Reactive Bandits with Attitude

Pedro A. Ortega, Kee-Eung Kim and Daniel D. Lee

Motivation

- Most of the literature on multi-armed bandits deals with either one of two general classes of bandits: **stochastic** and **adversarial**.
- Bubeck & Slivkins (2012) and Seldin & Slivkins (2014) have presented algorithms that achieve optimal performance in both classes of bandits.
- This unification is important: modelling and identifying the "**attitude**" of a bandit from data has applications in systems that are risk-sensitive, *e.g.* systems that must prevent attacks or adaptively build trust in its users.
- We introduce a bandit model that can instantiate the **full continuum** from **adversarial**, to **stochastic**, and even to **cooperative** bandits by varying a single **attitude** parameter.

Reactive Bandits

Model: In each round, the player issues action I from a (mixed) strategy $\vec{p}$. The bandit then replies with a reward $\vec{r}$ drawn from the **reactive distribution**

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

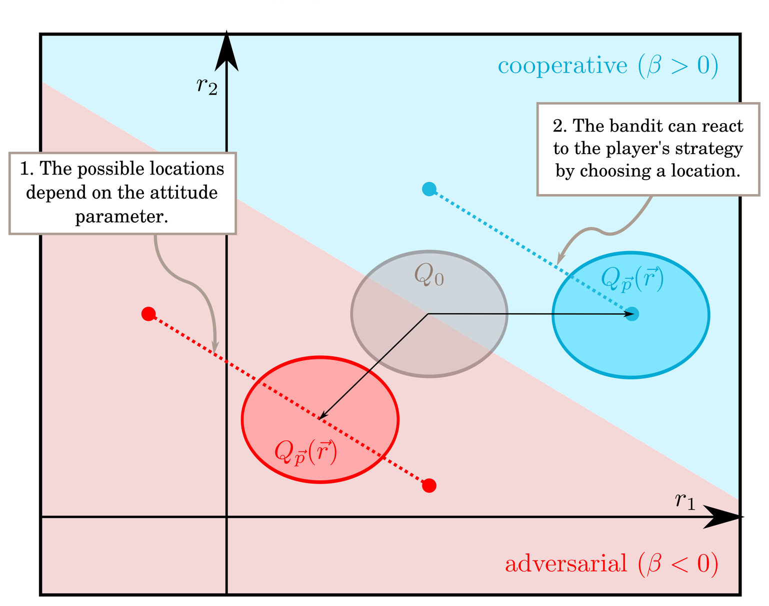
where Q_0 is a **reference distribution**, $\beta \in \mathbb{R}$ is an **attitude parameter** that controls the strength of the bandit's reaction to the player's policy, and $Z_{\vec{p}}$ is a normalizing constant.

Why? Because $Q_{\vec{p}}$ maximizes the **free energy**

$$F_{\vec{p}} = \max_{Q} \Big\{ \beta \cdot \underbrace{\mathbb{E}_Q\big[\vec{p}\cdot\vec{r}\big]}_{\text{Expected Reward}} - \underbrace{D_{\mathrm{KL}}\big[Q(\vec{r}) \big\| Q_0(\vec{r})\big]}_{\text{KL Regularization}} \Big\}$$

Example: Gaussian case

$$Q_{\vec{p}}(\vec{r}) = \prod_{k=1}^{K} \mathcal{N}(r_k; \mu_k + \underline{\beta\sigma_k^2 p_k}, \sigma_k^2)$$



The bandit reacts by shifting the mean either **towards** ($\beta > 0$) or **against** ($\beta < 0$) more probable actions.

Optimal Strategy

Goal: Maximize the expected reward:

$$\mathbb{E}_{Q_{\vec{p}}}[r|\vec{p}] = \sum_k p_k \Big[\int Q_{\vec{p}}(\vec{r}) r_k \, d\vec{r} \Big].$$

The **optimal strategy** depends on β:

Case $\beta > 0$: The optimal strategy is **deterministic**:

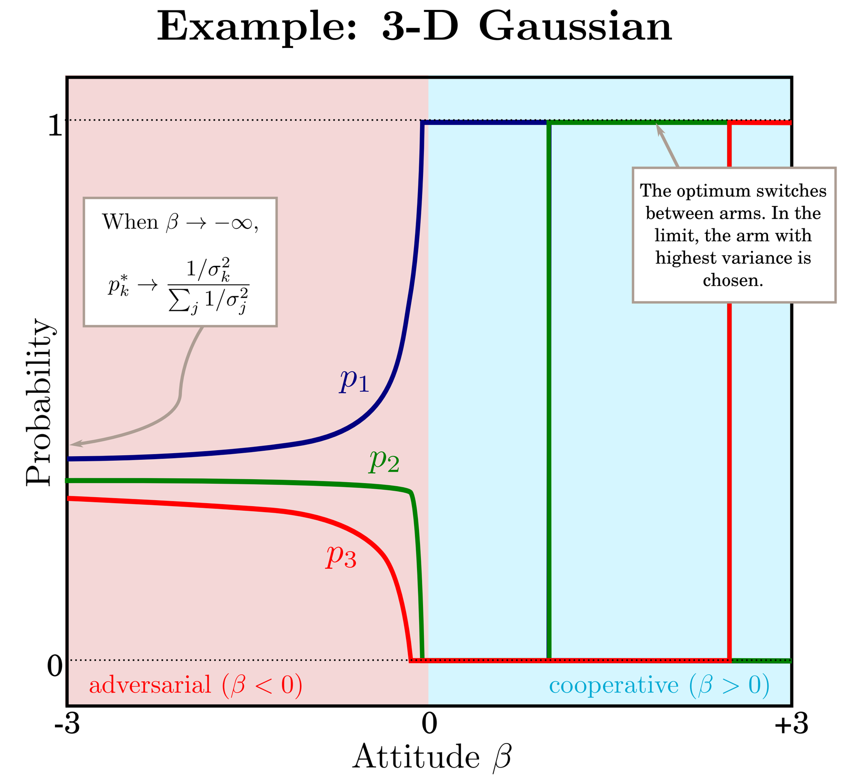
$$I^* = \arg\max_i (\mu_i + \beta\sigma_i^2)$$

Case $\beta < 0$: The optimal strategy is

$$p_k^* = \max\Big\{ \frac{\lambda - \mu_k}{2\beta\sigma_k^2}, 0 \Big\}$$

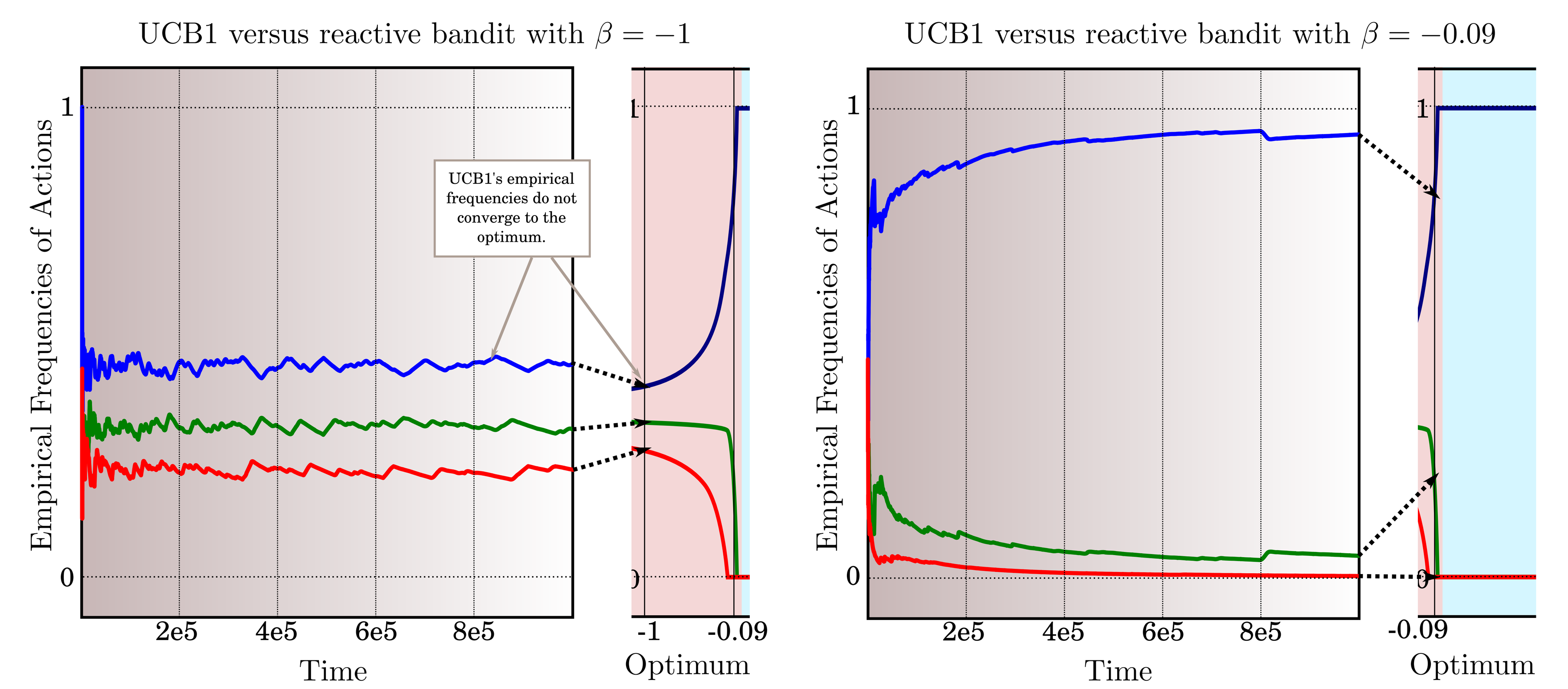
where λ ensures that $\sum_k p_k = 1$. Algorithmically, λ is obtained through a **water-filling** algorithm. In general, it is **stochastic**.

Example: 3-D Gaussian



Bandit Algorithms

Algorithms, like UCB1, do not learn the optimal strategy.



Learning

Model: The bandit's reactive distribution can be learned using a **Bayesian model**. We use the **conjugate prior**

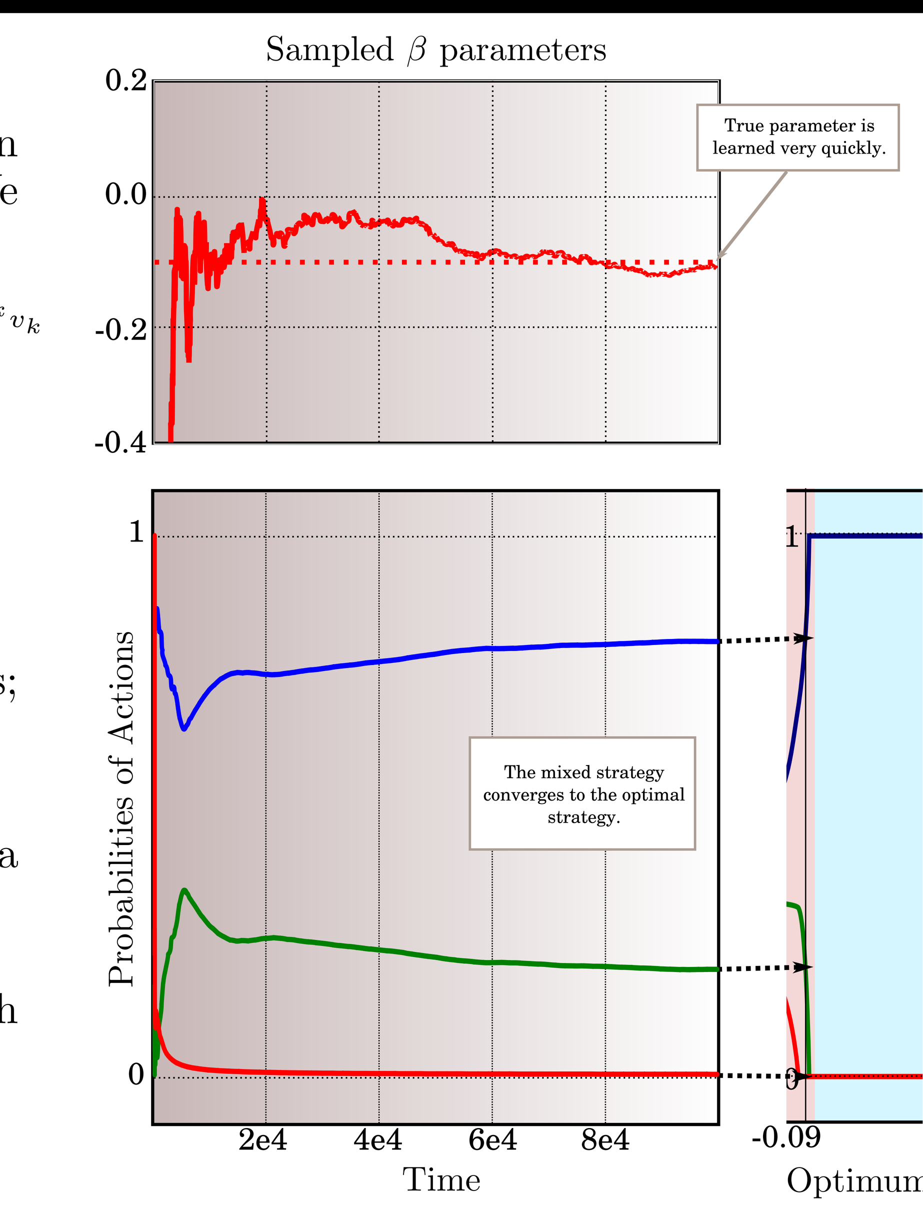
$$P(\mu_k, \tau_k, \beta | \{a_k, b_k, A^k\}) \propto \tau_k^{a_k - 1} e^{-b_k \tau_k - \frac{1}{2}\tau_k v_k^T A^k v_k}$$

on each arm, where:

- $\tau_k = 1/\sigma_k^2$ is the precision;
- $v_k = [\mu_k, \beta/\tau_k, 1]$;
- a_k and b_k are Gamma shape parameters;
- and A^k is a 3×3 symmetric matrix.

When $\beta = 0$ is known, this corresponds to a **Normal-Gamma** distribution.

Algorithm: We use the Bayesian model with **Thompson sampling**.



Conclusions

- We introduce a **class of reactive bandits** that modulate their reward distribution in response to the past actions of the player.
- For $\beta > 0$, rewards **partially align** with the player.
- For $\beta < 0$, rewards **partially counteract** the player's strategy.
- The Gaussian case has analytic solutions and a simple optimal policy, which is **mixed** in the adversarial case.
- Current bandits algorithms do not possess the necessary strategy space and thus cannot achieve sublinear regret.
- We show that these bandits can be played using a Bayesian model in combination with Thompson sampling.

GRASP Laboratory, School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia, U.S.A.