

Thompson Sampling Algorithm for Stochastic Differential Games

Yuqiong Wang
yuqw@umich.edu

Joint work with Asaf Cohen and Ruolan He

XIII Bachelier World Congress, Bologna

June 30, 2026

Motivation

In some games, many agents respond to a **common signal** that no one knows exactly.

- **Regulatory signal**
- **Market signal**

We consider a N -player competitive stochastic game where player i controls a private diffusion $X^i \in \mathbb{R}^d$ with:

$$dX_t^i = (AX_t^i - \alpha_t^i) dt + \sigma^i dW_t^i, \quad i = 1, \dots, N.$$

- a **common unknown drift** $A \in \mathbb{R}^{d \times d}$ entering every agent's dynamics,
- each agent observes **only its own trajectory**, all W 's are independent, and independent of A ;
- agents **compete through costs**, not by touching each other's state;
- an **ergodic** quadratic cost:

$$J^i(\mathbf{X}_0, \alpha) = \liminf_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}^i \left[\int_0^T \left((\mathbf{X}_t - \bar{\mathbf{x}}_i)^\top \mathcal{Q}^i (\mathbf{X}_t - \bar{\mathbf{x}}_i) + \frac{1}{2} (\alpha_t^i)^\top R^i \alpha_t^i \right) dt \right]$$

Questions

- Can we find learning-based strategies that give a Nash Equilibrium?
- What is the price of not knowing A ?

Background:

- Ergodic LQ N -player games with *known* A : explicit affine Nash feedbacks (Bardi–Priuli '14).
- Filtering of an unknown drift.
- Thompson sampling (TS) for *single-agent* LQ control: $O(\sqrt{T \log T})$ regret (Ouyang et al. '17/'20, Gagrani et al. '21).

What we do:

- Carry TS from single-agent control to an N -player game with partial information.
- Regret of the *same order* $O(\sqrt{T \log T})$ independent of N , under weaker assumptions.
- The TS profile is a *Nash equilibrium*.

The full-information game (Bardi–Priuli, 2014)

- Assume A is known to all players.
- Set the **Hamiltonian** for Player i :

$$H^i(x, p) := \max_{a \in \mathbb{R}^d} \left\{ -\frac{1}{2} a^\top R^i a - p^\top (Ax - a) \right\} = \frac{1}{2} p^\top (R^i)^{-1} p - p^\top Ax,$$

- Using ergodicity, integrate the others out against their stationary laws m^{-i} :

$$\tilde{f}^i(x; m^{-i}) = \int \tilde{F}^i(\xi^1, \dots, x, \dots, \xi^N) \prod_{j \neq i} dm^j(\xi^j).$$

- Denote $\varsigma^i = \frac{1}{2}(\sigma^i)(\sigma^i)^\top$. Then the **stationary HJB–KFP** system associated with the game is:

$$\begin{aligned} & -\operatorname{tr}(\varsigma^i D^2 v^i(x)) + H^i(x, \nabla v^i(x)) + \lambda^i = \tilde{f}^i(x; m^{-i}), \\ & -\operatorname{tr}(\varsigma^i D^2 m^i(x)) - \operatorname{div} \left(m^i(x) \frac{\partial H^i}{\partial p}(x, \nabla v^i(x)) \right) = 0, \\ & \int_{\mathbb{R}^d} m^i(x) dx = 1, \quad m^i(x) > 0. \end{aligned}$$

The full information game (Bardi–Priuli, 2014)

- Under some assumptions, the HJB–KFP system has a unique solution, explicit in A :

$$v^i(x) = \frac{1}{2}x^\top \Lambda^i x + (\rho^i)^\top x, \quad m^i = \mathcal{N}(\eta^i, (\Upsilon^i)^{-1}), \quad \lambda^i \in [0, \infty),$$

- where $\Upsilon^i(A)$ solves the Riccati equation $\frac{1}{2} \Upsilon \varsigma^i R^i \varsigma^i \Upsilon = \frac{1}{2} A^\top R^i A + Q_{ii}^i$.
- Affine Nash feedback

$$\bar{a}^i(x; A) = (\varsigma^i \Upsilon^i(A) + A)x - \varsigma^i \Upsilon^i(A) \eta^i(A).$$

- The resulting state process is Ornstein–Uhlenbeck:

$$dX_t^i = (-\varsigma^i \Upsilon^i X_t^i + \varsigma^i \Upsilon^i \eta^i) dt + \sigma^i dW_t^i..$$

Notations

- No hat = truth, full info. e.g., A, X_t^i, α_t^i
- Hat = sampled. e.g., $\hat{A}_k^i, \hat{X}_t^i, \hat{\alpha}_t^i$.

Partial information game: A unknown

- Now consider A unknown. Not solvable.
- **Goal:** Find an **equilibrium strategy** profile $\hat{\alpha}$ with a efficient learning of A such that as $T \rightarrow \infty$, the players behave as if they have full information. We use the Thompson Sampling (TS) algorithm.
- Idea of the algorithm: sample a reasonable $\hat{A}$, then play $\bar{a}^i(\cdot; \hat{A})$ as if $\hat{A}$ were true.

Prior Distribution of the Unknown Drift A

For any $i \in [N]$,

$$A \sim \mathcal{N}(\mu_0^i, \Sigma_0^i) \mid \Theta$$

- Θ is a compact set.
- Assume boundedness, stability. Relax on independence.

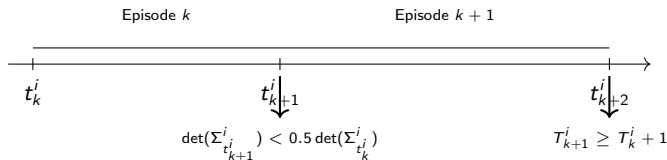
Each Player i Maintains a Posterior $\mathcal{N}(\mu_t^i, \Sigma_t^i)$ Over A :

- Updates posterior using its own observed trajectory $(\hat{X}_t^i)_{t \geq 0}$
- Filtering is done in real time, independently by each player

Thompson Sampling Algorithm

Initialization: Each player initializes prior (μ_0^i, Σ_0^i) and samples $\hat{A}_0^i \sim \mathcal{N}(\mu_0^i, \Sigma_0^i) \mid_{\Theta(v)}$.

- 1 During episode k , observe the state dynamics $\hat{X}_t^i$, by using the strategy $\hat{\alpha}_t^i = \bar{a}^i(\hat{X}_t^i; \hat{A}_k^i)$ via the latest sampled $\hat{A}_k^i$.
- 2 Update posterior $\mathcal{N}(\mu_t^i, \Sigma_t^i)$.
- 3 **(Stopping criteria):** If either
 - **Interval length stopping criterion:** $t - t_k^i \geq T_{k-1}^i + 1$
 - **Determinant stopping criterion:** $\det(\Sigma_t^i) < 0.5 \det(\Sigma_{t_k^i}^i)$



Two intermediate estimates: learning and coupling

1. Parameter learning estimate

$$\int_0^T \mathbb{E}^i \|A - \hat{A}_{k(t)}\|^2 dt \leq O(\log T).$$

- Heuristically, the posterior covariance decreases $\sim 1/t$
- This estimate controls how wrong the sampled model is over time.

2. State coupling estimate

$$\int_0^T \mathbb{E}^i \|\hat{X}_t^i - X_t^i\|^2 dt \leq O(\sqrt{T \log T}).$$

- $\hat{A}$ is close to A , the feedback coefficients depend regularly on A , the learned trajectory is close to the full-information trajectory:

$$\frac{1}{T} \int_0^T \mathbb{E}^i \|\hat{X}_t^i - X_t^i\|^2 dt \longrightarrow 0.$$

Main Theorem: Regret Bounds

Definition

The **Regret** of Player i up to time T (against the full information cost) is:

$$R^i(\hat{\alpha}^i, T) := \mathbb{E} \left[\int_0^T f^i(\hat{X}_t^i, \hat{\alpha}_t^i; \mathbf{m}^{-i}) dt - T \lambda^i(A) \right]$$

Theorem (Regret bound (Cohen, He, W.))

Player i 's regret $R^i(\hat{\alpha}^i, T) \leq O(\sqrt{T \log T})$.

- The regret bound is uniform with respect to the number of players N .
- We decompose

$$R^i = \underbrace{\mathbb{E} \int_0^T (\lambda^i(\hat{A}_{k(t)}) - \lambda^i(A)) dt}_{R_0 \text{ sampling error}} + \underbrace{\mathbb{E} [v^i(X_0) - v^i(\hat{X}_T)]}_{R_1 \text{ cumulative strategy effect}} + \underbrace{\mathbb{E} \int_0^T (\nabla v^i(\hat{X}_t))^T (A - \hat{A}_{k(t)}) \hat{X}_t dt}_{R_2 \text{ model mismatch}}.$$

- $R_0 \leq O(\sqrt{T \log T})$, $R_1 \leq O(1)$, $R_2 \leq O(\sqrt{T \log T})$.

Main Theorem: Equilibrium

Theorem (Nash Equilibrium (Cohen, He, W.))

If every player runs TS, then $(\hat{\alpha}^1, \dots, \hat{\alpha}^N)$ is a Nash equilibrium, and each $\hat{X}^i$ is ergodic with the same limiting distributions m^i as in the full information case.

Remark. The control itself converges sublinearly.

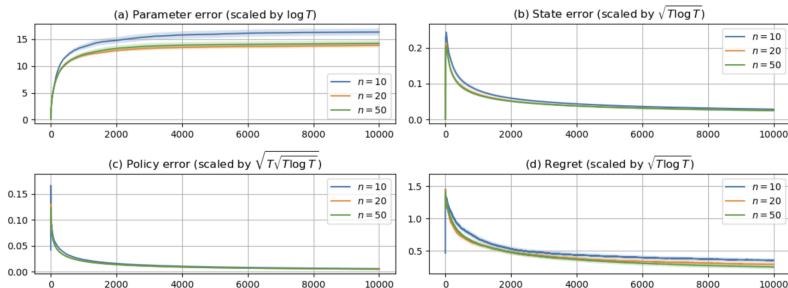
Experiment Setup:

- **State dimension:** $d = 2$; **Number of players:** $N = 10$
- **True drift matrix:** $A = \begin{bmatrix} -0.5 & 0 \\ 0 & -0.5 \end{bmatrix}$
- **Noise:** $\sigma^i = 0.5 \cdot I_d + 0.05 \cdot Z^i$, $Z^i \sim \mathcal{N}(0, I_d)$, independently sampled.
- **Prior:** $A \sim \mathcal{N}(0, 0.5I)$
- **Initial State:** $X_0^i = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}$.

Questions:

- 1 Does learning/state/control error and the regret scale as they should?
- 2 Does TS beat certainty-equivalent (posterior-mean) and blind baselines?
- 3 Is it robust to prior mis-specification?

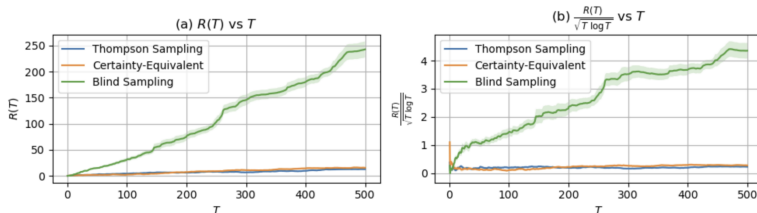
Numerics 1: learning, coupling, and regret



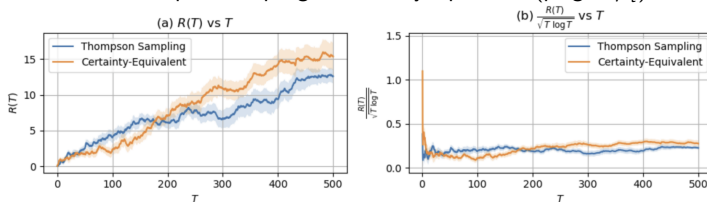
- **Learning:** parameter error scaled by $\log T$ stabilizes, consistent with $\int_0^T \mathbb{E} \|A - \hat{A}_{k(t)}\|^2 dt = O(\log T)$.
- **Coupling:** state and policy errors decay after normalization, so TS tracks the full-information benchmark in time average.
- **Regret:** regret scaled by $\sqrt{T \log T}$ stays bounded/decreases across sample sizes $n = 10, 20, 50$.

Numerics 2: Thompson sampling vs certainty equivalent

With blind sampling



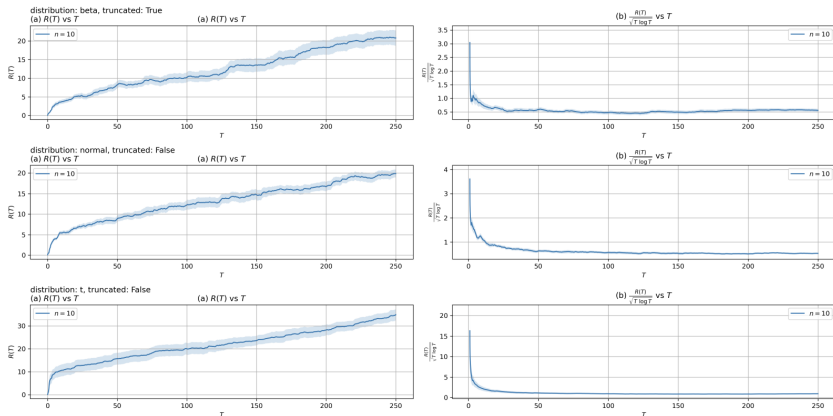
Thompson sampling vs certainty-equivalent (plug in μ_t^i)



- Blind sampling is much worse.
- CE strategy has slightly smaller regret earlier due to greedy exploitation, but fails exploration and accumulates larger regret.

Numerics 3: Robustness to prior mis-specification

Regret under different priors



- Same asymptotic scaling for Gaussian, t , exponential, beta.
- Same asymptotic scaling for both truncated and not truncated.

- **Coupling through dynamics**, one player's learning/control affects another player's observations
- **Closed-loop games**: players observe the *full* state vector and react to each other

Thank you!