochastic or deterministic.

11

3.4 Specialized limiting equations for Structured and Unstructured Bandits

In this section, we will derive the general limiting HJB equation for both unstructured and structured bandits. Then in Section 3.4.1, we will derive the continuous limit for some specific bandit problems which are common in the literature.

Unstructured Bandits In the unstructured bandit problem, action $a \in \mathcal{A}$ is completely uninformative of all other actions $b \neq a$. That is, when a is played, the learner gains no information about the reward distribution of the other arms (Chapter 4.3 of Lattimore and Szepesvári (2020)). Here we mainly discuss a specific kind of unstructured bandit problem in which all the arms belong to the same parametric family of distributions with different unknown parameters. For instance, the rewards of each arm may be normally distributed with an unknown mean. One will see that, in this case, the drift and diffusion terms of each arm in the limiting HJB equation have a similar form. Provided that the initial prior measure is determined by a hyperparameter β, then we claim that under certain conditions, the Bayesian bandit problem converges to the following HJB equation as the horizon $n \to \infty$:

$$
\partial_t v + \max_{\hat{\boldsymbol{\pi}}(t,\hat{\mathbf{s}},\hat{\mathbf{q}}) \in \Delta^K} \left[\sum_{k=1}^{K} \left(\partial_{\hat{q}_k} v + \hat{\mu}(\hat{s}_k, \hat{q}_k, \hat{\beta}_k) \partial_{\hat{s}_k} v + \frac{1}{2} \hat{\sigma}(\hat{s}_k, \hat{q}_k, \hat{\beta}_k)^2 \partial_{\hat{s}_k}^2 v \right. \right.
$$
$$
\left. \left. + \hat{\mu}(\hat{s}_k, \hat{q}_k, \hat{\beta}_k) \right) \hat{\pi}_k(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) \right] = 0,
$$
$$
v(1, \mathbf{s}, \mathbf{q}) = 0,
\tag{18}
$$

where the functions $\hat{\mu}$ and $\hat{\sigma}$ depend on the problem setting, and $\hat{\beta}$ is the rescaled hyperparameter. The main difference between the general form (14) and the unstructured version (19) is that the drift and diffusion functions $(\hat{\mu}_k, \hat{\sigma}_k)$ have the same form for each arm. Thus, the difference in drift and diffusion for each arm will be due only to differences in the arm histories $(\hat{s}_k, \hat{q}_k)$ at time t and the initial prior measure represented by $\hat{\beta}$.

Structured Bandits In a structured bandit problem, the learner can obtain information about other actions even if these actions are never played. That is, playing a particular arm $a \in \mathcal{A}$ may be informative of the reward distributions of other arms $b \neq a$ (Chapter 4.3 of Lattimore and Szepesvári (2020)). Specifically, we consider the following setting. Let the action space $\{\mathbf{a}_k\}_{k=1}^K = \mathcal{A} \subset \mathbb{R}^d$ be a set of real vectors, and assume that the reward $x \in \Omega$ on the arm $\mathbf{a}_k$ has density $p(x|\mathbf{a}_k, \boldsymbol{\nu})$ with an unknown parameter $\boldsymbol{\nu} \in \mathcal{E} \in \mathbb{R}^d$. Then we claim that under certain conditions, the Bayesian bandit problem described above converges to the following HJB equation as the horizon $n \to \infty$:

$$
\partial_t v + \max_{\hat{\boldsymbol{\pi}}(t,\hat{\mathbf{s}},\hat{\mathbf{q}}) \in \Delta^K} \left[\sum_{k=1}^{K} \left(\partial_{\hat{q}_k} v + \hat{\mu}(\hat{\mathbf{s}}, \hat{\mathbf{q}}, \mathbf{a}_k) \partial_{\hat{s}_k} v + \frac{1}{2} \hat{\sigma}(\hat{\mathbf{s}}, \hat{\mathbf{q}}, \mathbf{a}_k)^2 \partial_{\hat{s}_k}^2 v \right. \right.
$$
$$
\left. \left. + \hat{\mu}(\hat{\mathbf{s}}, \hat{\mathbf{q}}, \mathbf{a}_k) \right) \hat{\pi}_k(t, \mathbf{s}, \mathbf{q}) \right] = 0,
$$
$$
v(1, \mathbf{s}, \mathbf{q}) = 0,
\tag{19}
$$

Similar to the unstructured case, the main difference between (11) and (19) is that the drift and diffusion terms $(\hat{\mu}_k, \hat{\sigma}_k)$ have the same form. In the structured case, the two terms

depend on the history of all the arms $(\hat{\mathbf{s}}, \hat{\mathbf{q}})$ and the position of the arm $\mathbf{a}_k$, and the difference for each arm is only due to the position $\mathbf{a}_k$.

3.4.1 THREE COMMON BANDITS

Here we derive the HJB equation for three common examples: unstructured bandits with Bernoulli and normal arm rewards, and a structured linear bandit with normal rewards.

Unstructured Bernoulli rewards   Consider the environment class $\boldsymbol{\nu} \in [0,1]^K$ of Bernoulli distributions $\mathcal{B}_\gamma$ with horizon n. For an environment $\boldsymbol{\nu} \in \mathcal{E} = [0,1]^K$, the k-th arm follows a Bernoulli distribution taking values $\gamma(n)$ and $-\gamma(n)$ with probability ν_k and $1 - \nu_k$, respectively, where ν_k is the k-th component of $\boldsymbol{\nu}$. We set the initial prior measure for the k-th arm to be $\nu_k^1 \sim \text{Beta}(\alpha_k(n), \beta_k(n))$ for $1 \leq k \leq K$.

Lemma 3 *Let $f(n)$ be such that there exist real numbers $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\sigma})$ with*

$$\lim_{n \to \infty} \frac{\gamma(n)(\alpha_k(n) - \beta_k(n))}{f(n)} = \hat{\alpha}_k, \quad \lim_{n \to \infty} \frac{\alpha_k(n) + \beta_k(n)}{n} = \hat{\beta}_k, \quad \lim_{n \to \infty} \frac{\sqrt{n}}{f(n)}\gamma(n) = \hat{\sigma}.$$

Then as $n \to \infty$, the rescaled Bayesian bandit problem with scaling factor $f(n)$ converges to the HJB equation (18) *with*

$$\hat{\mu}(s, q, \hat{\alpha}_k, \hat{\beta}_k) = \frac{\hat{\alpha}_k + s}{\hat{\beta}_k + q}, \quad \hat{\sigma}(s, q, \hat{\alpha}_k, \hat{\beta}_k) \equiv \hat{\sigma}.$$

See Appendix A for the proof of the above lemma.

Unstructured normal rewards   Consider the environment class $\boldsymbol{\nu} \in \mathbb{R}^K$ of normally distributed arm rewards $\mathcal{N}_\sigma$ with horizon n. For environment $\boldsymbol{\nu} \in \mathcal{E} = \mathbb{R}^K$, the rewards of the k-th arm follow the normal distribution $\mathcal{N}(\nu_k, \sigma^2(n))$, where ν_k is the k-th component of $\boldsymbol{\nu}$. We set the initial prior measure to be $\nu_k^1 \sim \mathcal{N}(\alpha_k(n), \beta_k(n)^2)$ for $1 \leq k \leq K$.

Lemma 4 *Let $f(n)$ be such that there exist real numbers $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\sigma})$ with*

$$\lim_{n \to \infty} \frac{\sigma^2(n)\alpha_k(n)}{f(n)\beta_k^2(n)} = \hat{\alpha}_k, \quad \lim_{n \to \infty} \frac{\sigma^2(n)}{\beta_k^2(n)n} = \hat{\beta}_k, \quad \lim_{n \to \infty} \frac{\sqrt{n}}{f(n)}\sigma(n) = \hat{\sigma}.$$

Then as $n \to \infty$, the rescaled Bayesian bandit problem with the scaling factor $f(n)$ converges to the continuous control problem (19) *with*

$$\hat{\mu}(s, q, \hat{\alpha}_k, \hat{\beta}_k) = \frac{\hat{\alpha}_k + s}{\hat{\beta}_k + q}, \quad \hat{\sigma}(s, q, \hat{\alpha}_k, \hat{\beta}_k) \equiv \hat{\sigma}.$$

See Appendix B for the proof of the above lemma.

Linear bandits with normal rewards   Consider the case of stochastic linear bandits with horizon n, where the environment is encoded by a vector $\boldsymbol{\nu} \in \mathbb{R}^d$. For environment $\boldsymbol{\nu} \in \mathcal{E} = \mathbb{R}^d$, the reward X^i at round i depends linearly on the chosen action $A^i \in \mathcal{A} \subset \mathbb{R}^d$ in the following sense:

$$X^i = \langle A^i, \boldsymbol{\nu} \rangle + \eta^i, \tag{20}$$

13

where $(\eta^i)_{i=1}^n$ is a sequence of independent and identically distributed normal random variables $\mathcal{N}(0, \sigma^2(n))$ with given σ. We set the prior measure of $\boldsymbol{\nu}$ to be the normal distribution $\mathcal{N}(\boldsymbol{\alpha}(n), \Sigma(n))$.

Lemma 5 *Let $f(n)$ be such that there exist real constants $(\hat{\boldsymbol{\alpha}}, \hat{\Sigma})$ with*

$$\lim_{n \to \infty} \frac{\sigma^2(n)}{n} \Sigma^{-1}(n) = \hat{\Sigma}^{-1}, \quad \lim_{n \to \infty} \frac{\sigma^2(n)}{f(n)} \Sigma^{-1}(n) \boldsymbol{\alpha}(n) = \hat{\boldsymbol{\alpha}}, \quad \lim_{n \to \infty} \frac{\sqrt{n}}{f(n)} \sigma(n) = \hat{\sigma}.$$

Then as $n \to \infty$, the rescaled Bayesian bandit problem with scaling factor $f(n)$ converges to the continuous control problem (19) with

$$\hat{\mu}(\mathbf{s}, \mathbf{q}, \mathbf{b}) = \mathbf{b}^\top \left(\hat{\Sigma}^{-1} + \sum_k q_k \mathbf{a}_k (\mathbf{a}_k)^\top \right)^{-1} \left(\hat{\boldsymbol{\alpha}} + \sum_{k=1}^K s_k \mathbf{a}_k \right), \quad \hat{\sigma}(\mathbf{s}, \mathbf{q}) \equiv \hat{\sigma}.$$

See Appendix C for the proof of the above lemma.

4. Approximate Bayes-optimal policy

What can one do with the limiting HJB equations or optimal control problems? In Section 4.1, we propose a regularized version of the Bayesian bandit by adding a regularizer term to the objective function of the optimal control limit. In this way, the resulting optimal policy will be a stochastic policy instead of a deterministic one, which can encourage exploration and make the solution less sensitive to perturbations. In Section 4.2, we propose an approximate Bayes-optimal policy algorithm, which is based on the solution to the (regularized or unregularized) HJB equation.

4.1 Regularized Bayesian bandits

The optimal policy from the optimal control problem (14) is deterministic and can be sensitive to small perturbations to the problem. To encourage robustness, one can add regularization to the objective function in (14), resulting in a stochastic optimal policy and encouraging more exploration and stability. For example, entropy regularization for the policy $-\sum_k \pi_k \log \pi_k$ can be added to (14):

$$\max_{\hat{\boldsymbol{\pi}}(\tau) \in \Delta^K} \mathbb{E} \left[\int_t^1 (\hat{\boldsymbol{\mu}}(\hat{\mathbf{s}}(\tau), \hat{\mathbf{q}}(\tau)) - \lambda \log \hat{\boldsymbol{\pi}}(\tau)) \cdot \hat{\boldsymbol{\pi}}(\tau) d\tau \right]$$

$$s.t. \quad d\hat{q}_k(\tau) = \hat{\pi}_k(\tau) dt, \quad 1 \le k \le K; \tag{21}$$

$$d\hat{s}_k(\tau) = \hat{\mu}_k(\hat{\mathbf{s}}(\tau), \hat{\mathbf{q}}(\tau)) \hat{\pi}_k(\tau) d\tau + \hat{\sigma}_k(\hat{\mathbf{s}}(\tau), \hat{\mathbf{q}}(\tau)) \sqrt{\hat{\pi}_k(\tau)} dB_\tau, \quad 1 \le k \le K;$$

$$\hat{\mathbf{s}}(t) = \hat{\mathbf{s}}, \quad \hat{\mathbf{q}}(t) = \hat{\mathbf{q}}.$$

The resulting HJB equation is (Evans, 2010)

$$\partial_t v + \max_{\hat{\boldsymbol{\pi}} \in \Delta^K} \left[\sum_{k=1}^K \left(\hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) \partial_{\hat{s}_k} v + \partial_{\hat{q}_k} v + \frac{1}{2} \hat{\sigma}_k^2(\hat{\mathbf{s}}, \hat{\mathbf{q}}) \partial_{\hat{s}_k}^2 v + \hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) \right. \right.$$

$$\left. \left. - \lambda \log(\hat{\pi}_k(t, \hat{\mathbf{s}}, \hat{\mathbf{q}})) \right) \hat{\pi}_k(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) \right] = 0. \tag{22}$$

The maximum in the above equation can be computed explicitly (Ying and Zhu, 2022). Let

$$H_k(p, m, h) = \hat{\mu}_k(\mathbf{s}, \mathbf{q})p + m + \frac{1}{2}\hat{\sigma}_k^2(\mathbf{s}, \mathbf{q})h + \hat{\mu}_k(\mathbf{s}, \mathbf{q}),$$

then the optimal policy is

$$\hat{\pi}_k^*(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) = \frac{1}{Z} \exp\left(\frac{1}{\lambda} H_k(\partial_{\hat{s}_k} v, \partial_{\hat{q}_k} v, \partial_{\hat{s}_k}^2 v)\right), \tag{23}$$

with normalizing constant $Z = \sum_k \exp\left(\frac{1}{\lambda} H_k(\partial_{s_k} v, \partial_{q_k} v, \partial_{s_k}^2 v)\right)$. Hence, (22) can be equivalently written as

$$\partial_t v + \lambda \log\left(\sum_k \exp\left(\frac{1}{\lambda} H_k(\partial_{\hat{s}_k} v, \partial_{\hat{q}_k} v, \partial_{\hat{s}_k}^2 v)\right)\right) = 0. \tag{24}$$

There are several potential advantages of the regularized version. First, the resulting optimal policy is always stochastic for $\lambda > 0$, which will be less sensitive to perturbations compared with deterministic optimal policy. Second, regularization encourages more exploration, which helps the performance when the initial prior is significantly different from the underlying truth. Third, regularization will usually lead to a smoother solution with a differentiable policy and value function, making it easier to numerically approximate the solution. For a more comprehensive introduction to the use of regularization in bandit and reinforcement learning problems, see, e.g., the tutorial by Geist et al. (2019).

4.2 Approximating the Bayes-optimal policy

Based on the limiting equation, if one can obtain the optimal policy for the HJB equation, then one can approximate the optimal Bayesian bandit policy by rescaling $(t, \hat{\mathbf{s}}, \hat{\mathbf{q}})$ to $(i, \mathbf{s}, \mathbf{q})$. This is summarized by the pseudocode in Algorithm 1.

Algorithm 1 Approximate Bayes-optimal policy

    for $i = 1, \ldots, K$ **do**
        $A_i \leftarrow i$, $s_i \leftarrow X_i$, $q_i \leftarrow 1$
    end for
    for $i = K+1, \ldots, n$ **do**
        Pull $A_i \sim \boldsymbol{\pi}^i(\mathbf{s}, \mathbf{q}) = \hat{\pi}_k^*(\frac{i-1}{n}, \frac{\mathbf{s}}{f(n)}, \frac{\mathbf{q}}{n})$ $\triangleright$ $\hat{\boldsymbol{\pi}}^*(t, \hat{\mathbf{s}}, \hat{\mathbf{q}})$ given in (12) (unregularized) or (23) (regularized)
        Get reward X_i
        $s_{A_i} \leftarrow s_{A_i} + X_i$, $q_{A_i} \leftarrow q_{A_i} + 1$
    end for

---

5. Solving the limiting HJB equation

One of the difficulties of the Bayesian bandit problem is its large computational cost. The computational complexity for solving a K-armed bandit with horizon n via backward induction is $O(n^{2K})$, which is intractable when n or K is large. If one can obtain the exact

optimal policy for the limiting HJB equation, then one can use it to approximate the Bayes-optimal policy for the finite horizon problem with almost no additional computational cost. Section 5.1 shows one of the cases where the exact solution can be obtained. Even if the exact solution cannot be obtained directly, Section 5.2 shows a numerical scheme to approximate the solution. The computational cost of numerically solving the HJB equation is $O(N^{2K})$, where N depends on the mesh of the scheme. This can be much more efficient than the discrete Bayesian bandit algorithm when n is large and K is small.

5.1 Exact solution

Although solving the HJB equation can also be challenging in general, it turns out that if $\mu_k(s,q) = \frac{s_k + \hat{\alpha}}{q_k + \hat{\beta}}$, one can obtain the exact solution. By Lemmas 3 and 4, two common bandit problems are exactly in this form.

Theorem 6 *If the drift term in the HJB equation (11) is*

$$\hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \frac{s_k + \hat{\alpha}_k}{q_k + \hat{\beta}_k},$$

then for any constants $(\hat{\alpha}_k, \hat{\beta}_k)$, the optimal policy for the unregularized HJB equation (11) is

$$\hat{\pi}_k^*(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) = \begin{cases} 1, & k = \underset{k}{\text{argmax}} \dfrac{\hat{s}_k + \hat{\alpha}_k}{\hat{q}_k + \hat{\beta}_k} \ ; \\ 0, & o.w. \end{cases} \tag{25}$$

and the optimal policy for the regularized HJB equation (22) is

$$\hat{\pi}_k^*(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) \propto \exp\left(\frac{1}{\lambda} \frac{\hat{s}_k + \hat{\alpha}_k}{\hat{q}_k + \hat{\beta}_k}\right). \tag{26}$$

The proof is given in Appendix D. Based on the above theorem, the approximate optimal policy for the unregularized Bayesian bandit problem is given by

$$\tilde{\pi}_k^{*,i}(\mathbf{s}, \mathbf{q}) = \hat{\pi}_k^*\left(\frac{i-1}{n}, \frac{\mathbf{s}}{f(n)}, \frac{\mathbf{q}}{n}\right) = \begin{cases} 1, & k = \underset{k}{\text{argmax}} \dfrac{s_k}{q_k + n\hat{\beta}_k} + \dfrac{f(n)\hat{\alpha}_k}{q + n\hat{\beta}_k} \\ 0, & o.w. \end{cases} \tag{27}$$

and the approximate optimal policy for the regularized Bayesian bandit problem is given by

$$\tilde{\pi}_k^{*,i}(\mathbf{s}, \mathbf{q}) = \hat{\pi}_k^*\left(\frac{i-1}{n}, \frac{\mathbf{s}}{f(n)}, \frac{\mathbf{q}}{n}\right) \propto \exp\left(\frac{n}{\lambda f(n)} \frac{s_k + f(n)\hat{\alpha}_k}{q_k + n\hat{\beta}_k}\right). \tag{28}$$

Furthermore, note that the approximate optimal policy (27) for the unregularized Bayesian bandit is similar in form to UCB: the first term is an approximation to the empirical mean, and the second term measures the degree to which the arm has been explored. The approximate optimal policy (28) for the regularized Bayesian bandit has a form similar to the tempered greedy algorithm, where the term $\frac{s_k + f(n)\hat{\alpha}_k}{q_k + n\hat{\beta}_k}$ is an approximation to the empirical mean and $\frac{n}{\lambda f(n)}$ adjusts the exploration rate. When $\frac{n}{\lambda f(n)}$ is smaller (i.e., when the regularization constant λ is larger), there is more exploration.

5.2 Numerical solution

In the general case, an exact solution to the HJB equation (11) is not available, so we present a numerical scheme to approximate the solution in this section. In certain cases, the numerical scheme yields the exact optimal policy and value function (see Lemma 7). In general, when the horizon is large, the computational cost of numerically solving the PDE will be much less than that of classical Bayesian bandit algorithms while still yielding a good approximation to the optimal policy.

First, observe that one can directly compute the maximum in the HJB equation (11). Let

$$k^*(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) = \mathrm{argmax}_k \ \partial_{\hat{q}_k} v + \hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}})\partial_{\hat{s}_k} v + \frac{1}{2}\hat{\sigma}_k^2(\hat{\mathbf{s}}, \hat{\mathbf{q}})\partial_{\hat{s}_k}^2 v + \hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}),$$

then the optimal policy is

$$\pi_k^*(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) = \begin{cases} 1, & k = k^*(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) \\ 0, & k \neq k^*(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) \end{cases}$$

and the optimal value function satisfies

$$\partial_t v + \partial_{\hat{q}_{k^*}} v + \hat{\mu}_{k^*}(\hat{\mathbf{s}}, \hat{\mathbf{q}})\partial_{\hat{s}_{k^*}} v + \frac{1}{2}\hat{\sigma}_{k^*}^2(\hat{\mathbf{s}}, \hat{\mathbf{q}})\partial_{\hat{s}_{k^*}}^2 v + \hat{\mu}_{k^*}(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = 0. \tag{29}$$

All of the above results hold in the deterministic case when $\hat{\sigma}_k \equiv 0$ for all k. Based on (29), we present a finite difference method for solving the HJB equation (29).

HJB equation with diffusion First, consider the case where $\hat{\sigma} \not\equiv 0$. We discretize the time interval $[0, 1]$ via the grid points $0 = t^0 < t^1 < \cdots < t^{N_t} = 1$, where $t^l = l\delta_t$ and $\delta_t = 1/N_t$. We impose a cutoff on the cumulative reward in $\mathbb{R}^K$ so that it lies within $[-S, S]^K$, then further discretize this clipped interval into $-S \leq s_k^{\mathbf{i}} \leq S$. Here $\mathbf{i} \in \{\mathbb{Z}^K, -N_s \leq i_k \leq N_s\}$ is a K-dimensional index vector, and $\mathbf{s}^{\mathbf{i}} = \mathbf{i}\delta_s$ with $\delta_s = S/N_s$. Finally, we discretize the number of pulls to $0 \leq q_k^{\mathbf{j}} \leq 1$, where $\mathbf{j} \in \{\mathbb{Z}^K, 0 \leq j_k \leq N_q\}$ is a K-dimensional index vector, and $\mathbf{q}^{\mathbf{j}} = \mathbf{j}\delta_q$ with $\delta_q = 1/N_q$. Observe that at each time t^l, one has that $0 \leq q_k^j \leq t^l$ and $\sum_{k=1}^K q_k^{\mathbf{j}} = t^l$. We will typically set $N_q = N_t$ and $\delta_q = \delta_t$.

Let $\tilde{v}_{\mathbf{i},\mathbf{j}}^l$ be the approximation for $v(t^l, \mathbf{s}^{\mathbf{i}}, \mathbf{q}^{\mathbf{j}})$, where $\mathbf{i}, \mathbf{j} \in \mathbb{Z}^K$. The numerical approximation $\tilde{v}_{\mathbf{i},\mathbf{j}}^l$ satisfies the following equation:

$$
\begin{aligned}
&\frac{1}{\delta_t}(\tilde{v}_{\mathbf{i},\mathbf{j}}^{l+1} - \tilde{v}_{\mathbf{i},\mathbf{j}}^l) + \frac{1}{2\delta_s}\hat{\mu}_{k^*}^{\mathbf{i},\mathbf{j}}(\tilde{v}_{\mathbf{i}+\mathbf{e}_{k^*},\mathbf{j}+\mathbf{e}_{k^*}}^{l+1} - \tilde{v}_{\mathbf{i}-\mathbf{e}_{k^*},\mathbf{j}+\mathbf{e}_{k^*}}^{l+1}) + \frac{1}{\delta_q}(\tilde{v}_{\mathbf{i},\mathbf{j}+\mathbf{e}_{k^*}}^{l+1} - \tilde{v}_{\mathbf{i},\mathbf{j}}^{l+1}) \\
&+ \frac{1}{2\delta_s^2}\hat{\sigma}_{k^*}^{\mathbf{i},\mathbf{j}}(\tilde{v}_{\mathbf{i}+\mathbf{e}_{k^*},\mathbf{j}+\mathbf{e}_{k^*}}^{l+1} - 2\tilde{v}_{\mathbf{i}+\mathbf{e}_{k^*},\mathbf{j}+\mathbf{e}_{k^*}}^{l+1} + \tilde{v}_{\mathbf{i}-\mathbf{e}_{k^*},\mathbf{j}+\mathbf{e}_{k^*}}^{l+1}) + \hat{\mu}_{k^*}^{\mathbf{i},\mathbf{j}} = 0,
\end{aligned} \tag{30}
$$

where $\mathbf{e}_k$ is the k-th standard basis vector, $\hat{\mu}_k^{\mathbf{i},\mathbf{j}} = \hat{\mu}_k(\mathbf{s}^{\mathbf{i}}, \mathbf{q}^{\mathbf{j}})$, and $\hat{\sigma}_k^{\mathbf{i},\mathbf{j}} = \hat{\sigma}_k(\mathbf{s}^{\mathbf{i}}, \mathbf{q}^{\mathbf{j}})$. Since the boundary conditions at the terminal time specify that $\tilde{v}_{\mathbf{i},\mathbf{j}}^{N_t} = 0$, one can use the preceding equations to solve backward in time and compute all of the values $\tilde{v}_{\mathbf{i},\mathbf{j}}^l$.

In fact, if one defines $g_{\mathbf{i},\mathbf{j}}^l$ as a vector with k-th element given by

$$
\begin{aligned}
(g_{\mathbf{i},\mathbf{j}}^l)_k =& \tilde{v}_{\mathbf{i},\mathbf{j}}^{l+1} + \delta_t \left[\frac{1}{2\delta_s}\hat{\mu}_k^{\mathbf{i},\mathbf{j}}(\tilde{v}_{\mathbf{i}+\mathbf{e}_k,\mathbf{j}+\mathbf{e}_k}^{l+1} - \tilde{v}_{\mathbf{i}-\mathbf{e}_k,\mathbf{j}+\mathbf{e}_k}^{l+1}) + \frac{1}{\delta_q}(\tilde{v}_{\mathbf{i},\mathbf{j}+\mathbf{e}_k}^{l+1} - \tilde{v}_{\mathbf{i},\mathbf{j}}^{l+1}) \right. \\
&\left. + \frac{1}{2\delta_s^2}\hat{\sigma}_k^{\mathbf{i},\mathbf{j}}(\tilde{v}_{\mathbf{i}+\mathbf{e}_k,\mathbf{j}+\mathbf{e}_k}^{l+1} - 2\tilde{v}_{\mathbf{i},\mathbf{j}+\mathbf{e}_k}^{l+1} + \tilde{v}_{\mathbf{i}-\mathbf{e}_k,\mathbf{j}+\mathbf{e}_k}^{l+1}) + \hat{\mu}_k^{\mathbf{i},\mathbf{j}} \right],
\end{aligned} \tag{31}
$$

then one gets an equivalent form for the numerical scheme (30) with the correspondence given by

$$k^* = \mathrm{argmax}_k (g^l_{\mathbf{i,j}})_k, \quad \text{and} \quad \tilde{v}^l_{\mathbf{i,j}} = (g^l_{\mathbf{i,j}})_{k^*}, \quad (\tilde{\pi}^l_{\mathbf{i,j}})_k = \begin{cases} 1, & k = k^* \\ 0, & k \neq k^* \end{cases} \tag{32}$$

Note that for this scheme to be numerically stable, δ_t and δ_s must satisfy the inequality

$$\delta_t \leq \min(\hat{\sigma}^k_{\mathbf{i,j}})^2 \delta_s^2.$$

Since $\tilde{v}^l_{\mathbf{i,j}}$ is only defined on grid points, the continuous approximated solution $\tilde{v}(t, \hat{\mathbf{s}}, \hat{\mathbf{q}})$ is defined as

$$\tilde{v}(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) = \tilde{v}^l_{\mathbf{i,j}} \quad \text{for} \quad \begin{cases} l\delta_t \leq t < (l+1)\delta_t \\ i_k \delta_s \leq \hat{s}_k < (i_k + 1)\delta_s \\ j_k \delta_q \leq \hat{q}_k < (j_k + 1)\delta_q. \end{cases} \tag{33}$$

HJB equation without diffusion Second, we consider the case where $\hat{\sigma} \equiv 0$. Since there is no longer a diffusion term, one should use the upwind scheme for the transport term. Let

$$\hat{\mu}^{\mathbf{i,j}}_{k,+} = \max(\hat{\mu}^{\mathbf{i,j}}_k, 0), \qquad \hat{\mu}^{\mathbf{i,j}}_{k,-} = \min(\hat{\mu}^{\mathbf{i,j}}_k, 0),$$

$$(g^l_{\mathbf{i,j}})_k = \tilde{v}^{l+1}_{\mathbf{i,j}} + \delta_t \left[\frac{\hat{\mu}^{\mathbf{i,j}}_{k,+}}{\delta_s} (\tilde{v}^{l+1}_{\mathbf{i+e}_k,\mathbf{j+e}_k} - \tilde{v}^{l+1}_{\mathbf{i,j+e}_k}) + \frac{\hat{\mu}^{\mathbf{i,j}}_{k,-}}{\delta_s} (\tilde{v}^{l+1}_{\mathbf{i,j+e}_k} - \tilde{v}^{l+1}_{\mathbf{i-e}_k,\mathbf{j+e}_k}) \right.$$
$$\left. + \frac{1}{\delta_q} (\tilde{v}^{l+1}_{\mathbf{i,j+e}_k} - \tilde{v}^{l+1}_{\mathbf{i,j}}) + \hat{\mu}^{\mathbf{i,j}}_k \right]. \tag{34}$$

Then

$$k^* = \mathrm{argmax}_k (g^l_{\mathbf{i,j}})_k \quad \text{and} \quad \tilde{v}^l_{\mathbf{i,j}} = (g^l_{\mathbf{i,j}})_k, \quad (\tilde{\pi}^l_{\mathbf{i,j}})_k = \begin{cases} 1, & k = k^* \\ 0, & k \neq k^*. \end{cases} \tag{35}$$

In this case, the stability conditions imply that

$$\max(\hat{\mu}^{\mathbf{i,j}}_k)\delta_t \leq \delta_s.$$

Connection to the Bayesian bandit algorithm In certain cases, the numerical schemes (31)-(32) and (34)-(35) give the exact optimal value function for the finite horizon problem.

Lemma 7 *For the Bernoulli bandits introduced in Section 3.1, when the initial hyperparameters $(\alpha_k, \beta_k) = (c_1 n, c_2 n)$ for some constants (c_1, c_2), then the numerical scheme (34)-(35) for the limiting HJB equation based on the scaling factor $f(n) = n$ gives the exact optimal value function when $\delta_t = \delta_q = \delta_s = \frac{1}{n}$.*

Similarly, for the binomial bandits described in Section 3.4.1 with γ being a constant independent of n, when the initial hyperparameters $(\alpha_k, \beta_k) = (c_1 n + c_2\sqrt{n}, c_1 - c_2\sqrt{n})$ for some constants (c_1, c_2), then the numerical scheme (31)-(32) for the limiting HJB equation based on the scaling factor $f(n) = \sqrt{n}$ gives the exact optimal value function when $\delta_t = \delta_q = \frac{1}{n}, \delta_s = \frac{\gamma}{\sqrt{n}}$.

See Appendix E for the proof of the above lemma.

6. Numerical experiments

6.1 Convergence to the HJB equation

In this section, we will show the convergence of the Bayes-optimal solution to the HJB solution as the horizon goes to infinity. Namely, we would like to show the differences in the optimal policy and the rescaled optimal cumulative reward

$$\pi^{i,n}(\mathbf{s},\mathbf{q}) - \hat{\pi}^n\left(\frac{i-1}{n}, \frac{\mathbf{s}}{f(n)}, \frac{\mathbf{q}}{n}\right), \quad \frac{1}{f(n)}w^{i,n}(\mathbf{s},\mathbf{q}) - v\left(\frac{i-1}{n}, \frac{\mathbf{s}}{f(n)}, \frac{\mathbf{q}}{n}\right)$$

decay as the horizon $n \to \infty$, where $\pi^{i,n}(\mathbf{s},\mathbf{q}), w^i(\mathbf{s},\mathbf{q})$ are obtained by backward induction and $\hat{\pi}(t,\hat{\mathbf{s}},\hat{\mathbf{q}}), v(t,\hat{\mathbf{s}},\hat{\mathbf{q}})$ are obtained by solving the corresponding HJB equation. We show the convergence result in Figures 2 and 3. Below are the details of the plots.

Consider the one-armed Bernoulli bandit problem, where the first arm has a reward 1 with probability ν and -1 with probability $1 - \nu$, while the second arm has a deterministic reward μ_2. In this case, given a prior measure $\nu \sim \text{Beta}(\alpha, \beta)$, one can obtain the exact optimal policy and cumulative reward $\pi^{i,n}(s,q)$ and $w^{i,n}(s,q)$ via the equations

$$\pi_k^{i,n}(s,q) = \begin{cases} 1, & \text{for} \quad k = \underset{k}{\text{argmax}}\ \hat{w}_k^{i,n}(s,q) \\ 0, & o.w. \end{cases} \quad ; \quad w^{i,n}(s,q) = \max_k \hat{w}_k^{i,n}(s,q),$$

where

$$\hat{w}_1^{i,n}(s,q) = p(s,q)w^{i+1,n}(s+1,q+1) + (1 - p(s,q)w^{i+1,n}(s-1,q+1)),$$
$$\hat{w}_2^{i,n}(s,q) = w^{i+1,n}(s,q) + \mu_2,$$

with $w^{n+1,n}(s,q) = 0$ for all (s,q) and $p(s,q) = \frac{\alpha+s/2+q/2}{\alpha+\beta+q}$. The limiting HJB equation depends on the scaling factor $f(n)$. By Lemma 3, one arrives at a stochastic optimal control problem if $f(n) = \sqrt{n}$ and a deterministic one if $f(n) = n$. We will compare $\pi^{i,n}, w^{i,n}$ with the limiting HJB solution for both of these scenarios.

In Figures 2 and 3, we set $\mu_2 = 1/(3\sqrt{n})$. The hyperparameters (α, β) for the initial prior measure are set to be $(n, n - \sqrt{n})$, which implies that in the limiting HJB equation

$$\partial_t v + \max_{\pi(s,q)\in[0,1]}\left[\hat{\mu}\partial_{\hat{s}}v + \partial_{\hat{q}}v + \frac{1}{2}\hat{\sigma}^2\partial_{\hat{s}}^2 v + \hat{\mu} - \mu_2\right] + \mu_2 = 0,$$

$\hat{\mu} = \frac{1+s}{2+q}, \hat{\mu}_2 = 1/3, \hat{\sigma} = 1$ for $f(n) = \sqrt{n}$, and $\hat{\mu} = \frac{s}{2+q}, \hat{\mu}_2 = 0, \hat{\sigma} = 0$ for $f(n) = n$. By Theorem 6, one can obtain the exact optimal policy for the limiting HJB equation. In Figure 2, we plot the average difference over i, s, q. That is, we plot

$$\begin{aligned} e_\pi^n &= \frac{1}{Z}\sum_{i,s,q}\left|\pi^{i,n}(s,q) - \hat{\pi}\left(\frac{i-1}{n}, \frac{s}{f(n)}, \frac{q}{n}\right)\right|, \\ e_w^n &= \frac{1}{Z}\sum_{i,s,q}\left|\frac{1}{f(n)}w^{i,n}(s,q) - v\left(\frac{i-1}{n}, \frac{s}{f(n)}, \frac{q}{n}\right)\right|, \end{aligned} \tag{36}$$

where the summation is over $i \in \{1, \cdots, n\}, s \in \{-(i-1), \cdots, i-1\}, q \in \{0, \cdots, i-1\}$, and Z is the number of summations.

Next, we test the difference between the Bayes-optimal solution and the numerical solution to the HJB equation presented in Section 6. The following averaged differences are plotted in Figure 3:

$$e_\pi^{n,N} = \frac{1}{Z} \sum_{i,s,q} \left| \pi^{i,n}(s,q) - \tilde{\pi}^N \left(\frac{i-1}{n}, \frac{s}{f(n)}, \frac{q}{n} \right) \right|,$$

$$e_w^{n,N} = \frac{1}{Z} \sum_{i,s,q} \left| \frac{1}{f(n)} w^{i,n}(s,q) - \tilde{v}^N \left(\frac{i-1}{n}, \frac{s}{f(n)}, \frac{q}{n} \right) \right|, \tag{37}$$

where $\tilde{v}^N$ is obtained according to the scheme (31)-(32) with $\delta_t = \delta_q = N^{-1}, \delta_s = N^{-1/2}$ when $f(n) = \sqrt{n}$; and $\tilde{v}^N$ is according to the scheme (34) - (35) with $\delta_t = \delta_q = \delta_s = N^{-1}$ when $f(n) = n$. The difference $e_w^{n,N}$ based on $f(n) = \sqrt{n}$ is rescaled by $\frac{1}{\sqrt{n}}$ so that it is on the same scale as the scheme based on $f(n) = n$. Due to the different discretizations, $\tilde{v}^N \left(\frac{i-1}{n}, \frac{s}{f(n)}, \frac{q}{n} \right)$ is not necessarily on a grid point, so we define the continuous approximation solution $\tilde{v}(t,s,q)$ as in (33), which implies that

$$\tilde{\pi}^N \left(\frac{i-1}{n}, \frac{s}{f(n)}, \frac{q}{n} \right) = \tilde{\pi}_{m,j}^l, \quad \tilde{v}^N \left(\frac{i-1}{n}, \frac{s}{f(n)}, \frac{q}{n} \right) = \tilde{v}_{m,j}^l, \tag{38}$$

for

$$l = \left\lfloor \frac{i-1}{n\delta_t} \right\rfloor, m = \left\lfloor \frac{s}{f(n)\delta_s} \right\rfloor, j = \left\lfloor \frac{q}{n\delta_q} \right\rfloor,$$

where $\lfloor x \rfloor$ is the largest integer less than or equal to x.

Figure 2 shows that the difference e_π^n decays as n increases for both scaling factors. The stochastic limit according to the scaling factor $f(n) = \sqrt{n}$ is closer to the optimal Bayesian solution compared with the deterministic limit. Figure 3 shows that the difference $e^{n,N}$ decays as n and N increase. Note that $e^{n,N}$ has two components: model error and numerical error. Model error decreases as the horizon n increases. The numerical error decreases as the number of grid points N increases. We can see from Figure 3 that when both the horizon n and the number of grid points N increase, the differences decrease. We observe that when $N = 50$, the difference $e_v^{n,N}$ in the value functions decreases slower or does not decrease after n reaches some threshold. This is because the numerical error dominates over the model error in this regime.

6.2 The performance of the approximate Bayes-optimal policy

We compare the performance of the approximate Bayes-optimal policy (Algorithm 1) with Thompson sampling and UCB in terms of the expected regret. (See Appendix F for the details of Thompson sampling and UCB.)

For unstructured bandits, we consider normal arm rewards, in which case the exact policy for the limiting HJB equation can be directly obtained. The performance of the three algorithms is shown in Figure 4. Below are the details of the plots.

Consider the K-armed normal bandit problem with $K \geq 2$. Assume that the first arm follows $r_1 \sim N(\mu_1, 1)$ for $\mu_1 \equiv 0$, while the k-th arm follows $r_k \sim N(\mu, 1)$ for $2 \leq k \leq K$. Note that although $\mu_1 \equiv 1$, this is unknown to us. We define

$$\Delta = \mu - \mu_1 \tag{39}$$

20

Convergence to the solution of the limiting HJB equation
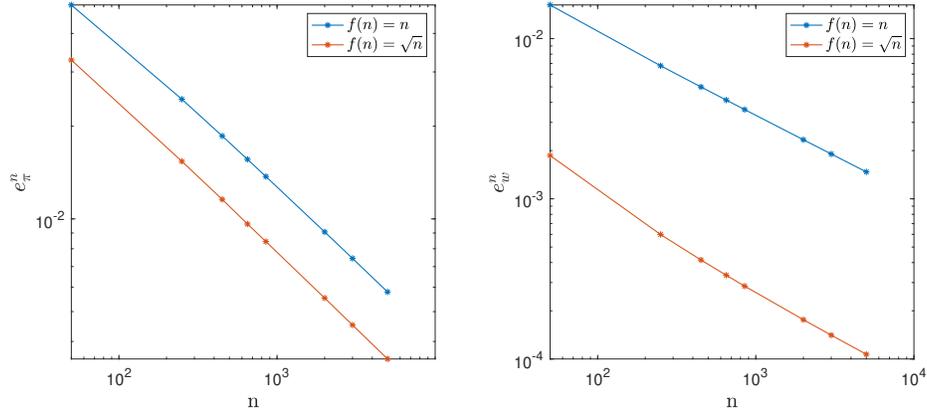
Figure 2: The above plot shows the decay of the difference between the Bayes-optimal solution and the solution to the HJB equation as n increases, i.e., e_π^n and e_w^n defined in (36). Here $f(n)$ is the scaling factor. When $f(n) = \sqrt{n}$, the resulting limit is a stochastic optimal control problem, while when $f(n) = n$, the resulting limit is a deterministic one.

Convergence to the numerical solution of the limiting HJB equation
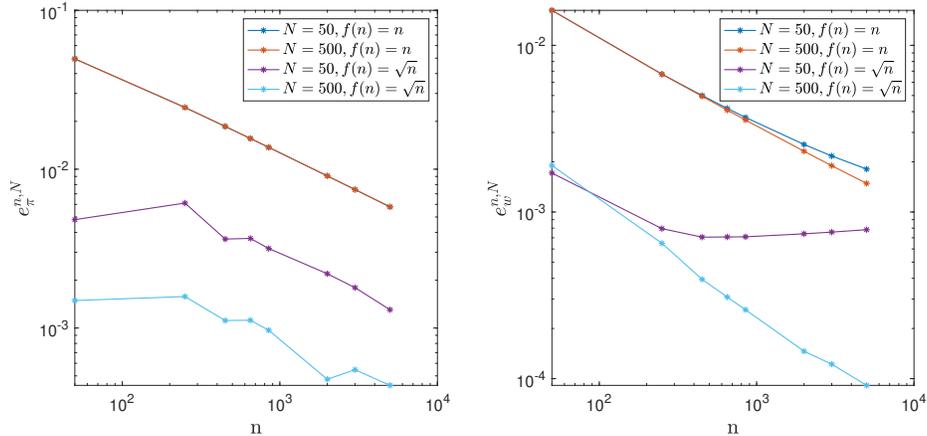
Figure 3: The above plot shows the decay of the difference between the Bayes-optimal solution and the numerical solution to the HJB equation as n and N increase, i.e., $e_\pi^{n,N}$ and $e_w^{n,N}$ defined in (37). Here N is the number of grid points when numerically solving the HJB equation, and $f(n)$ is the scaling factor. When $f(n) = \sqrt{n}$, the resulting limit is a stochastic optimal control problem, while when $f(n) = n$, the resulting limit is a deterministic one. The curve for $N = 50, f(n) = n$ is indistinguishable from the curve for $N = 500, f(n) = n$ in $e_\pi^{n,N}$, so it is not shown on the right plot.

to be the arm gap. The horizon is set to be $n = 10^3$. For the proposed method (Algorithm 1), we set the scaling factor $f(n) = \sqrt{n}$, that is, the limiting optimal control problem is stochastic. The exact solution to the limiting HJB equation can be obtained by Theorem 6. The initial prior measure for both the Bayes-optimal policy and Thompson sampling is $\nu_k \sim N(\frac{1}{\sqrt{n}}, \frac{1}{n})$ for all k. This implies that the limiting HJB equation is (11) with $\hat{\mu}_k(\mathbf{s}, \mathbf{q}) = \frac{s_k+1}{q_k+1}$ and $\hat{\sigma} \equiv 1$. In addition, $\delta = n^2$ for the UCB algorithm. Figure 4 shows the expected regret of the three algorithms for $\Delta \in [-1, 1]$ and $K = 5, 10, 20$. The expected regret is averaged over 10^3 simulations.

We can see from Figure 4 that the overall performance of the approximate Bayes-optimal policy is better than the other two algorithms, especially when the prior guess is close to the underlying environment. When Δ approaches -1, UCB is a bit better than the approximate Bayes-optimal policy because the prior guess is significantly different from the underlying truth. However, note that as the number of arms increases, the performance is almost the same, even around $\Delta = -1$.
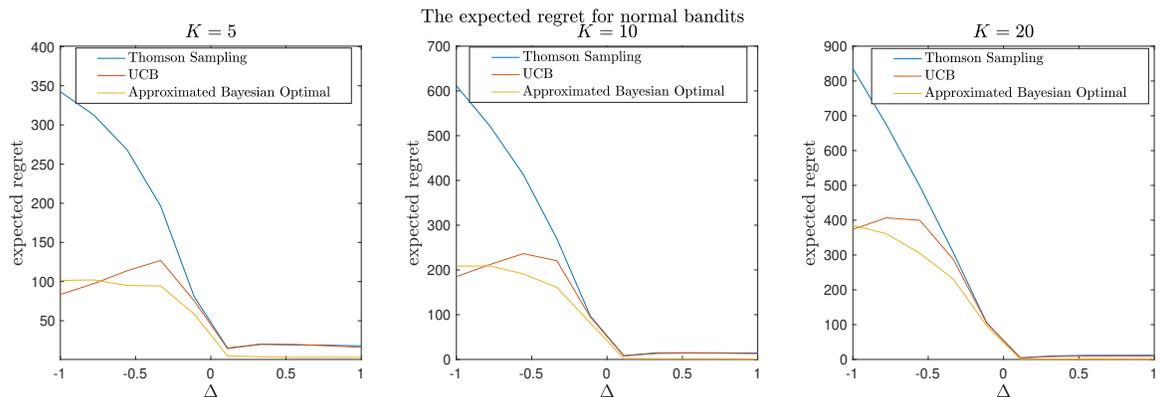
Figure 4: The above plot shows the expected regret in the K-armed normal bandit problem for the approximate Bayes-optimal policy, Thompson sampling, and UCB. The arm gap $\Delta \in [-1, 1]$ is defined in (39). The left, middle, and right plots correspond to $K = 5, 10, 20$.

For structured bandits, we consider the linear bandits described in Section 3.4.1. Assume there are two arms, and the reward for arm a_i follows $a_i\nu + \eta$ with unknown ν and $\eta \sim \mathcal{N}(0, 1)$. We set the initial measure for $\nu \sim \mathcal{N}(0, \frac{1}{n})$, and take the scaling factor $f(n) = \sqrt{n}$, then $\hat{\mu}, \hat{\sigma}$ for the limiting HJB equation can be obtained according to Lemma 5. We solve the limiting HJB equation by the numerical scheme (31)-(32) with $\delta_t = \delta_q = \frac{1}{N}$, $\delta_s = \frac{1}{\sqrt{N}}$ and $N = 100$. The performance in terms of the expected regret is plotted in Figure 5, where we test three different action positions $(a_1, a_2) = (0.1, -0.1), (0.1, -0.2), (0.1, 0.2)$.

Figure 5 shows that the overall performance of the approximate Bayes-optimal policy is more robust than the other two methods. First, the approximate Bayes-optimal always outperforms TS. When $(a_1, a_2) = (0.1, -0.1)$, the performance of UCB and approximate Bayes-optimal policy are similar. However, for the other two cases where $(a_1, a_2) =$

$(0.1, -0.2), (0.1, 0.2)$, UCB has much worse performance on one side, while the approximate Bayes-optimal policy has evener regret on both sides. If one measures the performance in the worst-case regret or in the averaged regret over the possible environments, the approximate Bayes-optimal policy outperforms the other two.
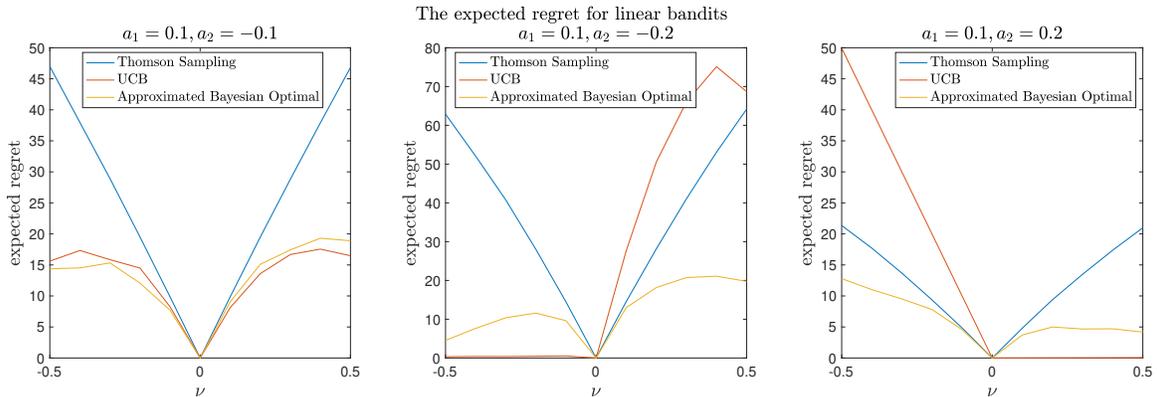
Figure 5: The above plot shows the expected regret in the 2-armed linear bandit problem for the approximate Bayes-optimal policy, Thompson sampling, and UCB. The arm gap of the normal bandits $\Delta \in [-1, 1]$ is defined in (39). The environment of the linear bandits $\nu \in [-1/2, 1/2]$ is defined in (20).

We also show the performance of the regularized approximate Bayes-optimal policy in Figure 6. We compare the regularized Bayes-optimal policy with the unregularized version for normal bandits and linear bandits. The setting of the two bandit problem is the same as Figures 4 and 5, but the initial Bayesian prior of the two bandits are worse (farther from the ground truth). We set $\nu_k \sim N(0.01\sqrt{n}, 1)$ for the normal bandits, and $\nu \sim \mathcal{N}(\sqrt{n}, 1)$ for the linear bandits. One can see from Figure 6 that the regularized version performs similarly to or better than the unregularized version when the initial prior measure is bad.

We remark that since the solution to the regularized HJB equation is always smooth when $\lambda > 0$, it is also potentially easier to break the curse of dimensionality. However, we leave the high-dimensional problem for future study.

7. Discussion

In this paper, we derived a continuous-in-time limit for the Bayesian bandit problem. We showed that the rescaled optimal cumulative reward converges to the solution of an HJB equation. We derive several benefits from the limiting PDE:

- **A single recipe for many Bayesian bandits.** For most multi-armed bandit problems, the classical Bayesian bandit algorithms yield a formulation that cannot be solved accurately. Different recipes are required to approximate the value function in different settings. On the other hand, the limiting PDE gives a single, unified recipe to solve the Bayes-optimal policy.

The expected regret of the regularized approximated bayesian algorithm
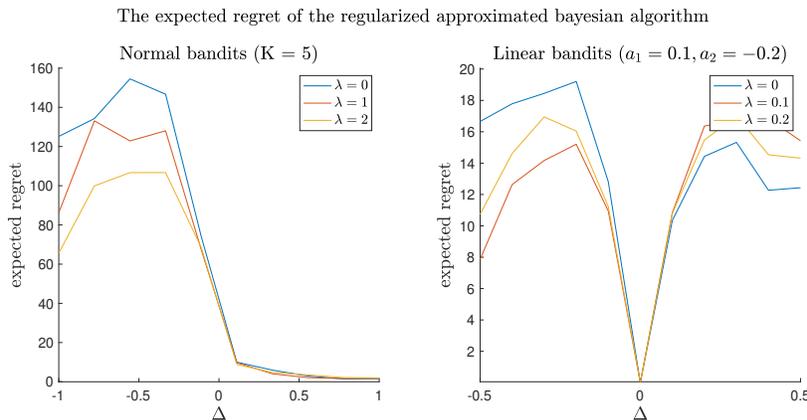
Figure 6: The above plot shows the expected regret of a 5-armed normal bandit problem and a 2-armed linear bandit problem for the regularized approximate Bayes-optimal policy. The environment $\nu \in [-1/2, 1/2]$ is defined in (20). The left, middle and right plots correspond to $(a_1, a_2) = (0.1, -0.1), (0.1, -0.2), (0.1, 0.2)$.

For example, one way to solve the one-armed bandit problem with normal distributions using a Bayesian bandit algorithm is to solve the following equation backward:

$$
W^i(\mu, q) = \max \left\{ W^{i+1}(\mu, q) + \mu_2, \mu + \frac{1}{2\pi} \int_{\mathbb{R}} \exp(-\frac{x}{2\sigma_i^2}) w^{i+1}(\mu + x, q + 1) dx \right\}
$$

for all $\mu \in \mathbb{R}$ and $i = 1, \cdots, n$ with $\sigma_i = (q^i + \sigma^{-2})^{-1}$. Since there is no closed-form solution for the integral in the above equation, one must approximate W^i using piecewise quadratic functions (see e.g. Section 35.3.2 of Lattimore and Szepesvári (2020) for details). This results in a completely different algorithm for solving this problem compared to solving the Bernoulli reward case. However, if one instead uses the HJB equation to solve the one-armed bandit problem with normally distributed rewards, the same formulation (18) which applies for Bernoulli rewards also applies for normal rewards. The only modification is the different forms of $\hat{\mu}$ according to Lemma 4.

- **Improved efficiency for large n.** The classical Bayesian bandit algorithm requires a computational cost of $O(n^{2K})$ to calculate the optimal policy, which can be prohibitive for large n.

On the other hand, as $n \to \infty$, the Bayesian bandit problem converges to the continuous HJB equation. The computational cost of solving the HJB equation is independent of the horizon, and it only depends on the numerical discretization of the PDE, which is $O(N^{2K})$. When $N \ll n$, one obtains huge computational savings by solving for the HJB value function instead. Since the accuracy of the approximation solution is $\tilde{v} - v = O(N^{-1})$, one retains an accurate approximation of the solution with much less computational cost.

- **Improved efficiency for large K** For the case where the exact solution can be obtained for the limiting HJB equation as stated in Theorem 6, there is no computational cost of solving the limiting equation. In this case, even if K is large, one can approximate the Bayes-optimal policy efficiently. For the case where the exact solution to the HJB equation cannot be obtained, it may be possible to break the curse of dimensionality numerically by using a non-linear function approximation, such as a deep neural network. We leave the high-dimensional problem for future study.

One can also extend the current framework of finite arms to infinite arms, for instance, with a continuous action space. The policy $\pi(t, \mathbf{s}, \mathbf{q}, \mathbf{a})$ is a probability density function such that $\int \pi(t, \mathbf{s}, \mathbf{q}, \mathbf{a}) d\mathbf{a} = 1$ for all $t, \mathbf{s}, \mathbf{q}$. The general limiting control problem takes the form

$$
\begin{aligned}
\max_{\int \hat{\pi} d\mathbf{a}=1} \quad & \mathbb{E}\left[\int_t^1 \int_{\mathcal{A}} \hat{\mu}(\hat{\mathbf{s}}, \hat{\mathbf{q}}, \mathbf{a})\hat{\pi}(\tau, \mathbf{a}) da d\tau\right] \\
s.t. \quad & d\hat{q}(\tau, \mathbf{a}) = \hat{\pi}(\tau, \mathbf{a})dt, \quad \forall \mathbf{a} \in \mathcal{A}; \\
& d\hat{s}(\tau, \mathbf{a}) = \hat{\mu}(\hat{\mathbf{s}}, \hat{\mathbf{q}}, \mathbf{a})\hat{\pi}(\tau, \mathbf{a})dt + \hat{\sigma}(\hat{\mathbf{s}}, \hat{\mathbf{q}}, \mathbf{a})\sqrt{\hat{\pi}(\tau, \mathbf{a})}dB_t, \quad \forall \mathbf{a} \in \mathcal{A}; \\
& \hat{\mathbf{s}}(t, \mathbf{a}) = \mathbf{s}, \quad \hat{\mathbf{q}}(t, \mathbf{a}) = \mathbf{q}.
\end{aligned}
\tag{40}
$$

We will leave the study of the above case to future research.

References

Victor F Araman and René A Caldentey. Diffusion approximations for a class of sequential experimentation problems. *Management Science*, 68(8):5958–5979, 2022.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Dirk Bergemann and Juuso Valimaki. Bandit problems. 2006.

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

Russell N Bradt, SM Johnson, and Samuel Karlin. On sequential designs for maximizing the sum of n observations. *The Annals of Mathematical Statistics*, 27(4):1060–1074, 1956.

Yeon-Koo Che and Johannes Hörner. Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics*, 133(2):871–925, 2018.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Zhengdao Chen, Grant Rotskoff, Joan Bruna, and Eric Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks. *Advances in Neural Information Processing Systems*, 33:22217–22230, 2020.

Nicole El Karoui, Monique Jeanblanc, and Vincent Lacoste. Optimal portfolio management with american capital guarantee. *Journal of Economic Dynamics and Control*, 29(3):449–468, 2005.

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Soc., 2010.

Lin Fan and Peter W Glynn. Diffusion approximations for thompson sampling. *arXiv preprint arXiv:2105.09232*, 2021.

Kris Johnson Ferreira, David Simchi-Levi, and He Wang. Online network revenue management using thompson sampling. *Operations research*, 66(6):1586–1602, 2018.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019.

John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

Yonatan Gur, Gregory Macnamara, and Daniela Saban. Sequential procurement with contractual and experimental learning. *Management Science*, 68(4):2714–2731, 2022.

Jiequn Han, Qianxiao Li, et al. A mean-field optimal control formulation of deep learning. *Research in the Mathematical Sciences*, 6(1):1–41, 2019.

Michael Kapralov and Rina Panigrahy. Prediction strategies without loss. *Advances in Neural Information Processing Systems*, 24, 2011.

Vladimir A Kobzar and Robert V Kohn. A pde-based analysis of the symmetric two-armed bernoulli bandit. *arXiv preprint arXiv:2202.05767*, 2022.

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

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

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Benoît Leloup and Laurent Deveaux. Dynamic pricing on the internet: Theory and simulations. *Electronic Commerce Research*, 1(3):265–276, 2001.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2017.

Qiang Liu. Stein variational gradient descent as gradient flow. *Advances in neural information processing systems*, 30, 2017.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Makiko Nisio. Stochastic control theory. *ISI Lecture Notes*, 9, 2015.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27, 2014.

Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

William R 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.

Stefan Wager and Kuang Xu. Diffusion asymptotics for sequential experiments. *arXiv preprint arXiv:2101.09855*, 2021.

Abraham Wald. *Sequential analysis*. Courier Corporation, 2004.

Zhilei Wang and Robert V Kohn. A new approach to drifting games, based on asymptotically optimal potentials. *arXiv preprint arXiv:2207.11405*, 2022.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

Lexing Ying and Yuhua Zhu. A note on optimization formulations of markov decision processes. *Communications in Mathematical Sciences*, 20(3):727–745, 2022.

Appendices

A Proof of Lemma 3

The expectation of the k-th arm in environment $\boldsymbol{\nu}$ is $\mu_k(\boldsymbol{\nu}) = \gamma(2\nu_k - 1)$. At the start of the i-th round, the posterior distribution of ν_k is uniquely determined by the cumulative reward $s_k^i = \sum_{j=1}^{i-1} X^j \mathbb{1}_{A^j=k}$ and the number of pulls $q_k^i = \sum_{j=1}^{i-1} \mathbb{1}_{A^j=k}$ through round $i-1$ according to

$$\nu_k^i \sim \text{Beta}(\alpha_k^i, \beta_k^i), \quad \alpha_k^i = \alpha + s_k^i/(2\gamma) + q_k^i/2, \quad \beta_k^i = \beta - s_k^i/(2\gamma) + q_k^i/2.$$

Hence the joint posterior $\rho^i(\boldsymbol{\nu})$ at round i is

$$\rho(\boldsymbol{\nu}|\mathbf{s}^i, \mathbf{q}^i) = \frac{1}{Z} \prod_{k=1}^{K} \nu_k^{\alpha_k^i - 1} (1 - \nu_k)^{\beta_k^i - 1} \tag{A.1}$$

where Z is a normalizing constant. This allows us to compute the posterior mean and variance of each arm:

$$\bar{\mu}_k(\mathbf{s}^i, \mathbf{q}^i) = \int_{[0,1]^K} \gamma(2\nu_k - 1)\rho(\boldsymbol{\nu}|\mathbf{s}^i, \mathbf{q}^i)d\boldsymbol{\nu} = \gamma\left(2\frac{\alpha_k^i}{\alpha_k^i + \beta_k^i} - 1\right) = \gamma\frac{\alpha_k - \beta_k + s_k^i/\gamma}{\alpha_k + \beta_k + q_k^i},$$

$$\bar{\sigma}_k^2(\mathbf{s}^i, \mathbf{q}^i) = \int_{[0,1]^K} \gamma^2 \rho(\boldsymbol{\nu}|\mathbf{s}^i, \mathbf{q}^i)d\boldsymbol{\nu} = \gamma^2,$$

$$\tag{A.2}$$

and the higher-order moments are

$$\bar{E}_k^p(\mathbf{s}^i, \mathbf{q}^i) = \int \gamma^p \nu_k + (-\gamma)^p(1 - \nu_k)\rho(\boldsymbol{\nu}|\mathbf{s}^i, \mathbf{q}^i)d\boldsymbol{\nu} = \begin{cases} \gamma^p, & p \text{ is even}, \\ \gamma^p \dfrac{\alpha_k - \beta_k + s_k^i/\gamma}{\alpha_k + \beta_k + q_k^i}, & p \text{ is odd}. \end{cases}$$

By the definition of $\hat{\mu}, \hat{\sigma}$ in (10), one has

$$\hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \lim_{n \to \infty} \frac{\frac{\gamma(\alpha_k - \beta_k)}{f(n)} + \hat{s}_k}{\frac{\alpha_k + \beta_k}{n} + \hat{q}_k}$$

$$\hat{\sigma}_k^2(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \lim_{n \to \infty} \frac{n}{f(n)^2}\gamma^2$$

$$\hat{E}_k^p = \lim_{n \to \infty} \frac{n}{f(n)^p} \bar{E}_k(f(n)\mathbf{s}, n\mathbf{q}) = \begin{cases} \lim_{n \to \infty} \left(\dfrac{\gamma}{f(n)}\right)^{p-2} \dfrac{n}{f(n)^2}\gamma^2, & p \text{ is even}, \\ \lim_{n \to \infty} \left(\dfrac{\gamma}{f(n)}\right)^{p-1} \dfrac{\frac{\gamma(\alpha_k - \beta_k)}{f(n)} + \hat{s}_k}{\frac{\alpha_k + \beta_k}{n} + \hat{q}_k}, & p \text{ is odd}. \end{cases}$$

By the assumption given in Lemma 3, one ends up with

$$\hat{\mu}(s, q, \hat{\alpha}_k, \hat{\beta}_k) = \lim_{n \to \infty} \frac{\hat{\alpha}_k + s}{\hat{\beta}_k + q}, \quad \hat{\sigma}(s, q) \equiv \hat{\sigma}, \quad \hat{E}_k^p \equiv 0.$$

The last equation is because $\lim_{n \to \infty} \frac{\sqrt{n}}{f(n)}\gamma = \hat{\sigma}$ implies that $\lim_{n \to \infty} \frac{1}{f(n)}\gamma = 0$.

Different limiting equations due to different scaling factors Let us consider the case where the reward value γ and $-\gamma$ are independent of the horizon n. We set $\gamma = 1$ and the prior hyperparameter to be $(\alpha, \beta) = (\frac{n}{2}, \frac{n}{2})$, i.e., the prior measure of $\nu_k \sim \text{Beta}(\frac{n}{2}, \frac{n}{2})$. In this case, one can rescale the cumulative reward s and c_n by $\sqrt{n}$. By Lemma 3, one has

$$\hat{\mu}(s, q) = \frac{s}{1 + q}; \quad \hat{\sigma}(s, q) \equiv 1.$$

One can also rescale the cumulative reward s and c_n by n^{-1}. By Lemma 3, one has

$$\hat{\mu}(s, q) = \frac{s}{1 + q}; \quad \hat{\sigma}(s, q) \equiv 0.$$

In this case, the Bayesian bandit problem converges to a deterministic control problem in the form of (18) with

$$\hat{\mu}(s, q) = \frac{s}{1 + q}, \quad \hat{\sigma} \equiv 0. \tag{A.3}$$

One can see from the above examples that the same Bayesian bandit problem may converge toward different control problems based on the chosen scaling.

B Proof of Lemma 4

The probability density function for the k-th arm in environment $\boldsymbol{\nu}$ is

$$P_k^\nu = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{|x - \nu_k|^2}{2\sigma^2}}.$$

(Note that here π denotes the constant 3.14159... rather than the policy.) Thus the expected reward for the k-th arm in environment $\boldsymbol{\nu}$ is $\mu_k(\boldsymbol{\nu}) = \nu_k$. At the start of the i-th round, the posterior distribution of ν_k is uniquely determined by the cumulative reward s_k^i and the number of pulls q_k^i through round $i - 1$ according to

$$\nu_k^i \sim \mathcal{N}(\alpha_k^i, (\beta_k^i)^2), \quad \alpha_k^i = \frac{\alpha\beta^{-2} + s_k^i\sigma^{-2}}{q_k^i\sigma^{-2} + \beta^{-2}}, \quad (\beta_k^i)^2 = \frac{1}{q_k^i\sigma^{-2} + \beta^{-2}}.$$

Hence the joint posterior $\rho^i(\boldsymbol{\nu})$ at round i is

$$\rho^i(\boldsymbol{\nu}) = \rho(\boldsymbol{\nu}|\mathbf{s}^i, \mathbf{q}^i) = \prod_{k=1}^K \frac{1}{\sqrt{2\pi}\beta_k^i} e^{-\frac{|\nu_k - \alpha_k^i|^2}{2(\beta_k^i)^2}}.$$

From this, it follows that

$$\bar{\mu}_k(\mathbf{s}^i, \mathbf{q}^i) = \int_{\mathbb{R}^K} \nu_k \rho^i(\boldsymbol{\nu})d\boldsymbol{\nu} = \alpha_k^i = \frac{\alpha\beta^{-2} + s_k^i\sigma^{-2}}{q_k^i\sigma^{-2} + \beta^{-2}},$$

$$\bar{\sigma}_k^2(\mathbf{s}^i, \mathbf{q}^i) = \int_{\mathbb{R}^K} (\sigma^2 + \nu_k^2)\rho^i(\boldsymbol{\nu})d\boldsymbol{\nu} = \sigma^2 + (\beta_k^i)^2 + (\alpha_k^i)^2 = \sigma^2 + \frac{1}{q_k^i\sigma^{-2} + \beta^{-2}} + (\bar{\mu}_k)^2,$$

Inserting $\bar{\mu}_k$ and $\bar{\sigma}_k$ into (10) yields

$$\hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \lim_{n \to \infty} \frac{n}{f(n)} \frac{\alpha \beta^{-2} + f(n)\hat{s}_k \sigma^{-2}}{n\hat{q}_k \sigma^{-2} + \beta^{-2}} = \lim_{n \to \infty} \frac{\hat{s}_k + \frac{\sigma^2 \alpha}{f(n)\beta^2}}{\hat{q}_k + \frac{\sigma^2}{n\beta^2}}.$$

$$(\hat{\sigma}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}))^2 = \lim_{n \to \infty} \frac{n}{f(n)^2} \sigma^2 + \frac{n}{f(n)^2(n\sigma^{-2}\hat{q}_k + \beta^{-2})} + \frac{1}{n}\left(\frac{n}{f(n)}\bar{\mu}_k\right)^2$$

$$= \lim_{n \to \infty} \frac{n}{f(n)^2} \sigma^2 \left(1 + \frac{1}{n} \frac{1}{\hat{q}_k + \frac{\sigma^2}{n\beta^2}}\right).$$

By the condition in Lemma 4, one ends up with

$$\hat{\mu}(s, q, \hat{\alpha}_k, \hat{\beta}_k) = \frac{\hat{\alpha}_k + s}{\hat{\beta}_k^{-2} + q}, \quad \hat{\sigma}(s, q) \equiv \hat{\sigma}.$$

For the higher-order moments, one can write the moments of the normal distribution in the following form,

$$\int x^p P_k^\nu(x) dx = \sum_{j=0}^{\lfloor p/2 \rfloor} C(p, j) \nu_k^{p-2j} \sigma^{2j}$$

where $C(p, j)$ is a constant depends on p, j. Then one has

$$\bar{E}^p(\mathbf{s}^i, \mathbf{q}^i) = \int \int x^p P_k^\nu(x) dx \rho(\nu_k) d\nu_k = \sum_{j=0}^{\lfloor p/2 \rfloor} C(p, j) \sigma^{2j} \int \nu_k^{p-2j} \rho(\nu_k) d\nu_k$$

$$= \sum_{j=0}^{\lfloor p/2 \rfloor} \sum_{l=0}^{\lfloor (p-2j)/2 \rfloor} C(p, j) C(p - 2j, l) \sigma^{2j} (\beta_k^i)^{2l} (\alpha_k^i)^{p-2j-2l}$$

$$= \sum_{j=0}^{\lfloor p/2 \rfloor} \sum_{l=0}^{\lfloor (p-2j)/2 \rfloor} C(p, j) C(p - 2j, l) \sigma^{2j} (\beta_k^i)^{2l} (\bar{\mu}_k)^{p-2j-2l}$$

Therefore, after rescaling, one has

$$\hat{E}_k^p(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \lim_{n \to \infty} \sum_{j=0}^{\lfloor p/2 \rfloor} \sum_{l=0}^{\lfloor (p-2j)/2 \rfloor} C(p, j) C(p - 2j, l) n^{1+j-p} \left(\frac{n}{f(n)^2}\sigma^2\right)^{j+l} \left(\frac{1}{\hat{q}_k + \frac{\sigma^2}{n\beta^2}}\right)^l \left(\frac{n}{f(n)}\bar{\mu}_k\right)^{p-2j-2l}$$

Since

$$n^{1+j-p} \le n^{-1}, \quad \lim_{n \to \infty} \frac{n}{f(n)^2}\sigma^2 = \hat{\sigma}, \quad \lim_{n \to \infty} \frac{1}{\hat{q}_k + \frac{\sigma^2}{n\beta^2}} = \frac{1}{\hat{q}_k + \hat{\beta}^{-2}}, \quad \lim_{n \to \infty} \frac{n}{f(n)}\bar{\mu}_k = \hat{\mu},$$

one has,

$$\hat{E}_k^p(\hat{\mathbf{s}}, \hat{\mathbf{q}}) \equiv 0$$

C Proof of Lemma 5

The expectation of the reward at round i in environment $\boldsymbol{\nu}$ is $\langle A^i, \boldsymbol{\nu}\rangle$, and the probability density function for k-th arm is

$$P_k^{\boldsymbol{\nu}} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{|x-(\mathbf{a}_k)^\top \boldsymbol{\nu}|}{2\sigma^2}}.$$

Then, the posterior distribution of $\boldsymbol{\nu}$ at round i is uniquely determined by the cumulative reward $(\mathbf{s}^i, \mathbf{q}^i)$ up to time $i-1$,

$$\boldsymbol{\nu}^i \sim \mathcal{N}(\boldsymbol{\alpha}^i, \Sigma^i), \quad \Sigma^i = (\Sigma^{-1} + \sigma^{-2}\sum_{k=1}^{K} q_k^i \mathbf{a}_k(\mathbf{a}_k)^\top)^{-1}, \quad \boldsymbol{\alpha}^i = \Sigma^i(\Sigma^{-1}\boldsymbol{\alpha} + \sigma^{-2}\sum_{k=1}^{K} s_k^i \mathbf{a}_k).$$

Note that the $\mathbf{a}_k \in \mathbb{R}^d$ in the above equation represents the action value of the k-th arm. Hence the posterior measure of $\boldsymbol{\nu}^i$ at round i is

$$\rho^i(\boldsymbol{\nu}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma^i|}} \exp\left(-\frac{1}{2}(\boldsymbol{\nu}-\boldsymbol{\alpha}^i)^\top \Sigma^i(\boldsymbol{\nu}-\boldsymbol{\alpha}^i)\right).$$

Therefore, one obtains

$$\begin{aligned}
\bar{\mu}_k(\mathbf{s}^i, \mathbf{q}^i) &= \int_{\mathbb{R}^K} (\mathbf{a}_k)^\top \boldsymbol{\nu}\rho^i(\boldsymbol{\nu})d\boldsymbol{\nu} = (\mathbf{a}_k)^\top \boldsymbol{\alpha}^i \\
&= (\mathbf{a}_k)^\top (\Sigma^{-1} + \sigma^{-2}\sum_{j=1}^{K} q_j^i \mathbf{a}_j(\mathbf{a}_j)^\top)^{-1}(\Sigma^{-1}\boldsymbol{\alpha} + \sigma^{-2}\sum_{j=1}^{K} s_j^i \mathbf{a}_j), \\
\bar{\sigma}_k^2(\mathbf{s}^i, \mathbf{q}^i) &= \int_{\mathbb{R}^K} (\sigma^2 + (\mathbf{a}_k)^\top \boldsymbol{\nu}\boldsymbol{\nu}^\top(\mathbf{a}_k)^\top)\rho^i(\boldsymbol{\nu})d\boldsymbol{\nu} = \sigma^2 + (\mathbf{a}_k)^\top(\Sigma^i + \boldsymbol{\alpha}^i(\boldsymbol{\alpha}^i)^\top)\mathbf{a}_k \\
&= \sigma^2 + (\mathbf{a}_k)^\top (\Sigma^{-1} + \sigma^{-2}\sum_{j=1}^{K} q_j^i \mathbf{a}_j(\mathbf{a}_j)^\top)^{-1}\mathbf{a}_k + (\bar{\mu}_k)^2.
\end{aligned} \tag{C.1}$$

Inserting $\bar{\mu}_k$ and $\bar{\sigma}_k$ into (10) yields

$$\hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \lim_{n\to\infty} (\mathbf{a}_k)^\top \left(\frac{\sigma^2}{n}\Sigma^{-1} + \sum_{j=1}^{K} \hat{q}_j^i \mathbf{a}_j(\mathbf{a}_j)^\top\right)^{-1} \left(\frac{\sigma^2}{f(n)}\Sigma^{-1}\boldsymbol{\alpha} + \sum_{j=1}^{K} \hat{s}_j^i \mathbf{a}_j\right).$$

$$(\hat{\sigma}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}))^2 = \lim_{n\to\infty} \left(\frac{n}{f(n)^2}\sigma^2\right)\left(1 + (\mathbf{a}_k)^\top \left(\frac{\sigma^2}{n}\Sigma^{-1} + \sum_{j=1}^{K} \hat{q}_j^i \mathbf{a}_j(\mathbf{a}_j)^\top\right)^{-1}\mathbf{a}_k\right) + \frac{1}{n}\left(\frac{n}{f(n)}\bar{\mu}_k\right)^2.$$

By the condition in Lemma 4, one ends up with

$$\hat{\mu}(\mathbf{s}, \mathbf{q}, \mathbf{b}) = \mathbf{b}^\top(\hat{\Sigma}^{-1} + \sum_k q_k \mathbf{a}_k(\mathbf{a}_k)^\top)^{-1}(\hat{\boldsymbol{\alpha}} + \sum_{k=1}^{K} s_k \mathbf{a}_k), \quad \hat{\sigma}(\mathbf{s}, \mathbf{q}) \equiv \hat{\sigma}.$$

For the higher-order moments, since

$$\Sigma(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \sigma^2 n^{-1} \left(\frac{\sigma^2}{n} \Sigma^{-1} + \sum_k \hat{q}_k \mathbf{a}_k (\mathbf{a}_k)^\top \right)^{-1}$$

$$\boldsymbol{\alpha}^i(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \frac{f(n)}{n} \left(\Sigma^{-1} + \sum_k \hat{q}_k \mathbf{a}_k (\mathbf{a}_k)^\top \right)^{-1} \left(\frac{\sigma^2 \Sigma^{-1} \boldsymbol{\alpha}}{f(n)} + \sum_k \hat{s}_k \mathbf{a}_k \right),$$

Therefore, similar to the normal bandit problem, the higher-order moments are in the following order:

$$\hat{E}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \lim_{n \to \infty} \frac{n}{f(n)^p} \sum_{j=0}^{\lfloor p/2 \rfloor} \sum_{l=0}^{\lfloor (p-2j)/2 \rfloor} O(\sigma^{2j}) O(\Sigma^l) O(\boldsymbol{\alpha}^{p-2j-2l})$$

$$= \lim_{n \to \infty} \sum_{j=0}^{\lfloor p/2 \rfloor} \sum_{l=0}^{\lfloor (p-2j)/2 \rfloor} O\left(\frac{n}{f(n)^p} \right) O(\sigma^{2j}) O(\sigma^{2l} n^{-l}) O\left(\frac{f(n)^{p-2j-2l}}{n^{p-2j-2l}} \right)$$

$$= \lim_{n \to \infty} \sum_{j=0}^{\lfloor p/2 \rfloor} \sum_{l=0}^{\lfloor (p-2j)/2 \rfloor} O\left(\left(\frac{\sqrt{n}\sigma}{f(n)} \right)^{2j+2l} n^{1+j-p} \right) = 0$$

where the last equality is because $1_j - p \leq -1$ for all $p \geq 3$ and $\lim_{n \to \infty} \frac{\sqrt{n}\sigma}{f(n)} = \hat{\sigma}$.

D Proof of Theorem 6

Look at the optimal control problem (21) with $\hat{\sigma} \equiv 0$, when $\hat{\mu}_k = \frac{\hat{s}_k + \alpha_k}{\hat{q}_k + \beta_k}$, note that

$$\frac{d}{d\tau} \hat{\mu}_k = \frac{1}{\hat{q}_k + \beta_k} \frac{d}{d\tau} \hat{s}_k - \frac{\hat{s}_k + \alpha_k}{(\hat{q}_k + \beta_k)^2} \frac{d}{d\tau} \hat{q}_k = 0 \quad \text{for} \quad \forall \hat{\pi}(\tau).$$

This implies that $\hat{\mu}_k(\tau) \equiv \hat{\mu}_k(t)$ for $\forall k$. Therefore, the objective function becomes

$$\int_t^1 (\hat{\boldsymbol{\mu}}(t) - \lambda \boldsymbol{\pi}(\tau)) \cdot \boldsymbol{\pi}(\tau) d\tau \tag{D.1}$$

Since $\hat{\boldsymbol{\mu}}(t)$ is a constant, the above objective function will be maximized at π^* given in (25) and (26) for the unregularized version, i.e., $\lambda = 0$ and the regularized version, i.e., $\lambda > 0$.

For the stochastic case, Note that

$$d\mathbb{E}[\hat{\mu}_k] = \mathbb{E}\left[\frac{1}{\hat{q}_k + \beta_k} d\hat{s}_k - \frac{\hat{s}_k + \alpha_k}{(\hat{q}_k + \beta_k)^2} d\hat{q}_k \right] = 0 + \mathbb{E}[\hat{\sigma}(\hat{\mathbf{s}}, \hat{\mathbf{q}})\sqrt{\pi(\hat{\mathbf{s}}, \hat{\mathbf{q}})} dB_t] = 0 \quad \text{for} \quad \forall \hat{\pi}(\tau).$$
$$\tag{D.2}$$

so the objective function for the stochastic case is the same as (D.1), which results in the same optimal policy.

E Proof of Lemma 7

Consider k-armed Bernoulli bandit, where the k-th arm gives reward 1 with probability ν_k and 0 with probability $1 - \nu_k$. The initial prior measure of ν_k follows $\text{Beta}(\alpha_k, \beta_l)$. Then

the rescaled optimal cumulative reward

$$v^i(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \frac{1}{n} w^i(\mathbf{s}, \mathbf{q})$$

with scaling factor $f(n) = n$ satisfies

$$v^i(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \max_k \left\{ \frac{1}{n} \tilde{p}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) + \tilde{p}_k(\hat{s}, \hat{q}) v^{i+1}(\hat{\mathbf{s}} + \frac{1}{n}\mathbf{e}_k, \hat{q} + \frac{1}{n}\mathbf{e}_k) + (1 - \tilde{p}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}})) v^{i+1}(\hat{\mathbf{s}}, \hat{\mathbf{q}} + \frac{1}{n}\mathbf{e}_k) \right\}$$

$$(E.1)$$

where

$$\tilde{p}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \frac{n^{-1}\alpha_k + s_k}{n^{-1}(\alpha_k + \beta_k) + q_k}.$$

The corresponding HJB equation under the scaling factor $f(n) = n$ is

$$\partial_t v + \max_{\hat{\boldsymbol{\pi}}(t, \hat{s}, \hat{q}) \in \Delta^K} \left(\hat{\mu}_k + \hat{\mu}_k \partial_{s_k} v + \partial_{q_k} v \right) \pi_k = 0, \qquad v(1, \hat{s}, \hat{q}) = 0.$$

where

$$\hat{\mu}_k = \frac{\hat{\alpha}_k + s_k}{\hat{\beta}_k + q_k}, \quad \text{with} \quad \hat{\alpha}_k = \lim_{n \to \infty} n^{-1} \alpha_k, \quad \hat{\beta}_k = \lim_{n \to \infty} n^{-1}(\alpha_k + \beta_k).$$

Since $\hat{\mu} \geq 0$, applying the numerical scheme (34)-(35) with $\delta_t = \delta_q = \delta_s = 1/n$ gives ,

$$\tilde{v}(\frac{l}{n}, \hat{\mathbf{s}}, \hat{\mathbf{q}}) = \max_k \left\{ \tilde{v}(\frac{l+1}{n}, \hat{\mathbf{s}}, \hat{\mathbf{q}}) + \frac{1}{n} \left[\frac{\hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}})}{n^{-1}} \left(\tilde{v}(\frac{l+1}{n}, \hat{\mathbf{s}} + \frac{1}{n}\mathbf{e}_k, \hat{q} + \frac{1}{n}\mathbf{e}_k) \right. \right. \right.$$
$$\left. \left. - \tilde{v}(\frac{l+1}{n}, \hat{\mathbf{s}}, \hat{\mathbf{q}} + \frac{1}{n}\mathbf{e}_k) \right) + \frac{1}{n^{-1}} \left(\tilde{v}(\frac{l+1}{n}, \hat{\mathbf{s}}, \hat{\mathbf{q}} + \frac{1}{n}\mathbf{e}_k) - \tilde{v}(\frac{l+1}{n}, \hat{\mathbf{s}}, \hat{\mathbf{q}}) \right) \right.$$
$$\left. + \hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}})] \right\}$$

$$(E.2)$$

By comparing (E.1) and (E.2), one can see if and only if

$$\hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \tilde{p}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}),$$

The numerical scheme is equivalent to the exact Bayes-optimal algorithm. The above condition holds if and only if

$$(\alpha_k, \beta_k) = (c_1 n, c_2 n)$$

for any constants (c_1, c_2), which completes the proof for the first part of the Lemma 7.

Consider the binomial bandits described in Section 3.4.1. First, the optimal cumulative reward satisfies,

$$w^i(\mathbf{s}, \mathbf{q}) = \max_k \left\{ \gamma(2p_k(\mathbf{s}, \mathbf{q}) - 1) + p_k(\mathbf{s}, \mathbf{q}) w^{i+1}(\mathbf{s} + \gamma\mathbf{e}_k, \mathbf{q} + \mathbf{e}_k) \right.$$
$$\left. + (1 - p_k(\mathbf{s}, \mathbf{q})) w^{i+1}(\mathbf{s} - \gamma\mathbf{e}_k, \mathbf{q} + \mathbf{e}_k) \right\}$$

with $p_k(\mathbf{s}, \mathbf{q}) = \frac{\alpha_k + s_k/(2\gamma) + q_k/2}{\alpha_k + \beta_k + q_k}$. Then the rescaled optimal cumulative reward

$$v^i(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \frac{1}{\sqrt{n}} w^i(\mathbf{s}, \mathbf{q})$$

with scaling factor $f(n) = \sqrt{n}$ satisfies

$$v^i(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \max_k \left\{ n^{-1/2}\gamma(2\hat{p}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) - 1) + \hat{p}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}})v^{i+1}(\hat{\mathbf{s}} + \gamma n^{-1/2}\mathbf{e}_k, \hat{\mathbf{q}} + n^{-1}\mathbf{e}_k) \right.$$
$$\left. + (1 - \hat{p}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}))v^{i+1}(\hat{\mathbf{s}} - \gamma n^{-1/2}\mathbf{e}_k, \hat{\mathbf{q}} + n^{-1}\mathbf{e}_k) \right\},$$

where

$$2\hat{p}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) - 1 = \frac{1}{\sqrt{n}\gamma} \frac{\frac{\gamma(\alpha_k - \beta_k)}{\sqrt{n}} + \hat{s}_k}{\frac{\alpha_k + \beta_k}{n} + \hat{q}_k}$$

By letting $\tilde{\mu}(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \sqrt{n}\gamma(2\hat{p}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) - 1)$, then

$$v^i(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \max_k \left\{ n^{-1}\tilde{\mu}(\hat{\mathbf{s}}, \hat{\mathbf{q}}) + \frac{1}{2}\left(\frac{\tilde{\mu}(\hat{\mathbf{s}}, \hat{\mathbf{q}})}{\sqrt{n}\gamma} + 1\right)v^{i+1}(\hat{\mathbf{s}} + \gamma n^{-1/2}\mathbf{e}_k, \hat{\mathbf{q}} + n^{-1}\mathbf{e}_k) \right.$$
$$\left. + \frac{1}{2}\left(1 - \frac{\tilde{\mu}(\hat{\mathbf{s}}, \hat{\mathbf{q}})}{\sqrt{n}\gamma}\right)v^{i+1}(\hat{\mathbf{s}} - \gamma n^{-1/2}\mathbf{e}_k, \hat{\mathbf{q}} + n^{-1}\mathbf{e}_k) \right\}, \tag{E.3}$$

On the other hand, the limiting HJB equation for $v^i(\hat{\mathbf{s}}, \hat{\mathbf{q}})$ is

$$\partial_t v + \max_{\hat{\boldsymbol{\pi}}(t, \hat{\mathbf{s}}, \hat{\mathbf{q}}) \in \Delta^K} \left(\hat{\mu}_k + \hat{\mu}_k \partial_{s_k} v + \partial_{q_k} v + \frac{1}{2}\hat{\sigma}^2 \partial_{s_k}^2 v \right) \pi_k = 0, \qquad v(1, \hat{s}, \hat{q}) = 0,$$

where

$$\hat{\mu}(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \frac{\hat{\alpha}_k + \hat{s}_k}{\hat{\beta}_k + \hat{q}_k}, \quad \hat{\sigma} = \gamma, \quad \text{with} \quad \hat{\alpha}_k = \lim_{n \to \infty} \frac{\gamma(\alpha_k - \beta_k)}{\sqrt{n}}, \quad \hat{\beta}_k = \lim_{n \to \infty} \frac{\alpha_k + \beta_k}{n}.$$

By comparing (E.1) and (E.2), one can see if and only if

$$\hat{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}) = \tilde{\mu}_k(\hat{\mathbf{s}}, \hat{\mathbf{q}}),$$

The numerical scheme is equivalent to the exact Bayes-optimal algorithm. The above condition holds if and only if

$$(\alpha_k, \beta_k) = (c_1 n + c_2\sqrt{n}, c_1 - c_2\sqrt{n})$$

for any constants (c_1, c_2), which completes the proof for the second part of the Lemma 7.

F Detailed algorithms

Algorithm 2 Thompson Sampling for unstructured bandits

Require: Input: $n, (\mathbf{s}, \mathbf{q}) = 0, \rho(\boldsymbol{\nu}|\beta)$
   for $i = 1, \ldots, K$ **do**
      $A_i \leftarrow i, \, s_i \leftarrow X_i, \, q_i \leftarrow 1$
   end for
   for $i = K+1, \ldots, n$ **do**
      Update $\rho(\boldsymbol{\nu}|\mathbf{s}, \mathbf{q})$ according to Bayesian rule
      Sample $\boldsymbol{\nu}^i$ according to the probability distribution $\rho(\boldsymbol{\nu}|\mathbf{s}, \mathbf{q})$
      $A_i \leftarrow \operatorname{argmax}_k \mu_k(\boldsymbol{\nu}^i), \, s_k \leftarrow s_k + X_i, \, q_k \leftarrow q_k + 1$
   end for

Algorithm 3 UCB for unstructured bandits

Require: Input: $n, (\mathbf{s}, \mathbf{q}) = 0, \delta$
   for $i = 1, \ldots, K$ **do**
      $A_i \leftarrow i, \, s_i \leftarrow X_i, \, q_i \leftarrow 1$
   end for
   for $i = K+1, \ldots, n$ **do**
      Update $\hat{\mu}_k \leftarrow \frac{s_k}{q_k} + \sqrt{\frac{2\log(\delta)}{q_k}}$
      $A_i \leftarrow \operatorname{argmax}_k \hat{\mu}_k, \, s_k \leftarrow s_k + X_i, \, q_k \leftarrow q_k + 1$
   end for

Algorithm 4 Thompson Sampling for linear bandits

Require: Input: $n, \mathbf{a}_k, (\mathbf{s}, \mathbf{q}) = 0, \rho(\boldsymbol{\nu}) \sim \mathcal{N}(\mu, \Sigma)$
   for $i = 1, \ldots, K$ **do**
      $A_i \leftarrow i, \, s_i \leftarrow X_i, \, q_i \leftarrow 1$
   end for
   for $i = K+1, \ldots, n$ **do**
      Sample $\boldsymbol{\nu}$ according to the probability distribution $\rho(\boldsymbol{\nu})$
      $A^i = \operatorname{argmax}_k \mathbf{a}_k^\top \boldsymbol{\nu}$
      $x = \mathbf{a}_{A^i}, \, y = \boldsymbol{\nu} * x + \eta, \text{ where } \eta \sim \mathcal{N}(0, \sigma^2)$
      $\mu = (\Sigma^{-1} + \sigma^{-2}xx^\top)^{-1}(\Sigma^{-1}\mu + \sigma^{-2}yx), \, \Sigma = (\Sigma^{-1} + \sigma^{-2}xx^\top)^{-1}$
   end for

Algorithm 5 UCB for linear bandits

Require: Input: $V = v_0, W = 0, \hat{\boldsymbol{\nu}} = 0, \lambda = 0.1$
   for $i = 1, \ldots, K$ **do**
      $A_i \leftarrow i, \, s_i \leftarrow X_i, \, q_i \leftarrow 1$
   end for
   for $i = K+1, \ldots, n$ **do**
      $\beta \leftarrow \sqrt{\lambda} + \sqrt{2\log(n^2) + \log(1 + (i-1)/\lambda)}$
      $A^i \leftarrow \operatorname{argmax}_k \mathbf{a}_k \hat{\boldsymbol{\nu}} + \beta * \sqrt{\mathbf{a}_k^2/V}$
      $x \leftarrow \mathbf{a}_{A^i}, \, y \leftarrow \boldsymbol{\nu} * x + \eta, \text{ where } \eta \sim \mathcal{N}(0, \sigma^2)$
      $V \leftarrow V + x^2, \, W \leftarrow W + xy, \, \hat{\boldsymbol{\nu}} \leftarrow \frac{W}{V}$
   end for