BelMan: An Information-Geometric Approach to Stochastic Bandits

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

[1]Data Science and AI Division, Chalmers University of Technology, Sweden
[2]DI, Ecole Normale Superieure, and INRIA, Paris, France
[3]School of Computing, National University of Singapore, Singapore

ECML PKDD 2019
September 19, 2019

Roadmap

1. Bandits: Primer

2. BelMan
    - Information Representation: Belief-Reward Manifold
    - Information Accumulation: Pseudobelief-reward Distribution
    - Information Exploitation: Focal and Pseudobelief-focal Distributions
    - Algorithm: Alternating Information Projections
    - Experiments: Logarithmic Regret Growth

3. Queuing Bandits: An Application

4. Wrap-up: Take-away

. . .

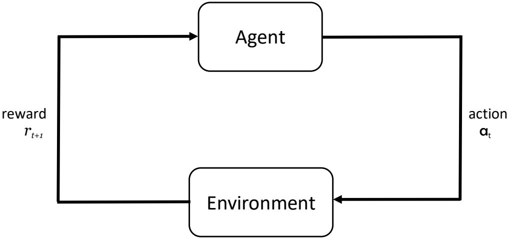
K one-armed bandits.

K one-armed bandits.

K independent arms with unknown stationary distributions such that the rewards are independently sampled from an identical distribution (I.I.D).

K one-armed bandits.

K independent arms with unknown stationary distributions such that the rewards are independently sampled from an identical distribution (I.I.D).



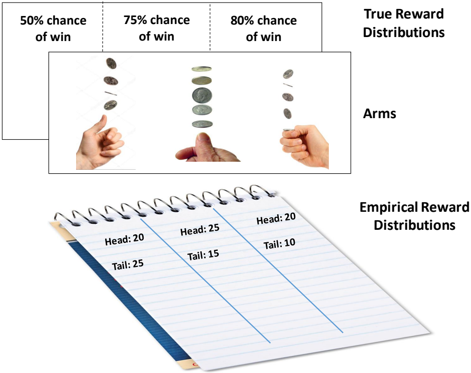
Agent

Environment

reward r_{t+1}

action a_t

A Bandit Game



50% chance of win

75% chance of win

80% chance of win

True Reward Distributions

Arms

Empirical Reward Distributions

Goal: Maximise the number of heads by 500 tosses.

The Stochastic Bandit

At each time step t of the bandit game,

- the agent $\mathcal{A}$ chooses an arm $a_t \in \{1, \ldots, K\}$,
- samples a reward R_t from the reward distribution of arm a_t.

Maximising the cumulative reward

The goal of the agent $\mathcal{A}$ is to compute a policy that *maximises the cumulative reward*, which is the sum of the expected rewards of the arms played till time T.

$$S(\mathcal{A}, T) \triangleq \sum_{a=1}^{K} \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\left(R_{A_t} \times \underbrace{\mathbb{1}(A_t = a)}_{\text{Arm } a \text{ is played}}\right)\right]}_{\text{Expected reward from arm } a \text{ by time } T}$$

Cumulative Regret: The Price of Incomplete Information

The optimal algorithm OPT knows the expected reward of all the arms and draws the arm a^* with maximum expected reward.

$$S(\mathrm{OPT}, T) \triangleq \underbrace{\mathbb{E}[R_{a^*}]}_{\text{Expected reward of arm } a^*} \times \underbrace{T}_{\text{Arm } a^* \text{ is played}}$$

Cumulative Regret: The Price of Incomplete Information

The optimal algorithm OPT knows the expected reward of all the arms and draws the arm a^* with maximum expected reward.

$$S(\mathrm{OPT}, T) \triangleq \underbrace{\mathbb{E}\left[R_{a^*}\right]}_{\text{Expected reward of arm } a^*} \times \underbrace{T}_{\text{Arm } a^* \text{ is played}}$$

Cumulative regret is the deficit of cumulative reward obtained by a bandit algorithm with respect to the optimal algorithm.

$$Reg(\mathcal{A}, T) = S(\mathrm{OPT}, T) - S(\mathcal{A}, T).$$

Minimising the cumulative regret

The goal of the agent $\mathcal{A}$ is to minimise the cumulative regret $Reg(\mathcal{A}, T)$ for a given time T.

Two Sides of a Bandit: Exploration and Exploitation

Pure Exploration
Draw each arm uniformly at random and accumulating empirical knowledge.

Pure Exploitation
Draw the arm with maximum expected reward as per present knowledge.

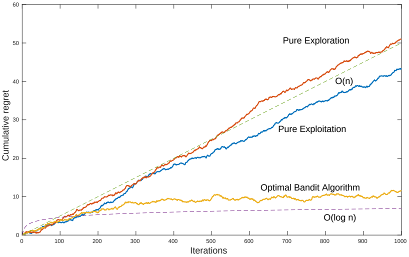
Explore Then Commit
Commit to exploitation after an initial window of exploration (Garivier et al., 2016).

Two Sides of a Bandit: Exploration and Exploitation

Pure Exploration
Draw each arm uniformly at random and accumulating empirical knowledge.

Pure Exploitation
Draw the arm with maximum expected reward as per present knowledge.

Explore Then Commit
Commit to exploitation after an initial window of exploration (Garivier et al., 2016).

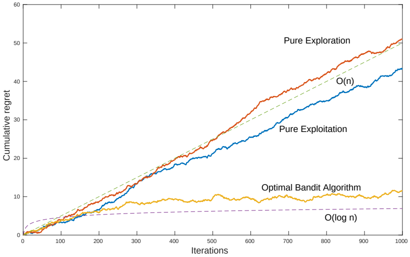


Two Sides of a Bandit: Exploration and Exploitation

Pure Exploration
Draw each arm uniformly at random and accumulating empirical knowledge.

Pure Exploitation
Draw the arm with maximum expected reward as per present knowledge.

Explore Then Commit
Commit to exploitation after an initial window of exploration (Garivier et al., 2016).



The Trade-off

Exploration and exploitation should be performed simultaneously, and adapted on-the-go to achieve the optimal logarithmic regret.

Roadmap

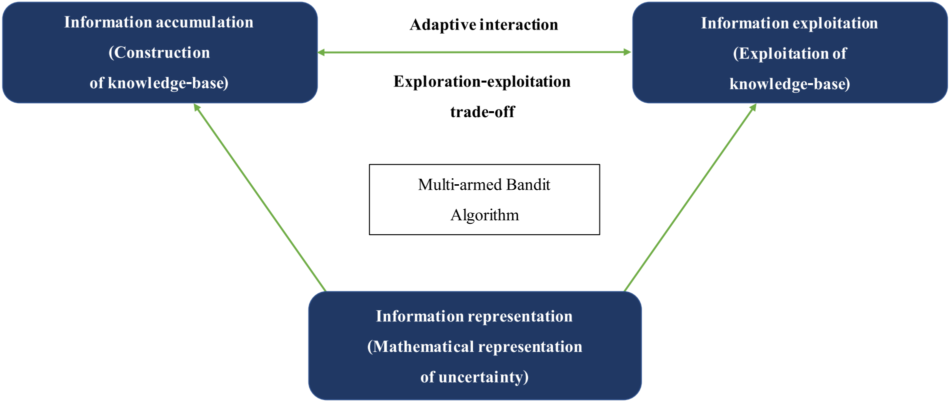
1. Bandits: Primer

2. BelMan
   - Information Representation: Belief-Reward Manifold
   - Information Accumulation: Pseudobelief-reward Distribution
   - Information Exploitation: Focal and Pseudobelief-focal Distributions
   - Algorithm: Alternating Information Projections
   - Experiments: Logarithmic Regret Growth

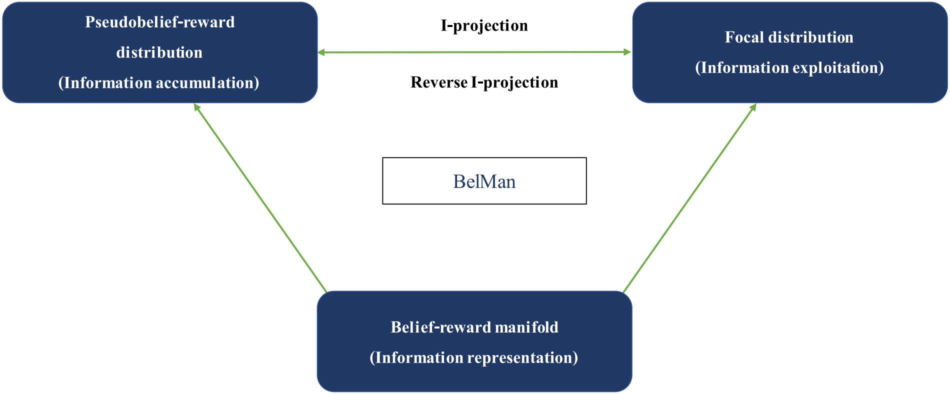
3. Queuing Bandits: An Application

4. Wrap-up: Take-away

Recipe of an Optimal Bandit Algorithm → BelMan



Recipe of an Optimal Bandit Algorithm → BelMan



Pseudobelief-reward distribution (Information accumulation)

I-projection

Reverse I-projection

Focal distribution (Information exploitation)

BelMan

Belief-reward manifold (Information representation)

Information Representation

Information Representation

We need to express the uncertainty regarding exploration and exploitation using a mathematical representation.

Information Representation

Information Representation

We need to express the uncertainty regarding exploration and exploitation using a mathematical representation.

- The uncertainty of obtaining reward R from arm a is represented by a reward distribution $f_{\boldsymbol{\theta}^a}(R)$ with unknown parameter $\boldsymbol{\theta}^a$.
- The uncertainty over the parameter $\boldsymbol{\theta}^a$ of the reward distribution of arm a is represented by a belief distribution $b_{\boldsymbol{\eta}^a}(\theta^a)$ with empirically computed parameter $\boldsymbol{\eta}^a$.

The belief and reward distributions represent the uncertainties of partial information, and the stochastic nature of reward generation respectively.

Information Representation

Information Representation

We need to express the uncertainty regarding exploration and exploitation using a mathematical representation.

- The uncertainty of obtaining reward R from arm a is represented by a reward distribution $f_{\boldsymbol{\theta}^a}(R)$ with unknown parameter $\boldsymbol{\theta}^a$.
- The uncertainty over the parameter $\boldsymbol{\theta}^a$ of the reward distribution of arm a is represented by a belief distribution $b_{\boldsymbol{\eta}^a}(\theta^a)$ with empirically computed parameter $\boldsymbol{\eta}^a$.

The belief and reward distributions represent the uncertainties of partial information, and the stochastic nature of reward generation respectively.

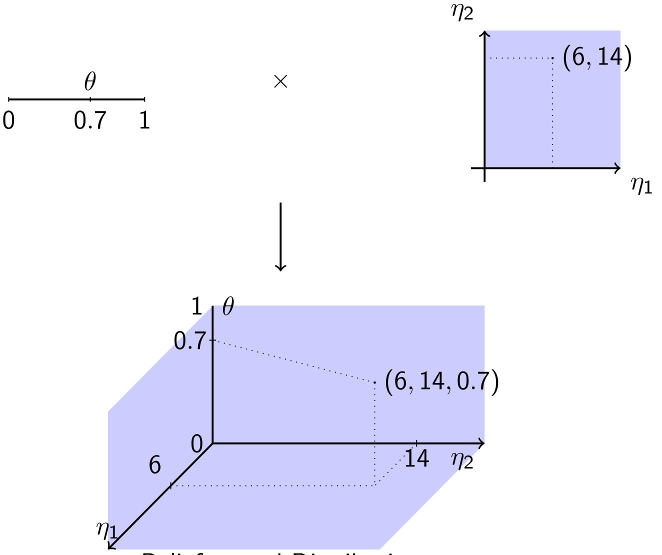
How can we represent and manipulate these uncertainties together?

Reward Distribution

$f_\theta(R) = Ber(R;\theta) = \theta^R(1-\theta)^{(1-R)}$

Belief Distribution

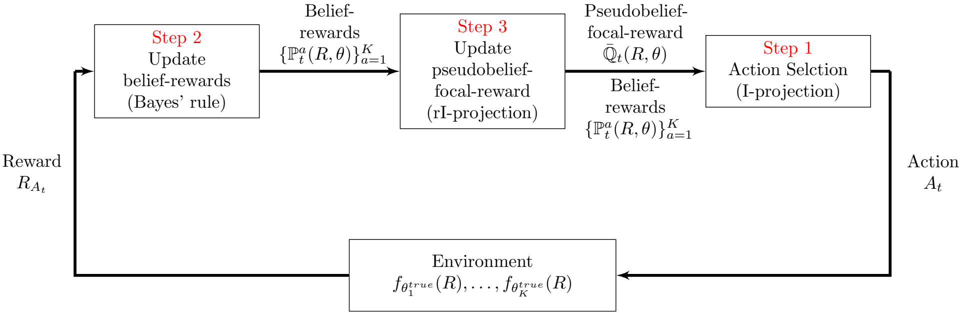
$b_\eta(\theta) = Beta(\theta;\eta_1,\eta_2) = \theta^{\eta_1}(1-\theta)^{\eta_2}$



Belief-reward Distribution

$\mathbb{P}(R,\theta) = Ber(R;\theta) \times Beta(\theta;\eta_1,\eta_2)$

BelMan: Trading-off Information Accumulation and Exploitation



Step 2
Update belief-rewards (Bayes' rule)
Belief-rewards $\{\mathbb{P}_t^a(R,\theta)\}_{a=1}^K$
Step 3
Update pseudobelief-focal-reward (rI-projection)
Pseudobelief-focal-reward $\bar{\mathbb{Q}}_t(R,\theta)$
Belief-rewards $\{\mathbb{P}_t^a(R,\theta)\}_{a=1}^K$
Step 1
Action Selction (I-projection)
Action A_t
Environment
$f_{\theta_1^{true}}(R),\dots,f_{\theta_K^{true}}(R)$
Reward R_{A_t}

Information Accumulation → Pseudobelief–reward

Information Accumulation

We construct a knowledge-base on the belief-reward distributions of the arms using exploration.

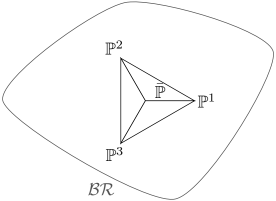
- In information geometry, the *barycenter* of a set of distributions with KL-divergence as a pseudo-distance is *a geometric summary of the accumulated information*.
- We compute the barycenter of as the belief-reward distribution *minimising the KL-divergence from the belief-reward distributions of the arms*.
- We refer to the computed barycenter in the belief-reward manifold as the pseudobelief-reward distribution $\bar{\mathbb{P}}(r, \theta)$.

Pseudobelief-reward: Properties

Pseudobelief-reward

We propose to use the barycenter in the belief-reward manifold as the summary of collective information and to call it pseudobelief-reward distribution.

$$\bar{\mathbb{P}}(R,\theta) \triangleq \operatorname*{argmin}_{\mathbb{P}\in\mathcal{B}_\theta\mathcal{R}} \sum_{a=1}^{K} D_{\mathrm{KL}}\left(\mathbb{P}^a(R,\theta) \| \mathbb{P}(R,\theta)\right)$$

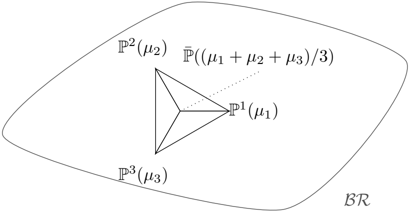


Pseudobelief-reward: Properties

Pseudobelief-reward

We propose to use the barycenter in the belief-reward manifold as the summary of collective information and to call it pseudobelief-reward distribution.

$$\bar{\mathbb{P}}(R,\theta) \triangleq \operatorname*{argmin}_{\mathbb{P}\in\mathcal{B}_\theta\mathcal{R}} \sum_{a=1}^{K} D_{\mathrm{KL}}\left(\mathbb{P}^a(R,\theta) \| \mathbb{P}(R,\theta)\right)$$

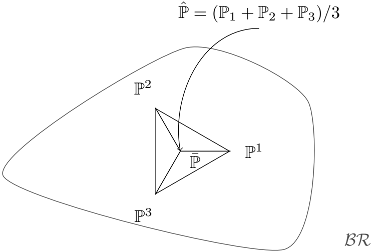


Pseudobelief-reward: Properties

Pseudobelief-reward

We propose to use the barycenter in the belief-reward manifold as the summary of collective information and to call it pseudobelief-reward distribution.

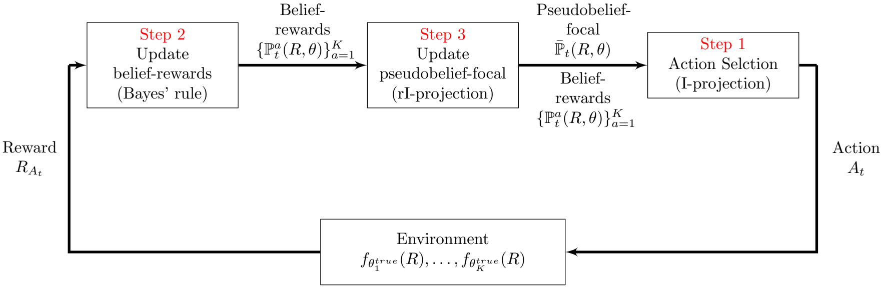
$$\bar{\mathbb{P}}(R,\theta) \triangleq \operatorname*{argmin}_{\mathbb{P}\in\mathcal{B}_\theta\mathcal{R}} \sum_{a=1}^{K} D_{\mathrm{KL}}\left(\mathbb{P}^a(R,\theta) \| \mathbb{P}(R,\theta)\right)$$



BelMan-Explore

Step 3: Update the pseudobelief-reward to the reverse I-projection of the belief-reward distributions of the arms.

$$\bar{\mathbb{P}}_t(R,\theta) = \operatorname*{argmin}_{\mathbb{P}\in\mathcal{B}_\theta\mathcal{R}} \sum_{a=1}^{K} D_{\mathrm{KL}}\left(\mathbb{P}_t^a(R,\theta)\|\bar{\mathbb{P}}(R,\theta)\right)$$

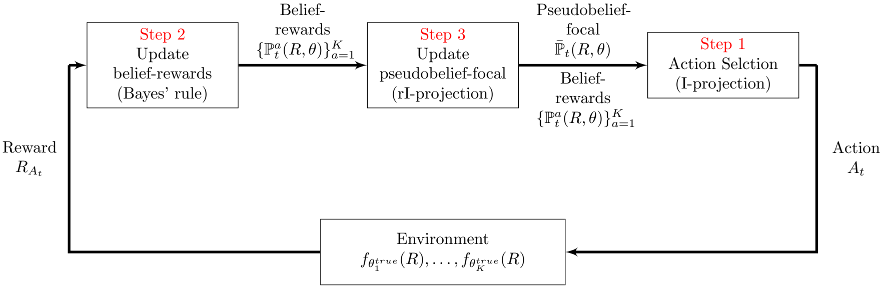


Step 2
Update belief-rewards (Bayes' rule)
Belief-rewards $\{\mathbb{P}_t^a(R,\theta)\}_{a=1}^K$
Step 3
Update pseudobelief-focal (rI-projection)
Pseudobelief-focal $\bar{\mathbb{P}}_t(R,\theta)$
Belief-rewards $\{\mathbb{P}_t^a(R,\theta)\}_{a=1}^K$
Step 1
Action Selction (I-projection)
Reward R_{A_t}
Action A_t
Environment
$f_{\theta_1^{true}}(R),\ldots,f_{\theta_K^{true}}(R)$

BelMan-Explore

Step 3: Update the pseudobelief-reward to the reverse I-projection of the belief-reward distributions of the arms.

$$\bar{\mathbb{P}}_t(R,\theta) = \operatorname*{argmin}_{\mathbb{P}\in\mathcal{B}_\theta\mathcal{R}} \sum_{a=1}^{K} D_{\mathrm{KL}}\left(\mathbb{P}_t^a(R,\theta)\|\bar{\mathbb{P}}(R,\theta)\right)$$



Step 1: Choose the arm a_t at time t whose belief-reward distribution is the information projection of the pseudobelief-reward distribution.

$$a_t = \operatorname*{argmin}_{a\in\{1,\ldots,K\}} D_{\mathrm{KL}}\left(\mathbb{P}_t^a(R,\theta)\|\bar{\mathbb{P}}_t(R,\theta)\right)$$

Information Exploitation → Pseudobelief-Focal Distribution

Information Exploitation

We need to exploit the accumulated information using the knowledge-base.

We need a mechanism that gradually concentrates on higher rewards.

Information Exploitation → Pseudobelief-Focal Distribution

Information Exploitation

We need to exploit the accumulated information using the knowledge-base.

We need a mechanism that gradually concentrates on higher rewards.

We consider the focal distribution of the form $L_t(R) \propto \exp\left(\frac{R}{\tau(t)}\right)$, where exposure $\tau(t)$ is a decreasing function of t.

Information Exploitation → Pseudobelief-Focal Distribution

Information Exploitation

We need to exploit the accumulated information using the knowledge-base.

We need a mechanism that gradually concentrates on higher rewards.

We consider the focal distribution of the form $L_t(R) \propto \exp\left(\frac{R}{\tau(t)}\right)$, where exposure $\tau(t)$ is a decreasing function of t.

Definition (Pseudobelief-focal distribution)

The pseudobelief-focal distribution is defined as the product of the pseudobelief-reward and the focal distribution.

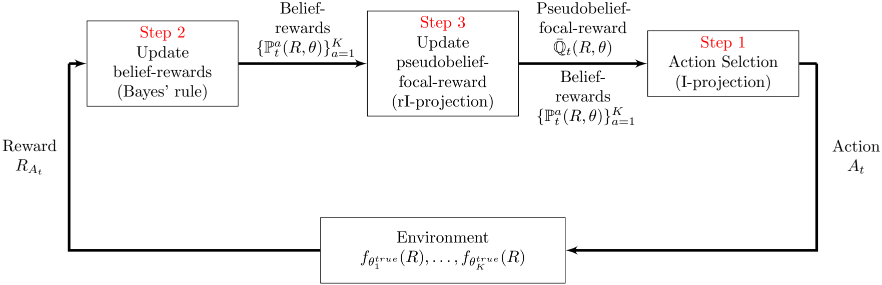
$$\bar{\mathbb{Q}}(R, \theta) \triangleq \frac{1}{\bar{Z}_t} \bar{\mathbb{P}}(R, \theta) \exp\left(\frac{R}{\tau(t)}\right)$$

Here, $\bar{Z}_t$ is the normalisation factor.

BelMan

Step 3: Update the pseudobelief-focal to the reverse I-projection of the belief-reward distributions of the arms.

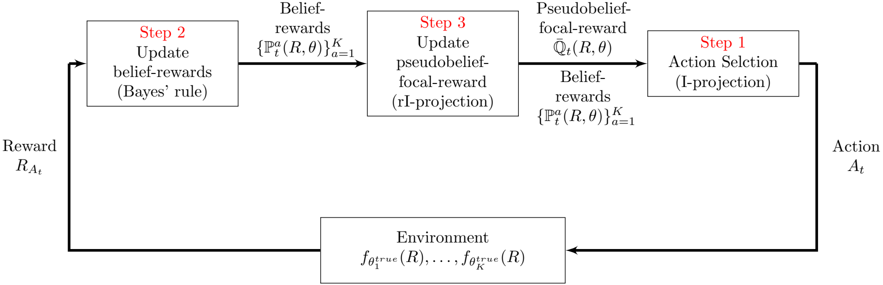
$$\bar{\mathbb{Q}}_t(R,\theta) = \operatorname*{argmin}_{\bar{\mathbb{Q}} \in \mathcal{B}_\theta \mathcal{R}} \sum_{a=1}^{K} D_{\mathrm{KL}}\left(\mathbb{P}_t^a(R,\theta) \| \bar{\mathbb{Q}}(R,\theta)\right)$$



BelMan

Step 3: Update the pseudobelief-focal to the reverse I-projection of the belief-reward distributions of the arms.

$$\bar{\mathbb{Q}}_t(R,\theta) = \operatorname*{argmin}_{\bar{\mathbb{Q}} \in \mathcal{B}_\theta \mathcal{R}} \sum_{a=1}^{K} D_{\mathrm{KL}}\left(\mathbb{P}_t^a(R,\theta) \| \bar{\mathbb{Q}}(R,\theta)\right)$$



Step 2
Update
belief-rewards
(Bayes' rule)

Belief-
rewards
$\{\mathbb{P}_t^a(R,\theta)\}_{a=1}^K$

Step 3
Update
pseudobelief-
focal-reward
(rI-projection)

Pseudobelief-
focal-reward
$\bar{\mathbb{Q}}_t(R,\theta)$

Belief-
rewards
$\{\mathbb{P}_t^a(R,\theta)\}_{a=1}^K$

Step 1
Action Selction
(I-projection)

Action
A_t

Environment
$f_{\theta_1^{true}}(R), \ldots, f_{\theta_K^{true}}(R)$

Reward
R_{A_t}

Step 1: Choose the arm a_t at time t whose belief-reward distribution is the information projection of the pseudobelief-focal-reward distribution.

$$a_t = \operatorname*{argmin}_{a \in \{1,\ldots,K\}} D_{\mathrm{KL}}\left(\mathbb{P}_t^a(R,\theta) \| \bar{\mathbb{Q}}_{t-1}(R,\theta)\right)$$

The Arm Selection and Exploration–exploitation Trade-off

$$a_t \triangleq \underset{a}{\operatorname{argmin}}\ D_{\mathrm{KL}}\left(\mathbb{P}^a_t(R,\boldsymbol{\theta}) \parallel \bar{\mathbb{Q}}_{t-1}(R,\boldsymbol{\theta})\right)$$

$$= \underset{a}{\operatorname{argmax}} \left[\underbrace{\mathbb{E}_{\mathbb{P}^a_t(R,\boldsymbol{\theta})}[R]}_{\text{First Term}} - \tau(t) \times \underbrace{D_{\mathrm{KL}}\left(\mathbb{P}^a_t(\boldsymbol{\theta}) \parallel \mathbb{P}_{\bar{\boldsymbol{\eta}}_t}(\boldsymbol{\theta})\right)}_{\text{Second Term}}\right]$$

- The first term represents the expected reward. Maximising the first term is analogous to greedily exploiting the present information about the arms.

- The second term quantifies the amount of uncertainty that can be decreased if the arm is chosen on the basis of the present pseudobelief. Minimising the second term is analogous to accumulating information to reduce uncertainty.

- Exposure $\tau(t)$ controls information vs. reward trade-off. Decrease in $\tau(t)$ increases the exploitation with time t.

Asymptotic Convergence

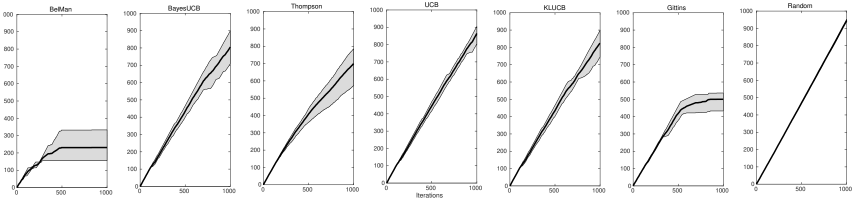
For a proper choice of exposure $\tau(t)$, BelMan asymptotically converges to choosing the optimal arm.

Theorem (Asymptotic Consistency)

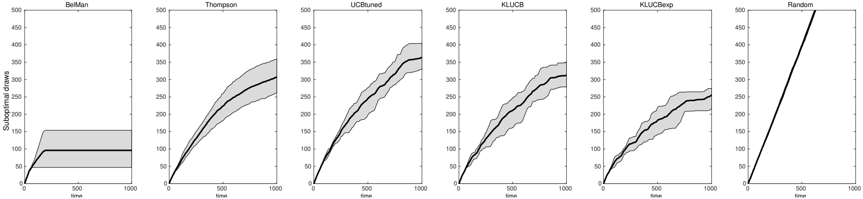
If all the arms have finite expected rewards $|\mu_a| < \pm\infty$ and finite variances $V(\mathbb{P}^a) < \infty$, there exists at least an optimal arm with expected reward $\mu^* \triangleq \max_a \mu(\theta_a)$, and the exposure grows to satisfy $\frac{1}{\tau(t)} = \Omega(\frac{1}{\sqrt{t}})$, then with high probability

$$\lim_{t\to\infty} \frac{S(\mathcal{A}, t)}{t} = \mu^*.$$

Experimental Results: Bernoulli and Exponential Bandits



Cumulative regret for 20-arm Bernoulli bandit.



Cumulative regret for 5-arm bounded exponential bandit.

Roadmap

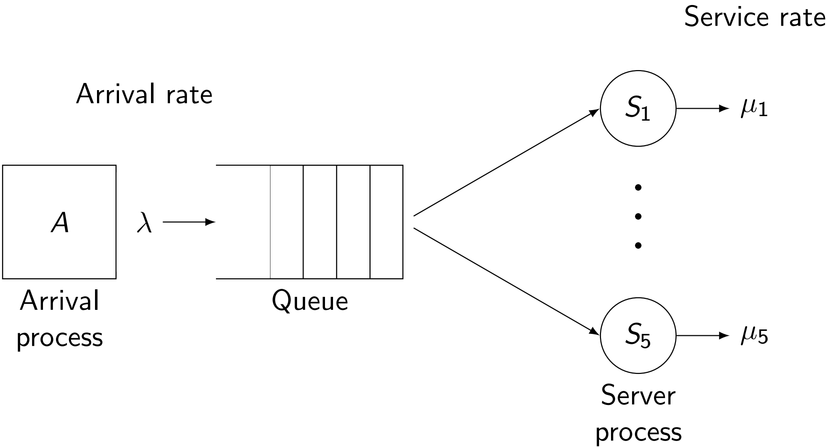
1. Bandits: Primer

2. BelMan
    - Information Representation: Belief-Reward Manifold
    - Information Accumulation: Pseudobelief-reward Distribution
    - Information Exploitation: Focal and Pseudobelief-focal Distributions
    - Algorithm: Alternating Information Projections
    - Experiments: Logarithmic Regret Growth

3. Queuing Bandits: An Application

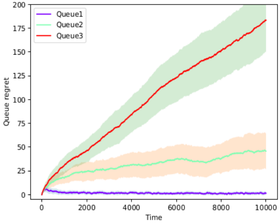
4. Wrap-up: Take-away

Queuing Bandits

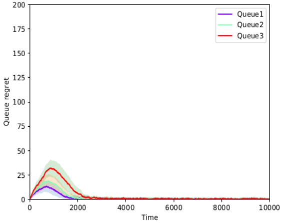


Multi-armed bandit framework models the online scheduling of jobs in a multi-server, multi-queue system with unknown arrival and service rates. We call this problem setup the Queuing Bandit.

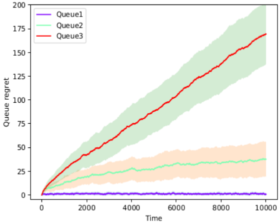
M/B/5: Single Queue Multiple Server



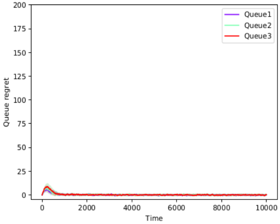
(a) Q-Thompson Sampling (Krishnasamy et al., 2016, NIPS)



(b) Q-UCB (Krishnasamy et al., 2016, NIPS)



(c) Thompson Sampling (Thompson, 1933, Biometrika)



(d) BelMan (Basu et al., 2018, arXiv)

URL : https ://github.com/Debabrota-Basu/QBelMan

Roadmap

1. Bandits: Primer

2. BelMan
   - Information Representation: Belief-Reward Manifold
   - Information Accumulation: Pseudobelief-reward Distribution
   - Information Exploitation: Focal and Pseudobelief-focal Distributions
   - Algorithm: Alternating Information Projections
   - Experiments: Logarithmic Regret Growth

3. Queuing Bandits: An Application

4. Wrap-up: Take-away

Closure

Theoretical Puzzle

How to effectively learn through exploration and efficiently optimize through exploitation in stochastic multi-armed bandits?

Proposed Solution

We propose an information geometric framework, BelMan, to balance the exploration and exploitation.

- **Information representation:** Belief-reward manifold that embeds the joint uncertainty of reward generation and incomplete information
- **Information accumulation:** Pesudeobelief-reward distribution that is the barycenter of belief-reward distributions of all the arms
- **Arm selection:** I-projection of the pseudobelief-focal-reward distribution that is equivalent to weighted maximisation of reward and information gain

Future Work

Conjecture

For stochastic bandits with bounded reward and $\frac{1}{\tau(t)} = \Theta(\frac{1}{\sqrt{t}})$, BelMan would achieve logarithmic regret bound.

$$Reg(\mathrm{BelMan}, T) = O(\log T).$$

Extension

We would extend this framework to resolve the exploration-exploitation trade-off in Markov decision processes which is a generalisation of multi-armed bandits.

Agrawal, S. and Goyal, N. (2012).
Analysis of thompson sampling for the multi-armed bandit problem.
In *Conference on Learning Theory*, pages 39–1.

Audibert, J.-Y. and Bubeck, S. (2009).
Minimax policies for adversarial and stochastic bandits.
In *COLT*, pages 217–226.

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

Basu, D., Senellart, P., and Bressan, S. (2018).
Belman: Bayesian bandits on the belief–reward manifold.
arXiv preprint arXiv:1805.01627.

Garivier, A. and Cappé, O. (2011).
The KL-UCB algorithm for bounded stochastic bandits and beyond.
In *COLT*, pages 359–376.

Kaufmann, É., Cappé, O., and Garivier, A. (2012).
On Bayesian upper confidence bounds for bandit problems.
In *AISTATS*, pages 592–600.

Lai, T. L. and Robbins, H. (1985).
Asymptotically efficient adaptive allocation rules.
Advances in applied mathematics, 6(1):4–22.

Osband, I., Russo, D., and Van Roy, B. (2013).
(More) efficient reinforcement learning via posterior sampling.
In *NIPS*, pages 3003–3011.

Thompson, W. R. (1933).
On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
Biometrika, 25(3–4):285.

Frequentist Bandit Algorithms

Act as if the empirically best choice is truly the best choice.

Algorithm 1 Optimism in Face of Uncertainty Framework

```
Initialisation: $N$ arms, horizon $T \gg N$.
for $t = 1, \ldots, N$ do
    Play arm $t$, $a_t = t$.
end for
for $t = N + 1, \ldots, T$ do
    Play arm
        $$a_t = \operatorname*{argmax}_{a \in \{1,\ldots,N\}} f(n_{a,t}, \hat{\mu}_a(t))$$
end for
```

Frequentist Bandit Algorithms

Act as if the empirically best choice is truly the best choice.

Algorithm 2 Optimism in Face of Uncertainty Framework

```
**Initialisation:** $N$ arms, horizon $T \gg N$.
**for** $t = 1, \dots, N$ **do**
    Play arm $t$, $a_t = t$.
**end for**
**for** $t = N+1, \dots, T$ **do**
    Play arm
        $a_t = \underset{a \in \{1,\dots,N\}}{\mathrm{argmax}}\ f(n_{a,t}, \hat{\mu}_a(t))$
**end for**
```

Small $n_{a,t}$ $\implies$ large $f(n_{a,t}, \hat{\mu}_a(t))$ $\implies$ more uncertainty $\implies$ needs EXPLORATION

Large $n_{a,t}$ $\implies$ small $f(n_{a,t}, \hat{\mu}_a(t))$ $\implies$ more certainty $\implies$ EXPLOIT depending on $\hat{\mu}_a(t)$

Bayesian Bandit Algorithms

- Bayesian bandit algorithms [Agrawal and Goyal, 2012] exploit prior knowledge about rewards.
- Bayesian bandit algorithms compute posterior distribution of rewards from the history and prior.
- Leverage posterior distribution to guide exploration.
  - Probability matching → Thompson sampling [Thompson, 1933]
  - Upper confidence bound → Bayes-UCB [Kaufmann et al., 2012]
  - Reward gain vs information gain ratio → Information-directed sampling [Osband et al., 2013]
  - **Divergence based information and reward gain with respect to accumulated knowledge-base** → *BelMan* [Basu et al., 2018]

Bandit Algorithm	Decision Function	Finite-time regret bound
UCB	$\widehat{\mu_a}(t) + \sqrt{\frac{2\log t}{n_a(t)}}$	$8\left[\sum_{a:\mu_a<\mu^*} \frac{\log t}{\Delta_a}\right] + (1+\frac{\pi^2}{3})(\sum_{a=1}^{K} \Delta_a)$
UCB-tuned	$\widehat{\mu_a}(t) + \sqrt{\frac{\log t}{n_a(t)} \min\{0.25, \hat{\sigma_a}(t)\}}$	–
MOSS	$\widehat{\mu_a}(t) + \frac{\log T}{Kn_a(t)}$	$25\sqrt{TK}$
KL-UCB	$D_{\mathrm{KL}}\left(\widehat{\mu_a}(t) \Vert M\right) \leq \frac{\log t + c\log\log t}{n_a(t)}$	$\left(\sum_{a=1}^{K} \frac{\Delta_a}{D_{\mathrm{KL}}(f_a \Vert f^*)}\right) \log t$ $+C\left(\sum_{a=1}^{K} \Delta_a\right) \log\log t$
Bayes-UCB	$Q(1-\frac{1}{t}, \mathrm{Posterior}_a^{t-1}(X))$	$\left(\sum_{a=1}^{K} \frac{\Delta_a}{D_{\mathrm{KL}}(f_a \Vert f^*)}\right)(\log t + C\log\log t)$

The Bouquet of Bandit Algorithms

Bandit Algorithms	Frequentist	Bayesian
Deterministic	UCB [Auer et al., 2002], KL-UCB [Garivier and Cappé, 2011] MOSS [Audibert and Bubeck, 2009]	Bayes-UCB [Kaufmann et al., 2012], **BelMan** [Basu et al., 2018]
Randomised	Adaptive ϵ greedy [Auer et al., 2002]	Thompson sampling [Thompson, 1933] Information-directed sampling [Osband et al., 2013]

How Good/Bad a Bandit Algorithm Can Be?[Lai and Robbins, 1985]

- **Consistency:** A bandit algorithm is asymptotically consistent if it detects and keeps on playing the optimal arm almost surely.

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\boldsymbol{\mu}}[\sum_{t=1}^{T} R_t] = \mu^*.$$

- **Upper bound:** A bandit algorithm is asymptotically efficient (optimal) if its regret grows logarithmically with time.

For any suboptimal arm a, $$\limsup_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[n_a(T)]}{\log T} \leq \frac{1}{D(\mu_a\|\mu^*)}.$$

- **Lower bound:** A bandit algorithm producing uniformly efficient strategy is fundamentally limited by logarithmic regret growth.

For any suboptimal arm a, $$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[n_a(T)]}{\log T} \geq \frac{1}{D(\mu_a\|\mu^*)}.$$

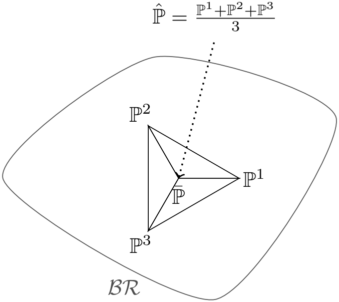
Properties of Pseudobelief-reward Distribution I

We prove that the pseudobelief-reward exists on the belief-reward manifold, is uniquely defined, and is a projection of average belief-reward of all arms.

Theorem (Existence and Uniqueness)

The pseudobelief-reward $\bar{\mathbb{P}}_t$ is unique, and its expectation parameter is average of the expectation parameters of the belief-reward distributions of the arms.

$$\bar{\mu}_t(\theta) = \frac{1}{K}\sum_{a=1}^{K} \mu_t^a(\theta)$$



Properties of Pseudobelief-reward Distribution II

Corollary (Summary of Belief-reward Distributions)

The pseudobelief-reward distribution $\bar{\mathbb{P}}_t$ is the point on the belief-reward manifold that has minimum KL-divergence from the average belief-reward distribution

$$\hat{\mathbb{P}}_t(R,\theta) \triangleq \frac{1}{N}\sum_{a=1}^{K}\mathbb{P}_t^a(R,\theta).$$

Theorem (Consistent Estimation of Belief-reward Distributions)

If $\tilde{\bar{\mu}}_T \triangleq \frac{1}{K}\sum_{a=1}^{K}\tilde{\mu}^a_{n_a(T)}$ 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 $\frac{T}{K^2 n_a(T)}$ tends to λ^j as $T \to \infty$.

Queueing Bandits

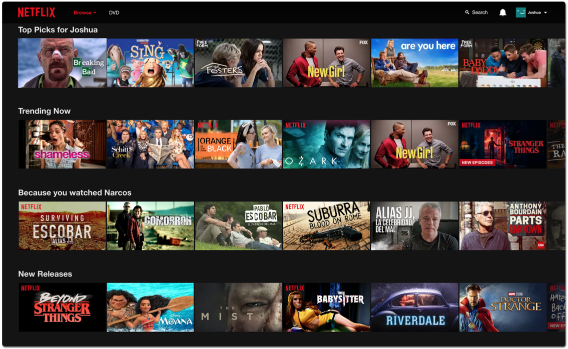
- **Queueing Bandits:** Multi-armed bandit framework models the online scheduling of jobs in a multi-server, multi-queue system with unknown arrival and service rates. We call this problem setup the Queuing Bandit.

- **Experimental Setup:** We experimentally verify BelMan for queueing system with Markovian arrival process and Bernoulli service distribution. We run 100 experiments for a time horizon of $10,000$ in case of one and three queues respectively.

- **Competing Algorithms:** We compare BelMan with the algorithms, Q-UCB and Q-Thompson Sampling, designed for queuing bandits (Krishnasamy et al., 2016, NIPS), and also the Thompson sampling (Thompson, 1933, Biometrika).

Real-life Applications

How to recommend which cinema you would prefer to watch?



For meta-optimisation in deep learning and so on...