Asymptotic Randomised Control with applications to bandits

Samuel N. Cohen* Tanut Treetanthiploet†

May 8, 2026

Abstract

We consider a general multi-armed bandit problem with correlated (and simple contextual and restless) elements, as a relaxed control problem. By introducing an entropy premium, we obtain a smooth asymptotic approximation to the value function. This yields a novel semi-index approximation of the optimal decision process. This semi-index can be interpreted as explicitly balancing an exploration–exploitation trade-off as in the UCB (Upper Confidence Bound) principle where the learning premium explicitly describes asymmetry of information available in the environment and non-linearity in the reward function.

Performance of the resulting Asymptotic Randomised Control (ARC) algorithm compares favourably well with other approaches to correlated multi-armed bandits.

Keywords: multi-armed bandit, stochastic control, asymptotic approximation
MSC2020: 60J20, 93E35, 90C40, 41A58

1 Introduction

In many situations, one needs to decide between acting to reveal data about a system and acting to generate profit; this is the trade-off between exploration and exploitation. A simple situation where we face this trade-off is a multi-armed bandit problem, where one has K 'bandits' or equivalently, a bandit with K arms, and one must choose which bandit to play at each time. Playing a bandit results in a reward generated from a fixed unknown distribution which must be inferred 'on-the-fly'. At each time, we need to decide whether to focus on playing the bandit which gives the best reward or on gaining information to exploit in the future.

The first theoretical results for the multi-armed bandit problem were proved by Gittins and Jones [14]. They formulated the bandit as an optimisation problem for a Markov Decision Process over an infinite time horizon with an underlying state corresponding to the posterior distribution. An optimal solution is described in terms of an index strategy, where we always play an arm with the maximum index, and the index of each arm can be computed by solving an optimal stopping problem. The crucial assumption guaranteeing the optimality of this index strategy is the independence between arms, which may not hold in more general settings.

More recently, multi-armed bandit problems are often formulated as a statistical problem, rather than an optimisation problem (see e.g. [4, 33, 19, 22, 11]). Novel algorithms for bandit problems are often proposed using heuristic justifications and are then shown to give theoretical guarantees in terms of a regret bound (either from a Bayesian or a frequentist perspective) in symmetric settings which typically assume that rewards are the only observation and are linear in the unknown parameter.

*Mathematical Institute, University of Oxford, UK, samuel.cohen@maths.ox.ac.uk
†ttreetanthiploet@gmail.com

In this paper, we aim to address three fundamental questions for learning. (i) How can we quantify the amount of information that each arm provides, especially when information from each arm is not symmetric? This distinguishes between uncertainty and the value of learning. (ii) What is the connection between information and reward? In particular, when rewards do not depend linearly on the parameters of our model, how will this affect learning? (iii) Can we use this analysis to construct a reasonable learning algorithm which can be applied to a wide class of learning problems?

To achieve our goal, we formulate the multi-armed bandit problem as an optimisation problem (in particular, a stochastic control problem) over an infinite time horizon using a Bayesian formulation, as in [14]. We extend the entropy regularised control approach of Reisinger and Zhang [20] and Wang, Zariphopoulou and Zhou [32] to the discrete time setting to obtain an asymptotic approximation to the value function. This approximation results in a randomised index strategy. This index strategy enjoys a natural interpretation as a sum of the instantaneous reward (exploiting) and the benefit of learning (exploring). Since our solution is obtained by approximating the value function using a randomised strategy, we address various limitations of other algorithms found in the literature.

The main contribution of this work is to develop conceptual insight into a general class of learning (bandit) problems from first principles. We aim to understand how to make decisions when information from each option available to us will result in us learning about multiple future options, in an asymmetric manner, and also to understand the connection between reward and observation. The resulting algorithm generated from our approximation performs well numerically, however we do not aim to provide a formal regret analysis.

The paper proceeds as follows. In Section 2, we describe how to formulate various classes of bandit problems in terms of a discrete-time diffusion process. By considering the behaviour of the diffusion dynamic in a regime with small uncertainty, we propose the Asymptotic Randomised Control (ARC) algorithm together with its heuristic derivation and summary of our main results. In Section 3, we give an overview of well-known approaches for bandit problems and discuss some limitations of these algorithms, and how the ARC algorithm addresses these limitations. We also show that the derived ARC algorithm results in a decision scheme connecting the Upper Confidence Bound (UCB) [1, 2, 15], the Knowledge Gradient (KG) [27] and Boltzmann Exploration (BE) [28, 5] principles through the discrete time version of Itô's lemma. In Section 4, we provide a formal derivation of the ARC approach; further technical results are given in the Appendix. Finally, in Section 5, we run a numerical experiment to show that when the uncertainty is small, ARC algorithm is numerically accurate. We also show that our algorithm performs well compared to other approaches.

Notation: We summarise here the notation that will be used in our discussion.

- For any function $h : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^n$, we write $D^k h$ (and $\partial_m^k \partial_d^l h$) for a tensor of degree $(1, k)$ (and $(1, k, l)$) with dimension $n(p+q)^k$ (and $np^k q^m$) corresponding to the derivative (and partial derivative) of h.
- We write $\langle \cdot ; \cdot \rangle$ for Euclidean tensor products. In particular, $\langle P; Q \rangle$ is a standard inner product when P, Q are vectors, $\langle P; Q \rangle = \mathrm{Tr}(PQ)$ when P and Q are square matrices.
- For any (Borel) random variables X, Y and a σ-algebra $\mathcal{G}$, we write $\mathcal{L}(Y|\mathcal{G})$ and $\mathcal{L}(Y|X, \mathcal{G})$ for the conditional law of Y given $\mathcal{G}$ and $\sigma(X) \vee \mathcal{G}$, respectively.
- We write e_i for the ith standard basis vector with appropriate dimension, $\mathbb{1}_n = (1, 1, ..., 1) \in \mathbb{R}^n$ and I_n for the identity matrix with dimension n.

- Throughout the proof, we shall introduce a generic constant $C \geq 0$ to quantify an upper bound. This constant does not depend on variables (m, d, λ, t, T) introduced throughout the paper.

2 Asymptotic Randomised Control (ARC) approach

To introduce the Asymptotic Randomised Control (ARC) approach, we first give a heuristic description of the propagation of the posterior distribution in various situations. We then describe the intuition behind the ARC algorithm, and a sketch of our main result. We defer a rigorous mathematical justification to Section 4.

Suppose that our bandit has K arms. At each time the controller will select one of these arms, and depending on this choice will receive an observation and a reward. We model this using an underlying (unknown) parameter $\boldsymbol{\theta} \sim \pi$, which we treat in a Bayesian way with prior π. The parameter $\boldsymbol{\theta}$ describes the distribution of our bandit, i.e., when the ith arm is chosen at time t, we observe a random variable $Y_t^{(i)} \sim \pi_i(\cdot|\boldsymbol{\theta})$ and obtain a reward $r_i(Y_t^{(i)})$. We allow that our observation can be more informative than the reward.

The objective of our problem is to find a sequence of decisions (A_t) taking values in $\{1, 2, ..., K\}$ to maximise

$$\mathbb{E}_\pi\Big[\sum_{t=1}^{\infty} \beta^{t-1} r_{A_t}\big(Y_t^{(A_t)}\big)\Big]. \tag{2.1}$$

The complexity of this problem arises due to the fact that A_t can depend on the previous observations $\{Y_s^{A_s}\}_{s\leq t}$, which depend on the actions previously taken.

In the following section, we will consider different bandit structures and describe (2.1) in terms of a Markov Decision Process (MDP) with infinite state space.

2.1 Bandit structures

2.1.1 A Gaussian-correlated structure

We begin with a classical example, and suppose the prior π for $\boldsymbol{\theta}$ is a multivariate normal $N(m, d)$ with dimension p, and the observation from the ith bandit at time t is given by a q-dimensional random vector $Y_t^{(i)}$, with distribution $\mathcal{L}(Y_t^{(i)}|\boldsymbol{\theta}) = N\big(c_i^\top \boldsymbol{\theta}, P_i^{-1}\big)$ for some $c_i \in \mathbb{R}^{p\times q}$ and P_i a $q \times q$ positive semi-definite matrix.

For simplicity, we will demonstrate the calculation when P_i is invertible. The positive semidefinite case can be studied either by simply taking the limit of the derived result or by using appropriate pseudo-inverses.

Let $\mathcal{G}_t^U$ denote the filtration describing our observation up to time t. Suppose that our posterior at time t is given by $\mathcal{L}(\boldsymbol{\theta}|\mathcal{G}_t^U) = N\big(M_t, D_t\big)$. By standard Bayesian inference, we can show that

$$\mathcal{L}(Y_{t+1}^{(i)}|\mathcal{G}_t^U) = N\big(c_i^\top M_t, c_i^\top D_t c_i + P_i^{-1}\big) \quad \text{and} \quad \mathcal{L}(\boldsymbol{\theta}|\mathcal{G}_{t+1}^U) = \mathcal{L}(\boldsymbol{\theta}|Y_{t+1}^{(i)}, \mathcal{G}_t^U) = N\big(M_{t+1}, D_{t+1}\big)$$

where $M_{t+1} = \big(D_t^{-1} + c_i P_i c_i^\top\big)^{-1}\big(D_t^{-1} M_t + c_i P_i Y_{t+1}^{(i)}\big)$ and $D_{t+1} = \big(D_t^{-1} + c_i P_i c_i^\top\big)^{-1}$.

Recall the Woodbury matrix identity that, for any invertible matrices $A \in \mathbb{R}^{p\times p}$ and $P \in \mathbb{R}^{q\times q}$, and $x \in \mathbb{R}^{p\times q}$, then $(A^{-1} + xPx^T)^{-1} = A - Ax(P^{-1} + x^\top Ax)^{-1}x^\top A$. Applying this identity to $(D_t^{-1} + c_i P_i c_i^\top)^{-1}$ and representing $\mathcal{L}(Y_{t+1}^{(i)}|\mathcal{G}_t^U) = N\big(c_i^\top M_t, c_i^\top D_t c_i + P_i^{-1}\big)$ in term of a standard

Gaussian, we see that

$$\begin{aligned} M_{t+1} - M_t &= \Big(D_t - D_t c_i (P_i^{-1} + c_i^\top D_t c_i)^{-1} c_i^\top D_t\Big) c_i P_i \Big(c_i^\top D_t c_i + P_i^{-1}\Big)^{1/2} Z_{t+1}^{(i)} \\ D_{t+1} - D_t &= -D_t c_i (P_i^{-1} + c_i^\top D_t c_i)^{-1} c_i^\top D_t. \end{aligned} \tag{2.2}$$

where $Z_{t+1}^{(i)} = (c_i^\top D_t c_i + P_i^{-1})^{-1/2}(Y_{t+1}^{(i)} - c_i^\top M_t) \sim N(0, I_q)$.

By taking the conditional expectation $\mathcal{G}_t^U$ together with the tower property and the dominated convergence theorem, we can rewrite (2.1) as an objective function for a classical stochastic control problem (MDP) with underlying state (M_t, D_t);

$$\mathbb{E}_{m,d}\Big[\sum_{t=0}^{\infty} \beta^t f_{A_{t+1}}(M_t, D_t)\Big] \quad \text{with} \quad f_i(m,d) = \int_{\mathbb{R}^q} r_i\big(c_i^\top m + (b_i^\top d c_i + P_i^{-1})^{1/2} z\big)\varphi_q(z)dz \tag{2.3}$$

where φ_q is the density of $N(0, I_q)$. The underlying dynamic of this control can be described as a discrete version of a diffusion process and the prior is encoded as an initial state.

It is worth pointing out that this set-up covers various classes of multi-armed bandit problem which can be found in the literature.

Example 2.1 (Linear bandit)**.** The case when $q = 1$ and $r_i(y) = y$ corresponds to the linear bandit considered in Russo and Roy [24, 25]. The case when r_i is a non-linear function is considered in Filippi et al. [11] with a different distribution assumption.

Example 2.2 (Classical bandit)**.** The case when $q = 1$, $K = p$, $c_i = e_i$ (the ith basis vector) and $r_i(y) = y$ corresponds to the classical stochastic bandit which often finds in most of the literature (see e.g. [19, 4]). However, they often relax the Gaussian assumption to sub-Gaussian and often restrict $P_i = 1$.

Example 2.3 (Structural bandit)**.** The case when $p = q = 1$, $c_i = 1$ and $r_i(y) = u_i y + v_i$ corresponds to the structural bandit studied in Rusmevichientong et al. [22].

Example 2.4 (Classical bandit with additional information)**.** Let consider when $q = p$, $K = p$, $c_i = I_q$, $r_i(y) = \tilde{r}_i(y_i)$ and P_i is diagonal. In this case, we can see that the reward of the ith arm depends only on the parameter $\boldsymbol{\theta}_i$ (as $r_i(y)$ only depends on the ith component of y). However, when the ith arm is chosen, we may also observe some information depending on $\boldsymbol{\theta}_j$, provided that the (j,j)th entry of P_i is positive. This can lead to asymmetry in the information flows, as there may be arms which are informative, while not having a particularly beneficial reward. We will modify this example to demonstrate the idea of learning under asymmetry in Section 3.3.

Example 2.5 (Semi-bandit feedback)**.** The case with P_i a diagonal matrix and $r_i(y) = \frac{1}{|\mathcal{A}_i|}\sum_{j\in\mathcal{A}_i} y_j$ where $\mathcal{A}_i := \{j : (P_i)_{jj} > 0\}$ corresponds to the semi-bandit feedback considered in Russo and Roy [24, 25].

2.1.2 A general-distribution independent structure

We now consider an alternative setting, where each arm is associated to different (independent) parameters.

Suppose that we can write $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_q)$, where $(\boldsymbol{\theta}_j)_{j=1}^q$ are independent under a prior of interest $\bigotimes_{j=1}^q \pi_j$. Suppose that when the ith arm is chosen at time t, we observe $Y_t^{(i)} := \big(Y_t^{(i,1)}, ..., Y_t^{(i,q)}\big) \sim \bigotimes_{j=1}^q \pi_i(\cdot|\boldsymbol{\theta}_j)$. Due to the product structure of the prior and observations, the

posterior of $\boldsymbol{\theta}$ also maintains independence of its components. Therefore, we can evaluate the posterior of each $\boldsymbol{\theta}_j$ separately.

For simplicity of discussion, we consider only the case when $\boldsymbol{\theta}_j$ is one-dimensional. This can be easily extended as long as π_j and $\pi_i(\cdot|\boldsymbol{\theta}_j)$ are a conjugate pair for all i and j. We implicitly allow $\pi_i(\cdot|\boldsymbol{\theta}_j)$ to be degenerate, corresponding to the case when no information regarding $\boldsymbol{\theta}_j$ is revealed when the ith arm is chosen.

In the following examples, we formulate the multi-armed bandit problem in terms of a stochastic control problem (equivalently, MDP) with an underlying discrete diffusion dynamic, as in (2.2). The process M is interpreted as an estimator of $\boldsymbol{\theta}$, while D is the inverse precision of the estimate M. Here, we will only focus on deriving the corresponding dynamics of the underlying state (M, D). The objective function can then be written in the same manner as in (2.3).

Example 2.6 (Binomial bandit)**.** Suppose that the prior π_j of $\boldsymbol{\theta}_j$ is $\text{Beta}(\alpha_j, \beta_j)$ and $\pi_i(\cdot|\boldsymbol{\theta}_j) = \text{Binomial}(n_i, \boldsymbol{\theta}_j)$. We again denote by $\mathcal{G}_t^U$ our observations up to time t and assume that $\mathcal{L}(\boldsymbol{\theta}_j|\mathcal{G}_t^U) = \text{Beta}\big(M_t^j/D_{j,t}, (1-M_t^j)/D_{j,t}\big)$. When the ith arm is chosen at time $t+1$, one can check that the posterior distribution becomes $\mathcal{L}(\boldsymbol{\theta}_j|\mathcal{G}_{t+1}^U) = \text{Beta}\big(M_{j,t+1}/D_{j,t+1}, (1-M_{j,t+1})/D_{j,t+1}\big)$ where

$$M_{j,t+1} - M_{j,t} = \left(\frac{D_{j,t}}{1+n_i D_{j,t}}\right)(Y_{t+1}^{(i,j)} - n_i M_{j,t}) \qquad \text{and} \qquad D_{j,t+1} - D_{j,t} = -\frac{n_i (D_{j,t})^2}{1+n_i D_{j,t}}.$$

We can show that $\mathbb{E}[M_{j,t+1} - M_{j,t}|\mathcal{G}_t^U] = 0$ and

$$\text{Var}[M_{j,t+1} - M_{j,t}|\mathcal{G}_t^U] = \left(\frac{D_{j,t}}{1+n_i D_{j,t}}\right)^2 \mathbb{E}\big[n_i \boldsymbol{\theta}_j(1-\boldsymbol{\theta}_j)\big|\mathcal{G}_t^U\big] = \left(\frac{D_{j,t}}{1+n_i D_{j,t}}\right)^2 \left(\frac{n_i M_{j,t}(1-M_{j,t})}{1+D_{j,t}}\right).$$

Therefore, we can represent the dynamics of the posterior parameter of $\boldsymbol{\theta}_j$, when the ith arm is chosen by

$$M_{j,t+1} - M_{j,t} = \left(\frac{D_{j,t}}{1+n_i D_{j,t}}\right)\sqrt{\left(\frac{n_i M_{j,t}(1-M_{j,t})}{1+D_{j,t}}\right)} Z_{t+1}^{(i,j)} \quad \text{and} \quad D_{j,t+1} - D_{j,t} = -\frac{n_i D_{j,t}^2}{1+n_i D_{j,t}}, \tag{2.4}$$

where $Z_{t+1}^{(i,j)} := (Y_{t+1}^{(i,j)} - n_i M_{j,t})/\sqrt{\left(\frac{n_i M_{j,t}(1-M_{j,t})}{1+D_{j,t}}\right)}$ satisfies $\mathbb{E}Z_{t+1}^{(i,j)} = 0, \text{Var}[Z_{t+1}^{(i,j)}|M_t, D_t] = 1$.

Example 2.7 (Poisson bandit)**.** Suppose that the prior π_j of $\boldsymbol{\theta}_j$ is $\text{Gamma}(\alpha_j, \beta_j)$ and $\pi_i(\cdot|\boldsymbol{\theta}_j) = \text{Poisson}(n_i\boldsymbol{\theta}_j)$. Assume that $\mathcal{L}(\boldsymbol{\theta}_j|\mathcal{G}_t^U) = \text{Gamma}\big(M_{j,t}/D_{j,t}, 1/D_{j,t}\big)$. When the ith arm is chosen at time $t+1$, the posterior distribution becomes $\mathcal{L}(\boldsymbol{\theta}_j|\mathcal{G}_{t+1}^U) = \text{Gamma}\big(M_{j,t+1}/D_{j,t+1}, 1/D_{j,t+1}\big)$, where

$$M_{j,t+1} - M_{j,t} = \left(\frac{D_{j,t}}{1+n_i D_{j,t}}\right)(Y_{t+1}^{(i,j)} - n_i M_{j,t}) \qquad \text{and} \qquad D_{j,t+1} - D_{j,t} = -\frac{n_i (D_{j,t})^2}{1+n_i D_{j,t}}.$$

Similarly to above, we can represent the dynamics of the posterior parameter of $\boldsymbol{\theta}_j$ when the ith arm is chosen by

$$M_{j,t+1} - M_{j,t} = \left(\frac{D_{j,t}}{1+n_i D_{j,t}}\right)\sqrt{n_i M_{j,t}} Z_{t+1}^{(i,j)} \quad \text{and} \quad D_{j,t+1} - D_{j,t} = -\frac{n_i D_{j,t}^2}{1+n_i D_{j,t}}. \tag{2.5}$$

where $Z_{t+1}^{(i,j)} := (Y_{t+1}^{(i,j)} - M_{j,t})/\sqrt{n_i M_{j,t}}$ satisfies $\mathbb{E}Z_{t+1}^{(i,j)} = 0, \text{Var}[Z_{t+1}^{(i,j)}|M_t, D_t] = 1$.

2.2 From multi-armed bandits to diffusive control problems

Inspired by (2.2), (2.4) and (2.5), we now give a formal set-up of a 'discrete-time diffusion control problem' which can be used to study the multi-armed bandit problem.

Let $(\Omega, \mathbb{P}, \mathcal{F})$ be a probability space equipped with two independent sequences of IID $U[0,1]$ random variables $(\boldsymbol{\xi}_t)$ and (ζ_t). We define the filtration $\mathcal{F}_t := \sigma(\boldsymbol{\xi}_s, \zeta_s : s \leq t)$. Here, (ζ_t) represents a random seed used to select a random decision in each time step, whereas $(\boldsymbol{\xi}_t)$ represents the randomness of the outcome.

A *random action* is a stochastic process $(A_t)_{t\in\mathbb{N}}$ taking values in a finite set $\mathcal{A} := \{1, 2, ..., K\}$ where, for each t, A_t is measurable with respect to $\mathcal{F}_{t-1} \vee \sigma(\zeta_t)$.

In our problem, the agent does not choose $(A_t)_{t\in\mathbb{N}}$ directly but instead chooses a relaxed control, that is, a family of conditional laws $(\Pi_t(\cdot))_{t\in\mathbb{N}}$ of A_t where $\Pi_t(\cdot) := \mathbb{P}(A_t \in \cdot|\mathcal{F}_{t-1})$. For convenience, we represent a relaxed policy Π, at each time t, by a probability vector U_t, that is, an $\mathcal{F}_{t-1}$-measurable random variable taking values in the K-dimensional simplex, $\Delta^K := \Big\{u \in [0,1]^K : \sum_{i=1}^K u_i = 1\Big\}$. The connection between Π_t and U_t can be given by $U_t = \big(\Pi_t(\{1\}), ..., \Pi_t(\{K\})\big)$. We will denote the space of the controls $(U_t)_{t\in\mathbb{N}}$ by $\mathcal{U}$. We prescribe the (random) action A_t in terms of U_t and ζ_t by $A_t = A(U_t, \zeta_t) := \sup\Big\{i : \sum_{k=1}^i U_{k,t} \geq \zeta_t\Big\}$.

The state of our system is represented by a Markov process (M, D) taking values in $\Theta \times \mathcal{D} \subseteq \mathbb{R}^p \times \mathbb{R}^q$. Suppose that, when the control $i \in \mathcal{A}$ is chosen, the underlying state evolves according to the transition map

$$\Phi(m, d, i, \xi) := \begin{pmatrix} m \\ d \end{pmatrix} + \begin{pmatrix} \mu_i(m, d) \\ b_i(m, d) \end{pmatrix} + \begin{pmatrix} \sigma_i(m, d)z(m, d, i, \xi) \\ 0 \end{pmatrix} \tag{2.6}$$

where $\mu_i : \Theta \times \mathcal{D} \to \mathbb{R}^p$, $b_i : \Theta \times \mathcal{D} \to \mathcal{D}$, $\sigma_i : \Theta \times \mathcal{D} \to \mathbb{R}^{p\times r}$, $z : \Theta \times \mathcal{D} \times \mathcal{A} \times [0,1] \to \mathbb{R}^r$.

We usually view m as an estimator of the parameters in our model, while d represents the variance, or *inverse precision*, of our estimates. (The decomposition of the underlying state considered here is implicitly referred to as the 'knowledge state' in the Knowledge Gradient (KG) considered by Ryzhov et al. [27].)

Inspired by (2.2), (2.4) and (2.5), we make the following assumption.

H.1. *$\mathcal{D}$ is a compact set and there exists a constant $C > 0$ and a norm $\|\cdot\|$ on $\mathcal{D}$ such that*

(i) For any $i \in \mathcal{A}$ and $(m, d) \in \Theta \times \mathcal{D}$, $\|d + b_i(m, d)\| \leq \|d\|$.

(ii) For any $i \in \mathcal{A}$ and $(m, d) \in \Theta \times \mathcal{D}$, $m + \mu_i(m, d) + \sigma_i(m, d)z(m, d, i, \xi) \in \Theta$.

(iii) For any $i \in \mathcal{A}$ and $(m, d) \in \Theta \times \mathcal{D}$ and $t \in \mathbb{N}$, $\mathbb{E}[z(m, d, i, \boldsymbol{\xi}_t)] = 0$, $Var[z(m, d, i, \boldsymbol{\xi}_t)] = I_r$.

(iv) For any $i \in \mathcal{A}$, $(m, d) \in \Theta \times \mathcal{D}$ and $t \in \mathbb{N}$, $\mathbb{E}\big[\big|z(m, d, i, \boldsymbol{\xi}_t)\big|^3\big] \leq C$.

(v) For any $\psi \in \big\{b_i, \mu_i, \big(\sigma_i\sigma_i^\top\big) : i \in \mathcal{A}\big\}$, and $(m, d) \in \Theta \times \mathcal{D}$, we have

$$\sup_{m\in\Theta} |\psi(m, d)| \leq C\|d\|^2, \quad \sup_{m\in\Theta} |\partial_m\psi(m, d)| \leq C\|d\|^2, \quad \text{and} \quad \sup_{m\in\Theta} |\partial_d\psi(m, d)| \leq C\|d\|.$$

It is worth emphasising that the dynamics of our examples (2.2), (2.4) and (2.5) can be written in the form (2.6) and satisfy (H.1)(i)-(H.1)(iv), with $\Theta = \mathbb{R}^p$, $\Theta = [0,1]^p$ and $\Theta = [0,\infty)^p$, respectively. Here, we see that (H.1) holds for (2.2) with the operator norm on positive definite matrices whereas (H.1) holds for (2.4) and (2.5) with a standard Euclidean norm. Moreover, we

can see that (2.2) and (2.4) also satisfy (H.1)(v). Unfortunately, (2.5) does not satisfy (H.1)(v) due to appearance of m in the dynamics, and Θ is not bounded in this case. However, we may consider such dynamics by applying a smooth truncation to σ_i to ensures that (H.1)(v) holds. We can interpret this as an approximation of the learning dynamics in a Poisson setting.

Here, (H.1)(i) says that our precision (i.e. our knowledge) of the parameter always improves with more observations. (H.1)(ii) says that the updated parameter estimate always lies in our parameter set Θ. (H.1)$(iii)-(iv)$ ensures that the process $z(M_t, D_t, A_t, \xi_t)$ is a white noise with bounded third moment, and in this sense these are structural assumptions which allow us to interpret our dynamics in terms of μ_i and σ_i. Finally, (H.1)(v) encodes stability assumptions which appear naturally through the propagation of the information as discussed in (2.2), (2.4) and (2.5). (H.1)(v) is the key assumption for the asymptotic analysis that will be considered in the later sections.

Remark 2.1. In fact, we may weaken (H.1) to consider an ergodic diffusion with small perturbation, which is closely related to the Kalman-filtering theory. We state the corresponding assumption (H.2) here for precision of our discussion. Nonetheless, we will only focus on (H.1) for clarity and then discuss how to extend our analysis to this framework in Remark 4.3 and Theorem 4.9.

H.2. *(H.1) holds with* (i) *replaced by*

(i') *For any* $i \in \mathcal{A}$ *and* $(m,d) \in \Theta \times \mathcal{D}$, $\|d + b_i(m,d)\| \in \mathcal{D}$.

Assumption (H.1) describes the dynamics of the state of our controlled system. As illustrated in (2.1), using notation inspired by (2.3), the objective of our control problem is to find $U \in \mathcal{U}$ to maximize $V^U(m,d)$, where

$$V^U(m,d) := \mathbb{E}_{m,d}\Big[\sum_{t=0}^{\infty} \beta^t f_{A_{t+1}}\big(M_t^U, D_t^U\big)\Big] = \mathbb{E}_{m,d}\Big[\sum_{t=0}^{\infty} \beta^t \Big(\sum_{i=1}^{K} f_i\big(M_t^U, D_t^U\big) U_{i,t+1}\Big)\Big], \tag{2.7}$$

using controls $A_t = A(U_t, \zeta_t)$. We will assume that the expected reward $f : \Theta \times \mathcal{D} \to \mathbb{R}^K$ satisfies the following assumption.

H.3. *$f \in \mathcal{C}_b^3$, that is, f is 3-times differentiable, and there exists a constant $C \geq 0$ such that $|Df|, |D^2 f|, |D^3 f| \leq C$.*

We recall that $f_i(m,d)$ represents the expected value of the reward $r_i(Y_t^{(i)})$ when the distribution of $\boldsymbol{\theta}$ is parameterised by (m,d). In most classical frameworks, we find that $\mathbb{E}[r_i(Y_t^{(i)})|\boldsymbol{\theta}]$ is linear in $\boldsymbol{\theta}$. If $M_t := \mathbb{E}_\pi[\boldsymbol{\theta}|\mathcal{G}_t^U]$ where $\mathcal{G}_t^U$ is the historical observations at time t, the corresponding function f is linear in m and does not depend on d.

Assumption (H.1) implies that the state (m,d) (which corresponds to the posterior parameter) will not change much when $\|d\|$ is small. Therefore, it will be reasonable to consider an approximate solution to (2.7) in the asymptotic regime when $\|d\|$ is small.

2.3 Learning premia and index strategies

Consider the classical multi-armed bandit problem, which corresponds to the case $p = K$ and $f_i(m,d) = m_i$, and assume the underlying distribution is Gaussian as in Example 2.2.

For independent multi-armed bandits (specifically where d is diagonal), many algorithms (e.g. Gittins index [14], UCB [2], Bayes-UCB [15], or Knowledge Gradient [27]) make a decision based on an index strategy – the choice made maximises an index α, which has the form

$$\alpha_i = (\text{Exploitation gain})_i + (\text{Learning premium})_i. \tag{2.8}$$

The exploitation gain is often simply the expected reward (or m_i in this case), while the learning premium (or exploration gain) differs between algorithms.

While the Gittins index is not immediately of the form (2.8) (as it depends on solving an optimal stopping problem for each arm), Brezzi and Lai [3] and Russo [23] give approximations to the Gittins index and show that, in the Gaussian case described above, it admits a decomposition (2.8) where the learning premium scales with the standard error of the estimated average reward. This suggests that we should take the learning premium as an additional reward proportional to the uncertainty (statistical error) of our estimate. This is often known as the *optimistic principle*.

Since we add uncertainty as an additional reward, one could interpret this as a claim that we should prefer an uncertain options, in order to encourage learning. In general this is a misleading conclusion, as the following example shows.

Example 2.8 (Uncertainty Preference)**.** Let consider a bandit with two arms. Suppose that the reward of the first arm is sampled from $N(\boldsymbol{\theta}, 1)$ where $\boldsymbol{\theta}$ is not known, while the reward of the second arm is fixed and always 1. Suppose that we only collect the reward of the arm that we choose, but we always observe the reward of the first arm. Hence, we do not have to play the first arm to learn $\boldsymbol{\theta}$. Therefore, most decision makers (without taking any risk/ uncertainty aversion) will choose arms purely based on their estimate m of $\boldsymbol{\theta}$. In particular, they will choose the first arm if the estimate $m > 1$ and choose the second arm otherwise.

In the above example, the reward of the first arm is more uncertain than the second arm, but a preference for uncertainty does not benefit our decision. In fact, in many behavioural models, people have a bias against risk and uncertainty (see e.g. [7], Knight [18] and Keynes [16]). In the situation described in Example 2.8, rational decision makers still prefer the second arm when $m > 1$, since they want to avoid risky outcomes.

Pessimism and optimism towards uncertainty do not necessarily contradict each other (see [8] for further discussion). A better interpretation of the optimistic principle is that we have a preference for information gain or the *reduction* in uncertainty (rather than the uncertainty itself). In the case when the arms are independent (as in a classical bandit), it happens to be the case that the information gain corresponds to the uncertainty of the current estimate, resulting in optimism being optimal.

The above observation leads us to ask how we can quantify information gain? Consider a simple Greedy approach, where we choose $A_t \equiv \arg\max_i m_i$, i.e. the probability U_t^{Greedy} is the kth basis vector corresponding to $k = \arg\max_i m_i$. Here, we have $\boldsymbol{\theta} \sim N(m, d)$ and

$$0 \leq V(m,d) - V^U(m,d) \leq \sum_{t=0}^{\infty} \beta^t \mathbb{E}\big(\max_i \boldsymbol{\theta}_i - \max_i m_i\big) \leq (1-\beta)^{-1}\sqrt{\|d\| \log K}, \tag{2.9}$$

the first inequality follows from the fact that we cannot make a better decision than the best decision with known $\boldsymbol{\theta}$, and the second inequality follows from the Gaussian maximal inequality (see e.g. Chernozhukov et al. [6, Theorem 1]).

We may see the Greedy strategy as a first-order approximation to the optimal solution, considering only Exploitation gain in (2.8). This decision introduces error (relative to the best strategy) with order $\mathcal{O}(\|d\|^{1/2})$. This suggests an index (like) strategy, where we treat the learning premium as a second-order approximation (with respect to d). Unfortunately, as discussed in Reisinger and Zhang [20] (in a continuous time setting), the optimal randomised policy over a discrete set of actions is typically a Dirac measure, i.e. an optimal decision will always choose a single action with probability 1. In particular, the optimal control involves the argmax function, which is, in general, non-smooth, and can lead to sensitivity of the control to perturbation of the coefficients

(b, μ, σ, f). To overcome this non-smoothness, we consider an entropy regularised control problem (see also [32, 29]). We will use this regularised control to approximate $V(m, d)$ up to second-order when $\|d\|$ is small. We then show that this approximation (and its corresponding strategy) introduces an error of order $\mathcal{O}(\|d\|)$ compared to $\mathcal{O}(\|d\|^{1/2})$ for the naive greedy approach.

2.4 From a regularised control problem to the ARC algorithm

In this section, we will use an entropy regularised control to construct an approximation to the value function and sketch a derivation of the corresponding ARC algorithm to solve the diffusive control problem, introduced in Section 2.2. Our derivation here will be fully heuristic; the proof of the error of the approximation will be provided in Section 4.

We first observe that for $a \in \mathbb{R}^K$, we can approximate the maximum function by

$$\max_i a_i \approx S^\lambda_{\max}(a) := \sup_{u \in \Delta^K} \Big(\sum_{i=1}^K u_i a_i + \lambda \mathcal{H}(u) \Big) \tag{2.10}$$

where $\mathcal{H}$ is a smooth entropy function and $\lambda > 0$ is a small regularisation parameter. We also write

$$\nu^\lambda(a) := \partial_a S^\lambda_{\max}(a) = \arg\max_{u \in \Delta^K} \Big(\sum_{i=1}^K u_i a_i + \lambda \mathcal{H}(u) \Big). \tag{2.11}$$

Remark 2.2 (Shannon Entropy). For simplicity of our discussion, we defer the formal assumptions on $\mathcal{H}$ to Definition 4.2 in Section 4. The reader may simply suppose $\mathcal{H}$ is the Shannon entropy, $\mathcal{H}(u) := -\sum_{i=1}^K u_i \ln u_i$. In this case, we have $S^\lambda_{\max}(a) = \lambda \ln \big(\sum_{i=1}^K \exp\big(a_i/\lambda\big)\big)$ and $\nu^\lambda_i(a) = \exp\big(a_i/\lambda\big) / \big(\sum_{j=1}^K \exp\big(a_j/\lambda\big)\big)$.

By introducing an entropy function, we obtain a smooth approximation of the maximum $S^\lambda_{\max}(a)$, where $S^\lambda_{\max}(a) \to \max_i a_i$ as $\lambda \downarrow 0$ uniformly (Theorem 4.2).

Definition 2.1. Let $\mathcal{H} : \Delta^K \to \mathbb{R}$ be a smooth entropy function (Definition 4.2) and $\lambda > 0$. We define the regularised value function to be

$$\begin{aligned} V^\lambda_\infty(m, d) &:= \sup_{U \in \mathcal{U}} V^{\lambda, U}_\infty(m, d), \qquad \text{where} \\ V^{\lambda, U}_\infty(m, d) &:= \mathbb{E}_{m,d}\Bigg[\sum_{t=0}^\infty \beta^t \bigg(\Big(\sum_{i=1}^K f_i\big(M^U_t, D^U_t\big) U_{i,t+1} \Big) + \lambda \mathcal{H}(U_{t+1}) \bigg) \Bigg]. \end{aligned} \tag{2.12}$$

Inspired by the Gittins index (discussed in Section 2.3), we consider the index as an incremental value of choosing each option. Hence, we would like to construct a function $\alpha^\lambda : \Theta \times \mathcal{D} \to \mathbb{R}^K$ where the ith component of α^λ corresponds to an 'incremental reward' over a single step for choosing the ith option. In particular, let us assume that when $\|d\|$ is sufficiently small, the value function is approximated by the maximum amongst the values of the available options, i.e.,

$$V^\lambda_\infty(m, d) \approx (1 - \beta)^{-1} \big(S^\lambda_{\max} \circ \alpha^\lambda \big)(m, d). \tag{2.13}$$

By the dynamic programming principle and the dynamics given in (2.6), we can then rewrite (2.12) as

$$\begin{aligned} V^\lambda_\infty(m, d) &= \sup_{u \in \Delta^K} \Bigg(\sum_{i=1}^K \bigg(u_i \Big(f_i(m, d) + \beta \mathbb{E}[V^\lambda_\infty(\Phi(m, d, i, \boldsymbol{\xi}))] \Big) \bigg) + \lambda \mathcal{H}(u) \Bigg) \\ &\approx \sup_{u \in \Delta^K} \Bigg(\sum_{i=1}^K \bigg(u_i \Big(f_i(m, d) + \beta (1 - \beta)^{-1} \mathbb{E}\Big[\big(S^\lambda_{\max} \circ \alpha^\lambda \big)\big(\Phi(m, d, i, \boldsymbol{\xi})\big) \Big] \Big) \bigg) + \lambda \mathcal{H}(u) \Bigg). \end{aligned} \tag{2.14}$$

Using the approximation (2.13) on the LHS and rearranging the above expression, we obtain

$$\big(S^\lambda_{\max}\circ\alpha^\lambda\big)(m,d) \approx S^\lambda_{\max}\left(f(m,d)+\beta(1-\beta)^{-1}\mathbb{E}\Big[\big(S^\lambda_{\max}\circ\alpha^\lambda\big)\big(\Phi(m,d,\cdot,\boldsymbol{\xi})\big)-\big(S^\lambda_{\max}\circ\alpha^\lambda\big)(m,d)\mathbf{1}_K\Big]\right).$$

In particular, this suggests the approximation

$$\alpha^\lambda(m,d) \approx f(m,d)+\beta(1-\beta)^{-1}\mathbb{E}\Big[\big(S^\lambda_{\max}\circ\alpha^\lambda\big)\big(\Phi(m,d,\cdot,\boldsymbol{\xi})\big)-\big(S^\lambda_{\max}\circ\alpha^\lambda\big)(m,d)\mathbf{1}_K\Big]. \tag{2.15}$$

Since our dynamics form a discrete diffusion process, a discrete-time version of Ito's lemma shows that a solution to the approximate fixed-point in (2.15) is given by

$$\alpha^\lambda(m,d) := f(m,d)+\beta(1-\beta)^{-1}L^\lambda(m,d), \tag{2.16}$$

where $L^\lambda:\Theta\times\mathcal{D}\to\mathbb{R}^K$ is given by

$$L^\lambda_i(m,d) := \langle\mathcal{B}^\lambda(m,d);b_i(m,d)\rangle+\langle\mathcal{M}^\lambda(m,d);\mu_i(m,d)\rangle+\frac{1}{2}\langle\Sigma^\lambda(m,d);\sigma_i\sigma_i^\top(m,d)\rangle. \tag{2.17}$$

Here, $\mathcal{B}^\lambda:\Theta\times\mathcal{D}\to\mathbb{R}^p$, $\mathcal{M}^\lambda:\Theta\times\mathcal{D}\to\mathcal{D}$ and $\Sigma^\lambda:\Theta\times\mathcal{D}\to\mathbb{R}^{p\times p}$ are given by

$$\mathcal{B}^\lambda(m,d) := \sum_{j=1}^K\nu^\lambda_j\big((f(m,d)\big)\partial_d f_j(m,d), \qquad \mathcal{M}^\lambda(m,d) := \sum_{j=1}^K\nu^\lambda_j\big((f(m,d)\big)\partial_m f_j(m,d), \tag{2.18}$$

$$\Sigma^\lambda(m,d) := \sum_{j=1}^K\Big(\nu^\lambda_j\big((f(m,d)\big)\partial^2_m f_j(m,d)\Big)+\frac{1}{\lambda}\sum_{i,j=1}^K\Big(\eta^\lambda_{ij}\big((f(m,d)\big)\big(\partial_m f_i(m,d)\big)\big(\partial_m f_j(m,d)\big)^\top\Big).$$

with $\eta^\lambda(a) := \lambda\partial^2_a S^\lambda_{\max}(a)$ (which is determined by $\eta^\lambda_{ij}(a)=\nu^\lambda_i(a)(\mathbb{I}(i=j)-\nu^\lambda_j(a))$ for the case of Shannon Entropy given in Remark 2.2). It is worth noting that the terms $\mathcal{B}^\lambda$, $\mathcal{M}^\lambda$ and Σ^λ are derivatives ∂_d,∂_m and ∂^2_m of $S^\lambda_{\max}\circ f$.

Using this approximation for α^λ in (2.14) and (2.15), we can write

$$V^\lambda_\infty(m,d) \approx \sup_{u\in\Delta^K}\left(\sum_{i=1}^K u_i\alpha^\lambda_i(m,d)+\lambda\mathcal{H}(u)\right)+\beta(1-\beta)^{-1}\big(S^\lambda_{\max}\circ\alpha^\lambda\big)(m,d), \tag{2.19}$$

with an error which is small for small $\|d\|$. In particular, an approximate solution to the dynamic program (2.14) is given by $u^{*,\lambda}(m,d)\approx\nu^\lambda\big(\alpha^\lambda(m,d)\big)$. This suggests we can choose an arm using a softmax function applied to the the index $\alpha^\lambda(m,d)$, and this gives an approximately optimal strategy.

Remark 2.3. While our approximation result is will focus on the case when $\|d\|$ is small, the algorithm we propose should operate well in a wide range of settings. This is because the key approximation step is the one-step-in-the-future approximation (2.13) of the value function. In other words, when we use the ARC algorithm, we account for the short-run consequences of our actions on our learning, and approximate these using a quadratic expansion (2.17), leading to computational efficiency. Even when $\|d\|$ is not particularly small, this provides an efficient first-order correction, leading to good performance.

2.5 Description of the main results

We summarise our main results here. The proofs can be found in Section 4.

As discussed in the previous section, by introducing an entropy regularisation to our control problem, we can obtain an approximation due to the smoothness of $S^{\lambda}_{\max}$. Unfortunately, derivatives of this smooth approximation explode as $\lambda \to 0$. This means that the approximation error of (2.19) may explode if we take $\lambda \to 0$ while fixing $\|d\|$. Hence, we propose an algorithm where λ is a function of (m, d), to optimise the apparent trade-off between these sources of error. We introduce the following function to give a precise description of our upper error bound,

$$P_{\lambda,d}(m,n) := \big(1+\lambda^{-m}\big)\|d\|^n \quad ; \quad \lambda \in (0,\infty), \quad d \in \mathcal{D}, \quad m,n \in \mathbb{N}. \tag{2.20}$$

Error bound in the regularised control problem (Theorem 4.6 and Theorem 4.7) : With α^{λ} as defined in (2.16), there exists a constant $C \geq 0$ such that the optimal value V^{λ}_{∞} in (2.12) satisfies

$$\big|V^{\lambda}_{\infty}(m,d) - (1-\beta)^{-1}\big(S^{\lambda}_{\max}\circ\alpha^{\lambda}\big)(m,d)\big| \leq C\big(P_{\lambda,d}(2,3)+P_{\lambda,d}(3,4)\big).$$

Moreover,

$$\big|V^{\lambda}_{\infty}(m,d) - V^{\lambda,U^{\lambda}}_{\infty}(m,d)\big| \leq C\big(P_{\lambda,d}(2,3)+P_{\lambda,d}(3,4)\big)$$

where $V^{U^{\lambda}}$ is the value (2.12) when using the control U^{λ} corresponding to the feedback strategy $(m,d) \mapsto \nu^{\lambda}\big(\alpha^{\lambda}(m,d)\big)$.

Our objective is to maximise the unregularised problem (2.7). Introducing the regulariser introduces an error of $\mathcal{O}(\lambda)$. On the other hand, our approximation introduces an error which possibly explodes as $\lambda \to 0$. Hence, we introduce $\boldsymbol{\lambda}$ as a function function of $(m,d) \in \Theta \times \mathcal{D}$ explore the trade-off between λ and the error of the approximation.

Definition 2.2. We say a function $\boldsymbol{\lambda} : \Theta \times \mathcal{D} \to (0,\infty)$ is *consistent* with order $\kappa \geq 0$ if there exist $\underline{c}, \bar{c} > 0$ such that for all $(m,d) \in \Theta \times \mathcal{D}$, $\underline{c}\|d\|^{\kappa} \leq \boldsymbol{\lambda}(m,d) \leq \bar{c}\|d\|^{\kappa}$.

Error bound in the unregularised control problem (Theorem 4.8) : For any consistent $\boldsymbol{\lambda}$ with order κ, there exists a constant $C \geq 0$ such that for any $(m,d) \in \Theta \times \mathcal{D}$,

$$V(m,d) - V^{U^{\boldsymbol{\lambda}}}(m,d) \leq C(\|d\|^{3-2\kappa} + \|d\|^{4-3\kappa} + \|d\|^{\kappa} + \|d\|^4) \tag{2.21}$$

where $V^{U^{\boldsymbol{\lambda}}}$ is the value (2.7) when using control $U^{\boldsymbol{\lambda}}$ corresponding to the feedback strategy $(m,d) \mapsto \nu^{\boldsymbol{\lambda}(m,d)}\big(\alpha^{\boldsymbol{\lambda}(m,d)}(m,d)\big)$ and V is the optimal value.

We see from (2.21) that when $\kappa = 1$, this asymptotic strategy introduces error of order $\mathcal{O}(\|d\|)$ compared with $\mathcal{O}(\|d\|^{1/2})$ for the naive greedy approach (Section 2.3).

Finally, we see that the strategy $U^{\boldsymbol{\lambda}}$ ensures that at the limit, we have a perfect knowledge of the parameter and hence can asymptotically achieve the best option.

Complete Learning (Theorem 4.10) : For any consistent $\boldsymbol{\lambda}$ with order $\kappa \in [0,2]$, $\|D_t^{U^{\boldsymbol{\lambda}}}\| \to 0$ almost surely as $t \to \infty$.

2.6 Summary of the ARC Algorithm

We now summarise how our approximation of the regularised control problem yields an explicit algorithm for a (correlated) multi-armed bandit.

Suppose that our bandit has K arms with an unknown parameter $\boldsymbol{\theta}$, as described at the beginning of Section 2. Suppose that when the ith arm is chosen, we observe a random variable $Y^{(i)} \sim \pi_i(\cdot|\boldsymbol{\theta})$ and obtain a reward $r_i(Y^{(i)})$.

We assume that the observation distribution π_i and the prior π form a conjugate pair for all i and we can parameterise the posterior distribution by (m, d), such that the dynamics of these posterior parameters satisfy (H.1). Broadly speaking, we think of m as a posterior mean of $\boldsymbol{\theta}$ and d as a posterior variance (or some quantity which is inversely proportional to the number of observations).

To use the ARC algorithm, we need to evaluate the following quantities explicitly:

(i) **The expected reward** when each arm is chosen, conditional on the posterior dynamics and its first two derivative, i.e. we need to evaluate $f_i(m, d) := \mathbb{E}_{m,d}[r_i(Y^{(i)})]$ and compute $\partial_d f_i(m, d), \partial_m f_i(m, d)$ and $\partial_m^2 f_i(m, d)$.

(ii) **Evolution of posterior parameters:** suppose that after choosing the ith arm, we observe $Y^{(i)}$ and obtain the new posterior parameter $(\boldsymbol{m}_i, \boldsymbol{d}_i) = \tilde{\Phi}(m, d, Y^{(i)}, i)$. We need to evaluate

$$\mu_i(m, d) := \mathbb{E}_{m,d}[\boldsymbol{m}_i - m], \quad \sigma_i\sigma_i^\top(m, d) := \mathrm{Var}_{m,d}(\boldsymbol{m}_i) \quad \text{and} \quad b_i(m, d) := \mathbb{E}_{m,d}[\boldsymbol{d}_i - d].$$

The ARC algorithm also has a few hyper-parameters tuned by the decision maker.

(i) **The discount factor β:** This parameter reflects how long are we considering the learning environment. A heuristic choice is to choose $\beta = 1 - 1/T$ where T is the number of rounds of decisions we need to make.

(ii) **Smooth max approximator S:** This is a function to approximate the maximum function. For simplicity, we propose the log-sum-exp function, $S(a) = \log\left(\sum_i \exp(a_i)\right)$ which has a derivative corresponding to the (arg)softmax function and is computationally convenient. A more general choice of S can be made by considering Definition 4.2.

(iii) **Regulariser function $\boldsymbol{\lambda}$:** We choose a regulariser (a function of m and d) to reflect our learning preferences. We see in (2.21) that choosing $\boldsymbol{\lambda}(m, d) = \mathcal{O}(\|d\|)$ gives the best order for the sub-optimal bound. A heuristic choice is thus to choose $\lambda(m, d) := |\langle r(m); d\rangle|$ where $r : \Theta \to \mathbb{R}^q$ is bounded above and away from 0. In fact, one may simply choose $\boldsymbol{\lambda}(m, d) = \rho\|d\|$.

Given the environment $(f, \mu_i, \sigma_i, b_i)$ and hyper-parameters $(\beta, S, \boldsymbol{\lambda})$, we can evaluate the function $(\lambda, m, d) \mapsto L^\lambda(m, d)$ as in (2.17) with $S^\lambda_{\max}(a) := \lambda S(a/\lambda)$ and write $\nu^\lambda(a) := \partial_a S^\lambda_{\max}(a) = \partial_a S(a/\lambda)$. We describe the procedure of the ARC algorithm as follow.

Algorithm 1: ARC Algorithm

Input m_0, d_0, T ;
Set $(m, d) \leftarrow (m_0, d_0)$;
for *t = 1, 2,..., T* **do**
  Evaluate $\nu := \nu^{\boldsymbol{\lambda}(m,d)}\big(f(m, d) + \beta(1 - \beta)^{-1} L^{\boldsymbol{\lambda}(m,d)}(m, d)\big)$;
  Sample $A \sim \mathrm{Random}(\{1, 2, ..., K\}, \nu)$;
  Choose the Ath arm, observe $Y^{(A)}$ and collect the reward $R^{(A)}(Y^{(A)})$;
  Update $(m, d) \leftarrow \tilde{\Phi}(m, d, Y^{(A)}, A)$;
end

3 Comparison with other approaches to bandit problems

3.1 General Approaches

There are various approaches to study bandit problems, and theoretical guarantees are typically proved in specific settings (see e.g. Lattimore and Szepesvári [19]). We summarise the broad idea of a few approaches and extend them to our setting using Bayesian inference when needed. For simplicity, we write $f_i(m,d) := \mathbb{E}_{m,d}[r_i(Y_t^{(i)})]$, $\mathbb{E}_\theta[\cdot] := \mathbb{E}[\cdot|\boldsymbol{\theta} = \theta]$ for the expected reward, (m,d) for the posterior parameter at a given time and $(\boldsymbol{m}_i, \boldsymbol{d}_i) = \tilde{\Phi}(m,d,Y^{(i)},i)$ for the posterior update of (m,d) when the ith arm is chosen.

- **ϵ-Greedy (ϵ-GD) [10, 31]:** At each time, we choose an arm with the maximal expected reward $A^{\epsilon-GD} = \arg\max_i f_i(m,d)$ with probability $1-\epsilon$; and choose uniformly at random with probability ϵ.

- **Boltzmann Exploration (BE) [28, 5]:** At each time, we choose an arm using the probability simplex $U^{BE} = \nu^{\boldsymbol{\lambda}(m,d)}\big(f(m,d)\big)$ where $\nu_i^\lambda(a) := \exp(a_i/\lambda)/\big(\sum_j \exp(a_j/\lambda)\big)$.

- **Thompson Sampling (TS) [30, 26, 24]:** At each time, a sample $\hat{\boldsymbol{\theta}}$ is taken from a posterior parameter $\pi_{(m,d)}$. We then choose the arm with $A^{TS} = \arg\max_i \mathbb{E}_{\hat{\boldsymbol{\theta}}}[r_i(Y^{(i)})]$.

- **Upper Confidence Bound (UCB) [1, 2, 15] :** At each time t, we choose an arm with the maximum index $A_t^{Bayes-UCB} = \arg\max_i \mathcal{Q}_{m,d}\big(1-t^{-1}(\log T)^{-c}, \mathbb{E}_{\boldsymbol{\theta}}[r_i(Y^{(i)})]\big)$ where T is the number of plays, $\mathcal{Q}_{m,d}(p,X)$ is the p-quantile of the random variable X conditional on the posterior parameter (m,d). Here, c is a hyper-parameter that can be chosen by the decision maker. Kaufmann et al. [15] prove a theoretical guarantee of optimal order for the Bernoulli bandit when $c \geq 5$; their simulations suggest that $c = 0$ performs best.

  NB. There are many variations of the UCB algorithm proposed in various settings. Bayes-UCB is one of this class which has a very clear extension to the general setting described in this paper.

- **Knowledge Gradient (KG) [27]:** At each time t, we choose an arm with the maximum index $A^{KG} = \arg\max_i \Big(f_i(m,d) + \beta(1-\beta)^{-1}\big(\mathbb{E}_{m,d}[\max_j f_j(\boldsymbol{m}_i,\boldsymbol{d}_i)] - \max_j f_j(m,d)\big)\Big)$.

  NB. In [27], an algorithm is proposed for the classical and linear Gaussian bandit together with an explicit expression for $\mathbb{E}_{m,d}[\max_j f_j(\boldsymbol{m}_i,\boldsymbol{d}_i)]$. In general, we may estimate this expression by using Monte-Carlo simulation, but this can be costly.

- **Information-Directed Sampling (IDS) [25, 17]:** At each time, for each probability vector $u \in \Delta^K$, we define a (one-step) regret by $\delta(u) := \sum_i u_i \mathbb{E}_{m,d}\big[r_{A^*}(Y^{(A^*)}) - r_i(Y^{(i)})\big]$ where $A^* := \arg\max_i \mathbb{E}_{\boldsymbol{\theta}}[r_i(Y^{(i)})]$, and define the information gain by $g(u) := \sum_i u_i\big(\mathcal{H}_{m,d}^{Sh}(\boldsymbol{\theta}) - \mathbb{E}_{m,d}[\mathcal{H}_{\boldsymbol{m}_i,\boldsymbol{d}_i}^{Sh}(\boldsymbol{\theta})]\big)$ where $\mathcal{H}_{m,d}^{Sh}(\boldsymbol{\theta}) := \int \big(\frac{\mathrm{d}\pi_{m,d}^{\boldsymbol{\theta}}}{\mathrm{d}\mu}\big)\log\big(\frac{\mathrm{d}\pi_{m,d}^{\boldsymbol{\theta}}}{\mathrm{d}\mu}\big)\mathrm{d}\mu$ is the (differential) entropy of the posterior $\pi_{m,d}^{\boldsymbol{\theta}}$ with parameter (m,d). Here, $\big(\frac{\mathrm{d}\pi_{m,d}^{\boldsymbol{\theta}}}{\mathrm{d}\mu}\big)$ is the Radon–Nikodym density of the posterior distribution $\pi_{m,d}^{\boldsymbol{\theta}}$ with respect to Lebesgue measure μ. We then choose an arm using the probability $U^{IDS} = \arg\min_{u\in\Delta^K}\big(\delta(u)^2/g(u)\big)$.

  NB. The information gain g considered above is used in Kirschner and Krause [17] which is different from the original proposed in Russo and Roy [25]; but is more computationally efficient.

ARC as a combination of the other algorithms We observe that the derived ARC algorithm appears as a combination of KG and BE through Itô's lemma, which results in a random index decision whose index can be decomposed as the sum between Exploitation gain and Learning premium, as in the UCB principle. This learning premium takes into account asymmetry of available information and a curvature term when the reward is non-linear (see section 3.3 for explicit evaluation of L^λ). This additional exploration gain is closely related to Boltzmann–Gumbel exploration (BGE) proposed by Cesa-Bianchi et al. [5], where they modify the BE by adding an external noise scaling with uncertainty to encourage learning. The ARC algorithm is similar to this, in the sense that we randomly make decisions, as in the BE algorithm, but a deterministic quantity corresponding to the reduction of uncertainty is added to encourage learning.

More precisely, recall that ARC chooses an arm based on the (arg)softmax $\nu^\lambda(\alpha(m,d))$ which is most likely to pick an arm with the maximum index $\alpha_i(m,d) = f_i(m,d) + \beta(1-\beta)^{-1}L_i^\lambda(m,d)$. The choice made is determined at random, as in BE, but using a modified index, as in BGE. In ARC the decision is modified through the learning term $L_i^\lambda(m,d)$, which can be seen as $L_i^\lambda(m,d) \approx \mathbb{E}_{m,d}[\max_j f_j(\boldsymbol{m}_i, \boldsymbol{d}_i)] - \max_j f_j(m,d)$. This also motivates the view of $\alpha(m,d)$ as an approximation of the KG index using a smooth max approximation and a second-order expansion through Itô's lemma. The decision is then made using the (arg-)softmax probability derived from $\alpha(m,d)$.

Computational Efficiency ϵ-GD, BE, TS, UCB and ARC are algorithms where we can often find an explicit expression and thus require negligible computational power. KG, on the other hand, requires evaluation of the expectation of the maximum involving a high-dimensional state (which only has an explicit expression in the Gaussian case). Implementing KG in general can be achieved by Monte-Carlo simulation, which can be costly. Similarly, the IDS requires evaluation of one-step regret and information gain, which is expensive in general.

3.2 Shortcomings of bandit algorithms

Even though many algorithms discussed in Section 3.1 perform well, they may fail to address a few phenomena which may appear in learning. For clarity, we shall illustrate these shortcomings in extreme scenarios. Many practical examples of these scenarios can be found in Russo and Roy [25].

Incomplete learning of time-homogeneous deterministic policies. Consider a policy which is deterministic and time-homogeneous, in that it depends only on the posterior parameter (m,d). If we start with a bad prior, leading us to initially choose an option which results in limited learning, we may never explore the system fully. This results in poor performance if our early experiences are misleading.

More concretely, consider a bandit with 2 arms: the first arm always gives a known reward; the second arm's reward is generated from an unknown distribution. Given a time-homogeneous deterministic policy, whenever this strategy decides to play the first arm, it will never play the second arm again. However, we can see that if the mean reward of the second arm has unbounded support, the probability that the mean reward of the first arm is smaller than the second arm is strictly positive. This probability never changes when the first arm is played, which means that we have a strictly positive probability of always playing sub-optimal options.

This is a problem for ϵ-GD (when $\epsilon = 0$) and KG algorithms.

Information Ignorance (Asymmetric Learning). Many works on bandit problems assume that all arms have an identical structure. Therefore, they fail to capture the setting where each arm provides different information.

Consider a bandit with 100 arms where every arm except the first always gives a strictly positive reward from an unknown distribution. The first arm is informative but costly; it always gives a reward 0, but will identify the parameters of all other arms after it is played once. In this case, the first arm never has the best reward and will be ignored by most bandit algorithms (since the information gain does not appear directly through its reward). Nevertheless, if the rewards of the remaining arms are sufficiently variable, it is optimal to play the first arm once, in order to eliminate uncertainty.

This is a problem for ϵ-GD (when $\epsilon = 0$), TS, and UCB algorithms. It is worth noting that BE and ϵ-GD (for general ϵ) will only choose the first arm at random. Hence, they will not optimize to only play the first arm once, despite it revealing any additional information.

Horizon effect A few algorithms are designed based on the principle that the decision should not vary when the terminal time is far away, and hence propose a stationary policy. When the horizon is short, these algorithms may still choose to explore, even if these explorations do not benefit future decisions. A trivial example of this is when there is only one choice remaining before the horizon, at which point the optimal action is a fully greedy policy, as there is no value to further information.

This is a problem for ϵ-GD, BE, TS and IDS algorithms.

3.3 Addressing shortcomings using ARC

The ARC algorithm (Section 2) automatically addresses the above flaws. Since we can vary a hyper-parameter β, which has a natural interpretation as the future weight, we directly address the *horizon effect* . Since the ARC algorithm is chosen based on (arg)soft-max, which takes values in the interior of the probability simplex Δ^K, it has the capacity to overcome incomplete learning. We also show (Theorem 4.10) that under appropriate conditions, ARC is a *complete learning* algorithm in the sense that the posterior (variance) parameter $D_t \to 0$ as $t \to \infty$.

To understand ARC in an *asymmetric learning* environment, and the effect of curvature, we shall illustrate explicit computation of the ARC algorithm for the Gaussian bandit with additional information:

Bandits with additional information Consider Example 2.4, when the rewards of the arms are uncorrelated but choosing the ith arm may allow us to observe information related to the reward of the jth arm.

We recall Example 2.4 for convenience of the reader. When the ith arm is chosen, we collect the reward $r_i(Y^{(i,i)})$ and observe a random variable $Y^{(i,j)} \sim N(\boldsymbol{\theta}_j, s_{ij}^{-1})$ for $j = 1, 2, ..., K$ where $s_{ij} \in [0, \infty)$ is known but $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^K$ is not known. Here, we see that the reward of the ith arm depends only on $\boldsymbol{\theta}_i$ and the reward of each arm also differs depending on the function r_i. Furthermore, we observe that when s_{ij} is small, the variance of the observation $Y^{(i,j)}$ is large. This means that it tells us very little about $\boldsymbol{\theta}_j$ (and vice versa for large s_{ij}).

We have parameters (m, d), corresponding to the posterior mean and posterior variance of $\boldsymbol{\theta}$. Assume that we use Shannon entropy (Remark 2.2) as our regulariser. The decision maker chooses using the (soft-)argmax of the index $\alpha^\lambda(m, d) = f(m, d) + \beta(1 - \beta)^{-1} L^\lambda(m, d)$ where we

can give an explicit expression (see Lemma **??**) for $L_i^\lambda(m,d)$:

$$L_i^\lambda(m,d) = \frac{1}{2\lambda}\sum_{j=1}^{K} \nu_j^\lambda(f(m,d))(1-\nu_j^\lambda(f(m,d)))d_{jj}^2\Big(\frac{s_{ij}}{1+d_{jj}s_{ij}}\Big)\big(\mathbb{E}_{m,d}[r_j'(Y^{(j,j)})]\big)^2. \tag{3.1}$$

The term $\beta(1-\beta)^{-1}L_i^\lambda(m,d)$ can be interpreted as a learning premium for choosing the ith arm. We can decompose this as follows: $\beta(1-\beta)^{-1}$ describes the importance of the future in our preferences, while $L_i^\lambda(m,d)$ comes from computing the total (learning) benefits of playing the ith arm to each jth arm:

- $d_{jj}^2\big(\frac{s_{ij}}{1+d_{jj}s_{ij}}\big)$ describes how much we can reduce uncertainty of the jth arm (see (2.4) or (2.5)); d_{jj} represents the uncertainty, while s_{ij} tells us how much information we would gain.
- $\big(\mathbb{E}_{m,d}[r_j'(Y^{(j,j)})]\big)^2$ which rescales the learning benefit to account for how our parameters impact the reward function (in particular, the curvature of the reward), as there is no benefit in learning parameters which do not yield a reward.
- $\nu_j^\lambda(f(m,d))(1-\nu_j^\lambda(f(m,d)))$ describes whether we see value in learning the parameter of the jth arm. In particular, when we have an arm i^* such that $f_{i^*}(m,d) \gg f_j(m,d)$ for $j \neq i^*$, the i^*th arm is much better than the others, and we probably do not need to learn further. In this case, we have $\nu^\lambda(f(m,d)) \approx e_{i^*}$ and $\nu_j^\lambda(f(m,d))(1-\nu_j^\lambda(f(m,d))) \approx 0$ for all j.

If we replace λ by a consistent $\boldsymbol{\lambda}$ with order $\kappa = 1$, we see that $L_i^\lambda(m,d) \sim \|d\|$ (which is slightly different from the Learning Premium of the UCB, which scales with $\|d\|^{1/2}$). Roughly speaking, in this setting d could also be interpreted as $(1/n_1, ..., 1/n_K)$ where n_i is the number of attempts on the ith arm.

This decomposition highlights that ARC is intrinsically flexible in its representation of the value of learning, as it accounts for different elements of how this learning will impact future decision making. This highlights how ARC is capable of overcoming the challenges of asymmetric learning.

4 Derivation of ARC approximation

We now flesh out our heuristic derivation in Section 2.4.

4.1 Convex Analysis and smooth max approximator

In (2.10) and (2.11), we briefly introduce $S_{\max}^\lambda$ and ν^λ as a smooth versions of the maximum and arg-maximum functions, obtained from a regularisation function $\mathcal{H}$. One well-known choice of $\mathcal{H}$ which gives us an analytical expression is the Shannon entropy. In general, we can consider other choices of $\mathcal{H}$ by constructing the smooth max approximator $S_{\max}^\lambda$ explicitly. Examples of explicit constructions can be found in Zhang and Reisinger [20, Remark 3.1].

In order to give a formal derivation of the ARC algorithm in a general regularisation framework, we recall some results from convex analysis.

Observe that $S_{\max}^\lambda$ in (2.10) can be expressed as $S_{\max}^\lambda(a) = \lambda S(a/\lambda)$ where

$$S(a) = \sup_{u\in\Delta^K}\Big(\sum_{i=1}^{K} u_i a_i + \mathcal{H}(u)\Big). \tag{4.1}$$

In particular, $-\mathcal{H}$ is the convex conjugate of S (see e.g. Rockafellar [21]). In fact, (4.1) is also known as a 'nonlinear expectation' [1] defined on a finite space (see Coquet et al. [9]) and (4.1) is known as the 'robust representation'.

Definition 4.1. We say a function $S : \mathbb{R}^K \to \mathbb{R}$ is a *convex nonlinear expectation* if it satisfies:

(i) **Monotonicity**: If $a \leq b$, then $S(a) \leq S(b)$;

(ii) **Translation Equivariance**: For all $c \in \mathbb{R}$, $S(a + c\mathbf{1}_K) = S(a) + c$;

(iii) **Convexity**: For any $\kappa \in [0, 1]$, $S(\kappa a + (1 - \kappa)b) \leq \kappa S(a) + (1 - \kappa)S(b)$;

where the inequalities are interpreted component-wise.

Theorem 4.1 (Robust Representation)**.** *A convex nonlinear expectation S admits a representation of the form $S(a) = \sup_{u \in \Delta^K} \left(\sum_{i=1}^K u_i a_i + \mathcal{H}_{\max}(u) \right)$, where $\mathcal{H}_{\max}(u) := -\sup_{a \in \mathcal{A}_S} \left(\sum_{i=1}^K u_i a_i \right)$ and $\mathcal{A}_S := \{a \in \mathbb{R}^K : S(a) \leq 0\}$.*

Furthermore, $\mathcal{H}_{\max}$ is the maximal function which represents S, i.e. if there exists $\mathcal{H}$ such that (4.1) holds with $\mathcal{H}$, then $\mathcal{H}(u) \leq \mathcal{H}_{\max}(u)$ for all $u \in \Delta^K$.

Proof. See Föllmer and Schied [12, Theorem 4.16] or Frittelli and Rosazza Gianin [13]. □

The following theorem shows that using $S^\lambda_{\max}$ as a smooth max approximator is equivalent to having $\mathcal{H}$ bounded. This will allow us to quantify the difference between a non-regularised (2.7) and regularised control problem (2.12) by $\mathcal{O}(\lambda)$. The proof, along with others from this section, is given in Appendix A.

Theorem 4.2. *Let S be a convex nonlinear expectation. The following are equivalent.*

(i) There exists $N \in \mathbb{R}$ such that $S(a) + N \geq \max_i a_i$ for all $a \in \mathbb{R}^K$.

(ii) There exists $N \in \mathbb{R}$ such that $\mathcal{A}_S := \{a \in \mathbb{R}^K : S(a) \leq 0\} \subseteq (-\infty, N]^K$.

(iii) There exists a bounded function $\mathcal{H} : \Delta^K \to \mathbb{R}$ such that (4.1) holds.

(iv) For $S^\lambda_{\max}(a) = \lambda S(a/\lambda)$, we have $\sup_{a \in \mathbb{R}} | S^\lambda_{\max}(a) - \max_i a_i| \to 0$ as $\lambda \downarrow 0$.

We now introduce a smooth entropy regulariser as the convex conjugate of a smooth max approximator.

Definition 4.2. We say a function $S : \mathbb{R}^K \to \mathbb{R}$ is a *smooth max approximator* if it is a 3-times differentiable convex nonlinear expectation with uniformly bounded derivatives such that Theorem 4.2 holds. We say a bounded function $\mathcal{H} : \Delta^K \to \mathbb{R}$ is a *smooth entropy* if $-\mathcal{H}$ is a convex conjugate of some smooth max approximator.[2].

For a smooth max approximator S, we write $S^\lambda_{\max}(a) := \lambda S(a/\lambda)$, $\nu^\lambda(a) := \partial_y S\big|_{y=a/\lambda} = \partial_a S^\lambda_{\max}(a)$ and $\eta^\lambda(a) := \partial^2_y S\big|_{y=a/\lambda} = \lambda \partial^2_a S^\lambda_{\max}(a)$.

We will hereafter assume that we are given a smooth entropy function and corresponding smooth max approximator.

[1] Nonlinear Expectations (or equivalently 'risk measures') are a classical tool in mathematical finance to study decision making under uncertainty.

[2] One can check that the Shannon Entropy (Remark 2.2) is a smooth entropy.

H.4. *The regularised function $\mathcal{H}$ is a smooth entropy (Definition 4.2) with corresponding smooth max approximator S.*

Remark 4.1. If S is a smooth max approximator, then ν^λ and η^λ are uniformly bounded. Moreover, it follows from Fenchel's inequality that $\nu^\lambda(a) = \arg\max_{u\in\Delta^K} \big(\sum_{i=1}^K u_i a_i + \lambda\mathcal{H}(u)\big)$. In particular, ν^λ can be interpreted as a smooth version of the argmax, as in (2.11).

4.2 Analysis of the regularised control problem over finite horizon

The objective of this section is to approximate the finite horizon value function as a sum of the (smooth) maximum of the incremental rewards. In particular, we show that the regularized value

$$\begin{aligned} V_T^{\lambda,U}(m,d) &:= \mathbb{E}_{m,d}\left[\sum_{t=0}^{T-1}\beta^t\Big(\Big(\sum_{i=1}^K f_i\big(M_t^U, D_t^U\big)U_{i,t+1}\Big) + \lambda\mathcal{H}(U_{t+1})\Big)\right] \\ V_T^{\lambda}(m,d) &:= \sup_{U\in\mathcal{U}} V_T^{\lambda,U}(m,d) \end{aligned} \tag{4.2}$$

satisfies $V_T^\lambda(m,d) \approx \sum_{t=0}^{T-1}\beta^t\big(S_{\max}^\lambda \circ \alpha_{T-t}^\lambda\big)(m,d)$, where the incremental reward with t-steps to go is given by

$$\alpha_t^\lambda(m,d) := f(m,d) + L^\lambda(m,d)\Big(\sum_{s=1}^{t-1}\beta^s\Big) \tag{4.3}$$

and L^λ is given in (2.17).

The idea behind the analysis is to consider an asymptotic expansion as $\|d\| \to 0$. Due to our learning structure (H.1), the change in the underlying state (m,d) are (in expectation) of order $\mathcal{O}(\|d\|^2)$. Hence, the global Lipschitz property of f implies that the instantaneous reward changes with $\mathcal{O}(\|d\|^2)$. Hence, we can use Taylor's theorem to obtain an asymptotic expansion in $\|d\|$ keeping the terms up to order $\mathcal{O}(\|d\|^2)$, and then quantify coefficients in terms of λ and β. We shall quantify the error bounds in terms of $P_{\lambda,d}(m,n) := \big(1+\lambda^{-m}\big)\|d\|^n$ as in (2.20).

We first quantify the perturbation error in the learning term L^λ. We will show that this error is $\mathcal{O}(\|d\|^3)$ and can be ignored in our approximation.

Lemma 4.3. *Suppose that (H.1), (H.3) and (H.4) hold. There exists a constant $C>0$ such that for any $\lambda>0$, $i\in\mathcal{A}$, $(m,d)\in\Theta\times\mathcal{D}$ and $t\in\mathbb{N}$,*

$$\mathbb{E}\big|L^\lambda\big(\Phi(m,d,i,\boldsymbol{\xi}_t)\big) - L^\lambda(m,d)\big| \le CP_{\lambda,d}(2,3)$$

where $P_{\lambda,d}$ is a polynomial defined in (2.20).

Now, we consider the second order approximation of the smooth maximum $S_{\max}^\lambda$ over the incremental reward α. We show that L^λ is the second order approximation (in expectation) of the (smooth) maximum incremental reward.

Lemma 4.4. *Suppose that (H.1), (H.3) and (H.4) hold. There exists a constant $C>0$ such that for any $\lambda>0$, $i\in\mathcal{A}$, $(m,d)\in\Theta\times\mathcal{D}$ and $t,T\in\mathbb{N}$,*

$$\Big|\mathbb{E}\big(S_{\max}^\lambda\circ\alpha_T^\lambda\big)\big(\Phi(m,d,i,\boldsymbol{\xi}_t)\big) - \big(S_{\max}^\lambda\circ\alpha_T^\lambda\big)(m,d) - L_i^\lambda(m,d)\Big| \le C\big(P_{\lambda,d}(2,3) + P_{\lambda,d}(3,4)\big)$$

where α_t^λ is defined in (4.3).

As discussed in (2.19), in this finite horizon case, we are interested in the approximate optimal strategy given by

$$U^{\lambda,T}_{t+1} = \nu^\lambda\Big(\alpha^\lambda_{T-t}(M^{U^{\lambda,T}}_t, D^{U^{\lambda,T}}_t)\Big); \quad \text{where } \nu^\lambda(a) = \partial_a\, S^\lambda_{\max}(a) \quad \text{and} \quad t = 0, ..., T-1. \tag{4.4}$$

Here (M^U_t, D^U_t) corresponds to the solution of the dynamics (2.6) corresponding to the randomised policy U.

We now show that $U^{\lambda,T}$ results in a value which is close to $\sum_{t=0}^{T-1} \beta^t\big(\, S^\lambda_{\max} \circ \alpha^\lambda_{T-t}\big)(m,d)$, and this is also close to the optimal value function.

Theorem 4.5. *Suppose that (H.1), (H.3) and (H.4) hold. There exists a constant $C > 0$ such that for any $\lambda > 0$, $i \in \mathcal{A}$, $(m,d) \in \Theta \times \mathcal{D}$ and $T \in \mathbb{N}$,*

$$\Big|V^{\lambda,U^{\lambda,T}}_T(m,d) - \sum_{t=0}^{T-1} \beta^t\big(\, S^\lambda_{\max} \circ \alpha^\lambda_{T-t}\big)(m,d)\Big| \le C\big(P_{\lambda,d}(2,3) + P_{\lambda,d}(3,4)\big), \quad \textit{and} \tag{4.5}$$

$$\Big|V^{\lambda}_T(m,d) - \sum_{t=0}^{T-1} \beta^t\big(\, S^\lambda_{\max} \circ \alpha^\lambda_{T-t}\big)(m,d)\Big| \le C\big(P_{\lambda,d}(2,3) + P_{\lambda,d}(3,4)\big), \tag{4.6}$$

where $U^{\lambda,T}$ is defined in (4.4) and $V^{\lambda,U}_T$ and V^λ_T are defined in (4.2).

Proof. We will prove upper-bounds in (4.5) and (4.6) by induction. For notational simplicity, we write $Q_{\lambda,d} = (1-\beta)^{-2} C\big(P_{\lambda,d}(2,3) + P_{\lambda,d}(3,4)\big)$, where $C \ge 0$ is the constant given in Lemma 4.4 (which is uniform over $T \in \mathbb{N}$).

We begin with the base case for induction, by fixing $T = 1$. It follows from (H.4) via (2.10) that $V^\lambda_1 = S^\lambda_{\max} \circ \alpha^\lambda_1$, so no error is introduced at this stage.

Now assume for induction that the required inequality holds at $T-1$. Define

$$\Delta V^{\lambda,U^{\lambda,T}}_T := V^{\lambda,U^{\lambda,T}}_T - \sum_{t=0}^{T-2} \beta^t\big(\, S^\lambda_{\max} \circ \alpha^\lambda_{T-t}\big)$$

and observe that

$$\begin{aligned}
&\beta\mathbb{E}\big[V^{\lambda,U^{\lambda,T-1}}_{T-1}\big(\Phi(m,d,i,\boldsymbol{\xi}_1)\big)\big] \\
&= \beta\mathbb{E}\big[\Delta V^{\lambda,U^{\lambda,T-1}}_{T-1}\big(\Phi(m,d,i,\boldsymbol{\xi}_1)\big)\big] + \sum_{t=1}^{T-1} \beta^t \mathbb{E}\big[\big(\, S^\lambda_{\max} \circ \alpha^\lambda_{T-1-t}\big)\big(\Phi(m,d,i,\boldsymbol{\xi}_1)\big)\big] \qquad (4.7)\\
&= \beta\mathbb{E}\big[\Delta V^{\lambda,U^{\lambda,T-1}}_{T-1}\big(\Phi(m,d,i,\boldsymbol{\xi}_1)\big)\big] + \sum_{t=1}^{T-1} \beta^t \Big(\big(\, S^\lambda_{\max} \circ \alpha^\lambda_{T-1-t}\big)(m,d) + L^\lambda_i(m,d) + \Delta^i_t(m,d)\Big).
\end{aligned}$$

where

$$\Delta^i_t(m,d) := \mathbb{E}\big(\, S^\lambda_{\max} \circ \alpha^\lambda_t\big)\big(\Phi(m,d,i,\boldsymbol{\xi}_1)\big) - \big(\, S^\lambda_{\max} \circ \alpha^\lambda_t\big)(m,d) - L^\lambda_i(m,d)$$

satisfies $|\Delta^i_t(m,d)| \le (1-\beta)^2 Q_{\lambda,d}$, by Lemma 4.4. Hence, $\sum_{t=1}^{T-1} \beta^t |\Delta^i_t(m,d)| \le (1-\beta) Q_{\lambda,d}$.

Now, define

$$R^i_T(m,d) := \beta\mathbb{E}\big[\Delta V^{\lambda,U^{\lambda,T-1}}_{T-1}\big(\Phi(m,d,i,\boldsymbol{\xi}_1)\big)\big] + \sum_{t=0}^{T-2} \beta^t \Delta^i_t(m,d).$$

By (H.1)(i) and our inductive hypothesis, $\mathbb{E}\big|\Delta V_{T-1}^{\lambda,U^{\lambda,T-1}}\big(\Phi(m,d,i,\boldsymbol{\xi}_1)\big)\big| \le Q_{\lambda,d+b^i(d)} \le Q_{\lambda,d}$. Hence,

$$|R_T^i(m,d)| \le \beta Q_{\lambda,d} + (1-\beta)Q_{\lambda,d} = Q_{\lambda,d}.$$

Using dynamic programming, we can express $V_T^{\lambda,U^{\lambda,T}}$ in terms of $V_{T-1}^{\lambda,U^{\lambda,T-1}}$, that is,

$$V_T^{\lambda,U^{\lambda,T}}(m,d) = \sum_{i=1}^K U_{i,1}^{\lambda,T}\Big(f_i(m,d) + \beta\mathbb{E}\big[V_{T-1}^{\lambda,U^{\lambda,T-1}}\big(\Phi(m,d,i,\boldsymbol{\xi}_1)\big)\big]\Big) + \lambda\mathcal{H}(U_1^{\lambda,T}).$$

Applying (4.7), we see that

$$\begin{aligned} V_T^{\lambda,U^{\lambda,T}}(m,d) &= \bigg(\sum_{i=1}^K U_{i,1}^{\lambda,T}\Big(\alpha_{T,i}^\lambda(m,d) + R_T^i(m,d)\Big) + \lambda\mathcal{H}(U_1^{\lambda,T})\bigg) + \beta\sum_{t=0}^{T-2}\beta^t\big(S_{\max}^\lambda \circ \alpha_{T-1-t}^\lambda\big)(m,d) \\ &= \bigg(\sum_{i=1}^K U_{i,1}^{\lambda,T} R_T^i(m,d)\bigg) + \big(S_{\max}^\lambda \circ \alpha_T^\lambda\big)(m,d) + \beta\sum_{t=0}^{T-2}\beta^t\big(S_{\max}^\lambda \circ \alpha_{T-1-t}^\lambda\big)(m,d) \end{aligned}$$

where the first equality follows from the fact that $\alpha_{T,i}^\lambda(m,d) = f_i(m,d) + \beta\big(\sum_{t=1}^{T-1}\beta^t\big)L_i^\lambda(m,d)$ and the second equality holds from (H.4) via (2.10). Since $|R_T^i(m,d)| \le Q_{\lambda,d}$ for all $i \in \mathcal{A}$, we know $\big|\sum_{i=1}^K U_{i,1}^{\lambda,T} R_T^i(m,d)\big| \le Q_{\lambda,d}$. Rearrangement completes the proof of (4.5) by induction.

Similarly, to prove (4.6), we apply the dynamic programming principle to observe that

$$V_T^\lambda(m,d) = \sup_{u\in\Delta^K}\bigg\{\sum_{i=1}^K u_i\Big(f_i(m,d) + \beta\mathbb{E}\big[V_{T-1}^\lambda\big(\Phi(m,d,i,\boldsymbol{\xi}_1)\big)\big]\Big) + \lambda\mathcal{H}(u)\bigg\}. \tag{4.8}$$

By defining $\tilde{R}_T^i(m,d) := \beta\mathbb{E}\big[\Delta V_{T-1}^\lambda\big(\Phi(m,d,i,\boldsymbol{\xi}_1)\big)\big] + \sum_{t=0}^{T-2}\beta^t\Delta_t^i(m,d)$, we can obtain the same equality as in (4.7) with $V^{\lambda,U^{\lambda,T}}$ replaced by V^λ and R_T replaced by $\tilde{R}_T$ with $|\tilde{R}_T^i(m,d)| \le Q_{\lambda,d}$. Therefore, we can see that

$$\begin{aligned} V_T^\lambda(m,d) &= \sup_{u\in\Delta^K}\bigg\{\sum_{i=1}^K u_i\Big(\alpha_{T,i}^\lambda(m,d) + R_T^i(m,d)\Big) + \lambda\mathcal{H}(u)\bigg\} + \sum_{t=1}^{T-1}\beta^t\big(S_{\max}^\lambda \circ \alpha_{T-1-t}^\lambda\big)(m,d) \\ &\le \sup_{u\in\Delta^K}\bigg\{\sum_{i=1}^K u_i\alpha_{T,i}^\lambda(m,d) + \lambda\mathcal{H}(u)\bigg\} + \sum_{t=1}^{T-1}\beta^t\big(S_{\max}^\lambda \circ \alpha_{T-t}^\lambda\big)(m,d) + \max_{i\in\mathcal{A}}\big(R_T^i(m,d)\big) \\ &= \big(S_{\max}^\lambda \circ \alpha_T^\lambda\big)(m,d) + \sum_{t=1}^{T-1}\beta^t\big(S_{\max}^\lambda \circ \alpha_{T-t}^\lambda\big)(m,d) + \max_{i\in\mathcal{A}} R_T^i(m,d) \end{aligned}$$

Since $\max_{i\in\mathcal{A}}|R_T^i(m,d)| \le Q_{\lambda,d}$ for all $i \in \mathcal{A}$, we obtain (4.6) by induction. □

4.3 Analysis over an infinite horizon

We have seen, in Theorem 4.5, an approximation for the finite-horizon control problem in the presence of learning. We observe that the error of our approximation is uniform in T, which suggests that the dependence on the horizon is limited. In this section, we consider an approximation for the infinite-horizon control problem, which has the advantage that the solution is time-homogeneous. We note that the definition of the value in (4.2) works equally well when we take $T = \infty$, so we do not need to define additional notation.

Theorem 4.6. *Suppose that (H.1), (H.3) and (H.4) hold. For any $(m,d) \in \Theta \times \mathcal{D}$ and $\lambda > 0$, $V_T^\lambda(m,d) \to V_\infty^\lambda(m,d)$ as $T \to \infty$ and*

$$\sum_{t=0}^{T-1} \beta^t \big(S^\lambda_{\max} \circ \alpha^\lambda_{T-t}\big)(m,d) \to (1-\beta)^{-1} \big(S^\lambda_{\max} \circ \alpha^\lambda\big)(m,d)$$

where α_t^λ is defined in (4.3) and α^λ is defined in (2.16).

In particular, taking $T \to \infty$ in Theorem 4.5, we can find a constant $C \geq 0$ such that for any $(m,d) \in \Theta \times \mathcal{D}$ and $\lambda > 0$

$$\big| V_\infty^\lambda(m,d) - (1-\beta)^{-1} \big(S^\lambda_{\max} \circ \alpha^\lambda\big)(m,d) \big| \leq C\big(P_{\lambda,d}(2,3) + P_{\lambda,d}(3,4)\big).$$

Proof. By (H.1) together with the Cauchy–Schwarz inequality, we can show that for any $U \in \mathcal{U}$, we have $\mathbb{E}_{m,d}|M_{t+1}^U - M_t^U| \leq C\|D_t^U\| \leq C\|d\|$. In particular, $\mathbb{E}_{m,d}|M_t^U| \leq |m| + CT\|d\|$. By (H.3), there exists a constant $C \geq 0$, such that

$$\begin{aligned} \sup_{U\in\mathcal{U}} \sum_{t=T}^\infty \beta^t \mathbb{E}\bigg| \sum_{i=1}^K f_i\big(M_t^U, D_t^U\big) U_{i,t+1} \bigg| &\leq C \sup_{U\in\mathcal{U}} \sum_{t=T}^\infty \beta^t \mathbb{E}\big[|M_t^U| + \|d\| + 1\big] \\ &\leq C \sum_{t=T}^\infty \beta^t \big((t+1)\|d\| + |m| + 1\big) \to 0 \qquad \text{as} \quad T \to \infty. \end{aligned} \tag{4.9}$$

Moreover, by (H.4) and Theorem 4.2, $\mathcal{H}$ is bounded, so $\sup_{U\in\mathcal{U}} \sum_{t=T}^\infty \beta^t \mathbb{E}_{m,d}|\mathcal{H}(U_{t+1})| \to 0$. Combining this with (4.9), we obtain

$$\big| V_T^\lambda(m,d) - V_\infty^\lambda(m,d)\big| \leq \sup_{U\in\mathcal{U}} \bigg| \mathbb{E}_{m,d}\bigg[\sum_{t=T}^\infty \beta^t \Big(\sum_{i=1}^K f_i\big(M_t^U, D_t^U\big) U_{i,t+1} + \lambda \mathcal{H}(U_{t+1})\Big)\bigg]\bigg| \to 0 \text{ as } T \to \infty.$$

This proves the first stated limit. The second limit follows from the Tauberian theorem (Theorem A.1) applied to the sequence $a_t := (S^\lambda_{\max} \circ \alpha_t)(m,d) \to (S^\lambda_{\max} \circ \alpha)(m,d)$ as $t \to \infty$. The final bound follows from Theorem 4.5 and these limits. □

As discussed in (2.19), we are interested in the approximate optimal strategy given by the feedback strategy

$$U_{t+1}^\lambda = \nu^\lambda\big(\alpha^\lambda(M_t^{U^\lambda}, D_t^{U^\lambda})\big) \qquad \text{where} \quad \nu^\lambda(a) = \partial_a S^\lambda_{\max}(a) \quad \text{and} \quad t = 0,1,2,... \tag{4.10}$$

where (M_t^U, D_t^U) corresponds to the state variables following (2.6) under the randomised policy U.

In (4.4), we gave a finite horizon version of this strategy and showed in Theorem 4.5 that the corresponding value of this strategy and the optimal strategy can be approximated by the same function. The following results shows the convergence of the finite horizon value function (4.4) to the infinite horizon value function for the strategy (4.10).

Theorem 4.7. *Suppose that (H.1), (H.3) and (H.4) hold. Let $U^{\lambda,T}$, U^λ be the policies corresponding to the finite and infinite horizon approximations (4.4), (4.10) respectively. Then, with $V_T^{\lambda,U}$, $V_\infty^{\lambda,U}$ as in (4.2), for any $(m,d) \in \Theta \times \mathcal{D}$ and $\lambda > 0$,*

$$V_T^{\lambda,U^{\lambda,T}}(m,d) \to V_\infty^{\lambda,U^\lambda}(m,d) \text{ as } T \to \infty.$$

In particular, taking $T \to \infty$ in Theorem 4.5 and combining with Theorem 4.6, we can find a constant $C \geq 0$ such that for any $(m,d) \in \Theta \times \mathcal{D}$ and $\lambda > 0$

$$V_\infty^\lambda(m,d) - V_\infty^{\lambda,U^\lambda}(m,d) \leq C\big(P_{\lambda,d}(2,3) + P_{\lambda,d}(3,4)\big).$$

Proof. By dominated convergence theorem, we can write

$$V_\infty^\lambda(m,d) = \sum_{t=0}^{\infty} \beta^t g^*(t) := \sum_{t=0}^{\infty} \beta^t \mathbb{E}_{m,d}\Big[\Big(\sum_{i=1}^{K} f_i\big(M_t^{U^\lambda}, D_t^{U^\lambda}\big)U_{i,t+1}^\lambda\Big) + \lambda\mathcal{H}(U_{t+1}^\lambda)\Big]$$
$$V_T^{\lambda,U^{\lambda,T}}(m,d) = \sum_{t=0}^{\infty} \beta^t g(t,T) := \sum_{t=0}^{T-1} \beta^t \mathbb{E}_{m,d}\Big[\Big(\sum_{i=1}^{K} f_i\big(M_t^{U^{\lambda,T}}, D_t^{U^{\lambda,T}}\big)U_{i,t+1}^{\lambda,T}\Big) + \lambda\mathcal{H}(U_{t+1}^{\lambda,T})\Big].$$

It follows from (H.1) that there exists $C \geq 0$ such that $|g(t,T)| \leq C\|d\|t$ for all $t, T \in \mathbb{N}$. Hence, it suffices to show that for any fixed $t \in \mathbb{N}$, $g(t,T) \to g^*(t)$. The required result then follows from a version of the Tauberian theorem (Theorem A.2).

Fix $t \in \mathbb{N}$ and the initial state of the underlying state $(m,d) \in \Theta \times \mathcal{D}$. Observe that by (H.1), (H.3) and (H.4), there exist constants $C, C' \geq 0$ such that for any $s = 0, 1, ..., t-1$,

$$\begin{aligned}
&\Big|\nu^\lambda\big(\alpha^\lambda(M_s^{U^\lambda}, D_s^{U^\lambda})\big) - \nu^\lambda\big(\alpha_{T-s}^\lambda(M_s^{U^\lambda}, D_s^{U^\lambda})\big)\Big| \\
&\leq C'\beta\Big(\sum_{r=T-s}^{\infty} \beta^r\Big)\big|L^\lambda(M_s^{U^\lambda}, D_s^{U^\lambda})\big| \\
&\leq C\beta^{T-s+1}(1-\beta)^{-1}\|D_s^{U^\lambda}\|^2 \\
&\leq C\beta^{T-t+1}(1-\beta)^{-1}\|d\|^2.
\end{aligned} \tag{4.11}$$

Fix $\epsilon \in (0,1)$ and choose T_0 such that for all $T \geq T_0$, $C\beta^{T-t+1}(1-\beta)^{-1}\|d\|^2 < \epsilon/K$. Recall the expression A_s^U for the action corresponding to the control U as discussed in section 2.2. By (4.11), for $T \geq T_0$,

$$\mathbb{P}(A_s^{U^\lambda} \neq A_s^{U^{\lambda,T}} | A_r^{U^\lambda} = A_r^{U^{\lambda,T}} \ \forall r = 1, 2, ..., s-1) \leq \epsilon.$$

Let $E_T := \{A_s^{U^\lambda} = A_s^{U^{\lambda,T}} \ \forall s = 1, ..., t\} \supseteq \big\{(M_t^{U^\lambda}, D_t^{U^\lambda}) = (M_t^{U^{\lambda,T}}, D_t^{U^{\lambda,T}}), U_{t+1}^\lambda = U_{t+1}^{\lambda,T}\big\}$. We then see that

$$\mathbb{P}(E_T) \geq (1-\epsilon)^t \geq 1 - t\epsilon, \quad \text{i.e. } \mathbb{P}(E_T^c) \leq t\epsilon.$$

By (H.1) and (H.3), we can find a constant $C \geq 0$ such that for any $U \in \mathcal{U}$, we know $\mathbb{E}_{m,d}|f\big(M_t^U, D_t^U\big)|^2 \leq Ct(|m|^2 + \|d\|^2 + 1)$. Moreover, by (H.4), we can assume (wlog) that $\sup_{u \in \Delta^K} \mathcal{H}(u) \leq C$. By decomposing using the event E_T and applying these bounds together with the Cauchy–Schwartz inequality, we obtain

$$|g(t,T) - g^*(t)| \leq 2 \sup_{U \in \mathcal{U}} \mathbb{E}\Big[\big(|f(M_t^U, D_t^U)| + \lambda\mathcal{H}(U_{t+1})\big)\mathbb{I}_{E_T^c}\Big] \leq 4Ct(|m|^2 + \|d\|^2 + 2)(t\epsilon).$$

As ϵ was arbitrary (provided T_0 is sufficiently large), we conclude that $g(t,T) \to g^*(t)$ as $T \to \infty$, as required. □

In earlier sections, we derived an approximate value function and corresponding feedback control (4.10) for the regularised control problem with regularisation parameter λ. However, our ultimate objective is to optimise the non-regularised version of the problem (i.e. (2.7)). By Theorem (4.2) and (H.4), our regularisation may possibly introduce an error of $\mathcal{O}(\lambda)$. We control the corresponding error in the following theorem.

Theorem 4.8. *Suppose that (H.1), (H.3) and (H.4) hold. For any $T \leq \infty$, and any consistent $\boldsymbol{\lambda}$ with order $\kappa \in [0,2]$ (Definition 2.2), there exists a constant $C \geq 0$ such that for any $(m,d) \in \Theta \times \mathcal{D}$,*

$$|V_T(m,d) - V_T^{U^{\boldsymbol{\lambda}}}(m,d)| \leq C(\|d\|^{3-2\kappa} + \|d\|^{4-3\kappa} + \|d\|^{\kappa} + \|d\|^4).$$

Proof. Let $C > 0$ be a constant which can vary from line to line. As $\boldsymbol{\lambda}$ is consistent with order κ, we know that $P_{\boldsymbol{\lambda}(m,d),d}(a,b) \leq (1 + \bar{c}\|d\|^{-\kappa a})\|d\|^b$. In particular, for $\kappa \in [0,2]$, we know $P_{\boldsymbol{\lambda}(m,d),d}(2,3) = (1+\bar{c}\|d\|^{-2\kappa})\|d\|^3 \leq C(\|d\|^3 + \|d\|^{3-2\kappa})$ and $P_{\boldsymbol{\lambda}(m,d),d}(3,4) \leq C(\|d\|^4 + \|d\|^{4-3\kappa})$. Substituting these into the bounds in Theorem 4.5 or 4.6, we see that the regularized problems satisfy

$$|V_T^{\boldsymbol{\lambda}(m,d)}(m,d) - V_T^{\boldsymbol{\lambda}(m,d),U^{\boldsymbol{\lambda}}}(m,d)| \leq C(\|d\|^3 + \|d\|^{3-2\kappa} + \|d\|^4 + \|d\|^{4-3\kappa}).$$

We then recall that our smooth entropy is bounded, so the differences

$$|V_T^{\boldsymbol{\lambda}(m,d)}(m,d) - V_T(m,d)| \leq C|\boldsymbol{\lambda}(m,d)| \leq C\|d\|^{\kappa},$$

and similarly for $|V_T^{\boldsymbol{\lambda}(m,d),U^{\boldsymbol{\lambda}}}(m,d) - V_T^{U^{\boldsymbol{\lambda}}}(m,d)|$. Combining these bounds (and omitting the dominated $\|d\|^3$ term) yields the result. □

Remark 4.2. For $\|d\| \to 0$, it is clear that the optimal order in the previous theorem is given by taking $\kappa = 1$.

Remark 4.3. Through our discussion, we prove all of our result using (H.1) which is inspired by the multi-armed bandit problem (Section 2.1) for learning. The nature of the learning results in the transient state. This is reflected in (H.1)(i) where our precision $\|D_t^U\|$ in non-increasing over $t \in \mathbb{N}$. In general, we can extend our approximation result to the case when our precision is recurrent and take value in a (small) compact set $\mathcal{D}$, as is typical when our learning is modelled using a Kalman-like filtering process. All of our earlier analysis follows by using $\|d\| \leq h := \sup_{d\in\mathcal{D}} \|d\|$ and $\|d + b_i(m,d)\| \leq h$ for all $i \in \mathcal{A}$. In particular, we have the following result.

Theorem 4.9. *Suppose that (H.2), (H.3) and (H.4) hold. Then Theorem 4.6, Theorem 4.7 and Theorem 4.8 holds with all $\|d\|$ in the upper-bounds replaced by $\sup_{d\in\mathcal{D}} \|d\|$.*

4.4 Complete Learning

We have discussed in (3.2) that many bandit algorithms e.g. Gittins' index, Knowledge Gradient and Thompson Sampling suffer from incomplete learning. In this section, we will show that the ARC algorithm overcomes this limitation in the sense that $\|D_t^{U^\lambda}\| \to 0$ as $t \to \infty$. We need the following Assumption, to ensure that our dynamics are non-degenerate.

H.5. *If every arm is chosen infinitely often, then the process $\|D_t^{U^\lambda}\| \to 0$. More precisely, for any $U \in \mathcal{U}$, we have the inclusion of events*

$$\{\|D_t^U\| \to 0\} \supseteq \Big\{ \sum_{t=1}^{\infty} \mathbb{I}(A_t^U = i) = \infty \ \forall i \in \mathcal{A} \Big\}$$

where A_t^U is the (random) action corresponding to the control U_t.

Theorem 4.10. *Suppose that (H.1), (H.3), (H.4), and (H.5) hold. Suppose further that f is uniformly bounded and the derivative of the smooth max approximator S does not vanish on compact sets[3], that is: for any compact set $K \subseteq \mathbb{R}^K$, there exists a non-empty open ball, $B(r)$, such that $B(r) \cap \partial_a S(K) = \emptyset$. Then for any consistent $\boldsymbol{\lambda}$ with $\kappa \in [0,2]$, we know $\|D_t^{U^{\boldsymbol{\lambda}}}\| \to 0$ a.s.*

[3] One can check that the Shannon Entropy (Remark 2.2) satisfies this property.

Proof. Without loss of generality, fix a bound on the initial value $\|D_0^U\|$. By (H.1) and (H.3), and the consistency of λ (Definition 2.2), there exist constants $C, \bar{C} \geq 0$ such that

$$|L^{\boldsymbol{\lambda}(m,d)}(m,d)| \leq C\big(1 + \boldsymbol{\lambda}(m,d)^{-1}\big)\|d\|^2 \leq C(\|d\|^2 + \|d\|^{2-\kappa}\underline{c}^{-1}) \leq \bar{C}.$$

As f is uniformly bounded and β is fixed, there also exists a constant $C \geq 0$ such that $|\alpha^{\boldsymbol{\lambda}(m,d)}(m,d)| \leq C$ for all $(m,d) \in \Theta \times \mathcal{D}$.

Fix $i \in \mathcal{A}$ and $\epsilon > 0$, and consider the events $E_\epsilon := \{\|D_t^{U^{\boldsymbol{\lambda}}}\| > \epsilon \quad \forall t \in \mathbb{N}\}$, $F_i := \{A_t^{U^{\boldsymbol{\lambda}}} = i \text{ for finitely many } t \in \mathbb{N}\}$ and $G_{i,t} := \{A_t^{U^{\boldsymbol{\lambda}}} = i, \|D_t^{U^{\boldsymbol{\lambda}}}\| > \epsilon \quad \forall s = 0, 1, ..., t-1\}$. By (H.5), $E_\epsilon \subseteq \cup_{i \in \mathcal{A}} F_i$. Moreover, we can see that $F_i \cap E_\epsilon \subseteq \{G_{i,t}^c \text{ eventually}\}$.

For $\|d\| > \epsilon$, we know that $\alpha^{\boldsymbol{\lambda}(m,d)}(m,d)/\boldsymbol{\lambda}(m,d)$ takes value in a compact set. Hence, there exists $r \in (0,1)$ such that $\nu_i^{\boldsymbol{\lambda}(m,d)}\big(\alpha^{\boldsymbol{\lambda}(m,d)}(m,d)\big) = (\partial_a S)_i\big(\alpha^{\boldsymbol{\lambda}(m,d)}(m,d)/\boldsymbol{\lambda}(m,d)\big) > r$. In particular, for any $n \in \mathbb{N}$, we have $\mathbb{P}(G_{i,t} | \bigcap_{s=n}^{t-1} G_{i,s}^c) > r$. Therefore,

$$\mathbb{P}\Big(\bigcap_{t=n}^{N} G_{i,t}^c\Big) = \prod_{t=n}^{N} \mathbb{P}\Big(G_{i,t}^c \Big| \bigcap_{s=n}^{t-1} G_{i,s}^c\Big) \leq (1-r)^{N-n+1} \to 0 \text{ as } N \to \infty.$$

Hence, $\mathbb{P}(F_i \cap E_\epsilon) \leq \mathbb{P}(G_{i,t}^c \text{ eventually}) = \mathbb{P}(\bigcup_{n=1}^{\infty} \bigcap_{t=n}^{\infty} G_{i,t}^c) \leq 0$. This implies that $\mathbb{P}(\bigcup_{i \in \mathcal{A}} F_i \cap E_\epsilon) = 0$ and thus $\mathbb{P}(E_\epsilon) = 0$. The result follows by considering $\bigcup_{n=1}^{\infty} E_{1/n}$. □

Remark 4.4. Variations of this result are possible. If f is not assumed bounded, then the proof still largely works, provided we can show that m does not go to infinity (almost surely). This can often be shown using additional assumptions on the dynamics of the problem.

Remark 4.5. A possible criticism of the ARC approach is that the approximation is only valid if D_t is small. This result shows that, at least in theory, this approximation will always be asymptotically true, and hence we can guarantee that the ARC algorithm will converge to an optimal strategy. In practice, we will see that the precise way in which we measure the convergence of D can have an impact on the performance of the algorithm, particularly in high-dimensional settings.

5 Numerical Experiments

We run numerical experiments to illustrate the performance of the ARC algorithm, in comparison with the true optimal solution, and with existing state-of-the-art methods.

5.1 Comparison to the optimal solution for $1\frac{1}{2}$ bandit

We first consider the '$1\frac{1}{2}$ bandit', where one arm gives a deterministic reward. This case has the advantage that the value function can be determined using a straightforward Monte Carlo method, allowing us to compare our approximate value function and its corresponding control to the exact value function.

Suppose that our bandit has two arms. The first arm gives the reward $Y \sim N(\theta, 1)$, for an unknown parameter $\theta \in \mathbb{R}$, whereas the second arm always gives a reward 1. We observe the reward of the first arm only when the first arm is chosen. Formulating this as a relaxed control problem (2.12) with dynamics as in (2.2) gives

$$V_\infty^\lambda(m,d) = \sup_{U \in \mathcal{U}} \mathbb{E}_{m,d}\Big[\sum_{t=0}^{\infty} \beta^t \Big(\big(U_{1,t+1} M_t^U + U_{2,t+1}\big) + \lambda \mathcal{H}(U_{t+1})\Big)\Big] \quad ; \quad \mathcal{H}(u) = -\sum_{i=1}^{2} u_i \ln u_i.$$

In this case, we have $\Theta = \mathbb{R}$ and $\mathcal{D} = (0, \infty)$. The transition (2.6) of the problem is given by

$$\Phi(m, d, i, \xi) := \begin{cases} (m, d) + \big(d(1+d)^{-1/2} z(\xi), \quad -d^2(1+d)^{-1}\big) & ; \ i = 1 \\ (m, d) & ; \ i = 2 \end{cases}$$

with innovations $z(\boldsymbol{\xi}_t) \sim_{IID} N(0,1)$. We may solve the above problem explicitly using classical Monte Carlo simulation. In particular, we start our iteration with $\tilde{V}_0^\lambda(m,d) = 0$ and iteratively compute on the grid (m, d),

$$\tilde{V}_{n+1}^\lambda(m,d) = \sup_{u \in \Delta^2} \left\{ u_1\Big(m + \frac{\beta}{N} \sum_{i=1}^N \tilde{V}_n^\lambda\big(\Phi(m,d,1,\boldsymbol{\xi}_{i,n})\big)\Big) + u_2\Big(1 + \beta \tilde{V}_n^\lambda(m,d)\Big) + \lambda \mathcal{H}(u) \right\}$$

where N is the number of Monte Carlo simulations and $(\boldsymbol{\xi}_{i,n})$ are such that $z(\boldsymbol{\xi}_{i,n}) \sim_{IID} N(0,1)$. We then interpolate $\tilde{V}_{n+1}^\lambda$ and repeat the procedure until $\tilde{V}_n^\lambda$ converges to $\tilde{V}^\lambda$. The corresponding optimal (feedback) probability to choose the first arm is then given by

$$\tilde{p}^{\lambda,*}(m,d) \approx \frac{\exp\Big(\frac{1}{\lambda}\big(m + \frac{\beta}{N}\sum_{i=1}^N \tilde{V}^\lambda(\Phi(m,d,1,\boldsymbol{\xi}_i))\big)\Big)}{\exp\Big(\frac{1}{\lambda}\big(m + \frac{\beta}{N}\sum_{i=1}^N \tilde{V}^\lambda(\Phi(m,d,1,\boldsymbol{\xi}_i))\big)\Big) + \exp\Big(1 + \beta \tilde{V}^\lambda(m,d)\Big)}$$

where $z(\boldsymbol{\xi}_i) \sim_{IID} N(0,1)$. On the other hand, our ARC approximation gives

$$V_\infty^{\lambda,ARC}(m,d) = (1-\beta)^{-1}\lambda \log\Big(\exp\Big(\frac{\alpha^\lambda(m,d)}{\lambda}\Big) + \exp\Big(\frac{1}{\lambda}\Big)\Big), \quad p^{\lambda,ARC} = \frac{\exp\Big(\frac{\alpha^\lambda(m,d)}{\lambda}\Big)}{\exp\Big(\frac{\alpha^\lambda(m,d)}{\lambda}\Big) + \exp\Big(\frac{1}{\lambda}\Big)}$$

where $\alpha^\lambda(m,d) = \Big(m + \frac{1}{2\lambda}\beta(1-\beta)^{-1}\nu_1^\lambda(m)(1-\nu_1^\lambda(m))d^2(1+d)^{-1}, 1\Big)$ and $\nu_1^\lambda(m) = \frac{\exp(\frac{1}{\lambda}m)}{\exp(\frac{1}{\lambda}m)+\exp(\frac{1}{\lambda})}$.

We now compare $\tilde{V}^\lambda, \tilde{p}^{\lambda,*}$ with $V_\infty^{\lambda,ARC}, p^{\lambda,ARC}$ when $\lambda = 0.1$ and $\beta = 0.99$. In the numerical experiment, we use $N = 1000$ and consider $m \in [0,2]$ and $d \in [1/100, 1/20]$. We see that the ARC approximation gives a reasonable estimate of the value function for the regularised problem (Figure 1. Observe that the error of the approximation is less than 1% in the worst case (when m is zero and $1/d$ is small), and is very small when $1/d$ is large. Moving to the action probabilities (Figure 2), we observe that the ARC algorithm introduces a smoothing of the boundary where it is optimal to switch arms, however captures the other principal features of the optimal action correctly.

5.2 General bandit results

We now focus on a variety of bandit problems, where our method can be used. In earlier sections we considered the value function (2.7) as our objective to optimise under a Bayesian formulation. While this quantity can be used to compare performance of the algorithm, an variation which is more commonly used in the multi-armed bandit literature [25, 11, 4, 15, 17] is the regret

$$R(A, T, \boldsymbol{\theta}) = \sum_{t=1}^T \Big(\max_{i \in \mathcal{A}} \mathbb{E}\big[R^{(i)}(Y^{(i)}) \big| \boldsymbol{\theta}\big] - \mathbb{E}\big[R^{(A_t)}(Y^{(A_t)}) \big| \boldsymbol{\theta}\big] \Big).$$

This has the advantage of removing the first-order effect of the optimal strategy, allowing more direct comparisons.

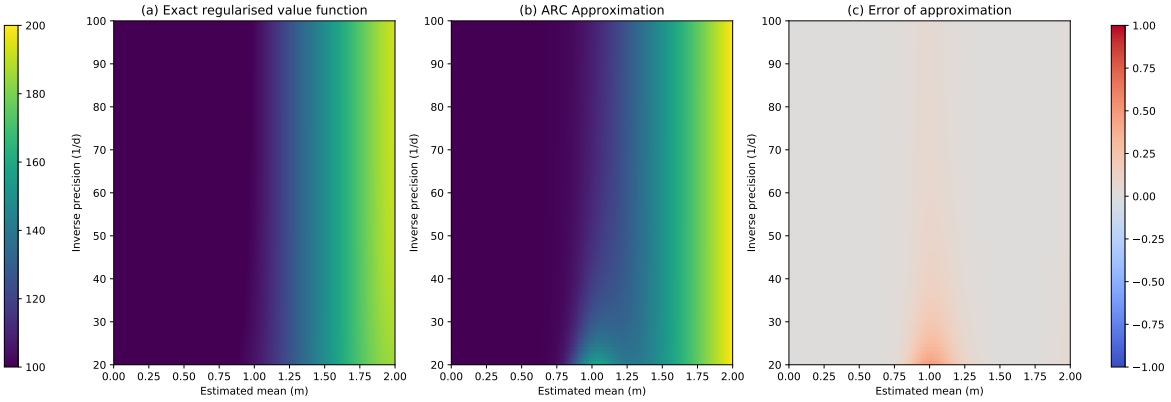

Figure 1: (a) $\tilde{V}^{\lambda}$ (b) $V_{\infty}^{\lambda,ARC}$ (c) $(V_{\infty}^{\lambda,ARC} - \tilde{V}^{\lambda})/\tilde{V}^{\lambda}$

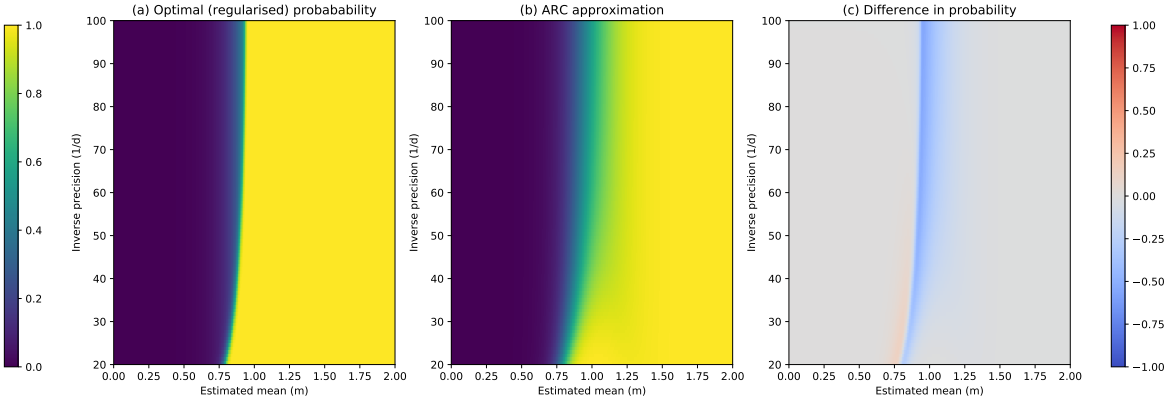

Figure 2: (a) $\tilde{p}^{\lambda,*}$ (b) $p_{\infty}^{\lambda,ARC}$ (c) $p_{\infty}^{\lambda,ARC} - \tilde{p}^{\lambda,*}$

We will compare the regret of the ARC algorithm with other approaches, as described in Section 3.1, under 3 environments; a classical bandit, a bandit with an informative arm and a linear bandit. In each of these environments, we consider decisions on the 50-armed bandit with horizon $T = 2 \times 10^3$ steps over 10^3 simulations. In the nth simulation, we sample the true parameter $\boldsymbol{\theta}^{(n)} \sim N(\mathbf{1}_{50}, I_{50})$ and then simulate the interaction between each of the algorithms and the environment with the parameter estimate $\boldsymbol{\theta}^{(n)}$ starting from initial belief $(m_0, d_0) = (0, 10^3 \times I_{50})$ corresponding to a non-informative prior. We then use our simulated output to compute statistics of the map $t \mapsto R(A, t, \boldsymbol{\theta})$ for each of our considered algorithms.

Simulation Environment: We consider the following environment for our simulation. Let $\boldsymbol{\theta}$ be an unknown parameter taking values in $\mathbb{R}^{50}$.

- *Classical bandit.* When choosing the ith option, we observe and receive the reward sampled from the distribution $N(\boldsymbol{\theta}_i, 5)$.
- *Bandit with an informative arm.* When choosing the ith option with $i \neq 1$, we observe and receive the reward sampled from the distribution $N(\boldsymbol{\theta}_i, 5)$. When the 1st option is chosen, we receive a reward sampled from $N(\boldsymbol{\theta}_1 - 1, 5)$ and in addition, we observe a sample from

$N(\boldsymbol{\theta}, 5 \times I_{50})$. In particular, playing the first arm allows us to observe rewards of other arms without choosing them, but this arm yields a smaller reward than others.

- *Linear bandit.* When choosing the ith option, we observe and receive the reward sampled from the distribution $N(b_i^\top \boldsymbol{\theta}, 5)$ where $b_i = e_i + e_{i+1}$ for $i \neq 50$ and $b_{50} = b_1 + b_{50}$. This introduces correlation between 'neighbouring' arms, affecting both the learning process and optimal strategies.

Hyper-parameter of bandit algorithms: We will consider the KG and IDS methods by introducing 100 Monte-Carlo samples to evaluate the required expectation appeared in the algorithm. We will set the parameter $\beta = 1 - 1/T = 0.9995$ for KG and ARC. The function $\boldsymbol{\lambda}(m, d)$ considered in the BE and ARC algorithms will be given in the form $\boldsymbol{\lambda}(m, d) = \rho \|d\|_{op}$ where $\|\cdot\|_{op}$ is the matrix operator norm and will take $\nu_i^{\lambda}(a) := \exp(a_i/\lambda)/\big(\sum_j \exp(a_j/\lambda)\big)$ which corresponds to the Shannon entropy as discussed in Remark 2.2. Unmentioned hyper-parameters will appear as a description of the algorithm in the regret plot.

ARC index strategy: For numerical efficiency, we also introduce a deterministic index strategy inspired by the ARC algorithm. In contrast to the ARC, instead of making decision based on the probability vector $\nu^{\boldsymbol{\lambda}(m,d)}\big(\alpha^{\boldsymbol{\lambda}(m,d)}(m, d)\big)$ with $\alpha^{\lambda}(m, d)$ given in (2.16), we simply choose the option i to maximise $\alpha_i^{\lambda}(m, d)$.

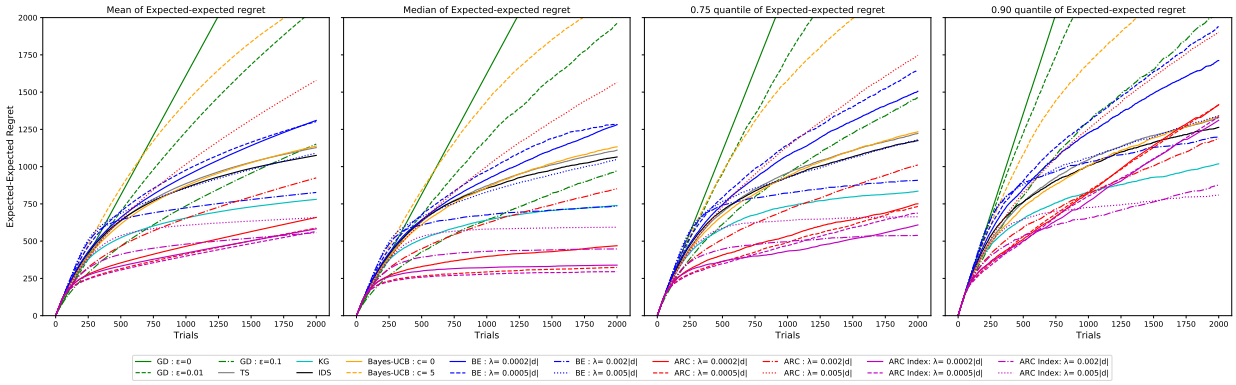

Figure 3: Regret for the classical bandit

Discussion: We can see in Figure 3, 4 and 5 that both ARC and ARC index strategies, with appropriate hyper-parameters, perform very well compared to other algorithms. We particularly see that the ARC approaches perform significantly better than other approaches in the setting for the bandit with an informative arm. This performance is as good as IDS, but requires significantly lower computational cost since every term can be evaluated explicitly. In fact, our implementation of IDS is too computationally expensive to demonstrate for the linear bandit with 50 arms and thus IDS is omitted in this case.

Even though, we find that the ARC algorithm derived in this paper performs well with appropriate hyper-parameters, it is worth commenting on the obstacles found in the derived ARC algorithm with many arms available. We see that taking $\boldsymbol{\lambda}(m, d) = \rho\|d\|$ may not allow $\|d\|$ to decay sufficiently fast (even though Theorem 4.10 ensures that this converges to 0). In particular, with many arms, we will often be able to identify some arms which are significantly worse than the

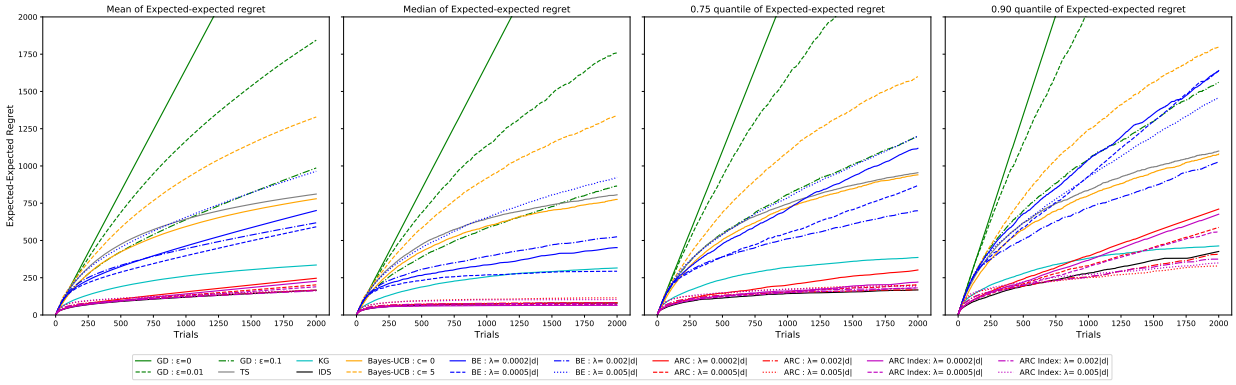

Figure 4: Regret for the bandit with an informative arm

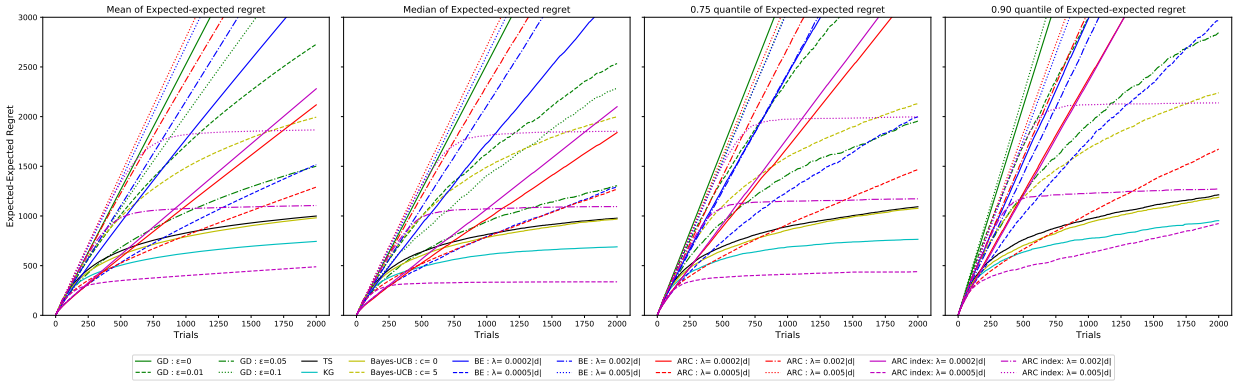

Figure 5: Regret for the linear bandit

others. These arm will be barely played, which leaves $\|d\|$ to be large for a long time. We found in the numerical simulation that following the ARC procedure, the agent identifies the reward of the very best few arms correctly (i.e. we obtain close estimates to the rewards of a few best options), but since $\|d\|$ does not decay sufficiently fast, this the ARC algorithm randomly chooses among a few best few options, even if it identifies the best option correctly (see Cesa-Bianchi et al. [5, Section 3.1] for related discussion). This explains why we observe linear trends in the regret plot for the simple ARC method.

To overcome this effect, we can consider the ARC index strategy, which does not require $\|d\|$ to decay to zero to terminate decision to a single (best) option. Here, we see in Figure 3, 4 and 5 that the gradient of the regret of the ARC index converges to zero which means that they eventually identify the best option. In general, one may also choose the function $\boldsymbol{\lambda}(m, d)$ depending on m to truncate our consideration to the best arm or one may also introduce a concatenation of function to allow the ARC to neglect the size of $\|d\|$ at later stages, as in Cesa-Bianchi et al. [5, Theorem 2]. This will naturally lead to a (small) risk of missing good arms; balancing these probabilities is significant in obtaining optimal asymptotic rates.

Conclusion: In this paper, we address a general class of learning (bandit) problems. We describe the interaction of the information between available options and study them in terms of the discrete-time diffusion process. We then use this setting to derive an approximate strategy which

takes those information interactions into consideration when making decisions. The derived strategy performs well and describes the role of the 'optimistic principle' in a more precise way than other algorithms, as it distinguishes between the benefits of learning and optimism. While we do not claim that our framework will provide the best learning strategy for general learning problems, we believe that this work provides a helpful perspective on learning with decision making.

References

[1] R. Agrawal, *Sample mean based index policies by $O(\log N)$ regret for the multi-armed bandit problem*, Advances in Applied Probability, (1995), pp. 1054–1078.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, *Finite-time analysis of the multiarmed bandit problem*, Machine Learning, (2002), pp. 235–256.

[3] M. Brezzi and T. Lai, *Optimal learning and experimentation in bandit problems*, Journal of Economic Dynamics and Control, (2002), pp. 87–108.

[4] G. Burtini, J. Loeppky, and R. Lawrence, *A Survey of Online Experiment Design with the Stochastic Multi-Armed Bandit*, arXiv:1510.00757v4, (2015).

[5] N. Cesa-Bianchi, C. Gentile, G. Lugosi, and G. Neu, *Boltzmann Exploration Done Right*, in NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.

[6] V. Chernozhukov, D. Chetverikov, and K. Kato, *Comparison and anti-concentration bounds for maxima of Gaussian random vectors*, Probability Theory and Related Fields, (2013).

[7] S. N. Cohen, *Data-driven nonlinear expectations for statistical uncertainty in decisions*, Electronic Journal of Statistics, (2016), pp. 1858–1889.

[8] S. N. Cohen and T. Treetanthiploet, *Gittins' theorem under uncertainty*, Electronic Journal of Probability, (2022), pp. 1–48.

[9] F. Coquet, Y. Hu, J. Mémin, and S. Peng, *Filtration consistent nonlinear expectations and related g-expectations*, Probability Theory and Related Fields, (2002), pp. 1–27.

[10] E. Even-Dar, S. Mannor, and Y. Mansour, *Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems*, Journal of Machine Learning Research, (2006).

[11] S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári, *Parametric bandits: the Generalized Linear case*, in NIPS'10: Proceedings of the 23rd International Conference on Neural Information Processing Systems, 2010.

[12] H. Föllmer and A. Schied, *Stochastic Finance: an introduction in discrete time*, De Gruyler, 2016.

[13] M. Frittelli and E. R. Gianin, *Putting order in risk measures*, Journal of Banking & Finance, (2002), pp. 1473–1486.

[14] J. C. Gittins and D. M. Jones, *A dynamic allocation index for the sequential design of experiments*, in Progress in Statistics, J. Gani, ed., Amsterdam: North Holland, 1974, pp. 241–266.

[15] E. Kaufmann, O. Cappé, and A. Garivier, *On Bayesian Upper Confidence Bounds for Bandit problems*, in Artificial intelligence and statistics, 2012, pp. 592–600.

[16] J. M. Keynes, *A Treatise on Probability*, Macmillan and Co., 1921. Reprint BN Publishing, 2008.

[17] J. Kirschner and A. Krause, *Information directed sampling and bandits with heteroscedastic noise*, Proceedings of Machine Learning Research, (2018), pp. 1–28.

[18] F. H. Knight, *Risk, Uncertainty and Profit*, Houghton Mifflin, 1921. reprint Dover 2006.

[19] T. Lattimore and C. Szepesvári, *Bandit Algorithms*, Cambridge University Press, 2019.

[20] C. Reisinger and Y. Zhang, *Regularity and stability of feedback relaxed control*, SIAM Journal on Control and Optimization, (2021), pp. 3118–3151.

[21] R. Rockafellar, *Convex Analysis*, Princeton University Press, 1972.

[22] P. Rusmevichientong, A. Mersereau, and J. N. Tsitsiklis, *A Structured Multiarmed Bandit Problem and the Greedy Policy*, in Proceedings of the IEEE Conference on Decision and Control, 2009.

[23] D. Russo, *A note on the equivalence of upper confidence bounds and gittins indices for patient agents*, Operations Research, (2021), pp. 273–278.

[24] D. Russo and B. V. Roy, *An information-theoretic analysis of thompson sampling*, Journal of Machine Learning Research, (2016), pp. 1–30.

[25] ——, *Learning to Optimize via Information-Directed Sampling*, Operations Research, (2017), pp. 1–23.

[26] D. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen, *A tutorial on Thompson sampling*, Foundations and Trends in Machine Learning, 11 (2018), pp. 1–96.

[27] I. O. Ryzhov, W. B. Powell, and P. I. Frazier, *The knowledge gradient algorithm for a general class of online learning problems*, Operations Research, (2012).

[28] S. P. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári., *Convergence results for single-stepon-policy reinforcement-learning algorithms*, Machine Learning, (2000).

[29] D. Šiška and Ł. Szpruch, *Gradient flows for regularized stochastic control problems*, SIAM Journal on Control and Optimization, (2024).

[30] W. R. Thompson, *on the likelihood that one unknown probability exceeds another in view of the evidence of two samples*, Biometrika, (1933), pp. 285–294.

[31] J. Vermorel and M. Mohri, *Multi-armed bandit algorithms and empirical evaluation*, in European Conference on Machine Learning., 2005.

[32] H. Wang, T. Zariphopoulou, and X. Zhou, *Reinforcement learning in continuous time and space:a stochastic control approach*, Journal of Machine Learning Research, (2020), pp. 1–34.

[33] L. Zhou, *A survey on contextual multi-armed bandits*, arXiv:1508.03326, (2016).

A Proofs of relevant results

Proof of Theorem 4.2. $(i) \Rightarrow (ii)$: Fix $i \in \mathcal{A}$. Consider $a = (N+\epsilon)e_i + \sum_{j\neq i} r_j e_j$ where $r_j \in \mathbb{R}$ for all $j \neq i$ and e_i is the ith basis vector in $\mathbb{R}^K$. By (i), $S(a) + N \geq \max(N+\epsilon, r_j) \geq N+\epsilon$. Hence, $S(a) \geq \epsilon > 0$.

As ϵ is arbitrary, it follows that $\mathbb{R}^{i-1} \times (N,\infty) \times \mathbb{R}^{K-i} \subseteq \mathcal{A}_S^c$. The result then follows by considering intersection over all i.

$(ii) \Rightarrow (iii)$**:** By Theorem 4.1, we can write $S(a) = \sup_{u\in\Delta^K} \Big(\sum_{i=1}^K u_i a_i + \mathcal{H}_{max}(u) \Big)$, where $\mathcal{H}_{max}(u) := -\sup_{a\in\mathcal{A}_S} \Big(\sum_{i=1}^K u_i a_i \Big)$ with $\mathcal{A}_S := \{a \in \mathbb{R}^K : S(a) \leq 0\}$. As $\mathcal{A}_S \subseteq (-\infty, N]^K$ and $u \in \Delta^K$, $\mathcal{H}_{max}(u) \geq -\sup_{a\in(-\infty,N]^K} \Big(\sum_{i=1}^K u_i a_i \Big) \geq -N$.

Moreover, by (4.1), we have $\sup_{u\in\Delta^K} \mathcal{H}_{max}(u) \leq S(0)$. Therefore, $\mathcal{H}_{max}$ is bounded.

$(iii) \Rightarrow (iv)$ **:** Fix $a \in \mathbb{R}^K$ and define $i^* \in \arg\max_i a_i$. Then

$$
\begin{aligned}
-\lambda \sup_{u\in\Delta^K} |\mathcal{H}(u)| \leq \lambda \mathcal{H}(e^{(i^*)}) &= \sum_{i=1}^K (e^{(i^*)})_i a_i + \lambda \mathcal{H}(e^{(i^*)}) - \max_i a_i \\
&\leq S^\lambda_{\max}(a) - \max_i a_i \leq \sup_{u\in\Delta^K} \Big(\sum_{i=1}^K u_i a_i \Big) + \lambda \sup_{u\in\Delta^K} |\mathcal{H}(u)| - \max_i a_i = \lambda \sup_{u\in\Delta^K} |\mathcal{H}(u)|.
\end{aligned}
$$

Hence, $\sup_{a\in\mathbb{R}} |\, S^\lambda_{\max}(a) - \max_i a_i| \leq \lambda \sup_{u\in\Delta^K} |\mathcal{H}(u)| \to 0$ as $\lambda \downarrow 0$.

$(iv) \Rightarrow (i)$ Find $N > 0$ such that

$$
1 \geq \sup_{a\in\mathbb{R}} |S^{1/N}_{\max}(a) - \max_i a_i| = \frac{1}{N} \sup_{a\in\mathbb{R}} |S(Na) - \max_i Na_i| = \frac{1}{N} \sup_{a\in\mathbb{R}} |S(a) - \max_i a_i|.
$$

By rearranging the inequality above, the result follows. □

Proof of Lemma 4.3. Define $(\Delta m, \Delta d) := \Phi(m, d, i, \boldsymbol{\xi}_t) - (m, d)$.

Recall expressions for $\mathcal{B}^\lambda$, $\mathcal{M}^\lambda$, Σ^λ from (2.18). Since f is 3-time differentiable with bounded derivatives (H.3) , the terms f, $\partial_m f$, $\partial^2_m f$ and $\partial_d f$ are differentiable with bounded derivative. Moreover, by (H.4), the corresponding smooth max approximator S has a bounded derivative. In particular, there exists a constant $C \geq 0$ such that $\big|\partial_a \nu^\lambda(a)\big| \leq C/\lambda$ and $\big|\partial_a \eta^\lambda(a)\big| \leq C/\lambda$. Hence, it follows from Taylor's approximation (or the mean value inequality) that

$$
\begin{aligned}
\big|\mathcal{B}^\lambda(m+\Delta m, d+\Delta d) - \mathcal{B}^\lambda(m,d)\big| &\leq C(1+\lambda^{-1})\big(|\Delta m| + \|\Delta d\|\big), & |\mathcal{B}^\lambda| &\leq C, \\
\big|\mathcal{M}^\lambda(m+\Delta m, d+\Delta d) - \mathcal{M}^\lambda(m,d)\big| &\leq C(1+\lambda^{-1})\big(|\Delta m| + \|\Delta d\|\big), & |\mathcal{M}^\lambda| &\leq C, \\
\big|\Sigma^\lambda(m+\Delta m, d+\Delta d) - \Sigma^\lambda(m,d)\big| &\leq C(1+\lambda^{-2})\big(|\Delta m| + \|\Delta d\|\big), \text{ and} & |\Sigma^\lambda| &\leq C(1+\lambda^{-1}).
\end{aligned}
$$

By similar arguments as above applying to (H.1), we show that for any $\psi \in \{b_i, \mu_i, (\sigma_i\sigma_i^\top) : i \in \mathcal{A}\}$, and $(m,d) \in \Theta \times \mathcal{D}$, $|\psi(m,d)| \le C\|d\|^2$.

$$\begin{aligned}\big|\psi(m+\Delta m, d+\Delta d) - \psi(m,d)\big| &\le \sup_{\tilde d \in [d, d+\Delta d], \tilde m \in \Theta} \Big(|\partial_d \psi(m, \tilde d)| \cdot \|\Delta d\| + |\partial_m \psi(\tilde m, d)| \cdot |\Delta m| \Big) \\ &\le C \sup_{\tilde d \in [d, d+\Delta d], \tilde m \in \Theta} \Big(\|\tilde d\| \cdot \|\Delta d\| + \|d\|^2 \cdot |\Delta m| \Big) \le C\big(\|d\| \cdot \|\Delta d\| + \|d\|^2 \cdot |\Delta m| \big)\end{aligned}$$

where $[d, d+\Delta d]$ is defined to be a rectangle on $\mathbb{R}^q$ and the final inequality follows from the convexity of the norms and (H.1)(i).

Substituting above inequalities into (2.17), we obtain

$$\big|L_i(m+\Delta m, d+\Delta d) - L_i(m,d)\big| \le C(1+\lambda^{-2})\Big(\big(|\Delta m| \cdot \|d\|^2 + \|\Delta d\| \cdot \|d\|^2 + \|d\| \cdot \|\Delta d\| \big) \Big).$$

Finally, by (H.1)$(iii)-(iv)$ and Cauchy–Schwarz inequality, $\mathbb{E}|\Delta m| \le C\|d\|$ and $|\Delta d| \le C\|d\|^2$. Substituting these bounds into the above inequality, the general result follows. □

Proof of Lemma 4.4. Let $g(m,d) = f(m,d) + c$ where $c \in \mathbb{R}$ is a given constant. By (H.3) and (H.4), $S^\lambda_{\max} \circ g$ is 3-times differentiable. Then consider the Taylor approximation

$$\begin{aligned}&\big(S^\lambda_{\max} \circ g\big)\big(\Phi(\cdot,\cdot,i,\boldsymbol{\xi}_t)\big) - \big(S^\lambda_{\max} \circ g\big) \\ &\quad = \big\langle \partial_d\big(S^\lambda_{\max} \circ g\big); \Delta d_i \big\rangle + \big\langle \partial_m\big(S^\lambda_{\max} \circ g\big); \Delta m_i \big\rangle + \frac{1}{2}\big\langle \partial_m^2\big(S^\lambda_{\max} \circ g\big); \Delta m_i \Delta m_i^\top \big\rangle + \Delta S_T^1\end{aligned} \tag{A.1}$$

where all derivatives are evaluated at (m,d).

By (H.3) and (H.4), the second and the third derivative of $\big(S^\lambda_{\max} \circ g\big)(m,d)$ are $\mathcal{O}(1+\lambda^{-2})$. Moreover, by (H.1), $\mathbb{E}|\Delta m_i| \le C\|d\|$ and $\|\Delta d_i\| \le C\|d\|^2$. Applying these bounds to the third order terms and the remaining second order term, we obtain $\mathbb{E}|\Delta S_T^1| \le C(1+\lambda^{-2})\|d\|^3$. Here, the bounded constant $C \ge 0$ is uniform over c.

Taking expectation over (A.1), we can see that

$$\begin{aligned}&\mathbb{E}\big[\big(S^\lambda_{\max} \circ g\big)\big(\Phi(\cdot,\cdot,i,\boldsymbol{\xi}_t)\big)\big] - \big(S^\lambda_{\max} \circ g\big) \\ &\quad = \big\langle \partial_d\big(S^\lambda_{\max} \circ g\big); b_i \big\rangle + \big\langle \partial_m\big(S^\lambda_{\max} \circ g\big); \mu_i \big\rangle + \frac{1}{2}\big\langle \partial_m^2\big(S^\lambda_{\max} \circ g\big); \sigma_i\sigma_i^\top \big\rangle + R_T^1\end{aligned} \tag{A.2}$$

where $|R_T^1| \le C(1+\lambda^{-2})\|d\|^3$. Now, write

$$\begin{aligned}\partial_d\big(S^\lambda_{\max} \circ g\big) &= \sum_{i=1}^K \big(\partial_a S^\lambda_{\max} \circ g\big)_i \big(\partial_d f_i\big), \qquad \partial_m\big(S^\lambda_{\max} \circ g\big) := \sum_{i=1}^K \big(\partial_a S^\lambda_{\max} \circ g\big)_i \big(\partial_m f_i\big) \\ \partial_m^2\big(S^\lambda_{\max} \circ g\big) &= \sum_{i=1}^K \big(\partial_a S^\lambda_{\max} \circ g\big)_i \big(\partial_m^2 f_i\big) + \sum_{i,j=1}^K \big(\partial_a^2 S^\lambda_{\max} \circ g\big)_{ij} \big(\partial_m f_i\big)\big(\partial_m f_i\big)^\top.\end{aligned} \tag{A.3}$$

From (H.4), $\partial_a^k S^\lambda_{\max}(a) = \mathcal{O}(\lambda^{1-k})$ for $k = 1,2,3$. By (H.3), any terms involving derivatives of f in (A.3) are uniformly bounded. By mean value theorem and (H.1), we may estimate (A.2) by

$$\begin{aligned}&\mathbb{E}\big[\big(S^\lambda_{\max} \circ g\big)\big(\Phi(\cdot,\cdot,i,\boldsymbol{\xi}_t)\big)\big] - \big(S^\lambda_{\max} \circ g\big) \\ &\quad = \big\langle \partial_d\big(S^\lambda_{\max} \circ f\big); b_i \big\rangle + \big\langle \partial_m\big(S^\lambda_{\max} \circ f\big); \mu_i \big\rangle + \frac{1}{2}\big\langle \partial_m^2\big(S^\lambda_{\max} \circ f\big); \sigma_i\sigma_i^\top \big\rangle + R_T^2 = L_i^\lambda + R_T^2\end{aligned} \tag{A.4}$$

where $|R_T^2| \le C(1+\lambda^{-2})\|d\|^3 + c(1+\lambda^{-2})\|d\|^2$.

Substitute $c = L^\lambda(m,d)\big(\sum_{s=1}^{T-1}\beta^s\big)$ and use the fact that $|L^\lambda(m,d)| \le (1+\lambda^{-1})\|d\|^2$, we obtain

$$\Big|\mathbb{E}\, S^\lambda_{\max}\left(f\big(\Phi(\cdot,\cdot,i,\boldsymbol{\xi}_t)\big) + \Big(\sum_{s=1}^{T-1}\beta^s\Big)L^\lambda\right) - \big(S^\lambda_{\max}\circ\alpha^\lambda_T\big) - L^\lambda_i\Big| \le C\Big(P_{\lambda,d}(2,3) + (1-\beta)^{-1}P_{\lambda,d}(3,4)\Big). \tag{A.5}$$

Finally, denote $a(m,d,i,\boldsymbol{\xi}_t) := f\big(\Phi(m,d,i,\boldsymbol{\xi}_t)\big) + L^\lambda(m,d)\big(\sum_{s=1}^{T-1}\beta^s\big)$, it follows from (H.4) that the first derivative of $S^\lambda_{\max}$ is bounded and does not depend on λ. Hence, it follows from the mean value theorem that

$$\begin{aligned}\mathbb{E}\Big|S^\lambda_{\max}\Big(\alpha^\lambda_T\big(\Phi(m,d,i,\boldsymbol{\xi}_t)\big)\Big) - S^\lambda_{\max}\Big(a(m,d,i,\boldsymbol{\xi}_t)\Big)\Big| &\le C\mathbb{E}\big|\alpha^\lambda_T\big(\Phi(m,d,i,\boldsymbol{\xi}_t)\big) - a(m,d,i,\boldsymbol{\xi}_t)\big| \\ = C\mathbb{E}\Big|\Big(\sum_{s=1}^{T-1}\beta^s\Big)L^\lambda\big(\Phi(m,d,i,\boldsymbol{\xi}_t)\big) - \Big(\sum_{s=1}^{T-1}\beta^s\Big)L^\lambda(m,d)\Big| &\le C(1-\beta)^{-1}(1+\lambda^{-2})\|d\|^3 \end{aligned} \tag{A.6}$$

where the final inequality follows from Lemma 4.3. Combining (A.5) and (A.6), the result follows. □

Theorem A.1 (Tauberian theorem 1)**.** *Let (a_t) be a real-value sequence converging to a. Then $\sum_{t=0}^{T-1}\beta^t a_{T-t} \to (1-\beta)^{-1}a$ as $T\to\infty$.*

Proof. Fix $\epsilon > 0$ and find $s > 0$ such that for all $t \ge s$, $|a_t - a| \le \epsilon$. Since (a_t) is also bounded, we can find $T_0 > s$ such that for any $t > T_0 - s$, $\beta^{t/2}|a_u - a| \le \epsilon$ for all $u \in \mathbb{N}$. Hence, for any $T \ge T_0$,

$$\begin{aligned}\Big|\sum_{t=0}^{T-1}\beta^t a_{T-t} - (1-\beta)^{-1}a\Big| &= \Big|\sum_{t=0}^{T-1}\beta^t(a_{T-t}-a) + \beta^T(1-\beta)^{-1}a\Big| \\ &\le \sum_{t=T-s}^{T-1}\beta^t|a_{T-t}-a| + \sum_{t=s+1}^{T}\beta^{T-t}|a_t - a| + \beta^T(1-\beta)^{-1}|a| \\ &\le \sum_{t=T-s}^{T-1}\beta^{t/2}\epsilon + \sum_{t=s+1}^{T}\beta^{T-t}\epsilon + \beta^T(1-\beta)^{-1}|a| \le (1-\sqrt{\beta})^{-1}\epsilon + (1-\beta)\epsilon + \beta^T(1-\beta)^{-1}|a|.\end{aligned}$$

Hence, $\limsup_{T\to\infty}\big|\sum_{t=0}^{T-1}\beta^t a_{T-t} - (1-\beta)^{-1}a\big| \le (1-\sqrt{\beta})^{-1}\epsilon + (1-\beta)\epsilon$. Since ϵ is arbitrary, we obtain the required result. □

Theorem A.2 (Tauberian theorem 2)**.** *Let $g : \mathbb{N}_0\times\mathbb{N}_0 \to \mathbb{R}$ and $g^* : \mathbb{N}_0 \to \mathbb{R}$ be functions with a constant $C \ge 0$ such that $|g(t,T)| \le Ct$ and $g(t,T) \to g^*(t)$ as $T\to\infty$ for all $t \in \mathbb{N}$. Then for any $\beta \in (0,1)$, $\sum_{t=0}^{T-1}\beta^t g(t,T) \to \sum_{t=0}^{\infty}\beta^t g^*(t)$ as $T\to\infty$.*

Proof. Fix $\epsilon > 0$ and find $N > 0$ such that for all $t \ge N$ and $T \in \mathbb{N}$, $\beta^{t/2}|g(t,T) - g^*(t)| \le \epsilon$ where such N exists due to the linear growth condition.

Since $g(t,T) \to g^*(t)$ as $T\to\infty$, we can find $T_0 \ge N$ such that for any $T \ge T_0$, $|g(t,T) - g^*(t)| \le \epsilon$ for all $t = 0,1,...,N-1$. For any $T \ge T_0$,

$$\begin{aligned}\Big|\sum_{t=0}^{T-1}\beta^t g(t,T) - \sum_{t=0}^{\infty}\beta^t g^*(t)\Big| &\le \sum_{t=0}^{N-1}\beta^t|g(t,T)-g^*(t)| + \sum_{t=N}^{T-1}\beta^t|g(t,T)-g^*(t)| + \sum_{t=T}^{\infty}\beta^t|g^*(t)| \\ &\le \sum_{t=0}^{N-1}\beta^t\epsilon + \sum_{t=N}^{T-1}\beta^{t/2}\epsilon + C\sum_{t=T}^{\infty}t\beta^t \le (1-\beta)^{-1}\epsilon + (1-\sqrt{\beta})^{-1}\epsilon + T\beta^{T-1}(1-\beta)^{-2}.\end{aligned}$$

Hence, $\limsup_{T\to\infty} \left| \sum_{t=0}^{T-1} \beta^t a_{T-t} - (1-\beta)^{-1} a \right| \leq (1-\sqrt{\beta})^{-1}\epsilon + (1-\beta)\epsilon$. Since ϵ is arbitrary, we obtain the required result. $\square$