BelMan: An Information-Geometric Approach to Stochastic Bandits

Debabrota Basu[1] (✉), Pierre Senellart[2,3], and Stéphane Bressan[4]

[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] 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 theoretically 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]. 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, 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 paramaters 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 manifolds 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. Information geometrically 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 allows 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 the 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, and Thompson sampling.

2 Methodology

Bandit Problem. We consider a finite number $K > 1$ of independent arms. An arm a corresponds to a reward distribution $f^a_\theta(X)$. We assume that the form of the probability distribution $f_.(X)$ is known to the algorithm but the parametrisation $\theta \in \Theta$ 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 $f^a_\theta(X)$ as $f_{\theta_a}(X)$ for specificity. The agent sequentially chooses an arm a_t at each time step 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 her 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 time 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 $\{b^a_t(.)\}_{a=1}^K$, such that $B_t(\theta_1, \ldots, \theta_K) = \prod_{a=1}^K b^a_t(\theta_a) \triangleq \prod_{a=1}^K \mathbb{P}(\theta_a \mid \mathcal{H}_t)$.*

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 iteration is given by*

$$b^a_{t+1}(\theta_a) \propto f_{\theta_a}(x_t) \times b^a_t(\theta_a) \tag{1}$$

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 the* t^{th} *iteration 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 parametrization, the $\mathcal{B}^a$'s can be represented by a single manifold $\mathcal{B}_\theta$.

Lemma 1 (Belief-Reward Manifold). *If the belief-reward distributions* $\mathbb{P}(X,\theta)$ *have smooth probability density functions, their set defines a manifold* $\mathcal{B}_\theta\mathcal{R}$ *. We refer to it as the* belief-reward manifold*. Belief-reward manifold is the product manifold of the belief manifold and the reward manifold, i.e.* $\mathcal{B}_\theta\mathcal{R} = \mathcal{B}_\theta \times \mathcal{R}$*.*

The Bayesian belief update after each of the iteration 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.*

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

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.

Theorem 1. *For 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 for both pure exploration and exploration– exploitation but not sufficient for maximising 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-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.*

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 the normalisation factor $\bar{Z}_t = \int_{X\in\mathbb{R}}\int_{\theta\in\Theta} \bar{\mathbb{P}}(X,\theta)\exp\left(\frac{X}{\tau(t)}\right) \mathrm{d}\theta\mathrm{d}x$. Following Equation (2), we compute the pseudobelief-focal-reward distribution as

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

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

The focal distribution gradually concentrates on higher rewards as the exposure $\tau(t)$ decreases with time. Thus, it constrains using KL-divergence to choose distributions with higher rewards and induces the exploitive bias. From Theorem 3, we obtain $\frac{1}{\tau(t)}$ has to grow in the order $\Omega(\frac{1}{\sqrt{t}})$ for exploration–exploitation bandit problem independent of the family of reward distribution. Following the bounds obtained in [14], we set the exposure $\tau(t) = [\log(t) + C \times \log(\log(t))]^{-1}$ for experimental evaluation, where C is a constant (we choose the value $C = 15$ in the experiments) . 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 $\mathbb{P}^a_t(X, \boldsymbol{\theta})$ 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

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

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 $\tilde{\bar{\mu}}_T \triangleq \frac{1}{K}\sum_{a=1}^K \tilde{\mu}^a_{t^a_T}$ is estimator of the expectation parameters of the pseudobelief distribution, $\sqrt{T}(\tilde{\bar{\mu}}_T - \bar{\mu})$ converges in distribution to a centered 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 $T \to \infty$.*

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.

We need an additional assumption before moving into the asymptotic consistency claim in Theorem 3.

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. Here, $h(b^a_t)$ denotes the entropy of belief distribution b^a_t of arm a at time t. 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 an affine 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 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

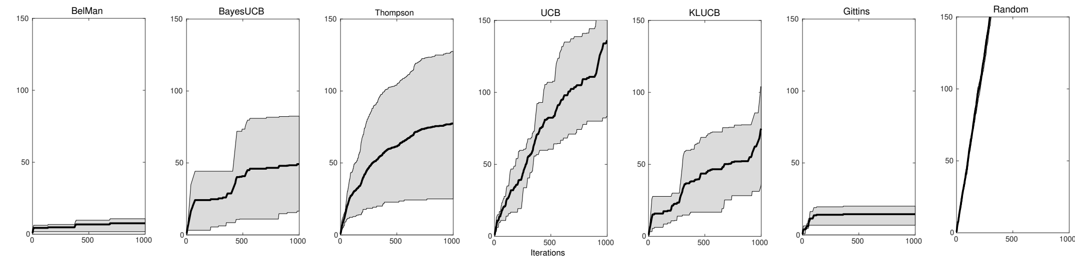


Fig. 1. Evolution of number of suboptimal draws for 2-arm Bernoulli bandit with expected rewards 0.8 and 0.9 for 1000 iterations. The dark black line shows the average over 25 runs. The grey area shows the 75 percentile.

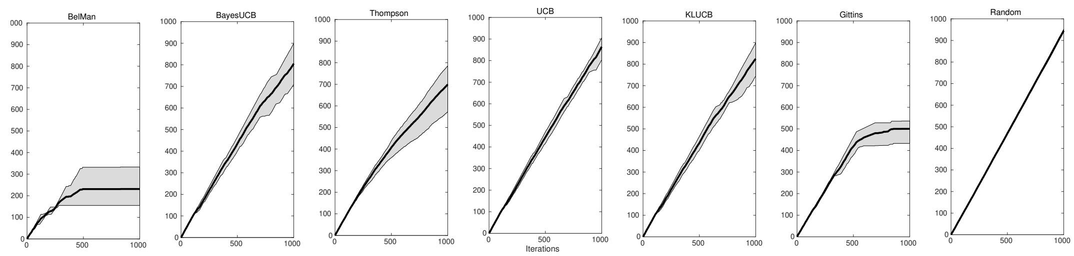


Fig. 2. Evolution of number of suboptimal draws for 20-arm Bernoulli bandit with expected rewards [0.25 0.22 0.2 0.17 0.17 0.2 0.13 0.13 0.1 0.07 0.07 0.05 0.05 0.05 0.02 0.02 0.02 0.01 0.01 0.01] for 1000 iterations.

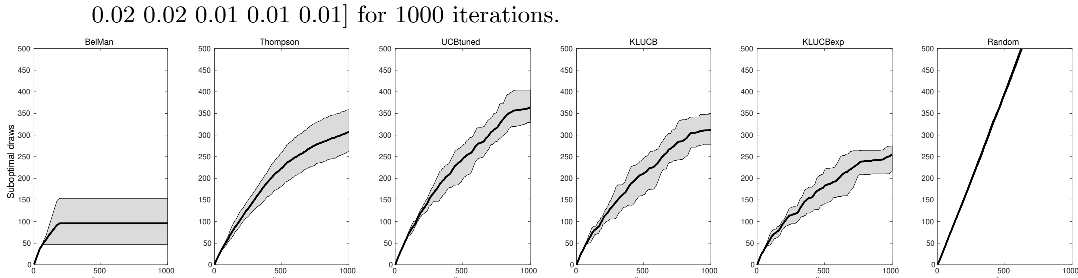


Fig. 3. Evolution of number of suboptimal draws for 5-arm bounded exponential bandit with expected rewards 0.2, 0.25, 0.33, 0.5, and 1.0 for 1000 iterations.

experiments for 25 runs each consisting of 1000 iterations. We begin with uniform distribution over corresponding parameters as the initial prior distribution for all the Bayesian algorithms.

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

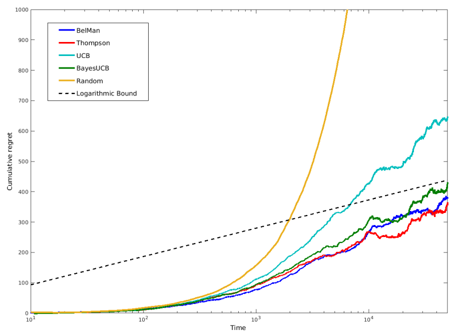


Fig. 4. Evolution of (mean) regret for exploration–exploitation 20-arm Bernoulli bandit setting of Figure 2 with horizon=50,000.

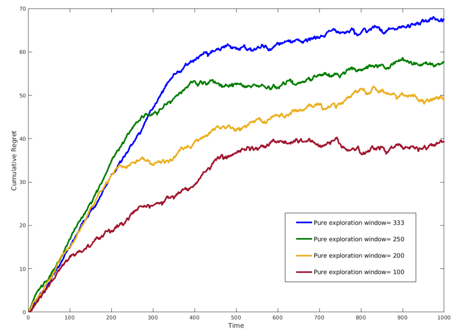


Fig. 5. Evolution of (mean) cumulative regret for two-phase 20-arm Bernoulli bandits.

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

From Figures 1, 2, and 3, 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 2 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 3 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 setting of Figure 2 to verify the asymptotic behaviour of BelMan. Figure **??** shows that BelMan's regret gradually becomes linear with respect to the logarithmic axis. Figure **??** empirically validates BelMan to achieve logarithmic regret like the competitors which are theoretically proven to reach logarithmic regret.

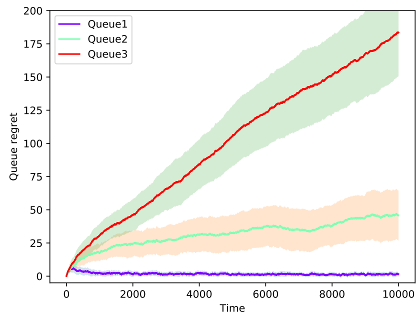
Two-phase reinforcement learning problem. In this experiment, we simulate a two-phase setup, as in [29]: the agent first does pure exploration for a fixed number of iterations, then move to exploration–exploitation. This is possible since BelMan supports both modes and can transparently switch. The setting is that of the 20-arm Bernoulli bandit in Figure 2. 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 5. 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 2, 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.

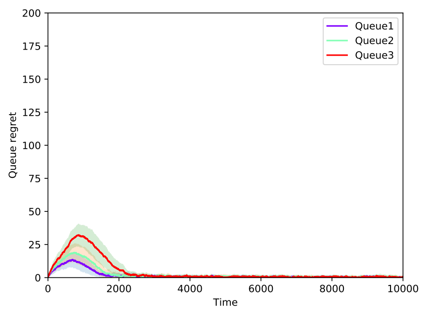
4 Application to Queueing Bandits

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].

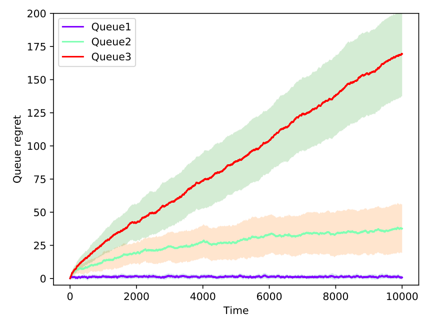
We consider a discrete-time queueing system with 1 queue and K servers. The servers are indexed by $a \in \{1, \dots, K\}$. Arrivals to the queue and service offered by the servers are assumed to be independent and identically distributed across time. The mean arrival rate is $\lambda \in \mathbb{R}^+$. 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, $\lambda < \max\limits_{a \in [K]} \mu_a$. 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



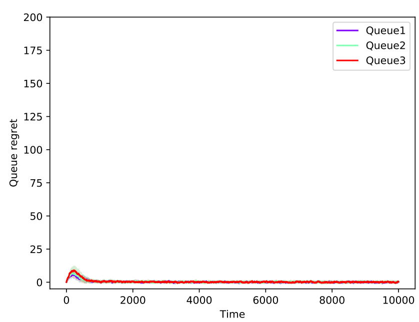
(a) Q-ThS



(b) Q-UCB



(c) Thompson sampling



(d) BelMan

Fig. 6. 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.

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 $Q : [T] \rightarrow \mathbb{R}^{\geqslant 0}$, $A : [T] \rightarrow \mathbb{R}^{\geqslant 0}$, and $S : [T] \rightarrow \mathbb{R}^{\geqslant 0}$. 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 service rates are unknown to it. This create 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 $A\left(t\right)$ 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 6 depict that BelMan is more stable and efficient than the competing algorithms: Q-UCB, Q-Thompson sampling, and Thompson sampling. We observe that in queues 2 and 3 the average service rates are lower than the corresponding arrival rates. Due to this inherent constraint, the queue 2 and 3 can have unstable queueing systems if the initial exploration of the algorithm does not damp fast enough. Though the randomisation of Thompson sampling is good for exploration but in this case playing the suboptimal servers can induce instability which affects the total performance in future.

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 available information on the reward generation process in form of a prior distribution. For 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.

Acknowledgement

We would like to thank Jonathan Scarlett for valuable discussions. This work is partially supported by WASP-NTU grant, the National University of Singapore Institute for Data Science project WATCHA, and Singapore Ministry of Education project Janus.

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)