Reactive bandits with attitude

Pedro A. Ortega
University of Pennsylvania
Philadelphia, PA 19104, U.S.A.

Kee-Eung Kim
Korea Advanced Institute
of Science and Technology
Daejeon 305-701, Korea

Daniel D. Lee
University of Pennsylvania
Philadelphia, PA 19104, U.S.A.

Abstract

We consider a general class of K-armed bandits that adapt to the actions of the player. A single continuous parameter characterizes the "attitude" of the bandit, ranging from stochastic to cooperative or to fully adversarial in nature. The player seeks to maximize the expected return from the adaptive bandit, and the associated optimization problem is related to the free energy of a statistical mechanical system under an external field. When the underlying stochastic distribution is Gaussian, we derive an analytic solution for the long run optimal player strategy for different regimes of the bandit. In the fully adversarial limit, this solution is equivalent to the Nash equilibrium of a two-player, zero-sum semi-infinite game. We show how optimal strategies can be learned from sequential draws and reward observations in these adaptive bandits using Bayesian filtering and Thompson sampling. Results show the qualitative difference in policy regret between our proposed strategy and other well-known bandit algorithms.

1 Introduction

As a standard model for sequential decision making, the multi-armed bandit has attracted much interest from the machine learning community in recent years. In both the stochastic and adversarial settings, there has been much progress in understanding the limits of achievable performance along with concrete algorithms that approach those limits [Bubeck and Cesa-Bianchi, 2012]. In this work, we introduce a class of multi-armed bandits that react to the actions of the player. We demonstrate how standard bandit algorithms can fail to maximize reward against this bandit and introduce a novel Bayesian bandit algorithm that explicitly models the reaction of the bandit.

The K-armed bandit problem can be described as a sequential game between a player and the environment. At each round t, the player chooses an arm I_t from the action set $\{1..K\}$, and the bandit chooses a vector of rewards $\vec{r}^t \in \Re^K$ from the distribution $\vec{r}^t \sim Q_t(\vec{r})$. In the partial information setting, the player only observes and receives the single reward $r^t_{I_t}$ and uses that information to update her strategy for subsequent rounds. The player's goal is to accumulate the largest amount of rewards over the rounds of play.

In the stochastic bandit model, the distribution of rewards that the bandit samples from is stationary and does not change with time: $Q_t(\vec{r}) = Q_0(\vec{r})$. In this simple model, the player needs to determine the arm with the greatest expected reward, and then repeatedly pull that arm to maximize her total reward. This illustrates the dilemma between *exploring* arms that have not been sampled enough, and *exploiting* arms that appear to give better rewards. The notion of regret, or more properly pseudo-regret [Bubeck and Cesa-Bianchi, 2012], is central to quantifying how efficiently the player performs both exploration and exploitation by measuring the difference in rewards from the optimal policy of always pulling the best arm:

$$R_T = \max_{i=1..K} \left\langle \sum_{t=1}^{T} r_i^t \right\rangle - \left\langle \sum_{t=1}^{T} r_{I_t}^t \right\rangle \qquad (1)$$

More recently, there have been a number of extensions to the K-armed stochastic bandit model. These include contextual bandits [Langford and Zhang, 2008, Li et al., 2010], (generalized) linear bandits [Dani et al., 2008, Filippi et al., 2010], $\mathcal{X}$-armed bandits [Bubeck et al., 2008], online convex programming [Zinkevich, 2003], and tree bandits [Kocsis and Szepesvári, 2006, Bubeck and Munos, 2010].

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

The adversarial bandit model drops the requirement that the reward distribution is stationary: $Q_t(\vec{r}) \neq Q_{t'}(\vec{r})$. In the oblivious setting, the distribution Q_t can change in time but cannot depend upon the past history of the player's actions $\{I_1, I_2, ..., I_{t-1}\}$. In this setting, it has been shown that it is possible to achieve sublinear regret $R_T = O(\sqrt{T})$ [Auer et al., 2002].

In our work we consider a reactive bandit, where the distribution may depend upon the history of the player: $Q_t(\vec{r}|\{I_1, I_2, ...I_{t-1}\})$. In general, it has previously been shown that this type of reactive bandit can always have linear regret [Pucci de Farias and Megiddo, 2006, Arora et al., 2012].

In this work, we introduce a class of reactive bandits, which modulates their reward distributions based upon the past actions of the player. The bandit changes its reward distribution depending upon the past player's actions by shifting the rewards to be adversarial, stationary stochastic, or cooperative. The bandit is governed by a continuous parameter β which shifts its "attitude" from fully adversarial in the zero-sum game sense to fully cooperative in the pure coordination game sense. We then analyze the optimal mixed strategy against such a bandit in the long run limit.

We describe our work in the following sections. In Section 2, we give a mathematical definition of the bandit which can be analyzed from a statistical mechanics perspective. In Section 3, we introduce a Gaussian version of the reactive bandit that can be solved analytically. Then in Section 4 we show how current bandit learning algorithms for the player fail to converge to the optimal mixed policy for a range of finite values of β. We then derive a Bayesian bandit algorithm that explicitly models the response of the bandit along with the means and variances of the rewards, and demonstrate how that algorithm can converge to the optimal player policies for different values of β in Section 5. We extend our analysis to correlated bandits and finish with concluding remarks and suggestions for future work in this area.

2 Bandits with attitude

We first review the *stochastic* multi-armed bandit, and describe the optimal policy for the player in terms of an optimization problem over mixed strategies. The rewards at each round $\vec{r}^t \in \Re^K$ are sampled from a stationary distribution $Q_0(\vec{r})$. Typically, this distribution is considered to be independent, i.e. $Q_0(\vec{r}) = \prod_i Q_{0,i}(r_i)$ but, in general, we can consider correlated reward distributions as well. The player chooses an arm I_t at round t, receiving the reward r_{I_t}. A general mixed strategy for the player is described by probabilities p_i such that $p_i \geq 0$ and $\sum_i p_i = 1$, and where p_{I_t} is the probability for selecting arm I_t. The optimal policy is then determined by maximizing the expected reward:

$$\max_{p_i} \sum_i p_i \langle r_i \rangle_{Q_0(\vec{r})} \tag{2}$$

where the expected reward is given by integrating over the distribution Q_0: $\langle r_i \rangle_{Q_0(\vec{r})} = \int d\vec{r}\, Q_0(\vec{r})\, r_i$.

For the stochastic bandit, the optimal policy is deterministic:

$$p_i^\star = \arg\max_{\vec{p}} \vec{p} \cdot \langle \vec{r} \rangle_{Q_0(\vec{r})} = \delta_{i,i^\star} \tag{3}$$

and the optimal arm $i^\star$ is given by:

$$i^\star \in \arg\max_i \langle r_i \rangle_{Q_0(\vec{r})} \tag{4}$$

2.1 Adaptive distribution

We now introduce our model for an adaptive bandit. At round t, the bandit estimates the policy of the agent $\hat{p}^t$ from the past history of player actions $\{I_1, I_2, ..., I_{t-1}\}$. The bandit responds by drawing the rewards $\vec{r}^t$ from the time-dependent distribution:

$$Q_{\hat{p}^t}(\vec{r}) = \frac{1}{Z_{\hat{p}^t}} Q_0(\vec{r}) e^{\beta \hat{p}^t \cdot \vec{r}} \tag{5}$$

where the normalization constant for the distribution is given by the partition function:

$$Z_{\hat{p}^t} = \int d\vec{r}\, Q_0(\vec{r}) e^{\beta \hat{p}^t \cdot \vec{r}} \tag{6}$$

The parameter $\beta \in (-\infty, +\infty)$ modulates the response of the bandit to the agent. In particular, when $\beta = 0$, we recover as a special case the definition of the stochastic bandit with stationary reward distribution $Q_0(\vec{r})$.

The objective of the agent playing the bandit is to maximize the expected reward under this adaptive reward distribution. Consider when the player plays a mixed strategy $\vec{p}$ and the bandit has fully adapted to this strategy $\hat{p}^t = \vec{p}$. The expected reward is then given by: $\langle r^t \rangle = \vec{p} \cdot \langle \vec{r} \rangle_{Q_{\vec{p}}(\vec{r})}$. Under these conditions, the optimal policy for the player is given by maximizing:

$$\vec{p}^\star = \arg\max_{\vec{p}} \vec{p} \cdot \langle \vec{r} \rangle_{Q_{\vec{p}}(\vec{r})} \tag{7}$$

2.2 Free energy interpretation

We can relate this bandit model to a statistical mechanical interpretation. The expected reward under

the adaptive reward distribution can be expressed as a derivative of the free energy $F_{\vec{p}} = \log Z_{\vec{p}}$:

$$\begin{aligned} \vec{p} \cdot \langle \vec{r} \rangle_{Q_{\vec{p}}(\vec{r})} &= \int d\vec{r}\, Q_{\vec{p}}(\vec{r})\, \vec{p} \cdot \vec{r} && (8) \\ &= \int d\vec{r}\, \frac{1}{Z_{\vec{p}}} Q_0(\vec{r}) e^{\beta \vec{p} \cdot \vec{r}}\, (\vec{p} \cdot \vec{r}) && (9) \\ &= \frac{1}{Z_{\vec{p}}} \frac{\partial Z_{\vec{p}}}{\partial \beta} && (10) \\ &= \frac{\partial}{\partial \beta} F_{\vec{p}} && (11) \end{aligned}$$

The reactive reward distribution can also be understood in terms of the variational principle. From the perspective of the bandit, the free energy is determined by maximizing over all reward distributions:

$$F_{\vec{p}} = \max_{Q(\vec{r})} \left[\beta \vec{p} \cdot \langle \vec{r} \rangle_{Q(\vec{r})} - D_{KL}(Q(\vec{r}) || Q_0(\vec{r}) \right] \quad (12)$$

where the Kullback-Leibler divergence is given by $D_{KL}(Q(\vec{r})||Q_0(\vec{r}) = \int d\vec{r} Q(\vec{r}) \log \frac{Q(\vec{r})}{Q_0(\vec{r})}$. In this interpretation, the free energy is determined by the bandit shifting the reward distribution to maximize the response to an external field given by the vector $\beta\vec{p}$. However, for finite β, the bandit is constrained to keep the reward distribution close to the stationary stochastic distribution $Q_0(\vec{r})$ as measured by the Kullback-Leibler divergence [Ortega and Braun, 2011]. The player seeks to extract the maximal reward from this shifted reward distribution.

2.3 Discrete distributions

We illustrate how the parameter β affects the reward distribution by first considering a discrete distribution $Q_0(\vec{r})$. In this case, the rewards consist of a discrete set of vectors:

$$Q_0(\vec{r}) = \sum_{l=1}^{L} q_l \delta(\vec{r} - \vec{r}_l) \quad (13)$$

That is, the reward vector can take on one of L potential vectors $\{\vec{r}_l\}$ with weights $q_l > 0$ that sum to unity. When $\beta = 0$, the expected reward is simply given by

$$\langle \vec{p} \cdot \vec{r} \rangle_{Q_0} = \sum_{l=1}^{L} q_l \vec{p} \cdot \vec{r}_l \quad (14)$$

and for arbitrary β,

$$\langle \vec{p} \cdot \vec{r} \rangle_{Q_{\vec{p}}} = \frac{\sum_{l=1}^{L} q_l e^{\beta \vec{p} \cdot \vec{r}_l} \vec{p} \cdot \vec{r}_l}{\sum_{l'=1}^{L} q_{l'} e^{\beta \vec{p} \cdot \vec{r}_{l'}}}. \quad (15)$$

In the limit that $\beta \to +\infty$, the expected reward is $\langle \vec{p} \cdot \vec{r} \rangle_{Q_{\vec{p}}} = \max_l \vec{p} \cdot \vec{r}_l$. In other words, the bandit chooses the reward vector that maximizes the expected reward of the player. In this limit, the optimal strategy for the player is given by:

$$\vec{p}^{\star} = \arg \max_{\vec{p}} \max_{l} \vec{p} \cdot \vec{r}_l. \quad (16)$$

In this case, the optimal policy for the player will be deterministic. This can be interpreted as a pure coordination game between the player and bandit with a shared normal form payoff matrix given by $r_{l,i}$, where the player has limited information about the payoff components and bandit's choices.

Similarly for $\beta \to -\infty$, the expected reward is $\langle \vec{p} \cdot \vec{r} \rangle_{Q_{\vec{p}}} = \min_l \vec{p} \cdot \vec{r}_l$. In this limit, the optimal strategy for the player is given by:

$$\vec{p}^{\star} = \arg \max_{\vec{p}} \min_{l} \vec{p} \cdot \vec{r}_l. \quad (17)$$

This can be interpreted as a zero-sum game between the player and bandit, where the payoffs for the player are given by the matrix $r_{l,i}$. From von Neumann's minimax theorem, we know that in general the optimal policy for the player will be a mixed strategy [Von Neumann and Morgenstern, 1945].

For intermediate values of β, we see the expected reward is a soft version of maximum or minimum depending upon the sign of β. When $\beta > 0$, the bandit adapts to partially cooperate with the player using softmax to jointly increase the expected reward. On the other hand, for $\beta < 0$ the bandit adapts to decrease the expected reward, antagonistically acting against the agent using the softmin function.

2.4 Semi-infinite S-game

Even with an infinite continuous distribution over rewards, we can relate our bandit model in the limit when $\beta \to -\infty$ to a two-player, zero-sum game. Let the support of Q_0 be $S_0 = \{\vec{r} : Q_0(\vec{r}) > 0\}$ and assume that it is bounded below: $\exists B\, \forall \vec{r} \in S_0,\ r_i > B$. Then finding the optimal policy in this scenario is equivalent to finding the Nash equilibrium between the agent and bandit in a semi-infinite S-game [Blackwell and Girshick, 1954]:

$$\vec{p}^{\star} = \arg \max_{\vec{p}} \min_{\vec{r} \in S_0} \vec{p} \cdot \vec{r} \quad (18)$$

The optimal policy for the player $\vec{p}^{\star}$ can be determined by an interesting geometric construction. It involves sliding an orthant in $\Re^K$ until it touches the convex hull of set S_0. The normal to the tangent at the intersection then gives the components of the optimal mixed strategy for the player.

2.5 Fictitious play

Here we provide a procedural description of how the reactive bandit chooses its rewards. The bandit is given the parameter β and the stationary distribution $Q_0(\vec{r})$. At each round t, the bandit tracks the past plays $\{I_1, I_2, ..., I_{t-1}\}$ of the player by computing the number of times each arm has been pulled: $\hat{N}_i^t = \sum_{t'<t} \mathbb{I}\{I_{t'} = i\}$. It computes the estimated policy vector $\hat{p}^t$ via fictitious play: $\hat{p}_i^t = \hat{N}_i^t/t$. The bandit then samples the reward vector from the distribution:

$$\vec{r}^t \sim \frac{1}{Z_{\hat{p}^t}} Q_0(\vec{r}) e^{\beta \hat{p}^t \cdot \vec{r}}. \tag{19}$$

Sampling from such a distribution can be performed using rejection or importance sampling. When the agent pulls arm I_t, the bandit provides the reward $r_{I_t}^t$ and repeats the procedure for the next round.

Rather than fictitious play, the bandit could use a weighted filter on the past plays $\{I_1, I_2, ..., I_{t-1}\}$ to estimate $\vec{p}^t$. Another option is for the bandit to track the player policy using a Dirichlet distribution, and to sample from this distribution to generate $\hat{p}^t$. In our experiments, we use fictitious play for the bandit, but other estimators will give asymptotically similar results as long as the beliefs converge in the time-average sense [Robinson, 1951]. Given any sequence of arm pulls $\{I_1, I_2, ...\}$ whose frequency counts converge: $\hat{N}^t/t \to \vec{p}_0$, then the estimator for the bandit $\hat{p}^t = f(\{I_1, I_2, ..., I_{t-1}\})$ should also converge $\hat{p}^t \to \vec{p}_0$.

3 Reactive independent Gaussian bandit

We now consider the independent Gaussian bandit with distribution:

$$Q_0(\vec{r}) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(r_i - \mu_i)^2}{2\sigma_i^2}\right] \tag{20}$$

so that for $\beta = 0$, rewards on arm i are drawn independently from a Gaussian with mean μ_i and variance σ_i^2. In this case, $Q_{\vec{p}}(\vec{r})$ in (5) is also Gaussian:

$$Q_{\vec{p}}(\vec{r}) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(r_i - (\mu_i + \beta\sigma_i^2 p_i))^2}{2\sigma_i^2}\right] \tag{21}$$

with shifted means $\mu_i' = \mu_i + \beta\sigma_i^2 p_i$ and variances σ_i^2.

3.1 Free energy computation

The free energy function can then be computed analytically:

$$F_{\vec{p}} = \sum_i \beta p_i \mu_i + \frac{\beta^2}{2} p_i^2 \sigma_i^2 \tag{22}$$

which is a quadratic function on the probability simplex.

The optimal mixed player strategy $\vec{p}^\star$ for varying β is determined by optimizing:

$$\vec{p}^\star = \arg\max_{\vec{p}} \frac{\partial F_{\vec{p}}}{\partial \beta} = \arg\max_{\vec{p}} \sum_i p_i \mu_i + \beta p_i^2 \sigma_i^2 \tag{23}$$

For the stochastic bandit with $\beta = 0$, the optimal policy is deterministic: $p_i^\star = \mathbb{I}\{i = i^\star\}$ where the optimal arm $i^\star$ is determined by maximizing the mean:

$$i^\star(\beta = 0) = \arg\max_i \mu_i. \tag{24}$$

3.2 Optimal policy ($\beta > 0$)

When $\beta > 0$, the optimal policy is also deterministic:

Proposition 1. *The optimal player strategy $\vec{p}^\star$ for $\beta > 0$ is deterministic, where the optimal arm is given by:*

$$i^\star = \arg\max_i \left(\mu_i + \beta\sigma_i^2\right) \tag{25}$$

Proof. Consider any strategy $\vec{p}$. Then

$$\sum_i p_i\mu_i + \beta p_i^2\sigma_i^2 \le \sum_i p_i\mu_i + \beta p_i\sigma_i^2 \le \mu_{i^\star} + \beta\sigma_{i^\star}^2 \tag{26}$$

where the final inequality is the expected reward of the optimal deterministic strategy. The inequality is strict for any stochastic policy where there is a p_i such that $0 < p_i < 1$. □

Note that the choice of the optimal arm depends upon the value of β. In particular, when $\beta \to +\infty$, the optimal strategy is to choose the arm with the highest variance rather than the arm with the highest mean. This is due to the cooperative nature of the bandit which favors arms that can be exploited in favor of the agent.

For intermediate values of $\beta > 0$, there will be discrete transitions where the optimal policy will switch from arms with higher means to arms with greater variance. This can be seen by plotting the linear functions $\mu_i + \beta\sigma_i^2$ as a function of β.

3.3 Optimal policy ($\beta < 0$)

On the other hand for $\beta < 0$, the optimization is more complex. We need to analyze the Lagrangian:

$$\begin{aligned} \max_{\vec{p}} \min_{\vec{\alpha}\ge 0, \lambda} L_\beta(\vec{p}) &= \sum_i \left[p_i\mu_i + \beta p_i^2\sigma_i^2\right] \\ &+ \sum_i \alpha_i p_i - \lambda\left(\sum_i p_i - 1\right) \end{aligned} \tag{27}$$

where the Lagrange multipliers $\alpha_i \geq 0$ and λ enforce the non-negativity and sum constraint on $\vec{p}$. Taking derivatives, we obtain the Karush-Kuhn-Tucker (KKT) conditions:

$$\mu_i + 2\beta\sigma_i^2 p_i = \lambda - \alpha_i \quad (28)$$

$$\alpha_i p_i = 0 \quad (29)$$

$$\sum_i p_i = 1 \quad (30)$$

Combining these expressions, we see that the optimal solution has the form:

$$p_i^\star(\beta < 0) = \max\left\{\frac{\lambda - \mu_i}{2\beta\sigma_i^2},\ 0\right\} \quad (31)$$

The Lagrange multiplier λ is chosen such that $\sum_i p_i^\star = 1$. This results in solving a piecewise-linear equation for λ, similar to the water-filling construction used for optimizing power allocations in communication channels [Proakis et al., 1994]. Near $\beta \simeq 0$, the optimal policy will be deterministic, choosing the arm with the largest mean exactly as in the stochastic bandit. However, when β is decreased, arms with lower mean rewards will begin to be mixed into the solution. Specifically, when $\beta < -\frac{1}{2}(\mu_{i^\star} - \mu_{i_2})/\sigma_{i^\star}^2$ where $\mu_{i^\star}$ is the largest mean and μ_{i_2} is the second largest mean, the optimal policy will no longer be deterministic.

As β is decreased, the optimal policy employs more and more arms into the mixed distribution. Finally, in the limit $\beta \to -\infty$, the optimal policy is given by:

$$p_i^\star(\beta \to -\infty) = \frac{\frac{1}{\sigma_i^2}}{\sum_j \frac{1}{\sigma_j^2}} \quad (32)$$

This solution shows that in the fully adversarial case, the optimal policy is stochastic but favors arms that have smaller variance regardless of the mean.

To summarize, depending on the parameter β we obtain a diverse set of optimal policies. For $\beta = 0$, we recover the stochastic bandit solution, where the policy selects the arm with the highest mean reward. However, the solution changes dramatically when $\beta > 0$ or $\beta < 0$. In the cooperative case, the optimal policy remains deterministic but shifts to arms with higher variance. In the adversarial case, the optimal policy becomes more stochastic by employing arms with lower expected rewards. Ultimately, the fully adversarial solution mixes all arms but favors those with smaller variance.

This is illustrated for a specific $K = 3$ Gaussian bandit in Figure 1. In the range $-0.05 < \beta < 1.0$, the optimal policy is to deterministically choose the first arm. Above $\beta > 1.0$, the deterministic optimal policy transitions from arm to arm until the third arm becomes optimal for $\beta > 2.5$. When $\beta < -0.05$, the optimal policy is a mixed strategy. It only mixes the first two arms for $-0.145 < \beta < -0.05$ and mixes all three arms when $\beta < -0.145$.

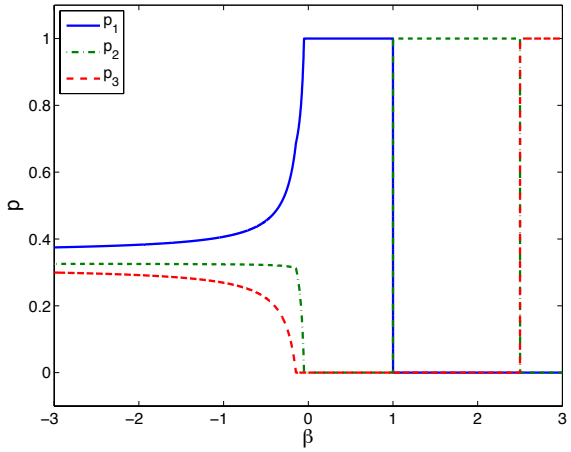

Figure 1: Optimal policies for K=3 reactive Gaussian bandit with $\mu = [0, -0.1, -0.2]$ and $\sigma^2 = [1.0, 1.1, 1.14]$ with deterministic and mixed strategy transitions as a function of β.

4 Bandit algorithms

We investigate how some current bandit algorithms perform in playing the reactive Gaussian bandit. The first algorithms we investigated are Upper-Confidence Bound (UCB) based algorithms which use the principle of optimism in the face of uncertainty. UCB algorithms compute an upper bound on the expected reward of an arm by summing the empirical mean reward and the uncertainty in the empirical mean, and then select the arm with the highest upper bound. The uncertainty factor increases for arms with less pulls, thereby encouraging exploration of arms that have not been sufficiently sampled.

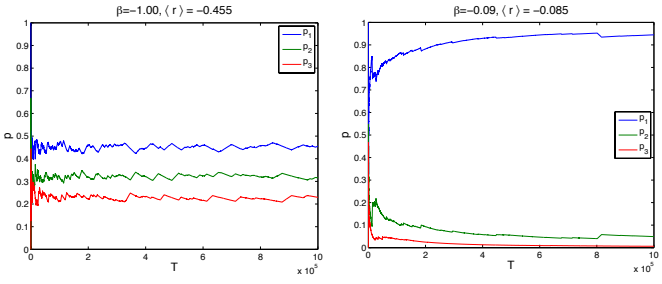

Figure 2: Convergence of UCB1 for $\beta = -1.0$ and $\beta = -0.09$.

Figure 2 shows the results of applying the UCB1 algorithm to the $K = 3$ Gaussian bandit with the same pa-

rameters as in Figure 1. The upper confidence bounds were calculated as a function of time t as:

$$B_i = \hat{\mu}_i + \eta \sqrt{\frac{\log t}{\hat{N}_i}} \tag{33}$$

where $\hat{\mu}_i$ is the empirical mean of the rewards on arm i and η is a scaling parameter on the uncertainty. For $\beta = -1.0$, the UCB1 algorithm mimics a mixed strategy with frequencies $\vec{p} = [0.45, 0.32, 0.22]$. This is close to but not quite equal to the optimal policy $\vec{p}^\star = [0.41, 0.32, 0.27]$. Compared to the optimal mixed policy, UCB1 suffers a small linear regret asymptotically.

The performance of UCB1 is even more clear for $\beta = -0.09$. The algorithm converges to a deterministic policy of only choosing the first arm whereas the optimal mixed policy is to choose the first two arms with frequencies $\vec{p}^\star = [0.79, 0.21, 0.0]$. This asymptotic behavior was observed even when tuning various parameters of the algorithm such as η in Eq. 33. For $\beta > 0$, UCB1 typically will converge to the optimal deterministic policy. However, there are situations where it gets trapped on a non-optimal arm. This is due to the non-convex nature of the optimization for $\beta > 0$ in Eq. 23.

We have also tried other bandit learning algorithms such as UCB algorithms which explicitly model the variances of the arms as well as EXP3 which carries theoretical guarantees on regret in the non-oblivious setting. However, we observed the same qualitative asymptotic behavior with those algorithms on this reactive Gaussian bandit. In particular, we see convergence to a near-deterministic strategy for $\beta = -0.09$, implying asymptotic linear regret compared to the optimal mixed strategy.

All these algorithms converge to a strategy $\vec{p}$ for which the expected rewards on arms with $p_i > 0$ are equalized so that the player is indifferent to the arms that she is playing. In the case of the Gaussian bandit, this implies that for arms with $p_i > 0$,

$$\mu_i + \beta p_i \sigma_i^2 = \lambda' \tag{34}$$

and $\mu_j < \lambda'$ for arms with $p_j = 0$. We note that these asymptotic conditions may be satisfied for $\beta > 0$ optimal deterministic policies. However, they do not match the necessary optimal conditions for mixed policies in Eqs. 28–30.

More specifically, for the Gaussian bandit with $\beta < 0$, we see that bandit algorithms that match expected rewards across the arms converge to the optimal mixed policy with attitude parameter $\beta' = \frac{1}{2}\beta$. This implies these algorithms will exhibit the correct asymptotic behavior in the fully adversarial situation where $\beta \to -\infty$. Convergence to optimality in the zero-sum game was previously known when the algorithms are Hannan consistent [Cesa-Bianchi and Lugosi, 2006]. However, we see that current bandit algorithms at $\beta = -0.09$ converge to the deterministic $\beta = -0.045$ solution, and bandit algorithms playing against the bandit at $\beta = -1$ approximate the $\beta = -0.5$ solution. Thus, they will all exhibit linear regret compared to the optimal mixed strategy.

5 Bayesian reactive bandit algorithm

In this section, we propose a novel Bayesian bandit algorithm for the player that models the bandit reaction, along with the expected means and variances of each arm. We employ the following conjugate prior on each arm:

$$P(\mu_i, \tau_i, \beta | \{a_i, b_i, A^i\}) \propto \tau_i^{a_i - 1} e^{-b_i \tau_i - \frac{\tau_i}{2} v_i^T A^i v_i} \tag{35}$$

where $\tau_i = 1/\sigma_i^2$ describes the precision of the arm and $v_i = [\mu_i, \beta/\tau_i, 1]$. The distribution is parameterized by shape a_i and scale b_i gamma parameters and a 3×3 symmetric matrix A^i. When $\beta = 0$ is known and constrained, the conditional distribution is equivalent to the conventional Normal-Gamma distribution.

For each time t, the frequency of past actions $\hat{p}^t$ is used along with the reward observation r_i^t to update the belief of the corresponding pulled arm:

$$\begin{aligned}
a_i &\leftarrow a_i + \frac{1}{2} && (36)\\
b_i &\leftarrow b_i + \frac{1}{2}(r_i^t)^2 && (37)\\
A^i_{\mu\mu} &\leftarrow A^i_{\mu\mu} + 1 && (38)\\
A^i_{\mu\beta} &\leftarrow A^i_{\mu\beta} + \hat{p}_i^t && (39)\\
A^i_{\beta\beta} &\leftarrow A^i_{\beta\beta} + (\hat{p}_i^t)^2 && (40)\\
A^i_{\mu 1} &\leftarrow A^i_{\mu 1} + r_i^t && (41)\\
A^i_{\beta 1} &\leftarrow A^i_{\beta 1} + \hat{p}_i^t r_i^t && (42)
\end{aligned}$$

These distributions are then used for Thompson sampling to generate a sample at each time from the posterior belief about the bandit: $\{\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_K, \hat{\sigma}_1^2, \hat{\sigma}_2^2, ..., \hat{\sigma}_K^2, \hat{\beta}\}$ We use Gibbs sampling to generate samples from this posterior distribution. These samples are then used to solve for the optimal policy $\vec{p}^t$:

$$\vec{p}^t = \arg\max_{\vec{p}} \sum_i \left[p_i \hat{\mu}_i + \hat{\beta} p_i^2 \hat{\sigma}_i^2 \right] \tag{43}$$

The solution to this quadratic programming optimization is given by our analysis in Section 3. Depending

upon the posterior samples $\hat{\beta}$ and mean and variances estimates of the arms, the optimal strategy $\vec{p}^t$ can be deterministic or mixed. If $\vec{p}^t$ is deterministic, the optimal arm is chosen to pull. Otherwise, if $\vec{p}^t$ is mixed, the action I_t is generated by sampling from the mixed distribution.

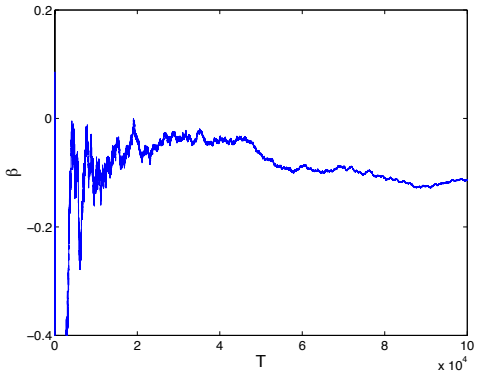

Figure 3: Bayesian estimates $\hat{\beta}$ over time for the $K = 3$, $\beta = -0.09$ Gaussian bandit.

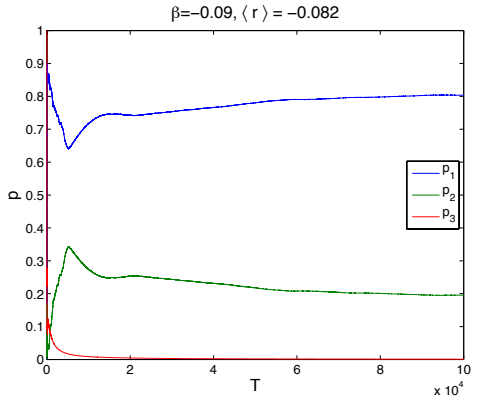

Figure 4: Convergence of the mixed strategy from the Bayesian reactive bandit algorithm over time for the $K = 3$, $\beta = -0.09$ Gaussian bandit.

In Figure 3, we show how our Bayesian algorithm is able to estimate the unknown β of the bandit. In this case, it is able to learn that the bandit is slightly adversarial with a $\beta = -0.09$. It uses that estimate to generate actions that correspond to the optimal mixed strategy. This is shown in Figure 4 where the Bayesian algorithm with Thompson sampling achieves the optimal reward by mixing the first two arms.

6 Correlated Gaussian bandit

We can extend our previous analysis to the case where the Gaussian distribution $Q_0(\vec{r})$ contains correlations:

$$Q_0(\vec{r}) = \frac{1}{\sqrt{(2\pi)^K \det C}} \exp\left[-\frac{1}{2}(\vec{r}-\vec{\mu})^T C^{-1}(\vec{r}-\vec{\mu})\right] \tag{44}$$

with means $\langle \vec{r} \rangle = \mu$ and positive definite covariance $\left\langle (\vec{r}-\vec{\mu})(\vec{r}-\vec{\mu})^T \right\rangle = C$.

The free energy can also be calculated analytically:

$$F_\beta(\vec{p}) = \beta \vec{p} \cdot \vec{\mu} + \frac{\beta^2}{2} \vec{p}^T C \vec{p} \tag{45}$$

With correlations, the optimal policy is given by analyzing the Lagrangian:

$$\max_{\vec{p}} \min_{\vec{\alpha} \geq 0, \lambda} L_\beta(\vec{p}) = \vec{p} \cdot \vec{\mu} + \beta \vec{p}^T C \vec{p} + \vec{\alpha} \cdot \vec{p} - \lambda \left(\sum_i p_i - 1 \right) \tag{46}$$

Taking derivatives yields the corresponding KKT conditions:

$$\vec{\mu} + 2\beta C \vec{p} = \lambda \vec{1} - \vec{\alpha} \tag{47}$$

$$\alpha_i p_i = 0 \tag{48}$$

$$\sum_i p_i = 1 \tag{49}$$

For the fully stochastic bandit, $\beta = 0$, the optimal policy will depend only upon maximizing the means μ_i. In the following, we analyze the optimal player strategies for non-zero β.

6.1 Optimal deterministic policy

First we show that even with correlations, the optimal policy for $\beta > 0$ is deterministic.

Proposition 2. *The optimal player strategy $\vec{p}^\star$ for $\beta > 0$ is deterministic, where the optimal arm is given by:*

$$i^\star = \arg\max_i \left(\mu_i + \beta C_{ii}\right). \tag{50}$$

Proof. We show that this deterministic strategy satisfies the KKT conditions. First, we have that:

$$2\beta C_{i^\star i^\star} = \lambda - \mu_{i^\star} \tag{51}$$

Then we must show that $\alpha_{i \neq i^\star} > 0$. This is obtained by considering:

$$\alpha_{i \neq i^\star} = \lambda - \mu_i - 2\beta C_{i i^\star} \tag{52}$$

$$= \mu_{i^\star} - \mu_i + 2\beta C_{i^\star i^\star} - 2\beta C_{i i^\star} \tag{53}$$

$$= \left[(\mu_{i^\star} + \beta C_{i^\star i^\star}) - (\mu_i + \beta C_{ii})\right] + \beta \left[C_{i^\star i^\star} - C_{i i^\star} - C_{i^\star i} + C_{ii}\right] \tag{54}$$

$$> 0 \tag{55}$$

The first bracketed term is positive because $i^\star$ is defined as the maximum in Eq. 50. The second bracketed term is also positive because C is symmetric positive definite: $v^T C v > 0$ where v is the sparse vector with $v_{i^\star} = 1$, $v_i = -1$ being the only non-zero components. Any mixing of arms will result in less reward, so the deterministic strategy of choosing the arm $i^\star$ is optimal. □

Thus, we see that for $\beta > 0$ the correlations are irrelevant with respect to determining the optimal policy $\vec{p}^\star$. The optimal policy is deterministic, given by maximizing the linear combination of mean μ_i and variance C_{ii}.

6.2 Optimal stochastic policy

What about for $\beta \to -\infty$? The optimal policy is given by minimizing

$$\vec{p}^\star = \arg\min_{\vec{p}} \vec{p}^T C \vec{p} \tag{56}$$

on the probability simplex. Naively, we may expect the optimal policy to be given by $\vec{p}^\star = (C^{-1}\vec{1})/(\vec{1}^T C^{-1}\vec{1})$ as for the independent Gaussian distribution, but this solution need not contain all positive components. In general, the optimal $\vec{p}^\star$ could be sparse with zero components.

We need to solve Eq. 56 for a particular covariance matrix C to know exactly how many arms will be mixed in the optimal solution. Similarly, for intermediate $\beta < 0$, the optimal policy will be obtained by solving a convex quadratic program over the probability simplex. This can also be done using projected gradient descent, or with multiplicative updates [Sha et al., 2007].

7 Conclusions

In this manuscript, we have introduced a class of reactive bandits that modulate their reward distributions in response to the past actions of the player. We can relate this bandit model to a statistical mechanical description of the rewards reacting to an external field generated by the player history. For $\beta > 0$, the rewards partially align with the player actions, and for $\beta < 0$ are anti-aligned with the player actions in the adversarial regime. We give analytic solutions for this model when the underlying stochastic distribution is Gaussian. In particular, we completely characterize the optimal solution space and show that current bandit algorithms achieve linear regret for finite $\beta < 0$ compared to the optimal mixed player strategy. We propose a novel algorithm with Bayesian estimation and Thompson sampling and show that it is able to asymptotically achieve optimal performance on these bandits.

A more formal analysis of the convergence of the Bayesian bandit algorithm for this class of bandit models is left for future work. Further experiments with other stochastic distributions such as Bernoulli or Poisson distributions also need to be performed. However, we anticipate that the analysis of the Gaussian model will open new avenues for investigation into the problem of reactive bandits.

Acknowledgements

We thank the anonymous reviewers for their valuable comments and suggestions for improving this manuscript. This study was funded by grants from the U.S. National Science Foundation, Office of Naval Research, and Department of Transportation.

References

R. Arora, O. Dekel, and A. Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of ICML*, 2012.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1), 2002.

D. Blackwell and M. A. Girshick. *Theory of games and statistical decisions*. John Wiley & Sons, 1954.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5:1–122, 2012.

S. Bubeck and R. Munos. Open loop optimistic planning. In *Proceedings of COLT*, number 1, page 15, 2010.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. Online optimization in X-armed bandits. In *Advances in Neural Information Processing Systems (NIPS) 21*, pages 201–208, 2008.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Computational Learning Theory (COLT)*, pages 355–366, 2008.

S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems (NIPS) 23*, pages 586–594, 2010.

L. Kocsis and C. Szepesvári. Bandit based Monte-Carlo planning. In *Proceedings of ECML*, pages 282–203, 2006.

J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of NIPS*, pages 1–8, 2008.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web (WWW)*, pages 661–670, New York, New York, USA, Apr. 2010. ACM Press.

P. Ortega and D. Braun. Information, utility and bounded rationality. In *Artificial General Intelligence*, pages 269–274, 2011.

J. G. Proakis, M. Salehi, N. Zhou, and X. Li. *Communication systems engineering*, volume 2. Prentice-Hall, Englewood Cliffs, 1994.

D. Pucci de Farias and N. Megiddo. Combining expert advice in reactive environments. *Journal of the ACM*, 53(5):762–799, Sept. 2006.

J. Robinson. An iterative method of solving a game. *Annals of mathematics*, pages 296–301, 1951.

F. Sha, Y. Lin, L. K. Saul, and D. D. Lee. Multiplicative updates for nonnegative quadratic programming. *Neural Computation*, 19(8):2004–2031, 2007.

J. Von Neumann and O. Morgenstern. Theory of games and economic behavior. *Bull. Amer. Math. Soc*, 51(7):498–504, 1945.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of ICML*, pages 421–422, 2003.