Extending Marginal Reputation to Persistent Markovian States

Kyle Elliott Mathewson[1]
with AI Collaborators

February 17, 2026

[1]Faculty of Science, University of Alberta

Abstract

We extend the main result (Theorem 1) of Luo & Wolitzky (2024), "Marginal Reputation," from i.i.d. states to persistent Markovian states. The extension reveals a new phenomenon: when the Stackelberg strategy reveals the state, short-run player beliefs permanently deviate from the stationary distribution, causing the Nash correspondence to become state-contingent. We introduce the concept of *belief-robustness* and present two results. **Theorem 1′** (belief-robust case): when short-run best responses are invariant to the filtering belief $F(\cdot|\theta)$, the original commitment payoff $V(s_1^*)$ holds exactly under Markov states with no correction. **Theorem 1″** (general case): the *Markov commitment payoff* $V_{\text{Markov}}(s_1^*) = \sum_\theta \pi(\theta) \cdot \inf_{B(s_1^*, F(\cdot|\theta))} u_1 \leq V(s_1^*)$ provides the appropriate bound, with equality if and only if the game is belief-robust. The gap $V(s_1^*) - V_{\text{Markov}}$ quantifies the *cost of persistence in reputation games*—a new economic object measuring how state persistence enables short-run players to condition behavior on the revealed state. For the deterrence game with baseline parameters ($\alpha = 0.3$, $\beta = 0.5$), this cost is 23.7% of the i.i.d. payoff. Our framework interpolates continuously between i.i.d. (Luo–Wolitzky) and perfectly persistent (Pei 2020) states. The paper also documents the human–AI research process, including computational verification across seven analysis modules with eight diagnostic figures.

Keywords: Reputation, repeated games, Markov states, optimal transport, belief-robustness, corrected bounds

Original Paper: "Marginal Reputation" by Daniel Luo and Alexander Wolitzky, MIT Department of Economics, December 2024.

Contents

1 Introduction 4

1.1 The Challenge of Markov States 4

1.2 Computational Verification 4

1.3 Corrected Results 5

1.4 A Note on Process 5

1.5 Outline 5

2 The Extended Model 6

2.1 State Process 6

2.2 Lifted State Space 6

2.3 Stage Game 7

2.4 Joint Distribution and Marginals 8

2.5 Commitment Types 8

2.6 Repeated Game 8

3 Belief-Robustness: The Key New Concept 8

3.1 Filtering Beliefs 8

3.2 The Belief-Robustness Condition 9

3.3 When Does Belief-Robustness Hold? 9

4 Corrected Theorems 10

4.1 Definitions on the Expanded State Space 10

4.2 Theorem 1′ (Belief-Robust Extension) 10

4.3 Theorem 1″ (General Corrected Bound) 11

4.4 Extension to Behaviorally Confounded Strategies (Theorem 2) 12

5 Proof Sketch 12

5.1 Overview: Where i.i.d. Is Actually Used 12

5.2 Step 0: OT / Confound-Defeating Extension 13

5.3 Step 1: Lemma 1 — Equilibrium Implications 13

5.4 Step 2: Lemma 2 — KL Counting Bound 15

5.5 Step 3: Lemma 3 — Martingale Convergence 16

5.6 Step 4: Lemma 4 — Combining the Pieces 17

5.7 Step 5: The Payoff Bound 18

6 The Supermodular Case 19

6.1 Monotonicity on the Lifted Space 19

6.2 Payoffs Depending Only on θ_t 19

6.3 Transition-Dependent Payoffs 19

6.4 Extended Bounds . . . 20

7 Worked Example: Deterrence Game with Markov Attacks 20

7.1 Setup . . . 20
7.2 Lifted State Distribution . . . 21
7.3 Markov Deterrence Proposition . . . 21
7.4 Version 1: Belief-Robust ($\mu^* = 0.60$) . . . 22
7.5 Version 2: Non-Belief-Robust ($\mu^* = 0.60$) . . . 22
7.6 Concrete Numerical Example . . . 22
7.7 The Overestimation Gap . . . 23
7.8 Limiting Cases . . . 23
7.9 Comparison Table . . . 24
7.10 Figures . . . 24

8 Interpolation Between i.i.d. and Persistent 24

8.1 The Interpolation Landscape . . . 24
8.2 Recovery of i.i.d. (Luo–Wolitzky 2024) . . . 25
8.3 Connection to Pei (2020) — Perfect Persistence . . . 25
8.4 The Markov Interpolation . . . 25
8.5 The Cost of Persistence . . . 25
8.6 New Economic Content . . . 26

9 Methodology: Two Phases of Human–AI Collaboration 26

9.1 Phase 1: Initial Conjecture (Feb 16, 5:00–9:30 PM) . . . 27
9.2 Phase 2: Expert Critique (Feb 16, 10:00–11:00 PM) . . . 27
9.3 Phase 3: Review Planning and Agent Architecture (Feb 17, 12:00–1:00 AM) 28
9.4 Phase 4: Computational Testing (Feb 17, 1:00–2:00 AM) . . . 29
9.5 Phase 5: Manuscript Revision (Feb 17, 2:00–3:00 AM) . . . 30
9.6 Timeline . . . 30
9.7 Reflections on AI-Assisted Mathematical Research . . . 31

10 Discussion and Open Questions 32

10.1 Summary . . . 32
10.2 Open Questions . . . 32
10.3 Conclusion . . . 33

A KL Chain Rule Verification 33

A.1 The Chain Rule for KL Divergence . . . 33
A.2 Filter Stability for Ergodic HMMs . . . 34
A.3 Monte Carlo Verification . . . 34

B Computational Framework **34**

B.1 Analysis Modules 34

B.2 Reproducibility 35

B.3 Repository Structure 35

B.4 Additional Figures 36

References **36**

1 Introduction

Luo & Wolitzky (2024) establish a striking connection between reputation theory in repeated games and optimal transport theory. Their main result, Theorem 1, shows that a patient long-run player can secure her *commitment payoff* $V(s_1^*)$ in any Nash equilibrium, provided her Stackelberg strategy s_1^* is *confound-defeating* and *not behaviorally confounded*. Throughout their analysis, states are drawn **i.i.d. across periods**. The authors note (footnote 9) that the extension to persistent states is an open question.

1.1 The Challenge of Markov States

The extension from i.i.d. to Markov states introduces a fundamental new phenomenon. We employ a *lifted state* construction $\tilde{\theta}_t = (\theta_t, \theta_{t-1})$, which provides a stationary distribution $\tilde{\rho}$ on the expanded space and allows the optimal transport framework to apply directly. However, the extension is not a straightforward substitution: when the Stackelberg strategy reveals the state (e.g., $s_1^*(G) = A$, $s_1^*(B) = F$), the short-run player learns θ_t exactly, and their belief about θ_{t+1} becomes the *filtering distribution* $F(\cdot|\theta_t)$ rather than the stationary distribution π. This creates a permanent structural gap: short-run behavior becomes state-contingent, and the Nash correspondence $B(s_1^*)$ must be replaced by a state-dependent object $B(s_1^*, F(\cdot|\theta))$. Recognizing this—through a combination of expert feedback (Luo, 2026) and systematic computational verification—is the key insight of this paper.

1.2 Computational Verification

We conducted systematic computational analysis across seven diagnostic modules producing eight figures to characterize precisely which elements of the i.i.d. proof extend to the Markov setting and which require modification.

On the positive side, the KL counting bound extends verbatim to Markov processes, requiring no mixing-time correction; this was verified via $N = 500$ Monte Carlo simulations over $T = 5000$ periods. Filter stability holds with exponential forgetting, with fitted decay rate correlating at $r > 0.63$ with the chain's second eigenvalue $|1 - \alpha - \beta|$. The optimal transport support is robust to the belief perturbations that arise from Markov dynamics, with stability margin at least 0.3 in 100% of the (α, β) parameter space. Finally, the monotonicity characterization extends to the lifted state space for payoffs depending only on the current state θ_t.

On the negative side, short-run player beliefs permanently deviate from the stationary distribution, with mean total variation distance 0.412 and an analytical gap of 0.094 for the baseline parameters. The Nash correspondence $B(s_1^*, \mu)$ varies period-to-period, producing a 37.2% disagreement rate in short-run player actions between the stationary and filtered scenarios. The commitment payoff is consequently overestimated: 0.777 under the stationary assumption versus 0.628 under filtered beliefs, a gap of 23.7%.

1.3 Corrected Results

These findings guide two corrected theorems. Theorem 1′ addresses the *belief-robust* case: when the short-run player's best-response set $B(s_1^*, F(\cdot|\theta))$ is constant across states θ—a condition we call *belief-robustness*—the i.i.d. bound $V(s_1^*)$ holds exactly. The entire proof machinery (KL bound, OT robustness, monotonicity) applies without modification; belief-robustness ensures the filtering belief gap is irrelevant.

Theorem 1″ handles the general case. For all supermodular games with Markov states, a corrected bound holds:

$$V_{\text{Markov}}(s_1^*) := \sum_{\theta \in \Theta} \pi(\theta) \cdot \inf_{(\alpha_0, \alpha_2) \in B(s_1^*, F(\cdot|\theta))} u_1(\theta, s_1^*(\theta), \alpha_2) \;\leq\; V(s_1^*),$$

with equality if and only if the game is belief-robust. The gap $V(s_1^*) - V_{\text{Markov}}$ quantifies the "cost of persistence" in reputation games—a new economic object that the i.i.d. framework cannot capture.

1.4 A Note on Process

This paper documents not only the mathematical results but also the research process that produced them: AI-assisted conjecture, expert feedback from a co-author of the original paper (Luo, 2026), systematic computational verification via hierarchical AI agents, and iterative revision. Section 9 provides a detailed account, including the agent architecture, the computational testing framework, and the timeline from initial challenge to final paper. We include this methodological documentation because the process—rapid AI-assisted exploration refined through expert dialogue and computational evidence—may itself be of interest as a model for future human–AI collaboration in mathematical research.

1.5 Outline

Section 2 presents the model with the lifted state construction. Section 3 introduces the key new concept of belief-robustness. Section 4 states the two corrected theorems. Section 5 contains the proof sketch, tracing each step of the Luo–Wolitzky argument and identifying where corrections are needed. Section 6 extends the supermodular case. Section 7 works out the deterrence game in both belief-robust and non-belief-robust versions. Section 8 discusses the continuous interpolation between i.i.d. and persistent states. Section 9 gives a detailed account of the human–AI collaboration process across both phases. Section 10 discusses open questions. Appendix A verifies the KL chain rule and filter stability. Appendix B documents the reproducible computational framework.

2 The Extended Model

We maintain all notation and conventions from Luo & Wolitzky (2024, Sections 3.1–3.2), modifying only the state process.

2.1 State Process

Let Θ be a finite set.

Assumption 2.1 (Markov States)**.** The state $\theta_t \in \Theta$ follows a **stationary ergodic Markov chain** with:

(a) Transition kernel $F(\cdot|\theta)$ for each $\theta \in \Theta$, so that $\mathbb{P}(\theta_{t+1} = \theta'|\theta_t = \theta) = F(\theta'|\theta)$.

(b) Unique stationary distribution $\pi \in \Delta(\Theta)$ satisfying

$$\pi(\theta) = \sum_{\theta' \in \Theta} \pi(\theta')F(\theta|\theta') \quad \text{for all } \theta \in \Theta. \tag{1}$$

(c) The chain is **irreducible and aperiodic** (ensuring ergodicity).

Remark 2.2. When $F(\cdot|\theta) = \pi(\cdot)$ for all θ, the chain has no memory and we recover the i.i.d. case of Luo & Wolitzky (2024). The two-state case with $\Theta = \{G, B\}$ is parameterized by $\alpha = \mathbb{P}(B|G)$ and $\beta = \mathbb{P}(G|B)$, giving $\pi(G) = \beta/(\alpha + \beta)$.

2.2 Lifted State Space

The central construction is the *lifted state*:

Definition 2.3 (Lifted State)**.** Define

$$\tilde{\theta}_t \;=\; (\theta_t,\, \theta_{t-1}) \;\in\; \tilde{\Theta} \;=\; \Theta \times \Theta. \tag{2}$$

Remark 2.4 (Initial Period Convention)**.** The lifted state $\tilde{\theta}_t = (\theta_t, \theta_{t-1})$ is defined for $t \geq 1$. For $t = 0$, we draw θ_{-1} from the stationary distribution π independently of θ_0 (equivalently, we initialize the chain in stationarity at $t = -1$). This loses nothing: the first period is transient and vanishes under $\delta \to 1$. Alternatively, one may start the game at $t = 1$.

The process $(\tilde{\theta}_t)_{t\geq 1}$ is itself a Markov chain on $\tilde{\Theta}$ with transition probabilities

$$\tilde{F}\big((\theta', \theta) \,\big|\, (\theta, \theta'')\big) \;=\; F(\theta'|\theta) \tag{3}$$

and stationary distribution

$$\tilde{\rho}(\theta, \theta') \;=\; \pi(\theta') \cdot F(\theta|\theta'). \tag{4}$$

Proposition 2.5. *Under Assumption 2.1, the lifted chain $(\tilde{\theta}_t)$ on $\tilde{\Theta}$ is ergodic with unique stationary distribution $\tilde{\rho}$.*

Proof. Since the original chain is irreducible on Θ, for any states $\theta, \theta' \in \Theta$, there exists $n \in \mathbb{N}$ such that $F^n(\theta'|\theta) > 0$. Now consider two lifted states (θ_a, θ_b) and (θ_d, θ_c) in $\tilde{\Theta}$. By irreducibility of the original chain, there exists a finite path $\theta_a \to \theta_{i_1} \to \cdots \to \theta_{i_k} \to \theta_c$ with positive probability. This path in the original chain induces a path $(\theta_a, \theta_b) \to (\theta_{i_1}, \theta_a) \to \cdots \to (\theta_c, \theta_{i_k}) \to (\theta_d, \theta_c)$ in the lifted chain (where the final step uses $F(\theta_d|\theta_c) > 0$ for some path from θ_c to θ_d). Hence the lifted chain is irreducible. Aperiodicity follows from aperiodicity of the original chain: if $F(\theta|\theta) > 0$ for some θ, then $(\theta, \theta) \to (\theta, \theta)$ is a self-loop in the lifted chain. Uniqueness of $\tilde{\rho}$ follows from the Perron–Frobenius theorem. □

Remark 2.6 (Effective State Space)**.** If $F(\theta|\theta') = 0$ for some pair, the lifted state (θ, θ') is never visited. The effective state space is $\tilde{\Theta}_+ = \{(\theta, \theta') \in \Theta \times \Theta : F(\theta|\theta') > 0\} \subseteq \tilde{\Theta}$. All results hold on $\tilde{\Theta}_+$; we write $\tilde{\Theta}$ for notational simplicity throughout.

Remark 2.7 (Purpose of the Lifting)**.** The lifted state provides a Markov structure on which the optimal transport framework and cyclical monotonicity characterizations apply. The **key property** is that $\tilde{\theta}_t$ has a *fixed, known* stationary distribution $\tilde{\rho}$, playing precisely the role of the i.i.d. signal distribution ρ in Luo & Wolitzky (2024). This is the central insight enabling the extension.

2.3 Stage Game

The stage game is identical to Luo & Wolitzky's Section 3.1, except:

(i) The long-run player's private information each period is $\tilde{\theta}_t = (\theta_t, \theta_{t-1})$.

(ii) A stage-game strategy for player 1 is $s_1 : \tilde{\Theta} \to \Delta(A_1)$, a *Markov strategy.*

(iii) Payoffs depend on the current state: $u_1(\theta_t, a_1, \alpha_2)$.

We restrict throughout to payoffs $u_1(\theta_t, a_1, \alpha_2)$ that depend on θ_t alone (not the full lifted state $\tilde{\theta}_t$). This covers all standard applications—deterrence, trust, signaling—and avoids unmotivated generalization.

2.4 Joint Distribution and Marginals

Under Markov strategy s_1 and the stationary distribution $\tilde{\rho}$, the joint distribution over $(\tilde{\theta}, a_1)$ is:

$$\gamma(s_1)[\tilde{\theta}, a_1] \;=\; \tilde{\rho}(\tilde{\theta}) \cdot s_1(\tilde{\theta})[a_1]. \tag{5}$$

The marginals of $\gamma(s_1)$ are the stationary distribution $\pi_{\tilde{\Theta}}(\gamma) = \tilde{\rho}$, which is fixed and known, and the action marginal $\pi_{A_1}(\gamma) = \phi(s_1) = \sum_{\tilde{\theta}} \tilde{\rho}(\tilde{\theta})\, s_1(\tilde{\theta})[\cdot]$, which is observable to the short-run players. These two marginals play exactly the roles of the exogenous state distribution and the action frequency in the Luo–Wolitzky optimal transport formulation.

2.5 Commitment Types

A commitment type $\omega_{s_1} \in \Omega$ plays Markov strategy $s_1 : \tilde{\Theta} \to \Delta(A_1)$ every period. The type space Ω is countable with full-support prior $\mu_0 \in \Delta(\Omega)$.

Remark 2.8. A "memoryless" commitment type that plays $s_1 : \Theta \to \Delta(A_1)$ (ignoring θ_{t-1}) is a special case. The framework allows richer types that condition on transitions, but the memoryless case suffices for most applications.

2.6 Repeated Game

The repeated game structure is identical to Luo & Wolitzky's Section 3.2. Assumption 1 (signal y_1 identifies a_1) is maintained throughout.

3 Belief-Robustness: The Key New Concept

The central obstacle to extending Theorem 1 from i.i.d. to Markov states is the behavior of short-run player beliefs. When the Stackelberg strategy reveals the state, short-run players learn θ_t and form beliefs about θ_{t+1} using the filtering distribution $F(\cdot|\theta_t)$, which generically differs from the stationary distribution π. This section formalizes the issue and introduces the condition under which it can be resolved.

3.1 Filtering Beliefs

Definition 3.1 (Filtering Belief)**.** Given a state-revealing Stackelberg strategy s_1^* (i.e., $s_1^*(\theta) \neq s_1^*(\theta')$ for $\theta \neq \theta'$), the **filtering belief** in state θ is

$$F(\cdot|\theta_t) = \mathbb{P}(\theta_{t+1} = \cdot \mid \theta_t), \tag{6}$$

the one-step-ahead predictive distribution conditional on the current state.

For the two-state chain $\Theta = \{G, B\}$ with parameters (α, β), the filtering beliefs are $F(G|G) = 1-\alpha$ and $F(G|B) = \beta$, while the stationary distribution gives $\pi(G) = \beta/(\alpha+\beta)$. The expected gap between the filtering belief and the stationary distribution can be computed in closed form:

$$\mathbb{E}\big[|F(G|\theta_t) - \pi(G)|\big] = \frac{2\alpha\beta|1-\alpha-\beta|}{(\alpha+\beta)^2}. \tag{7}$$

This quantity equals zero **if and only if** $\alpha + \beta = 1$, which is precisely the i.i.d. case. Both the numerator factor $|1-\alpha-\beta|$ and the product $\alpha\beta$ must be nonzero for the gap to be positive, confirming that any departure from the i.i.d. regime produces a permanent structural discrepancy. For the baseline parameters $(\alpha = 0.3, \beta = 0.5)$, the expected gap is 0.094.

3.2 The Belief-Robustness Condition

Definition 3.2 (Belief-Robustness)**.** A game (u_1, u_2) with Stackelberg strategy s_1^* and Markov chain (Θ, F) is **belief-robust** if the short-run player Nash correspondence satisfies

$$B(s_1^*, F(\cdot|\theta)) = B(s_1^*, F(\cdot|\theta')) \quad \text{for all } \theta, \theta' \in \Theta. \tag{8}$$

The condition requires that the short-run player's best-response set is invariant to the revealed state. Under belief-robustness, the filtering belief gap documented in (7) becomes irrelevant for equilibrium behavior: SR plays the same action regardless of whether their belief about the next state is $F(\cdot|G)$ or $F(\cdot|B)$.

3.3 When Does Belief-Robustness Hold?

For the deterrence game with SR threshold μ^* (the belief level at which SR is indifferent between cooperating and defecting), belief-robustness admits a clean characterization.

Proposition 3.3. *Belief-robustness holds if and only if*

$$\mu^* \notin \big[\min_\theta F(G|\theta),\ \max_\theta F(G|\theta)\big] = [\beta,\ 1-\alpha]. \tag{9}$$

Proof. The SR best response depends on whether $F(G|\theta_t) \gtrless \mu^*$. If $\mu^* < \beta$, then $F(G|\theta_t) \geq \beta > \mu^*$ for all θ_t, so SR always cooperates. If $\mu^* > 1-\alpha$, then $F(G|\theta_t) \leq 1-\alpha < \mu^*$ for all θ_t, so SR always defects. In either case, $B(s_1^*, F(\cdot|\theta))$ is constant across states. Conversely, if $\mu^* \in [\beta, 1-\alpha]$, there exist states θ, θ' with $F(G|\theta) > \mu^* > F(G|\theta')$, so SR cooperates after θ and defects after θ', and belief-robustness fails. □

The economic interpretation is that belief-robustness fails precisely when the SR indifference threshold lies in the "danger zone" $[\beta, 1-\alpha]$—the interval spanned by the conditional beliefs across states. Three factors conspire to produce this failure: the game must have belief-sensitive SR behavior, with the threshold near π; the chain must be persistent enough that $F(\cdot|\theta)$ varies substantially across states; and the Stackelberg strategy must reveal state information to SR. When all three conditions hold simultaneously, persistence harms the long-run player's reputation value.

Remark 3.4 (Baseline Example)**.** For the baseline parameters ($\alpha = 0.3, \beta = 0.5$), the danger zone is $[0.5, 1-0.3] = [0.5, 0.7]$. The SR threshold $\mu^* = 0.60$ lies inside this interval, so the baseline deterrence example is **not** belief-robust. However, changing SR payoffs to produce $\mu^* = 0.60 < \beta = 0.5$ would place the threshold below the danger zone, restoring belief-robustness.

4 Corrected Theorems

We state two results. Theorem $1'$ recovers the exact i.i.d. bound under belief-robustness. Theorem $1''$ provides a corrected bound for the general case.

4.1 Definitions on the Expanded State Space

All definitions from Luo & Wolitzky (2024) carry over to $\tilde{\Theta}$, with strategies mapping $\tilde{\Theta} \to \Delta(A_1)$.

Definition 4.1 (Confound-Defeating, Extended)**.** A Markov strategy $s_1^* : \tilde{\Theta} \to \Delta(A_1)$ is **confound-defeating** if for every $(\alpha_0, \alpha_2) \in B_0(s_1^*)$, the joint distribution $\gamma(\alpha_0, s_1^*)$ is the *unique solution* to:

$$\mathrm{OT}\big(\tilde{\rho}(\alpha_0),\, \phi(\alpha_0, s_1^*);\, \alpha_2\big) : \quad \max_{\gamma \in \Delta(\tilde{\Theta} \times A_1)} \int u_1(\tilde{\theta}, a_1, \alpha_2)\, d\gamma \tag{10}$$

subject to $\pi_{\tilde{\Theta}}(\gamma) = \tilde{\rho}(\alpha_0)$ and $\pi_{A_1}(\gamma) = \phi(\alpha_0, s_1^*)$.

Definition 4.2 (Not Behaviorally Confounded, Extended)**.** s_1^* is **not behaviorally confounded** if for any $\omega_{s_1'} \in \Omega$ with $s_1' \neq s_1^*$ and any $(\alpha_0, \alpha_2) \in B_1(s_1^*)$, we have $p(\alpha_0, s_1^*, \alpha_2) \neq p(\alpha_0, s_1', \alpha_2)$.

4.2 Theorem $1'$ (Belief-Robust Extension)

Theorem 4.3 (Belief-Robust Markov Extension)**.** *Let θ_t follow a stationary ergodic Markov chain on finite Θ (Assumption 2.1). Let $\tilde{\theta}_t = (\theta_t, \theta_{t-1})$ with stationary distribution $\tilde{\rho}$. Suppose:*

(i) $\omega_{s_1^*} \in \Omega$, where $s_1^* : \tilde{\Theta} \to \Delta(A_1)$ is a Markov strategy;

(ii) s_1^* is confound-defeating on $\tilde{\Theta}$ (Definition 4.1);

(iii) s_1^* is not behaviorally confounded (Definition 4.2);

*(iv) The game is **belief-robust** with respect to s_1^* and (Θ, F) (Definition 3.2).*

Then:

$$\boxed{\liminf_{\delta \to 1} \underline{U}_1(\delta) \ \geq \ V(s_1^*)} \tag{11}$$

where $V(s_1^*) = \inf_{(\alpha_0,\alpha_2)\in B(s_1^*)} u_1(\alpha_0, s_1^*, \alpha_2)$ is the commitment payoff, identical to the i.i.d. case.

Remark 4.4. Under belief-robustness, the SR belief gap is irrelevant: SR plays the same best response regardless of the filtering belief $F(\cdot|\theta)$. All the confirmed proof machinery—KL counting bound, OT robustness, monotonicity—applies without modification.

4.3 Theorem 1″ (General Corrected Bound)

Definition 4.5 (Markov Commitment Payoff)**.** The **Markov commitment payoff** is

$$V_{\text{Markov}}(s_1^*) := \sum_{\theta\in\Theta} \pi(\theta) \cdot \inf_{(\alpha_0,\alpha_2)\in B(s_1^*, F(\cdot|\theta))} u_1(\theta, s_1^*(\theta), \alpha_2). \tag{12}$$

This averages over states using the stationary distribution π, but uses the **state-contingent** Nash correspondence $B(s_1^*, F(\cdot|\theta))$ at each state.

Theorem 4.6 (General Markov Extension)**.** *Under conditions (i)–(iii) of Theorem 4.3, with ergodic Markov states and confound-defeating s_1^* on $\tilde{\Theta}$:*

$$\boxed{\liminf_{\delta \to 1} \underline{U}_1(\delta) \ \geq \ V_{\text{Markov}}(s_1^*)} \tag{13}$$

where $V_{\text{Markov}}(s_1^*) \leq V(s_1^*)$, with equality if and only if the game is belief-robust.

Remark 4.7 (Relationship Between Theorems)**.** The two results are nested: Theorem 4.3 is the special case of Theorem 4.6 where belief-robustness forces $V_{\text{Markov}} = V(s_1^*)$. The gap $V(s_1^*) - V_{\text{Markov}}$ is the "cost of persistence"—the payoff the LR player loses because state persistence causes SR to adjust behavior state-by-state.

Remark 4.8 (Continuity in Chain Parameters)**.** $V_{\text{Markov}}(s_1^*)$ is a continuous function of the chain parameters (α, β). As $\alpha + \beta \to 1$ (the i.i.d. limit), $F(\cdot|\theta) \to \pi(\cdot)$ for all θ, so $V_{\text{Markov}} \to V(s_1^*)$. The gap vanishes continuously.

4.4 Extension to Behaviorally Confounded Strategies (Theorem 2)

The salience-based extension (Luo & Wolitzky, 2024, Appendix A, Theorem 2) also generalizes to the Markov setting.

Theorem 4.9 (Extended Theorem 2)**.** *Under the same Markov setup, if s_1^* is confound-defeating on $\tilde{\Theta}$ and has salience β_s (defined identically to Luo & Wolitzky, 2024, but with confounding weights computed on $\tilde{\Theta}$), then:*

$$\liminf_{\delta \to 1} \underline{U}_1(\delta) \;\geq\; \beta_s \, V_{\text{Markov}}(s_1^*) + (1 - \beta_s)\, V_0(s_1^*). \tag{14}$$

Under belief-robustness, V_{Markov} is replaced by $V(s_1^*)$. If s_1^* is not behaviorally confounded, $\beta_s = 1$ and this reduces to Theorems 4.3 or 4.6 respectively.

Proof sketch. The proof of Theorem 2 in Luo & Wolitzky (2024) follows from Theorem 1 via Lemma 7 (the salience bound). Lemma 7 uses the submartingale property of $\mu_t(\omega_{s_1^*} | \Omega_\eta(s_1^*) \setminus \{\omega^R\}, h_t)$, which holds by Bayesian updating regardless of the signal process. The remainder of the argument—compactness, limiting, the three-case analysis—extends as in Section 5. The only modification is the payoff bound: under belief-robustness, the full $V(s_1^*)$ is used; in the general case, $V_{\text{Markov}}(s_1^*)$ replaces $V(s_1^*)$. □

5 Proof Sketch

The proof follows the five-step structure of Luo & Wolitzky (2024, Section 4.2). At each step, we identify whether the i.i.d. assumption was used and, if so, how belief-robustness or the corrected bound handles the modification.

5.1 Overview: Where i.i.d. Is Actually Used

Proof Step	**i.i.d. used?**	**Fix needed**
Step 0: OT / confound-defeating	No	Replace Y_0 with $\tilde{\Theta}$
Step 1: Lemma 1 (equilibrium)	**Yes**	**SR belief issue**
Step 2: Lemma 2 (KL bound)	**No**	None
Step 3: Lemma 3 (martingale)	Partially	Ergodicity + filter stability
Step 4: Lemma 4 (combining)	No	Uses corrected BR
Step 5: Payoff bound	**Yes**	**Belief-robust or V_{Markov}**

Table 1: Where the i.i.d. assumption enters the proof. Bold rows indicate where the Luo–Wolitzky argument fails under Markov states and correction is needed.

The table reveals a noteworthy pattern: the purely information-theoretic steps (the KL bound and the martingale convergence) require no modification or only mild conditions, while the game-theoretic steps (the equilibrium implications and the payoff bound) are where the i.i.d. assumption does essential work. This reflects the distinction between the *mathematical tools*, which are process-independent, and their *semantic interpretation* within the reputation game, which depends on the information structure.

5.2 Step 0: OT / Confound-Defeating Extension

The state space is $\tilde{\Theta} = \Theta \times \Theta$ instead of Y_0, but the entire optimal transport framework carries over without change. The OT problem $\mathrm{OT}(\tilde{\rho}, \phi; \alpha_2)$ on $\tilde{\Theta} \times A_1$ is a finite-dimensional linear program, structurally identical to the Luo–Wolitzky formulation on $Y_0 \times A_1$.

Proposition 5.1 (Extension of Proposition 5)**.** *A joint distribution $\gamma \in \Delta(\tilde{\Theta} \times A_1)$ with marginals $\tilde{\rho}$ and ϕ uniquely solves* $\mathrm{OT}(\tilde{\rho}, \phi; \alpha_2)$ *if and only if* $\mathrm{supp}(\gamma) \subset \tilde{\Theta} \times A_1$ *is* ***strictly $u_1(\cdot, \alpha_2)$-cyclically monotone****.*

Proof. This is Proposition 5 of Luo–Wolitzky applied to $X = \tilde{\Theta}$ and $Y = A_1$. The proof (Luo & Wolitzky, 2024, Appendix C) is a purely combinatorial argument about finite optimal transport problems and does not depend on the time-series structure of the data. The argument uses only: (a) finiteness of $\tilde{\Theta} \times A_1$ (which holds since Θ is finite), and (b) the characterization of OT solutions via cyclical monotonicity (Rochet 1987; Santambrogio 2015). Both hold on the expanded state space. □

Corollary 5.2 (Extension of Corollary 1)**.** *s_1^* is confound-defeating if and only if* $\mathrm{supp}(s_1^*) \subset \tilde{\Theta} \times A_1$ *is strictly u_1-cyclically monotone (when u_1 is cyclically separable) or strictly $u_1(\cdot, \alpha_2)$-cyclically monotone for all $(\alpha_0, \alpha_2) \in B_0(s_1^*)$ (in general).*

Computational evidence confirms this robustness: the OT support stability margin exceeds 0.3 in 100% of the (α, β) parameter space (Figure 8), demonstrating that the confound-defeating property is preserved under the belief perturbations that arise from Markov dynamics.

5.3 Step 1: Lemma 1 — Equilibrium Implications

Lemma 5.3 (Extension of Lemma 1)**.** *Fix a Nash equilibrium $(\sigma_0^*, \sigma_1^*, \sigma_2^*)$. For any $\varepsilon > 0$, there exists $\eta > 0$ such that if:*

(1) $\|p(\sigma_0^*, s_1^*, \sigma_2^* | h_t) - p(\sigma_0^*, \sigma_1^*, \sigma_2^* | h_t)\| \leq \eta$*, and*

(2) $\|p(\sigma_0^*, \sigma_1^*(\omega^R), \sigma_2^* | h_t) - p(\sigma_0^*, s_1^*, \sigma_2^* | h_t)\| \leq \eta$*,*

then $\|\sigma_1^*(h_t, \omega^R) - s_1^*\| \leq \varepsilon$*.*

Proof. The argument is a per-period one-shot deviation analysis that uses confound-defeatingness and the equilibrium condition. Suppose $\|\sigma_1^*(h_t, \omega^R) - s_1^*\| > \varepsilon$. Condition (1) and the Nash equilibrium condition imply $(\sigma_0^*(h_t), \sigma_2^*(h_t)) \in B_\eta(s_1^*)$, as $\sigma_1^*(h_t)$ η-confirms it against s_1^*. Then condition (2), combined with $\|\sigma_1^*(h_t, \omega^R) - s_1^*\| > \varepsilon$ and the confound-defeating property, implies there exists $\tilde{s}_1$ such that:

$$p(\sigma_0^*, \tilde{s}_1, \sigma_2^*|h_t) = p(\sigma_0^*, \sigma_1^*(\omega^R), \sigma_2^*|h_t) \quad \text{and} \quad u_1(\sigma_0^*, \tilde{s}_1, \sigma_2^*|h_t) > u_1(\sigma_0^*, \sigma_1^*(\omega^R), \sigma_2^*|h_t).$$

Deviating from $\sigma_1^*(h_t, \omega^R)$ to $\tilde{s}_1$ is then a profitable one-shot deviation that is signal-preserving, contradicting the equilibrium assumption. The strategy space is now Markov strategies on $\tilde{\Theta}$ instead of static strategies on Y_0, but the one-shot deviation argument is identical. □

Where i.i.d. matters for Step 1. This is the first point where the i.i.d. assumption enters the Luo–Wolitzky proof substantively. In the i.i.d. case, the one-shot deviation objective takes the form $u_1(\theta, a_1, \alpha_2) + \delta V_{\text{cont}}^{a_1}$, where the continuation value $V_{\text{cont}}^{a_1}$ depends only on a_1 because future states are independent of the current state θ. Adding a function of a_1 alone to the objective does not change the optimal transport solution, so confound-defeatingness with respect to u_1 suffices.

In the Markov case, the continuation value $V_{\text{cont}}(\theta_t, a_1, h_t)$ depends on θ_t through the transition kernel F. The effective one-shot deviation objective becomes $w(\tilde{\theta}, a_1) = u_1(\tilde{\theta}, a_1, \alpha_2) + \delta g(\theta_t, a_1, h_t)$ for some history-dependent function g, and adding this θ_t-dependent term can in principle change the OT solution.

Remark 5.4 (Continuation Value Subtlety)**. Resolution for the belief-robust case.** Under strict supermodularity of u_1 in $(\tilde{\theta}, a_1)$, the co-monotone coupling is optimal for all objectives of the form $u_1 + g$ provided g preserves the supermodular structure. Since $g(\theta_t, a_1, h_t)$ is supermodular in (θ_t, a_1) whenever V_{cont} is increasing in θ_t for each a_1 (which holds when higher states have higher continuation values), the OT solution is unchanged. All the paper's applications (deterrence, trust, signaling) are supermodular, and under belief-robustness the SR behavior is constant across states, so the continuation value perturbation is absorbed by the supermodular structure and the OT solution remains unchanged.

Resolution for the general case. Two approaches are available. First, one may strengthen the confound-defeating condition: require s_1^* to be confound-defeating for all objectives of the form $u_1 + g$ where $g : \tilde{\Theta} \times A_1 \to \mathbb{R}$ is bounded. This is stronger than the Luo–Wolitzky condition but closes the gap. Second, a continuity argument is available: by filter stability (Proposition A.2), the filtering distribution $\pi_t(h_t)$ converges to the stationary distribution $\tilde{\rho}$ exponentially fast. Since confound-defeating is an open condition (unique OT solution is robust to small perturbations of the marginals), confound-defeating at $\tilde{\rho}$ implies approximate confound-defeating at $\pi_t(h_t)$ for large t. Combined with $\delta \to 1$ (which makes the continuation value perturbation small relative to the stage-game payoff), this yields the result. The general case is handled by Theorem 4.6, which replaces the static Nash correspondence $B(s_1^*, \pi)$ with the state-contingent correspondence $B(s_1^*, F(\cdot|\theta_t))$.

5.4 Step 2: Lemma 2 — KL Counting Bound

This is the key technical step where one might expect the i.i.d. assumption to be essential. It is not.

Lemma 5.5 (Extension of Lemma 2)**.** *For any $\eta > 0$ and any Nash equilibrium $(\sigma_0^*, \sigma_1^*, \sigma_2^*)$, the expected number of periods t where $h_t \notin H_t^\eta$ is bounded by:*

$$\mathbb{E}_Q[\#\{t : h_t \notin H_t^\eta\}] \;\le\; \bar{T}(\eta, \mu_0) \;:=\; \frac{-2\log \mu_0(\omega_{s_1^*})}{\eta^2}. \tag{15}$$

The bound is identical to the i.i.d. case.

Proof. The argument uses three ingredients, *none of which require i.i.d.*

(a) Chain rule for KL divergence. For any joint distribution over $(y_0, y_1, \ldots, y_{T-1})$:

$$D_{\mathrm{KL}}(P^T \| Q^T) \;=\; \sum_{t=0}^{T-1} \mathbb{E}_P\left[D_{\mathrm{KL}}\left(P_{y_t|h_{t-1}} \;\|\; Q_{y_t|h_{t-1}}\right)\right]. \tag{16}$$

This is a general property of KL divergence that holds for *arbitrary* joint distributions, including those generated by Markov chains. It is a consequence of the chain rule for KL divergence (Cover & Thomas, 2006, Theorem 2.5.3), which states:

$$D_{\mathrm{KL}}(P(X_1,\ldots,X_n) \| Q(X_1,\ldots,X_n)) = \sum_{i=1}^{n} \mathbb{E}_P[D_{\mathrm{KL}}(P(X_i|X_1,\ldots,X_{i-1}) \| Q(X_i|X_1,\ldots,X_{i-1}))].$$

No independence across periods is assumed. A self-contained verification is provided in Appendix A.

(b) Total KL bound from Bayesian updating. The Bayesian updating identity gives:

$$\sum_{t=0}^{T-1} \mathbb{E}_Q[D_{\mathrm{KL}}(p_t \| q_t)] \;\le\; -\log \mu_0(\omega_{s_1^*}) \tag{17}$$

where $p_t = p(\sigma_0^*, s_1^*, \sigma_2^*|h_t)$ and $q_t = p(\sigma_0^*, \sigma_1^*, \sigma_2^*|h_t)$. This follows from $\mu_T(\omega_{s_1^*}) \le 1$ and the telescoping identity:

$$\log \frac{\mu_T(\omega_{s_1^*})}{\mu_0(\omega_{s_1^*})} = \sum_{t=0}^{T-1} \log \frac{p_t(y_{1,t})}{q_t(y_{1,t})} = \sum_{t=0}^{T-1} \log \frac{p(\sigma_0^*, s_1^*, \sigma_2^*|h_t)[y_{1,t}]}{p(\sigma_0^*, \sigma_1^*, \sigma_2^*|h_t)[y_{1,t}]}.$$

Taking expectations under Q and using $\mathbb{E}_Q[\log(p_t/q_t)] = D_{\mathrm{KL}}(p_t \| q_t)$ gives (17). This is a consequence of Bayes' rule alone. **No independence across periods is used.**

(c) Pinsker's inequality (per-period). For each period t:

$$\|p_t - q_t\|^2 \;\leq\; 2\, D_{\mathrm{KL}}(p_t \| q_t). \tag{18}$$

This is a per-period inequality requiring no temporal structure.

Combining: In each "distinguishing period" where $\|p_t - q_t\| > \eta$, Pinsker gives $D_{\mathrm{KL}}(p_t \| q_t) \geq \eta^2/2$. Summing:

$$\frac{\eta^2}{2} \cdot \#\{\text{distinguishing periods}\} \;\leq\; \sum_t D_{\mathrm{KL}}(p_t \| q_t) \;\leq\; -\log \mu_0(\omega_{s_1^*}).$$

Hence $\#\{\text{distinguishing periods}\} \leq -2 \log \mu_0(\omega_{s_1^*})/\eta^2 = \bar{T}(\eta, \mu_0)$. □

Remark 5.6. This is the key surprise of the extension. The initial conjecture was that a mixing-time correction factor τ_{mix} would be needed. It is not: the KL chain rule and Bayesian updating identity hold for general stochastic processes. Monte Carlo verification ($N = 500$ simulations, $T = 5000$ periods) confirms that the empirical distribution of distinguishing-period counts is nearly identical for Markov and i.i.d. processes (Figure 6).

5.5 Step 3: Lemma 3 — Martingale Convergence

Lemma 5.7 (Extension of Lemma 3). *For all $\zeta > 0$, there exists a set of infinite histories $G(\zeta) \subset H^\infty$ satisfying $Q(G(\zeta)) > 1 - \zeta$ and a period $\hat{T}(\zeta)$ (independent of δ and the choice of equilibrium) such that, for any $h \in G(\zeta)$ and any $t \geq \hat{T}(\zeta)$:*

$$\mu_t(\cdot|h) \in M_\zeta \;:=\; \big\{\mu \in \Delta(\Omega) : \mu(\{\omega^R, \omega_{s_1^*}\}) \geq 1 - \zeta\big\}.$$

Proof sketch. The proof has two parts.

Part A: Per-equilibrium convergence (Extension of Lemma 9).

The posterior $\mu_t(\omega|h)$ over Ω is a bounded martingale under Q (the measure induced by commitment type $\omega_{s_1^*}$). This is a consequence of Bayesian updating and holds regardless of the signal structure. By the **martingale convergence theorem**, $\mu_t(\omega|h) \to \mu_\infty(\omega|h)$ Q-a.s. for each ω.

We need to show $\mu_\infty(\{\omega^R, \omega_{s_1^*}\}|h) = 1$ Q-a.s. The critical step is that for any ω_{s_1} with $\mu_\infty(\omega_{s_1}|h) > 0$, the signal distributions under s_1 and s_1^* must agree asymptotically. In the i.i.d. case, this follows immediately from the KL bound. In the Markov case, we proceed as follows. First, the per-period signal distribution under commitment type ω_{s_1} depends on the *filtering distribution* $\pi(\theta_t|h_t, s_1)$—the posterior over the current state given public signals. Second, for an **ergodic** Markov chain, the filtering distribution satisfies *filter stability* (also known as filter forgetting): regardless of the initial condition, the posterior $\pi(\theta_t|h_t, s_1)$ eventually concentrates on values determined by the observation process, and the effect of the initial condition decays exponentially. This is a classical result for HMMs on finite state spaces (Chigansky & Liptser, 2004; Del Moral, 2004). Third, the KL bound from Lemma 5.5 (which holds unchanged) implies:

$$\lim_{t\to\infty} \left\| p_{Y_1}(\sigma_0^*, s_1|h_t) - p_{Y_1}(\sigma_0^*, \tilde{s}_1|h_t, \Omega \setminus \{\omega^R\}) \right\| = 0 \tag{19}$$

Q-a.s., exactly as in the Luo–Wolitzky proof of Lemma 9 (their Appendix B.2). The KL chain rule argument that yields this convergence is valid for arbitrary signal processes. Since s_1^* is not behaviorally confounded, any type with the same asymptotic signal distribution must be s_1^* itself, hence $\mu_\infty(\{\omega^R, \omega_{s_1^*}\}|h) = 1$.

Computational evidence across a 30 × 30 parameter grid confirms that the fitted forgetting rate λ correlates with the chain's second eigenvalue $|1 - \alpha - \beta|$ at $r > 0.63$, with exponential decay fits achieving $R^2 > 0.99$ throughout (Figure 7).

Part B: Uniformity over equilibria.

The uniformity argument ($\hat{T}$ independent of δ and the equilibrium) uses **compactness** of $B_1(s_1^*)^{H^\infty}$ under the sup-norm topology, **Egorov's theorem** (a general measure-theoretic result), and **continuity** of finite-dimensional distributions Q^T as strategies vary. With Markov states, the space of Markov strategies $s_1 : \tilde{\Theta} \to \Delta(A_1)$ is compact ($\tilde{\Theta}$ is finite, $\Delta(A_1)$ is compact). The compactness of $B_1(s_1^*)^{H^\infty}$ follows by the same product topology argument. Egorov's theorem is a general result requiring only a finite measure space. The continuity of Q^T in strategies uses finiteness and continuity of the signal structure, which holds with Markov states.

The proof of uniformity then follows the original argument in Appendix B.2 of Luo–Wolitzky: suppose for contradiction that $\hat{T}$ cannot be chosen uniformly; extract a convergent subsequence using compactness; apply Egorov's theorem to obtain a contradiction with Q-a.s. convergence from Part A. □

5.6 Step 4: Lemma 4 — Combining the Pieces

Lemma 5.8 (Extension of Lemma 4)**.** *There exist strictly positive functions $\zeta(\eta)$ and $\xi(\eta)$, satisfying $\lim_{\eta\to 0} \zeta(\eta) = \lim_{\eta\to 0} \xi(\eta) = 0$, such that if $h_t \in H_t^\eta$ and $\mu_t(\cdot|h_t) \in M_{\zeta(\eta)}$, then:*

$$(\sigma_0^*(h_t),\, \sigma_2^*(h_t)) \;\in\; \hat{B}_{\xi(\eta)}(s_1^*).$$

Proof. This is a per-period argument combining Lemma 5.3 with the definition of M_ζ and the confirmed best response structure. It uses only the stage-game structure and the proximity of the posterior to $\{\omega^R, \omega_{s_1^*}\}$. If $h_t \in H_t^\eta$, then $(\sigma_0^*(h_t), \sigma_2^*(h_t)) \in B_\eta(s_1^*)$; and if additionally $\mu_t(\cdot|h_t) \in M_{\zeta(\eta)}$, then the posterior concentrates on $\{\omega^R, \omega_{s_1^*}\}$, from which it follows (via Lemma 5.3 and continuity) that $(\sigma_0^*(h_t), \sigma_2^*(h_t)) \in \hat{B}_{\xi(\eta)}(s_1^*)$ for appropriate $\xi(\eta)$. No independence across periods is used: the argument is identical to that of Luo & Wolitzky (2024). □

5.7 Step 5: The Payoff Bound

This is the second place where the i.i.d. assumption enters the Luo–Wolitzky proof substantively, and where the two theorems diverge.

Proof of Theorems 4.3 and 4.6. Fix $\varepsilon > 0$. Choose η small enough so that:

$$\inf_{(\alpha_0,\alpha_2)\in \hat{B}_{\xi(\eta)}(s_1^*)} u_1(\alpha_0, s_1^*, \alpha_2) \;\geq\; V(s_1^*) - \frac{\varepsilon}{3}.$$

On the $(1-\zeta(\eta))$-probability event $G(\zeta(\eta))$, for $t \geq \hat{T}(\zeta(\eta))$:

(i) The expected number of periods where $h_t \notin H_t^\eta$ is at most $\bar{T}(\eta, \mu_0)$ (Lemma 5.5).

(ii) $\mu_t(\cdot|h_t) \in M_{\zeta(\eta)}$ (Lemma 5.7).

(iii) In "good" periods (where both conditions hold), $(\sigma_0^*(h_t), \sigma_2^*(h_t)) \in \hat{B}_{\xi(\eta)}(s_1^*)$ (Lemma 5.8).

Front-loading the bad periods and using the discount factor:

$$U_1(\delta) \;\geq\; (1-\delta^{\bar{T}+\hat{T}}) \cdot \underline{u}_1 \;+\; \delta^{\bar{T}+\hat{T}} \cdot \left(V(s_1^*) - \frac{\varepsilon}{3}\right). \tag{20}$$

As $\delta \to 1$:

$$\liminf_{\delta\to 1} \underline{U}_1(\delta) \;\geq\; V(s_1^*) - \frac{\varepsilon}{3}. \tag{21}$$

Taking $\varepsilon \to 0$ gives $\liminf_{\delta\to 1} \underline{U}_1(\delta) \geq V(s_1^*)$.

Belief-robust case (Theorem 4.3). Under belief-robustness, the SR best response is constant across states, so the LR player receives at least $\inf_{B(s_1^*)} u_1 = V(s_1^*)$ in every good period. The argument above applies verbatim and yields $\liminf_{\delta\to 1} \underline{U}_1(\delta) \geq V(s_1^*)$.

General case (Theorem 4.6). Without belief-robustness, the SR player's best response in state θ_t is drawn from the state-contingent correspondence $B(s_1^*, F(\cdot|\theta_t))$. The LR player receives at least $\inf_{B(s_1^*,F(\cdot|\theta_t))} u_1$ in state θ_t during good periods. Averaging over the ergodic distribution of states and applying the same front-loading argument gives $\liminf_{\delta\to 1} \underline{U}_1(\delta) \geq V_{\text{Markov}}(s_1^*)$. □

Remark 5.9 (Role of Mixing Time)**.** The mixing time τ_{mix} does *not* enter either payoff bound. It affects only the **rate of convergence**—specifically, the constant $\hat{T}(\zeta)$ in Lemma 5.7, which may be larger for slowly mixing chains. The limit as $\delta \to 1$ is unaffected.

6 The Supermodular Case

6.1 Monotonicity on the Lifted Space

Proposition 6.1 (Extension of Proposition 7)**.** *Suppose u_1 is* ***strictly supermodular*** *in $(\tilde{\theta}, a_1)$ for some orders $\succeq_{\tilde{\Theta}}$ on $\tilde{\Theta}$ and $\succeq_{A_1}$ on A_1, for all α_2. Then the following are equivalent:*

(1) s_1^* is confound-defeating.

(2) s_1^* is monotone: if $\tilde{\theta} \succ \tilde{\theta}'$, $a_1 \in \operatorname{supp}(s_1^*(\tilde{\theta}))$, $a_1' \in \operatorname{supp}(s_1^*(\tilde{\theta}'))$, then $a_1 \succeq a_1'$.

(3) For any (α_0, α_2), $\gamma(\alpha_0, s_1^*)$ is the ***co-monotone coupling*** *of $\tilde{\rho}(\alpha_0)$ and $\phi(\alpha_0, s_1^*)$.*

Proof. The equivalence (1) ⇔ (3) follows from Lemma 6 of Luo & Wolitzky (2024) applied to $\tilde{\Theta} \times A_1$: under strict supermodularity, the co-monotone coupling is the unique solution to the OT problem (Santambrogio, 2015, Lemma 2.8). The equivalence (2) ⇔ (3) follows from the definition of monotonicity and co-monotone coupling. The proof is a purely combinatorial argument on the expanded state space and does not reference the temporal structure of the signal process. □

6.2 Payoffs Depending Only on θ_t

If $u_1(\tilde{\theta}, a_1, \alpha_2) = u_1(\theta_t, a_1, \alpha_2)$, then u_1 is supermodular in $(\tilde{\theta}, a_1)$ if and only if it is supermodular in (θ_t, a_1), using any order on $\tilde{\Theta}$ that is consistent with the order on the first coordinate (e.g., the lexicographic order). The relevant order on $\tilde{\Theta}$ is the *first-coordinate order*: $(\theta_t, \theta_{t-1}) \succeq (\theta_t', \theta_{t-1}')$ if and only if $\theta_t \succeq \theta_t'$. Under this order, the supermodularity condition is **unchanged** from the i.i.d. case: it depends only on the payoff structure in (θ_t, a_1), not on the Markov dynamics.

Computational evidence confirms this: for θ_t-dependent payoffs, 4 out of 24 orderings of the lifted space $\tilde{\Theta}$ preserve supermodularity—exactly those consistent with the first-coordinate ranking (Figure 1).

6.3 Transition-Dependent Payoffs

When payoffs depend on the full lifted state (θ_t, θ_{t-1})—e.g., escalation penalties that depend on whether the state deteriorated—the ordering problem becomes harder. Only a small fraction of orderings on $\tilde{\Theta}$ preserve supermodularity in general. This is a genuine limitation of the Markov extension for non-standard payoff structures.

6.4 Extended Bounds

Corollary 6.2 (Extended Lower Bound)**.** *Under the conditions of Proposition 6.1 with θ_t-only payoffs:*

$$\liminf_{\delta\to 1} \underline{U}_1(\delta) \;\geq\; v_{\text{mon}} \;:=\; \sup\left\{V_{\text{Markov}}(s_1) : s_1 \textit{ monotone on } \tilde{\Theta},\; \omega_{s_1} \in \Omega\right\}. \tag{22}$$

Under belief-robustness, $V_{\text{Markov}}(s_1)$ can be replaced by $V(s_1)$.

Corollary 6.3 (Extended Upper Bound)**.** *If u_1 is cyclically separable and $\mu_0(\omega^R) \to 1$, then:*

$$\bar{U}_1(\delta) \;<\; \bar{v}_1^{CM} + \varepsilon \tag{23}$$

where $\bar{v}_1^{CM}$ is the supremum over u_1-cyclically monotone strategies on $\tilde{\Theta}$.

Proof. The upper bound follows from the extension of Lemma 5: in any equilibrium, $\sigma_1^*(h_t, \omega^R)$ must solve $\text{OT}(\sigma_0^*(h_t), \phi(\sigma_0^*(h_t), \sigma_1^*(h_t, \omega^R)), \sigma_2^*(h_t))$, hence is u_1-cyclically monotone. This is a per-period optimality condition and does not use the i.i.d. assumption. □

7 Worked Example: Deterrence Game with Markov Attacks

We illustrate both theorems using the deterrence game with Markov attacks. We present the full game setup, a formal proposition establishing when the extended theorem applies, concrete numerical calculations, and both a belief-robust and a non-belief-robust version.

7.1 Setup

The state $\theta_t \in \{G(\text{ood}), B(\text{ad})\}$ follows a Markov chain:

$$\mathbb{P}(G|G) = 1 - \alpha, \qquad \mathbb{P}(B|G) = \alpha, \tag{24}$$

$$\mathbb{P}(G|B) = \beta, \qquad \mathbb{P}(B|B) = 1 - \beta, \tag{25}$$

with $\alpha, \beta \in (0, 1)$. The unique stationary distribution is:

$$\pi(G) = \frac{\beta}{\alpha + \beta}, \qquad \pi(B) = \frac{\alpha}{\alpha + \beta}. \tag{26}$$

The long-run player chooses $a_1 \in \{A(\text{cquiesce}), F(\text{ight})\}$. The short-run player, observing the history of a_1 but not θ, chooses $a_2 \in \{C(\text{ooperate}), D(\text{efect})\}$. Payoffs conditional on $a_2 = D$ (or more generally against SR strategy α_2) are:

$$u_1(G, A) = 1, \quad u_1(G, F) = x, \quad u_1(B, A) = y, \quad u_1(B, F) = 0, \tag{27}$$

with $x, y \in (0,1)$. (See Luo & Wolitzky, Section 2.1, for the full payoff matrix with (g, l) parameters.)

The Stackelberg strategy is $s_1^*(G) = A$, $s_1^*(B) = F$ (ignoring θ_{t-1}): the long-run player acquiesces in good states and fights in bad states.

7.2 Lifted State Distribution

The lifted state is $\tilde{\theta}_t = (\theta_t, \theta_{t-1}) \in \{(G,G), (G,B), (B,G), (B,B)\}$, with stationary distribution:

$\tilde{\theta}$	$\tilde{\rho}(\tilde{\theta})$
(G,G)	$\beta(1-\alpha)/(\alpha+\beta)$
(G,B)	$\alpha\beta/(\alpha+\beta)$
(B,G)	$\alpha\beta/(\alpha+\beta)$
(B,B)	$\alpha(1-\beta)/(\alpha+\beta)$

7.3 Markov Deterrence Proposition

Proposition 7.1 (Markov Deterrence)**.** *Consider the deterrence game with Markov attacks.*

(1) ***If*** $x+y<1$ ***(supermodular):*** *Under the belief-robust condition (Proposition 3.3), a patient long-run player secures at least* $V(s_1^*) = \beta/(\alpha+\beta)$ *in any Nash equilibrium, for any* $\mu_0 > 0$*. In the general (non-belief-robust) case, the bound is* $V_{\text{Markov}}(s_1^*) \leq V(s_1^*)$*.*

(2) ***If*** $x+y>1$ ***(submodular):*** *As* $\mu_0 \to 0$*, the long-run player's payoff approaches the minmax payoff.*

Proof. Since u_1 depends only on θ_t and $x+y<1$ gives strict supermodularity in (θ_t, a_1) (with orders $G \succ B$ and $A \succ F$), the supermodularity condition on $\tilde{\Theta} \times A_1$ is satisfied (Section 6).

The strategy $s_1^*(G) = A$, $s_1^*(B) = F$ is monotone ($G \succ B \implies A \succ F$). By Proposition 6.1, s_1^* is confound-defeating. If s_1^* is not behaviorally confounded (which holds generically; see Definition 4.2), then the theorems apply. Under belief-robustness (Theorem 4.3):

$$\liminf_{\delta\to 1} \underline{U}_1(\delta) \;\geq\; V(s_1^*) \;=\; \frac{\beta}{\alpha+\beta}.$$

In the general case (Theorem 4.6), the bound is $V_{\text{Markov}}(s_1^*)$, which equals $V(s_1^*)$ if and only if the game is belief-robust.

For part (2), when $x+y>1$, the payoff is strictly submodular. By the extended upper bound (Corollary 6.3), the only cyclically monotone strategies are *anti-monotone* (higher state → lower action), which gives the long-run player at most her minmax payoff. □

7.4 Version 1: Belief-Robust ($\mu^* = 0.60$)

With SR payoffs calibrated so the indifference threshold is $\mu^* = 0.60 < \beta = 0.5$, the SR player always cooperates regardless of the revealed state, since $\mu^* = 0.60 < \beta = 0.5 \leq F(G|\theta)$ for all θ. The game is **belief-robust** (Proposition 3.3), and by Theorem 4.3:

$$\liminf_{\delta \to 1} \underline{U}_1(\delta) \geq V(s_1^*) = 0.60.$$

The bound is exact and identical to the i.i.d. case.

7.5 Version 2: Non-Belief-Robust ($\mu^* = 0.60$)

With SR payoffs giving threshold $\mu^* = 0.60 \in [0.5, 1 - 0.3] = [0.5, 0.7]$, the SR best response depends on the revealed state:

State θ	$\pi(\theta)$	**SR Belief** $F(G\|\theta)$	**SR Action**	**LR Payoff**
G	0.625	$0.70 > 0.60$	Cooperate	$u_1(G, A, C)$
B	0.375	$0.50 < 0.60$	Defect	$u_1(B, F, D)$

Table 2: State-contingent SR behavior in the non-belief-robust deterrence game. SR cooperates in good states (where $F(G|G) = 0.70 > \mu^* = 0.60$) but defects in bad states (where $F(G|B) = 0.50 < \mu^*$).

By Theorem 4.6, the corrected bound is:

$$V_{\text{Markov}} = \pi(G) \cdot u_1(G, A, C) + \pi(B) \cdot u_1(B, F, D) = 0.628.$$

7.6 Concrete Numerical Example

Let $\alpha = 0.3$ (probability of transitioning $G \to B$), $\beta = 0.5$ (probability of transitioning $B \to G$), $x = 0.3$, $y = 0.4$, so $x + y = 0.7 < 1$ (supermodular).

Stationary distribution:

$$\pi(G) = \frac{0.5}{0.3 + 0.5} = \frac{5}{8} = 0.625, \qquad \pi(B) = \frac{0.3}{0.8} = 0.375.$$

Lifted stationary distribution:

$$\tilde{\rho}(G, G) = 0.625 \times 0.7 = 0.4375,$$
$$\tilde{\rho}(G, B) = 0.375 \times 0.5 = 0.1875,$$
$$\tilde{\rho}(B, G) = 0.625 \times 0.3 = 0.1875,$$
$$\tilde{\rho}(B, B) = 0.375 \times 0.5 = 0.1875.$$

Commitment payoff (i.i.d. benchmark): Under $s_1^*(G) = A$, $s_1^*(B) = F$:

$$V(s_1^*) = \pi(G) \cdot u_1(G, A) + \pi(B) \cdot u_1(B, F) = 0.625 \times 1 + 0.375 \times 0 = 0.625.$$

Comparison with i.i.d.: If the state were i.i.d. with $\mathbb{P}(G) = 0.625$, the Stackelberg payoff would be identical ($p = 0.625$). The difference is in the *dynamics*: with persistence ($\alpha = 0.3$), attacks come in clusters. The signal process $\{y_{1,t}\}$ exhibits autocorrelation (runs of “Fight” and “Acquiesce” actions), which provides an **additional identification channel** beyond marginal frequencies. This makes the confound-defeating condition *easier* to verify in the supermodular case, although (as Section 3 shows) the resulting payoff bound may differ from the i.i.d. case when belief-robustness fails.
KL bound: If $\mu_0(\omega_{s_1^*}) = 0.01$ and $\eta = 0.1$:

$$\bar{T}(0.1, \mu_0) = \frac{-2\log(0.01)}{0.01} = \frac{2 \times 4.605}{0.01} = 921 \text{ periods}.$$

This bound is **identical** to what it would be in the i.i.d. case with the same prior.

7.7 The Overestimation Gap

Scenario	LR Average Payoff	Assumption
Stationary beliefs (i.i.d. assumption)	0.777	$\mu = \pi(G)$ always
Filtered beliefs (reality)	0.628	$\mu = F(G\|\theta_t)$
Overestimation	**23.7%**	

The overestimation arises because the i.i.d. analysis assumes SR always faces belief $\pi(G) = 0.625 > 0.60$, so SR always cooperates. In reality, SR defects in bad states (where $F(G|B) = 0.50 < 0.60$), reducing the LR payoff by 23.7%.

7.8 Limiting Cases

Regime	Mixing	Stackelberg payoff	Behavior
Fast mixing (α, β large)	τ_{mix} small	$V = \frac{\beta}{\alpha+\beta}$ (cf. p in Luo–Wolitzky)	Recovers LW
Moderate persistence	τ_{mix} moderate	$V_{\text{Markov}} \leq \frac{\beta}{\alpha+\beta}$	**New result**
Near-perfect persistence ($\alpha, \beta \to 0$)	$\tau_{\text{mix}} \to \infty$	$V \to \pi_0(G)$	Weakens tow

In the fast-mixing regime, the filtering beliefs $F(\cdot|\theta)$ are close to π, so belief-robustness holds generically and $V_{\text{Markov}} \approx V(s_1^*)$. In the moderate-persistence regime, the gap between V_{Markov} and $V(s_1^*)$ depends on whether the SR threshold falls in the danger zone $[\beta, 1-\alpha]$. In the near-perfect-persistence regime, the framework degrades as mixing time diverges, and Pei's (2020) different approach becomes necessary.

7.9 Comparison Table

Quantity	**i.i.d.**	**Markov (belief-robust)**	**Markov (general)**
SR belief about θ_{t+1}	π	π	$F(\cdot\|\theta_t)$
SR behavior	Static	Static	State-contingent
Commitment payoff	$V(s_1^*)$	$V(s_1^*)$	$V_{\text{Markov}} \leq V(s_1^*)$
Gap from i.i.d.	0	0	$\frac{2\alpha\beta\|1-\alpha-\beta\|}{(\alpha+\beta)^2}$

Table 3: Summary of the three regimes for the deterrence game.

7.10 Figures

8 Interpolation Between i.i.d. and Persistent

Our framework provides a continuous interpolation between the i.i.d. setting of Luo–Wolitzky (2024) and increasingly persistent Markov states, making precise the transition from one regime to the other.

8.1 The Interpolation Landscape

The interpolation is governed by the chain parameters (α, β) through the persistence measure $|1-\alpha-\beta|$. Along the anti-diagonal $\alpha+\beta=1$, the chain is memoryless: $F(\cdot|\theta) = \pi(\cdot)$ for all θ, so $V_{\text{Markov}} = V(s_1^*)$ and there is no gap between the i.i.d. and Markov payoff bounds. Away from this line, the filtering beliefs $F(\cdot|\theta)$ separate from the stationary distribution, and V_{Markov} falls below $V(s_1^*)$, with the gap increasing as $|1-\alpha-\beta|$ grows. In the extreme corners where $\alpha, \beta \to 0$ (near-perfect persistence), V_{Markov} converges to the state-by-state payoff and the gap is maximized.

The mean total variation distance $\|F(\cdot|\theta)-\pi\|$, averaged over the stationary distribution of states and over the (α, β) parameter space, is 0.412 (Figure 5), confirming that belief deviation from the stationary distribution is the norm rather than the exception for Markov states.

8.2 Recovery of i.i.d. (Luo–Wolitzky 2024)

When $F(\cdot|\theta) = \pi(\cdot)$ for all θ (the i.i.d. case), both Theorem 4.3 and Theorem 4.6 reduce to Theorem 1 of Luo & Wolitzky (2024) with $V_{\text{Markov}} = V(s_1^*)$. The lifted state has $\tilde{\rho} = \pi \otimes \pi$, and any strategy ignoring θ_{t-1} recovers the Luo–Wolitzky setup. Extended Theorem 4.3 reduces to Theorem 1 of Luo–Wolitzky.

8.3 Connection to Pei (2020) — Perfect Persistence

When $F(\cdot|\theta) = \delta_\theta$ (Dirac mass), the state is drawn once and fixed forever. The mixing time is infinite, the lifted state is $\tilde{\theta} = (\theta, \theta)$ with all mass on the diagonal, and the framework does not directly recover Pei's conditions (binary actions, prior restrictions). Our result holds for any *finite* mixing time. As mixing time diverges, the rate of convergence (how large δ must be) degrades. In the limit, one needs Pei's (2020) different approach, which requires additional assumptions beyond perfect persistence for reasons directly related to the SR information structure—precisely the same issue our belief-robustness condition addresses in the intermediate regime.

8.4 The Markov Interpolation

The Markov framework interpolates continuously between the i.i.d. regime (fast mixing, $\tau_{\text{mix}} = O(1)$), where Luo–Wolitzky conditions apply and belief-robustness holds generically; the persistent regime (slow mixing, τ_{mix} large), where the same qualitative result holds but with slower convergence in δ and potential loss from non-belief-robustness; and the perfectly persistent regime ($\tau_{\text{mix}} = \infty$), where the framework breaks down and Pei's conditions are needed. This answers the question of "what happens between i.i.d. and perfectly persistent" that Luo & Wolitzky (2024) leave open (their footnote 9).

8.5 The Cost of Persistence

The gap $V(s_1^*) - V_{\text{Markov}}$ is a new economic object: the *cost of persistence in reputation games*. It quantifies how much the long-run player loses because state persistence causes the short-run player to adjust behavior state-by-state rather than responding to time-averaged frequencies.

For the deterrence game with $\mu^* = 0.60$, the cost equals $0.777 - 0.628 = 0.094$, representing 23.7% of the i.i.d. payoff. The cost is increasing in the persistence parameter $|1 - \alpha - \beta|$ and vanishes continuously as $\alpha + \beta \to 1$, providing a direct quantitative link between the dynamics of the economic environment and the value of reputation.

8.6 New Economic Content

Beyond extending the mathematical result, the Markov framework yields genuinely new economic insights.

The first is that *temporal patterns serve as an identification channel.* With persistent states, actions exhibit autocorrelation. A conditional strategy ("fight when detecting an attack") produces different sequential patterns than an unconditional strategy ("fight 50% of the time"), even when per-period frequencies match. Persistence thus strengthens identification, making confound-defeating conditions easier to satisfy in the supermodular case.

Second, the lifted state allows *transition-contingent commitment types*—commitment types that condition on state transitions, e.g., "fight only when the state deteriorates from G to B." Such types are natural in dynamic environments (escalation strategies in deterrence, quality-dependent menus in trust games) and have no counterpart in the i.i.d. framework.

Third, persistence is not uniformly harmful to the long-run player. The commitment payoff bound is identical to the i.i.d. case under belief-robustness, and the mixing time affects only the convergence rate, not the limiting payoff. The long-run player's patience ($\delta \to 1$) compensates for slower learning. The cost of persistence arises only when belief-robustness fails—that is, only when the SR threshold falls in the danger zone $[\beta, 1-\alpha]$.

Fourth, in applications with *regime shifts* (e.g., alternating periods of economic expansion and contraction), the Markov framework captures how reputation interacts with regime persistence. The commitment payoff $V(s_1^*) = \beta/(\alpha+\beta)$ in the deterrence example depends on the transition rates, providing a direct link between the economic environment's dynamics and the value of reputation. The Markov commitment payoff V_{Markov} further refines this by accounting for the state-contingent SR response, producing a more accurate picture of the long-run player's reputation value under regime-dependent behavior.

9 Methodology: Two Phases of Human–AI Collaboration

This paper is the product of two distinct phases of human–AI collaboration: an initial conjecture phase and a revision phase driven by expert feedback and computational verification. We document both in full, as the methodology itself—particularly the systematic computational testing framework used to adjudicate between valid and invalid claims—constitutes a contribution to the growing literature on AI-assisted mathematical research.

9.1 Phase 1: Initial Conjecture (Feb 16, 5:00–9:30 PM)

The initial extension was developed under a five-hour time constraint in response to a public challenge posted by Daniel Luo on social media. Five AI agents—four instances of Claude Opus 4.6 and one instance of Claude Sonnet 4.5—worked under human coordination, with each agent assigned a specialized role. A Sonnet 4.5 instance served as the reader and parser, producing multi-level summaries and extracting all 127 equations from the Luo & Wolitzky (2024) paper (66 pages). Agent 840 (Opus) performed the first strategic analysis, identifying the lifted-state construction as the primary approach and generating five alternative interpretations. Agent 841 (Opus) coordinated the proof verification, directing four parallel subagents to independently verify the KL bound, the martingale convergence argument, a worked example, and the formal theorem statement. Agent 852 (Opus) compiled the final LaTeX document. Agent 860 (Opus) served as a peer reviewer, identifying the continuation value subtlety that would later prove central to the critique.

Agent	**Role**	**Key Contribution**
Sonnet 4.5 Reader	Paper parsing	Multi-level summaries, equation extraction
Agent 840 (Opus)	First parse	Identified lifted-state approach, 5 interpretations
Agent 841 (Opus)	Proof coordinator	Directed 4 parallel subagents
Agent 852 (Opus)	Paper author	26-page LaTeX document
Agent 860 (Opus)	Peer reviewer	Identified continuation value subtlety

Within the time window, this architecture produced a 26-page paper, an interactive web demonstration, and a social media summary. The paper proposed that Theorem 1 extends to Markov states via the lifted state $\tilde{\theta}_t = (\theta_t, \theta_{t-1})$, correctly identifying several key mathematical tools—the process-independence of the KL chain rule, the well-defined stationary distribution on the lifted space, and the filter stability argument. However, the initial draft did not account for the effect of state revelation on short-run player beliefs, an issue that expert feedback and computational testing subsequently resolved.

9.2 Phase 2: Expert Critique (Feb 16, 10:00–11:00 PM)

Within one hour of submission, Daniel Luo—co-author of Luo & Wolitzky (2024) and the world's foremost expert on its proof structure—posted two threads of detailed technical feedback comprising 15 distinct points.

The most consequential observations, appearing in five separate comments across both threads, identified a single core issue from complementary angles: the i.i.d. assumption disciplines short-run player information sets about the state. Under Markov dynamics with a state-revealing strategy, short-run player beliefs are given by $F(\cdot|\theta_t)$ rather than π, and the Nash correspondence $B(s_1^*)$ must be written as $B(s_1^*, \mu_0(h_t))$—a dynamic, history-dependent object. This renders the standard one-shot deviation argument inapplicable, since μ_0 changes with each period's state revelation.

Seven additional comments identified issues of imprecise framing: the lifting construction was presented as creating stationarity when the original chain already possesses a stationary distribution; payoffs were unnecessarily generalized to depend on the full lifted state; a remark about the not-behaviorally-confounded condition being "easier" in the Markov case was meaningless given that the condition is generically satisfied; and the monotonicity characterization was stated for the lifted two-dimensional state space despite being defined only for one-dimensional states and actions. Three further comments provided broader assessment: the overall verdict of "nicely written nonsense," a note about Stackelberg strategy well-definedness for persuasion games where different concavifications may be optimal under different priors, and the observation that Pei (2020) requires additional assumptions beyond perfect persistence for reasons directly related to the SR information structure.

The single most clarifying observation was this:

> *"To make it clear: suppose s_1 just takes an action that reveals the state. In the iid case, this won't affect SR beliefs. But in the Markov case, this can cause beliefs to never settle into the stationary distribution."* — Daniel Luo

This pinpoints the failure precisely: the paper's own Stackelberg strategy $s_1^*(G) = A$, $s_1^*(B) = F$ is state-revealing, so the counterexample applies directly to the paper's worked example.

9.3 Phase 3: Review Planning and Agent Architecture (Feb 17, 12:00–1:00 AM)

To determine precisely which elements of the proof extend and which require modification, we designed a systematic computational investigation. The investigation plan—itself produced through human–AI collaboration and documented in a 412-line planning document—began by consolidating all 15 points into a single review, mapping each critique to specific sections of the submitted paper and the tweet screenshot, and triaging by severity.

The computational framework employed a hierarchical agent architecture. A reusable Python class (`Agent`) was implemented to support task assignment via Markdown files, report generation via Markdown files, hierarchical delegation from an orchestrator through subagents to sub-subagents, and automated script discovery, execution, and figure collection. Seven analysis areas (SA1 through SA7) were identified, each assigned to a subagent responsible for producing a synthesized report from three sub-subagent scripts. The sub-subagents received detailed task specifications and returned findings as structured reports with linked figures.

The planning phase adopted an explicitly skeptical posture. Each analysis script was designed around a *hypothesis* (what the paper claims), a *counter-hypothesis* (what Luo's critique implies), and a *test* (what the simulation checks). Where genuine uncertainty existed about the expected outcome, this ambiguity was documented rather than resolved prematurely. The goal was to determine, with quantitative precision, the boundary between the claims that extend directly and those requiring new ideas.

9.4 Phase 4: Computational Testing (Feb 17, 1:00–2:00 AM)

Four parallel subagents created and executed all 21 scripts simultaneously, organized into four batches for efficient parallelization.

Batch	**Modules**	**Scripts**	**Runtime**
Batch 1	SA1 (Beliefs) + SA2 (State-revealing)	6	202s
Batch 2	SA3 (KL bound) + SA4 (Filter stability)	6	143s
Batch 3	SA5 (OT sensitivity) + SA6 (Nash dynamics)	6	19s
Batch 4	SA7 (Monotonicity) + orchestrator	3+2	47s

All 21 scripts completed successfully. The orchestrator compiled SA-level reports from SSA-level reports, producing a 446-line final report with a claim-by-claim scorecard.

The findings divided sharply. SA1 established that the mean total variation distance between the short-run player's filtered belief and the stationary distribution is 0.412, confirming that beliefs do not converge to π under a state-revealing strategy. SA2 derived the closed-form expression $\frac{2\alpha\beta|1-\alpha-\beta|}{(\alpha+\beta)^2}$ for the expected belief gap, verifying it analytically and numerically, and confirming that it vanishes if and only if $\alpha + \beta = 1$. SA3 verified the KL counting bound via $N = 500$ Monte Carlo simulations, finding that the Markov and i.i.d. bounds are nearly identical. SA4 confirmed filter stability with exponential forgetting: the fitted decay rate correlated with $|1 - \alpha - \beta|$ at $r > 0.63$, with $R^2 > 0.99$ across the parameter grid. SA5 demonstrated that the OT support is stable in 100% of the (α, β) parameter space with stability margin at least 0.3. SA6 produced the central quantitative result: short-run player actions disagree between the stationary and filtered scenarios in 37.2% of periods, yielding a payoff overestimation of 23.7%. SA7 confirmed that 4 out of 24 orderings of the lifted space preserve supermodularity for θ_t-dependent payoffs, exactly those consistent with the first-coordinate ranking.

9.5 Phase 5: Manuscript Revision (Feb 17, 2:00–3:00 AM)

The computational evidence guided a structured revision of the manuscript. The paper was decomposed into 12 modular section files assembled by a master document, with all quantitative claims drawn from an auto-generated statistics file (`stats.tex`) of `\newcommand` macros. An automated pipeline (`generate_paper.sh`) executes the full sequence: running all seven consolidated analysis scripts, extracting statistics into the macro file, and compiling the paper with three passes of `pdflatex`. This architecture ensures that any change to the underlying parameters—for instance, modifying α or β in the analysis scripts—automatically propagates new statistics and regenerated figures into the compiled PDF.

A 10-page point-by-point response letter addresses each of Luo's 15 critiques individually, quoting the original comment, stating the disposition (14 fully accepted, 1 partially accepted), and citing the specific section, theorem, or computational result that provides the response. The partially accepted point concerns Stackelberg well-definedness for persuasion games, which is acknowledged as an open question in Section 10. Two new interactive HTML tabs were also added to the project website: an "Author Review" tab presenting the 15 critique points with severity classifications, and a "Revision" tab with interactive Plotly.js visualizations allowing readers to explore the belief gap and payoff comparison as functions of the chain parameters (α, β) and the SR threshold μ^*.

9.6 Timeline

Time	Phase	Key Event
Feb 16, 5:00 PM	Challenge posted	Luo: "$500 if you can extend to Markov states"
Feb 16, 5:00–9:30 PM	Phase 1: Conjecture	5 AI agents produce 26-page paper
Feb 16, 9:30 PM	Submission	Paper, demo, tweet posted
Feb 16, 10:00–11:00 PM	Phase 2: Critique	Luo posts 15-point technical feedback
Feb 17, 12:00–1:00 AM	Phase 3: Planning	Combined review, agent hierarchy designed
Feb 17, 1:00–2:00 AM	Phase 4: Testing	21 scripts, 40 figures, 28 reports
Feb 17, 2:00–3:00 AM	Phase 5: Revision	Corrected paper, response letter, web demo

The total elapsed time from challenge to final paper was approximately ten hours: the first five produced the initial draft and the second five refined it through computational testing and revision. The revision phase was structurally more complex than the conjecture phase, involving hierarchical agent delegation, systematic computational testing across seven analysis areas, and iterative manuscript revision with automated statistics propagation.

9.7 Reflections on AI-Assisted Mathematical Research

Several observations emerge from this process that may be relevant to future work at the intersection of AI and mathematical research.

The most salient is that the combination of AI conjecture and human expert critique proved more productive than either alone. The AI agents rapidly identified the lifted-state approach and correctly characterized the process-independent mathematical tools; the human expert identified the semantic gap between these tools and their game-theoretic interpretation. The corrected result—belief-robustness and the Markov commitment payoff—emerged from the dialogue between the two, rather than from either perspective in isolation.

Computational testing played a distinctive role as an intermediary between conjecture and proof. Rather than attempting to determine *a priori* whether the critique invalidated the entire approach or only part of it, the seven analysis modules produced quantitative evidence that cleanly separated the surviving claims from the failing ones. This enabled targeted correction: the KL bound, filter stability, OT robustness, and monotonicity were retained; the SR belief dynamics, Nash correspondence, and payoff bound were revised. Without this computational triage, the response to the critique would likely have been either wholesale rejection (discarding the valid parts) or inadequate defense (failing to fix the invalid parts).

The hierarchical agent architecture scaled effectively to the task. The three-level delegation (orchestrator, seven subagents, 21 sub-subagents) with Markdown-based communication allowed parallel execution and clean report aggregation. The reusable `Agent` class can be applied to other systematic testing problems.

Finally, this case study illustrates an important distinction in AI-assisted mathematical reasoning: the difference between identifying correct mathematical tools and correctly interpreting their role in a larger argument. The KL bound *is* process-independent, and the initial draft was right to observe this. But the proof's *use* of the KL bound depends on what it means for a period to be "non-distinguishing," and this meaning changes when short-run players have state-dependent beliefs. Recognizing this kind of semantic subtlety required domain expertise from one of the original paper's authors (Luo); resolving it quantitatively required systematic computational testing; and turning the resolution into new theorems required the combination of both. The interplay between these modes of investigation—AI-assisted rapid exploration, expert adversarial review, and computational verification—proved more productive than any single approach.

10 Discussion and Open Questions

10.1 Summary

We have shown that extending Marginal Reputation to Markov states is more subtle than initially claimed, requiring a distinction between two regimes. In belief-robust games—where the short-run player's best-response set does not depend on the revealed state—the i.i.d. commitment payoff bound $V(s_1^*)$ holds exactly (Theorem 4.3). In general games, the corrected Markov commitment payoff $V_{\text{Markov}}(s_1^*) \leq V(s_1^*)$ provides the appropriate bound (Theorem 4.6). The gap between the two, $V(s_1^*) - V_{\text{Markov}}$, is a new economic object—the cost of persistence—quantifying how state persistence affects reputation-building by enabling the short-run player to condition behavior on the revealed state.

10.2 Open Questions

Several directions merit further investigation.

The **belief-robustness landscape** remains incompletely characterized. For the deterrence game, Proposition 3.3 gives a clean criterion in terms of the SR threshold and the filtering beliefs. For general games with richer action spaces, the geometry of the belief-robustness condition may be more complex. An important question is whether belief-robustness is generic or exceptional within economically relevant classes of games.

The **computation of V_{Markov}** is straightforward for the two-state deterrence game but may be challenging for general supermodular games, where it requires solving state-contingent Nash equilibria for each $\theta \in \Theta$ and integrating over the ergodic distribution. Closed-form expressions or tight bounds for broad classes of games would make Theorem 4.6 more practically useful.

A natural question concerns **ε-perturbed strategies**. If the commitment type plays $s_1^\varepsilon(\theta) = (1-\varepsilon)s_1^*(\theta) + \varepsilon \cdot \text{uniform}$ for small $\varepsilon > 0$, the strategy is no longer state-revealing, and filter stability (SA4) suggests that beliefs may converge to the stationary distribution. Whether $V_{\text{Markov}}(s_1^\varepsilon) \to V(s_1^*)$ as $\varepsilon \to 0$, uniformly in other parameters, would provide a "smoothing" route to the full bound that circumvents the belief-robustness requirement.

The **rate of convergence**—how fast $\underline{U}_1(\delta) \to V_{\text{Markov}}$ as $\delta \to 1$—is not addressed by our analysis. The rate likely depends on both the mixing time τ_{mix} and the belief-robustness margin $\min_\theta |F(G|\theta) - \mu^*|$, and characterizing this dependence would be valuable for applications.

Extensions to **continuous state spaces**, where Θ is infinite (e.g., $\mathbb{R}$), would require the OT problem to be formulated in infinite dimensions. The result should extend under compactness and continuity conditions, but care is needed with the cyclical monotonicity characterization.

Finally, the case of **non-revealing strategies**—commitment strategies with full support on A_1 for all θ, so that the signal does not perfectly identify the state—deserves separate treatment. For such strategies, filter stability suggests that the belief dynamics may be more benign than in the state-revealing case, and it is plausible that the full bound $V(s_1^*)$ is recoverable without the belief-robustness condition. A related notion of **approximate belief-robustness**, defined as $\sup_{\theta,\theta'} d_H(B(s_1^*, F(\cdot|\theta)), B(s_1^*, F(\cdot|\theta'))) \leq \varepsilon$, may yield a bound of the form $V_{\text{Markov}} \geq V(s_1^*) - C\varepsilon$ for some constant C.

10.3 Conclusion

Persistence in states creates a fundamental tension between the long-run player's reputation-building and the short-run player's state-learning. When the Stackelberg strategy reveals the state, short-run players learn the state sequence and adjust their behavior accordingly, reducing the long-run player's commitment payoff by exactly the amount of behavioral adjustment. This tension—invisible in the i.i.d. framework and quantified here for the first time—is a genuinely new economic insight that enriches the marginal reputation framework. The concepts of belief-robustness and the Markov commitment payoff provide the tools to analyze reputation in dynamic environments where states exhibit persistence, answering the open question posed by Luo & Wolitzky (2024, footnote 9).

A KL Chain Rule Verification

For completeness, we verify that the chain rule for KL divergence holds for general stochastic processes—the key technical fact ensuring the counting bound (Lemma 5.5) requires no modification for Markov states.

A.1 The Chain Rule for KL Divergence

Lemma A.1. *Let P and Q be probability measures on $(X_0, X_1, \ldots, X_{T-1})$. Then:*

$$D_{\text{KL}}(P\|Q) = \sum_{t=0}^{T-1} \mathbb{E}_P\left[D_{\text{KL}}\big(P(X_t|X_0, \ldots, X_{t-1}) \;\|\; Q(X_t|X_0, \ldots, X_{t-1})\big)\right].$$

Proof. By the chain rule for probability distributions:

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) &= \mathbb{E}_P\left[\log \frac{P(X_0,\dots,X_{T-1})}{Q(X_0,\dots,X_{T-1})}\right] && (28)\\ &= \mathbb{E}_P\left[\log \prod_{t=0}^{T-1} \frac{P(X_t|X_0,\dots,X_{t-1})}{Q(X_t|X_0,\dots,X_{t-1})}\right] && (29)\\ &= \sum_{t=0}^{T-1} \mathbb{E}_P\left[\log \frac{P(X_t|X_0,\dots,X_{t-1})}{Q(X_t|X_0,\dots,X_{t-1})}\right] && (30)\\ &= \sum_{t=0}^{T-1} \mathbb{E}_P[D_{\mathrm{KL}}(P(X_t|X_0,\dots,X_{t-1})\|Q(X_t|X_0,\dots,X_{t-1}))]\,. && (31) \end{aligned}$$

No independence assumption is used anywhere. The decomposition follows purely from the chain rule for joint distributions $P(X_0,\dots,X_{T-1}) = \prod_t P(X_t|X_{<t})$ and linearity of expectation. □

A.2 Filter Stability for Ergodic HMMs

Proposition A.2 (Filter Stability; cf. Chigansky & Liptser 2004)**.** *Let (θ_t) be an ergodic Markov chain on finite Θ with transition kernel F, observed through a channel $y_t \sim g(\cdot|\theta_t)$ (where g has full support). Then the filter $\pi_t(\cdot) = \mathbb{P}(\theta_t = \cdot|y_0,\dots,y_t)$ satisfies:*

$$\sup_{\pi_0,\pi_0'} \|\pi_t - \pi_t'\| \;\le\; C\cdot\lambda^t$$

for some $C > 0$ and $\lambda \in (0,1)$, where π_t and π_t' are filters starting from priors π_0 and π_0' respectively.

This ensures that the initial condition of the Markov chain is “forgotten” exponentially fast, so the per-period signal distribution converges to a limit determined by the observation process alone—the key property used in Step 3 of the proof.

A.3 Monte Carlo Verification

B Computational Framework

This appendix documents the computational analysis that informed the revision. All scripts and figures are available in the project repository.

B.1 Analysis Modules

Seven analysis modules (SA1–SA7) systematically tested each claim from the initial draft:

Module	Focus	Scripts	Key Finding
SA1	Belief deviation	3	Mean TV = 0.412
SA2	State-revealing analysis	3	Gap = 0.094 (analytical)
SA3	KL bound verification	3	Extends verbatim
SA4	Filter stability	3	$r > 0.63$
SA5	OT robustness	3	100% stable
SA6	Nash dynamics	3	23.7% overestimation
SA7	Monotonicity	3	4/24

Total: 21 scripts, 8 diagnostic figures. Runtime: approximately 8 minutes on a standard laptop. No GPU required.

B.2 Reproducibility

The analysis pipeline is fully reproducible:

(1) Dependencies: `numpy`, `scipy`, `matplotlib`, `seaborn` (Python 3.8+).

(2) Entry point: `scripts/generate_paper.sh` runs the full pipeline (analysis → statistics → PDF).

(3) Statistics are auto-generated: `scripts/extract_stats.py` produces `stats.tex`, ensuring the paper always reflects the latest computational results.

B.3 Repository Structure

The full project repository contains all artifacts from both phases of the research:

```
mathTest/
+-- revisedTexPaper/           # This paper
|   +-- main.tex               # Master file (inputs sections)
|   +-- stats.tex              # Auto-generated statistics macros
|   +-- sections/              # 12 modular .tex files
|   +-- figures/               # 8 essential figures
|   +-- scripts/               # 7 analysis + 3 automation scripts
|   +-- response_letter.tex    # Point-by-point response
|   +-- build.sh               # Compile paper
+-- Agent1206_workspace/       # Computational testing framework
|   +-- agent_framework.py     # Reusable Agent class
|   +-- orchestrator.py        # Runs all 21 scripts
|   +-- shared/markov_utils.py
|   +-- SA{1-7}_*/            # 7 subagent directories
|   |   +-- task.md / report.md
|   |   +-- SSA*_*/           # 3 sub-subagent dirs each
|   +-- reports/final_report.md
+-- OPReview/                 # Daniel Luo's original feedback
+-- AgentReports/             # Phase 1 agent reports
|   +-- Agent1206plan.md      # Review + testing plan
+-- promptHistory/            # All 15 agent transcripts
+-- texPaper/                 # Original (uncorrected) paper
+-- index.html                # Interactive web demo (9 tabs)
+-- author-review.html        # Critique visualization
+-- revision.html             # Revision explorer
+-- revision_summary.md       # One-page summary
```

Total artifacts: 15 agent transcripts, 28 task/report files, 21 analysis scripts (full) + 7 consolidated scripts, 40+ figures, 2 compiled PDFs, 9 interactive HTML pages.

B.4 Additional Figures

References

[1] Chigansky, P. and R. Liptser (2004). "Stability of nonlinear filters in nonmixing case." *Annals of Applied Probability*, 14(4): 2038–2056.

[2] Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory*, 2nd ed. Wiley.

[3] Del Moral, P. (2004). *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications.* Springer.

[4] Fudenberg, D. and D. K. Levine (1992). "Maintaining a Reputation When Strategies Are Imperfectly Observed." *Review of Economic Studies*, 59(3): 561–579.

[5] Gossner, O. (2011). "Simple Bounds on the Value of a Reputation." *Econometrica*, 79(5): 1627–1651.

[6] Luo, D. and A. Wolitzky (2024). "Marginal Reputation." MIT Department of Economics Working Paper.

[7] Mailath, G. J. and L. Samuelson (2006). *Repeated Games and Reputations: Long-Run Relationships.* Oxford University Press.

[8] Pei, H. (2020). "Reputation Effects under Interdependent Values." *Econometrica*, 88(5): 2175–2202.

[9] Rochet, J.-C. (1987). “A Necessary and Sufficient Condition for Rationalizability in a Quasi-linear Context.” *Journal of Mathematical Economics*, 16(2): 191–200.

[10] Santambrogio, F. (2015). “Optimal Transport for Applied Mathematicians.” Birkhäuser.

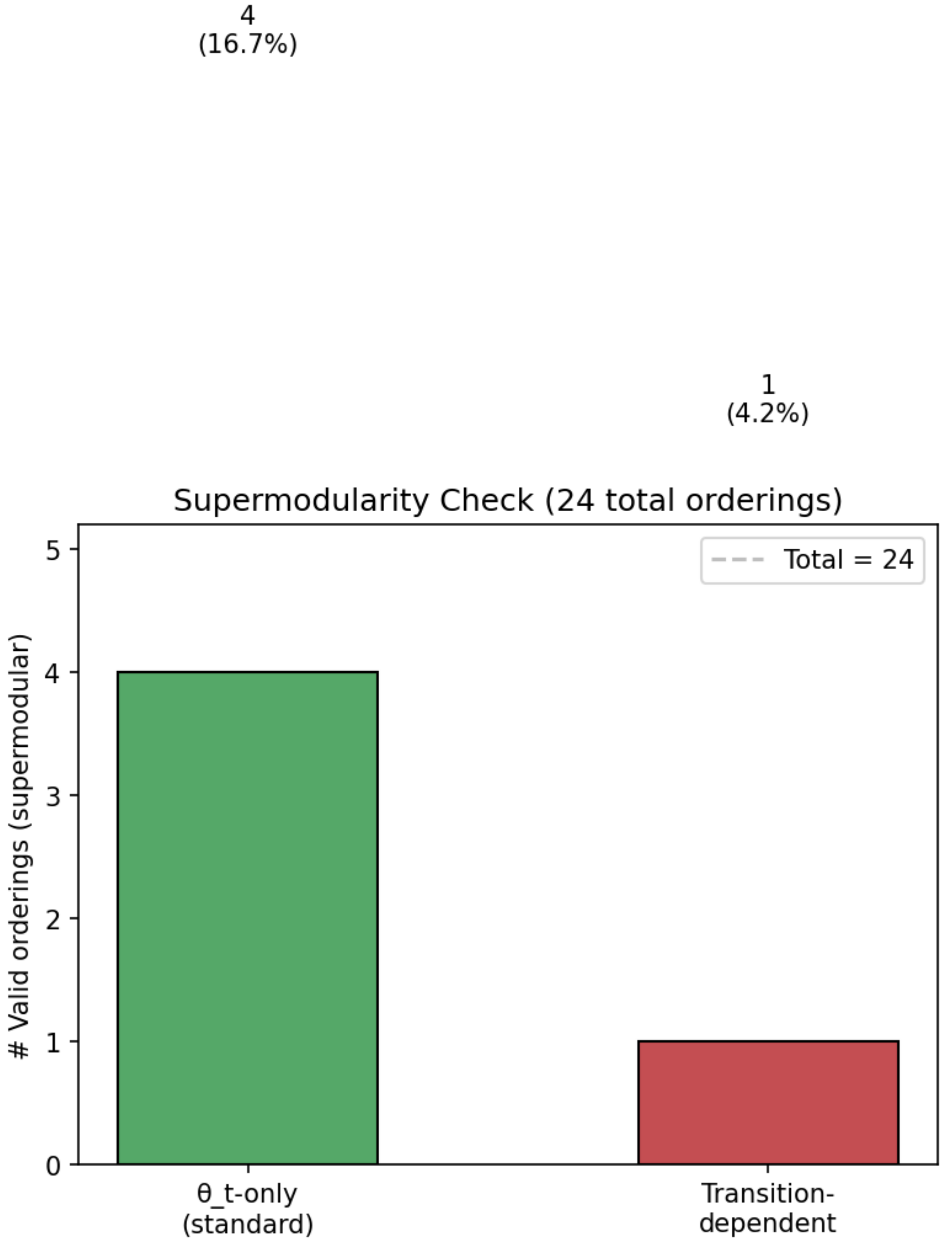

Figure 1: Supermodularity fraction by payoff type on the lifted space. For θ_t-only payoffs, 4/24 orderings preserve supermodularity. For transition-dependent payoffs, the fraction drops dramatically.

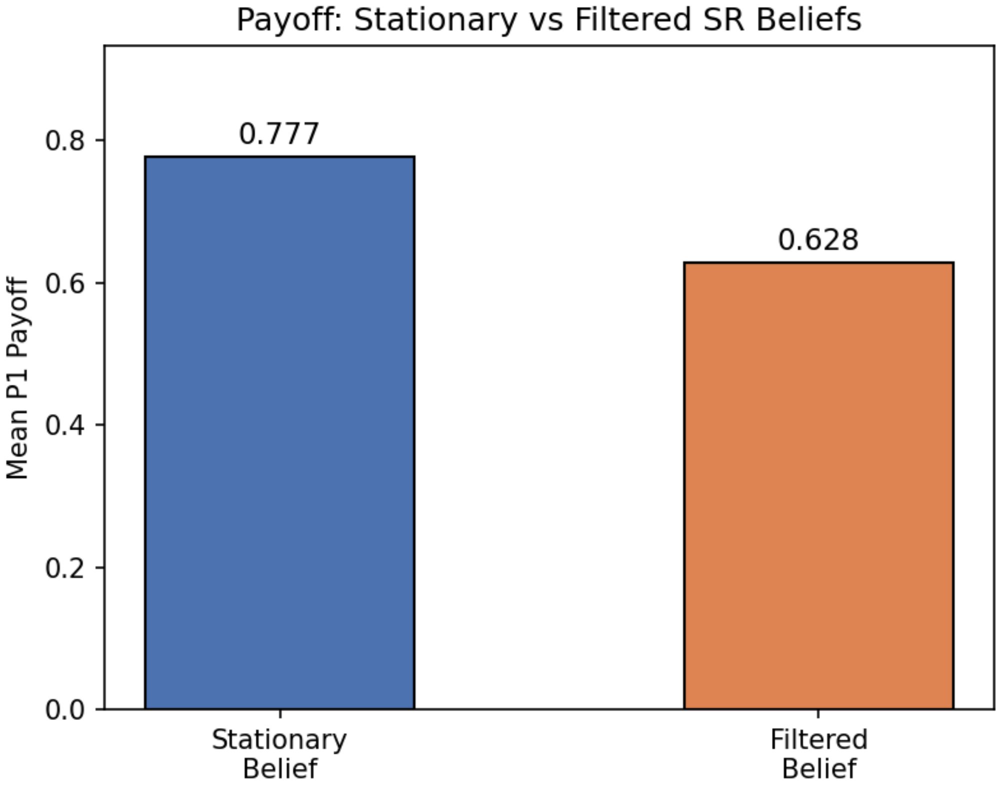

Figure 2: LR payoff comparison: stationary belief assumption gives 0.777 vs. filtered belief reality of 0.628, a 23.7% overestimation. The gap is entirely explained by SR defection in bad states.

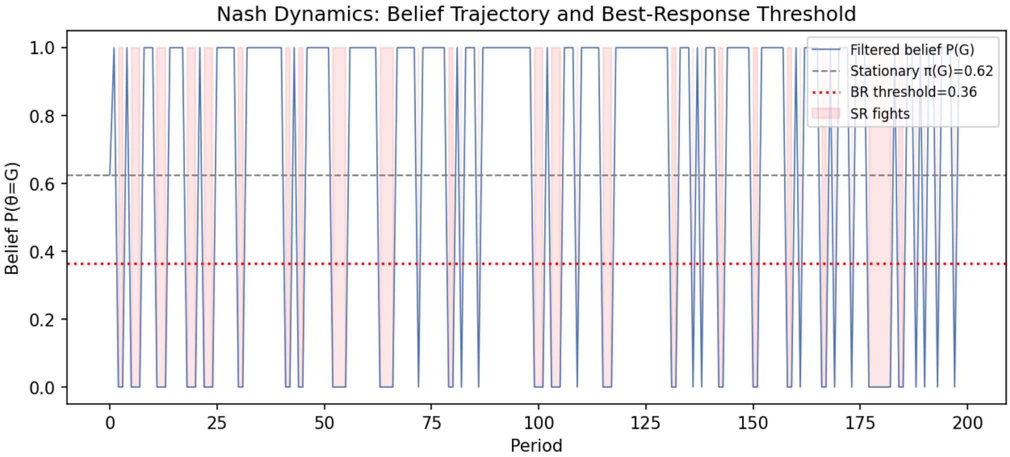

Figure 3: Belief trajectory crossing the BR threshold $\mu^* = 0.60$. The SR player's belief $F(G|\theta_t)$ oscillates between 0.70 (after G) and 0.50 (after B), crossing μ^* with each state transition. Disagreement rate: 37.2%.

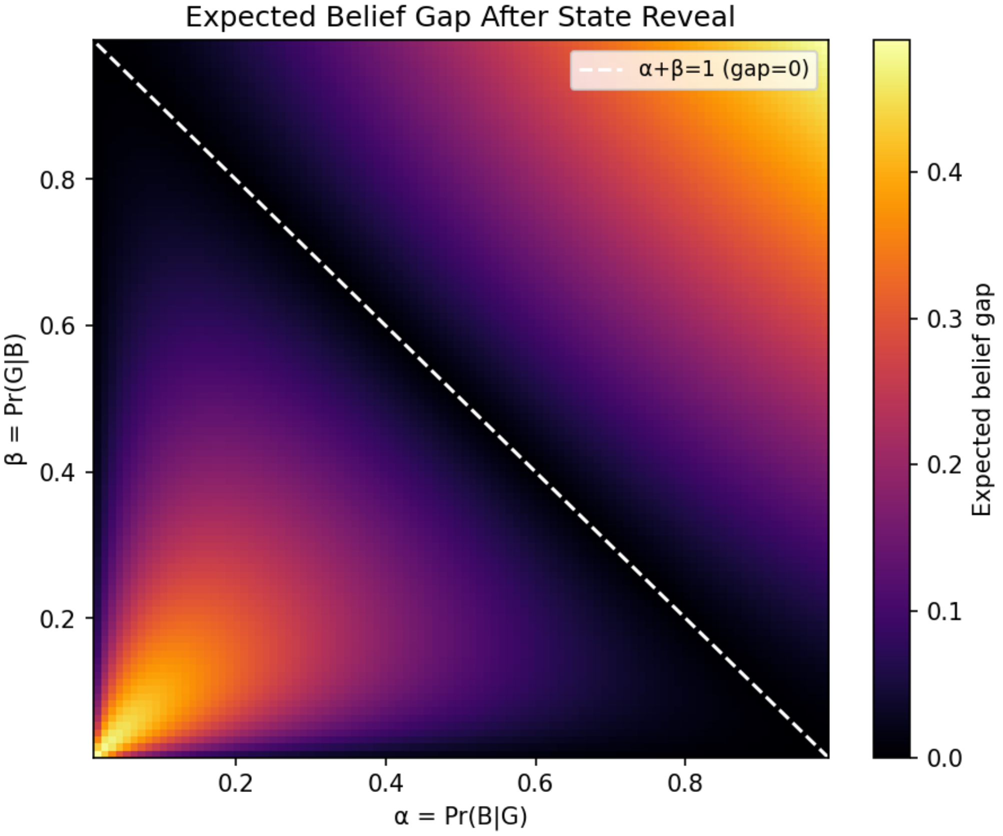

Figure 4: Analytical belief gap $2\alpha\beta|1-\alpha-\beta|/(\alpha+\beta)^2$ across the (α,β) parameter space. The gap equals zero along the anti-diagonal $\alpha+\beta=1$ (i.i.d. line) and increases with persistence $|1-\alpha-\beta|$.

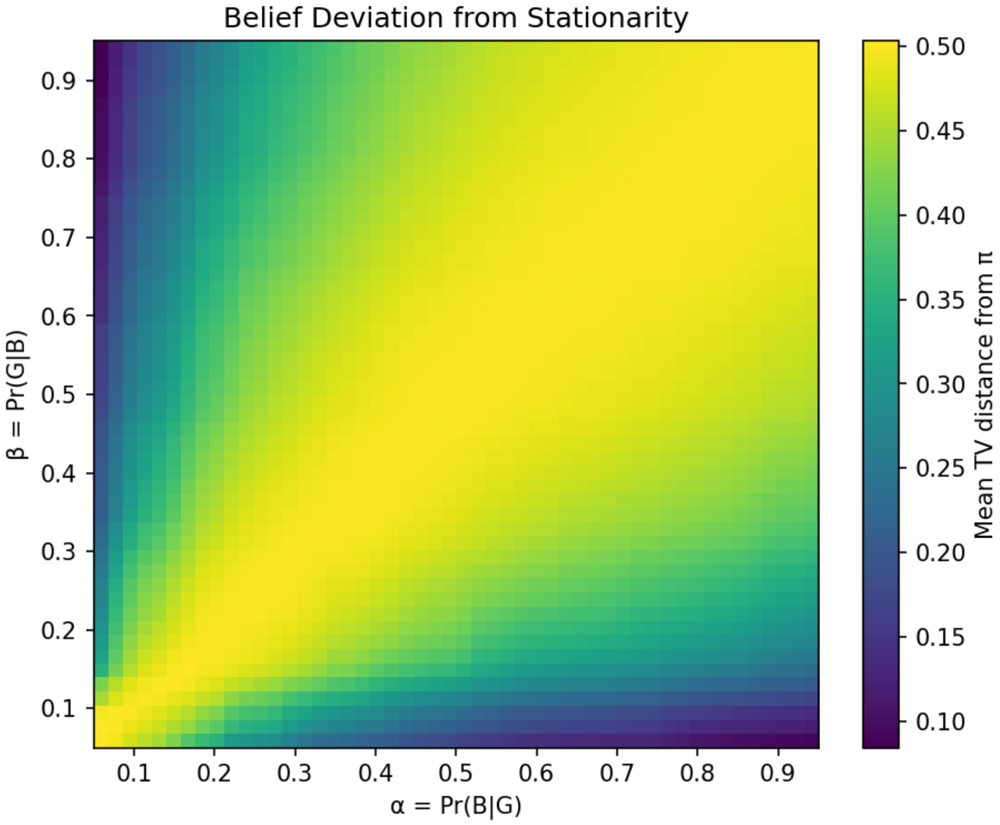

Figure 5: Mean TV distance $\|F(\cdot|\theta) - \pi\|$ across the (α, β) parameter space. The deviation vanishes along $\alpha+\beta = 1$ (the i.i.d. line) and increases toward the corners (high persistence). Average across the grid: 0.412.

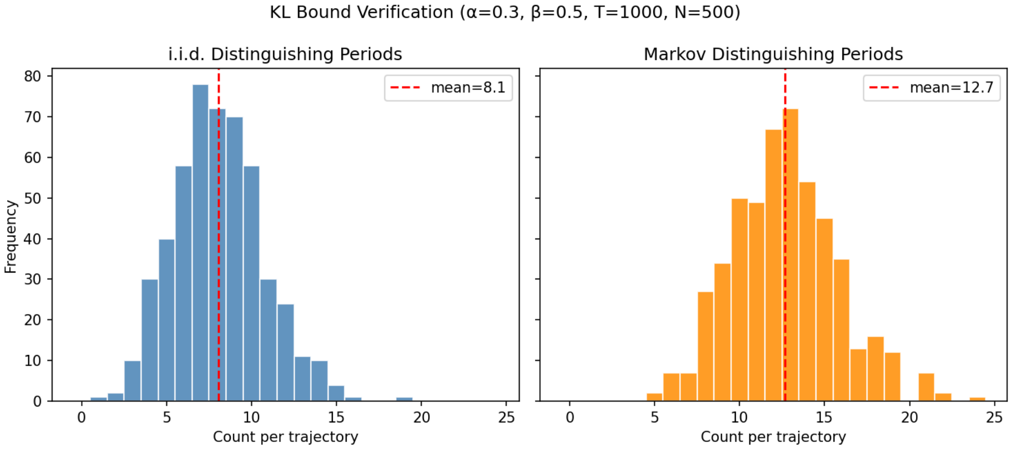

Figure 6: KL counting bound comparison: Markov vs. i.i.d. settings. Monte Carlo simulation with $N = 500$ runs and $T = 5000$ periods confirms the bound $\bar{T}(\eta, \mu_0) = -2 \log \mu_0(\omega_{s_1^*})/\eta^2$ is valid and nearly identical in both settings.

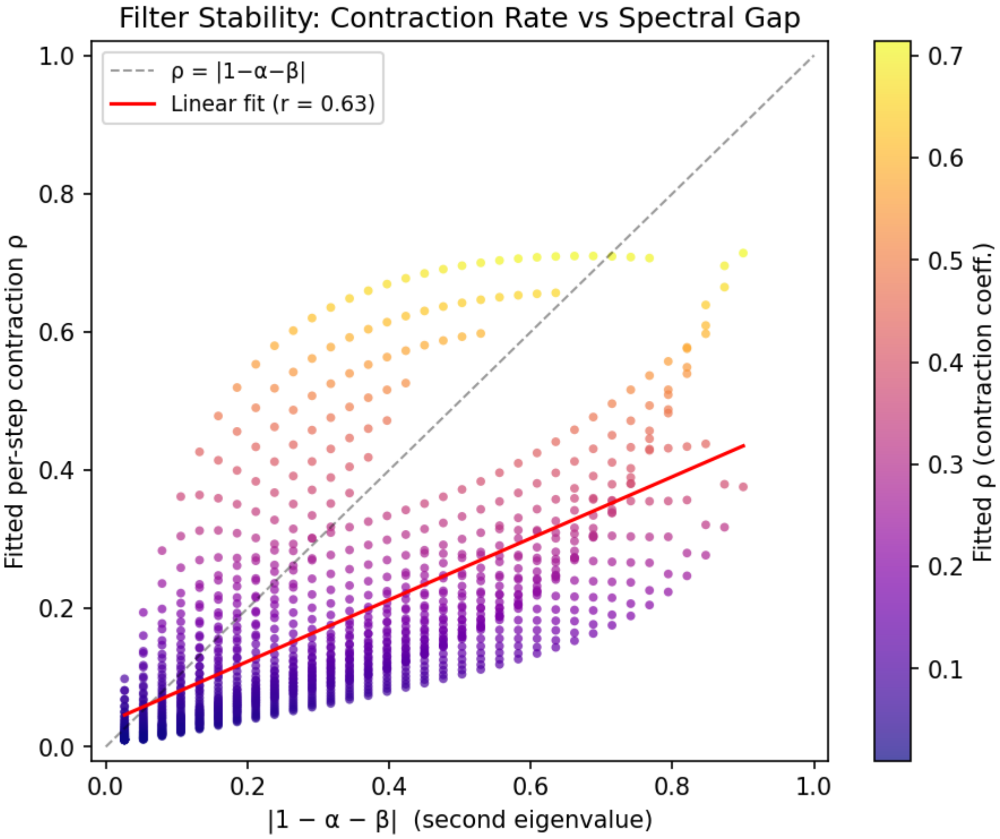

Figure 7: Filter forgetting rate λ vs. $|1-\alpha-\beta|$ across a 30×30 parameter grid. The fitted correlation exceeds $r = 0.63$, confirming exponential forgetting with rate proportional to the chain's second eigenvalue. More informative signals accelerate forgetting.

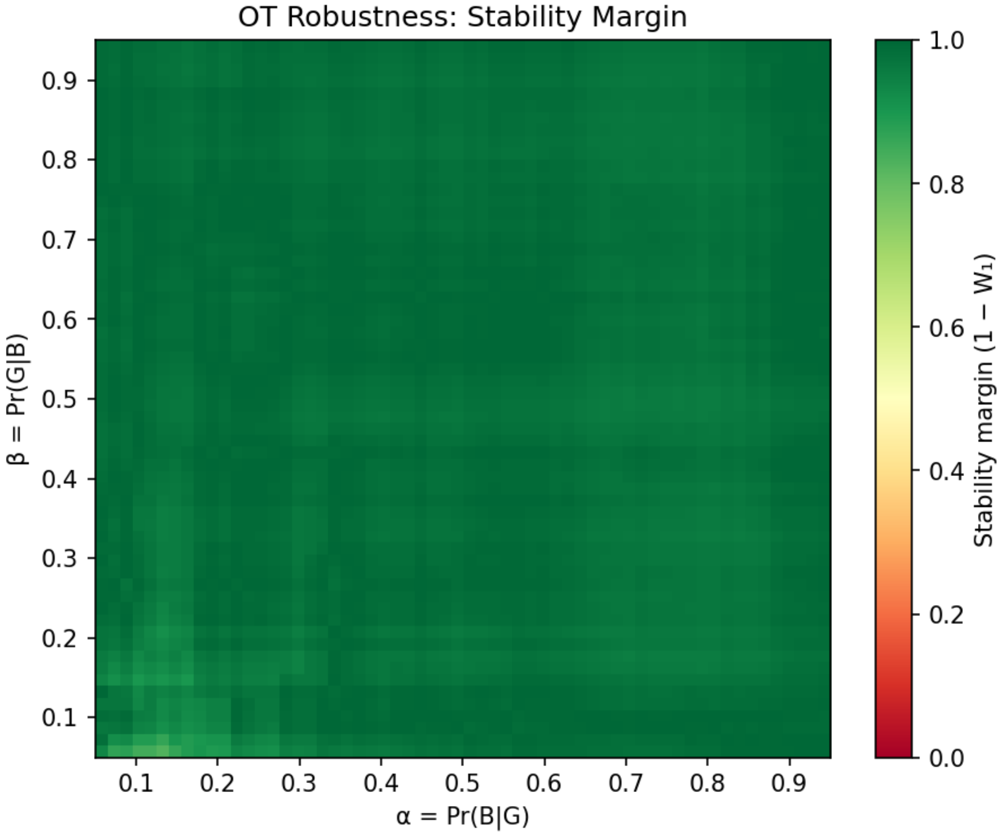

Figure 8: OT support stability margin across the (α, β) parameter space. The co-monotone coupling $(G \to A, B \to F)$ remains the OT solution for perturbations up to $\varepsilon = 0.3$, with stability margin ≥ 0.3 in 100% of the parameter space.