

# BelMan: Bayesian Bandits on the Belief–Reward Manifold

Debabrota Basu\*  
National University of Singapore  
Singapore

Pierre Senellart†  
École Normale Supérieure  
Paris, France

Stéphane Bressan‡  
National University of Singapore  
Singapore

## ABSTRACT

We propose a generic, Bayesian, information geometric approach to the exploration–exploitation trade-off in multi-armed bandit problems. Our approach, BelMan, uniformly supports pure exploration, exploration–exploitation, and two-phase bandit problems. The knowledge on bandit arms and their reward distributions is summarised by the barycentre of the joint distributions of beliefs and rewards of the arms, the *pseudobelief-reward*, within the beliefs-rewards manifold. BelMan alternates *information projection* and *reverse information projection*, i.e., projection of the pseudobelief-reward onto beliefs-rewards to choose the arm to play, and projection of the resulting beliefs-rewards onto the pseudobelief-reward. It introduces a mechanism that infuses an exploitative bias by means of a *focal distribution*, i.e., a reward distribution that gradually concentrates on higher rewards. Comparative performance evaluation with state-of-the-art algorithms shows that BelMan is not only competitive but can also outperform other approaches in specific setups, for instance involving many arms and continuous rewards.

## KEYWORDS

Multi-armed bandit; Statistical manifolds; Bayesian bandit; Alternating information projection

## 1 INTRODUCTION

*Multi-armed bandits* [33] are a class of sequential decision-making problems [15] where an agent with incomplete information learns about a set of statistical populations and decides to sample from one of them depending on its goal. Such problems are found in a wide variety of real-life applications, from design of ethical clinical trials [37] to task assignment in crowdsourcing markets [21].

In the classical version, a gambler comes across a slot machine with multiple arms. Each arm has a probability distribution of rewards. She plays an arm at a time and receives a reward sampled from the arm’s distribution. The goal of the gambler is to find a strategy that will let her decide the arm to pull to maximise cumulative reward. In the literature, this setup is known as the *stochastic bandit* [3]. This is a simple, archetypal setting of reinforcement learning [35]. We refer to it in this paper as the *exploration–exploitation bandit*. In another variant of the bandit problem, the gambler plays in order to accumulate more information about the arms rather than to maximise cumulative reward. This is called the *pure exploration*

*bandit* [8]. This is linked to efficient identification of the best set of arms [7] and to scenarios where information accumulation has external constraints [10]. We elaborate on both problems in Section 2.

Thus, bandit problems deal with two issues [29]: accumulation of information to reduce uncertainty of decision making (*exploration*) and leveraging present knowledge to gain higher rewards (*exploitation*). Since the agent starts with incomplete information about the stochastic reward structure and gradually discovers more through actions, exploration is necessary. Investigating the pure exploration problem allows us to focus on these aspects, equally significant for exploration–exploitation bandits. On the other hand, in the exploration–exploitation problem, exploration alone is not sufficient: the gambler has to exploit available information to draw the optimal arm. The trade-off between exploration and exploitation emerges as a central question.

*Frequentist* algorithms with *optimism in the face of uncertainty* such as UCB [3] and KL-UCB [18] are state-of-the-art for the exploration–exploitation setup. Though these algorithms work considerably well, their frequentist nature prevents assimilation of *a priori* knowledge about the arms or the underlying process. Such prior can improve performance in applications where an underlying model for the reward distributions can be constructed [25]. Bayesian algorithms, such as Thompson sampling [36] and Bayes-UCB [23], leverage a prior distribution that summarizes the *a priori* knowledge. Though Thompson sampling is popular due to its generality and simplicity, it does not shed any light on the pure exploration problem and thus fails to unify both the variants of bandits [24]. The same holds for Bayes-UCB. Indeed, [24] states that formulation of a Bayesian algorithm for pure exploration is yet to be satisfactorily solved. We propose here a unified Bayesian approach (Section 3) to model this underlying uncertainty and use it to address both pure exploration and exploration–exploitation bandits. Our approach leverages elements of information geometry [1] to address the questions of information representation, accumulation, and exploration–exploitation trade-off.

We maintain a joint distribution over the parameter of the reward distribution and the reward itself. We refer to the distribution over the parameter as the *belief distribution*, and to the joint one as the *belief-reward distribution*. It quantifies the total uncertainty of the underlying process. We further investigate the *belief-reward manifold* of all possible belief-reward distributions. This statistical manifold structure enables information-theoretic and geometric analysis of bandit problems. As rewards are accumulated, belief-reward distributions are updated using Bayes’ theorem. This amounts to a displacement in the belief-reward space.

Exploiting this structure to update the belief-reward distribution of the arms would lead to an effective estimate of the reward distribution of the most played arm but would get myopically stuck in it. Thus, efficient exploration requires a collective representation of the

\*D. Basu is PhD student in School of Computing, National University of Singapore.

†P. Senellart is Professor in DI, École Normale Supérieure, Paris.

‡S. Bressan is Associate Professor in School of Computing, National University of Singapore.

knowledge accumulated by the agent. We propose a *pseudobelief-reward* as the geometric representation of this collective knowledge-base. It is a distribution that belongs to the convex hull created by the belief-reward distributions of all arms and minimises the sum of KL-divergences. We show the pseudobelief-reward is a weighted barycentre of the belief-reward distributions of the arms. Though pseudobelief-reward deals with exploration, in the exploration–exploitation problem, it is essential to exploit the present knowledge of rewards and to gradually increase exploitation in order to achieve higher cumulative reward [29]. We construct a *focal distribution* in the reward space that incrementally focuses on higher rewards with each iteration. This evolution towards higher values of reward depends on a time-variant factor, called *exposure*. Exposure decreases with time and its variation decides the exploration–exploitation trade-off.

We develop an algorithm, BelMan, based on *alternating information projection* [12]. BelMan alternates *information* (I-) and *reverse information* (rI-) projections between the set of belief-reward distributions of the arms and the pseudobelief-focal distributions. Thus, BelMan iteratively updates its knowledge about the reward distributions of the arms and decides the arm to be explored to maximise the cumulative reward as well as the information. By convergence of alternating information projection [12], BelMan asymptotically estimates the ‘true’ reward distributions for the arms and converges to the choice of the optimal arm. Since BelMan solves both pure exploration and exploration–exploitation bandit problems, it provides a single framework to deal with representation of underlying uncertainty, accumulation of observables and information, pure exploration, and exploration–exploitation trade-off.

Though the approach is independent of the family of probability distributions, I- and rI-projections are unique and smoothly computable for the very general exponential family of distributions [1]. Thus, in Section 4, we evaluate the performance of BelMan on two reward distributions of the exponential family – Bernoulli and exponential – on various numbers of arms and parameters. These experiments validate the applicability of BelMan for both discrete and continuous rewards. BelMan exhibits comparable and sometimes better performance with respect to state-of-art bandit algorithms, in both pure exploration and exploration–exploitation scenarios. We also apply BelMan to the two-phase reinforcement learning problem [32], demonstrating its flexibility.

## 2 PROBLEM FORMULATION

We fix a finite number, say  $k > 1$ , of independent real-valued statistical populations  $\{X_j\}_{j=1}^k$ . Each of the populations is specified by corresponding probability density functions  $\{f_\theta^j(x)\}_{j=1}^k$  with respect to a base measure  $\nu$ . We assume that the form of the probability distribution  $f$  is known but the parametrization  $\theta \in \Theta \subseteq \mathbb{R}^d$  is unknown. For example, for Bernoulli bandits  $f_\theta(x) \triangleq \theta^x (1-\theta)^{(1-x)}$ . We call each of the populations an *arm* and the corresponding density function the *reward distribution* of the arm. Following the bandit literature [28], we assume the expectation of the reward distribution  $\mu(\theta) \triangleq \int_{-\infty}^{\infty} x f_\theta(x) d\nu(x)$  well-defined and finite. Sequentially drawing the arms return a sequence of rewards  $[x_n]_{n \geq 0}$ . The agent specifies a *policy* or strategy that will sequentially draw a set of arms depending on her previous actions, observations and intended goal.

*Exploration–exploitation bandit problem.* In exploration–exploitation bandits, the agent searches for a policy that maximises the expected value of *cumulative reward*  $S_n \triangleq \sum_{i=1}^n x_i$  as  $n \rightarrow \infty$ . A policy is *asymptotically consistent* [33] if it asymptotically tends to choose the arm with maximum expected reward, i.e.,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_\theta[S_n] = \mu^*(\theta) \quad \forall \theta \in \Theta^k \quad (1)$$

where  $\mu^*(\theta) \triangleq \max_{1 \leq j \leq k} \mu(\theta_j)$ . The *cumulative regret*  $R_n(\theta)$  [28] is the amount of extra reward the agent can obtain if it follows the optimal policy rather than the present sequence:

$$R_n(\theta) \triangleq n\mu^*(\theta) - \mathbb{E}_\theta[S_n] = \sum_j [\mu^*(\theta) - \mu(\theta_j)] \mathbb{E}_\theta[T_{n,j}],$$

where  $T_{n,j}$  is the number of times arm  $j$  is pulled till the  $n$ th iteration. [28] proved that for all algorithms satisfying  $R_n(\theta) = o(n^a)$  for a non-negative  $a$  and given  $\theta$ , the cumulative regret increases asymptotically in  $\Omega(\log n)$ . Such algorithms are called *asymptotically efficient*. Based on this bound, [3] extensively studied the upper confidence bound (UCB) family of algorithms. Later on, this family of algorithms was analysed and improved to propose algorithms such as KL-UCB [18].

Frequentist approaches implicitly assume an optimal parametrization  $\theta^*$ . In contrast, Bayesians model the uncertainty on the parameter using another probability distribution  $B(\theta)$  [15, 34], the *belief distribution*. We begin with a prior  $B_0(\theta)$  over the parameters and eventually try to find out a posterior distribution such that the Bayesian sum of rewards  $\int \mathbb{E}_\theta[S_n] dB(\theta)$  is maximised or the Bayesian risk  $\int R_n(\theta) dB(\theta)$  is minimised. Though there exists an optimal algorithm for discounted Bayesian bandits based on the Gittins index [20], explicit computation of the indices is not always tractable and does not provide clear insights into what they look like and how they change as sampling proceeds [30].

Thus, researchers developed approximation algorithms [27] and sequential sampling schemes like Thompson sampling [36]. At any iteration, the latter samples  $k$  parameter values from the belief distributions and chooses the arm that has maximum expected reward for them. [23] also proposed a Bayesian analogue of the UCB algorithm. Unlike the original, it uses belief distributions to keep track of arm uncertainty and update them using Bayes’ theorem, computes UCBs for each arm using the belief distributions, and chooses the arm accordingly.

*Pure exploration bandit problem.* In this variant of the bandit problem, the agent aims to gain more information about the arms. [8] formulated this notion of gaining information as minimisation of the simple regret rather than cumulative regret. *Simple regret*  $r_n(\theta)$  is the expected difference between the maximum achievable reward  $\mu^*(\theta)$  and the achieved reward  $\mathbb{E}_\theta[x_n]$ . Unlike cumulative regret, minimising simple regret depends only on exploration and the number of available rounds to do so. [8] proved that, for Bernoulli bandits, if an exploration–exploitation algorithm achieves an upper-bounded regret, it cannot reduce the expected simple regret by more than a fixed lower bound. This establishes the fundamental difference between exploration–exploitation bandits and pure exploration bandits. Existing frequentist algorithms [2, 7, 24] do not provide an intuitive and rigorous explanation of how a unified framework would work for both the pure exploration and the exploration–exploitation scenario. As discussed in Section 1, both Thompson sampling and

Bayes-UCB also lack this feature of constructing a single successful structure for both pure exploration and exploration–exploitation.

### 3 METHODOLOGY

In this section, we formulate the bandit problem in terms of belief-reward distributions and define the belief-reward manifold. Following this, we propose an alternating information projection scheme, BelMan, on the belief-reward manifold. In this context, we construct pseudobelief-reward and focal distributions. Finally, we instantiate BelMan to the exponential family of reward distributions.

#### 3.1 The Belief-Reward Manifold

As mentioned in Section 2, the probabilistic nature of the reward corresponding to the  $j^{\text{th}}$  arm is represented using *reward distributions*  $f_{\theta}^j(x)$ . In the parametric setting, we assume them to have the same form but vary on the parametrisation  $\theta_j \in \mathbb{R}$ . Thus,  $f_{\theta}^j(x) \equiv f_{\theta_j}(x)$ . For a given smooth probability density function  $f_{\cdot}$ , the space of all reward distributions constructs a  $k$ -dimensional smooth statistical manifold [1]  $\mathcal{R}$ . We call  $\mathcal{R}$  the *reward manifold*. Since the agent plays with partial information, the ‘true’ parameter vector of the arms,  $\theta = [\theta_1, \dots, \theta_k]$  is not certainly known. The uncertainty over  $\theta$  is represented using another probability distribution  $B(\theta)$ . We call  $B(\theta)$  the *belief distribution*.

In the Bayesian bandit process, the agent starts with a prior belief distribution  $B_0(\theta)$ . This structure of prior distribution is flexible, both to be uninformative or to carry some *a priori* information [22]. The agent sequentially chooses an arm  $a_n$  at each time step  $n$ . The agent samples a reward  $x_n$  from  $f_{\theta_{a_n}}$  with expected value  $\mu_{a_n}$ . The actions taken and rewards obtained by the player till time  $n$  create the history of the bandit process,  $\mathcal{H}_n \triangleq [(a_1, x_1), (a_2, x_2), \dots, (a_{n-1}, x_{n-1})]$ . This history  $\mathcal{H}_n$  sequentially constructs the belief distribution over the parameter vector as  $B_n(\theta) \triangleq \mathbb{P}(\theta \mid \mathcal{H}_n)$ . We define the space consisting of all such distributions over  $\theta$  as the *belief space*  $\mathcal{B}$ . This space  $\mathcal{B}$  for a smooth probability density function  $B$  is a  $(k \times d)$ -dimensional statistical manifold, where  $d$  is the dimension of the parameter space  $\Theta$ . We call  $\mathcal{B}$  the *belief manifold* of the multi-armed bandit process. In order to consider uncertainties of partial information along with the stochastic nature of reward using a single representation, we define belief-reward distributions.

*Definition 1 (Belief-reward distribution).* The joint distribution on reward and parameter for the  $j^{\text{th}}$  arm the  $n^{\text{th}}$  iteration as  $\mathbb{P}_n^j(x, \theta)$  is defined as the *belief-reward distribution*.

$$\mathbb{P}_n^j(x, \theta) \triangleq \frac{b_n^j(\theta) f_{\theta}(x)}{\int_{x \in \mathbb{R}} \int_{\theta \in \Theta} b_n^j(\theta) f_{\theta}(x) d\theta dx} = \frac{1}{Z} b_n^j(\theta) f_{\theta}(x).$$

LEMMA 2. *The set of belief-reward distributions  $\mathbb{P}(x, \theta)$  defines a manifold  $\mathcal{BR}$ , such that  $\mathcal{BR} = \mathcal{B} \times \mathcal{R}$ . We call it the *belief-reward manifold*.*

Following the trend of the bandit literature and Bayesian methods, we construct our work by assuming arm independence and a Bayesian evolution of belief distribution.

ASSUMPTION 1 (INDEPENDENCE OF ARMS). *The parameters  $\{\theta_j\}_{1 \leq j \leq k}$  are drawn independently from  $k$  belief distributions  $\{b_n^j(\cdot)\}_{1 \leq j \leq k}$*

such that

$$B_n(\theta) = \prod_{j=1}^k b_n^j(\theta_j) \triangleq \prod_{j=1}^k \mathbb{P}(\theta_j \mid \mathcal{H}_n).$$

This implies that the belief manifold  $\mathcal{B}$  is a product of  $k$  manifolds  $\mathcal{B}^j \triangleq \{b^j(\theta_j)\}$ . Here,  $\mathcal{B}^j$  is the statistical manifold of belief distributions for the  $j^{\text{th}}$  arm. Due to the common parametrization, the  $\mathcal{B}^j$ ’s can be represented by a single manifold  $\mathcal{B}_{\theta}$ . Thus, we operate on  $\mathcal{B}_{\theta} \mathcal{R}$  and represent the belief-reward distribution of each arm as a point in that space. We slightly misuse the terminology to call each  $\mathcal{B}_{\theta} \mathcal{R}$  also a belief-reward manifold. Though Assumption 1 is followed throughout this paper, it is not essential to develop this framework. It is assumed to make calculations easier.

ASSUMPTION 2 (BAYESIAN EVOLUTION). *When conditioned over  $\{\theta_j\}_{1 \leq j \leq k}$  and the choice of arm, the sequence of rewards  $[x_1, \dots, x_n]$  is jointly independent.*

Thus, the Bayesian update at the  $n^{\text{th}}$  iteration is given by:

$$b_{n+1}^j(\theta_{a_n}) \propto f_{\theta_{a_n}}(x_n) b_n^j(\theta_j) \quad (2)$$

if the  $j^{\text{th}}$  arm is drawn and a reward  $x_n$  is obtained. For all other arms, the belief remains unchanged. We can then deduce the belief update after each of the iteration as a movement on the belief manifold from a point  $B_n$  to another point  $B_{n+1}$  with *maximum information gain*. Thus, the process of bandit games and the evolution of belief-reward distributions creates a set of trajectories on the belief-reward manifold. The goal of such trajectories is to reach the points in the belief-reward manifold which resembles the ‘true’ reward distributions well enough either to estimate the true distributions well or to decrease the regrets accumulated in the path as much as possible.

#### 3.2 BelMan: An Alternating Projection Scheme

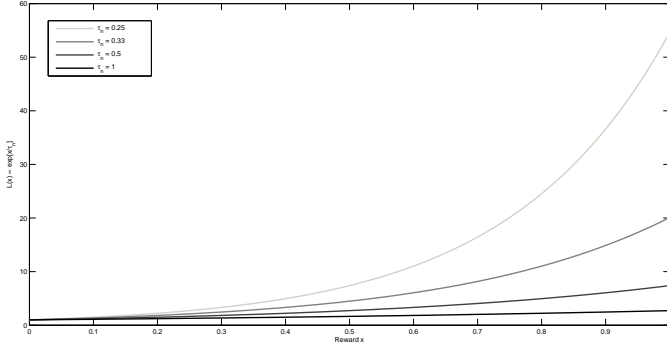
Now, the central puzzle is how to represent exploration and exploitation on the belief-reward manifold and how to incentivise it to form an algorithm.

*Pseudobelief: Summarising the explored knowledge.* The agent aims to establish a collective knowledge on the arms via exploration. In the belief-reward manifold, we represent this using the pseudobelief-reward distribution.

*Definition 3 (Pseudobelief-reward distribution).* A *pseudobelief-reward distribution*  $\bar{\mathbb{P}}(x, \theta)$  is a point in the belief-reward manifold that minimises the sum of KL-divergences from the belief-reward distributions  $\mathbb{P}^j(x, \theta)$  of all the arms.

$$\bar{\mathbb{P}}(x, \theta) \triangleq \arg \min_{\mathbb{P} \in \mathcal{B}_{\theta} \mathcal{R}} \sum_{j=1}^k D_{\text{KL}} \left( \mathbb{P}^j(x, \theta) \parallel \mathbb{P}(x, \theta) \right). \quad (3)$$

Since the belief-reward distribution of each arm is a point on the belief-reward manifold, the pseudobelief-reward is the barycentre of their convex hull with respect to KL-divergence. Since any point inside a polygon in a flat space is representable as a linear combination of the vertices, we take these individual distributions as a basis of our representation, and introduce coefficients  $(\lambda_j)_{1 \leq j \leq k}$  for the pseudobelief-reward. Since KL-divergence is e-flat i.e., linear with respect to scalars in exponent, we express pseudobelief-reward as a log-linear combination of them. Thus,  $\log \bar{\mathbb{P}}(x, \theta) \triangleq$



**Figure 1: Evolution of the focal distribution over  $x \in [0, 1]$  for  $\tau(n) = 1, 0.5, 0.33$  and  $0.25$ .**

$\sum_{j=1}^k \lambda_j \log \mathbb{P}_n^j(x, \theta)$ , where  $\lambda_j \in [0, 1]$ , and  $\sum_{j=1}^k \lambda_j = 1$ . This implies that  $\bar{\mathbb{P}}(x, \theta) = \prod_{j=1}^k \mathbb{P}_n^j(x, \theta)^{\lambda_j}$ . For  $\lambda_j = 0$ , the pseudobelief-reward becomes oblivious to arm  $j$ . For  $\lambda_j = 1$ , the pseudobelief-reward coincides with the belief-reward distribution of arm  $j$ .  $\sum_{j=1}^k \lambda_j = 1$  keeps the pseudobelief-reward normalised. These constraints keep the pseudobelief-reward a summary of belief-reward distributions of all arms.

*Focal distribution: Infusing exploitative bias.* Creating a succinct pseudobelief-reward is essential for both pure exploration and exploration–exploitation but not sufficient for minimising the cumulative regret in exploration–exploitation. Thus, we introduce the focal distribution that gradually focuses on higher rewards and infuses a bias towards choosing the optimal arm.

*Definition 4 (Focal distribution).* A focal distribution is a reward distribution of the form  $L_n(x) \propto \exp\left(\frac{x}{\tau(n)}\right)$ , where  $\tau(n)$  is a decreasing function of  $n \geq 1$ . We call  $\tau(n)$  the *exposure* of the focal distribution.

The focal distribution gradually concentrates on higher rewards as the exposure  $\tau(n)$  decreases with time. We see this feature in Figure 1. Thus, it constrains using KL-divergence to choose distributions with higher rewards and infuses the exploitive bias. Following the bounds obtained in [18], we set the focal distribution to  $\tau(n) \triangleq [\log(n) + C \times \log(\log(n))]^{-1}$  with  $C$  a constant (we choose the value  $C = 15$  in the experimental evaluation) for the exploration–exploitation bandit problem. In the pure exploration setup, we simply take  $\tau(n) = \infty$  to remove any bias towards exploitation.

We amalgamate the pseudobelief-reward and the focal distribution to form the *pseudobelief-focal distribution*

$$\bar{\mathbb{Q}}(x, \theta) \triangleq \frac{1}{\bar{Z}_n} \bar{\mathbb{P}}(x, \theta) \exp\left(\frac{x}{\tau(n)}\right).$$

Here,  $\bar{Z}_n = \int_{x \in \mathbb{R}} \int_{\theta \in \Theta} \bar{\mathbb{P}}(x, \theta) \exp\left(\frac{x}{\tau(n)}\right) d\theta dx$  is a normalisation factor. We use the pseudobelief-focal distribution as the representative of explored knowledge and exploitation bias in our algorithm. Following Equation (3), we define the pseudobelief-focal as  $\bar{\mathbb{Q}}(x, \theta) \triangleq \arg \min_{\mathbb{Q}} \sum_{j=1}^k D_{\text{KL}}(\mathbb{P}^j(x, \theta) \parallel \mathbb{Q}(x, \theta))$ .

*An Alternating Projection Scheme.* The main idea of BelMan is to alternately minimise the KL-divergence  $D_{\text{KL}}(\cdot \parallel \cdot)$  [26] between

the belief-reward distributions of the arms and the pseudobelief-focal distribution.

[13] introduced the concept of minimisation of KL-divergence with respect to a participating distribution as a projection to the set of the other distribution.

*Definition 5 (I-projection).* The *information projection* (or *I-projection*) of a distribution  $q \in \mathcal{Q}$  onto a non-empty, closed, convex set  $\mathcal{P}$  of probability distributions on a fixed support set is defined by the probability distribution  $p^* \in \mathcal{P}$  that has minimum KL-divergence to  $q$ :  $p^* \triangleq \arg \min_{p \in \mathcal{P}} D_{\text{KL}}(p \parallel q)$ .

Since  $D_{\text{KL}}(p(s) \parallel q(s)) = -h(p(s)) + H(p(s), q(s))$ , we observe that the I-projection  $p^*$  is the distribution in  $\mathcal{P}$  that maximises the entropy  $h(p)$  of  $\mathcal{P}$ , while minimising the mutual information  $H(p, q)$ : it is the distribution in  $\mathcal{P}$  which is most similar to  $q$ . This implies that the I-projection  $p^*$  captures at least the first moment, i.e., the expectation of the fixed distribution  $q$ .

*Definition 6 (rI-projection).* The *reverse information projection* (or *rI-projection*) of a distribution  $p \in \mathcal{P}$  onto  $\mathcal{Q}$ , which is also a non-empty, closed, convex set of probability distributions on a fixed support set, is defined by the distribution  $q^* \in \mathcal{Q}$  that has minimum KL-divergence from  $p$ :  $q^* \triangleq \arg \min_{q \in \mathcal{Q}} D_{\text{KL}}(p \parallel q)$ .

The rI-projection finds the distribution  $q^*$  from a space of candidate distributions  $\mathcal{Q}$  that encodes maximum information of the distribution  $p$ . If the set of candidate distributions is engendered by a statistical model, the rI-projection of the empirical distribution formed from samples to the model is equivalent to finding the *maximum likelihood estimate*. Since rI-projection aims to maximise the complete likelihood rather than finding a distribution with similar entropy,  $q^*$  also captures higher moments of the fixed distribution  $p$ . Thus, it is computationally more demanding but more informative than I-projection.

Due to the underlying minimisation operation, if we begin from  $p_0 \in \mathcal{P}$  and  $q_0 \in \mathcal{Q}$  and alternately perform I-projection and reverse I-projection, it will lead to two distributions  $p^*$  and  $q^*$  for which the KL-divergence between sets  $\mathcal{P}$  and  $\mathcal{Q}$  are minimum. [13]

We now present BelMan (Algorithm 1), which implements this idea of alternate projection for the bandit problems. We are working with distributions in the belief-reward manifold. On the one hand, the algorithm constructs empirical belief-reward distributions  $\mathbb{P}_n(x, \theta)$  by accumulation of rewards. Thus, here the set of all such empirical distributions forms  $\mathcal{P} \subset \mathcal{BR}$ . On the other hand, it computes and updates a representation of knowledge and exploitive bias in the form of the pseudobelief-focal distributions  $\bar{\mathbb{Q}}_n(x, \theta)$ . The set of all pseudobelief-focal distributions constitutes  $\mathcal{Q} \subset \mathcal{BR}$ .

The algorithm is initially provided (Line 1) with a prior belief distribution  $B_0(\theta)$  and reward distributions  $\{f_{\theta_j}(x)\}_{j=1}^k$  for each of the arms. They form the initial empirical point  $\mathbb{P}_0(x, \theta) \triangleq \prod_{j=1}^k b_0^j(\theta) f_{\theta_j}(x)$  in  $\mathcal{P}$ . Similarly, we begin with  $\bar{\mathbb{Q}}_0(x, \theta)$  as the initial point of  $\mathcal{Q}$ . Following the initialisation step, Algorithm 1 processes iteratively, with each iteration formed of three functional parts.

In the first part (Lines 3–4), it decides which arm to pull by an I-projection of the pseudobelief-focal onto the beliefs-rewards of

---

**Algorithm 1** BelMan
 

---

- 1: **Input:** Time horizon  $T$ , Number of arms  $k$ , Prior on parameters  $B_0$ , Reward function  $f$ , Exposure  $\tau(n)$ .
  - 2: **for**  $n = 1$  **to**  $T$  **do**
  - 3:   /\* I-projection \*/
  - 4:   Draw arm  $a_n$  such that
 
$$a_n = \arg \min_j \sum_{j=1}^k D_{\text{KL}} \left( \mathbb{P}_n^j(x, \theta) \parallel \bar{\mathbb{Q}}_{n-1}(x, \theta) \right). \quad (4)$$
  - 5:   /\* Accumulation of observables \*/
  - 6:   Sample a reward  $x_n$  out of  $f_{\theta_{a_n}}$ .
  - 7:   Update the belief-reward distribution of  $a_n$  using Bayes' theorem.
  - 8:   /\* Reverse I-projection \*/
  - 9:   Update the pseudobelief-reward distribution to
 
$$\bar{\mathbb{Q}}_n(x, \theta) = \arg \min_{\bar{\mathbb{Q}} \in \mathcal{B}_\theta \mathcal{R}} \sum_{j=1}^k D_{\text{KL}} \left( \mathbb{P}_n^j(x, \theta) \parallel \bar{\mathbb{Q}}(x, \theta) \right). \quad (5)$$
  - 10: **end for**
- 

each of the arms. It amounts to computing

$$\begin{aligned} a_n &\triangleq \arg \min_j \sum_{j=1}^k D_{\text{KL}} \left( \mathbb{P}_n^j(x, \theta) \parallel \bar{\mathbb{Q}}_{n-1}(x, \theta) \right) \\ &= \arg \min_j \sum_{j=1}^k \left( \mathbb{E}_{\mathbb{P}_n^j(x, \theta)} \left[ \frac{-x}{\tau(n)} \right] + D_{\text{KL}} \left( b_n^j(\theta) \parallel b_{\bar{n}_n}(\theta) \right) \right). \end{aligned}$$

The first term symbolises the expected reward of arm  $j$ . Maximising this term alone is analogous to greedily exploiting the present information about the arms. The second term quantifies the amount of uncertainty that can be decreased if arm  $j$  is chosen on the basis of the present pseudobelief. The exposure  $\tau(n)$  of the focal distribution keeps a weighted balance between exploration and exploitation. Increasing  $\frac{1}{\tau(n)}$  increases the exploitation with time which is quite an intended property of an exploration–exploitation algorithm.

In the next part (Line 5–7), the agent plays the chosen arm  $a_n$  and samples a reward  $x_n$ . This observation is incorporated in the belief of the arm using Bayes' rule of Equation (2).

In the last part (Lines 8–9), the updated beliefs are used to obtain the pseudobelief-focal distribution using rI-projection. If we substitute the log-linear form of pseudobelief in Equation (5), we observe that computing  $\bar{\mathbb{Q}}_n$  is equivalent to finding  $\arg \max_{\lambda_1, \dots, \lambda_k} \sum_{j=1}^k (1 - \lambda_j) h(b_n^j)$ . Thus, we compute such a vector of  $\lambda_j$ 's that allows the updated pseudobelief to encode as much information as possible from all the belief distributions. The normalisation factor that depends on  $\tau(n)$  creates additional constraints in this minimisation. Here, BelMan is infusing the exploitative bias. It keeps the pseudobelief-focal distribution away from the 'actual' barycentre of the belief-reward distributions and pushes it towards the arms with higher expected reward. Increasing exploitative bias eventually merges the pseudobelief-focal distribution to the 'true' reward distribution of the optimal arm that has the highest expected reward.

**THEOREM 7 (ASYMPTOTIC CONSISTENCY).** *For a large  $n$ , BelMan will asymptotically converge to choosing the optimal arm. Mathematically,*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_\theta [S_n] = \mu^*(\theta) \quad (6)$$

where  $\mu^*(\theta) \triangleq \max_{1 \leq j \leq k} \mu(\theta_j)$ .

**PROOF.** Without loss of generality, let us consider there exists at least one optimal arm and it is identified as the arm 1. At the I-projection, we choose the arm that has minimum KL-divergence  $D_{\text{KL}}(\mathbb{P}_n^j(x, \theta) \parallel \bar{\mathbb{Q}}(x, \theta))$  from the pseudobelief–lens distribution. Thus, we have to prove that for large  $n$ ,  $D_{\text{KL}}(\mathbb{P}_n^1(x, \theta) \parallel \bar{\mathbb{Q}}(x, \theta)) - D_{\text{KL}}(\mathbb{P}_n^a(x, \theta) \parallel \bar{\mathbb{Q}}(x, \theta))$  is non-positive for any  $a \neq 1$ . We begin as follows,

$$\begin{aligned} &D_{\text{KL}}(\mathbb{P}_n^1(x, \theta) \parallel \bar{\mathbb{Q}}(x, \theta)) - D_{\text{KL}}(\mathbb{P}_n^a(x, \theta) \parallel \bar{\mathbb{Q}}(x, \theta)) \\ &= \underbrace{\left[ h(\mathbb{P}_n^a(x, \theta)) - h(\mathbb{P}_n^1(x, \theta)) \right]}_{\text{T1}} \\ &\quad + \underbrace{\int_x \int_\theta \left[ \mathbb{P}_n^a(x, \theta) - \mathbb{P}_n^1(x, \theta) \right] \log \bar{\mathbb{Q}}(x, \theta) d\theta dx}_{\text{T2}} \end{aligned}$$

The first term T1 is the difference in entropy in two of the arms.

$$\begin{aligned} \text{T1} &= \mathbb{E}_{\mathbb{P}_n^1(x, \theta) - \mathbb{P}_n^a(x, \theta)} [\log \mathbb{P}_n^1(x, \theta)] - D_{\text{KL}}(\mathbb{P}_n^a(x, \theta) \parallel \mathbb{P}_n^1(x, \theta)) \\ &\leq \sup_{(a) \ x, \theta} \left[ \mathbb{P}_n^1(x, \theta) - \mathbb{P}_n^a(x, \theta) \right] \log(1 - \varepsilon) - D_{\text{KL}}(\mathbb{P}_n^a(x, \theta) \parallel \mathbb{P}_n^1(x, \theta)) \\ &\leq \log(1 - \varepsilon) \sqrt{\frac{D_{\text{KL}}(\mathbb{P}_n^a(x, \theta) \parallel \mathbb{P}_n^1(x, \theta))}{2}} - D_{\text{KL}}(\mathbb{P}_n^a(x, \theta) \parallel \mathbb{P}_n^1(x, \theta)) \\ &\stackrel{(b)}{\leq} \log(1 - \varepsilon) \sqrt{0.5 D_{\text{KL}}(\mathbb{P}_n^a(x, \theta) \parallel \mathbb{P}_n^1(x, \theta))}. \end{aligned}$$

Since we consider that all the distributions has well-defined, non-empty supports, we have  $1 > \mathbb{P}_n^a(x, \theta) > 0$  for all  $a, x, \theta$ . Let us assume the maximum of all such values is  $1 - \varepsilon \in (0, 1)$ . This gives us inequality (a). Inequality (b) is derived from Pinsker's inequality [11]. (c) is valid due to the non-negativity of KL-divergence. Similarly, we get for the second term T2:

$$\begin{aligned} \text{T2} &= \int_x \int_\theta \left[ \mathbb{P}_n^a(x, \theta) - \mathbb{P}_n^1(x, \theta) \right] \log \bar{\mathbb{Q}}(x, \theta) d\theta dx \\ &= \int_x \int_\theta \left[ \mathbb{P}_n^a(x, \theta) - \mathbb{P}_n^1(x, \theta) \right] \log \left( \prod_j \mathbb{P}_n^j(x, \theta)^{\lambda_j} \right) d\theta dx \\ &\quad - \frac{1}{\tau(n)} \mathbb{E}_{\mathbb{P}_n^1(x, \theta) - \mathbb{P}_n^a(x, \theta)} [x] + \log \bar{Z}_n \mathbb{E}_{\mathbb{P}_n^1(x, \theta) - \mathbb{P}_n^a(x, \theta)} [1] \\ &\stackrel{(d)}{\leq} \log(1 - \varepsilon) \sqrt{0.5 D_{\text{KL}}(\mathbb{P}_n^a(x, \theta) \parallel \mathbb{P}_n^1(x, \theta))} - \frac{\Delta_n^a}{\tau(n)}. \end{aligned}$$

Here,  $\Delta_n^a = \mu_n^1 - \mu_n^a$  i.e, the deviation of expected reward of the arm  $a$  from the optimal arm. Inequality (d) is obtained if we apply AM-GM inequality, inequality (a) and (b) in sequence. Thus,  $\text{T1} + \text{T2} \leq \log(1 - \varepsilon) \sqrt{2 D_{\text{KL}}(\mathbb{P}_n^a(x, \theta) \parallel \mathbb{P}_n^1(x, \theta))} - \frac{\Delta_n^a}{\tau(n)}$  and it is non-positive if

$$\frac{1}{\tau(n)} \geq \log(1 - \varepsilon) \sqrt{2} \sqrt{\frac{D_{\text{KL}}(\mathbb{P}_n^a(x, \theta) \parallel \mathbb{P}_n^1(x, \theta))}{\Delta_n^a}}.$$

RHS tends to  $\frac{\sqrt{D_{\text{KL}}(\mathbb{P}^a(x, \theta) \parallel \mathbb{P}^1(x, \theta))}}{\Delta^a}$ , as  $n \rightarrow \infty$  and both the arms gather more samples. Thus, RHS is a finite value for a given setup. Hence for any  $n$  greater than  $N$  that satisfies the equality in above inequality BelMan would always choose the optimal arm. This proves that BelMan is asymptotically consistent for any bounded bandit problem.  $\square$

This lower bound on  $\frac{1}{\tau(N)}$  being inversely proportional to  $\Delta^a$  indicates that we have to induce higher exploitative bias to reach higher rewards if the difference between the optimal and the suboptimal arm is minute. It is intuitive as the algorithm would need more samples to distinguish between the optimal and the suboptimal similar to it.

We can also intuitively validate this claim. We know the KL-divergence between belief-reward of any arm and the pseudobelief-reward  $D_{\text{KL}}(\mathbb{P}_n^j(x, \theta) \parallel \mathbb{Q}(x, \theta)) = (1 - \lambda_j)h(b_n^j) - \frac{1}{\tau(n)}\mu_n^j$ . As  $n \rightarrow \infty$ , the entropy of belief on each arm reduces to a constant dependent on its internal entropy. Thus, when  $\frac{1}{\tau(n)}$  exceeds the entropy term for a large  $n$ , BelMan greedily chooses the arm with highest expected reward. Hence, BelMan is asymptotically consistent for bounded finite arm bandits.

### 3.3 BelMan for Exponential Family Distributions

The *exponential family* [6] is a class of probability distributions which can be defined using a set of *natural parameters*  $\omega(\theta)$  and a given natural *sufficient statistics*  $T(x)$  as follows:

$$f_{\theta}(x) \triangleq h(x) \exp(\langle \omega(\theta), T(x) \rangle - A(\theta)).$$

Here,  $h(x)$  is the *base measure* on reward  $x$  and  $A(\theta)$  is called the *log-partition function*. The exponential family includes the majority of the distributions found in the bandit literature such as Bernoulli, beta, Gaussian, Poisson, exponential, and chi-squared.

We choose the exponential family to instantiate our framework not only because of its wide range and applicability but also due to its well behaving Bayesian and information geometric properties. From a Bayesian point of view, the most useful property of the exponential family is the existence of *conjugate distributions* which also belong to this family [6]. Two parametric distributions  $f_{\theta}(x)$  and  $b_{\eta}(\theta)$  are conjugate if the posterior distribution  $\mathbb{P}(\theta|x)$  formed by multiplying them has the same form as  $b_{\eta}(\theta)$ . Thus, if the reward distribution belongs to the exponential family, the belief distribution is represented as:  $b_{\eta}(\theta) \triangleq h(\theta) \exp(\langle \eta, T(\theta) \rangle - A(\eta))$  with the natural parameters  $\eta$ .

Since exponential family distributions are flat with respect to KL-divergence [1], both I-and rI-projections in BelMan are well-defined and unique. Thus, at each iteration, we obtain an optimal and unambiguous choice of the arm and pseudobelief respectively. [1] also stated that the necessary and sufficient condition for a parametric probability distribution to have an efficient estimator is that the distribution belongs to the exponential family and has an expectation parametrization. Thus, working with exponential family distributions implicitly supports the well-defined nature and possibility of getting an efficient estimation. Being a member of the exponential family, the belief distributions  $b_{\eta}(\theta)$  construct a statistical manifold with local co-ordinates  $\eta$  [1]. Thus, we identify each of the

arm's belief distributions as points  $\eta_1, \dots, \eta_k$  and the pseudobelief as  $\bar{\eta} = \sum_{j=1}^k \lambda_j \eta_j$ .

## 4 EXPERIMENTAL EVALUATION

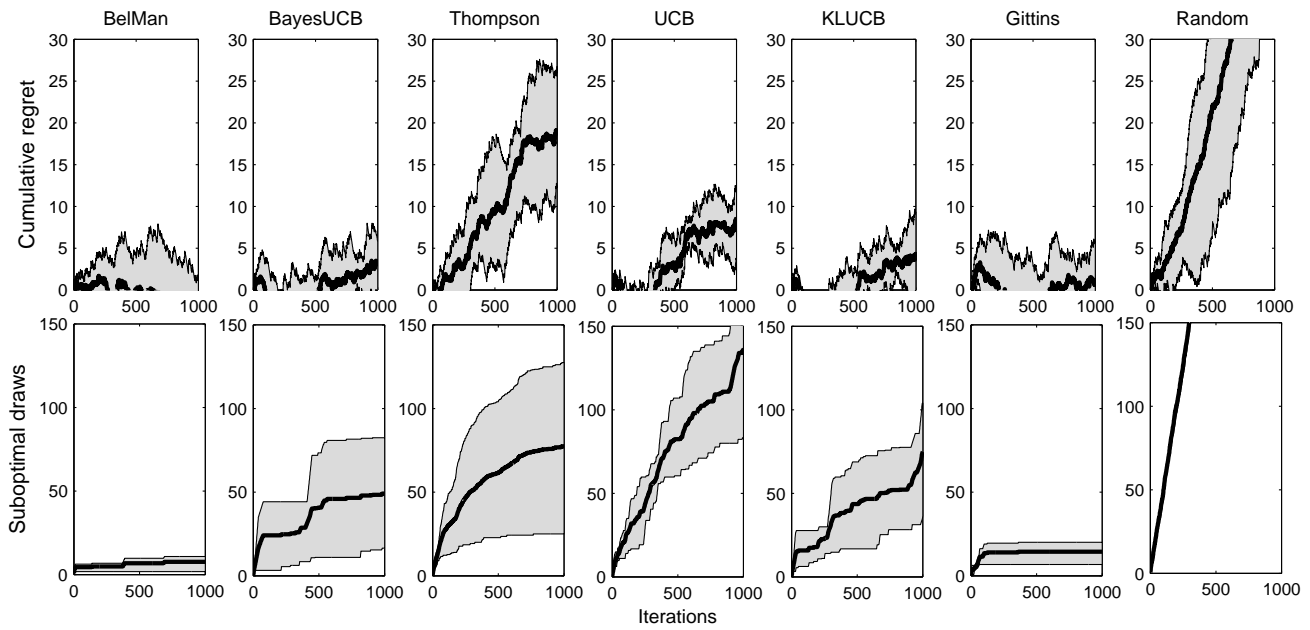
*Exploration-exploitation bandit problem.* We evaluate the performance of BelMan for two exponential family distributions – Bernoulli and exponential. They represent two instances of discrete and continuous rewards respectively. We use the *pymaBandits* library [9] for implementation of all the algorithms except ours, and run it on MATLAB 2013a. We plot the evolution of the mean and the 75 percentile of cumulative regret and number of suboptimal draws. For each instance, we run experiments for 25 runs each consisting of 1000 iterations.

We run experiments for BelMan on two instances of Bernoulli bandits (Figures 2 and 3). We compare the performance of BelMan with frequentist methods like UCB [3] and KL-UCB [18], and Bayesian methods like Thompson sampling [36] and Bayes-UCB [23]. We compare with Gittins index [20] which is the optimal algorithm for Markovian finite arm independent bandits with discounted rewards. Though we are not interested in the discounted case, but the algorithm is indeed transferable to the finite horizon setting with slight manipulation. Though it is often computationally intractable, we use it as the optimal baseline for Bernoulli bandits. We also plot performance of the uniform sampling method (*Random*), as a naïve baseline.

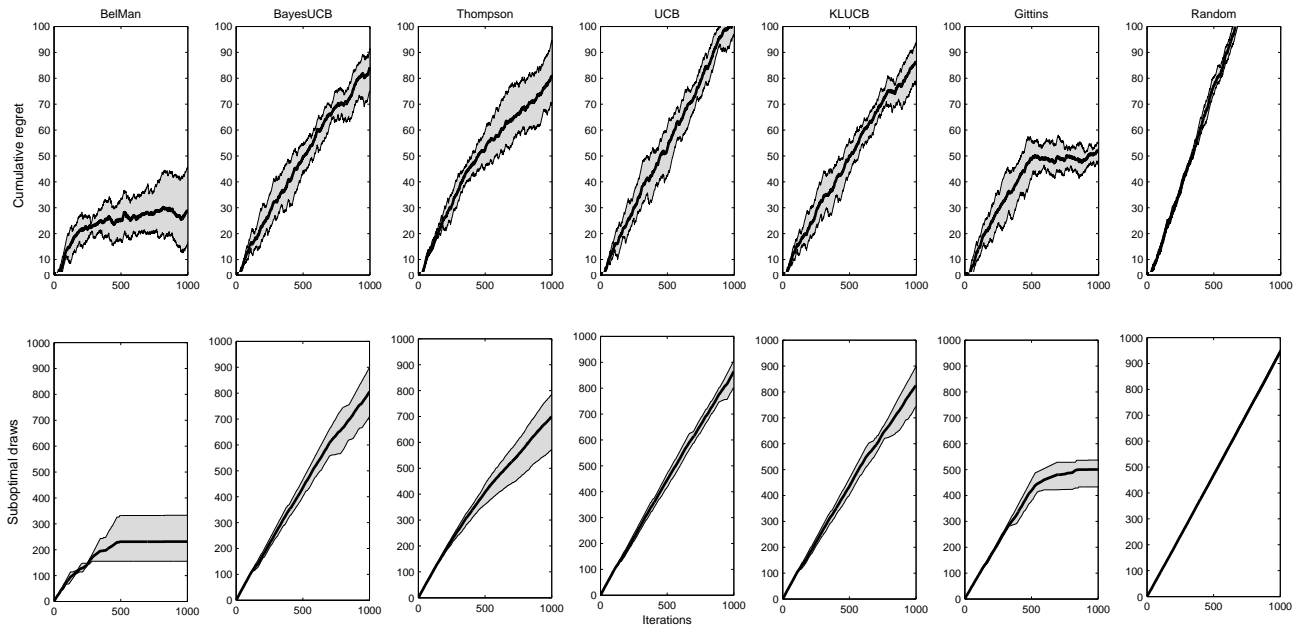
For the 2-arm bandit of Figure 2 ( $\theta_1 = 0.8, \theta_2 = 0.9$ ), we observe that at the very beginning the cumulative regret of BelMan grows linearly and then transitions to a state of slow growth. This initial linear growth of suboptimal draws followed by a logarithmic growth is an intended property of any optimal bandit algorithm as can be seen in the performance of competing algorithms and also pointed out by [19]: an initial phase dominated by exploration and a second phase dominated by exploitation. The phase change indicates the ability of the algorithm to reduce uncertainty after a certain number of iterations and to find a trade-off between exploration and exploitation. BelMan performs comparatively well with respect to the contending algorithms, achieving the phase of exploitation faster than others, with significantly less variance.

Figure 3 depicts similar features of BelMan for 20-arm bandits (with means 0.25, 0.22, 0.2, 0.17, 0.17, 0.2, 0.13, 0.13, 0.1, 0.07, 0.07, 0.05, 0.05, 0.05, 0.02, 0.02, 0.02, 0.02, 0.01, 0.01, and 0.01). Since more arms ask for more exploration and more iterations drawing the suboptimal arms, all algorithms show higher regret values. We note that on all experiments performed, BelMan behaved competitively w.r.t. competing approaches.

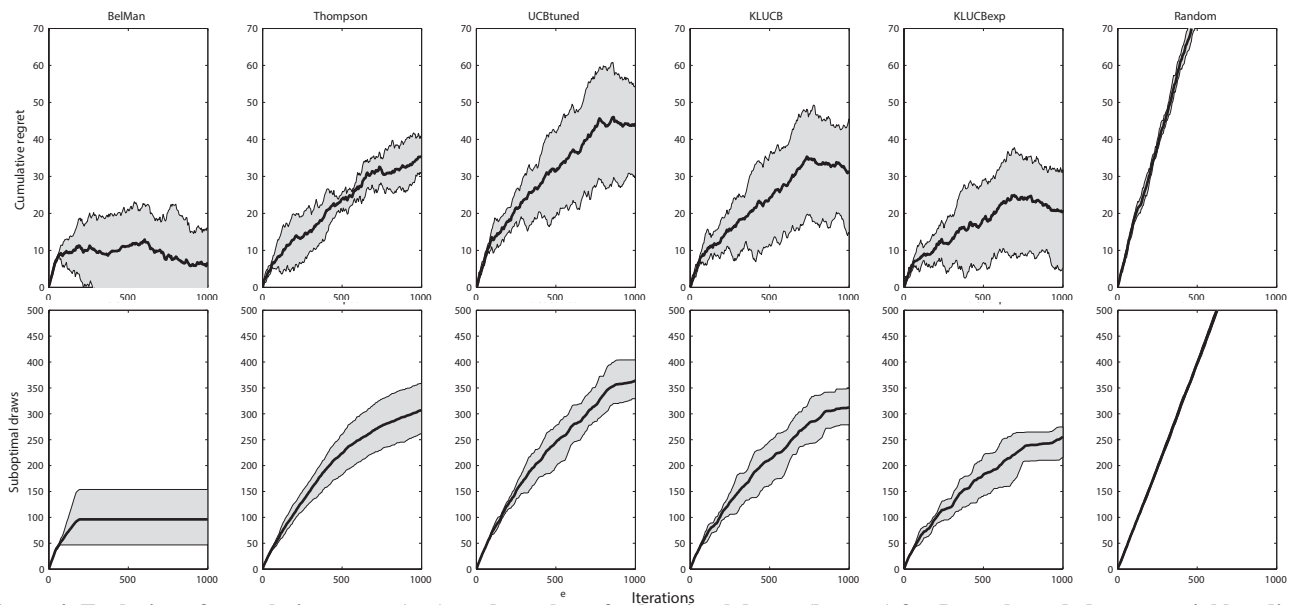
We have also run the experiments 50 times with horizon=50000 for the 20 arms to verify the asymptotic behaviour of BelMan. Figure 5 shows that BelMan's regret gradually becomes linear with respect to the logarithmic axis. Thus, Figure 5 empirically validates BelMan to reach logarithmic regret like its competitors which are theoretically proven to reach logarithmic regret bound.



**Figure 2: Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 2-arm Bernoulli bandit. The dark black line shows the average over 25 runs. The grey area shows the 75 percentile.**

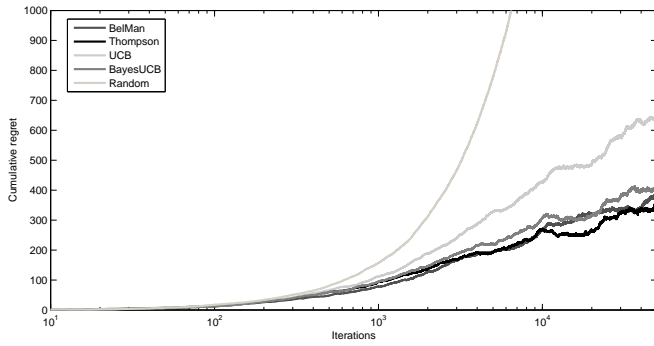


**Figure 3: Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 20-arm Bernoulli bandit.**



**Figure 4: Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 5-arm bounded exponential bandit.**



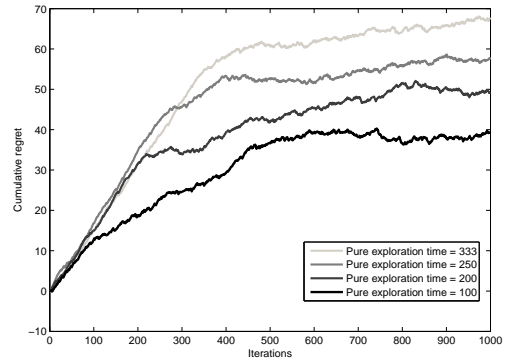


**Figure 5: Evolution of (mean) cumulative regret for exploration-exploitation 20-arm Bernoulli bandits with horizon=50,000.**

We also test BelMan on exponential bandits: 5 arms with expected rewards  $\{0.2, 0.25, 0.33, 0.5, 1.0\}$ . Figure 4 shows that BelMan performs more efficiently than state-of-the-art methods for exponential distributions: Thompson sampling, UCBtuned [3], KL-UCB, and KL-UCB-exp, a method tailored for exponential distribution of rewards [18]. This validates BelMan’s applicability to different reward structures.

*Two-Phase reinforcement learning.* Two-phase reinforcement learning problems append the pure exploration and the exploration-exploitation problems together. In this family of problems, the agent gets a first phase of a given window for pure exploration. In this phase, the agent collects more information about the underlying reward distributions. Following this, the agent goes through the exploration-exploitation phase. In this phase, it solves the exploration-exploitation problem and focuses on maximising the cumulative reward. This setup is perceivable as an initial online model building or ‘training’ phase followed by an online problem solving or ‘testing’ phase. This problem setup often emerges in applications [16] where the decision maker explores for an initial phase to create a knowledge base and another phase to take decisions by leveraging this pre-build knowledge base. Thus, two-phase reinforcement learning gives us a middle ground between model-free and model-dependent approaches in decision making which is often the path taken by a practitioner.

Formally, this knowledge base is a prior distribution built from the agent’s experience. Since Bayesian methods naturally accommodate and leverage prior distributions, our Bayesian formulation adopts this problem without any modification. [32] approached this problem with a technique amalgamating a sampling technique, PSPE, and an extension of Thompson sampling, PSRL [31], for episodic fixed horizon Markov decision processes (MDPs) [14]. PSPE uses Bayesian update to create a posterior distribution for the reward distribution of a policy. Then, it samples from the distribution in order to evaluate the policies. These two steps are performed iteratively for the initial pure exploration phase. PSRL [31] is an extension of Thompson sampling for episodic MDPs. Unlike Thompson sampling, they also use Markov chain Monte Carlo method for creating the posteriors corresponding to each of the policies. Though the amalgamation of these two methods for the two phase problems in episodic MDPs perform reasonably, they lack a reasonable unified structure attacking the problem and a natural cause to pipeline them.



**Figure 6: Evolution of (mean) cumulative regret for two-phase 20-arm Bernoulli bandits.**

We approach the two-phase reinforcement learning from this point. Following the trend of this paper, we focus on the two phase setup of multi-armed bandits. Since our framework tackles both the pure exploration and the exploration-exploitation problems, and stands on a Bayesian framework inherently leveraging prior distributions, it stands as a legitimate candidate to address the two phase problem. The two phase algorithm is exactly BelMan (Algorithm 1) with  $\frac{1}{\tau(n)} = 0$  for an initial phase of length  $T$  followed by the increasing function of  $n$  previously indicated. Thus, BelMan gives us a single algorithm for three setup of bandit problem – pure exploration, exploration-exploitation, and two-phase learning. We only have to choose a different  $\tau(n)$  depending on the problem we want to address. This also supports BelMan’s claim as a generalised, unified framework for bandit problems.

In this experiment, we simulate a two-phase setup, as in [32]: the agent first does pure exploration for a fixed number of iterations, then move to exploration-exploitation. This is possible since Belman supports both modes and can transparently switch. The setting is that of the 20-arm Bernoulli from Figure 3.

We observe a sharp phase transition in Figure 6. While the pure exploration version acts in the designated window length, it explores almost uniformly to gain more information about the reward distributions. We know for such pure exploration the cumulative regret grows linearly with iterations. Following this, the growth of cumulative regret decreases and becomes sublinear. If we also compare it with the initial growth in cumulative regret and suboptimal draws of BelMan in Figure 3, we observe that the regret for the exploration-exploitation phase is less than that of regular BelMan exploration-exploitation. Also, with increase in the window length the phase transition becomes sharper as the growth in regret becomes very small. In brief, there are two major lessons of this experiment. First, Bayesian methods provide an inherent advantage in leveraging *a priori* knowledge (here, from the first phase). Second, a pure exploration phase helps in improving the performance during the exploration-exploitation phase.

## 5 CONCLUSION

BelMan is a generic Bayesian approach for solving pure exploration, exploration-exploitation, and two-phase bandit problems. BelMan, when instantiated to rewards modelled by any distribution of the exponential family, conveniently leads to analytical forms



that allow to derive a well-defined and unique projection as well as to devise an effective and fast computation. BelMan is asymptotically consistent. Proof of consistency indicates that growth of exposure transforms the exploration–exploitation duel into pure exploitation after accumulating large enough samples. We empirically, with Bernoulli and exponential distributions, and comparatively, with the state-of-the-art bandit algorithms, show that BelMan is not only competitive but can also be leading on specific problem instances involving many arms and continuous rewards. Experiments validate that BelMan asymptotically achieves logarithmic regret. BelMan is a Bayesian information geometric approach able to incorporate *a priori* knowledge. Experiments for two-phase reinforcement learning problem explicate that BelMan not only spontaneously adapts with but also leverages explored information to escalate performance efficiency.

We are now trying and determining the form that exposure has to satisfy as a time dependent function. We are investigating the asymptotic efficiency and stability of BelMan. We are also investigating how BelMan can be extended to address the exploration–exploitation dilemma for bandit problems involving non-parametric domain, continuous or dependent arms, and problems as general as MDPs.

## REFERENCES

- [1] Shun-Ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191 of *Translations of mathematical monographs*. American Mathematical Society, 2007.
- [2] Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.
- [3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2–3):235–256, 2002.
- [4] Richard Bellman. A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics (1933–1960)*, 16(3/4):221–229, 1956.
- [5] Jose M Bernardo. Algorithm AS 103: Psi (digamma) function. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 25(3):315–317, 1976.
- [6] L. D. Brown. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, 1986.
- [7] Sébastien Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *ICML*, pages 258–265, 2013.
- [8] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *ALT*, pages 23–37. Springer, 2009.
- [9] Olivier Cappé, Aurelien Garivier, and Emilie Kaufmann. *pymaBandits*, 2012. <http://mloss.org/software/view/415/>.
- [10] Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *NIPS*, pages 379–387, 2014.
- [11] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [12] I Csiszár and Gábor Tusnády. Information geometry and alternating minimization procedures. *Statistics and decisions*, Supplement issue No. 1:205–237, 1984.
- [13] Imre Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, 12(3):768–793, 1984.
- [14] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *NIPS*, pages 2818–2826, 2015.
- [15] Morris H DeGroot. *Optimal statistical decisions*, volume 82 of *Wiley Classics Library*. John Wiley & Sons, 2005.
- [16] Muhammad Faheem and Pierre Senellart. Adaptive web crawling through structure-based link classification. In *Proc. ICADL*, pages 39–51, Seoul, South Korea, December 2015.
- [17] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *NIPS*, pages 3212–3220, 2012.
- [18] Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *COLT*, pages 359–376, 2011.
- [19] Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *arXiv preprint arXiv:1602.07182*, 2016.
- [20] John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.
- [21] Chien-Ju Ho and Jennifer Wortman Vaughan. Online task assignment in crowd-sourcing markets. In *AAAI*, pages 45–51, 2012.
- [22] Edwin T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4:227–241, 1968.
- [23] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On Bayesian upper confidence bounds for bandit problems. In *AISTATS*, pages 592–600, 2012.
- [24] Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *COLT*, pages 228–251, 2013.
- [25] Jaya Kawale, Hung H Bui, Branislav Kveton, Long Tran-Thanh, and Sanjay Chawla. Efficient Thompson sampling for online matrix-factorization recommendation. In *NIPS*, pages 1297–1305, 2015.
- [26] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.
- [27] T. L. Lai. Asymptotic solutions of bandit problems. In Wendell Fleming and Pierre-Louis Lions, editors, *Stochastic differential systems, stochastic control theory and applications*, pages 275–292. Springer, 1988.
- [28] T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, March 1985.
- [29] William G Macready and David H Wolpert. Bandit problems and the exploration/exploitation tradeoff. *IEEE Transactions on evolutionary computation*, 2(1):2–22, 1998.
- [30] José Nino-Mora. Computing a classic index for finite-horizon bandits. *INFORMS Journal on Computing*, 23(2):254–267, 2011.
- [31] Ian Osband, Dan Russo, and Benjamin Van Roy. (More) efficient reinforcement learning via posterior sampling. In *NIPS*, pages 3003–3011, 2013.
- [32] Sudeep Raja Putta and Theja Tulabandhula. Pure exploration in episodic fixed-horizon Markov decision processes. In *AAMAS*, pages 1703–1704, 2017.
- [33] Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5):527–535, 09 1952.
- [34] Steven L Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.
- [35] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.
- [36] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3–4):285, 1933.
- [37] Sofia S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

## A SUPPLEMENTARY MATERIAL FOR SECTION 2 (PROBLEM FORMULATION)

*Exploration–exploitation bandit.* The Lai–Robbins bound can be mathematically formulated, for policies with  $R_n(\theta) = o(n^a)$ , as

$$\liminf_{n \rightarrow \infty} \frac{R_n(\theta)}{\log n} \geq \frac{\sum_{j: \mu^*(\theta) > \mu(\theta_j)} [\mu^*(\theta) - \mu(\theta_j)]}{\inf_j D_{\text{KL}}(f_{\theta_j}(x) \parallel f_{\theta^*}(x))}, \quad (7)$$

where  $f_{\theta^*}(x)$  is the reward distribution of the optimal arm. This states that the best we can achieve is a logarithmic growth of cumulative regret. It also implies that this optimality is harder to achieve as the minimal KL-divergence between the optimal arm and any other arm decreases. This is intuitive because in such scenario the agent has to explore these two arms more to distinguish between them and to choose the optimal arm. [28] have also shown policies satisfying this lower bound will also satisfy the asymptotic property of Equation (1). [28] also showed that for specific families of distributions, the expected number of draws of any suboptimal arm  $j$  satisfies

$$T_n^j(\theta) \leq \left( \frac{1}{\inf_j D_{\text{KL}}(f_{\theta_j}(x) \parallel f_{\theta^*}(x))} + o(1) \right) \log n. \quad (8)$$

Equation (7) and (8) together claim that the best achievable number of draws of suboptimal arms is  $\Theta(\log n)$ .

Another variation of the Bayesian formulation was introduced by [4] with a discounted reward setting. Unlike  $S_n$ , the discounted sum of rewards  $D_\gamma \triangleq \sum_{j=0}^{\infty} [\gamma^j x_{j+1}]$  ensures convergence of the sequential sum for  $\gamma \in [0, 1)$ . Intuitively, the discounted sum implies the effect of an action decay with each time step by the discount factor  $\gamma$ . This setting assumes  $k$  independent priors on each of the arms and also models the process of choosing the next arm as a Markov process. Thus, we can reformulate the bandit problem as maximising

$$\int \dots \int \mathbb{E}_\theta [D_\gamma] db^1(\theta_1) \dots db^k(\theta_k)$$

where,  $b^j$  is the independent prior distribution on  $\theta_j$  for  $j = 1, \dots, k$ . [20] showed the agent can have an optimally indexed policy by sampling from the arm with largest Gittins index

$$G^j(s^j) \triangleq \sup_{\tau > 0} \frac{\mathbb{E} \left[ \sum_{n=0}^{\tau} \gamma^n x^j(S_n^j) \mid S_0^j = s^j \right]}{\mathbb{E} \left[ \sum_{n=0}^{\tau-1} \gamma^n \mid S_0^j = s^j \right]}$$

where  $s^j$  is the state of arm  $j$  and  $\tau$  is referred to as the stopping time i.e. the first time when the index is no greater than its initial value.

*Pure exploration bandit.* [2] identified the pure exploration problem as *best arm identification* and proposed the Successive Rejects algorithm under fixed budget constraints. [7] extended this algorithm for finding  $m$ -best arms and proposed the Successive Accepts and Rejects algorithm. In another endeavour to adapt the UCB family to pure exploration scenario, the LUCB family of frequentist algorithms are proposed [17, 24]. In the beginning, they sample all the arms. Following that, they sample both the arm with maximum expected reward and the one with maximum upper-confidence bound till the algorithm can identify each of them separately.

## B SUPPLEMENTARY MATERIAL FOR SECTION 3 (METHODOLOGY)

### B.1 KL-divergence on the Manifold.

Kullback-Liebler divergence (or KL-divergence) [26] is a premetric measure of dissimilarity between two probability distributions.

*Definition 1 (KL-divergence).* If there exist two probability measures  $P$  and  $Q$  defined over a set  $S$  and  $P$  is absolutely continuous with respect to  $Q$ , we define the KL-divergence between them as

$$D_{\text{KL}}(P \parallel Q) \triangleq \int_S \log \frac{dP}{dQ} dP.$$

$\frac{dP}{dQ}$  is the Radon-Nikodym derivative of  $P$  with respect to  $Q$ .

Since it represents the expected information lost if  $P$  is encoded using  $Q$ , it is also called *relative entropy*. Depending on the applications,  $P$  acts as the representative of ‘true’ underlying distribution obtained from observations or data or natural law, and  $Q$  represents the model or approximation of  $P$ . For two probability density functions  $p(s)$  and  $q(s)$  defined over a set  $S$ , the KL-divergence can be rewritten as

$$D_{\text{KL}}(p(s) \parallel q(s)) = \int_{s \in S} p(s) \log \frac{p(s)}{q(s)} ds = -h(p(s)) + H(p(s), q(s)). \quad (9)$$

Here,  $h(p(s))$  is entropy of  $p$  and  $H(p(s), q(s))$  is the mutual information between  $p$  and  $q$ . Thus, from an information-theoretic perspective, we perceive KL-divergence as the natural divergence function on the belief-reward manifold when we analyse the dynamics of the entropy

function on it. Except that, any general  $\alpha$ -divergence function on the statistical manifold is a convex combination of  $\pm 1$ -divergences. Mathematically, for  $\alpha \in (-1, +1)$ ,

$$\begin{aligned} D^{(\alpha)}(p \parallel q) &\triangleq \frac{1+\alpha}{2} D^{(+1)}(p \parallel q) + \frac{1-\alpha}{2} D^{(-1)}(p \parallel q) \\ &= \frac{1+\alpha}{2} D_{\text{KL}}(q \parallel p) + \frac{1-\alpha}{2} DD_{\text{KL}}(p \parallel q). \end{aligned} \quad (10)$$

From a manifold perspective, it seems that the divergence function for the  $\pm 1$ -connections on the belief-reward manifolds and a convex mixture of  $D_{\text{KL}}$  divergences form the general notion of movement on any such space. Thus, KL-divergence between two belief-reward distributions is an effective and natural quantifier of movement, and also of information accumulation during Bayesian update. Hence, for updating the beliefs in an almost optimal manner to decrease the uncertainty and to improve the reward, we have to express the observations and a representation of knowledge-base and exploiting scheme using the belief-reward distributions, and to minimise the KL-divergence between these distributions with each iteration. This allows us to conclude our query about the notion of optimal accumulation of information for precise estimation of arms' distributions is by mapping them into KL-divergence and alternately minimizing it. If  $P$  are the candidate belief-reward distributions of the arms formed by accumulation of actions and rewards, and  $Q$  are the pseudobelief or pseudobelief-focal distributions, the alternating minimisation scheme looks for the most succinct representation  $Q$  of the knowledge or the exploitation bias while choosing such arms whose belief-reward distributions resemble their true reward distributions as much as possible.

## B.2 Condition for Existence of Alternating Projection Scheme

Both I- and rI-projections are valid and well-defined if the KL-divergence between any two distributions in  $\mathcal{P}$  and  $\mathcal{Q}$  is defined and finite.

ASSUMPTION 3 ((ABSENCE OF SINGULARITIES)). *The distribution families  $\mathcal{P}$  and  $\mathcal{Q}$  are defined over the sets  $\text{Supp}(\mathcal{P}) \triangleq \{a : p(a) > 0, \forall p \in \mathcal{P}\}$  and  $\text{Supp}(\mathcal{Q}) \triangleq \{a : q(a) > 0, \forall p \in \mathcal{P}\}$  respectively. Moreover, none of the supports are empty and  $\text{Supp}(\mathcal{P}) \subseteq \text{Supp}(\mathcal{Q})$ .*

Assumption 3 avoids any singularity in both I- and reverse I- projections and keep them finite.

## B.3 BelMan for Exponential Family Distributions

*Bernoulli Bandits.* In the case of Bernoulli bandits, we assume that drawing an arm returns the rewards 1 and 0 with probability  $\theta$  and  $1 - \theta$  respectively. Thus, the reward distribution of the  $j^{\text{th}}$  arm is  $f_{\theta_j}(x) \triangleq \text{Ber}(\theta_j)$ . Following the Bayesian approach, we choose the conjugate prior to begin with. Thus, we keep the prior belief over each arm as a beta distribution with shape parameters  $\{\alpha^j\}_{j=1}^k$  and  $\{\beta^j\}_{j=1}^k$ . After  $n$ -iterations the prior over the probability of success of the  $j^{\text{th}}$  arm is

$$b_n^j(\theta_j) \triangleq \text{Beta}(\theta_j; \alpha_n^j, \beta_n^j) = \frac{1}{B(\alpha_n^j, \beta_n^j)} \theta_j^{\alpha_n^j - 1} (1 - \theta_j)^{\beta_n^j - 1},$$

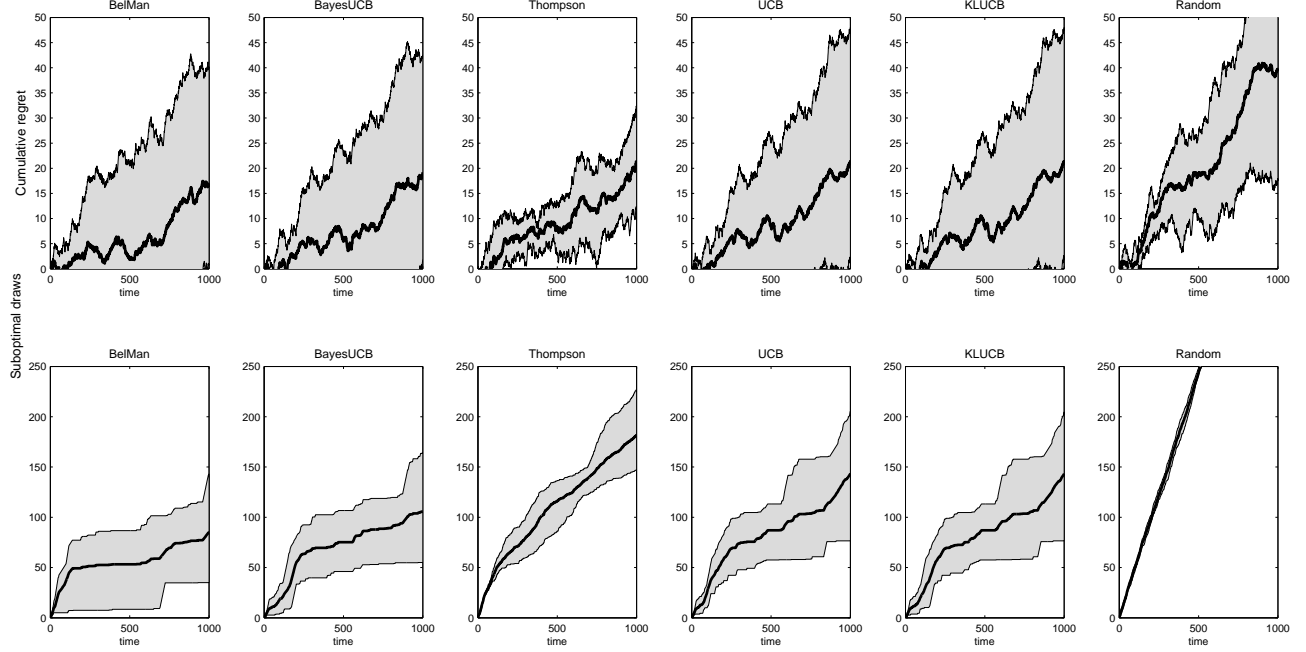
for  $\alpha_n^j, \beta_n^j > 0$  and  $\theta_j \in (0, 1)$ . Here,  $\alpha_n^j$  and  $\beta_n^j$  are the number of successes and failures, respectively, for the arm  $j$  till iteration  $n$ . We begin with both  $\alpha_0^j$  and  $\beta_0^j$  to be 1 for all arms. This amounts to the uniform distribution over 0 and 1. This initialization allows us to choose all the arms with equal probability and without any initial bias. We update this belief eventually as we further draw the arms and compute it using BelMan. Under this specific setting of beta prior and Bernoulli reward, we compute the targeted KL-divergence of BelMan as

$$\begin{aligned} \sum_{j=1}^k D_{\text{KL}} \left( \mathbb{P}_n^j(x, \theta) \parallel \bar{\mathbb{Q}}_{n-1}(x, \theta) \right) &= \sum_{j=1}^k \left[ -\frac{1}{\tau(n)} \frac{\alpha_n^j}{N_n^j} - \log \left( B \left( \alpha_n^j, \beta_n^j \right) \right) + (\alpha_n^j - \bar{\alpha}_{n-1}) \Psi(\alpha_n^j) + (\beta_n^j - \bar{\beta}_{n-1}) \Psi(\beta_n^j) \right. \\ &\quad \left. - (N_n^j - \bar{N}_{n-1}) \Psi(N_n^j) \right] + k \log \left( \frac{\bar{\alpha}_{n-1} \exp\left(\frac{1}{\tau(n)}\right) + \bar{\beta}_{n-1}}{\bar{N}_{n-1}} \right) + k \log \left( B \left( \bar{\alpha}_{n-1}, \bar{\beta}_{n-1} \right) \right). \end{aligned}$$

Here,  $N_n^j = \alpha_n^j + \beta_n^j$  is the total number of times the  $j^{\text{th}}$  arm is played till the  $n^{\text{th}}$  iteration,  $\bar{N} = \bar{\alpha} + \bar{\beta}$  and  $\Psi$  is the digamma function [5] defined as the derivative of the logarithm of gamma function, i.e.  $\frac{d}{da} (\log \Gamma(a))$ .

In Line 4 of Algorithm 1, we first perform the I-projection to decide which arm  $a_n$  to draw to minimize the KL-divergence. Following this, we update the pseudobelief using I-projection in Line 9 of Algorithm 1. In order to perform this update, we find out such  $\bar{\alpha}$  and  $\bar{\beta}$  that minimize the objective and update the pseudobelief accordingly. The presence of pseudobelief offers BelMan a chance to explore the less successful arms to minimize the entropy, while the Focal distribution creates the scope of exploiting the present information of the best arm.

*Exponential Bandits.* The *exponential distribution* is another member of the exponential family. For a given positive *rate parameter*  $\theta_j$ , the reward distribution of arm  $j$  of exponential bandit is  $f_{\theta_j}(x) \triangleq \theta_j \exp(-\theta_j x)$  for  $x \in [0, \infty)$ . Following the structure of Sections 3.3 and the previous Bernoulli case, we obtain the gamma distribution, another member of the exponential family, as the conjugate prior. After the  $n^{\text{th}}$



**Figure 7: Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 500 iterations for 2-arm Bernoulli bandit with means 0.45 and 0.55. The dark line shows the average over 25 runs. The grey area shows 75 percentile.**

iteration, the belief distribution corresponding to  $j^{\text{th}}$  arm is expressed as

$$b_n^j(\theta_j) \triangleq \text{Gamma}(\theta_j; \alpha_n^j, \beta_n^j) = \frac{\beta_n^j \alpha_n^j}{\Gamma(\alpha_n^j)} \theta_j^{\alpha_n^j - 1} \exp(-\theta_j \beta_n^j),$$

for both shape and rate parameters  $\alpha_n^j, \beta_n^j > 0$ . Here,  $\alpha_n^j$  and  $\beta_n^j$  are, respectively, the number of times the arm  $j$  is played and sum of the rewards obtained by playing the arm till iteration  $n$ . As we update using Equation (2), we get gamma distributions with parameters  $\alpha_{n+1}^j = \alpha_n^j + 1$ , and  $\beta_{n+1}^j = \beta_n^j + x_n$  if the arm  $j$  is played and a reward  $x_n$  is obtained. Under this specific setting of gamma prior and exponential reward, we compute the targeted KL-divergence of BelMan as

$$\begin{aligned} \sum_{j=1}^k D_{\text{KL}} \left( \mathbb{P}_n^j(x, \theta) \parallel \bar{\mathbb{Q}}(x, \theta) \right) &= \sum_{j=1}^k \left[ -\frac{1}{\tau(n)} \frac{\alpha_n^j}{\beta_n^j} - \log \left( \Gamma \left( \alpha_n^j \right) \right) + (\alpha_n^j - \bar{\alpha}_{n-1}) \Psi(\alpha_n^j) - \frac{\alpha_n^j}{\beta_n^j} (\beta_n^j - \bar{\beta}_{n-1}) \right. \\ &\quad \left. + \bar{\alpha}_{n-1} \log \beta_n^j \right] + k \log \bar{Z}_n + k \log \left( \Gamma \left( \bar{\alpha}_{n-1} \right) \right) - k \bar{\alpha}_{n-1} \log \bar{\beta}_{n-1}. \end{aligned}$$

We incorporate this analytical form in Algorithm 1 and update it as mentioned in the Bernoulli case.

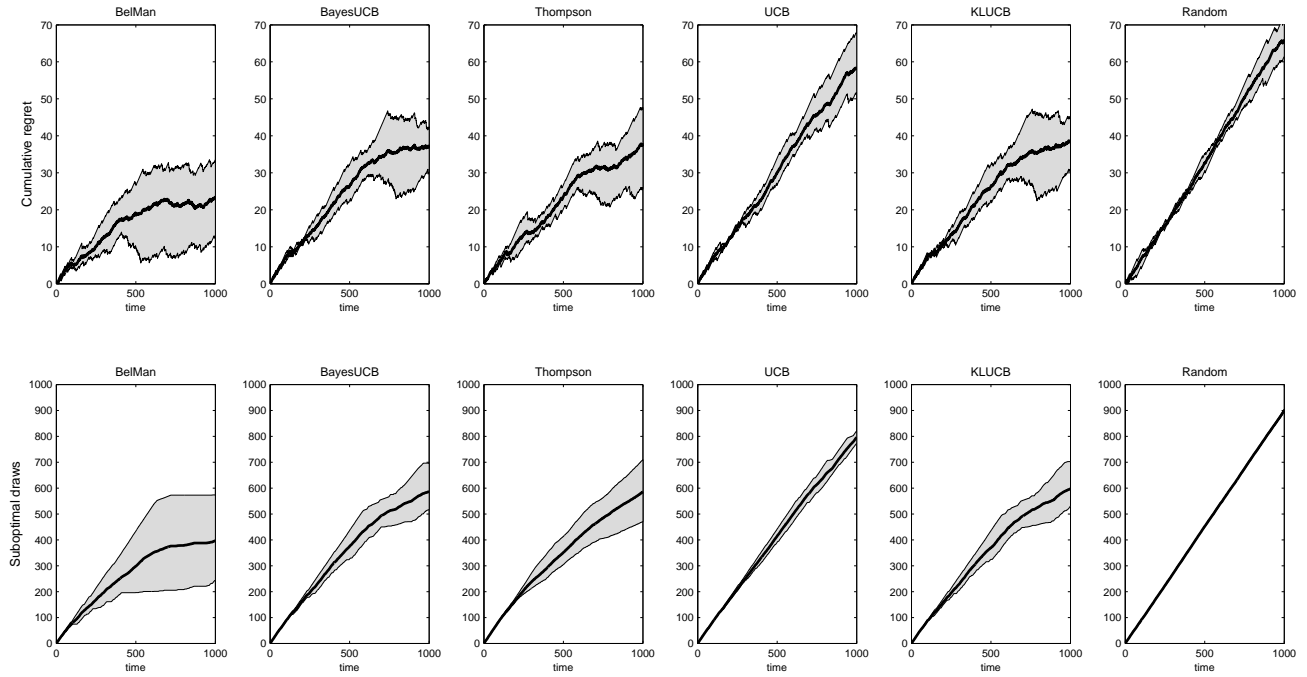
## C SUPPLEMENTARY MATERIAL FOR SECTION 4 (EXPERIMENTAL EVALUATION)

### C.1 Exploration–Exploitation Setup.

We also experimented on another 2-arm bandit scenario with means 0.45 and 0.55. Figures 7 depicts the evolution of cumulative regret and suboptimal draws for BelMan and the other competing algorithms. Similar to Figure 7, we observe the cumulative regret of BelMan grows at first linearly and then it transits to a state of slow growth. Except showing this ideal behaviour, BelMan performs competitively with the contending algorithms. This shows its efficiency as a candidate solution to the exploration–exploitation bandit.

Figure 8 shows performance for 10-arm Bernoulli bandit. For this setup, BelMan outperforms other algorithms. We also observe though the number of arms increases from Figure 7 to Figure 8 that performance of all algorithms is comparatively better in the first case. This is explainable from the fact that hardness of minimising cumulative regret increases as the number of arms increases. Beside that, as more arms with identical or almost identical distributions appear, the algorithm requires more exploration to separate them and to determine which one is optimal. The difference in performance between Figure 7 and 2 indicates this.

We finally tested BelMan on an exponential bandit consisting of 5-arms with expected rewards  $\{0.2, 0.25, 0.33, 0.5, 1.0\}$ . We compare performance of BelMan with state-of-the-art frequentist method tailored for exponential distribution of rewards, called KL-UCBExp [18].



**Figure 8: Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 500 iterations for 10-arm Bernoulli bandit with means  $\{0.1, 0.05, 0.05, 0.05, 0.02, 0.02, 0.02, 0.01, 0.01, 0.01\}$ . The dark black line shows the average. The grey area shows 75 percentile.**

We also compare it with Thompson sampling, UCBtuned and uniform sampling method (Random). The results are shown in Figure 9 and 10. Since the formulation is oblivious to boundedness of the distribution, we choose to validate also on unbounded rewards. In Figure 9, it outperforms all the other algorithms. In Figure 10, though KL-UCBexp performs the best, performance of BelMan is still competitive with it. These results validate BelMan’s claim as a generic solution to a wide range of bandit problems.

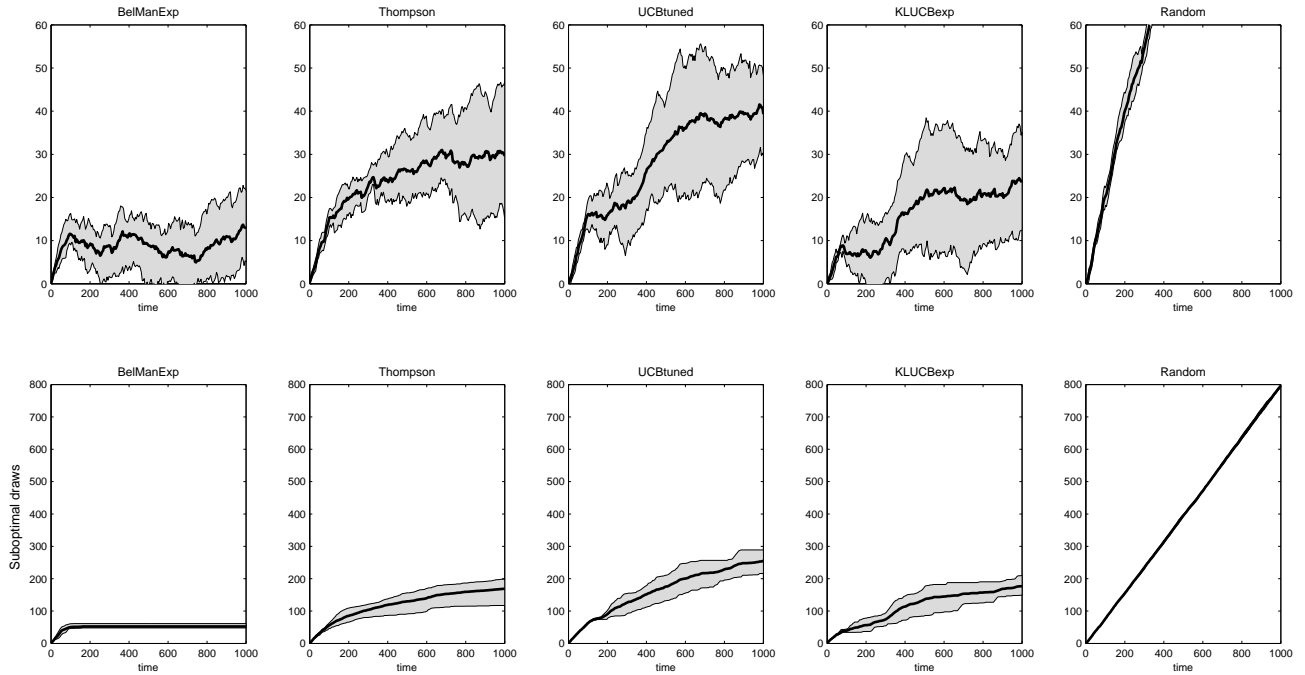


Figure 9: Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 1000 iterations for 5-arm unbounded exponential bandit with parameters  $\{0.2, 0.25, 0.33, 0.5, 1.0\}$ .

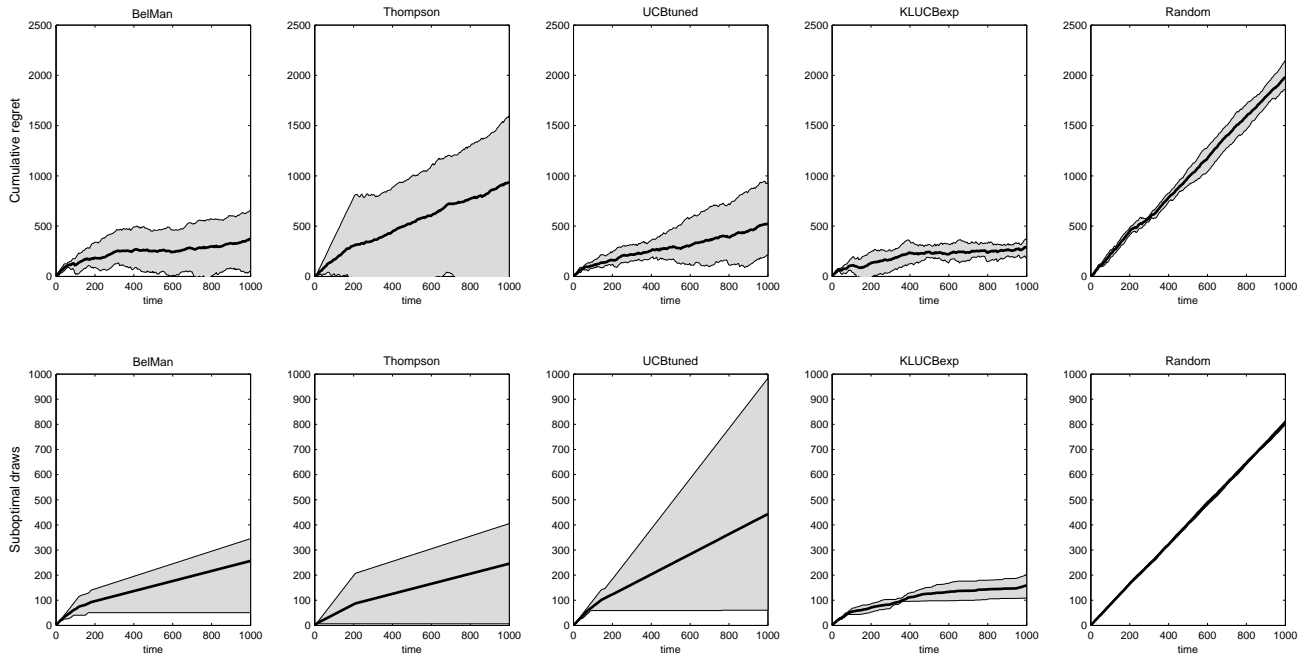


Figure 10: Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 1000 iterations for 5-arm unbounded exponential bandit with parameters  $\{1, 2, 3, 4, 5\}$ .