A Proofs

This section provides the proof sketch of Lemmas and Theorems mentioned in the main paper.

Lemma 1. For the K-armed stochastic MAB problem, Thompson Sampling has expected regret: $\mathbb{E}[R_T^{\text{MAB}}] = \mathcal{O}\left(\frac{K}{\Delta}\ln T\right)$, where Δ is the difference between expected rewards of the best two arms.

Proof. This lemma is a direct result from Theorem 2 of Agrawal & Goyal (2012) and Theorem 1 of Kaufmann et al. (2012). □

Lemma 2. Running INDSELFSPARRING with infinite time horizon will sample each arm infinitely often.

Proof. Proof by contradiction.
Let $B(x;\alpha,\beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt$. Then the CDF of Beta distribution with parameters (α,β) is

$$F(x;\alpha,\beta) = \frac{B(x;\alpha,\beta)}{B(1;\alpha,\beta)}.$$

Suppose arm b can only be sampled in finite number of iterations. Then there exists finite upper bound T_b for $\alpha_b + \beta_b$. For any given $x \in (0,1)$, the probability of sampling values of arm b θ_b greater than x is

$$P(\theta_b > x) = 1 - F(x;\alpha_b,\beta_b)$$

$$\geq 1 - F(x;1,T_b-1) = (1-x)^{T_b-1} > 0$$

Then by running INDSELFSPARRING, the probability of choosing arm b after it has been chosen T_b times:

$$P(\theta_b \geq max_i\{\theta_{b_i}\}) \geq \prod_i P(\theta_b \geq \theta_{b_i})$$

is strictly non-zero. That violates any fixed upper bound T_b. □

Theorem 1. Under Approximate Linearity, INDSELFSPARRING converges to the optimal arm b_1 as running time $t \to \infty$: $\lim_{t\to\infty}\mathbb{P}(b_t = b_1) = 1$.

Proof. INDSELFSPARRING keeps one Beta distribution $Beta(\alpha_i(t),\beta_i(t))$ for each arm b_i at time step t. Let $\hat{\mu}_i(t) = \frac{\alpha_i(t)}{\alpha_i(t)+\beta_i(t)}$, $\hat{\sigma}_i^2(t) = \frac{\alpha_i(t)\beta_i(t)}{(\alpha_i(t)+\beta_i(t))^2(\alpha_i(t)+\beta_i(t)+1)}$ be the empirical mean and variance for arm b_i.
Obviously, $\hat{\sigma}_i^2(t) \to 0$ as $(\alpha_i(t)+\beta_i(t)) = (S_i(t)+F_i(t)) \to \infty$. By Lemma 2 we have $(S_i(t)+F_i(t)) \to \infty$ as $t \to \infty$. That shows every Beta distribution is concentrating to a Dirac function at $\hat{\mu}_i(t)$ when $t \to \infty$.

Define $\hat{\mu}(t) = [\hat{\mu}_1(t),\cdots,\hat{\mu}_K(t)]^T \in [0,1]^K$to be the vector of means of all arms. Then $\mu = \{\mu_i = P(b_i \succ b_1)\}_{i=1,\cdots,K}$ is a stable point for INDSELFSPARRING in the K dimensional mean space.

Suppose there exists another stable point $\nu \in [0,1]^K (\nu \neq \mu)$ for INDSELFSPARRING, consider the following two possibilities: (1) $\nu_1 = max_i\{\nu_i\}$ and (2) $\nu_1 < max_i\{\nu_i\} = \nu_j$.

Since the Beta distributions for each arm b_i is concentrating to Dirac functions at ν_i, $P(\theta_i > \theta_j) \in [\mathbb{I}(\nu_i > \nu_j) - \delta, \mathbb{I}(\nu_i > \nu_j) + \delta]$ for any fixed $\delta > 0$ with high probability.

If (1) holds, then ν_1 will converge to $\frac{1}{2} = \mu_1$ and ν_i will converge to $P(b_i \succ b_1) = \mu_i$. Thus $\nu = \mu$. Contradict to $\nu \neq \mu$.

If (2) holds, then ν_j will converge to $\frac{1}{2} = \mu_1$ and $\nu_1 \in [P(b_1 \succ b_j) - \delta, P(b_1 \succ b_j) + \delta]$ for any fixed $\delta > 0$ with high probability. Since $P(b_1 \succ b_j) \geq \frac{1}{2} + \Delta$, $\nu_1 \in [P(b_1 \succ b_j) - \delta, P(b_1 \succ b_j) + \delta] \geq \frac{1}{2} + \Delta - \delta$. Since δ can be arbitrarily small, we have $\nu_1 \geq \frac{1}{2} + \Delta - \delta > \frac{1}{2} + \delta > \nu_j$. That contradict to $\nu_1 < \nu_j$.

In summary, $\mu = \{\mu_i = P(b_i \succ b_1)\}_{i=1,\cdots,K}$ is the only stable point in the mean space. As $\hat{\mu}(t) \to \mu$, $\mathbb{P}(b_t = b_1) \to 1$.

Define $\mathbb{P}_t = [P_1(t), P_2(t), ..., P_K(t)]$ as the probabilities of picking each arm at time t. Let $\mathbb{P} = \{\mathbb{P}_t\}_{t=1,2,...}$ be the sequence of probabilities w.r.t. time. Assume INDSELFSPARRING is non-convergent. It is equivalent to say that $\mathbb{P}$ is not converging to a fixed distribution. Then $\exists\delta > 0$ and arm i s.t. the sequence of probabilities $\{P_i(t)\}_t$ satisfies:

$$\limsup_{t\to\infty} P_i(t) - \liminf_{t\to\infty} P_i(t) > \delta$$

w.h.p. which is equivalent of having:

$$\limsup_{t\to\infty} \hat{\mu}_i(t) - \liminf_{t\to\infty} \hat{\mu}_i(t) > \epsilon$$

w.h.p. for some fixed $\epsilon > 0$. This violates the stability of INDSELFSPARRING in the K dimensional mean space as shown above. So as $t \to \infty$, $\hat{\mu}(t) \to \mu$, $\mathbb{P}(b_t = b_1) \to 1$.

□

Lemma 3. Under Approximate Linearity, selecting only one arm via Thompson sampling against a fixed distribution over the remaining arms leads to optimal regret w.r.t. choosing that arm.

Proof. We first prove the results for $m = 2$. Results for any $m > 2$ can be proved in a similar way.

Consider Player 1 drawing arms from a fixed distribution L. Player 2's drawing strategy is an MAB algorithm $\mathcal{A}$.

Let $R_A(T)$ be the regret of algorithm $\mathcal{A}$ within horizon T. $B(T) = \sup \mathbb{E}[R_A(T)]$ is the supremum of the expected regret of $\mathcal{A}$.

The reward of Player 2 at iteration t is $\phi(b_{2t}, b_{1t})$. Reward of keep playing the optimal arm is $\phi(b_1, b_{1t})$. So the total regret after T rounds is

$$R_A(T) = \sum_{t=1}^{T} [\phi(b_1, b_{1t}) - \phi(b_{2t}, b_{1t})]$$

Since Approximate Linearity yields

$$\phi(b_1, b_{1t}) - \phi(b_{2t}, b_{1t}) \geq \gamma \cdot \phi(b_1, b_{2t})$$

We have

$$\mathbb{E}[R_A(T)] = \mathbb{E}\mathbb{E}_{b_{1t} \sim L} \left[\sum_{t=1}^{T} [\phi(b_1, b_{1t}) - \phi(b_{2t}, b_{1t}) \right]$$

$$\geq \mathbb{E}\mathbb{E}_{b_{1t} \sim L} \left[\sum_{t=1}^{T} \gamma \cdot \phi(b_1, b_{2t}) \right]$$

$$= \gamma \cdot \mathbb{E} \left[\sum_{t=1}^{T} \phi(b_1, b_{2t}) \right] = \gamma \cdot \mathbb{E}[R(T)]$$

So the total regret of Player 2 is bounded by

$$\mathbb{E}[R(T)] \leq \frac{1}{\gamma} \mathbb{E}[R_A(T)] \leq \frac{1}{\gamma} \sup \mathbb{E}[R_A(T)] = \frac{1}{\gamma} B(T)$$

□

Corollary 1. *If approximate linearity holds, competing with a drifting but converging distribution of arms guarantees the one-side convergence for Thompson Sampling.*

Proof. Let D_t be the drifting but converging distribution and $D_t \to D$ as $t \to \infty$. Let b_T be the drifting mean bandit of D_T after T iterations. Since D_t is convergent, $\exists T > K$ such that

$$\phi(\sup_{t>T} b_T, \inf_{t>T} b_T) < \phi(b_1, b_2)$$

where $\phi(b_1, b_2)$ is the preference between the best two arms. The mean value of feedback by playing arm i is $\phi(b_i, b_T)$. If b_T is fixed, by Lemma3, Thompson sampling converges to the arm: $i^* = \text{argmax}_i \, \phi(b_i, b_T)$. For drifting b_T, define $b^+ = \sup_{t>T} b_T$ and $b^- = \inf_{t>T} b_T$.

Thompson sampling convergence to the optimal arm implies that:

$$\phi(b_1, b^+) > \phi(b_i, b^-)$$

for all $i \neq 1$. Consider:

$$\begin{aligned} &\phi(b_1, b^+) - \phi(b_2, b^-) \\ &= \phi(b_1, b^+) - \phi(b_2, b^-) + \phi(b_1, b^-) - \phi(b_1, b^-) \\ &= \phi(b_1, b^-) - \phi(b_2, b^-) + \phi(b_2, b^+) - \phi(b_1, b^-) \\ &\geq \gamma \cdot [\phi(b_1, b_2) - \phi(b^+, b^-)] > 0 \end{aligned}$$

by approximate linearity.

So we have $\phi(b_1, b^+) > \phi(b_2, b^-)$. Since $\phi(b_2, b^-) > \phi(b_i, b^-)$ for $i > 2$. Then we have

$$\phi(b_1, b^+) > \phi(b_i, b^-)$$

holds for all $i \neq 1$. So Thompson sampling converge to the optimal arm. □

Theorem 2. Under Approximate Linearity, IndSelfSparring converges to the optimal arm with asymptotically optimal no-regret rate of $\mathcal{O}(K \ln(T)/\Delta)$. Where Δ is the difference between the rewards of the best two arms.

Proof. Theorem 1 provides the convergence guarantee of IndSelfSparring. Corollary 1 shows one-side convergence for playing against a converging distribution.

Since IndSelfSparring converges to the optimal arm b_1 as running time $t \to \infty$: $\lim_{t\to\infty} \mathbb{P}(b_t = b_1) = 1$. For $\forall \delta > 0$, there exists $C(\delta) > 0$ such that for any $t > C(\delta)$, the following condition holds w.h.p.: $P(b_t = b_1) \geq 1 - \delta$.

For the triple of bandits $b_1 \succ b_i \succ b_K$, Approximate Linearity guarantees:

$$\phi(b_i, b_K) < \phi(b_1, b_K) \leq \omega$$

holds for some fixed $\omega > 0$ and $\forall i \in \{2, \cdots, K-1\}$. With small δ, the competing environment of any Player p is bounded. If $\delta < \frac{\Delta}{\Delta+\omega}$, $(1-\delta)\cdot(-\Delta)+\delta\cdot\phi(b_2, b_K) < 0 = 1 \cdot \phi(b_1, b_1)$. The competing environment can be considered as unbiased and the theoretical guarantees for Thompson sampling for stochastic multi-armed bandit is valid (up to a constant factor).

Then IndSelfSparring has an no-regret guarantee that asymptotically matches the optimal rate of $\mathcal{O}(K \ln(T)/\Delta)$ up to constant factors, which proves Theorem 2.

□

B Further Experiments

We run further experiments on 16-arm synthetic datasets. The distributions of utilities of arms are shown in Table 2. We compared the performances of 8 algorithms and 15 scenarios as shown in Figure 12.

Name	Distribution of Utilities of arms
1good	1 arm with utility 0.8, 15 arms with utility 0.2
2good	1 arm with utility 0.8, 1 arms with utility 0.7, 14 arms with utility 0.2
6good	1 arm with utility 0.8, 5 arms with utility 0.7, 10 arms with utility 0.2
arith	1 arm with utility 0.8, 15 arms forming an arithmetic sequence between 0.7 and 0.2
geom	1 arm with utility 0.8, 15 arms forming a geometric sequence between 0.7 and 0.2

Table 2: 16-arm synthetic datasets used for experiments.

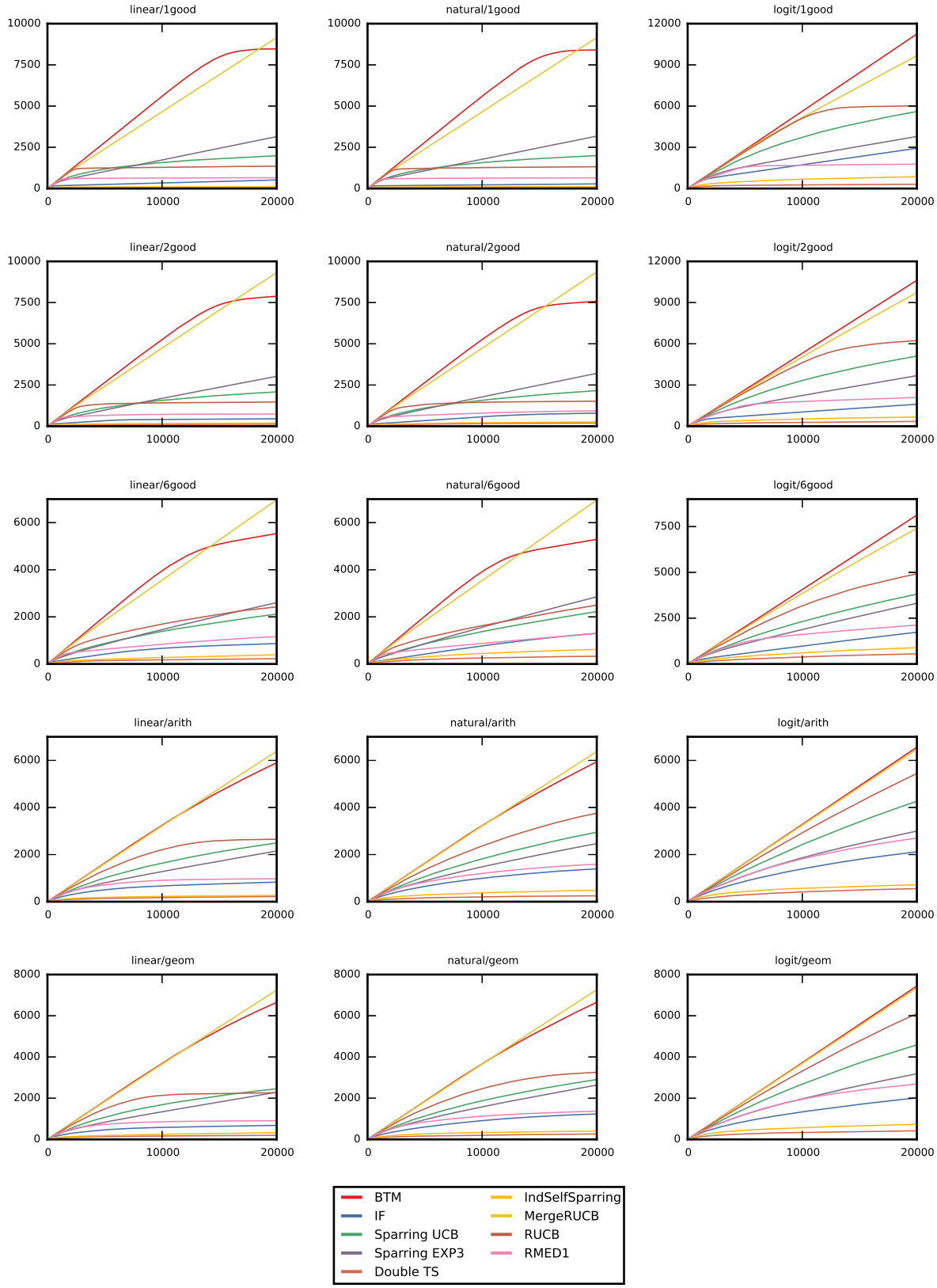

Figure 12: Average regret vs iterations for each of 8 algorithms and 15 scenarios.