BelMan: An Information-Geometric Approach to Stochastic Bandits

Debabrota Basu[1], Pierre Senellart[2,3,4], and Stéphane Bressan[5]

[1] Data Science and AI Division, Chalmers University of Technology, Göteborg, Sweden
[2] DI ENS, ENS, CNRS, PSL University, Paris, France
[3] Inria, Paris, France
[4] LTCI, Télécom ParisTech, Paris, France
[5] School of Computing, National University of Singapore, Singapore

Abstract. We propose a Bayesian information-geometric approach to the exploration–exploitation trade-off in stochastic multi-armed bandits. The uncertainty on reward generation and belief is represented using the manifold of joint distributions of rewards and beliefs. Accumulated information is summarised by the barycentre of joint distributions, the *pseudobelief-reward*. While the pseudobelief-reward facilitates information accumulation through exploration, another mechanism is needed to increase exploitation by gradually focusing on higher rewards, the *pseudobelief-focal-reward*. Our resulting algorithm, BelMan, alternates between projection of the pseudobelief-focal-reward onto belief-reward distributions to choose the arm to play, and projection of the updated belief-reward distributions onto the pseudobelief-focal-reward. We prove BelMan to be asymptotically optimal and to incur a sublinear regret growth. We instantiate BelMan to stochastic bandits with Bernoulli and exponential rewards, and to a real-life application of scheduling queueing bandits. Comparative evaluation with the state of the art shows that BelMan is not only competitive for Bernoulli bandits but in many cases also outperforms other approaches for exponential and queueing bandits.

1 Introduction

The *multi-armed bandit* problem [30] is a sequential decision-making problem [11] in which a gambler plays a set of arms to obtain a sequence of rewards. In the *stochastic bandit* problem [7], the rewards are obtained from reward distributions on arms. These reward distributions belong to the same family of distributions but vary in the parameters. These parameters are unknown to the gambler. In the classical setting, the gambler devises a strategy, choosing a sequence of arm draws, that maximises the *expected cumulative reward* [30]. In an equivalent formulation, the gambler devises a strategy that minimises the *expected cumulative regret* [26], that is the expected cumulative deficit of reward caused by the gambler not always playing the optimal arm. In order to achieve this goal, the gambler must simultaneously learn the parameters of the reward distributions of arms. Thus, solving the stochastic bandit problem consists in devising strategies that combine both the accumulation of information to reduce the uncertainty of decision

making, *exploration*, and the accumulation of rewards, *exploitation* [27]. We refer to the stochastic bandit problem as the *exploration–exploitation bandit* problem to highlight this trade-off. If a strategy relies on independent phases of exploration and exploitation, it necessarily yields a suboptimal regret bound [15]. The gambler has to adaptively balance and intertwine exploration and exploitation [3].

In a variant of the stochastic bandit problem, called the *pure exploration bandit* problem [8], the goal of the gambler is solely to accumulate information about the arms. In another variant of the stochastic bandit problem, the gambler interacts with the bandit in two consecutive phases of pure exploration and exploration–exploitation. The authors of [29] named this variant the *two-phase reinforcement learning* problem.

Although frequentist algorithms with optimism in the face of uncertainty such as UCB [3] and KL-UCB [14] work considerably well for the exploration–exploitation bandit problem, their frequentist nature prevents effective assimilation of a priori knowledge about the reward distributions of the arms [23]. Bayesian algorithms for the exploration–exploitation problem, such as Thompson sampling [34] and Bayes-UCB [21], leverage a prior distribution that summarises a priori knowledge. However, as argued in [22], there is a need for Bayesian algorithms that also cater for pure exploration. Neither Thompson sampling nor Bayes-UCB are able to do so.

Our contribution. We propose a unified Bayesian approach to address the exploration–exploitation, pure exploration, and two-phase reinforcement learning problems. We address these problems from the perspective of information representation, information accumulation, and balanced induction of bias. Here, the uncertainty is two fold. Sampling reward from the reward distributions is inherently stochastic. The other layer is due to the incomplete information about the true parameters of the reward distributions. Following Bayesian algorithms [34], we maintain a parameterised *belief* distribution for each arm representing the uncertainty on the parameter of its reward distribution. Extending this representation, we use a joint distribution to express the two-fold uncertainty induced by both the belief and the reward distributions of each arm. We refer to these joint distributions as the *belief-reward distributions* of the arms. We set the learning problem in the statistical manifold [2] of the belief-reward distributions, which we call the *belief-reward manifold*. The belief-reward manifold provides a representation for controlling pure exploration and exploration–exploitation, and to design a unifying algorithmic framework.

The authors of [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 drives us to first devise a pure exploration algorithm, which requires a collective representation of the accumulated knowledge about the arm. From an information-geometric point of view [4,1], the barycentre of the belief-reward distributions in the belief-reward manifold serves as a succinct summary. We refer to this barycentre as the *pseudobelief-reward*. We prove the pseudobelief-reward to be a unique representation in the manifold. Though pseudobelief-reward facilitates the accumulation

of knowledge, it is essential for the exploration–exploitation bandit problem to also incorporate a mechanism that gradually concentrates on higher rewards [27]. We introduce a distribution that induces such an increasing exploitative bias. We refer to this distribution as the *focal distribution*. We incorporate it into the definition of the pseudobelief-reward distribution to construct the *pseudobelief-focal-reward distribution*. This pushes the summarised representation towards the arms having higher expected rewards. We implement the focal distribution using an exponential function of the form $\exp(X/\tau(t))$, where X is the reward, and a parameter $\tau(t)$ dependent on time t and named as *exposure*. Exposure controls the exploration–exploitation trade-off.

In Section 2, we apply these information-geometric constructions to develop the BelMan algorithm. BelMan projects the pseudobelief-focal-reward onto belief-rewards to select an arm. As it is played and a reward is collected, BelMan updates the belief-reward distribution of the corresponding arm by projecting of the updated belief-reward distributions onto the pseudobelief-focal-reward. In terms of information geometry, these two projections are studied as information (I-) and reverse information (rI-) projections [10], respectively. BelMan alternates I- and rI-projections between belief-reward distributions of the arms and the pseudobelief-focal-reward distribution for arm selection and information accumulation. We prove the law of convergence of the pseudobelief-focal-reward distribution for BelMan, and that BelMan asymptotically converges to the choice of the optimal arm. BelMan can be tuned, using the exposure, to support a continuum from pure exploration to exploration–exploitation, as well as two-phase reinforcement learning.

We instantiate BelMan for distributions of the exponential family [6]. These distributions lead to analytical forms that allow derivation of well-defined and unique I- and rI-projections as well as to devise an effective and fast computation. In Section 3, we empirically evaluate the performance of BelMan on different sets of arms and parameters for Bernoulli and exponential distributions, thus showing its applicability to both discrete and continuous rewards. Experimental results validate that BelMan asymptotically achieves logarithmic regret. We compare BelMan with state-of-the-art algorithms: UCB [3], KL-UCB, KL-UCB-Exp [14], Bayes-UCB [21], Thompson sampling [34], and Gittins index [17], in these different settings. Results demonstrate that BelMan is not only competitive but also outperforms existing algorithms for challenging setups such as those involving many arms and continuous rewards. For two-phase reinforcement learning, results show that BelMan spontaneously adapts to the explored information, improving the efficiency.

We also instantiate BelMan to the application of queueing bandits [24]. Queueing bandits represent the problem of scheduling jobs in a multi-server queueing system with unknown service rates. The goal of the corresponding scheduling algorithm is to minimise the number of jobs in hold while also learning the service rates. A comparative performance evaluation for queueing systems with Bernoulli service rates show that BelMan performs significantly better than the existing algorithms, such as Q-UCB, Q-ThS [24], and Thompson sampling.

An extended discussion of the related work, full proofs, and additional experiments are available in a supplementary material.

2 Methodology

Bandit Problem. We consider a finite number $K > 1$ of independent arms. An arm a corresponds to a reward distribution $f_\theta^a(X)$. We assume that the form of the probability distribution $f_\cdot(X)$ is known to the algorithm but the parameterisation $\theta \in \Theta$ is unknown. We assume the reward distributions of all arms to be identical in form but to vary over the parametrisation θ. Thus, we refer to $f_\theta^a(X)$ as $f_{\theta_a}(X)$ for specificity. The agent sequentially chooses an arm a_t at each timestep t that generates a sequence of rewards $[x_t]_{t=1}^T$, where $T \in \mathbb{N}$ is the time horizon. The algorithm computes a *policy* or strategy that sequentially draws a set of arms depending on its previous actions, observations and intended goal. The algorithm does not know the 'true' parameters of the arms $\{\theta_a^{\text{true}}\}_{a=1}^K$ a priori. Thus, the uncertainty over the estimated parameters $\{\theta_a\}_{a=1}^K$ is represented using a probability distribution $B(\theta_1, \ldots, \theta_K)$. We call $B(\theta_1, \ldots, \theta_K)$ the *belief distribution*. In the Bayesian approach, the algorithm starts with a prior belief distribution $B_0(\theta_1, \ldots, \theta_K)$ [19]. The actions taken and rewards obtained by the algorithm till timestep t create the history of the bandit process, $\mathcal{H}_t \triangleq [(a_1, x_1), \ldots, (a_{t-1}, x_{t-1})]$. This history $\mathcal{H}_t$ is used to sequentially update the belief distribution over the parameter vector as $B_t(\theta_1, \ldots, \theta_K) \triangleq \mathbb{P}(\theta_1, \ldots, \theta_K \mid \mathcal{H}_t)$. We define the space consisting of all such distributions over $\{\theta_a\}_{a=1}^K$ as the *belief space* $\mathcal{B}$. Following the stochastic bandit literature, we assume the arms to be independent, and perform Bayesian updates of beliefs.

Assumption 1 (Independence of Arms). *The parameters $\{\theta_a\}_{a=1}^K$ are drawn independently from K belief distributions $\{b_t^a(.)\}_{a=1}^K$, such that $B_t(\theta_1, \ldots, \theta_K) = \prod_{a=1}^K b_t^a(\theta_a) \triangleq \prod_{a=1}^K \mathbb{P}(\theta_a \mid \mathcal{H}_t)$.*

Though Assumption 1 is followed throughout this paper, we note it is not essential to develop the framework BelMan relies on, though it makes calculations easier.

Assumption 2 (Bayesian Evolution). *When conditioned over $\{\theta_a\}_{a=1}^K$ and the choice of arm, the sequence of rewards $[x_1, \ldots, x_t]$ is jointly independent. Thus, the Bayesian update at the t-th timestep is given by*

$$b_{t+1}^a(\theta_a) \propto f_{\theta_a}(x_t) \times b_t^a(\theta_a) \tag{1}$$

if $a_t = a$ and a reward x_t is obtained. For all other arms, the belief remains unchanged.

Belief-reward manifold. We use the joint distributions $\mathbb{P}(X, \theta)$ on reward X and parameter θ in order to represent the uncertainties of partial information about the reward distributions along with the stochastic nature of reward.

Definition 1 (Belief-reward distribution). *The joint distribution $\mathbb{P}_t^a(X, \theta)$ on reward X and parameter θ_a for the a^{th} arm at timestep t is defined as the* belief-reward distribution.

$$\mathbb{P}_t^a(X, \theta) \triangleq \frac{b_t^a(\theta) f_\theta(X)}{\int\limits_{X \in \mathbb{R}} \int\limits_{\theta \in \Theta} b_t^a(\theta) f_\theta(X) \mathrm{d}\theta \mathrm{dX}} = \frac{1}{Z} b_t^a(\theta) f_\theta(X).$$

If $f_\cdot(X)$ is a smooth function of θ_a's, the space of all reward distributions constructs a smooth statistical manifold [2], $\mathcal{R}$. We call $\mathcal{R}$ the *reward manifold*. If belief B is a smooth function of its parameters, the belief space $\mathcal{B}$ constructs another statistical manifold. We call $\mathcal{B}$ the *belief manifold* of the multi-armed bandit process. Assumption 1 implies that the belief manifold $\mathcal{B}$ is a product of K manifolds $\mathcal{B}^a \triangleq \{b^a(\theta_a)\}$. Here, $\mathcal{B}^a$ is the statistical manifold of belief distributions for the ath arm. Due to the identical parametrisation, the $\mathcal{B}^a$'s can be represented by a single manifold $\mathcal{B}_\theta$.

Lemma 1 (Belief-Reward Manifold). *The set of belief-reward distributions* $\mathbb{P}(X, \theta)$ *defines a manifold* $\mathcal{B}_\theta\mathcal{R}$*, such that* $\mathcal{B}_\theta\mathcal{R} = \mathcal{B}_\theta \times \mathcal{R}$*. We refer to it as the* belief-reward manifold*.*

The Bayesian belief update after each timestep is a movement on the belief manifold from a point b_t^a to another point b_{t+1}^a *with maximum information gain* from the obtained reward. Thus, the belief-reward distributions of the played arms evolve to create a set of trajectories on the belief-reward manifold. The goal of pure exploration is to control such trajectories collectively such that after a long enough time each of the belief-rewards accumulate enough information to resemble the 'true' reward distributions well enough. The goal of exploration–exploitation is to gain enough information about the 'true' reward distributions while increasing the cumulative reward in the path, i.e, by inducing a bias towards playing the arms with higher expected rewards.

Pseudobelief: Summary of explored knowledge. In order to control the exploration, the algorithm has to construct a summary of the collective knowledge on the belief-rewards of the arms. Since the belief-reward distribution of each arm is a point on the belief-reward manifold, geometrically their barycentre on the belief-reward manifold represents a valid summarisation of the uncertainty over all the arms [1]. Since the belief-reward manifold is a statistical manifold, we obtain from information geometry that this barycentre is the point on the manifold that minimises the sum of KL-divergences from the belief-rewards of all the arms [4,2]. We refer to this minimising belief-reward distribution as the pseudobelief-reward distribution of all the arms.

Definition 2 (Pseudobelief-reward distribution). *A* pseudobelief-reward distribution $\bar{\mathbb{P}}_t(X, \theta)$ *is a point in the belief-reward manifold that minimises the sum of KL-divergences from the belief-reward distributions* $\mathbb{P}_t^a(X, \theta)$ *of all the arms.*

$$\bar{\mathbb{P}}_t(X, \theta) \triangleq \underset{\mathbb{P} \in \mathcal{B}_\theta \mathcal{R}}{\arg\min} \sum_{a=1}^{K} D_{\mathrm{KL}} \left(\mathbb{P}_t^a(X, \theta) \,\|\, \mathbb{P}(X, \theta) \right). \tag{2}$$

We prove existence and uniqueness of the pseudobelief-reward for K given belief-reward distributions. This proves the pseudobelief-reward to be an unambiguous representative of collective knowledge. We also prove that the pseudobelief-reward distribution $\bar{\mathbb{P}}_t$ is the projection of the average belief-reward distribution $\hat{\mathbb{P}}_t(X, \theta) = \sum_a \mathbb{P}_t^a(X, \theta)$ on the belief-reward manifold. This result validates the claim of pseudobelief-reward as the summariser of the belief-rewards of all the arms.

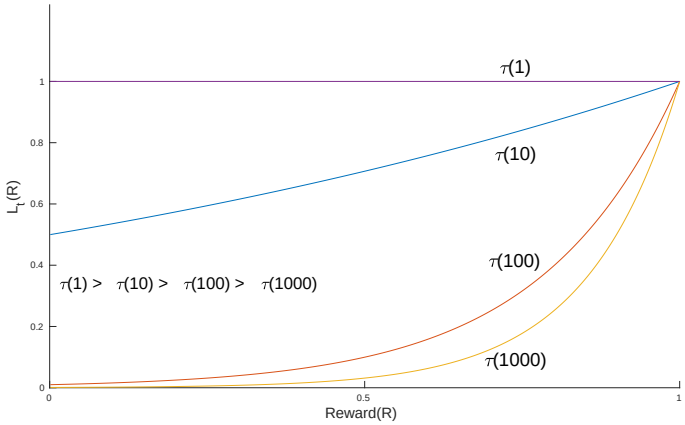


Fig. 1. Evolution of the focal distribution over $X \in [0, 1]$ for $t = 1, 10, 100$ and 1000.

Theorem 1. *For a given set of belief-reward distributions* $\{\mathbb{P}_t^a\}_{a=1}^K$ *defined on the same support set and having a finite expectation,* $\bar{\mathbb{P}}_t$ *is uniquely defined, and is such that its expectation parameter verifies* $\hat{\mu}_t(\theta) = \frac{1}{K}\sum_{a=1}^K \mu_t^a(\theta)$.

Hereby, we establish as a unique summariser of all the belief–reward distributions. Using this uniqueness proof, we can prove that the pseudobelief–reward distribution $\bar{\mathbb{P}}$ is projection of the average belief–reward distribution $\hat{\mathbb{P}}$ on the belief–reward manifold.

Corollary 1. *The pseudobelief-reward distribution* $\bar{\mathbb{P}}_t(X, \theta)$ *is the unique point on the belief-reward manifold that has minimum KL-divergence from the distribution* $\hat{\mathbb{P}}_t(X, \theta) \triangleq \frac{1}{K}\sum_{a=1}^K \mathbb{P}_t^a(X, \theta)$.

Focal distribution: Inducing exploitative bias. Creating a succinct pseudobelief-reward is essential both for pure exploration and for exploration–exploitation but not sufficient to maximise the cumulative reward in case of exploration–exploitation. If a reward distribution having such increasing bias towards higher rewards is amalgamated with the pseudobelief-reward, the resulting belief-reward distribution provides a representation in the belief-reward manifold to balance the exploration–exploitation. Such a reward distribution is called the *focal distribution*. The product of the pseudobelief-reward and the focal distribution jointly represents the summary of explored knowledge and exploitation bias using a single belief-reward distribution. We refer to this as the *pseudobelief-focal-reward distribution* In this paper, we use $\exp\left(\frac{X}{\tau(t)}\right)$ with a time-dependent and controllable parameter $\tau(t)$ as the reward distribution inducing increasing exploitation bias.

Definition 3 (Focal distribution). *A* focal distribution *is a reward distribution of the form* $L_t(X) \propto \exp\left(\frac{X}{\tau(t)}\right)$*, where* $\tau(t)$ *is a decreasing function of* $t \geqslant 1$*. We term* $\tau(t)$ *the* exposure *of the focal distribution.*

Algorithm 1 BelMan

Input: Time horizon T, Number of arms K, Prior on parameters B_0, Reward function f, Exposure $\tau(t)$.
for $t = 1$ **to** T **do**
  /* *I-projection* */
  Draw arm a_t such that

$$a_t = \arg\min_a D_{\mathrm{KL}}\left(\mathbb{P}^a_{t-1}(X,\theta) \,\|\, \bar{\mathbb{Q}}_{t-1}(X,\theta)\right).$$

  /* *Accumulation of observables* */
  Sample a reward x_t out of $f_{\theta_{a_t}}$.
  Update the belief-reward distribution of a_t to $\mathbb{P}^a_t(X,\theta)$ using Bayes' theorem.
  /* *Reverse I-projection* */
  Update the pseudobelief-reward distribution to

$$\bar{\mathbb{Q}}_t(X,\theta) = \arg\min_{\bar{\mathbb{Q}} \in \mathcal{B}_\theta \mathcal{R}} \sum_{a=1}^{K} D_{\mathrm{KL}}\left(\mathbb{P}^a_t(X,\theta) \,\|\, \bar{\mathbb{Q}}(X,\theta)\right).$$

end for

Thus, the pseudobelief-focal-reward distribution-reward distribution is represented as $\bar{\mathbb{Q}}(X,\theta) \triangleq \frac{1}{\bar{Z}_t}\bar{\mathbb{P}}(X,\theta)\exp\left(\frac{X}{\tau(t)}\right)$, where $\bar{Z}_t = \int_X \int_\theta \bar{\mathbb{P}}(X,\theta)\exp\left(\frac{X}{\tau(t)}\right) \mathrm{d}\theta \mathrm{dX}$ is the normalisation factor. Following Equation (2), we compute the pseudobelief-focal-reward distribution as

$$\bar{\mathbb{Q}}_t(X,\theta) \triangleq \arg\min_{\bar{\mathbb{Q}}} \sum_{a=1}^{K} D_{\mathrm{KL}}\left(\mathbb{P}^a_{t-1}(X,\theta) \,\|\, \bar{\mathbb{Q}}(X,\theta)\right).$$

The focal distribution gradually concentrates on higher rewards as the exposure $\tau(t)$ decreases with time. We see this feature in Figure 1. Thus, it constrains using KL-divergence to choose distributions with higher rewards and induces an exploitative bias. As we shall later see (Theorem 3), $\tau(t)$ is chosen to grow in the order of $O(\frac{1}{\log(t)+c\times\log(\log(t))})$ for the exploration–exploitation bandit problem, independently of the family of reward distribution (c is a constant). Following the similarities of bounds obtained in [14] and the corresponding experiments in [14], we set the exposure as $\tau(t) \coloneqq [\log(t)+15\times\log(\log(t))]^{-1}$ for experimental evaluation. As the exposure $\tau(t)$ decreases with t, the focal distribution gets more concentrated on higher reward values. For the pure exploration bandits, we set the exposure $\tau(t) = \infty$ to remove any bias towards higher reward values i.e, exploitation.

BelMan: An Alternating Projection Scheme. A bandit algorithm performs three operations in each step– chooses an arm, samples from the reward distribution of the chosen arm and incorporate the sampled reward to update the knowledge-base. BelMan (Algorithm 1) performs the first and the last operations by alternately minimising the KL-divergence $D_{\mathrm{KL}}(.\,\|\,.)$ [25] between the belief-reward distributions of the arms and the pseudobelief-focal-reward distribution-reward distribution. BelMan chooses to play the arm whose belief-reward incurs minimum KL-divergence with respect to the pseudobelief-focal-reward distribution. Following that, BelMan uses the reward collected from the played arm

to do Bayesian update of the belief-reward and to update the pseudobelief-focal-reward distribution-reward distribution to the point minimising the sum of KL-divergences from the belief-rewards of all the arms. [10] geometrically formulated such minimisation of KL-divergence with respect to a participating distribution as a projection to the set of the other distributions. For a given t, the belief-reward distributions of all the arms $\mathbb{P}_t^a(X, \boldsymbol{\theta})$ form a set $\mathcal{P} \subset \mathcal{B}_\theta\mathcal{R}$ and the pseudobelief-focal-reward distribution-reward distributions $\bar{\mathbb{Q}}_t(X, \boldsymbol{\theta})$ constitute another set $\mathcal{Q} \subset \mathcal{B}_\theta\mathcal{R}$.

Definition 4 (I-projection). *The* information projection *(or* I-projection*) of a distribution $\bar{\mathbb{Q}} \in \mathcal{Q}$ onto a non-empty, closed, convex set $\mathcal{P}$ of probability distributions, $\mathbb{P}^a$'s, defined on a fixed support set is defined by the probability distribution $\mathbb{P}^{a*} \in \mathcal{P}$ that has minimum KL-divergence to q: $\mathbb{P}^{a*} \triangleq \arg\min_{\mathbb{P}^a \in \mathcal{P}} D_{\mathrm{KL}}(\mathbb{P}^a \,\|\, \bar{\mathbb{Q}})$.*

BelMan decides which arm to pull by an I-projection of the pseudobelief-focal-reward distribution onto the beliefs-rewards of each of the arms (Lines 3–4). This operation amounts to computing

$$\begin{aligned} a_t &\triangleq \arg\min_a D_{\mathrm{KL}}\left(\mathbb{P}_{t-1}^a(X,\theta) \,\|\, \bar{\mathbb{Q}}_{t-1}(X,\theta)\right) \\ &= \arg\max_a \left(\mathbb{E}_{\mathbb{P}_{t-1}^a(X,\theta)}\left[\frac{X}{\tau(t)}\right] - D_{\mathrm{KL}}\left(b_{t-1}^a(\theta) \,\|\, b_{\bar{\boldsymbol{\eta}}_{t-1}}(\theta)\right)\right) \end{aligned}$$

The first term symbolises the expected reward of arm a. 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 a is chosen on the basis of the present pseudobelief. The exposure $\tau(t)$ of the focal distribution keeps a weighted balance between exploration and exploitation. Decreasing $\tau(t)$ decreases the exploration with time which is quite an intended property of an exploration–exploitation algorithm.

Following that (Line 5–7), the agent plays the chosen arm a_t and samples a reward x_t. This observation is incorporated in the belief of the arm using Bayes' rule of Equation (1).

Definition 5 (rI-projection). *The* reverse information projection *(or* rI-projection*) of a distribution $\mathbb{P}^a \in \mathcal{P}$ 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 $\bar{\mathbb{Q}}^* \in \mathcal{Q}$ that has minimum KL-divergence from $\mathbb{P}^a$: $\bar{\mathbb{Q}}^* \triangleq \arg\min_{\bar{\mathbb{Q}} \in \mathcal{Q}} D_{\mathrm{KL}}(\mathbb{P}^a \,\|\, \bar{\mathbb{Q}})$.*

Theorem 2 (Central limit theorem). *If $\tilde{\bar{\mu}}_T \triangleq \frac{1}{K}\sum_{a=1}^K \tilde{\mu}_{t_T^a}^a$ is estimator of the expectation parameters of the pseudobelief distribution, $\sqrt{T}(\tilde{\bar{\mu}}_T - \bar{\mu})$ converges in distribution to a centred normal random vector in $\mathcal{N}(0, \bar{\Sigma})$. The covariance matrix $\bar{\Sigma} = \sum_{a=1}^K \lambda_a \Sigma^a$ such that $\frac{T}{K^2 t_T^a}$ tends to λ^a as $T \to \infty$.*

Theorem 2 shows that the parameters of pseudobelief can be constantly estimated and their estimation would depend on the accuracy of the estimators

of individual arms with a weight on the number of draws on the corresponding arms. Thus, the uncertainty in the estimation of the parameter is more influenced by the arm that is least drawn and less influenced by the arm most drawn. In order to decrease the uncertainty corresponding to pseudobelief, we have to draw the arms less explored.

In Theorem 3, we prove BelMan to be asymptotically consistent. We need an additional assumption on the posterior distributions of the arms to claim this.

Assumption 3 (Bounded log-likelihood ratios). *The log-likelihood of the posterior belief distribution at time t with respect to the true posterior belief distribution is bounded such that* $\lim_{t\to\infty} \left|\log \frac{\mathbb{P}^a(X,\theta)}{\mathbb{P}^a_t(X,\theta)}\right| \leqslant C < \infty$ *for all a.*

This assumption helps to control the convergence of sample KL divergences in to the true KL-divergences as the number of samples grow infinitely. This is a relaxed version of Assumption 2 employed in [18] to bound the regret of Thompson sampling. This is also often used in the statistics literature to control the convergence rate of posterior distributions [33,35].

Theorem 3 (Asymptotic consistency). *Given $\tau(t) = \frac{1}{\log t + c \times \log\log t}$ for any $c \geqslant 0$, BelMan will asymptotically converge to choosing the optimal arm in case of a bandit with bounded reward and finite arms. Mathematically, if there exists $\mu^* \triangleq \max_a \mu(\theta_a)$,*

$$\lim_{T\to\infty} \frac{1}{T}\, \mathbb{E}\left[\sum_{t=1}^{T} X_{a_t}\right] = \mu^*. \tag{3}$$

We intuitively validate this claim. We can show the KL-divergence between belief-reward of arm a and the pseudobelief-focal-reward is $D_{\mathrm{KL}}(\mathbb{P}^a_t(X,\theta) \,\|\, \bar{\mathbb{Q}}(X,\theta)) = (1-\lambda^a)h(b^a_t) - \frac{1}{\tau(t)}\mu^a_t$, for λ^a computed as per Theorem 2. As $t \to \infty$, the entropy of belief on each arm reduces to a constant dependent on its internal entropy. Thus, when $\frac{1}{\tau(t)}$ exceeds the entropy term for a large t, BelMan greedily chooses the arm with highest expected reward. Hence, BelMan is asymptotically consistent.

BelMan is applicable to any belief-reward distribution for which KL-divergence is computable and finite. Additionally for reward distributions belonging to the exponential family of distributions, the belief distributions, being conjugate to the reward distributions, also belong to the exponential family [6]. This makes belief-reward distributions flat with respect to KL-divergence. Thus, both I-and rI-projections in BelMan are well-defined and unique for exponential family reward distributions. Furthermore, if we identify the belief-reward distributions with expectation parameters, we obtain the pseudobelief as a weighted sum of them. This allows us to compute belief-reward distribution directly instead of computing its dependence on each belief-reward separately. The exponential family includes the majority of the distributions found in the bandit literature such as Bernoulli, beta, Gaussian, Poisson, exponential, and χ^2.

3 Empirical Performance Analysis

Exploration–exploitation bandit problem. We evaluate the performance of BelMan for two exponential family distributions – Bernoulli and exponential. They

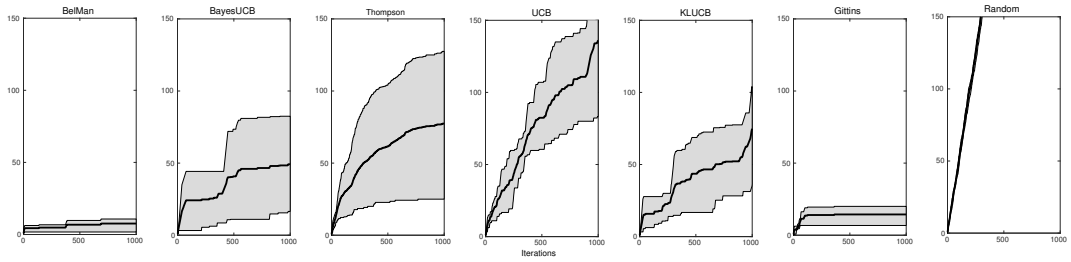


Fig. 2. Evolution of number of suboptimal draws for 2-arm Bernoulli bandit with expected rewards 0.8 and 0.9 for 1000 iterations. The dark black line shows the average over 25 runs. The grey area shows the 75 percentile.

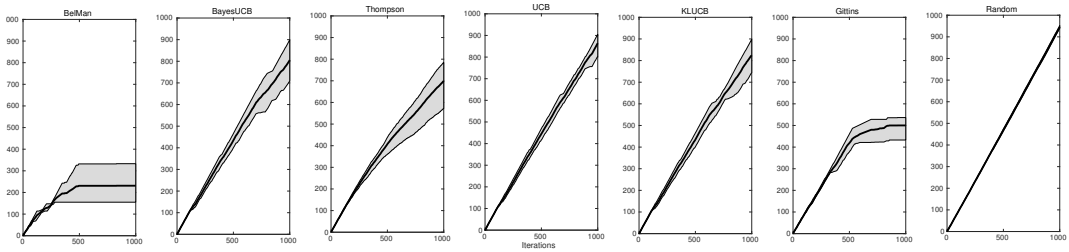


Fig. 3. Evolution of number of suboptimal draws for 20-arm Bernoulli bandit with expected rewards [0.25 0.22 0.2 0.17 0.17 0.2 0.13 0.13 0.1 0.07 0.07 0.05 0.05 0.05 0.02 0.02 0.02 0.01 0.01 0.01] for 1000 iterations.

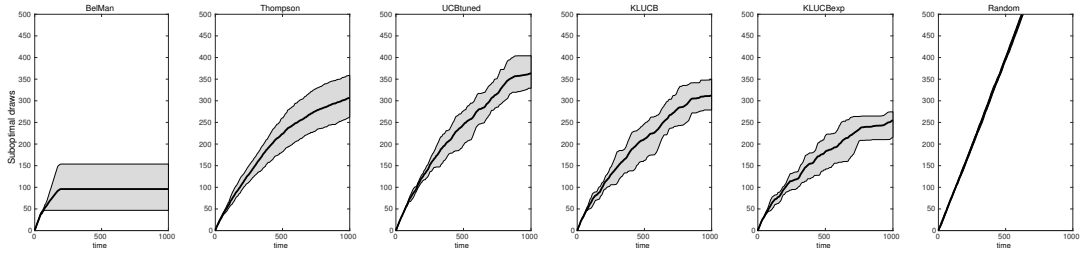


Fig. 4. Evolution of number of suboptimal draws for 5-arm bounded exponential bandit with expected rewards 0.2, 0.25, 0.33, 0.5, and 1.0 for 1000 iterations.

stand for discrete and continuous rewards respectively. We use the pymaBandits library [9] for implementation of all the algorithms except ours, and run it on MATLAB 2014a. 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 compare the performance of BelMan with frequentist methods like UCB [3] and KL-UCB [14], and Bayesian methods like Thompson sampling [34] and Bayes-UCB [21]. For Bernoulli bandits, we also compare with Gittins index [17] which is the optimal algorithm for Markovian finite arm independent bandits with discounted rewards. Though we are not specifically interested in the discounted case, Gittins' algorithm is indeed transferable to the finite horizon setting with slight manipulation. Though it is often computationally intractable, we use it

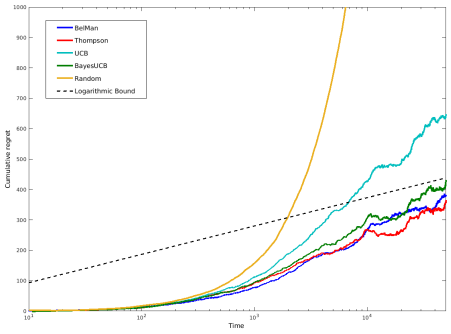


Fig. 5. Evolution of (mean) regret for exploration–exploitation 20-arm Bernoulli bandits with horizon=50,000.

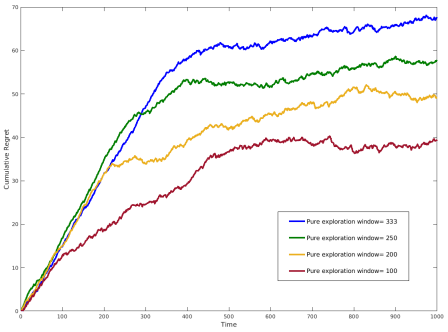


Fig. 6. Evolution of (mean) cumulative regret for two-phase 20-arm Bernoulli bandits.

as the optimal baseline for Bernoulli bandits. We also plot performance of the uniform sampling method (*Random*), as a naïve baseline.

From Figures 2, 3, and 4, we observe that at the very beginning the number of suboptimal draws 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 [16]: 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 by learning after a certain number of iterations, and to find a trade-off between exploration and exploitation. For the 2-arm Bernoulli bandit ($\theta_1 = 0.8, \theta_2 = 0.9$), 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 Bernoulli 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, 0.01, 0.01, and 0.01). Since more arms ask for more exploration and more suboptimal draws, all algorithms show higher regret values. On all experiments performed, BelMan outperforms the competing approaches. We also simulated 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 reward distributions- Thompson sampling, UCBtuned [3], KL-UCB, and KL-UCB-exp, a method tailored for exponential distribution of rewards [14]. This demonstrates BelMan's broad applicability and efficient performance in complex scenarios.

We have also run the experiments 50 times with horizon 50 000 for the 20 arm Bernoulli bandit to verify the asymptotic behaviour of BelMan. Figure 5 shows that BelMan's regret gradually becomes linear with respect to the logarithmic axis. Figure 5 empirically validates BelMan to achieve logarithmic regret like the competitors which are theoretically proven to reach logarithmic regret bound illustrated by the dashed line.

Two-phase reinforcement learning problem. In this experiment, we simulate a two-phase setup, as in [29]: the agent first does pure exploration

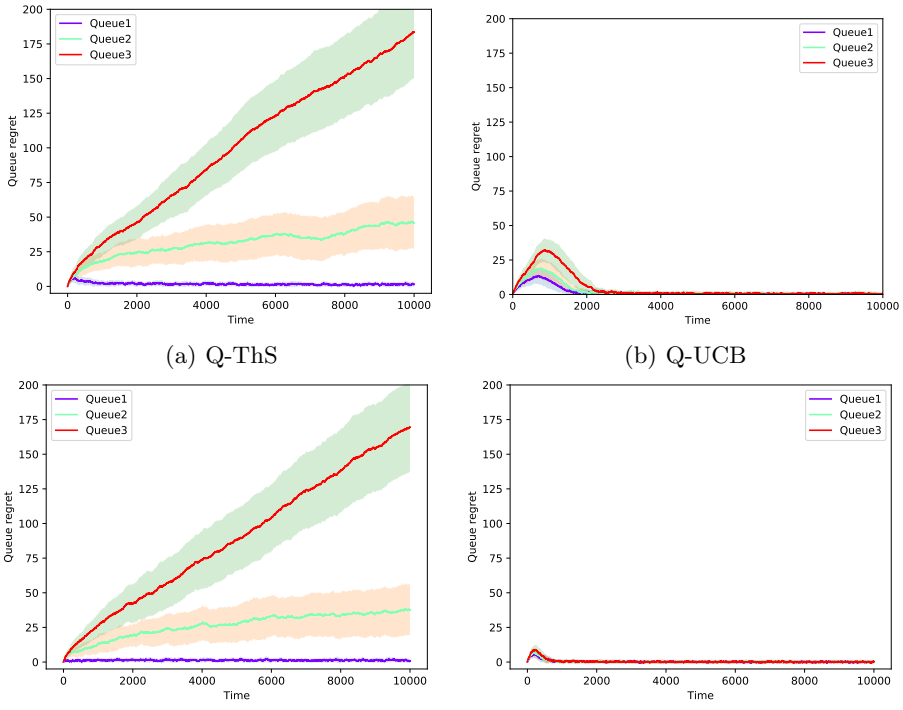
for a fixed number of iterations, then move to exploration–exploitation. This is possible since BelMan supports both modes and can transparently switch. The setting is that of the 20-arm Bernoulli bandit in Figure 3. The two-phase algorithm is exactly BelMan (Algorithm 1) with $\tau(t) = \infty$ for an initial phase of length T_{EXP} followed by the decreasing function of t as indicated previously. Thus, BelMan gives us a single algorithmic framework for three setups of bandit problems– pure exploration, exploration–exploitation, and two-phase learning. We only have to choose a different $\tau(t)$ depending on the problem addressed. This supports BelMan's claim as a generalised, unified framework for stochastic bandit problems.

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 three major lessons of this experiment. First, Bayesian methods provide an inherent advantage in leveraging prior knowledge (here, accumulated in the first phase). Second, a pure exploration phase helps in improving the performance during the exploration–exploitation phase. Third, we can leverage the exposure to control the exploration–exploitation trade-off.

4 Application to Queueing Bandits

We instantiate BelMan for the problem of scheduling jobs in a multiple-server multiple-queue system with known arrival rates and unknown service rates. The goal of the agent is to choose such a server for the given system such that the total queue length, i.e. the jobs waiting in the queue, will be as less as possible. This problem is referred as the queueing bandit [24].

We consider a discrete-time queueing system with 1 queue and K servers. The servers are indexed by $a \in \{1, \ldots, K\}$. Arrivals to the queue and service offered by the servers are assumed to be independent and identically distributed across time. The mean arrival rate is $\lambda \in \mathbb{R}^+$. The mean service rates are denoted by $\boldsymbol{\mu} \in \{\mu_a\}_{a=1}^K$, where μ_a is the service rate of server a. At a time, a server can serve the jobs coming from a queue only. We assume the queue to be stable i.e, $\lambda < \max_{a \in [K]} \mu_a$. Now, the problem is to choose a server at each time $t \in [T]$ such that the number of jobs waiting in queues is as less as possible. The number of jobs waiting in queues is called the *queue length* of the system. If the number of arrivals to the queues at time t is $A(t)$ and $S(t)$ is the number of jobs served, the queue length at time t is defined as $Q(t) \triangleq Q(t-1) + A(t) - S(t)$, where $Q : [T] \to \mathbb{R}^{\geqslant 0}$, $A : [T] \to \mathbb{R}^{\geqslant 0}$, and $S : [T] \to \mathbb{R}^{\geqslant 0}$. The agent, which is the scheduling algorithm in this case, tries to minimise this queue length for a given horizon $T > 0$. The arrival rates are known to the scheduling algorithm but the



(a) Q-ThS (b) Q-UCB

(c) Thompson sampling (d) BelMan

Fig. 7. Queue regret for single queue and 5 server setting with Poisson arrival with arrival rate 0.35 and Bernoulli service distribution with service rates [0.5,0.33,0.33,0.33,0.25], [0.33,0.5,0.25,0.33,0.25], and [0.25,0.33,0.5,0.25,0.25] respectively. Each experiment is performed 50 times for a horizon of 10,000. (Code: `https://github.com/debabrota-basu/qbelman`)

service rates are unknown to it. This creates the need to learn about the service distributions, and in turn, engenders the exploration-exploitation dilemma.

Following the bandit literature, [24] proposed to use *queue regret* as the performance measure of a queueing bandit algorithm. Queue regret is defined as the difference in the queue length if a bandit algorithm is used instead of an optimal algorithm with full information about the arrival and service rates. Thus, the *optimal algorithm* OPT knows all the arrival and service rates, and allocates the queue to servers with the best service rate. Hence, we define the queue regret of a queueing bandit algorithm $\Psi(t) \triangleq \mathbb{E}\left[Q(t) - Q^{\mathrm{OPT}}(t)\right]$. In order to keep the bandit structure, we assume that both the queue length $Q(t)$ of algorithm $\mathcal{A}$ and that of the optimal algorithm $Q^{\mathrm{OPT}}(t)$ starts with the same stationary state distribution $\nu(\lambda, \boldsymbol{\mu})$.

We show experimental results for the $M/B/K$ queueing bandits. We assume the arrival process to be Markovian, and the service process to be Bernoulli. The arrival process being Markovian implies that the stochastic process describing the number of arrivals is therefore $A(t)$ have increments independent of time. This makes the distribution of $A(t)$ to be a Poisson distribution [12] with mean arrival rate λ. We denote $B_a(\mu_a)$ is the Bernoulli distribution of the service time of

server a. It implies that the server processes a job with probability $\mu_a \in (0, 1)$ and refuses to serve it with probability $1 - \mu_a$. The goal is to perform the scheduling in such a way that the queue regret will be minimised. The experimental results in Figure 7 depict that BelMan is more stable and efficient than the competing algorithms: Q-UCB, Q-Thompson sampling, and Thompson sampling.

5 Related Work

[5] posed the problem of discounted reward bandits with infinite horizon as a single-state Markov decision process [17] and proposed an algorithm for computing deterministic Gittins indices to choose the arm to play. Though Gittins index is proven to be optimal for discounted Bayesian bandits with Bernoulli rewards [17], 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 [28]. This motivated researchers to design computationally tractable algorithms [7] that still retain the asymptotic efficiency [26].

These algorithms can be classified into two categories: frequentist and Bayesian. Frequentist algorithms use the history obtained as the number of arm plays and corresponding rewards obtained to compute point estimates of the fitness index to choose an arm. UCB [3], UCB-tuned [3], KL-UCB [14], KL-UCB-Exp [14], KL-UCB$^+$ [20] are examples of frequentist algorithms. These algorithms are designed by the philosophy of optimism in face of uncertainty. This methodology prescribes to act as if the empirically best choice is truly the best choice. Thus, all these algorithms overestimate the expected reward of the corresponding arms in form of frequentist indices.

Bayesian algorithms encode the available information on the reward generation process in form of a prior distribution. In case of stochastic bandits, this prior consists of K belief distributions on the arms. The history obtained by playing the bandit game is used to update the posterior distribution. This posterior distribution is further used to choose the arm to play. Thompson sampling [34], information-directed sampling [32], Bayes-UCB [20], and BelMan are Bayesian algorithms.

In a variant of the stochastic bandit problem, called the *pure exploration bandit* problem [8], the goal of the gambler is solely to accumulate information about the arms. In another variant of the stochastic bandit problem, the gambler interacts with the bandit in two consecutive phases of pure exploration and exploration–exploitation. [29] named this variant the *two-phase reinforcement learning* problem. 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 [13]. As frequentist methods are well-tuned for exploration-exploitation bandits, a different set of algorithms need to be developed for pure exploration bandits [8]. [23] pointed out the lack of Bayesian methods to do so. This motivated recent developments of Bayesian algorithms [31] which are modifications of their exploration–exploitation counterparts such as Thompson sampling. BelMan leverages its geometric insight to manage the pure exploration bandits only by turning the exposure to infinity. Thus, it provides a single framework to manage the pure exploration, exploration–exploitation, and two-phase reinforcement learning problems only by tuning the exposure.

6 Conclusion

BelMan implements a generic Bayesian information-geometric approach for stochastic multi-armed bandit problems. It operates in a statistical manifold constructed by the joint distributions of beliefs and rewards. Their barycentre, the pseudobelief-reward, summaries the accumulated information and forms the basis of the exploration component. The algorithm is further extended by composing the pseudobelief-reward distribution with a reward distribution that gradually concentrates on higher rewards by means of a time-dependent function, the exposure. In short, BelMan addresses the issue of the adaptive balance of exploration–exploitation from the perspective of information representation, accumulation, and balanced induction of exploitative bias. Consequently, BelMan can be uniformly tuned to support pure exploration, exploration–exploitation, and two-phase reinforcement learning problems. BelMan, when instantiated to rewards modelled by any distribution of the exponential family, conveniently leads to analytical forms that allow derivation of a well-defined and unique projection as well as to devise an effective and fast computation. In queueing bandits, the agent tries and minimises the queue length while also learning the unknown service rates of multiple servers. Comparative performance evaluation shows BelMan to be more stable and efficient than existing algorithms in the queueing bandit literature.

We are investigating the analytical asymptotic efficiency and stability of BelMan. We are also investigating how BelMan can be extended to other settings such as dependent arms, non-parametric distributions and continuous arms.

References

1. Agueh, M., Carlier, G.: Barycenters in the wasserstein space. SIAM Journal on Mathematical Analysis **43**(2), 904–924 (2011)
2. Amari, S.I., Nagaoka, H.: Methods of information geometry, Translations of mathematical monographs, vol. 191. American Mathematical Society (2007)
3. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Machine learning **47**(2–3), 235–256 (2002)
4. Barbaresco, F.: Information geometry of covariance matrix: Cartan-siegel homogeneous bounded domains, mostow/berger fibration and frechet median. In: Matrix Information Geometry, pp. 199–255. Springer (2013)
5. Bellman, R.: A problem in the sequential design of experiments. Sankhyā: The Indian Journal of Statistics (1933–1960) **16**(3/4), 221–229 (1956)
6. Brown, L.D.: Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory. Institute of Mathematical Statistics (1986)
7. Bubeck, S., Cesa-Bianchi, N., et al.: Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends in Machine Learning **5**(1), 1–122 (2012)
8. Bubeck, S., Munos, R., Stoltz, G.: Pure exploration in multi-armed bandits problems. In: ALT. pp. 23–37. Springer (2009)
9. Cappé, O., Garivier, A., Kaufmann, É.: pymaBandits (2012), `http://mloss.org/software/view/415/`
10. Csiszár, I.: Sanov property, generalized I-projection and a conditional limit theorem. The Annals of Probability **12**(3), 768–793 (1984)

11. DeGroot, M.H.: Optimal statistical decisions, Wiley Classics Library, vol. 82. John Wiley & Sons (2005)
12. Durrett, R.: Probability: theory and examples. Cambridge University Press (2010)
13. Faheem, M., Senellart, P.: Adaptive web crawling through structure-based link classification. In: Proc. ICADL. pp. 39–51. Seoul, South Korea (Dec 2015)
14. Garivier, A., Cappé, O.: The KL-UCB algorithm for bounded stochastic bandits and beyond. In: COLT. pp. 359–376 (2011)
15. Garivier, A., Lattimore, T., Kaufmann, E.: On explore-then-commit strategies. In: Advances in Neural Information Processing Systems 29, pp. 784–792. Curran Associates, Inc. (2016)
16. Garivier, A., Ménard, P., Stoltz, G.: Explore first, exploit next: The true shape of regret in bandit problems. arXiv preprint arXiv:1602.07182 (2016)
17. Gittins, J.C.: Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society. Series B (Methodological) **41**(2), 148–177 (1979)
18. Gopalan, A., Mannor, S.: Thompson sampling for learning parameterized markov decision processes. In: Conference on Learning Theory. pp. 861–898 (2015)
19. Jaynes, E.T.: Prior probabilities. IEEE Transactions on Systems Science and Cybernetics **4**, 227–241 (1968)
20. Kaufmann, E.: On bayesian index policies for sequential resource allocation. Annals of Statistics **46**(2), 842–865 (April 2018)
21. Kaufmann, E., Cappé, O., Garivier, A.: On Bayesian upper confidence bounds for bandit problems. In: AISTATS. pp. 592–600 (2012)
22. Kaufmann, E., Kalyanakrishnan, S.: Information complexity in bandit subset selection. In: COLT. pp. 228–251 (2013)
23. Kawale, J., Bui, H.H., Kveton, B., Tran-Thanh, L., Chawla, S.: Efficient Thompson sampling for online matrix-factorization recommendation. In: NIPS. pp. 1297–1305 (2015)
24. Krishnasamy, S., Sen, R., Johari, R., Shakkottai, S.: Regret of queueing bandits. In: Advances in Neural Information Processing Systems. pp. 1669–1677 (2016)
25. Kullback, S.: Information theory and statistics. Courier Corporation (1997)
26. Lai, T.L., Robbins, H.: Asymptotically efficient adaptive allocation rules. Adv. Appl. Math. **6**(1), 4–22 (Mar 1985)
27. Macready, W.G., Wolpert, D.H.: Bandit problems and the exploration/exploitation tradeoff. IEEE Transactions on evolutionary computation **2**(1), 2–22 (1998)
28. Nino-Mora, J.: Computing a classic index for finite-horizon bandits. INFORMS Journal on Computing **23**(2), 254–267 (2011)
29. Putta, S.R., Tulabandhula, T.: Pure exploration in episodic fixed-horizon Markov decision processes. In: AAMAS. pp. 1703–1704 (2017)
30. Robbins, H.: Some aspects of the sequential design of experiments. Bull. Amer. Math. Soc. **58**(5), 527–535 (09 1952)
31. Russo, D.: Simple bayesian algorithms for best arm identification. In: Conference on Learning Theory. pp. 1417–1418 (2016)
32. Russo, D., Van Roy, B.: An information-theoretic analysis of Thompson sampling. Journal of Machine Learning Research (2014)
33. Shen, X., Wasserman, L., et al.: Rates of convergence of posterior distributions. The Annals of Statistics **29**(3), 687–714 (2001)
34. Thompson, W.R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika **25**(3–4), 285 (1933)
35. Wong, W.H., Shen, X.: Probability inequalities for likelihood ratios and convergence rates of sieve mles. The Annals of Statistics **23**(2), 339–362 (1995)

Supplementary Material: BelMan

We provide the following supplementary material:

- in Section A, an extended discussion of the related work and of the setting of bandits, beyond what could fit in the main paper;
- in Section B, proofs and technical details that complement the methodology section (Section 2);
- in Section C, additional experiments in the exploration–exploitation setup.

A Extended Discussion of Related Work

Exploration–exploitation bandit problem. In the exploration–exploitation bandits, the agent searches for a policy that maximises the expected value of *cumulative reward* $S_T \triangleq \mathbb{E}\left[\sum_{t=1}^{T} X_{a_t}\right]$ as $T \to \infty$. A policy is *asymptotically consistent* [Rob52] if it asymptotically tends to choose the arm with maximum expected reward $\mu^* \triangleq \max_{1 \leqslant a \leqslant K} \mu(\theta_a)$, i.e.,

$$\lim_{T\to\infty} \frac{1}{T} S_T = \mu^*. \tag{4}$$

The *cumulative regret* R_T [LR85] is the amount of extra reward the gambler can obtain if she knows the optimal arm a^* and always plays it instead of the present sequence:

$$\begin{aligned} R_T &\triangleq T\mathbb{E}\left[X_{a^*}\right] - \mathbb{E}\left[\sum_{t=1}^{T} \left(X_{a_t}\right)\right] \\ &= T\mu^* - \sum_{a=1}^{K} \mathbb{E}\left[\sum_{t=1}^{T} \left(X_{a_t} \times \mathbf{1}(a_t = a)\right)\right] \\ &= \sum_{a=1}^{K} \left[\mu^* - \mu^a\right] \mathbb{E}[t_T^a], \end{aligned}$$

where t_T^a is the number of times arm a is pulled till the Tth iteration. [LR85] proved that for all algorithms satisfying $R_T = o(T^c)$ for a non-negative c, the cumulative regret increases asymptotically in $\Omega(\log T)$. Such algorithms are called *asymptotically efficient*. The Lai–Robbins bound can be mathematically formulated as

$$\liminf_{T\to\infty} \frac{R_T}{\log T} \geqslant \frac{\sum\limits_{a:\mu^*(\boldsymbol{\theta})>\mu(\theta_a)} \left[\mu^*(\boldsymbol{\theta}) - \mu(\theta_a)\right]}{\inf_a D_{\mathrm{KL}}(f_{\theta_a}(x) \parallel f_{\theta^*}(x))}, \tag{5}$$

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. [LR85] also showed that for specific reward distributions, the expected number of draws of any suboptimal arm a should satisfy

$$t_T^a \leqslant \left(\frac{1}{\inf_a D_{\mathrm{KL}}(f_{\theta_a}(x) \parallel f_{\theta^*}(x))} + o(1) \right) \log T. \tag{6}$$

Equation (5) and (6) together claim that the best achievable number of draws of suboptimal arms is $\Theta(\log T)$. Based on this bound, [ACBF02] extensively studied the upper confidence bound (UCB) family of algorithms. These algorithms operate on the philosophy of optimism in face of uncertainty. They compute the upper confidence bounds of each of the arm's distributions in a frequentist way and choose the one with the maximum upper confidence bound optimistically expecting that one to be the arm with maximum expected reward. Later on, this family of algorithms was analysed and improved to propose algorithms such as KL-UCB [GC11] and DMED [HT11].

Frequentist approaches implicitly assume a 'true' parametrisation of reward distributions $(\theta_1^{true}, \ldots, \theta_K^{true})$. In contrast, Bayesians model the uncertainty on the parameter using another probability distribution $B(\theta_1, \ldots, \theta_K)$ [DeG05,Sco10] which is referred to as the *belief distribution*. Bayesian algorithms begin with a prior $B_0(\theta_1, \ldots, \theta_K)$ over the parameters and eventually try to find out a posterior distribution such that the Bayesian sum of rewards $\int S_T \mathrm{d}B(\theta_1, \ldots, \theta_K)$ is maximised, or equivalently the Bayesian risk $\int R_T \mathrm{d}B(\theta_1, \ldots, \theta_K)$ is minimised.

Another variant of the Bayesian formulation was introduced by [Bel56] with a discounted reward setting. Unlike S_T, the discounted sum of rewards $D_\gamma \triangleq \sum_{t=0}^{\infty} [\gamma^t x_{t+1}]$ is calculated over infinite horizon. Here, $\gamma \in [0, 1)$ ensures convergence of the sequential sum of rewards for infinite horizon. Intuitively, the discounted sum implies the effect of an action decay with each time step by the discount factor γ. 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, the bandit problem is reformulated as maximising

$$\int \ldots \int \mathbb{E}_{\boldsymbol{\theta}}[D_\gamma] \mathrm{d}b^1(\theta_1) \ldots \mathrm{d}b^K(\theta_K)$$

where, b^a is the independent prior distribution on the parameter θ_a for $a = 1, \ldots, K$. [Git79] showed the agent can have an optimally indexed policy by sampling from the arm with largest Gittins index

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

where s^a is the state of arm a and τ is referred to as the stopping time i.e, the first time when the index is no greater than its initial value. Though Gittins index

[Git79] is proven to be optimal for discounted Bayesian bandits with Bernoulli rewards, 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 [NM11].

Thus, researchers developed approximation algorithms [Lai88] and sequential sampling schemes like Thompson sampling [Tho33]. At any iteration, the latter samples K parameter values from the belief distributions and chooses the arm that has maximum expected reward for them. [KCG12] 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. [BMS09] formulated this notion of gaining information as minimisation of the simple regret rather than cumulative regret. *Simple regret* $r_t(\boldsymbol{\theta})$ at time t is the expected difference between the maximum achievable reward X_{a^*} and the sampled reward X_{a_t}. Unlike cumulative regret, minimising simple regret depends only on exploration and the number of available rounds to do so. [BMS09] 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. [AB10] identified the pure exploration problem as *best arm identification* and proposed the Successive Rejects algorithm under fixed budget constraints. [BWV13] 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 [KK13]. 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. Existing frequentist algorithms [AB10,BWV13,KK13] 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.

Two-Phase reinforcement learning. Two-phase reinforcement learning problems append the exploration–exploitation problem after the pure exploration problem. The agent gets an initial phase of pure exploration for a given window. 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 [FS15] 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. In applications, this way of beginning the exploration–exploitation is called a warm start. 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, Bayesian formulation provide the scope to approach this problem without any modification. [PT17] approached this problem with a technique amalgamating a sampling technique, PSPE, and an extension of Thompson sampling, PSRL [ORVR13], for episodic fixed horizon Markov decision processes (MDPs) [DB15]. PSPE uses Bayesian update to create a posterior distribution for the reward distribution of a policy. Then, PSPE samples from the distribution in order to evaluate the policies. These two steps are performed iteratively for the initial pure exploration phase. PSRL [ORVR13] 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.

B Supplementary Material for Section 2 (Methodology)

B.1 KL-divergence on the Manifold.

Kullback-Liebler divergence (or KL-divergence) [Kul97] is a pre-metric measure of dissimilarity between two probability distributions.

Definition 6 (KL-divergence). *If there exist two probability measures P and Q defined over a support set S and P is absolutely continuous with respect to Q, we define the KL-divergence between them as*

$$D_{\mathrm{KL}}(P \,\|\, Q) \triangleq \int_S \log \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}P.$$

$\frac{\mathrm{d}P}{\mathrm{d}Q}$ 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 support set S, the KL-divergence can be rewritten as

$$D_{\mathrm{KL}}(p(s) \,\|\, q(s)) = \int_{s \in S} p(s) \log \frac{p(s)}{q(s)} ds = -h(p(s)) + H(p(s), q(s)). \quad (7)$$

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 α-divergence function on the statistical manifold is a convex combination of ± 1-divergences. Mathematically, for $\alpha \in (-1, +1)$,

$$\begin{aligned} D^{(\alpha)}(p \,\|\, q) &\triangleq \frac{1+\alpha}{2} D^{(+1)}(p \,\|\, q) + \frac{1-\alpha}{2} D^{(-1)}(p \,\|\, q) \\ &= \frac{1+\alpha}{2} D_{\mathrm{KL}}(q \,\|\, p) + \frac{1-\alpha}{2} D_{\mathrm{KL}}(p \,\|\, q). \end{aligned} \tag{8}$$

From a manifold perspective, it seems that the divergence function for the ± 1-connections on the belief-reward manifolds and a convex mixture of D_{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 optimal manner, and to decrease the uncertainty, we have to represent the observations using a knowledge-base, and to minimise the KL-divergence between the knowledge-base and other distributions respectively. If $\mathcal{P}$ are the candidate belief-reward distributions of the arms formed by accumulation of actions and rewards, and $\mathcal{Q}$ are the pseudobelief or pseudobelief-focal-reward distribution-reward distributions, the alternating minimisation scheme looks for the most succinct representation $\mathcal{Q}$ of the knowledge and the exploitation bias while choosing such arms whose belief-reward distributions resemble their true reward distributions as much as possible.

B.2 Exponential Family

Use of KL-divergence as a divergence measure on the statistical manifolds and also the issue of representation of a random variable using sufficient statistics provoked the study of the exponential family of distributions. Interesting properties of exponential family distributions, such as existence of finite representation of sufficient statistics, convenient mathematical form, and existence of moments, provided them a central stage in the field of mathematical statistics [Bro86][Dar35][Kau18][Koo36].

The *exponential family* [Bro86] is a class of probability distributions which is defined by a set of *natural parameters* $\omega(\theta)$ and a *sufficient statistics* $T(X)$ of the random variable X as follows:

$$f_\theta(X) \triangleq g(X) \exp\left(\langle \omega(\theta), T(x) \rangle - A(\theta)\right).$$

Here, $g(X)$ 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. For $T(X) = X$, the log-partition function is logarithm of the Laplace transform of the base measure.

Example 1. Bernoulli distribution with probability of success $\theta \in (0,1)$ is defined as

$$f_\theta(X) \triangleq \text{Ber}(\theta) = \theta^X (1-\theta)^{(1-X)} \\ = \exp\left(X \log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right)$$

for $X \in \{0,1\}$. Here, the base measure $g(x)$ is 1. The sufficient statistics is $T(X) = X$. The natural parameter is $\omega(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$. The log-partition function is $A(\theta) = -\log(1-\theta) = \log(1+\exp(\omega))$.

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 sampling and uncertainty representation point of view, the exponential family is useful because of its finite representation of sufficient statistics. Specifically, sufficient statistics of exponential family can represent any arbitrary number of independent identically distributed samples using a finite number of variables [Koo36]. This keeps the uncertainty representation tractable for exponential family distributions.

From a Bayesian point of view, the useful property of the exponential family is the existence of *conjugate distributions* which also belong to this family [Bro86]. Two parametric distributions $f_{\boldsymbol{\theta}}(x)$ and $b_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ are conjugate if the posterior distribution $\mathbb{P}(\boldsymbol{\theta}|x)$ formed by multiplying them has the same form as $b_{\boldsymbol{\eta}}(\boldsymbol{\theta})$. Mathematically, the conjugate distribution of the distribution of Equation B.2 is given by $b_{\boldsymbol{\eta}}(\boldsymbol{\theta}) \triangleq \mathbb{P}(\boldsymbol{\theta}|\boldsymbol{\eta}, v) = f(\boldsymbol{\eta}, v)\exp(\langle \boldsymbol{\eta}, \boldsymbol{\theta}\rangle - vA(\boldsymbol{\theta})) = f(\boldsymbol{\eta}, v)g(\boldsymbol{\theta})^v \exp(\langle \boldsymbol{\eta}, \boldsymbol{\theta}\rangle)$. Here, η is the parameter of the conjugate prior and $v > 0$ corresponds to the effective number of observations that the prior contributes. Thus, if the reward distribution belongs to the exponential family, the belief distribution is represented as: $b_{\boldsymbol{\eta}}(\theta) \triangleq h(\theta)\exp\left(\langle \eta, T(\theta)\rangle - A(\eta)\right)$ with the natural parameters $\eta \in \mathbb{R}^{d'}$.

From information geometric point of view, exponential family distributions are flat with respect to KL-divergence [AN07]. Thus, both information and reverse information projections [Csi84] that we would use in BelMan are well-defined and unique. Thus, at each timestep, we obtain an optimal and unambiguous computation of the decision variables of BelMan. [AN07] 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 parametrisation. Thus, working with exponential family distributions implicitly supports the well-defined nature and possibility of getting an efficient estimation.

B.3 Pseudobelief–reward: Existence, Uniqueness and Consistency

In order to establish pseudobelief–reward as a valid knowledge-base for all the arms, we have to prove that it exists uniquely and its parameters can be consistently estimated.

The proofs require only two assumptions. Firstly, the belief–reward manifold can be described by a unique chart. This implies that pdf of the belief–reward distributions is a bijective function of parameters. Secondly, there exist unique geodesics between any two points of the belief–reward manifold. This implies that the divergence function between any two belief–reward distributions is uniquely defined. Instead of having such modest requirement, we represent our proofs in form of the exponential family distributions due to ease of presentation and our limited interest.

Theorem 1. *For a given set of belief-reward distributions* $\{\mathbb{P}_t^a\}_{a=1}^K$ *defined on the same support set and having a finite expectation,* $\bar{\mathbb{P}}_t$ *is uniquely defined, and is such that its expectation parameter verifies* $\hat{\mu}_t(\theta) = \frac{1}{K}\sum_{a=1}^K \mu_t^a(\theta)$.

Proof. For belief–reward distributions $\mathbb{P}^a$ and $\mathbb{P}$, the KL-divergence is defined as

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathbb{P}_t^a \,\|\, \mathbb{P}\right) &= \int_\theta \int_X \mathbb{P}_t^a(X,\theta) \log \frac{\mathbb{P}_t^a(X,\theta)}{\mathbb{P}(X,\theta)} \mathrm{d}x \mathrm{d}\theta \\
&= \int_\theta \int_X f_\theta(X) b_{\xi_t}^a(\theta) \log \frac{b_{\xi_t}^a(\theta)}{b_\xi(\theta)} \mathrm{d}x \mathrm{d}\theta \\
&= \int_\theta b_{\xi_t}^a(\theta) \log \frac{b_{\xi_t}^a(\theta)}{b_\xi(\theta)} \left[\int_X f_\theta(X) \mathrm{d}x\right] \mathrm{d}\theta \\
&= \int_\theta b_{\xi_t}^a(\theta) \log \frac{b_{\xi_t}^a(\theta)}{b_\xi(\theta)} \mathrm{d}\theta \\
&= \mathbb{E}_{b_t^a}\left[\langle \xi_t^a, \Theta(\theta)\rangle - \Psi_t^a(\xi_t^a) - \langle \xi, \Theta(\theta)\rangle + \Psi(\xi)\right] \\
&= \langle \xi_t^a - \xi, \mu_t^a(\theta)\rangle - \Psi_t^a(\xi_t^a) + \Psi(\xi).
\end{aligned}$$

Thus, the objective function that $\bar{\mathbb{P}}$ minimises is given by

$$F(\mathbb{P}) \triangleq \frac{1}{K}\sum_{a=1}^K D_{\mathrm{KL}}\left(\mathbb{P}_t^a \,\|\, \mathbb{P}\right) = \frac{1}{K}\sum_{a=1}^K \langle \xi_t^a - \xi, \mu^a(\theta)\rangle - \frac{1}{K}\sum_{a=1}^K \Psi_t^a(\xi_t^a) + \Psi(\xi). \tag{9}$$

Since the exponential family distributions are dually flat [AN07], we get a unique expectation parametrisation $\mu(\theta)$ of the belief distributions for a given natural parametrisation ξ. The expectation parameter is defined as $\mu(\theta) \triangleq \mathbb{E}_b[\Theta(\theta)] = \nabla_\xi \Psi(\xi)$. $\mu(\theta)$ dually expresses a natural parametrisation as its dual. Mathematically, $\xi = \nabla_\mu(\langle \xi, \mu\rangle - \Psi(\xi)) = \nabla_\mu \Phi(\mu)$. $\Psi(\xi)$ and $\Phi(\mu)$ are log-normalisers under two parametrisations and are convex conjugate to each other. If we define $\hat{\mu}_t(\theta) \triangleq \frac{1}{K}\sum_{a=1}^K \mu_t^a$, we get a unique natural parameter $\hat{\xi}_t \triangleq \xi(\hat{\mu}_t)$. This allows us to rewrite Equation 9 as

$$\begin{aligned}
F(\mathbb{P}) &= \left[\langle \hat{\xi}_t - \xi, \hat{\mu}_t(\theta)\rangle - \Psi(\hat{\xi}_t) + \Psi(\xi)\right] + \frac{1}{K}\sum_{a=1}^K (\langle \xi_t^a, \mu_t^a(\theta)\rangle - \Psi_t^a(\xi)) - (\langle \xi(\hat{\mu}_t), \hat{\mu}_t(\theta)\rangle - \Psi(\xi(\hat{\mu}_t)) \\
&= D_{\mathrm{KL}}\left(\mathbb{P}_{\hat{\mu}_t} \,\|\, \mathbb{P}\right) + \frac{1}{K}\sum_{a=1}^K \Phi(\mu_t^a) - \Phi(\hat{\mu}_t) \geqslant \frac{1}{K}\sum_{a=1}^K \Phi(\mu_t^a) - \Phi(\hat{\mu}_t).
\end{aligned}$$

Since $D_{\mathrm{KL}}\left(\mathbb{P}_{\hat{\mu}_t} \,\|\, \mathbb{P}\right) = 0$ for $\mathbb{P} = \mathbb{P}_{\hat{\mu}_t}$, $F(\mathbb{P})$ reaches unique minimum $F(\mathbb{P}_{\hat{\mu}})$ for the belief–reward distribution with expectation parameter $\hat{\mu_t}(\theta) \triangleq \frac{1}{K}\sum_{a=1}^{K} \mu_t^a$. Thus, for a given set of belief–reward distributions the pseudobelief–reward distribution $\bar{\mathbb{P}}_t(X,\theta) \triangleq \mathbb{P}_{\hat{\mu}_t}(X,\theta)$ is a unique distribution in belief–reward manifold.

Corollary 1. *The pseudobelief-reward distribution $\bar{\mathbb{P}}_t(X,\theta)$ is the unique point on the belief-reward manifold that has minimum KL-divergence from the distribution $\hat{\mathbb{P}}_t(X,\theta) \triangleq \frac{1}{K}\sum_{a=1}^{K} \mathbb{P}_t^a(X,\theta)$.*

Proof. KL-divergence from $\hat{\mathbb{P}}_t(X,\theta)$ to any pseudobelief–reward distribution $\mathbb{P}(X,\theta)$is

$$D_{\mathrm{KL}}\left(\hat{\mathbb{P}}_t \,\|\, \mathbb{P}\right) = D_{\mathrm{KL}}\left(\hat{\mathbb{P}}_t \,\|\, \bar{\mathbb{P}}_t\right) + \langle \hat{\xi}_t - \xi, \hat{\mu}_t \rangle - \Psi(\hat{\xi}_t) + \Psi(\xi) = D_{\mathrm{KL}}\left(\hat{\mathbb{P}} \,\|\, \bar{\mathbb{P}}\right) + D_{\mathrm{KL}}\left(\bar{\mathbb{P}} \,\|\, \mathbb{P}\right).$$

Here, $\bar{\mathbb{P}}_t$ is the pseudobelief distribution with $\hat{\xi}_t$ and $\hat{\mu}_t$ as defined in Theorem 1. Since $\hat{\mathbb{P}}_t$ is a mixture of belief–reward distributions, it does not belong to the belief–reward manifold. Thus, $\hat{\mathbb{P}}_t \neq \bar{\mathbb{P}}_t$ and $D_{\mathrm{KL}}\left(\hat{\mathbb{P}}_t \,\|\, \bar{\mathbb{P}}_t\right) > 0$. Hence, $D_{\mathrm{KL}}\left(\hat{\mathbb{P}}_t \,\|\, \mathbb{P}\right)$ attends unique minimum for $\mathbb{P} = \bar{\mathbb{P}}_t$.

B.4 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 4 (Absence of singularities). *The distribution families $\mathcal{P}$ and $\mathcal{Q}$ are defined over the sets $Supp(\mathcal{P}) \triangleq \{a : p(a) > 0, \forall p \in \mathcal{P}\}$ and $Supp(\mathcal{Q}) \triangleq \{a : q(a) > 0, \forall p \in \mathcal{P}\}$ respectively. Moreover, none of the supports are empty and $Supp(\mathcal{P}) \subseteq Supp(\mathcal{Q})$.*

B.5 Implications of Alternating Projections

Definition 4 (I-projection). *The* information projection *(or* I-projection*) of a distribution $\bar{\mathbb{Q}} \in \mathcal{Q}$ onto a non-empty, closed, convex set $\mathcal{P}$ of probability distributions, $\mathbb{P}^a$'s, defined on a fixed support set is defined by the probability distribution $\mathbb{P}^{a*} \in \mathcal{P}$ that has minimum KL-divergence to q: $\mathbb{P}^{a*} \triangleq \arg\min_{\mathbb{P}^a \in \mathcal{P}} D_{\mathrm{KL}}(\mathbb{P}^a \,\|\, \bar{\mathbb{Q}})$.*

Since $D_{\mathrm{KL}}(p(s) \,\|\, q(s)) = -h(p(s)) + H(p(s), q(s))$, 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.

In the last part (Lines 8–9), the updated beliefs are used to obtain the pseudobelief-focal-reward distribution using rI-projection. Following Theorem 1, rI-projection would lead to a unique pseudobelief-focal-reward distribution for a given set of belief-rewards and exposure $\tau(t)$. Here, BelMan is inducing the exploitative bias. It keeps the pseudobelief-focal-reward 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-reward distribution to the distribution of the arm having

Definition 5 (rI-projection). *The* reverse information projection *(or* rI-projection*) of a distribution $\mathbb{P}^a \in \mathcal{P}$ 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 $\bar{\mathbb{Q}}^* \in \mathcal{Q}$ that has minimum KL-divergence from $\mathbb{P}^a$:* $\bar{\mathbb{Q}}^* \triangleq \arg\min_{\bar{\mathbb{Q}} \in \mathcal{Q}} D_{\mathrm{KL}}(\mathbb{P}^a \,\|\, \bar{\mathbb{Q}})$.

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 [Csi84].

B.6 Law of Convergence for the Pseudobelief-reward Distribution

We are simultaneously approximating the belief–reward parameters as well as the pseudobelief–reward parameters. If we look into the belief update step (Equation 1), we observe that the belief distribution of each arm $b^a_{\xi_t}(\theta)$ is updated by incorporating i.i.d samples obtained from the reward distribution of that arm. Let us assume that BelMan has played total T times and any arm a for t^a_T times. Since we are doing naïve Bayesian updates with i.i.d. samples, the belief distributions will follow central limit theorem. This means that if $\tilde{\mu}^a_{t^a}$ is the estimate of the expectation parameters of the belief distribution of arm a constructed from samples $\{X^a_i\}_{i=1}^{t^a_T}$, $\sqrt{t^a_T}(\tilde{\mu}^a_{t^a_T} - \mu^a)$ converges in distribution to a centred normal random vector in $\mathcal{N}(0, \Sigma^a)$. In Theorem 2, we show that the estimator of the mean parameters of pseudobelief is also consistent with these estimators and satisfies central limit theorem.

Theorem 2 (Central limit theorem). *If $\tilde{\bar{\mu}}_T \triangleq \frac{1}{K}\sum_{a=1}^K \tilde{\mu}^a_{t^a_T}$ is estimator of the expectation parameters of the pseudobelief distribution, $\sqrt{T}(\tilde{\bar{\mu}}_T - \bar{\mu})$ converges in distribution to a centred normal random vector in $\mathcal{N}(0, \bar{\Sigma})$. The covariance matrix $\bar{\Sigma} = \sum_{a=1}^K \lambda_a \Sigma^a$ such that $\frac{T}{K^2 t^a_T}$ tends to λ^a as $T \to \infty$.*

Proof. The characteristics function for $\sqrt{N}(\tilde{\bar{\mu}}_N - \bar{\mu})$ is

$$\begin{aligned}
\boldsymbol{\Phi}_{\sqrt{T}(\tilde{\bar{\mu}}_T-\bar{\mu})}(t) &= \mathbb{E}\left[\exp(\iota\langle t, \sqrt{T}(\tilde{\bar{\mu}}_T - \bar{\mu})\rangle)\right] \\
&= \mathbb{E}\left[\exp(\iota\langle t, \frac{\sqrt{T}}{K}\sum_{a=1}^{K}(\tilde{\mu}^a_{t^a_T} - \mu^a)\rangle)\right] \\
&= \prod_{a=1}^{K}\mathbb{E}\left[\exp(\iota\langle t, \frac{\sqrt{T}}{K}(\tilde{\mu}^a_{t^a_T} - \mu^a)\rangle)\right] \\
&= \prod_{a=1}^{K}\mathbb{E}\left[\exp(\iota\langle \frac{\sqrt{T}}{K\sqrt{t^a_T}}t, \sqrt{t^a_T}(\tilde{\mu}^a_{t^a_T} - \mu^a)\rangle)\right] \\
&= \prod_{a=1}^{K}\boldsymbol{\Phi}_{\sqrt{t^a_T}(\tilde{\mu}^a_{t^a_T}-\mu^a)}\left(\frac{\sqrt{T}}{K\sqrt{t^a_T}}t\right)
\end{aligned}$$

Since each of the $\sqrt{t^a_T}(\tilde{\mu}^a_{t^a_T} - \mu^a)$ converges in distribution to a random vector that follows $\mathcal{N}(0, \Sigma^a)$, the covariance matrix for $\sqrt{T}(\tilde{\bar{\mu}}_T - \bar{\mu})$ would be $\lim_{T\to\infty}\sum_{a=1}^{K}\left(\frac{\sqrt{T}}{K\sqrt{t^a_T}}\right)^2\Sigma^a = \sum_{a=1}^{K}\lambda^a\Sigma^a \triangleq \bar{\Sigma}$.

B.7 Proof of Theorem 3

Theorem 3 (Asymptotic consistency). *Given $\tau(t) = \frac{1}{\log t + c\times\log\log t}$ for any $c \geqslant 0$, BelMan will asymptotically converge to choosing the optimal arm in case of a bandit with bounded reward and finite arms. Mathematically, if there exists $\mu^* \triangleq \max_a \mu(\theta_a)$,*

$$\lim_{T\to\infty}\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T}\,X_{a_t}\right] = \mu^*. \tag{3}$$

We reformulate this result more precisely using Lemma 2.

Lemma 2. *If Assumption 3 is true and there exists at least an optimal arm with expected reward $\mu^* \triangleq \max_a \mu(\theta_a)$, and the exposure satisfies $\lim_{t\to\infty}\tau(t) \leqslant \frac{1}{\sqrt{2C}}$, then BelMan would satisfy asymptotic consistency*

$$\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(X_{A_t})\right] = \mu^*. \tag{10}$$

Proof. Without loss of generality, let us consider that there exists at least one optimal arm and it is identified as the arm $a = 1$. At the I-projection step, we choose the arm that has minimum KL-divergence $D_{\mathrm{KL}}\left(\mathbb{P}^a_t(X,\theta)\,\|\,\bar{\mathbb{Q}}(X,\theta)\right)$ from the pseudobelief–focal distribution. Thus, we have to prove that for large t and for all $a \neq 1$,

$$\lim_{t\to\infty}\mathbb{P}(D_{\mathrm{KL}}\left(\mathbb{P}^1_t(X,\theta)\,\|\,\bar{\mathbb{Q}}(X,\theta)\right) - D_{\mathrm{KL}}\left(\mathbb{P}^a_t(X,\theta)\,\|\,\bar{\mathbb{Q}}(X,\theta)\right) < 0) = 1.$$

This is equivalent to proving that almost surely

$$\lim_{t\to\infty} \left(D_{\mathrm{KL}} \left(\mathbb{P}_t^1(X,\theta) \,\|\, \bar{\mathbb{Q}}(X,\theta) \right) - D_{\mathrm{KL}} \left(\mathbb{P}_t^a(X,\theta) \,\|\, \bar{\mathbb{Q}}(X,\theta) \right) \right) < 0. \tag{11}$$

We begin as follows,

$$\begin{aligned}
&D_{\mathrm{KL}} \left(\mathbb{P}_t^1(X,\theta) \,\|\, \bar{\mathbb{Q}}(X,\theta) \right) - D_{\mathrm{KL}} \left(\mathbb{P}_t^a(X,\theta) \,\|\, \bar{\mathbb{Q}}(X,\theta) \right) \\
&= \underbrace{\int_X \int_\theta \mathbb{P}_t^1(X,\theta) \log \mathbb{P}_t^1(X,\theta) \; \mathrm{d}\theta \; \mathrm{dX} - \int_X \int_\theta \mathbb{P}_t^a(X,\theta) \log \mathbb{P}_t^a(X,\theta) \; \mathrm{d}\theta \; \mathrm{dX}}_{\mathbf{T1}} \\
&\quad + \underbrace{\int_X \int_\theta \left[\mathbb{P}_t^a(X,\theta) - \mathbb{P}_t^1(X,\theta) \right] \log \bar{\mathbb{Q}}(X,\theta) \; \mathrm{d}\theta \; \mathrm{dX}}_{\mathbf{T2}}
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{T1} &= \int_X \int_\theta \mathbb{P}_t^1(X,\theta) \log \mathbb{P}_t^1(X,\theta) \; \mathrm{d}\theta \; \mathrm{dX} - \int_X \int_\theta \mathbb{P}_t^a(X,\theta) \log \mathbb{P}_t^a(X,\theta) \; \mathrm{d}\theta \; \mathrm{dX} \\
&= \int_X \int_\theta \left[\mathbb{P}_t^a(X,\theta) - \mathbb{P}_t^1(X,\theta) \right] \log \mathbb{P}_t^1(X,\theta) \; \mathrm{d}\theta \; \mathrm{dX} \; - \; D_{\mathrm{KL}} \left(\mathbb{P}_t^a(X,\theta) \,\|\, \mathbb{P}_t^1(X,\theta) \right) \\
&\underset{(a)}{\leqslant} \int_X \int_\theta \left[\mathbb{P}_t^a(X,\theta) - \mathbb{P}_t^1(X,\theta) \right] \log \mathbb{P}_t^1(X,\theta) \; \mathrm{d}\theta \; \mathrm{dX} \\
&\underset{(b)}{\leqslant} \int_X \int_\theta \left| \left[\mathbb{P}_t^a(X,\theta) - \mathbb{P}_t^1(X,\theta) \right] \log \mathbb{P}_t^1(X,\theta) \right| \; \mathrm{d}\theta \; \mathrm{dX} \\
&\underset{(c)}{\leqslant} \sup_{X,\theta} \left| \log \mathbb{P}_t^1(X,\theta) \right| \int_X \int_\theta \left| \mathbb{P}_t^a(X,\theta) - \mathbb{P}_t^1(X,\theta) \right| \; \mathrm{d}\theta \; \mathrm{dX} \\
&\underset{(d)}{\leqslant} \sup_{X,\theta} \left| \log \mathbb{P}_t^1(X,\theta) \right| \sqrt{\frac{\log 2}{2} \; D_{\mathrm{KL}} \left(\mathbb{P}_t^a(X,\theta) \,\|\, \mathbb{P}_t^1(X,\theta) \right)}
\end{aligned}$$

The inequality (a) is due to the non-negativity of KL-divergence. Inequality (b) is derived from the monotonicity of integrals. This means that if $f \leqslant g$ for all $w \in W$ then $\int_{w\in W} f(w) \; \mathrm{dw} \; \leqslant \int_{w\in W} \; g(w) \; \mathrm{dw}$. Boundedness of the log-likelihood function of the pseudobelief-reward as stated in Proposition 3 results to inequality (c). Inequality (d) is obtained from Pinsker's inequality [CT12].

Similarly, we get for the second term **T2**:

$$\begin{aligned}
\mathbf{T2} &= \int_X \int_\theta \left[\mathbb{P}_t^a(X,\theta) - \mathbb{P}_t^1(X,\theta)\right] \log \bar{\mathbb{Q}}(X,\theta)\ \mathrm{d}\theta\ \mathrm{dX}) \\
&= \int_X \int_\theta \left[\mathbb{P}_t^a(X,\theta) - \mathbb{P}_t^1(X,\theta)\right] \log \left(\prod_a \mathbb{P}_t^a(X,\theta)^{\lambda_t^a}\right)\ \mathrm{d}\theta\ \mathrm{dX} \\
&- \frac{1}{\tau(t)} \mathbb{E}_{\mathbb{P}_t^1(X,\theta)-\mathbb{P}_t^a(X,\theta)}\left[X\right] + \log \bar{Z}_t \times \mathbb{E}_{\mathbb{P}_t^1(X,\theta)-\mathbb{P}_t^a(X,\theta)}\left[1\right] \\
&\underset{(e)}{\leqslant} \sup_{X,\theta} \left|\log \mathbb{P}_t^1(X,\theta)\right| \sqrt{\frac{\log 2}{2}}\ \sqrt{D_{\mathrm{KL}}\left(\mathbb{P}_t^a(X,\theta) \,\|\, \mathbb{P}_t^1(X,\theta)\right)} - \frac{\Delta_t^a}{\tau(t)}.
\end{aligned}$$

Here, $\Delta_t^a \triangleq \mu_t^1 - \mu_t^a$, which means the difference between the expected reward of the optimal arm and the suboptimal arm a. Inequality (e) is obtained by applying AM-GM inequality, inequalities (a), (b), (c), and (d) in sequence. Thus,

$$\begin{aligned}
\mathbf{T1} + \mathbf{T2} &\leqslant \sup_{X,\theta} \left|\log \mathbb{P}_t^1(X,\theta)\right| \sqrt{2\log 2}\sqrt{D_{\mathrm{KL}}\left(\mathbb{P}_t^a(X,\theta) \,\|\, \mathbb{P}_t^1(X,\theta)\right)} - \frac{\Delta_t^a}{\tau(t)} \\
&= \sqrt{2\log 2}\sqrt{D_{\mathrm{KL}}\left(\mathbb{P}_t^a(X,\theta) \,\|\, \mathbb{P}_t^1(X,\theta)\right)} \left(\sup_{X,\theta} \left|\log \mathbb{P}_t^1(X,\theta)\right| - \frac{1}{\tau(t)} \frac{\Delta_t^a}{\sqrt{D_{\mathrm{KL}}\left(\mathbb{P}_t^a(X,\theta) \,\|\, \mathbb{P}_t^1(X,\theta)\right)}}\right) \\
&\leqslant \sqrt{2\log 2}\sqrt{D_{\mathrm{KL}}\left(\mathbb{P}_t^a(X,\theta) \,\|\, \mathbb{P}_t^1(X,\theta)\right)} \left(\sup_{X,\theta} \left|\log \mathbb{P}_t^1(X,\theta)\right| - \frac{1}{\sqrt{2}\tau(t)}\right)
\end{aligned}$$

If we consider $\lim_{t\to\infty}$ for both sides of the inequality, we observe Equation 11 is true if

$$\lim_{t\to\infty} \left(\sup_{X,\theta} \left|\log \mathbb{P}_t^1(X,\theta)\right| - \frac{1}{\sqrt{2}\tau(t)}\right) < 0.$$

This holds as $D_{\mathrm{KL}}\left(\mathbb{P}_t^a(X,\theta) \,\|\, \mathbb{P}_t^1(X,\theta)\right) > 0$ for all a and t. By Assumption 4, we get $\lim_{t\to\infty} \sup_{X,\theta} \left|\log \mathbb{P}_t^1(X,\theta)\right| \leqslant C + \log \mathbb{P}_t^1(X,\theta) = C'(\text{say})$. Thus, we get in order to satisfy the inequality $\lim_{t\to\infty} \tau(t) < \frac{1}{\sqrt{2}C'}$ which is in our premise.

Lemma 3. *For* $\tau(t) = \frac{1}{\log t + c \log\log t}$ *with* $c \geqslant 0$, $\lim_{t\to\infty} \tau(t) < \frac{1}{C}$ *for any* $C < \infty$.

Proof. Since $\lim_{t\to\infty} \frac{1}{\log t + c\log\log t} = 0$, the aforementioned claim holds true.

Lemma 2 and 3 together prove Theorem 3. This proves that BelMan is asymptotically consistent for finite-arm stochastic bandit problems.

For exploration–exploitation bandit problem, we observe that $\tau(t)$ has to be a positive valued function of time t that asymptotically decreases with time. Such decay in the value of exposure $\tau(t)$ adaptively increases the importance of reward maximisation over minimising the KL-divergence between the belief-reward of selected arm and the pseudobelief-reward. This mechanism allows BelMan to adaptively balance between the exploration and exploitation components.

The growth rate proposed for exposure, $O(\frac{1}{\log t})$, is a loose bound. Beside this, it is also distribution independent. Thus, we observe a gap between the bound on exposure growth obtained here, and the one used in practice. It would be interesting to find out tighter bounds with more specific constants for given reward distributions.

B.8 BelMan for Exponential Family Distributions

As mentioned in Section B.2, *exponential family* [Bro86] 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:

$$f_\theta(X) \triangleq h(X) \exp\left(\langle \omega(\theta), T(X)\rangle - A(\theta)\right).$$

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 [Bro86]. Two parametric distributions $f_{\boldsymbol{\theta}}(X)$ and $b_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ are conjugate if the posterior distribution $\mathbb{P}(\boldsymbol{\theta}|X)$ formed by multiplying them has the same form as $b_{\boldsymbol{\eta}}(\boldsymbol{\theta})$. Thus, if the reward distribution belongs to the exponential family, the belief distribution is represented as: $b_{\boldsymbol{\eta}}(\theta) \triangleq h(\theta) \exp\left(\langle \boldsymbol{\eta}, T(\theta)\rangle - A(\boldsymbol{\eta})\right)$ with the natural parameters $\boldsymbol{\eta}$.

Since exponential family distributions are flat with respect to KL-divergence [AN07], both I-and rI-projections in BelMan are well-defined and unique. Thus, at each timestep, we obtain an optimal and unambiguous choice of the arm and pseudo-belief respectively. [AN07] 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 parametrisation. 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 $b_{\boldsymbol{\eta}}(\theta)$ construct a statistical manifold with local co-ordinates $\boldsymbol{\eta}$ [AN07]. Theorem 1 and 2 validate these claims in case of BelMan.

Bernoulli Bandits. In the case of Bernoulli bandits, we assume that drawing an arm returns the rewards 1 and 0 with probability θ and $1-\theta$ respectively. Thus, the reward distribution of the a^{th} arm is $f_{\theta_a}(X) \triangleq \text{Ber}(\theta_a)$. 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 $\{\alpha^a\}_{a=1}^K$ and $\{\beta^a\}_{a=1}^K$. After t-timesteps the prior over the probability of success of the a^{th} arm is

$$b_t^a(\theta_a) \triangleq \text{Beta}(\theta_a; \alpha_t^a, \beta_t^a) = \frac{1}{B(\alpha_t^a, \beta_t^a)} \theta_a^{\alpha_t^a - 1} (1-\theta_a)^{\beta_t^a - 1},$$

for $\alpha_t^a, \beta_t^a > 0$ and $\theta_a \in (0,1)$. Here, α_t^a and β_t^a are the number of successes and failures, respectively, for the arm a till timestep t. We begin with both α_0^a and β_0^a to be 1 for all arms. This amounts to the uniform distribution over 0 and 1. This initialisation 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

$$\sum_{a=1}^{K} D_{\mathrm{KL}}\left(\mathbb{P}_t^a(X,\theta) \,\|\, \bar{\mathbb{Q}}_{t-1}(X,\theta)\right)$$
$$= \sum_{a=1}^{K}[-\frac{1}{\tau(t)}\frac{\alpha_t^a}{N_t^a} - \log\left(B\left(\alpha_t^a, \beta_t^a\right)\right) + (\alpha_t^a - \bar{\alpha}_{t-1})\Psi(\alpha_t^a) + (\beta_t^a - \bar{\beta}_{t-1})\Psi(\beta_t^a)$$
$$- (N_t^a - \bar{N}_{t-1})\Psi(N_t^a)] + K\log\left(\frac{\bar{\alpha}_{t-1}\exp(\frac{1}{\tau(t)}) + \bar{\beta}_{t-1}}{\bar{N}_{t-1}}\right) + K\log\left(B\left(\bar{\alpha}_{t-1}, \bar{\beta}_{t-1}\right)\right).$$

Here, $N_t^a = \alpha_t^a + \beta_t^a$ is the total number of times the jth arm is played till the nth timestep, $\bar{N} = \bar{\alpha} + \bar{\beta}$ and Ψ is the digamma function [Ber76] defined as the derivative of the logarithm of gamma function, i.e. $\frac{d}{da}\left(\log \Gamma(a)\right)$.

In Line 4 of Algorithm 1, we first perform the I-projection to decide which arm a_t to draw to minimise 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 minimise the objective and update the pseudobelief accordingly. The presence of pseudobelief offers BelMan a chance to explore the less successful arms to minimise 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* θ_a, the reward distribution of arm a of exponential bandit is $f_{\theta_a}(X) \triangleq \theta_a \exp(-\theta_a X)$ for $X \in [0, \infty)$. Following the structure of Sections B.8 and the previous Bernoulli case, we obtain the gamma distribution, another member of the exponential family, as the conjugate prior. After the t^{th} timestep, the belief distribution corresponding to a^{th} arm is expressed as

$$b_t^a(\theta_a) \triangleq \mathrm{Gamma}(\theta_a; \alpha_t^a, \beta_t^a) = \frac{{\beta_t^a}^{\alpha_t^a}}{\Gamma(\alpha_t^a)} {\theta_a}^{\alpha_t^a - 1} \exp(-\theta_a \beta_t^a),$$

for both shape and rate parameters $\alpha_t^a, \beta_t^a > 0$. Here, α_t^a and β_t^a are, respectively, the number of times the arm a is played and sum of the rewards obtained by playing the arm till timestep t. As we update using Equation (1), we get gamma distributions with parameters $\alpha_{t+1}^a = \alpha_t^a + 1$, and $\beta_{t+1}^a = \beta_t^a + x_t$ if the arm a is played and a reward x_t is obtained. Under this specific setting of gamma prior

and exponential reward, we compute the targeted KL-divergence of BelMan as

$$\sum_{a=1}^{K} D_{\mathrm{KL}}\left(\mathbb{P}_t^a(X,\theta) \,\|\, \bar{\mathbb{Q}}(X,\theta)\right)$$
$$= \sum_{a=1}^{K}[-\frac{1}{\tau(t)}\frac{\alpha_t^a}{\beta_t^a} - \log\left(\Gamma\left(\alpha_t^a\right)\right) + (\alpha_t^a - \bar{\alpha}_{t-1})\Psi(\alpha_t^a) - \frac{\alpha_t^a}{\beta_t^a}(\beta_t^a - \bar{\beta}_{t-1})$$
$$+ \bar{\alpha}_{t-1}\log\beta_t^a] + K\log\bar{Z}_t + K\log\left(\Gamma\left(\bar{\alpha}_{t-1}\right)\right) - K\bar{\alpha}_{t-1}\log\bar{\beta}_{t-1}.$$

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

C Supplementary Material for Section 3 (Empirical Performance Analysis)

Figure 8, 9, and 10 show the evolution of cumulative regret with number of iterations for the three cases whose number of suboptimal arm draws are reported in Figure 2, 3, and 4, respectively.

We also experimented on another 2-arm bandit scenario with means 0.45 and 0.55. Figures 11 depicts the evolution of cumulative regret and suboptimal draws for BelMan and the other competing algorithms. Similar to Figure 11, 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 12 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 11 to Figure 12 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 11 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 [GC11]. We also compare it with Thompson sampling, UCBtuned and uniform sampling method (Random). The results are shown in Figure 13 and 14. Since the formulation is oblivious to boundedness of the distribution, we choose to validate also on unbounded rewards. In Figure 13, it outperforms all the other algorithms. In Figure 14, 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.

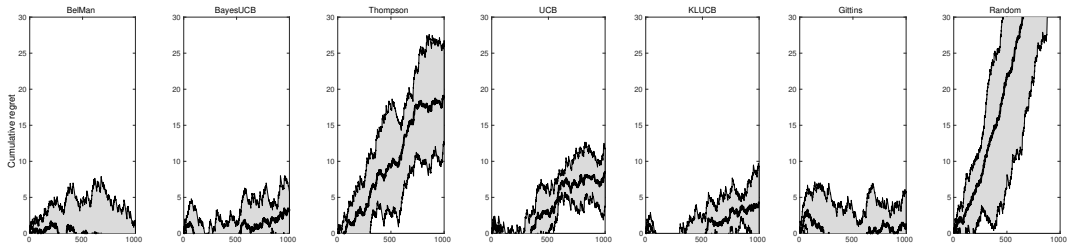


Fig. 8. Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 2-arm Bernoulli bandit with expected rewards 0.8 and 0.9 for 1000 iterations. The dark black line shows the average over 25 runs. The grey area shows the 75 percentile.

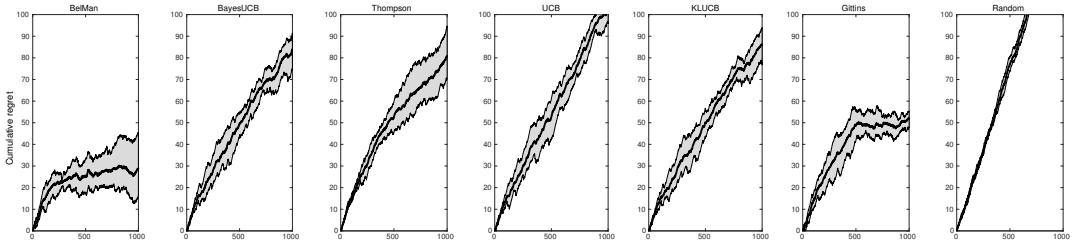


Fig. 9. Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 20-arm Bernoulli bandit with expected rewards [0.25 0.22 0.2 0.17 0.17 0.2 0.13 0.13 0.1 0.07 0.07 0.05 0.05 0.05 0.02 0.02 0.02 0.01 0.01 0.01] for 1000 iterations.

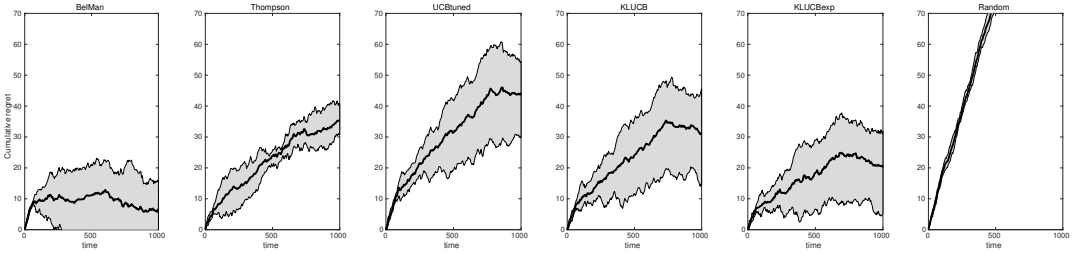


Fig. 10. Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 5-arm bounded exponential bandit with expected rewards 0.2, 0.25, 0.33, 0.5, and 1.0 for 1000 iterations.

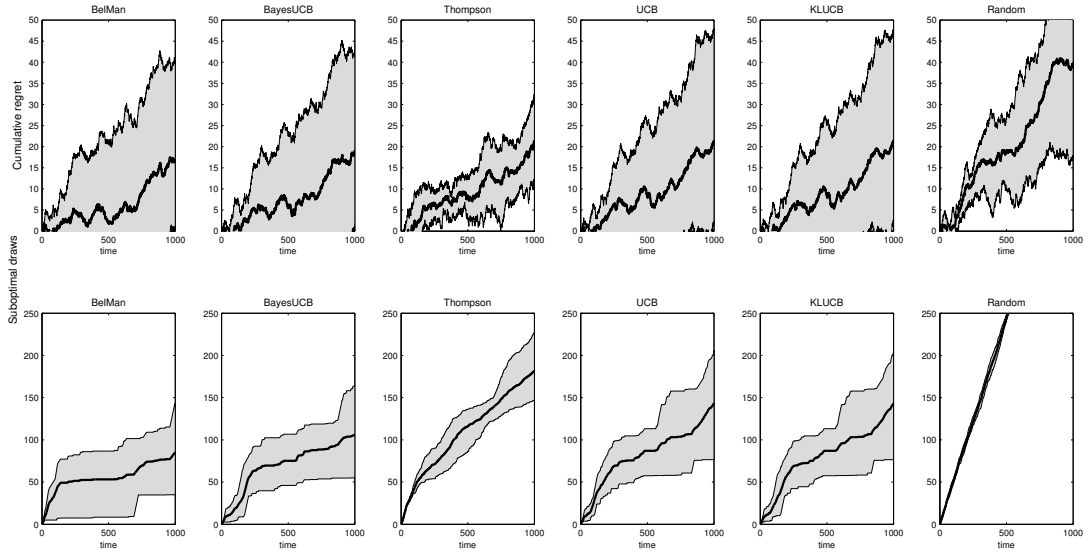


Fig. 11. Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 500 iterations for 2-arm Bernoulli bandit with means 0.45 and 0.55.

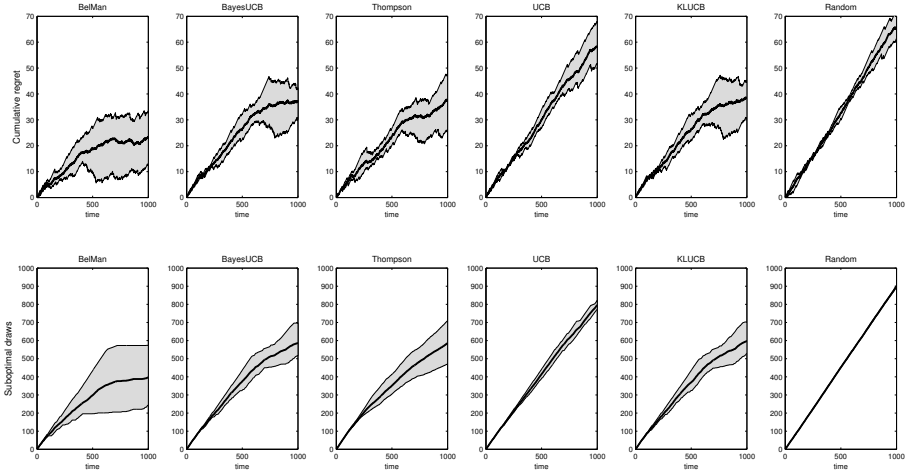


Fig. 12. Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 500 iterations for 10-arm Bernoulli bandit with means $\{0.1, 0.05, 0.05, 0.05, 0.02, 0.02, 0.02, 0.01, 0.01, 0.01\}$. The dark black line shows the average. The grey area shows 75 percentile.

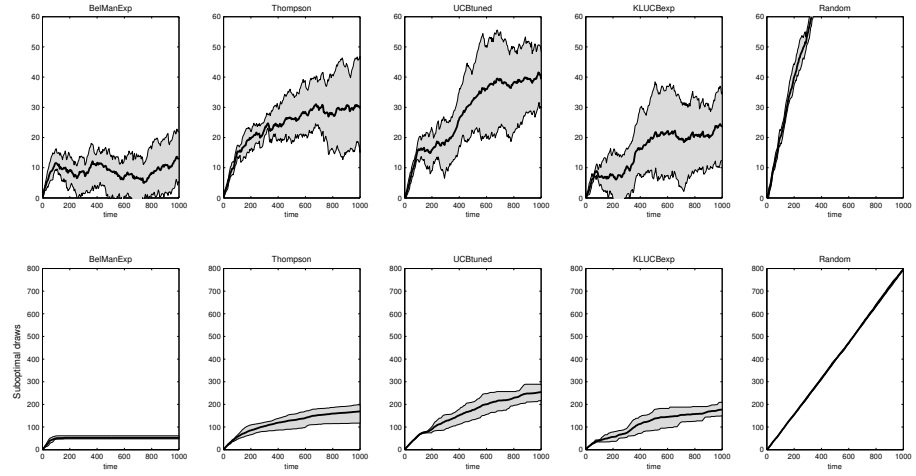


Fig. 13. Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 1000 iterations for 5-arm unbounded exponential bandit with parameters $\{0.2, 0.25, 0.33, 0.5, 1.0\}$.

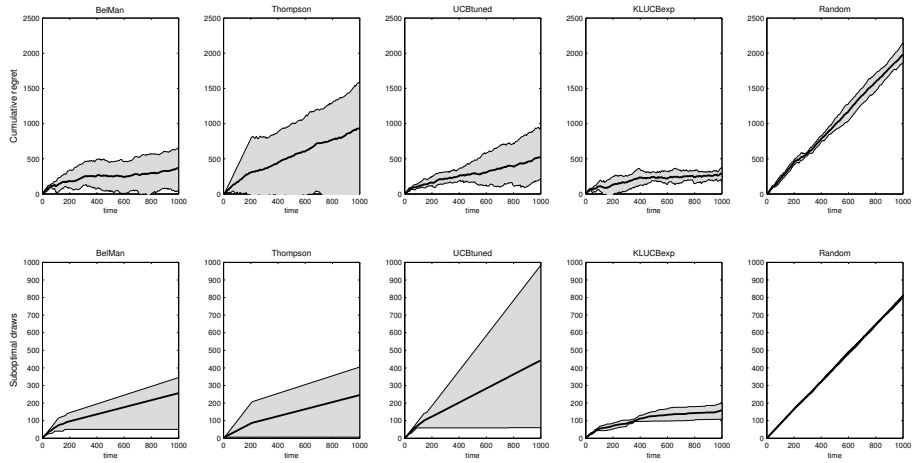


Fig. 14. Evolution of cumulative regret (top), and number of suboptimal draws (bottom) for 1000 iterations for 5-arm unbounded exponential bandit with parameters $\{1, 2, 3, 4, 5\}$.

References for the Appendix

AB10. Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.

ACBF02. Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2–3):235–256, 2002.

AN07. Shun-Ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191 of *Translations of mathematical monographs*. American Mathematical Society, 2007.

Bel56. Richard Bellman. A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics (1933–1960)*, 16(3/4):221–229, 1956.

Ber76. Jose M Bernardo. Algorithm AS 103: Psi (digamma) function. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 25(3):315–317, 1976.

BMS09. Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *ALT*, pages 23–37. Springer, 2009.

Bro86. L. D. Brown. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, 1986.

BWV13. Sébastian Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *ICML*, pages 258–265, 2013.

Csi84. Imre Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, 12(3):768–793, 1984.

CT12. Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Dar35. G. Darmois. Sur les lois de probabilites a estimation exhaustive. *C. R. Acad. Sci. Paris*, 200:1265–1266, 1935.

DB15. Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *NIPS*, pages 2818–2826, 2015.

DeG05. Morris H DeGroot. *Optimal statistical decisions*, volume 82 of *Wiley Classics Library*. John Wiley & Sons, 2005.

FS15. Muhammad Faheem and Pierre Senellart. Adaptive web crawling through structure-based link classification. In *Proc. ICADL*, pages 39–51, Seoul, South Korea, December 2015.

GC11. Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *COLT*, pages 359–376, 2011.

Git79. John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

HT11. Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011.

Kau18. Emilie Kaufmann. On bayesian index policies for sequential resource allocation. *Annals of Statistics*, 46(2):842–865, April 2018.

KCG12. Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On Bayesian upper confidence bounds for bandit problems. In *AISTATS*, pages 592–600, 2012.

KK13. Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *COLT*, pages 228–251, 2013.

Koo36. Bernard Osgood Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409, 1936.

Kul97. Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

Lai88. 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.

LR85. T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, March 1985.

NM11. José Nino-Mora. Computing a classic index for finite-horizon bandits. *INFORMS Journal on Computing*, 23(2):254–267, 2011.

ORVR13. Ian Osband, Dan Russo, and Benjamin Van Roy. (More) efficient reinforcement learning via posterior sampling. In *NIPS*, pages 3003–3011, 2013.

PT17. Sudeep Raja Putta and Theja Tulabandhula. Pure exploration in episodic fixed-horizon Markov decision processes. In *AAMAS*, pages 1703–1704, 2017.

Rob52. Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5):527–535, 09 1952.

Sco10. Steven L Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

Tho33. 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.