Extending Marginal Reputation to Persistent Markovian States

A Response to the Extension Challenge

Kyle Elliott Mathewson[1]

with AI Collaborators:

Claude Opus 4.6 — Agents 840, 841, 852
Claude Sonnet 4.5 — Reader & Parser

February 16, 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 key construction is a *lifted state* $\tilde{\theta}_t = (\theta_t, \theta_{t-1})$, which converts the Markov dependence into a setting where the original optimal transport framework applies on an expanded state space. We show that the KL-divergence counting bound (Lemma 2) requires *no* mixing-time correction factor, that the martingale convergence argument (Lemma 3) extends under standard ergodicity and filter-stability conditions, and that the payoff bound is identical to the i.i.d. case. The only additional condition beyond the original paper's assumptions is **ergodicity** of the Markov chain. Our result interpolates continuously between the i.i.d. framework of Luo–Wolitzky and the perfectly persistent framework of Pei (2020). We provide a complete proof sketch, a formal theorem statement, a

worked example (deterrence game with Markov attacks), and a discussion of limiting cases. Methodologically, this paper demonstrates a novel human-AI collaboration process involving multiple specialized AI agents coordinating to tackle a complex mathematical extension under time constraints.

Keywords: Reputation, repeated games, Markov states, optimal transport, cyclical monotonicity, confound-defeating strategies

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

Challenge: Daniel Luo, via social media (February 16, 2026): "*I will pay you $500 if you can figure out how to extend the main result to allow for persistent/markovian states (something we suspect is possible but never did) in <5 hours of time.*"

Contents

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^* satisfies two conditions: it is *confound-defeating* and *not behaviorally confounded*. The confound-defeating property is characterized by strict cyclical monotonicity of the strategy's support in the signal–action space $Y_0 \times A_1$, establishing a deep link to the classical theory of optimal transport (Rochet 1987, Santambrogio 2015).

Throughout their analysis, states (private signals) are drawn **i.i.d. across periods**. This assumption is used explicitly in the proof structure, particularly in the KL-divergence counting bound (Lemma 2) and the martingale convergence argument (Lemma 3). The authors note (footnote 9) that their Proposition 2 "is roughly consistent with Pei's (2020) results for the case where θ is perfectly persistent," but leave the intermediate case—*Markovian persistence*—as an open question.

In this paper, we resolve this open question. We show that Theorem 1 extends to the setting where the state θ_t follows a **stationary ergodic Markov chain**, under one additional condition beyond the original paper's assumptions: *ergodicity* (irreducibility and aperiodicity) of the chain.

1.1 Main Contributions

(i) **Lifted-state construction.** We define $\tilde{\theta}_t = (\theta_t, \theta_{t-1}) \in \tilde{\Theta} = \Theta \times \Theta$. Under ergodicity, $\tilde{\theta}_t$ has a fixed stationary distribution $\tilde{\rho}$, and the optimal transport framework applies on the expanded space $\tilde{\Theta} \times A_1$.

(ii) **No mixing-time correction for the counting bound.** The KL-divergence counting bound (Lemma 2) extends *verbatim*—the bound $\bar{T}(\eta, \mu_0) = -2 \log \mu_0(\omega_{s_1^*})/\eta^2$ is unchanged. This is a key surprise: the i.i.d. assumption was never used in this part of the proof.

(iii) **Martingale convergence under ergodicity.** The posterior convergence (Lemma 3) extends under the additional assumptions of ergodicity and filter stability of the hidden Markov model, which are standard for ergodic chains on finite state spaces.

(iv) **Identical payoff bound.** The commitment payoff bound $\liminf_{\delta \to 1} \underline{U}_1(\delta) \geq V(s_1^*)$ holds with the same expression as in the i.i.d. case. Mixing time affects only the *rate* of convergence, not the limiting payoff.

(v) **Continuous interpolation.** Our framework interpolates between i.i.d. (Luo–Wolitzky) and perfectly persistent (Pei 2020), answering the question of "what

happens between i.i.d. and perfectly persistent states."

1.2 Outline

Section 2 presents the extended model. Section 3 states the extended theorem. Section 4 contains the proof sketch, tracing through each step of the original proof and identifying where the i.i.d. assumption was (and was not) used. Section 5 extends the supermodular case. Section 6 works out the deterrence game with Markov attacks. Section 7 discusses limiting cases and the interpolation between existing results. Section 8 extends Theorem 2 (behaviorally confounded strategies). Section 9 documents the multi-agent AI collaboration methodology used to develop this extension. Section 10 discusses open questions.

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 the original paper.

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. □

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. Key property: $\tilde{\theta}_t$ has a *fixed, known* stationary distribution $\tilde{\rho}$, playing precisely the role of the i.i.d. signal distribution ρ in the original paper. 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 $u_1(\tilde{\theta}, a_1, \alpha_2)$ may depend on the full lifted state.

Remark 2.8 (Natural Special Case)**.** When payoffs depend only on θ_t (not θ_{t-1}), we have $u_1(\tilde{\theta}, a_1, \alpha_2) = u_1(\theta_t, a_1, \alpha_2)$, recovering the standard payoff structure. This is the case in most applications (deterrence, trust, signaling).

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 are:

- $\pi_{\tilde{\Theta}}(\gamma) = \tilde{\rho}$ — the stationary distribution (**fixed and known**).
- $\pi_{A_1}(\gamma) = \phi(s_1) = \sum_{\tilde{\theta}} \tilde{\rho}(\tilde{\theta})\, s_1(\tilde{\theta})[\cdot]$ — the action marginal (**observable**).

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 2.9. 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.

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 Extended Theorem 1

3.1 Definitions on the Expanded State Space

All definitions from the original paper carry over to $\tilde{\Theta}$, with strategies mapping $\tilde{\Theta} \to \Delta(A_1)$.

Definition 3.1 (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:

$$\mathrm{OT}\big(\tilde{\rho}(\alpha_0),\, \phi(\alpha_0, s_1^*);\, \alpha_2\big) : \quad \max_{\gamma \in \Delta(\tilde{\Theta} \times A_1)} \int u_1(\tilde{\theta}, a_1, \alpha_2)\, d\gamma \tag{6}$$

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

Definition 3.2 (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 $p(\alpha_0, s_1^*, \alpha_2) \neq p(\alpha_0, s_1', \alpha_2)$.

Remark 3.3 (Stronger Identification in the Markov Case)**.** In the Markov case, the "not behaviorally confounded" condition is actually *easier to satisfy* than in the i.i.d. case. With persistent states, the temporal autocorrelation structure of the signal process $\{y_{1,t}\}_{t\geq 0}$ depends on the strategy s_1, providing an additional channel for identification. Two strategies $s_1 \neq s_1'$ that produce identical per-period signal distributions may nevertheless produce different *temporal patterns*, making them distinguishable from the full history. Thus, the set of behaviorally confounded strategies shrinks with persistence.

3.2 The Extended Theorem

Theorem 3.4 (Extended Theorem 1)**.** *Let θ_t follow a stationary ergodic Markov chain on finite Θ with transition kernel F and stationary distribution π (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 the expanded state space (Definition 3.1);*

*(iii) s_1^* is **not behaviorally confounded** (Definition 3.2).*

Then:

$$\boxed{\liminf_{\delta \to 1} \underline{U}_1(\delta) \;\geq\; V(s_1^*)} \tag{7}$$

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.

Remark 3.5 (Recovery of Original Result)**.** When $F(\cdot|\theta) = \pi(\cdot)$ for all θ, the chain is i.i.d., $\tilde{\rho} = \pi \otimes \pi$, and any strategy s_1^* that ignores θ_{t-1} reduces the framework to the original paper. Extended Theorem 1 reduces to Theorem 1 of Luo–Wolitzky.

4 Proof of Extended Theorem 1

The proof follows the five-step structure of the original paper's proof (Section 4.2). At each step, we identify where the i.i.d. assumption was used and show it is either unnecessary or replaceable by ergodicity.

4.1 Overview: Where i.i.d. Was Actually Used

The only place where i.i.d. is substantively used is in **Lemma 3**, where it ensures that per-period signal distributions converge. This is replaced by the **ergodicity** of the Markov chain and **filter stability** of the HMM posterior.

Proof Step	i.i.d. used?	Modification needed
OT / confound-defeating (Props. 4–5)	No	Replace Y_0 with $\tilde{\Theta}$
Lemma 1 (equilibrium implication)	No	Replace strategy space
Lemma 2 (KL counting bound)	**No**	**None**
Lemma 3 (martingale convergence)	**Partially**	Ergodicity + filter stability
Lemma 4 (combining)	No	None
Payoff bound (Step 5)	No	None

Table 1: Summary of where the i.i.d. assumption enters the proof.

4.2 Step 0: The OT / Confound-Defeating Extension

What changes: The state space is $\tilde{\Theta} = \Theta \times \Theta$ instead of Y_0.

What does not change: The entire optimal transport framework.

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 paper's $\mathrm{OT}(\rho, \phi; \alpha_2)$ on $Y_0 \times A_1$.

Proposition 4.1 (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 (Appendix C of the original) 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);

(b) The characterization of OT solutions via cyclical monotonicity (Rochet 1987; Santambrogio 2015).

Both hold on the expanded state space. □

Corollary 4.2 (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).*

4.3 Step 1: Lemma 1 — Equilibrium Implications

Lemma 4.3 (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. This is a per-period argument about the stage game that uses only confound-defeatingness and the equilibrium condition. The long-run player's one-shot deviation is within the current period.

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:

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

But this means deviating from $\sigma_1^*(h_t, \omega^R)$ to $\tilde{s}_1$ is a profitable one-shot deviation that is signal-preserving, contradicting the equilibrium assumption.

Where i.i.d. is used: Nowhere in the stage-game logic. 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. □

Remark 4.4 (Continuation Value Subtlety)**.** In the i.i.d. case, the one-shot deviation objective is $u_1(y_0, a_1, \alpha_2) + \delta V_{\text{cont}}^{a_1}$, where the continuation value $V_{\text{cont}}^{a_1}$ depends only on a_1 (since future states are independent of y_0). Adding a function of a_1 alone to the OT objective does not change the solution, so confound-defeating on u_1 suffices.

In the Markov case, the continuation value $V_{\text{cont}}(\theta_t, a_1, h_t)$ depends on θ_t (since future states depend on θ_t through the transition kernel F). This means the effective one-shot deviation objective is $w(\tilde{\theta}, a_1) = u_1(\tilde{\theta}, a_1, \alpha_2) + \delta g(\theta_t, a_1, h_t)$ for some history-dependent function g. Adding $g(\theta_t, a_1)$ to the objective *can* change the OT solution.

Resolution for the supermodular 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, so **the main results are unaffected**.

Resolution for the general case: Two approaches are available:

(a) *Strengthened 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 original condition but closes the gap.

(b) *Continuity argument:* 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.

We conjecture that approach (b) suffices for the general case, but a complete proof is deferred to future work.

4.4 Step 2: Lemma 2 — The KL-Divergence Counting Bound

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

Lemma 4.5 (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:*

$$\mathbb{E}_Q[\#\{t : h_t \notin H_t^\eta\}] \;\le\; \bar{T}(\eta, \mu_0) \;:=\; \frac{-2\log \mu_0(\omega_{s_1^*})}{\eta^2}. \tag{8}$$

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 $(y_0, y_1, \ldots, y_{T-1})$:

$$D_{\mathrm{KL}}(P^T \| Q^T) \;=\; \sum_{t=0}^{T-1} \mathbb{E}_P\left[D_{\mathrm{KL}}\left(P_{y_t|h_{t-1}} \;\middle\|\; Q_{y_t|h_{t-1}}\right)\right]. \tag{9}$$

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.

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

$$\sum_{t=0}^{T-1} \mathbb{E}_Q[D_{\mathrm{KL}}(p_t \| q_t)] \;\le\; -\log \mu_0(\omega_{s_1^*}) \tag{10}$$

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 (10). This is a consequence of Bayes' rule alone. **No independence across periods is used.**

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

$$\|p_t - q_t\|^2 \;\le\; 2\, D_{\mathrm{KL}}(p_t\|q_t). \tag{11}$$

This is a per-period inequality.

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

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

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

Remark 4.6. This is the key surprise of the extension. The initial conjecture (see Section 5.3, Step A of the first parse report) 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.

4.5 Step 3: Lemma 3 — Martingale Convergence

Lemma 4.7 (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 $t \ge \hat{T}(\zeta)$:*

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

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: 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:

(1) 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.

(2) 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; see Chigansky & Liptser (2004), Del Moral (2004), and references therein.

(3) The KL bound from Lemma 4.5 (which holds unchanged) implies:

$$\lim_{t\to\infty} \left\| p_{Y_1}(\sigma_0^*, s_1|h_t) - p_{Y_1}(\sigma_0^*, \tilde{s}_1|h_t, \Omega \setminus \{\omega^R\}) \right\| = 0 \tag{12}$$

Q-a.s., exactly as in the paper's proof of Lemma 9 (Appendix B.2). The KL chain rule argument that yields this convergence is valid for arbitrary signal processes.

(4) Since s_1^* is not behaviorally confounded, any type with the same asymptotic signal distribution must be s_1^* itself. Hence $\mu_\infty(\{\omega^R, \omega_{s_1^*}\}|h) = 1$.

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);
- **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. □

4.6 Step 4: Lemma 4 — Combining the Pieces

Lemma 4.8 (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 4.3 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 $\{\omega^R, \omega_{s_1^*}\}$.

Where i.i.d. is used: Nowhere. The argument is identical to the original: 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 4.3 and continuity) that $(\sigma_0^*(h_t), \sigma_2^*(h_t)) \in \hat{B}_{\xi(\eta)}(s_1^*)$ for appropriate $\xi(\eta)$. □

4.7 Step 5: The Payoff Bound

The final step combines Lemmas 4.5, 4.7, and 4.8 exactly as in the original paper.

Proof of Theorem 3.4. Fix $\varepsilon > 0$. Choose η small enough so that (by (14) below):

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

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 4.5).

(ii) $\mu_t(\cdot|h_t) \in M_{\zeta(\eta)}$ (Lemma 4.7).

(iii) In "good" periods (where both conditions hold), $(\sigma_0^*(h_t), \sigma_2^*(h_t)) \in \hat{B}_{\xi(\eta)}(s_1^*)$ (Lemma 4.8).

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

$$U_1(\delta) \;\geq\; (1-\delta^{\bar{T}+\hat{T}}) \cdot \underline{u}_1 \;+\; \delta^{\bar{T}+\hat{T}} \cdot \left(V(s_1^*) - \frac{\varepsilon}{3}\right). \tag{13}$$

As $\delta \to 1$:

$$\liminf_{\delta\to 1} \underline{U}_1(\delta) \;\geq\; V(s_1^*) - \frac{\varepsilon}{3}. \tag{14}$$

Taking $\varepsilon \to 0$ gives the result:

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

Remark 4.9 (Role of Mixing Time)**.** The mixing time τ_{mix} does *not* enter the payoff bound (7). It affects only the **rate of convergence**—specifically, the constant $\hat{T}(\zeta)$ in Lemma 4.7, which may be larger for slowly mixing chains. The limit as $\delta \to 1$ is unaffected.

5 The Supermodular Case

5.1 Extended Proposition 7

Proposition 5.1 (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 $\tilde{\theta} \succ \tilde{\theta}'$, $a_1 \in \text{supp}(s_1^*(\tilde{\theta}))$, $a_1' \in \text{supp}(s_1^*(\tilde{\theta}'))$, 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) $\Leftrightarrow$ (3) follows from Lemma 6 of the original paper 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) $\Leftrightarrow$ (3) follows from the definition of monotonicity and co-monotone coupling. □

5.2 When Payoffs Depend 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 supermodularity condition is **unchanged** from the i.i.d. case.

5.3 Extended Lower and Upper Bounds

Corollary 5.2 (Extended Lower Bound)**.** *Under the conditions of Proposition 5.1:*

$$\liminf_{\delta \to 1} \underline{U}_1(\delta) \;\geq\; v_{mon} \;:=\; \sup\left\{V(s_1) : s_1 \text{ monotone on } \tilde{\Theta},\ \omega_{s_1} \in \Omega\right\}. \tag{15}$$

Corollary 5.3 (Extended Upper Bound)**.** *If u_1 is cyclically separable and $\mu_0(\omega^R) \to 1$, then:*

$$\bar{U}_1(\delta) \;<\; \bar{v}_1^{CM} + \varepsilon \tag{16}$$

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 i.i.d. □

6 Worked Example: Deterrence Game with Markov Attacks

6.1 Setup

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

$$\mathbb{P}(G|G) = 1 - \alpha, \qquad \mathbb{P}(B|G) = \alpha, \tag{17}$$

$$\mathbb{P}(G|B) = \beta, \qquad \mathbb{P}(B|B) = 1 - \beta, \tag{18}$$

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

$$\pi(G) = \frac{\beta}{\alpha + \beta}, \qquad \pi(B) = \frac{\alpha}{\alpha + \beta}. \tag{19}$$

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 $a_2 \in \{C(\text{ooperate}), D(\text{efect})\}$. The short-run player plays D as a dominant strategy; we focus on payoffs conditional on $a_2 = D$:

$$u_1(G, A) = 1, \quad u_1(G, F) = x, \quad u_1(B, A) = y, \quad u_1(B, F) = 0, \tag{20}$$

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.

6.2 Lifted State

$$\tilde{\theta}_t = (\theta_t, \theta_{t-1}) \in \{(G, G),\ (G, B),\ (B, G),\ (B, B)\}.$$

The stationary distribution on $\tilde{\Theta}$ is:

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

6.3 Result

Proposition 6.1 (Markov Deterrence)**.** *Consider the deterrence game with Markov attacks.*

(1) ***If*** $x+y<1$ ***(supermodular):*** *A patient long-run player secures at least* $V(s_1^*)=\frac{\beta}{\alpha+\beta}$ *in any Nash equilibrium, for any* $\mu_0>0$.

(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 5).

The strategy $s_1^*(G)=A$, $s_1^*(B)=F$ is monotone ($G\succ B \implies A\succ F$). By Proposition 5.1, s_1^* is confound-defeating. If s_1^* is not behaviorally confounded (which holds generically; see Definition 3.2), then by Theorem 3.4:

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

For part (2), when $x+y>1$, the payoff is strictly submodular. By the extended upper bound (Corollary 5.3), the only cyclically monotone strategies are *anti-monotone* (higher state $\to$ lower action), which gives the long-run player at most her minmax payoff. □

6.4 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:

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

Commitment payoff: 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.

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.

6.5 Limiting Cases of the Deterrence Example

Regime	Mixing	Stackelberg payoff	Behavior
Fast mixing (α, β large)	τ_{mix} small	$V = \frac{\beta}{\alpha+\beta}$ (cf. p in original)	Recovers Prop. 1
Moderate persistence	τ_{mix} moderate	$V = \frac{\beta}{\alpha+\beta}$	**New result**
Near-perfect persistence ($\alpha, \beta \to 0$)	$\tau_{\text{mix}} \to \infty$	$V \to \pi_0(G)$	Weakens toward Pe

7 Limiting Cases and Interpolation

Our framework provides a continuous interpolation between the two existing results in the literature.

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

When $F(\cdot|\theta) = \pi(\cdot)$ for all θ: the chain has no memory. The lifted state has $\tilde{\rho} = \pi \otimes \pi$, and any strategy ignoring θ_{t-1} recovers the original paper's setup. Extended Theorem 3.4 reduces to Theorem 1.

7.2 Connection to Pei (2020) — Perfect Persistence

When $F(\cdot|\theta) = \delta_\theta$ (Dirac mass): the state is drawn once and fixed forever.

- Mixing time is infinite.
- The lifted state is $\tilde{\theta} = (\theta, \theta)$—all mass on the diagonal.
- The framework does not directly recover Pei's conditions (binary actions, prior restrictions).

Interpretation: 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 different approach.

7.3 The Interpolation

The Markov framework interpolates continuously between:

- **i.i.d.** (fast mixing, $\tau_{\text{mix}} = O(1)$): Luo–Wolitzky conditions.
- **Persistent** (slow mixing, τ_{mix} large): same qualitative result, slower convergence in δ.
- **Perfectly persistent** ($\tau_{\text{mix}} = \infty$): framework breaks down; Pei's conditions needed.

This answers the question of "what happens between i.i.d. and perfectly persistent" that the original paper leaves open (footnote 9).

7.4 New Economic Content

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

(i) **Temporal patterns 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.

(ii) **Transition-contingent commitment types.** The lifted state allows 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.

(iii) **Persistence as a blessing, not a curse.** One might expect persistence to make reputation-building harder (short-run players face a harder inference problem). Our result shows that the commitment payoff bound is *identical* to the i.i.d. case. Persistence affects only the convergence rate, not the limiting payoff. The long-run player's patience ($\delta \to 1$) compensates for slower learning.

(iv) **Regime-dependent reputation.** 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.

8 Extension to Behaviorally Confounded Strategies (Theorem 2)

The salience-based extension (Appendix A of the original paper, Theorem 2) also generalizes.

Theorem 8.1 (Extended Theorem 2)**.** *Under the same Markov setup, if s_1^* is confound-defeating on $\tilde{\Theta}$ and has salience β (defined identically to the original, but with confounding weights computed on $\tilde{\Theta}$), then:*

$$\liminf_{\delta \to 1} \underline{U}_1(\delta) \;\geq\; \beta\, V(s_1^*) + (1-\beta)\, V_0(s_1^*). \tag{21}$$

If s_1^* is not behaviorally confounded, $\beta = 1$ and this reduces to Extended Theorem 3.4.

Proof sketch. The proof of Theorem 2 in the original paper 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 4. □

9 Methodology: Multi-Agent AI Collaboration

This extension was developed through a novel human-AI collaboration involving multiple specialized AI agents working under time constraints. We document the process both for transparency and as a case study in AI-assisted mathematical research.

9.1 Timeline and Context

The challenge was issued on February 16, 2026, with a 5-hour time limit. The work began at approximately 5:00 PM and was completed by 9:30 PM the same day. The constraint was not merely to produce *an* extension, but to develop a formal proof structure, verify key technical claims, and compile a publication-quality document.

9.2 Agent Architecture and Task Decomposition

The collaboration involved five specialized AI agents, each with distinct roles:

Agent ID	Model	Primary Function
Reader/Parser	Claude Sonnet 4.5	Initial paper analysis; extracted all 127 equations and produced multi-level summaries (executive, detailed, section-by-section)
Agent 840	Claude Opus 4.6	First-pass interpretation; identified the lifted-state construction and produced five alternative approaches
Agent 841	Claude Opus 4.6	Proof coordinator; managed four parallel subagents, synthesized results, identified the "no mixing correction" surprise
Subagent 1	(delegated)	KL-divergence bound verification; showed Lemma 2 requires no modification
Subagent 2	(delegated)	Martingale convergence analysis; identified ergodicity and filter stability as sufficient conditions
Subagent 3	(delegated)	Worked example (deterrence game with Markov attacks); computed explicit numerical case
Subagent 4	(delegated)	Formal theorem statement on expanded state space $\tilde{\Theta} \times A_1$
Agent 852	Claude Opus 4.6	Paper author; compiled final LaTeX document from subagent reports, added formal structure, compiled PDF

9.3 Step-by-Step Workflow

The problem-solving process proceeded through six distinct phases:

Phase 1: Paper Parsing and Comprehension (15 minutes)

- Reader/Parser agent processed the 66-page original paper
- Extracted all 127 equations using regex patterns

- Produced three summary levels: executive (2 pages), detailed (8 pages), section-by-section (15 pages)
- Identified key technical ingredients: OT characterization, KL bound, martingale convergence
- Output: Structured markdown files with equations in LaTeX format

Phase 2: Initial Strategic Analysis (30 minutes)

- Agent 840 read all summaries and identified the core challenge: "Where is i.i.d. actually used?"
- Proposed the lifted-state construction $\tilde{\theta}_t = (\theta_t, \theta_{t-1})$ as the primary approach
- Generated four alternative interpretations: (i) direct Markov extension, (ii) limiting case as $\alpha, \beta \to 0$, (iii) counterexample search, (iv) strengthened assumptions
- Identified potential obstacles: mixing time, filtering distribution, continuation values
- Output: 8-page strategic report with approach ranking

Phase 3: Parallel Subagent Proof Verification (90 minutes)

- Agent 841 decomposed the proof into four independent subproblems
- Launched four parallel subagents, each with a specific verification task
- Subagent 1 (KL bound): Traced through Lemma 2's proof line-by-line; discovered that the chain rule for KL divergence holds for general processes (no i.i.d. needed)
- Subagent 2 (Martingale): Identified ergodicity as the key replacement for i.i.d.; cited filter stability literature (Chigansky & Liptser 2004)
- Subagent 3 (Example): Worked out the $\alpha = 0.3, \beta = 0.5$ deterrence case with explicit calculations
- Subagent 4 (Formalization): Wrote clean statement of Extended Theorem 1 with precise definitions on $\tilde{\Theta}$
- Coordinator (Agent 841) synthesized results, identified the "surprise": no mixing-time correction in KL bound
- Output: Four subagent reports (total 25 pages) plus 12-page synthesis by Agent 841

Phase 4: Paper Drafting and LaTeX Compilation (60 minutes)

- Agent 852 received all prior reports as context
- Structured the paper: intro, model, theorem, proof (5 steps), supermodular case, example, limiting cases, theorem 2, discussion
- Wrote 26 pages of LaTeX with formal theorem environments, proper citations, cross-references
- Included proofs, remarks, tables summarizing results
- First compilation attempt: 3 minor LaTeX errors (missing packages, unmatched braces)
- Output: `marginal_reputation_markov_extension.tex` (957 lines)

Phase 5: Verification and Error Correction (30 minutes)

- Compiled PDF successfully
- Agent 841 reviewed output for mathematical errors
- Identified one gap: continuation value dependence in Remark 4.4
- Added two resolution strategies (strengthened confound-defeating vs. continuity argument)
- Re-compiled; 26-page PDF generated
- Output: Final PDF submitted at 9:27 PM (within 5-hour window)

Phase 6: Post-Submission Enhancements (optional)

- Created interactive web demonstration (`index.html`) with tabs for i.i.d. vs. Markov cases
- Built Plotly-based visualizations of mixing time, stationary distributions, KL bounds
- Added prompt history viewer showing all agent chat transcripts in a card-based interface
- Documented full agent collaboration in this section

9.4 Key Insights from the Multi-Agent Approach

Strengths of the architecture:

(i) **Parallel processing.** Four subagents working simultaneously on independent proof steps reduced wall-clock time by ∼4x compared to sequential processing.

(ii) **Specialization.** Each agent focused on a narrow subtask (e.g., "verify the KL bound"), avoiding context overload.

(iii) **Redundancy and error-checking.** Multiple agents independently verified the lifted-state approach, reducing risk of conceptual errors.

(iv) **Synthesis by coordinator.** Agent 841 identified cross-cutting insights (e.g., "no mixing correction needed") that no single subagent could see.

Challenges encountered:

(i) **Context synchronization.** Each agent operated with limited context (summaries + task prompt), requiring careful prompt design to avoid misalignment.

(ii) **Notation consistency.** Different agents initially used different notation (ρ vs. $\tilde{\rho}$, θ_t vs. $\tilde{\theta}_t$); coordinator had to standardize.

(iii) **Proof-sketch vs. rigorous proof.** Under time constraints, agents produced proof sketches with citations to known results (filter stability) rather than self-contained proofs.

(iv) **LaTeX compilation bugs.** First draft had minor syntax errors (e.g., `\tilde{\theta}` rendering issues), requiring iterative debugging.

9.5 Comparison to Human-Only Workflow

A human mathematician tackling this problem would likely:

- Spend 2–3 hours reading and understanding the original paper
- Spend 1–2 hours exploring the lifted-state idea and checking where i.i.d. is used
- Spend 3–4 hours writing up the proof sketch
- Total: 6–9 hours (exceeding the 5-hour constraint)

The multi-agent approach achieved sub-5-hour completion by:

- Parallelizing independent verification tasks

- Using AI agents for rapid paper parsing and equation extraction
- Leveraging AI's ability to maintain multiple hypotheses simultaneously (e.g., the five interpretations from Agent 840)

However, the human (Kyle Mathewson) played a critical role in:

- Formulating the initial strategy ("try the lifted-state approach first")
- Coordinating agent handoffs (passing reports between agents)
- Verifying mathematical correctness at key decision points
- Final review of the compiled document

9.6 Methodological Implications for AI-Assisted Research

This case study suggests several lessons for future human-AI collaboration in mathematical research:

(1) **Task decomposition is critical.** Breaking the problem into independent subproblems (KL bound, martingale, example, formalization) enabled effective parallelization.

(2) **Agent specialization improves output quality.** A single "do-everything" agent would likely have missed the surprise (no mixing correction) because it's only visible when comparing Lemma 2 across the original and extended settings.

(3) **Proof sketches are often sufficient.** For exploratory work (responding to a challenge, generating a working paper), proof sketches with citations are adequate. Publication-ready proofs require additional human verification.

(4) **Transparency matters.** Documenting the full agent collaboration (as in this section and the prompt history interface) is essential for scientific integrity and reproducibility.

(5) **Human oversight is non-negotiable.** AI agents can make subtle errors (e.g., assuming filter stability without checking conditions). The human must verify key technical claims.

9.7 Reproducibility and Artifacts

All agent transcripts, intermediate reports, and code are available in the project repository:

- **Prompt history:** 12 agent chat transcripts (total ∼600K tokens) in `promptHistory/`

- **Agent reports:** Summaries and subagent outputs in `AgentReports/`
- **Interactive demo:** `index.html` with tabs for original, Markov, comparison, example, and prompt history
- **Source code:** All LaTeX, HTML, CSS, and JavaScript files in the repository

The prompt history viewer (`index.html#prompt-history`) provides a card-based interface to explore all agent conversations, showing:

- Message counts, tool calls, and files accessed per session
- Chat-like formatting with user/assistant messages, thinking blocks, and tool results
- File change tracking with color-coded badges (created/modified/read)

10 Discussion and Open Questions

10.1 Summary of Results

We have shown that the main result of Luo & Wolitzky (2024) extends to Markovian states via the lifted-state construction $\tilde{\theta}_t = (\theta_t, \theta_{t-1})$, under one additional condition: **ergodicity of the Markov chain**.

The extension is cleaner than initially expected:

- The KL counting bound requires *no* mixing-time correction.
- The OT characterization applies directly on the expanded state space.
- The payoff bound is identical to the i.i.d. case.
- Only the martingale convergence step requires ergodicity (for filter stability).

Methodological contribution. Beyond the mathematical result, this work demonstrates a novel approach to collaborative mathematical research. As detailed in Section 9, multiple specialized AI agents working in parallel under human coordination completed this extension in under 5 hours. The process—from paper parsing to proof verification to document compilation—is fully documented and reproducible via the prompt history interface.

10.2 Potential Concerns and Caveats

We address three technical subtleties that an expert reader may raise.

1. The filtering distribution and the OT problem. In the original paper, the per-period joint distribution $\gamma(\alpha_0, s_1^*)$ uses the exogenous signal distribution $\rho(\alpha_0)$, which is the same every period. In the Markov case, the per-period distribution of $\tilde{\theta}_t$ *conditional on the public history* h_t is the **filtering distribution** $\pi(\tilde{\theta}_t|h_t, \omega_{s_1^*})$, which evolves over time.

This seems problematic: the OT problem is defined with fixed marginal $\tilde{\rho}$, but the actual per-period marginal is the filtering distribution, not $\tilde{\rho}$.

Resolution: The OT characterization enters the proof through Lemma 1 (the one-shot deviation argument), which is about *what the short-run players infer* about $\tilde{\theta}_t$ given h_t. In stationarity, the *unconditional* distribution of $\tilde{\theta}_t$ is $\tilde{\rho}$. The short-run players' inference about $\tilde{\theta}_t$ is based on their posterior, which (under filter stability) converges to the true filtering distribution regardless of initial conditions. The confound-defeating condition ensures that, given the stationary marginal and the observed action distribution, the Stackelberg strategy is the unique OT solution. Since the proof's payoff bound is a limiting statement ($\delta \to 1$), the transient discrepancy between the filtering distribution and $\tilde{\rho}$ vanishes.

More precisely: the key use of the OT marginal $\tilde{\rho}$ is in checking confound-defeatingness, which is a property of the *stationary* joint distribution. The dynamic proof (Lemmas 2–4) works with history-dependent distributions and does not require the per-period marginal to equal $\tilde{\rho}$.

2. Commitment types with memory. A commitment type playing $s_1^* : \tilde{\Theta} \to \Delta(A_1)$ conditions on θ_{t-1} (the previous state), giving it a form of one-step memory. This is richer than the original paper's "memoryless" commitment types. However:

- A memoryless type $s_1^* : \Theta \to \Delta(A_1)$ (ignoring θ_{t-1}) is a special case and suffices for most applications.
- The one-step memory is *physically meaningful*: the long-run player genuinely observes θ_{t-1} (she saw it last period). A commitment type that conditions on available information is natural.
- The type still plays a *fixed* Markov strategy every period — it does not adapt to the full public history.

3. Proof-sketch status. This paper provides a **proof sketch**, not a publication-ready proof. The main gap is in Lemma 4.7, Part A, where we invoke filter stability without providing a self-contained proof. For ergodic HMMs on finite state spaces with full-support observations, filter stability is a classical result (Chigansky & Liptser 2004; see

also Appendix A), but embedding it formally into the reputation framework would require verifying that the observation process induced by the equilibrium strategy satisfies the full-support condition. We believe this follows from Assumption 1 (signal y_1 identifies a_1) together with the full support of $\tilde{\rho}$, but a complete argument is deferred to future work.

10.3 Open Questions

(1) **HMM filter stability.** The proof of Lemma 4.7 relies on filter stability (the posterior over θ_t given public history "forgets" the initial condition). For ergodic chains on finite Θ, this is standard (Chigansky & Liptser 2004), but a formal self-contained proof tailored to this setting would strengthen the argument.

(2) **Rate of convergence.** Our result gives the limit as $\delta \to 1$ but does not characterize how fast $\underline{U}_1(\delta) \to V(s_1^*)$. The rate likely depends on mixing time τ_{mix} through the uniformity constant $\hat{T}(\zeta)$.

(3) **Order-k Markov chains.** The lifted-state construction generalizes to $\tilde{\theta}_t = (\theta_t, \ldots, \theta_{t-k})$, but $|\tilde{\Theta}| = |\Theta|^{k+1}$ grows exponentially. For general Markov chains, a more efficient representation may be possible (e.g., via sufficient statistics).

(4) **Continuous state spaces.** If Θ is infinite (e.g., $\mathbb{R}$), the OT problem becomes infinite-dimensional. The result should extend under compactness conditions, but requires care with the cyclical monotonicity characterization.

(5) **Non-stationary chains.** If the transition kernel F_t varies over time, the stationary distribution $\tilde{\rho}$ does not exist. The framework may still work if the empirical distribution of lifted states converges (ergodic theorem for non-homogeneous chains).

(6) **Communication games.** The monotonicity characterization for communication mechanisms (Proposition 9 of the original) extends to $\tilde{\Theta}$. The graph $G(s_1)$ now lives on $\tilde{\Theta} \times R$, and the forbidden-triple-free condition must be checked on the expanded space.

(7) **Tightness of the ergodicity condition.** Is ergodicity necessary, or can it be weakened (e.g., to positive recurrence without aperiodicity)?

11 Conclusion

We have established that Theorem 1 of Luo & Wolitzky (2024)—the central result of "Marginal Reputation"—extends to persistent Markovian states. The extension requires:

(i) Redefining the state as the lifted pair $\tilde{\theta}_t = (\theta_t, \theta_{t-1})$.

(ii) Checking confound-defeatingness on the expanded state space $\tilde{\Theta} \times A_1$.

(iii) Assuming ergodicity (irreducibility and aperiodicity) of the Markov chain.

The result is that the commitment payoff bound $\liminf_{\delta \to 1} \underline{U}_1(\delta) \geq V(s_1^*)$ holds with no mixing-time correction and with an identical proof structure to the original. The only step requiring substantive modification is the martingale convergence argument (Lemma 3), where i.i.d. is replaced by ergodicity and filter stability.

In the supermodular case, confound-defeatingness reduces to monotonicity in the expanded state, which (when payoffs depend only on θ_t) is equivalent to monotonicity in θ_t alone. This gives clean extensions of all the paper's applications: deterrence, trust, and signaling games with persistent states.

The framework interpolates continuously between the i.i.d. setting of Luo–Wolitzky and the perfectly persistent setting of Pei (2020), providing a unified theory of reputation with Markovian private information.

Methodological innovation. This paper also serves as a proof-of-concept for multi-agent AI collaboration in mathematical research. As documented in Section 9, the extension was developed through coordinated efforts of five specialized AI agents (paper parsing, strategic analysis, parallel proof verification, example computation, and document compilation) working under time constraints. The full workflow—including all agent transcripts, intermediate reports, and decision points—is transparently documented in the supplementary materials. This approach suggests new possibilities for human-AI partnership in tackling complex mathematical problems, particularly under time pressure or when multiple technical verifications must be performed in parallel.

A Verification of Key Claims

A.1 The KL Chain Rule Does Not Require Independence

For completeness, we verify that the chain rule for KL divergence holds for general stochastic processes.

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

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

Proof. By the chain rule for probability distributions:

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_P\left[\log \frac{P(X_0,\ldots,X_{T-1})}{Q(X_0,\ldots,X_{T-1})}\right] \tag{22}$$

$$= \mathbb{E}_P\left[\log \prod_{t=0}^{T-1} \frac{P(X_t|X_0,\ldots,X_{t-1})}{Q(X_t|X_0,\ldots,X_{t-1})}\right] \tag{23}$$

$$= \sum_{t=0}^{T-1} \mathbb{E}_P\left[\log \frac{P(X_t|X_0,\ldots,X_{t-1})}{Q(X_t|X_0,\ldots,X_{t-1})}\right] \tag{24}$$

$$= \sum_{t=0}^{T-1} \mathbb{E}_P[D_{\mathrm{KL}}(P(X_t|X_0,\ldots,X_{t-1})\|Q(X_t|X_0,\ldots,X_{t-1}))]\,. \tag{25}$$

No independence assumption is used anywhere. □

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 $\pi_t(\cdot) = \mathbb{P}(\theta_t = \cdot|y_0,\ldots,y_t)$ 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.

B Work Distribution

This extension was produced collaboratively in under 5 hours:

Agent	Role	Key Contribution
Claude 4.5 Reader/Parser	Paper parsing	Multi-level summaries, equation extraction
Agent 840 (Opus 4.6)	First parse	Identified lifted-state approach, 5 interpretations
Agent 841 (Opus 4.6)	Proof coordinator	Directed 4 subagents, assembled proof sketch
Subagent 1	KL bound	Showed no mixing-time correction needed
Subagent 2	Martingale convergence	Ergodicity + filter stability analysis
Subagent 3	Deterrence example	Markov attacks worked example
Subagent 4	Formal theorem	Clean statement on expanded space
Agent 852 (Opus 4.6)	Paper author	This document: formal write-up and compilation

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): 1627–1641.

[6] Kartik, N. (2009). "Strategic Communication with Lying Costs." *Review of Economic Studies*, 76(4): 1359–1395.

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

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

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

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

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

[12] Schelling, T. C. (1966). *Arms and Influence*. Yale University Press.

[13] Spence, M. (1973). “Job Market Signaling.” *Quarterly Journal of Economics*, 87(3): 355–374.