Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Time-Local Conditional Equilibria for N-Player Stochastic Differential Games with Exploratory Policies

Jisheng Liu

jsliu23@m.fudan.edu.cn

Fudan University

Research Article

Keywords: Stochastic differential games, time-local conditional equilibrium, entropy regularization, conditional distributions, potential games, q-learning

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11534523/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Time-Local Conditional Equilibria for N-Player Stochastic Differential Games with Exploratory Policies

Jisheng Liu[1*]

[1]School of Mathematical Sciences, Fudan University, Shanghai, China.

Corresponding author(s). E-mail(s): jsliu23@m.fudan.edu.cn;

Abstract

We study entropy-regularized responses in $\boldsymbol{N}$-player stochastic differential games under a prediction rule that holds the opponents' currently observed actions fixed over the continuation horizon. Each player's control problem produces a Gibbs response conditional on that action profile. We ask when these responses are the full conditionals of one joint action law, called a time-local conditional equilibrium. For smooth positive responses on a simply connected action domain, compatibility is characterized by symmetry of the entropy-scaled Hamiltonian cross-partials and normalizability of the associated potential. A decomposition identifies how continuation-value sensitivity enters this condition. On compact action intervals, a coordinate-path construction yields approximate joint laws whose conditional Kullback–Leibler errors are quadratic in a uniform compatibility gap. Under admissibility and verification conditions, these errors also bound the value loss in each frozen control problem. We obtain stationary discounted counterparts and specialize policy-improvement and martingale methods to estimate the conditional response maps. A two-player linear-quadratic example provides explicit compatible Gaussian laws and near-compatible error calculations, followed by coefficient regression with the value known.

Keywords: Stochastic differential games, time-local conditional equilibrium, entropy regularization, conditional distributions, potential games, $\boldsymbol{q}$-learning

MSC Classification: 91A15 , 93E20 , 60H10 , 49L20 , 91A10

1 Introduction

Continuous-time reinforcement learning for diffusion models uses randomized policies to represent exploration. Wang et al. (2020) formulate this randomization through stochastic relaxed controls: a policy specifies a distribution over actions, the state generator averages the drift and covariance over that distribution, and entropy regularizes the reward. The resulting exploratory Hamilton–Jacobi–Bellman (HJB) equation has a Gibbs optimizer. Tang et al. (2022) establish well-posedness and regularity for a class of exploratory HJB equations and prove convergence of the exploratory value to the classical-control value as exploration vanishes.

Learning these policies also requires evaluating and improving them from data. Jia and Zhou (2022a) characterize policy evaluation through martingale conditions and derive estimation equations from martingale orthogonality. Jia and Zhou (2022b) develop policy-gradient and actor–critic methods within this framework. Jia and Zhou (2023) introduce the continuous-time q-function, related to the instantaneous advantage rate and the Hamiltonian, and obtain policy-improvement and on- and off-policy martingale characterizations. Their erratum (Jia and Zhou 2025) formulates the sampled-data results using actions drawn on a time grid. Subsequent convergence and regret analyses include Huang et al. (2025); Tang and Zhou (in press).

Entropy regularization also appears in finite-game learning dynamics (Mertikopoulos and Sandholm 2016), mean field games (Guo et al. 2022), and discrete-time zero-sum stochastic games (Savas et al. 2019). In continuous time, Firoozi and Jaimungal (2022) obtain exploratory LQG mean field controls and finite-population approximate Nash equilibria, while Sun and Jia (2023) derive Gaussian saddle policies for exploratory LQ two-person zero-sum stochastic differential games. These equilibrium problems evaluate a player's response against the opponents' continuation behavior.

For a finite N-player game, a Nash best response likewise depends on the opponents' continuation policies. Entropy regularization gives a Gibbs formula for the optimizing action density, but the value derivatives appearing in its Hamiltonian still depend on those future policies. Specifying the opponents' current actions alone therefore requires an additional prediction rule before the player's dynamic control problem is determined.

We study the persistence forecast that extends the observed opponent actions unchanged over the remaining horizon. A player knows the physical dynamics and its own objective, observes the joint state and the current opponent actions, and uses this forecast without estimating an opponent response rule from past actions. Current-action prediction is familiar from best-reply adjustment in Cournot models (Hommes et al. 2018, Section 2); here it is an explicit behavioral assumption for the continuation-control problem. The states continue to evolve, and the player optimizes its own controls dynamically.

At each state-time pair, this produces one conditional Gibbs response for each player. We ask whether all these responses are the full conditionals of a common joint action law. Such a law, called a *time-local conditional equilibrium*, is preserved when any one player redraws its action according to its response. The consistency is

imposed at the given state and time; the underlying control assessment uses the full continuation horizon.

Full conditional densities need not admit a common joint law (Arnold and Press 1989; Arnold et al. 1992; Chen 2010). For smooth positive Gibbs responses, we characterize compatibility by a conservative field of entropy-scaled Hamiltonian action gradients and a normalizable potential. On simply connected action domains, this gives a cross-partial criterion and reconstructs the joint density. The criterion retains both direct action interactions and the sensitivity of the frozen continuation values to opponent actions; we separate these terms explicitly. The potential condition connects the construction with exact potential games (Monderer and Shapley 1996) and with Gibbs models of strategic interaction (Blume 1993; Blume and Durlauf 2003; Alós-Ferrer and Netzer 2010).

Correlated-equilibrium incentives are evaluated conditional on a player's own recommendation (Aumann 1974); regularized variants include Ortiz et al. (2007); Černý et al. (2022); Li et al. (2024). Here the opponents' current actions are observed, and the player's response is specified conditional on that entire profile.

Conditional-ratio reconstruction and approximation of incompatible conditional models have been studied in Lowd (2012); Muré (2019); Torroba Hennigen and Kim (2023). For the present smooth responses on compact action intervals, a uniform Hamiltonian cross-partial gap gives a coordinate-path joint construction with explicit conditional Kullback–Leibler (KL) bounds. The KL error is quadratic in the gap when the gap is small. The Gibbs variational identity and Itô's formula then bound the value loss in each frozen control problem, provided the constructed conditional policy is admissible and the verification conditions hold. We also obtain stationary discounted counterparts.

For estimation, we specialize the policy-improvement and martingale methods of Jia and Zhou (2022a, 2023, 2025) to control problems parametrized by frozen opponent commands. A two-player LQ benchmark gives explicit compatible Gaussian laws and near-compatible KL and value losses. A regression with the analytic value supplied checks the action-dependent coefficients.

Section 2 defines the game, prediction rule and equilibrium. Sections 3 and 4 give the compatibility and approximation results for finite and stationary discounted horizons. Section 5 develops the estimation equations, and Section 6 presents the LQ calculations and regression.

2 Model and Conditional Responses

We work on a filtered probability space $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\in[0,T]}, \mathbb{P})$ satisfying the usual conditions. There are N players, a finite horizon $T > 0$, a joint state $x \in \mathbb{R}^N$ observed by all players, and an action space $U \subset \mathbb{R}^m$ for each player. For $u = (u^1, \dots, u^N) \in U^N$, write u^{-i} for the opponents' action profile. Player i has discount rate $\beta^i \geq 0$ and exploration weight $\gamma^i > 0$.

Each player knows the physical state dynamics and its own reward criterion. The forecast of opponent actions is specified in Section 2.4.

We use ∂_x or ∇_x for the state gradient, and ∂_x^2 or D_x^2 for the Hessian. The exploratory performance and value are denoted by $\tilde{J}^i$ and $\tilde{V}^i$.

2.1 Controlled Dynamics

For an adapted control profile $\boldsymbol{\alpha}$, the joint state satisfies

$$dX_t^i = b^i(t, X_t, \boldsymbol{\alpha}_t)dt + \sigma^i(t, X_t, \boldsymbol{\alpha}_t)dW_t^i, \quad i = 1, \dots, N, \tag{1}$$

where $\boldsymbol{\alpha}_\cdot := (\alpha_\cdot^1, \dots, \alpha_\cdot^N)^\intercal$ is the $\mathbb{F}$-adapted, U^N-valued control vector, and $W^1, \dots, W^N$ are mutually independent one-dimensional standard Brownian motions. The coefficients $b^i : [0,T] \times \mathbb{R}^N \times U^N \to \mathbb{R}$ and $\sigma^i : [0,T] \times \mathbb{R}^N \times U^N \to \mathbb{R}$ are measurable in all variables. Write $W = (W^1, \dots, W^N)^\intercal$ and $\sigma = \mathrm{diag}(\sigma^1, \dots, \sigma^N)$ for the vector Brownian motion and diffusion matrix.

Equivalently, for $s \in [t,T]$,

$$X_s^{t,x,\boldsymbol{\alpha},i} = x^i + \int_t^s b^i(r, X_r^{t,x,\boldsymbol{\alpha}}, \boldsymbol{\alpha}_r)dr + \int_t^s \sigma^i(r, X_r^{t,x,\boldsymbol{\alpha}}, \boldsymbol{\alpha}_r)dW_r^i.$$

The running and terminal rewards are $f^i(t,x,u^i)$ and $g^i(x)$. The state dynamics may depend on the full control profile.

We impose the following standing conditions on the state coefficients and rewards.

Assumption 2.1 1. b^i, σ^i, f^i, g^i are continuous in their respective arguments.

2. There exists a constant $L > 0$ such that for all $(t,u) \in [0,T] \times U^N$ and $x, x' \in \mathbb{R}^N$,
$$\sum_{i=1}^N \left[|b^i(t,x,u) - b^i(t,x',u)| + |\sigma^i(t,x,u) - \sigma^i(t,x',u)|\right] \le L|x - x'|.$$

3. There exists a constant $L_u > 0$ such that for all $(t,x) \in [0,T] \times \mathbb{R}^N$ and $u, v \in U^N$,
$$\sum_{i=1}^N \left[|b^i(t,x,u) - b^i(t,x,v)| + |\sigma^i(t,x,u) - \sigma^i(t,x,v)|\right] \le L_u|u - v|.$$

4. There exists a constant $C > 0$ such that for all $(t,x,u) \in [0,T] \times \mathbb{R}^N \times U^N$,
$$\sum_{i=1}^N \left[|b^i(t,x,u)| + |\sigma^i(t,x,u)|\right] \le C(1 + |x| + |u|).$$

5. There exist constants $C > 0$, $p_f \ge 2$, and $q_f \ge 1$ such that, for all $(t,x,u^i) \in [0,T] \times \mathbb{R}^N \times U$ and $1 \le i \le N$,
$$|f^i(t,x,u^i)| \le C(1 + |x|^{p_f} + |u^i|^{q_f}), \quad |g^i(x)| \le C(1 + |x|^{p_f}).$$

For the classical HJB arguments, we use the following stronger conditions. Here $C^{k+\alpha}$ has its usual Hölder meaning, and $C^{1,2+\alpha,\ell}$ denotes one continuous time derivative, α-Hölder continuous second state derivatives, and ℓ continuous action derivatives; see Krylov (1987); Gilbarg and Trudinger (2001).

Assumption 2.2 In addition to Assumption 2.1, the diffusion is uniformly elliptic: there exists $\sigma_0 > 0$ such that $\sum_{k=1}^{N} [\sigma^k(t,x,u)]^2 \xi_k^2 \geq \sigma_0^2 |\xi|^2$ for all $(t,x,u,\xi) \in [0,T] \times \mathbb{R}^N \times U^N \times \mathbb{R}^N$. For some $\alpha \in (0,1)$,

$$b^i, \sigma^i \in C^{1,2+\alpha,2}([0,T] \times \mathbb{R}^N \times U^N), \qquad f^i \in C^{1,2+\alpha,1}([0,T] \times \mathbb{R}^N \times U),$$

and $g^i \in C^{2+\alpha}(\mathbb{R}^N)$, with all indicated derivatives bounded.

2.2 Exploratory Policies

Following Wang et al. (2020), we represent exploratory controls by action densities in a relaxed-control formulation. The dynamics average the drift and covariance over these densities, while entropy regularizes the reward.

Denote by $\mathcal{P}(U)$ the set of Borel probability measures on U that are absolutely continuous with respect to Lebesgue measure, and identify each such measure with its density. A *stochastic policy* π maps $(t,x) \in [0,T] \times \mathbb{R}^N$ to a density $\pi(\cdot;t,x) \in \mathcal{P}(U)$, jointly measurable in (t,x,a), with full support on U. For a standard profile $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^N)$, the players randomize independently and we write

$$\boldsymbol{\pi}(du;t,x) := \bigotimes_{j=1}^{N} \pi^j(du^j;t,x).$$

The time-local joint law introduced below may be non-product. We impose the following regularity and integrability conditions on policy densities.

Definition 2.3 We call a stochastic policy π *admissible*, and write $\pi \in \Pi$, when the following conditions hold.

1. $\pi(u;t,x)$ is continuous in (t,x): as $(t',x') \to (t,x)$, $\int_U |\pi(u;t,x) - \pi(u;t',x')| du \to 0$; and Lipschitz continuous in x: there exists $C > 0$ such that $\int_U |\pi(u;t,x) - \pi(u;t,x')| du \leq C|x - x'|$ for all $x, x' \in \mathbb{R}^N$.
2. For all (t,x): $\int_U |\log \pi(u;t,x)| \pi(u;t,x) du \leq C(1 + |x|^p)$ for some $p \geq 2$, and $\int_U |u|^p \pi(u;t,x) du \leq C_p(1 + |x|^p)$ for any $p \geq 1$.
3. There is a jointly measurable sampling map G^π such that $G^\pi(t,x,Z)$ has density $\pi(\cdot;t,x)$ for Z uniform on $[0,1]^m$, and, for every $p \geq 2$, $\int_{[0,1]^m} |G^\pi(t,x,z) - G^\pi(t,x',z)|^p dz \leq L_p |x - x'|^p$.

Averaging the drift and variance as in Wang et al. (2020) defines the exploratory state $X^{t,x,\boldsymbol{\pi}}$ by

$$X_s^{t,x,\boldsymbol{\pi},i} = x^i + \int_t^s \tilde{b}^i(r, X_r^{t,x,\boldsymbol{\pi}}, \boldsymbol{\pi}_r) dr + \int_t^s \tilde{\sigma}^i(r, X_r^{t,x,\boldsymbol{\pi}}, \boldsymbol{\pi}_r) dW_r^i, \tag{2}$$

where $\boldsymbol{\pi}_\cdot = \{\pi_\cdot^1, \ldots, \pi_\cdot^N\}$, and the averaged coefficients are

$$\tilde{b}^i(t,x,\boldsymbol{\pi}) := \int_{U^N} b^i(t,x,u) \boldsymbol{\pi}(u;t,x) du,$$

$$\tilde{\sigma}^i(t,x,\boldsymbol{\pi}) := \sqrt{\int_{U^N} [\sigma^i(t,x,u)]^2 \, \boldsymbol{\pi}(u;t,x) du}.$$

The exploratory performance functional of player i is

$$\tilde{J}^i(t,x;\boldsymbol{\pi}) := \mathbb{E}\Big[\int_t^T e^{-\beta^i(s-t)} \int_U \big[f^i(s, X_s^{t,x,\boldsymbol{\pi}}, u^i) - \gamma^i \log \pi^i(u^i; s, X_s^{t,x,\boldsymbol{\pi}}) \big] \\ \times \pi^i(u^i; s, X_s^{t,x,\boldsymbol{\pi}}) du^i \, ds + e^{-\beta^i(T-t)} g^i(X_T^{t,x,\boldsymbol{\pi}}) \Big], \quad (3)$$

with corresponding best-response value

$$\tilde{V}^i(t,x;\pi^{-i}) := \sup_{\pi^i \in \Pi} \tilde{J}^i(t,x;\pi^i,\pi^{-i}).$$

We write

$$\tilde{f}^i(t,x,\pi^i) := \int_U f^i(t,x,u^i)\pi^i(u^i;t,x)du^i,$$
$$\mathcal{E}(t,x,\pi^i) := \int_U \log \pi^i(u^i;t,x)\pi^i(u^i;t,x)du^i,$$

so that $\mathcal{E}$ is the negative differential entropy of the policy density.

Under Assumption 2.1, the exploratory SDE (2) is well-posed for each admissible Markov profile $\boldsymbol{\pi}$. Let $Z = (Z^1, \dots, Z^N)$ be independent uniform seeds and set

$$G^{\boldsymbol{\pi}}(t,x,Z) := \big(G^{\pi^1}(t,x,Z^1), \dots, G^{\pi^N}(t,x,Z^N) \big).$$

Then

$$\tilde{b}^i(t,x,\boldsymbol{\pi}) = \mathbb{E}_Z[b^i(t,x,G^{\boldsymbol{\pi}}(t,x,Z))], \qquad \tilde{\sigma}^i(t,x,\boldsymbol{\pi}) = \|\sigma^i(t,x,G^{\boldsymbol{\pi}}(t,x,Z))\|_{L^2(Z)}.$$

For $x, x' \in \mathbb{R}^N$, Assumption 2.1(2)–(3), Jensen's inequality, and Definition 2.3(3) give

$$|\tilde{b}^i(t,x,\boldsymbol{\pi}) - \tilde{b}^i(t,x',\boldsymbol{\pi})| \le L|x-x'| + L_u \, \|G^{\boldsymbol{\pi}}(t,x,Z) - G^{\boldsymbol{\pi}}(t,x',Z)\|_{L^1(Z)} \\ \le C_{\boldsymbol{\pi}} |x-x'|.$$

For the diffusion coefficient, the reverse triangle inequality in $L^2(Z)$ yields

$$|\tilde{\sigma}^i(t,x,\boldsymbol{\pi}) - \tilde{\sigma}^i(t,x',\boldsymbol{\pi})| \le \|\sigma^i(t,x,G^{\boldsymbol{\pi}}(t,x,Z)) - \sigma^i(t,x',G^{\boldsymbol{\pi}}(t,x',Z))\|_{L^2(Z)} \\ \le C_{\boldsymbol{\pi}} |x-x'|.$$

The linear-growth bound follows in the same way from Assumption 2.1(4) and the moment condition in Definition 2.3(2). Standard SDE estimates therefore give a unique strong solution and $\mathbb{E}[\sup_{s\in[t,T]} |X_s^{t,x,\boldsymbol{\pi}}|^p] < \infty$ for every $p \ge 2$.

2.3 Control Values and the Nash Benchmark

For $p \in \mathbb{R}^N$ and a symmetric matrix A, define

$$H^i(t,x,u,p,A) := \sum_{k=1}^{N} \left[b^k(t,x,u)p^k + \tfrac{1}{2}\sigma^k(t,x,u)^2 A^{k,k} \right] + f^i(t,x,u^i), \tag{4}$$

and, for a fixed opponent profile,

$$\tilde{H}^i(t,x,u^i,\pi^{-i},p,A) := \int_{U^{N-1}} H^i(t,x,u^i,u^{-i},p,A)\,\pi^{-i}(du^{-i};t,x). \tag{5}$$

Here $\pi^{-i}(du^{-i};t,x) = \bigotimes_{j\neq i}\pi^j(du^j;t,x)$. The candidate HJB equation for the best-response value is

$$\begin{aligned} 0 = {} & \partial_t \tilde{V}^i(t,x;\pi^{-i}) - \beta^i \tilde{V}^i(t,x;\pi^{-i}) \\ & + \sup_{\mu^i \in \mathcal{P}(U)} \int_U \Big[\tilde{H}^i\big(t,x,u^i,\pi^{-i},\partial_x\tilde{V}^i,\partial_x^2\tilde{V}^i\big) - \gamma^i \log \mu^i(u^i) \Big] \mu^i(u^i) du^i. \end{aligned} \tag{6}$$

Whenever the partition function is finite, the entropy variational formula gives the Gibbs density

$$\mathrm{BR}_i[\pi^{-i}](u^i;t,x) = \frac{\exp\{\tilde{H}^i(t,x,u^i,\pi^{-i},\partial_x\tilde{V}^i,\partial_x^2\tilde{V}^i)/\gamma^i\}}{\int_U \exp\{\tilde{H}^i(t,x,v^i,\pi^{-i},\partial_x\tilde{V}^i,\partial_x^2\tilde{V}^i)/\gamma^i\}\,dv^i}. \tag{7}$$

The following verification result identifies a classical HJB solution with the value over the admissible Markov policies and gives the DPP in that class.

Proposition 2.4 (Classical verification and dynamic programming) *Fix $\pi^{-i} \in \Pi^{N-1}$. Let $w \in C^{1,2}([0,T]\times\mathbb{R}^N)$ solve (6) with w in place of $\tilde{V}^i$ and terminal value $w(T,x) = g^i(x)$. Assume w and the derivatives used in Itô's formula have polynomial growth, uniformly in time. Suppose the partition function in (7) is finite and the resulting Gibbs feedback π^i_w, formed with w, belongs to Π. Then $w = \tilde{V}^i(\cdot;\pi^{-i})$, π^i_w is optimal, and, for every $[t,T]$-valued stopping time τ,*

$$\begin{aligned} \tilde{V}^i(t,x;\pi^{-i}) = \sup_{\pi^i\in\Pi} \mathbb{E}\Bigg[& \int_t^\tau e^{-\beta^i(r-t)} \big(\tilde{f}^i(r,X_r,\pi^i_r) - \gamma^i \mathcal{E}(r,X_r,\pi^i_r) \big) dr \\ & + e^{-\beta^i(\tau-t)} \tilde{V}^i(\tau,X_\tau;\pi^{-i}) \Bigg], \end{aligned} \tag{8}$$

where $X = X^{t,x,(\pi^i,\pi^{-i})}$. Taking $\tau \equiv s$ gives the deterministic-time identity.

Proof For a density μ, let $\mathcal{L}_i^\mu$ be the averaged state generator with the opponents using π^{-i}, and set

$$r_i^\mu(s,y) = \tilde{f}^i(s,y,\mu) - \gamma^i \mathcal{E}(s,y,\mu).$$

The HJB equation and its Gibbs optimizer give

$$\partial_t w - \beta^i w + \mathcal{L}_i^{\mu} w + r_i^{\mu} \le 0, \qquad \partial_t w - \beta^i w + \mathcal{L}_i^{\pi_w^i} w + r_i^{\pi_w^i} = 0.$$

Fix $\pi^i \in \Pi$ and apply Itô's formula along X to

$$Y_s = e^{-\beta^i(s-t)} w(s, X_s) + \int_t^s e^{-\beta^i(r-t)} r_i^{\pi_r^i}(r, X_r) dr.$$

Its drift is nonpositive. If $|\nabla_x w(s,y)| \le C(1+|y|^q)$, the averaged coefficient growth and state moments established above imply

$$\mathbb{E} \int_t^T \left| e^{-\beta^i(r-t)} (\nabla_x w)^{\intercal} \tilde{\sigma}(r, X_r, \boldsymbol{\pi}_r) \right|^2 dr \le C_{\boldsymbol{\pi}} \int_t^T \mathbb{E}(1 + |X_r|^{2q+2}) dr < \infty.$$

The stochastic integral is therefore a true martingale. Polynomial growth, the state supremum moments, and admissibility make the terminal and running terms integrable, also at bounded stopping times. Thus $\mathbb{E} Y_\tau \le w(t,x)$, with equality under π_w^i.

At $\tau = T$, the terminal condition gives $\tilde{J}^i(t, x; \pi^i, \pi^{-i}) \le w(t,x)$ for every $\pi^i \in \Pi$, with equality for π_w^i. Hence $w = \tilde{V}^i$. Substituting this identity into the same inequalities at the arbitrary stopping time τ and taking the supremum proves (8). □

Definition 2.5 A profile $\boldsymbol{\pi}^* = (\pi^{1,*}, \dots, \pi^{N,*}) \in \Pi^N$ is a standard exploratory Nash equilibrium if, for every i,

$$\tilde{J}^i(t, x; \pi^{i,*}, \pi^{-i,*}) \ge \tilde{J}^i(t, x; \pi^i, \pi^{-i,*}) \quad \text{for all } \pi^i \in \Pi$$

and every initial condition (t, x).

The next assessment holds the observed opponent actions fixed over the horizon in place of using their continuation policies.

2.4 Frozen Responses and Time-Local Equilibrium

For the compatibility results in Sections 3–4, the scalar action set $U \subset \mathbb{R}$ is an interval with nonempty interior. Response families are taken to be jointly measurable in the action profile. Equalities of densities are understood almost everywhere, and action derivatives are taken on the interior. The vector-action version replaces scalar mixed partials by the corresponding block-Jacobian symmetry conditions.

At a revision, player i observes the current state x and the opponents' actions $c = u^{-i}$ already in effect, before choosing its new action. We use c as the sole input about opponent behavior: no action history is retained to estimate a trend or a response rule. Extending this observation unchanged gives the persistence forecast

$$\widehat{u}^j(s; t, c) = c^j, \qquad s \in [t, T], \quad j \ne i.$$

The player optimizes its own controls over $[t, T]$, while all state coordinates evolve under the physical dynamics with the forecast opponent commands.

Under this forecast, player i faces an entropy-regularized single-agent control problem, with performance $\tilde{J}^i(t, x; \pi^i, u^{-i})$ and value $\tilde{V}^i(t, x; u^{-i})$. Its classical HJB equation is

$$0 = \partial_t \tilde{V}^i(t, x; u^{-i}) - \beta^i \tilde{V}^i(t, x; u^{-i})$$

$$+ \sup_{\pi^i \in \mathcal{P}(U)} \int_U \Big[H^i(t,x,u^i,u^{-i},\partial_x \tilde{V}^i, \partial_x^2 \tilde{V}^i) - \gamma^i \log \pi^i(u^i;t,x) \Big] \pi^i(u^i;t,x) du^i, \tag{9}$$

or, after carrying out the inner maximization,

$$0 = \partial_t \tilde{V}^i - \beta^i \tilde{V}^i + \gamma^i \log \int_U \exp\Big\{ \frac{1}{\gamma^i} H^i(t,x,u^i,u^{-i},\partial_x \tilde{V}^i, \partial_x^2 \tilde{V}^i) \Big\} du^i. \tag{10}$$

In the finite-horizon classical arguments below, we assume a $C^{1,2}$ solution of this equation with terminal value g^i, polynomial growth of the solution and its Itô derivatives, and an admissible Gibbs optimizer. The verification in Proposition 2.4, with u^{-i} fixed, identifies the solution with the frozen optimal value.

Set $v^i(t,x;u^{-i}) := \tilde{V}^i(t,x;u^{-i})$ and call $v = (v^1,\dots,v^N)$ the continuation assessment. Each v^i is specified by its frozen control problem, independently of a joint action law.

Evaluating the Hamiltonian at the derivatives of v^i gives

$$H_v^i(t,x,u) := H^i\big(t,x,u,\partial_x v^i(t,x;u^{-i}), \partial_x^2 v^i(t,x;u^{-i})\big). \tag{11}$$

For a density μ on U, define the local choice criterion at (t,x) by

$$\mathcal{L}_i(\mu;u^{-i}) := \int_U [H_v^i(t,x,a,u^{-i}) - \gamma^i \log \mu(a)]\mu(a) da.$$

Player i's time-local response maximizes $\mathcal{L}_i(\cdot;u^{-i})$. The entropy variational formula gives

$$\pi_*^{i,v}(u^i;t,x,u^{-i}) := \frac{\exp\{H_v^i(t,x,u^i,u^{-i})/\gamma^i\}}{\int_U \exp\{H_v^i(t,x,z^i,u^{-i})/\gamma^i\}\, dz^i}, \tag{12}$$

whenever the denominator is finite. Fix (t,x) and abbreviate $\pi_i(\cdot \mid u^{-i}) = \pi_*^{i,v}(\cdot;t,x,u^{-i})$. This has the exponential form of the logit choice rule in Alós-Ferrer and Netzer (2010, Section 2.1). There the score is a static payoff; here H_v^i includes both the running reward and the effect of the current action on the frozen continuation value.

For a density μ for which the criteria are finite, the loss from replacing π_i by μ is

$$\mathcal{L}_i(\pi_i;u^{-i}) - \mathcal{L}_i(\mu;u^{-i}) = \gamma^i D_{\mathrm{KL}}\big(\mu \,\|\, \pi_i(\cdot \mid u^{-i})\big), \tag{13}$$

where $D_{\mathrm{KL}}(\mu\|\pi) = \int_U \mu(a)\log[\mu(a)/\pi(a)]da$. This identity establishes individual optimality in the local criterion. We now ask whether the playerwise responses can be realized as the full conditionals of one joint action law.

Definition 2.6 Fix (t, x) and the continuation assessment v. A positive joint density $\psi^v(\cdot; t, x)$ on U^N is a *time-local conditional equilibrium* if, for every i, its full conditional density satisfies

$$\psi^v(u^i \mid u^{-i}; t, x) = \pi_*^{i,v}(u^i; t, x, u^{-i}) \quad \text{for a.e. } u^{-i}.$$

The joint-law requirement has an equivalent interpretation through individual revisions. Keeping (t, x) fixed, let player i leave u^{-i} unchanged and draw a new action from $\pi_i(\cdot \mid u^{-i})$. Denote this update by K_i: for bounded measurable $F : U^N \to \mathbb{R}$,

$$(K_i F)(u) = \int_U F(a, u^{-i}) \pi_i(a \mid u^{-i})\, da.$$

For a probability law ν, write νK_i for the law after this update. In Gibbs sampling, one starts from a target joint law and updates a coordinate using its full conditional; each such update preserves the target law (Roberts and Rosenthal 2004, Section 2.3). Here the control problems supply the conditional responses first.

Proposition 2.7 *A positive joint density ψ is a time-local conditional equilibrium at (t, x) if and only if $\psi K_i = \psi$ for every i.*

Proof Let ψ_{-i} be the marginal density of u^{-i}. Integrating out the old action of player i gives

$$(\psi K_i)(u) = \psi_{-i}(u^{-i}) \pi_i(u^i \mid u^{-i}). \tag{14}$$

Since $\psi_{-i} > 0$ almost everywhere, equality with $\psi(u)$ is equivalent to $\psi(u^i \mid u^{-i}) = \pi_i(u^i \mid u^{-i})$. Apply this identity to each i. □

3 Compatibility and Approximation

3.1 Compatibility and Weighted Potentials

Fix (t, x). We first characterize when positive conditional densities determine a joint law, then apply the criterion to the Gibbs responses (12).

Theorem 3.1 *Let $\pi_*^{i,v}(u^i \mid u^{-i})$ be positive conditional densities, twice continuously differentiable on the interior of U^N, and assume the interior of U^N is connected and simply connected. The following are equivalent.*

1. *A time-local conditional equilibrium exists; equivalently, there is a positive integrable function ψ such that*

$$\pi_*^{i,v}(u^i \mid u^{-i}) = \frac{\psi(u)}{\int_U \psi(z^i, u^{-i})\, dz^i}, \qquad i = 1, \ldots, N. \tag{15}$$

2. *There are positive functions $\phi^i : U^{N-1} \to (0, \infty)$ such that the common function*

$$\psi(u) := \pi_*^{i,v}(u^i \mid u^{-i}) \phi^i(u^{-i})$$

is independent of i and satisfies $0 < \int_{U^N} \psi(u) du < \infty$.

3. The cross-partial identities

$$\frac{\partial^2}{\partial u^j \partial u^i} \log \pi_*^{i,v}(u^i \mid u^{-i}) = \frac{\partial^2}{\partial u^i \partial u^j} \log \pi_*^{j,v}(u^j \mid u^{-j}), \qquad i \neq j, \tag{16}$$

hold, and a potential Φ *of the closed one-form*

$$\omega := \sum_{i=1}^{N} \partial_{u^i} \log \pi_*^{i,v}(u^i \mid u^{-i})\, du^i$$

is normalizable: $0 < \int_{U^N} e^{\Phi(u)} du < \infty$.

When these conditions hold, the joint density is unique.

Proof (1) $\Rightarrow$ (2). Put $\phi^i(u^{-i}) = \int_U \psi(z^i, u^{-i}) dz^i$. Then (15) gives $\psi = \pi_*^{i,v}\phi^i$ for every i, and integrability is inherited from the joint density.

(2) $\Rightarrow$ (1). Let ψ denote the common integrable function in (2) and normalize it by $Z = \int_{U^N} \psi(u) du$. Since $\int_U \pi_*^{i,v}(u^i \mid u^{-i}) du^i = 1$,

$$\int_U \psi(u) du^i = \phi^i(u^{-i}).$$

The normalization by Z cancels in the conditional ratio, so the density $Z^{-1}\psi$ has the prescribed full conditionals.

(1) $\Rightarrow$ (3). We first justify differentiating a joint realization ψ. By Fubini's theorem, one can fix an interior reference point z for which the conditional factorizations hold along the coordinate path for almost every u. Put

$$w^{(k)} = (u^1, \ldots, u^{k-1}, z^{k+1}, \ldots, z^N).$$

Cancelling the common marginal at each coordinate change gives the conditional-ratio reconstruction of Lowd (2012, Section 4):

$$\psi(u) = \psi(z) \prod_{k=1}^{N} \frac{\pi_*^{k,v}(u^k \mid w^{(k)})}{\pi_*^{k,v}(z^k \mid w^{(k)})} \quad \text{for a.e. } u.$$

This positive C^2 expression gives a version of ψ on the interior.

For this smooth version and almost every u^{-i}, the conditional factorization

$$\log \psi(u) = \log \pi_*^{i,v}(u^i \mid u^{-i}) + \log \phi^i(u^{-i})$$

gives $\partial_{u^i} \log \psi = \partial_{u^i} \log \pi_*^{i,v}$ along those fibres. Continuity extends this derivative identity throughout the interior. Differentiating with respect to u^j and interchanging mixed derivatives of $\log \psi$ gives (16). Taking $\Phi = \log \psi$ gives the normalizability condition.

(3) $\Rightarrow$ (1). The identities (16) say that ω is closed. By the Poincaré lemma, there is a C^2 function Φ such that $d\Phi = \omega$. Let

$$\widehat{\psi}(u) = Z^{-1} e^{\Phi(u)}, \qquad Z = \int_{U^N} e^{\Phi(v)} dv < \infty.$$

For fixed u^{-i},

$$\partial_{u^i} \log \widehat{\psi}(u) = \partial_{u^i} \Phi(u) = \partial_{u^i} \log \pi_*^{i,v}(u^i \mid u^{-i}).$$

Hence $\log \widehat{\psi}(u) - \log \pi_*^{i,v}(u^i \mid u^{-i})$ is independent of u^i. Thus

$$\widehat{\psi}(u) = c_i(u^{-i}) \pi_*^{i,v}(u^i \mid u^{-i}).$$

Integrating with respect to u^i shows $c_i(u^{-i}) = \int_U \widehat{\psi}(z^i, u^{-i})dz^i$, which is exactly (15).

Finally, if two positive joint densities have the same full conditionals, the derivative of the logarithm of their ratio with respect to every coordinate u^i is zero. Connectedness implies that the ratio is constant, and normalization makes it one. □

On compact U^N, the normalizability condition in part (3) is automatic when Φ extends continuously to U^N. For vector actions on a non-simply-connected domain, the cross-partial identities give only local exactness; global compatibility additionally requires the periods of ω along closed loops to vanish.

For the Gibbs responses, Theorem 3.1 gives the Hamiltonian criterion below.

Theorem 3.2 *Assume the hypotheses of Theorem 3.1. The responses in* (12) *admit a joint realization exactly when*

$$\frac{1}{\gamma^i}\frac{\partial^2 H_v^i}{\partial u^j \partial u^i}(t,x,u) = \frac{1}{\gamma^j}\frac{\partial^2 H_v^j}{\partial u^i \partial u^j}(t,x,u), \qquad i \neq j, \tag{17}$$

and the resulting potential is normalizable. If U is compact and the potential extends continuously to U^N, normalizability is automatic.

Proof The denominator in (12) is independent of u^i, so

$$\partial_{u^i} \log \pi_*^{i,v} = \frac{1}{\gamma^i}\partial_{u^i} H_v^i.$$

Thus (17) is the closedness condition in Theorem 3.1. That theorem gives the joint density $\psi^v = Z_v^{-1} e^{\Phi_v}$ whenever $Z_v = \int_{U^N} e^{\Phi_v(u)}du < \infty$. □

On a simply connected action domain, condition (17) is equivalent to the existence of a potential satisfying

$$\partial_{u^i}\Phi_v(t,x,u) = \frac{1}{\gamma^i}\partial_{u^i} H_v^i(t,x,u), \qquad i = 1,\ldots,N. \tag{18}$$

This is the potential-game criterion for the entropy-scaled Hamiltonians; cf. Monderer and Shapley (1996, Lemma 4.4 and Theorem 4.5). In particular, if $\gamma^i = \gamma$ for every i and the Hamiltonian game has exact potential P_v, then $\Phi_v = P_v/\gamma$ and $\psi^v \propto e^{P_v/\gamma}$, provided $\int_{U^N} e^{P_v/\gamma}du < \infty$.

3.2 Continuation Effects and Explicit Cases

The mixed derivative in (17) contains the explicit action dependence of the coefficients and the sensitivity of $v^i = \tilde{V}^i(\cdot; u^{-i})$ to the frozen parameter. The following decomposition separates these effects.

Proposition 3.3 *For $j \neq i$, assume the classical regularity used in Section 2.4 and that $\tilde{V}^i$, $\partial_x \tilde{V}^i$, and $D_x^2 \tilde{V}^i$ are continuously differentiable in u^{-i}. For the optimal q-function*

$$q_*^i = \partial_t \tilde{V}^i + H^i(t, x, u, \partial_x \tilde{V}^i, \partial_x^2 \tilde{V}^i) - \beta^i \tilde{V}^i,$$

we have

$$\frac{\partial^2 q_*^i}{\partial u^j \partial u^i} = \underbrace{\sum_{k=1}^N \left[\frac{\partial^2 b^k}{\partial u^j \partial u^i} (\nabla_x \tilde{V}^i)^k + \left(\frac{\partial \sigma^k}{\partial u^i} \frac{\partial \sigma^k}{\partial u^j} + \sigma^k \frac{\partial^2 \sigma^k}{\partial u^j \partial u^i} \right) (D_x^2 \tilde{V}^i)^{k,k} \right]}_{direct\ effect}$$
$$+ \underbrace{\sum_{k=1}^N \left[\frac{\partial b^k}{\partial u^i} \frac{\partial (\nabla_x \tilde{V}^i)^k}{\partial u^j} + \sigma^k \frac{\partial \sigma^k}{\partial u^i} \frac{\partial (D_x^2 \tilde{V}^i)^{k,k}}{\partial u^j} \right]}_{indirect\ effect}. \quad (19)$$

Proof Differentiate the definition of q_*^i first with respect to u^i. The terms $\partial_t \tilde{V}^i$ and $-\beta^i \tilde{V}^i$ do not depend on u^i, because u^{-i} is the frozen parameter. Hence

$$\frac{\partial q_*^i}{\partial u^i} = \sum_{k=1}^N \left[\frac{\partial b^k}{\partial u^i} (\nabla_x \tilde{V}^i)^k + \sigma^k \frac{\partial \sigma^k}{\partial u^i} (D_x^2 \tilde{V}^i)^{k,k} \right] + \frac{\partial f^i}{\partial u^i}.$$

Now differentiate with respect to u^j, $j \neq i$. The term $\partial f^i / \partial u^i$ is independent of u^j. Applying the product rule to the remaining terms gives the derivatives of the coefficients displayed in the first line of (19) and the derivatives of the frozen-value jets displayed in the second line. This is exactly (19). □

With separated state dynamics and rewards, the response laws have a product structure.

Proposition 3.4 (Separated rewards) *Suppose that, for every k,*

$$b^k(t, x, u) = b^k(t, x^k, u^k), \qquad \sigma^k(t, x, u) = \sigma^k(t, x^k, u^k),$$

and

$$f^i(t, x, u^i) = F^i(t, x^i, u^i) + h^i(t, x^{-i}), \qquad g^i(x) = G^i(x^i) + k^i(x^{-i}). \quad (20)$$

Assume the classical verification hypotheses for the separated control and expectation problems below, including admissibility of the own-control Gibbs optimizers. Then player i's frozen response depends only on (t, x^i, u^i), and the product of these responses is a time-local conditional equilibrium.

Proof Fix $c = u^{-i}$ and let $\mathcal{A}_{-i}^c$ be the generator of the opponent coordinates with these constant controls. The value of player i's opponent-state rewards is

$$R^i(t, x^{-i}; c) = \mathbb{E}_{t,x^{-i}}^c \left[\int_t^T e^{-\beta^i (s-t)} h^i(s, X_s^{-i})\, ds + e^{-\beta^i (T-t)} k^i(X_T^{-i}) \right].$$

It satisfies $(\partial_t + \mathcal{A}_{-i}^c - \beta^i) R^i + h^i = 0$ and terminal condition k^i. Let $w^i(t, x^i)$ be the entropy-regularized value for the own-coordinate dynamics and rewards F^i, G^i. Define

$$A_i(t, x^i, a) = b^i(t, x^i, a) \partial_{x^i} w^i + \tfrac{1}{2} [\sigma^i(t, x^i, a)]^2 \partial_{x^i x^i} w^i + F^i(t, x^i, a).$$

The one-coordinate HJB is

$$\partial_t w^i - \beta^i w^i + \gamma^i \log \int_U e^{A_i(t,x^i,a)/\gamma^i}\, da = 0, \qquad w^i(T,\cdot) = G^i.$$

Substituting $w^i + R^i$ into the full frozen HJB separates its left side into this equation and the equation for R^i. Verification therefore gives $v^i = w^i + R^i$, with the prescribed terminal value. Its Hamiltonian is

$$H_v^i(t,x,a,c) = A_i(t,x^i,a) + \mathcal{A}_{-i}^c R^i(t,x^{-i};c) + h^i(t,x^{-i}).$$

The last two terms are independent of a and cancel in Gibbs normalization. Thus

$$\pi_*^{i,v}(a;t,x,c) = \frac{e^{A_i(t,x^i,a)/\gamma^i}}{\int_U e^{A_i(t,x^i,z)/\gamma^i}\, dz}.$$

Its product over players has the required full conditionals. □

For two players the potential can be written explicitly by integrating along the two coordinate directions.

Theorem 3.5 *Let $N = 2$, and let $U \subset \mathbb{R}$ be a compact interval of positive length containing the origin. If $H_v^1, H_v^2 \in C^2(U^2)$, then a time-local conditional equilibrium exists if and only if*

$$\frac{1}{\gamma^1}\frac{\partial^2 H_v^1}{\partial u^2 \partial u^1} = \frac{1}{\gamma^2}\frac{\partial^2 H_v^2}{\partial u^1 \partial u^2}. \tag{21}$$

When this holds, its joint density is proportional to

$$\exp\left\{\int_0^{u^1} \frac{1}{\gamma^1}\frac{\partial H_v^1}{\partial v^1}(v^1,0)dv^1 + \int_0^{u^2} \frac{1}{\gamma^2}\frac{\partial H_v^2}{\partial v^2}(u^1,v^2)dv^2\right\}.$$

Proof Necessity follows from Theorem 3.2. Let Φ be the exponent in the displayed joint density. Then $\partial_{u^2}\Phi = (\gamma^2)^{-1}\partial_{u^2} H_v^2$, while

$$\begin{aligned}\partial_{u^1}\Phi &= \frac{1}{\gamma^1}\partial_{u^1} H_v^1(u^1,0) + \int_0^{u^2} \frac{1}{\gamma^2}\partial_{u^1}\partial_{v^2} H_v^2(u^1,v^2)dv^2 \\ &= \frac{1}{\gamma^1}\partial_{u^1} H_v^1(u^1,0) + \int_0^{u^2} \frac{1}{\gamma^1}\partial_{v^2}\partial_{u^1} H_v^1(u^1,v^2)dv^2 = \frac{1}{\gamma^1}\partial_{u^1} H_v^1(u^1,u^2),\end{aligned}$$

by (21) and the fundamental theorem of calculus. Hence $d\Phi$ is the required score field. Continuity on compact U^2 makes e^Φ integrable, and Theorem 3.2 gives the joint realization. □

3.3 Compatibility Gaps and Approximate Realization

For $i \neq j$, define the compatibility gap by

$$\Delta^{ij}(t,x,u) := \frac{1}{\gamma^i}\frac{\partial^2 H_v^i}{\partial u^j \partial u^i}(t,x,u) - \frac{1}{\gamma^j}\frac{\partial^2 H_v^j}{\partial u^i \partial u^j}(t,x,u). \tag{22}$$

The criterion (17) is $\Delta^{ij} = 0$, with a normalizable potential. The next estimate bounds the gap locally in the state variable under uniform frozen-value derivative bounds.

Proposition 3.6 *Suppose $U \subset \mathbb{R}$ is compact and Assumptions 2.1–2.2 hold. For each $R > 0$, assume that the classical frozen values satisfy, uniformly for sufficiently large γ^i,*

$$\sup_{[0,T]\times B_R\times U^{N-1}} \Big(|\partial_x \tilde{V}^i| + |D_x^2 \tilde{V}^i| + \sum_{j\neq i} \big[|\partial_x \partial_{u^j} \tilde{V}^i| + |D_x^2 \partial_{u^j} \tilde{V}^i|\big]\Big) \le C_R. \tag{23}$$

Then, with $\gamma = \min_i \gamma^i$,

$$\sup_{[0,T]\times B_R\times U^N} \left|\pi_*^i(u^i \mid u^{-i}; t, x) - \frac{1}{|U|}\right| = O_R((\gamma^i)^{-1}),$$

$$\max_{i\neq j} \sup_{[0,T]\times B_R\times U^N} |\Delta^{ij}(t, x, u)| = O_R(\gamma^{-1}).$$

Proof Fix R. By compactness of U, Assumption 2.1, and the first two bounds in (23), there is $M_R < \infty$, independent of sufficiently large γ^i, such that

$$|H^i(t, x, u, \partial_x \tilde{V}^i, D_x^2 \tilde{V}^i)| \le M_R \quad \text{on } [0, T] \times B_R \times U^N.$$

This bound implies

$$|U| e^{-M_R/\gamma^i} \le \int_U e^{H^i(t,x,z^i,u^{-i},\partial_x \tilde{V}^i, D_x^2 \tilde{V}^i)/\gamma^i} dz^i \le |U| e^{M_R/\gamma^i}.$$

The Gibbs formula therefore gives

$$\frac{e^{-2M_R/\gamma^i}}{|U|} \le \pi_*^i(u^i \mid u^{-i}; t, x) \le \frac{e^{2M_R/\gamma^i}}{|U|},$$

and the first estimate follows from $e^{2M_R/\gamma^i} - 1 = O_R((\gamma^i)^{-1})$.

For the second estimate, Proposition 3.3 expresses $\partial_{u^j} \partial_{u^i} q_*^i$ as the sum of direct coefficient derivatives and terms containing $\partial_x \partial_{u^j} \tilde{V}^i$ and $D_x^2 \partial_{u^j} \tilde{V}^i$. Assumption 2.2 and (23) thus give

$$\sup_{[0,T]\times B_R\times U^N} |\partial_{u^j} \partial_{u^i} q_*^i| \le C_R, \qquad i \neq j,$$

with C_R independent of sufficiently large exploration weights. Hence

$$|\Delta^{ij}| \le \frac{C_R}{\gamma^i} + \frac{C_R}{\gamma^j} \le \frac{2C_R}{\gamma},$$

which proves the second estimate. □

Perturbation bounds for logit dynamics in finite near-potential games are developed in Candogan et al. (2013, Section 4). Here a uniform bound on the continuous-action cross-partial gaps yields a joint law whose full conditionals approximate the frozen Gibbs responses.

Theorem 3.7 *Suppose $U \subset \mathbb{R}$ is a compact interval of positive length containing the origin, and H_v^i is twice continuously differentiable in u for each i. If*

$$\sup_{(t,x,u)\in[0,T]\times\mathbb{R}^N\times U^N} |\Delta^{ij}(t, x, u)| \le \varepsilon, \quad \forall i \neq j,$$

then there exists a joint density $\hat{\psi}(\cdot; t, x) \in \mathcal{P}(U^N)$ for each $(t, x) \in [0, T] \times \mathbb{R}^N$ such that

1. *the conditional distributions* $\hat{\pi}^i(\cdot \mid u^{-i}; t, x)$ *of* $\hat{\psi}$ *satisfy*

$$D_{\mathrm{KL}}(\hat{\pi}^i(\cdot|u^{-i}; t, x) \| \pi^i_*(\cdot|u^{-i}; t, x)) \le (e^{\delta} - 1)^2, \quad \forall i, u^{-i}, (t, x) \in [0, T] \times \mathbb{R}^N,$$

where $\delta := 2(N-1)\varepsilon|U|^2$. *If* $\delta \le 1$, *then*

$$D_{\mathrm{KL}}(\hat{\pi}^i(\cdot|u^{-i}; t, x) \| \pi^i_*(\cdot|u^{-i}; t, x)) \le 16(N-1)^2\varepsilon^2|U|^4, \quad \forall (t, x) \in [0, T] \times \mathbb{R}^N.$$

2. *for the frozen one-player response problem, define*

$$\varepsilon^i_{\mathrm{loc}} := \sup_{(t,x,u^{-i})} \gamma^i D_{\mathrm{KL}}(\hat{\pi}^i(\cdot \mid u^{-i}; t, x) \| \pi^i_*(\cdot \mid u^{-i}; t, x)).$$

Then $\varepsilon^i_{\mathrm{loc}} \le \gamma^i(e^{\delta} - 1)^2$; *if* $\delta \le 1$, *this becomes* $\varepsilon^i_{\mathrm{loc}} \le 16\gamma^i(N-1)^2\varepsilon^2|U|^4$.

When $\varepsilon = 0$, $\hat{\psi}$ *realizes all frozen Gibbs conditionals exactly.*

Proof Fix (t, x). The piecewise-linear coordinate path from 0 to u remains in U^N by convexity and the assumption $0 \in U$. Along this path, set

$$\Phi(u) := \sum_{k=1}^{N} \int_0^{u^k} a^k(u^1, \dots, u^{k-1}, v^k, 0, \dots, 0) dv^k,$$

where $a^k(u) := \frac{1}{\gamma^k} \frac{\partial H^k_v}{\partial u^k}(t, x, u)$. Set $\hat{\psi}(u; t, x) := e^{\Phi(u)}/Z(t, x)$ with $Z(t, x) = \int_{U^N} e^{\Phi(u)} du < \infty$ (since U is compact and Φ is continuous).

For $1 \le k \le N$, differentiating the kth path segment and the later segments gives

$$\begin{aligned} \partial_{u^k} \Phi(u) &= a^k(u^1, \dots, u^k, 0, \dots, 0) \\ &\quad + \sum_{j>k} \int_0^{u^j} \partial_{u^k} a^j(u^1, \dots, u^{j-1}, v^j, 0, \dots, 0) dv^j. \end{aligned}$$

The fundamental theorem of calculus also gives the telescoping identity

$$a^k(u) - a^k(u^1, \dots, u^k, 0, \dots, 0) = \sum_{j>k} \int_0^{u^j} \partial_{u^j} a^k(u^1, \dots, u^{j-1}, v^j, 0, \dots, 0) dv^j.$$

Subtracting and using $\partial_{u^k} a^j - \partial_{u^j} a^k = \Delta^{jk}$ yields

$$\partial_{u^k} \Phi(u) - a^k(u) = \sum_{j>k} \int_0^{u^j} \Delta^{jk}(u^1, \dots, u^{j-1}, v^j, 0, \dots, 0) dv^j.$$

Since $|\Delta^{jk}| \le \varepsilon$, this gives the uniform bound

$$\left| \tfrac{\partial \Phi}{\partial u^k}(u) - a^k(u) \right| \le (N-1)\varepsilon|U|. \tag{24}$$

For the log-ratio bound, the conditional density $\hat{\pi}^i(u^i|u^{-i}) = \hat{\psi}(u)/\int \hat{\psi}(v^i, u^{-i}) dv^i$ satisfies $\frac{\partial}{\partial u^i} \log \hat{\pi}^i = \frac{\partial \Phi}{\partial u^i}$, while $\frac{\partial}{\partial u^i} \log \pi^i_* = a^i(u)$. By (24), $\left| \frac{\partial}{\partial u^i} \log \frac{\pi^i_*}{\hat{\pi}^i} \right| \le (N-1)\varepsilon|U|$. Integrating from a reference point $u^i_0 \in U$ gives $\left| \log \frac{\pi^i_*}{\hat{\pi}^i}(u^i|u^{-i}) - c(u^{-i}) \right| \le (N-1)\varepsilon|U|^2 =: \eta$, where $c(u^{-i}) = \log(\pi^i_*(u^i_0|u^{-i})/\hat{\pi}^i(u^i_0|u^{-i}))$. Since both π^i_* and $\hat{\pi}^i$ integrate to 1 and $\pi^i_*/\hat{\pi}^i \in [e^{c-\eta}, e^{c+\eta}]$, normalization forces $|c(u^{-i})| \le \eta$, giving $|\log(\pi^i_*/\hat{\pi}^i)| \le 2\eta =: \delta = 2(N-1)\varepsilon|U|^2$.

The KL bound via χ^2-divergence follows: setting $h := \log(\pi^i_*/\hat{\pi}^i)$ with $|h| \le \delta$, we have $D_{\mathrm{KL}}(\hat{\pi}^i \| \pi^i_*) \le \int_U \pi^i_*(e^{-h}-1)^2\, du^i \le (e^\delta - 1)^2$.

For the local objective $\mathcal{L}_i$ in Section 2.4, the Gibbs identity (13) gives

$$\mathcal{L}_i(\pi^i_*; u^{-i}) - \mathcal{L}_i(\hat{\pi}^i; u^{-i}) = \gamma^i D_{\mathrm{KL}}(\hat{\pi}^i \| \pi^i_*).$$

The conditional KL bound therefore yields

$$\gamma^i D_{\mathrm{KL}}(\hat{\pi}^i \| \pi^i_*) \le \gamma^i (e^\delta - 1)^2 \le 16\gamma^i (N-1)^2 \varepsilon^2 |U|^4$$

for $\delta \le 1$. Taking the supremum over (t, x, u^{-i}) gives the stated bound for $\varepsilon^i_{\mathrm{loc}}$. □

The conditional KL bound also controls how much the approximate joint law changes under a single revision. Fix (t, x), let K_i be the revision kernel of Section 2.4, write $\hat{\psi}_{-i}$ for the opponent marginal, and use $\|\mu - \nu\|_{\mathrm{TV}} = \sup_A |\mu(A) - \nu(A)|$. Equation (14) and Pinsker's inequality give

$$\begin{aligned} \|\hat{\psi} K_i - \hat{\psi}\|_{\mathrm{TV}} &= \int \hat{\psi}_{-i}(c)\, \|\pi^i_*(\cdot \mid c) - \hat{\pi}^i(\cdot \mid c)\|_{\mathrm{TV}}\, dc \\ &\le (e^\delta - 1)/\sqrt{2}. \end{aligned}$$

To obtain a performance guarantee over the remaining horizon, we now accumulate the local KL loss along the frozen controlled state.

Corollary 3.8 *Under the hypotheses of Theorem 3.7, assume in addition that the constructed conditional policy $\hat{\pi}^i(\cdot \mid u^{-i}; t, x)$ is admissible for the frozen control problem, and that the classical frozen value and its derivatives used in Itô's formula have polynomial growth in x, uniformly in t for each fixed u^{-i}. Let $\tilde{V}^i(t, x; u^{-i})$ be the frozen-opponent optimal value and let $\tilde{V}^i_{\hat{\pi}}(t, x; u^{-i})$ denote the value under $\hat{\pi}^i$ with the same frozen u^{-i}. Then*

$$0 \le \tilde{V}^i(t, x; u^{-i}) - \tilde{V}^i_{\hat{\pi}}(t, x; u^{-i}) \le \frac{1 - e^{-\beta^i (T-t)}}{\beta^i}\, \varepsilon^i_{\mathrm{loc}} \le \frac{1}{\beta^i}\, \varepsilon^i_{\mathrm{loc}}, \quad \text{if } \beta^i > 0,$$

with the convention that for $\beta^i = 0$ the bound is interpreted in the limiting sense, giving $\tilde{V}^i - \tilde{V}^i_{\hat{\pi}} \le (T-t)\, \varepsilon^i_{\mathrm{loc}}$. In particular, along $\gamma := \min_k \gamma^k \to \infty$, if the global compatibility-gap bound satisfies $\varepsilon = O(1/\gamma)$ and $\gamma^i = O(\gamma)$, then the value-function loss is $O(1/\gamma)$.

Proof Fix u^{-i}, let $X = X^{\hat{\pi}}$ be the state under $(\hat{\pi}^i, u^{-i})$, and write $\tilde{V}^i_{\hat{\pi}} = \tilde{J}^i(\cdot; \hat{\pi}^i, u^{-i})$. The frozen HJB (10) and Gibbs formula (12) give

$$H^i_v = \gamma^i \log \pi^i_* - \partial_t \tilde{V}^i + \beta^i \tilde{V}^i.$$

Integrating against $\hat{\pi}^i$ yields

$$\partial_t \tilde{V}^i - \beta^i \tilde{V}^i + \int_U [H^i_v - \gamma^i \log \hat{\pi}^i] \hat{\pi}^i\, du^i = -\gamma^i D_{\mathrm{KL}}(\hat{\pi}^i \| \pi^i_*). \tag{25}$$

Here $\int_U H^i_v \hat{\pi}^i\, du^i$ is the sum of the controlled generator applied to $\tilde{V}^i$ and the averaged running reward.

Apply Itô's formula to $e^{-\beta^i(s-t)}\tilde{V}^i(s, X_s; u^{-i})$ up to $\tau_R = \inf\{s \geq t : |X_s| \geq R\} \wedge T$. The state moment bounds, coefficient growth, and polynomial gradient bound imply

$$\mathbb{E}^{\hat{\pi}} \int_t^T \left| e^{-\beta^i(s-t)} (\nabla_x \tilde{V}^i)^\intercal \tilde{\sigma}^{\,\hat{\pi}}(s, X_s) \right|^2 ds < \infty,$$

where $\tilde{\sigma}^{\,\hat{\pi}}$ is the averaged frozen diffusion matrix. Thus the stopped stochastic integrals converge in L^2. Polynomial growth and state supremum moments dominate the stopped terminal values; admissibility and the uniform KL bound make the running terms integrable. Sending $R \to \infty$ in the stopped identity and using the common terminal value g^i gives

$$\begin{aligned} &\tilde{V}^i(t, x; u^{-i}) - \tilde{V}^i_{\hat{\pi}}(t, x; u^{-i}) \\ &\quad = \mathbb{E}^{\hat{\pi}} \int_t^T e^{-\beta^i(s-t)} \gamma^i D_{\mathrm{KL}}\big(\hat{\pi}^i(\cdot \mid u^{-i}; s, X_s) \,\|\, \pi^i_*(\cdot \mid u^{-i}; s, X_s)\big)\, ds. \end{aligned} \tag{26}$$

The integrand without the discount factor lies in $[0, \varepsilon^i_{\mathrm{loc}}]$ by Theorem 3.7. Integration of $e^{-\beta^i(s-t)}$ proves the stated bound, including its limit at $\beta^i = 0$. □

If the compatibility-gap estimate is available only on $[0, T] \times B_R \times U^N$, the coordinate-path construction and conditional KL bound remain valid there, but a dynamic value bound requires either control along the whole state trajectory or stopping at the first exit from B_R.

4 Stationary Discounted Frozen Responses

4.1 Frozen Control and Compatibility

We next extend the same current-action forecast and its compatibility and frozen-value bounds to an infinite-horizon discounted problem. Consider time-homogeneous coefficients $f^i = f^i(x, u^i)$, $b^i = b^i(x, u)$, $\sigma^i = \sigma^i(x, u)$, and U compact. Let Π_{stat} denote stationary Markov densities $\pi(u; x)$ satisfying the stationary analogues of Definition 2.3, whose discounted regularized running reward is absolutely integrable.

For fixed u^{-i}, let X^π be the averaged diffusion (2) with initial state x, player i using π^i, and the opponents' controls held at u^{-i}. Each player discounts at rate $\beta^i > 0$. The stationary discounted frozen value is

$$\begin{aligned} V^i(x; u^{-i}) := \sup_{\pi^i \in \Pi_{\mathrm{stat}}} \mathbb{E}\Bigg[\int_0^\infty e^{-\beta^i s} \int_U \Big[f^i(X^\pi_s, u^i) - \gamma^i \log \pi^i(u^i; X^\pi_s) \Big] \\ \times \pi^i(u^i; X^\pi_s)\, du^i\, ds \Bigg]. \end{aligned}$$

The stationary HJB equation is

$$\beta^i V^i(x; u^{-i}) = G^i(x, \nabla_x V^i, D^2_x V^i; u^{-i}, \gamma^i), \tag{27}$$

where $G^i(x, p, A; u^{-i}, \gamma^i) := \gamma^i \log \int_U \exp\{H^i(x, (u^i, u^{-i}), p, A)/\gamma^i\} du^i$, with the time argument suppressed. The conditional optimal policy is $\pi^i_*(u^i \mid x, u^{-i}) = \exp\{q^i_*(x, u)/\gamma^i\}$, where $q^i_*(x, u) := H^i(x, (u^i, u^{-i}), \nabla_x V^i, D^2_x V^i) - \beta^i V^i(x; u^{-i})$ and

$\int_U e^{q_*^i/\gamma^i} du^i = 1$. In this section the value is assumed to be a classical solution of this elliptic equation. We use Δ^{ij} as in (22), with the stationary value V^i replacing $\tilde{V}^i$.

Proposition 4.1 *Under Assumption 2.2 with time-homogeneous coefficients, assume that V^i, $\partial_x V^i$, and $D_x^2 V^i$ are continuously differentiable in u^{-i}. Then the decomposition* (19) *holds with $\tilde{V}^i$ replaced by V^i.*

Proof Since $V^i(x; u^{-i})$ is independent of u^i,

$$\partial_{u^i} q_*^i = \partial_{u^i} H^i(x, u, \partial_x V^i, D_x^2 V^i).$$

Differentiating in u^j gives exactly the product-rule calculation in Proposition 3.3, with $\tilde{V}^i$ replaced by V^i. The assumed parameter regularity supplies all derivatives in (19). □

4.2 Approximation and Value Bounds

Proposition 4.2 *Suppose $U \subset \mathbb{R}$ is compact and the coefficient and parameter regularity in Proposition 4.1 holds. Assume the stationary frozen HJB has a classical solution and, for each $R > 0$, the stationary analog of* (23) *is bounded by a constant C_R independent of sufficiently large γ^i. Then*

$$\sup_{B_R \times U^N} \left| \pi_*^i(u^i \mid x, u^{-i}) - \frac{1}{|U|} \right| = O_R((\gamma^i)^{-1}),$$

$$\max_{i \neq j} \sup_{B_R \times U^N} |\Delta^{ij}(x,u)| = O_R(\gamma^{-1}), \qquad \gamma := \min_k \gamma^k.$$

Proof The stationary jet bounds, Proposition 4.1, and compactness of U give

$$|H^i(x, u, \partial_x V^i, D_x^2 V^i)| \le M_R, \qquad |\partial_{u^j} \partial_{u^i} q_*^i| \le C_R$$

on $B_R \times U^N$, with constants independent of sufficiently large exploration weights. The partition-function estimate in Proposition 3.6 therefore gives

$$\frac{e^{-2M_R/\gamma^i}}{|U|} \le \pi_*^i \le \frac{e^{2M_R/\gamma^i}}{|U|}, \qquad |\Delta^{ij}| \le C_R\big((\gamma^i)^{-1} + (\gamma^j)^{-1}\big) \le 2C_R/\gamma.$$

Expansion of the exponentials gives the first conclusion, and the second bound gives the compatibility-gap estimate. □

Corollary 4.3 *Under the hypotheses of Proposition 4.2, assume in addition that U is an interval containing the origin, and suppose the stationary compatibility gaps satisfy*

$$\sup_{(x,u) \in \mathbb{R}^N \times U^N} |\Delta^{ij}(x,u)| \le \varepsilon, \qquad i \neq j.$$

Then, for each $x \in \mathbb{R}^N$, there exists $\hat{\psi}(\cdot; x) \in \mathcal{P}(U^N)$ with the following properties.

1. *$D_{\mathrm{KL}}(\hat{\pi}^i(\cdot|u^{-i}; x) \| \pi_*^i(\cdot|u^{-i}; x)) \le (e^\delta - 1)^2$, where $\delta = 2(N-1)\varepsilon|U|^2$.*
2. *With $\varepsilon_{\mathrm{loc}}^i := \sup_{(x,u^{-i})} \gamma^i D_{\mathrm{KL}}(\hat{\pi}^i(\cdot \mid x, u^{-i}) \| \pi_*^i(\cdot \mid x, u^{-i}))$, one has $\varepsilon_{\mathrm{loc}}^i \le \gamma^i(e^\delta - 1)^2 \le 16\gamma^i(N-1)^2\varepsilon^2|U|^4$ for $\delta \le 1$.*

3. *If the constructed conditional policy $\hat{\pi}^i(\cdot \mid x, u^{-i})$ is admissible for the stationary frozen problem, V^i and its spatial derivatives used in Itô's formula have polynomial growth, and $\lim_{T\to\infty} \mathbb{E}^{\hat{\pi}}[e^{-\beta^i T}(|V^i(X_T; u^{-i})| + |V^i_{\hat{\pi}}(X_T; u^{-i})|)] = 0$, then $0 \le V^i(x; u^{-i}) - V^i_{\hat{\pi}}(x; u^{-i}) \le \varepsilon^i_{\mathrm{loc}}/\beta^i$.*

If only a gap bound on $B_R \times U^N$ is available, conclusions (1)–(2) hold for $x \in B_R$, with the supremum defining $\varepsilon^i_{\mathrm{loc}}$ restricted to $B_R \times U^{N-1}$. A dynamic value bound then requires either a global gap estimate or stopping at the first exit from B_R.

Proof Fix x and apply the coordinate-path construction of Theorem 3.7 to

$$a^k(u) = \frac{1}{\gamma^k}\partial_{u^k} q^k_*(x,u) = \frac{1}{\gamma^k}\partial_{u^k} H^k(x, u, \partial_x V^k, D^2_x V^k).$$

The coordinate calculation uses only the continuous scores and their cross derivatives, which are supplied by Proposition 4.1. The gap bound is precisely the one used in that calculation. Its conditional KL estimate gives (1); multiplying by γ^i and taking the stated supremum gives (2), with the same δ and constants.

For the value bound, use the stationary identity

$$q^i_*(x,u) = H^i(x, u, \partial_x V^i, D^2_x V^i) - \beta^i V^i(x; u^{-i}) = \gamma^i \log \pi^i_*(u^i \mid x, u^{-i}).$$

Let $\mathcal{L}^{\hat{\pi}}$ be the averaged frozen generator and write

$$r^i_{\hat{\pi}}(x) = \int_U [f^i(x,a) - \gamma^i \log \hat{\pi}^i(a \mid x, u^{-i})]\hat{\pi}^i(a \mid x, u^{-i})da.$$

Integrating the preceding Gibbs identity against $\hat{\pi}^i$ gives

$$\mathcal{L}^{\hat{\pi}} V^i - \beta^i V^i + r^i_{\hat{\pi}} = -\gamma^i D_{\mathrm{KL}}(\hat{\pi}^i \| \pi^i_*).$$

For fixed $T < \infty$, apply Itô's formula up to $\tau_R = \inf\{s \ge 0 : |X^{\hat{\pi}}_s| \ge R\} \wedge T$. The stationary admissibility assumptions give all state supremum moments on $[0,T]$. Together with the polynomial gradient and coefficient growth, they imply

$$\mathbb{E}^{\hat{\pi}} \int_0^T |e^{-\beta^i s}(\nabla_x V^i)^{\intercal} \tilde{\sigma}^{\,\hat{\pi}}(X_s)|^2 ds < \infty.$$

The stopped stochastic integrals converge in L^2. Polynomial growth dominates the stopped terminal values by an integrable function of $\sup_{s\le T}|X_s|$; admissibility and the KL bound dominate the running terms. Letting $R \to \infty$ therefore yields

$$V^i(x; u^{-i}) = \mathbb{E}^{\hat{\pi}}\left[e^{-\beta^i T} V^i(X_T; u^{-i}) + \int_0^T e^{-\beta^i s}\big(r^i_{\hat{\pi}}(X_s) + \gamma^i D_{\mathrm{KL}}(\hat{\pi}^i \| \pi^i_*)(X_s)\big)ds\right].$$

The Markov property and integrability of the discounted performance give the fixed-policy decomposition

$$V^i_{\hat{\pi}}(x; u^{-i}) = \mathbb{E}^{\hat{\pi}}\left[\int_0^T e^{-\beta^i s} r^i_{\hat{\pi}}(X_s)ds + e^{-\beta^i T} V^i_{\hat{\pi}}(X_T; u^{-i})\right].$$

Subtracting gives

$$V^i(x; u^{-i}) - V^i_{\hat{\pi}}(x; u^{-i}) = \mathbb{E}^{\hat{\pi}} \int_0^T e^{-\beta^i s}\gamma^i D_{\mathrm{KL}}(\hat{\pi}^i(\cdot \mid X_s, u^{-i}) \| \pi^i_*(\cdot \mid X_s, u^{-i}))ds$$

$$+\,\mathbb{E}^{\hat{\pi}}\big[e^{-\beta^i T}(V^i(X_T;u^{-i})-V^i_{\hat{\pi}}(X_T;u^{-i}))\big].$$

The last expectation tends to zero by the transversality assumption in the corollary. Letting $T\to\infty$ and using the uniform KL bound gives

$$0\le V^i(x;u^{-i})-V^i_{\hat{\pi}}(x;u^{-i})\le \varepsilon^i_{\mathrm{loc}}\int_0^\infty e^{-\beta^i s}ds=\frac{\varepsilon^i_{\mathrm{loc}}}{\beta^i}.$$

The local statement for conclusions (1)–(2) follows by applying the same construction on the stated compact region. □

5 Estimating Frozen Response Maps

To estimate the frozen responses, we use policy improvement and martingale conditions to identify their value and q-functions. The data are collected under fixed opponent commands; with the physical dynamics specified, these episodes can be simulated for each command profile. For fixed u^{-i} these are single-agent control problems, so we specialize the characterizations of Jia and Zhou (2023, 2025). The policy values used below are assumed to solve their fixed-policy evaluation equations classically. Values and candidate values, together with the derivatives used in Itô's formula, have polynomial growth in x, uniformly in time for each fixed u^{-i}. Section 5.2 verifies the state moments needed for grid-sampled data.

5.1 Policy Evaluation and Improvement

Fix player i and treat u^{-i} as a deterministic frozen parameter. The drift and diffusion may depend on the full control vector $u=(u^i,u^{-i})$.

Definition 5.1 For $u^{-i}\in U^{N-1}$ fixed and $\pi\in\Pi$, the *q^i-function* associated with π is defined by

$$\begin{aligned}q^i(t,x,u^i,u^{-i};\pi)&:=\partial_t\tilde{J}^i(t,x;\pi,u^{-i})\\&\quad+H^i\big(t,x,u^i,u^{-i},\partial_x\tilde{J}^i(t,x;\pi,u^{-i}),\partial_x^2\tilde{J}^i(t,x;\pi,u^{-i})\big)-\beta^i\tilde{J}^i(t,x;\pi,u^{-i}),\end{aligned}\tag{28}$$

where $\tilde{J}^i(\cdot;\pi,u^{-i})$ is the frozen performance of π.

For every (t,x,u^{-i}), write

$$\overline{H}^{i,\pi}:=\int_U\big[H^i(t,x,z^i,u^{-i},\partial_x\tilde{J}^i,D_x^2\tilde{J}^i)-\gamma^i\log\pi(z^i;t,x)\big]\pi(z^i;t,x)dz^i.$$

The fixed-policy evaluation equation is $\partial_t\tilde{J}^i-\beta^i\tilde{J}^i+\overline{H}^{i,\pi}=0$. Substitution into (28) gives the normalization and centering identities

$$\int_U\big[q^i(t,x,u^i,u^{-i};\pi)-\gamma^i\log\pi(u^i;t,x)\big]\pi(u^i;t,x)du^i=0,\tag{29}$$

$$q^i(t,x,u^i,u^{-i};\pi)=H^i(t,x,u^i,u^{-i},\partial_x\tilde{J}^i,D_x^2\tilde{J}^i)-\overline{H}^{i,\pi}.\tag{30}$$

The q^i-function is jointly continuous whenever the Hamiltonian and the derivatives in (28) are jointly continuous in their arguments. Since $\overline{H}^{i,\pi}$ is independent of u^i, exponentiating q^i/γ^i gives the same normalized update as exponentiating the Hamiltonian.

Theorem 5.2 *For $\pi \in \Pi$ satisfying the classical evaluation hypotheses above, define*

$$\mathcal{I}^i_{u^{-i}}(\pi)(u^i; t, x) := \frac{\exp\{q^i(t, x, u^i, u^{-i}; \pi)/\gamma^i\}}{\int_U \exp\{q^i(t, x, z^i, u^{-i}; \pi)/\gamma^i\}\, dz^i}. \tag{31}$$

If, for each frozen parameter u^{-i}, the policy $\pi' = \mathcal{I}^i_{u^{-i}}(\pi)$ belongs to Π, then

$$\tilde{J}^i(t, x; \pi', u^{-i}) \geq \tilde{J}^i(t, x; \pi, u^{-i})$$

for all (t, x, u^{-i}). For a fixed u^{-i}, any fixed point $\mathcal{I}^i_{u^{-i}}(\pi) = \pi$ is the optimal frozen policy π^i_*.

Proof We specialize the verification argument of Jia and Zhou (2023, Theorem 2 and its proof). Fix u^{-i} and abbreviate $J = \tilde{J}^i(\cdot; \pi, u^{-i})$. Put

$$Z_i(s, y) = \int_U e^{q^i(s,y,a,u^{-i};\pi)/\gamma^i} da.$$

Since $q^i - \gamma^i \log \pi' = \gamma^i \log Z_i$, the normalization (29) gives

$$\begin{aligned} D_i(s, y) &:= \int_U [q^i(s, y, a, u^{-i}; \pi) - \gamma^i \log \pi'(a; s, y)]\pi'(a; s, y) da \\ &= \gamma^i \log Z_i(s, y) = \gamma^i D_{\mathrm{KL}}(\pi(\cdot; s, y) \| \pi'(\cdot; s, y)) \geq 0. \end{aligned}$$

This is the Gibbs variational identity of Jia and Zhou (2023, Lemma 13) in the present normalization.

Let X be the averaged state under (π', u^{-i}), and let $r^i_{\pi'} = \int_U (f^i - \gamma^i \log \pi')\pi' da$. Integrating (28) against π' gives

$$\partial_t J - \beta^i J + \mathcal{L}^{\pi'} J + r^i_{\pi'} = D_i.$$

Consequently, Itô's formula yields

$$\begin{aligned} &e^{-\beta^i(T-t)} g^i(X_T) - J(t, x) + \int_t^T e^{-\beta^i(s-t)} r^i_{\pi'}(s, X_s) ds \\ &\quad = \int_t^T e^{-\beta^i(s-t)} D_i(s, X_s) ds + \int_t^T e^{-\beta^i(s-t)} (\nabla_x J)^\intercal \tilde{\sigma}^{\pi'}(s, X_s) dW_s. \end{aligned}$$

The polynomial gradient bound and admissible state moments give $\mathbb{E}\int_t^T |(\nabla_x J)^\intercal \tilde{\sigma}^{\pi'}|^2 ds < \infty$. All running terms are integrable by admissibility and polynomial growth. Taking expectations thus gives

$$\tilde{J}^i(t, x; \pi', u^{-i}) - \tilde{J}^i(t, x; \pi, u^{-i}) = \mathbb{E}^{\pi'} \int_t^T e^{-\beta^i(s-t)} D_i(s, X_s) ds \geq 0.$$

At a fixed point, π attains the supremum in the frozen HJB (9). Its value therefore solves (10) with terminal value g^i. Classical verification gives $\tilde{J}^i = \tilde{V}^i$ and $\pi = \pi^i_*$. □

5.2 Martingale Characterizations

The performance values are defined using the averaged exploratory diffusion. For the martingale data we instead sample an action at each grid point and hold it until the next one, as in Jia and Zhou (2025).

Given a grid $\mathcal{G}_{t:T} = \{t = s_0 < s_1 < \cdots < s_M = T\}$, take independent random seeds $\xi_0, \ldots, \xi_{M-1}$, uniform on $[0,1]^m$ and independent of the original diffusion filtration, on a product extension. Reveal ξ_k at s_k and use the usual augmentation of $\mathscr{F}_s \vee \sigma(\xi_k : s_k \le s)$ for the grid-sampled martingale statements; we again denote this filtration by $\mathscr{F}_s$.

Using the sampling map G^π from Definition 2.3, set, on $[s_k, s_{k+1})$,

$$a_s^{\mathcal{G},\pi} = G^\pi(s_k, X_{s_k}^{\mathcal{G},\pi}, \xi_k),$$

and let $X^{t,x,\mathcal{G},\pi,u^{-i}}$ solve the original state equation with player i's action equal to $a^{\mathcal{G},\pi}$ and the opponents frozen at u^{-i}. The state remains continuous while the action is sampled only at grid points.

For fixed $c = u^{-i}$, the sampled state has the finite-horizon moment bounds needed below. Write $a_r = a_r^{\mathcal{G},\pi}$, $X_r = X_r^{t,x,\mathcal{G},\pi,c}$, and let $\kappa(r) = s_k$ on $[s_k, s_{k+1})$. Conditional sampling and Definition 2.3(2) give, for $p \ge 2$,

$$\mathbb{E}|a_r|^p = \mathbb{E}\int_U |a|^p \pi(a; \kappa(r), X_{\kappa(r)})da \le C_p(1 + \mathbb{E}|X_{\kappa(r)}|^p).$$

On each grid cell the action is constant and has finite pth moment; coefficient growth and the SDE moment estimate therefore give finite state supremum moments by induction over the finite grid. To obtain a bound independent of the grid, set $F_p(s) = \mathbb{E}\sup_{t\le v\le s}|X_v|^p$. The drift estimate and the Burkholder–Davis–Gundy inequality give

$$\begin{aligned} F_p(s) &\le C_{p,T,c}\left(1 + |x|^p + \int_t^s [F_p(r) + \mathbb{E}|a_r|^p]dr\right) \\ &\le C_{p,T,c,\pi}\left(1 + |x|^p + \int_t^s F_p(r)dr\right). \end{aligned}$$

Gronwall's inequality and the action moment bound imply

$$\mathbb{E}\sup_{t\le r\le T}|X_r|^p + \sup_{t\le r<T}\mathbb{E}|a_r|^p \le C_{p,T,c,\pi}(1 + |x|^p), \tag{32}$$

where the constant is independent of the grid. Thus, if a classical function w satisfies $|\nabla_x w(r,y)| \le C(1 + |y|^q)$, coefficient growth and Young's inequality give

$$\begin{aligned} &\mathbb{E}\int_t^T |e^{-\beta^i r}(\nabla_x w)^\intercal \sigma(r, X_r, a_r, c)|^2 dr \\ &\qquad \le C_c \int_t^T \mathbb{E}(1 + |X_r|^{4q+4} + |a_r|^{4q+4})dr < \infty. \end{aligned} \tag{33}$$

We use (33) for the values and candidate values in the next three statements.

The following statement is the fixed-u^{-i} form of Jia and Zhou (2025, Theorem 6).

Theorem 5.3 *Let $\pi \in \Pi$ satisfy the classical evaluation hypotheses above, and let $\hat{q}^i : [0,T] \times \mathbb{R}^N \times U^N \to \mathbb{R}$ be continuous. Then $\hat{q}^i = q^i(\cdot;\pi)$ if and only if, for every (t, x, u^{-i}) and every time grid $\mathcal{G}_{t:T}$, the process*

$$\begin{aligned} M_s^{\mathcal{G}} &:= e^{-\beta^i s}\tilde{J}^i(s, X_s^{t,x,\mathcal{G},\pi,u^{-i}}; \pi, u^{-i}) \\ &\quad + \int_t^s e^{-\beta^i r}\big[f^i(r, X_r^{t,x,\mathcal{G},\pi,u^{-i}}, a_r^{\mathcal{G},\pi}) - \hat{q}^i(r, X_r^{t,x,\mathcal{G},\pi,u^{-i}}, a_r^{\mathcal{G},\pi}, u^{-i})\big]dr \end{aligned} \tag{34}$$

is a martingale.

Proof Assume first that $\hat{q}^i = q^i$. On each grid interval $[s_k, s_{k+1})$ the sampled action $a^{\mathcal{G},\pi}$ is constant and adapted. Itô's formula therefore applies to $e^{-\beta^i r}\tilde{J}^i(r, X_r^{t,x,\mathcal{G},\pi,u^{-i}}; \pi, u^{-i})$. Suppressing the fixed arguments (π, u^{-i}) and writing $X_r = X_r^{t,x,\mathcal{G},\pi,u^{-i}}$, we obtain

$$\begin{aligned} &d\big(e^{-\beta^i r}\tilde{J}^i(r, X_r)\big) + e^{-\beta^i r}\big[f^i(r, X_r, a_r^{\mathcal{G},\pi}) - q^i(r, X_r, a_r^{\mathcal{G},\pi}, u^{-i}; \pi)\big]dr \\ &\quad = e^{-\beta^i r}\Big[\partial_t\tilde{J}^i + H^i(r, X_r, a_r^{\mathcal{G},\pi}, u^{-i}, \partial_x\tilde{J}^i, \partial_x^2\tilde{J}^i) - \beta^i\tilde{J}^i - q^i(r, X_r, a_r^{\mathcal{G},\pi}, u^{-i}; \pi)\Big]dr \\ &\qquad + e^{-\beta^i r}(\nabla_x\tilde{J}^i(r, X_r))^{\intercal}\sigma(r, X_r, a_r^{\mathcal{G},\pi}, u^{-i})\,dW_r. \end{aligned}$$

The drift is zero by (28). Equation (33), applied to $w = \tilde{J}^i$, makes the stochastic integral square integrable. Integrating from t to s proves the martingale property of (34).

Conversely, assume (34) is a martingale for every grid. Applying Itô's formula as above, now with $\hat{q}^i$, shows that

$$A_s := \int_t^s e^{-\beta^i r}\big[q^i(r, X_r, a_r^{\mathcal{G},\pi}, u^{-i}; \pi) - \hat{q}^i(r, X_r, a_r^{\mathcal{G},\pi}, u^{-i})\big]dr$$

is both a continuous finite-variation process and a local martingale. Hence $A_s = 0$ a.s. for every s.

Set $F = q^i - \hat{q}^i$. Suppose $F(t_0, x_0, a_0, u^{-i}) > 0$ at some $t_0 < T$. By continuity, there are $h > 0$, neighborhoods B_x of x_0 and B_a of a_0, and $c_0 > 0$ such that $t_0 + h < T$ and

$$F(r, y, a, u^{-i}) \geq c_0, \qquad (r, y, a) \in [t_0, t_0 + h] \times B_x \times B_a.$$

Start at (t_0, x_0) and choose a grid with no resampling before $t_0 + h$. Full support gives $\mathbb{P}(a_{t_0}^{\mathcal{G},\pi} \in B_a) > 0$. Define

$$\tau = \inf\{r \geq t_0 : X_r \notin B_x\} \wedge (t_0 + h).$$

Continuity of X gives $\tau > t_0$ a.s. Taking a countable intersection over rational times and using continuity, A is zero simultaneously at all times on one event of probability one. On its intersection with $\{a_{t_0}^{\mathcal{G},\pi} \in B_a\}$, the action is held fixed and

$$A_\tau = \int_{t_0}^{\tau} e^{-\beta^i r}F(r, X_r, a_{t_0}^{\mathcal{G},\pi}, u^{-i})dr \geq c_0\int_{t_0}^{\tau} e^{-\beta^i r}dr > 0,$$

contradicting $A \equiv 0$. Applying the same argument to $-F$ excludes negative values. Hence $F = 0$ for $t < T$, and continuity gives the equality also at T.

□

Theorem 5.3 identifies q^i when the policy value is known. With a terminal condition and the policy normalization (29), the martingale condition identifies both functions; compare Jia and Zhou (2025, Theorem 7).

Corollary 5.4 *Fix an admissible policy $\pi \in \Pi$ satisfying the classical evaluation hypotheses above. For each fixed u^{-i}, let $\hat{J}^i(\cdot,\cdot;u^{-i}) \in C^{1,2}([0,T]\times\mathbb{R}^N)$ have polynomial growth, and let $\hat{q}^i$ be continuous. Suppose*

$$\hat{J}^i(T,x,u^{-i}) = g^i(x), \qquad \int_U [\hat{q}^i(t,x,u^i,u^{-i}) - \gamma^i \log \pi(u^i;t,x)]\pi(u^i;t,x)du^i = 0.$$

Then $\hat{J}^i(\cdot,\cdot;u^{-i}) = \tilde{J}^i(\cdot,\cdot;\pi,u^{-i})$ and $\hat{q}^i = q^i(\cdot;\pi)$ if and only if, for every (t,x,u^{-i}) and every grid $\mathcal{G}_{t:T}$, the process in (34), with $\tilde{J}^i$ replaced by $\hat{J}^i$, is a martingale.

Proof If $(\hat{J}^i,\hat{q}^i) = (\tilde{J}^i,q^i)$, the claim is Theorem 5.3.

For the converse define

$$q^i_{\hat{J}}(t,x,u) := \partial_t \hat{J}^i(t,x,u^{-i}) + H^i(t,x,u,\partial_x\hat{J}^i,\partial_x^2\hat{J}^i) - \beta^i \hat{J}^i(t,x,u^{-i}).$$

Applying Itô's formula to $e^{-\beta^i s}\hat{J}^i(s,X_s^{\mathcal{G},\pi})$, with its stochastic integral justified by (33), and subtracting the assumed martingale shows that

$$\int_t^s e^{-\beta^i r}\big[q^i_{\hat{J}}(r,X_r^{\mathcal{G},\pi},a_r^{\mathcal{G},\pi},u^{-i}) - \hat{q}^i(r,X_r^{\mathcal{G},\pi},a_r^{\mathcal{G},\pi},u^{-i})\big]dr = 0$$

a.s. for every s and every grid. The continuity and full-support argument in the proof of Theorem 5.3, applied to $q^i_{\hat{J}} - \hat{q}^i$, therefore gives $\hat{q}^i = q^i_{\hat{J}}$ pointwise. Substituting this identity into the normalization constraint and using the fact that $\partial_t\hat{J}^i - \beta^i\hat{J}^i$ is independent of u^i gives

$$\begin{aligned} 0 &= \partial_t\hat{J}^i(t,x,u^{-i}) - \beta^i\hat{J}^i(t,x,u^{-i}) \\ &\quad + \int_U \big[H^i(t,x,u^i,u^{-i},\partial_x\hat{J}^i,\partial_x^2\hat{J}^i) - \gamma^i\log\pi(u^i;t,x)\big]\pi(u^i;t,x)du^i. \end{aligned}$$

Together with the terminal condition, this is the fixed-policy evaluation equation for π. Apply Itô's formula along the averaged state X under (π,u^{-i}). The polynomial derivative and state moment bounds make its stochastic integral square integrable, so taking expectations gives

$$\begin{aligned} \hat{J}^i(t,x;u^{-i}) = \mathbb{E}^\pi\Bigg[& e^{-\beta^i(T-t)}g^i(X_T) \\ &+ \int_t^T e^{-\beta^i(s-t)}\big(\tilde{f}^i(s,X_s,\pi) - \gamma^i\mathcal{E}(s,X_s,\pi)\big)ds\Bigg] = \tilde{J}^i(t,x;\pi,u^{-i}). \end{aligned}$$

The identity $\hat{q}^i = q^i_{\hat{J}}$ then gives $\hat{q}^i = q^i(\cdot;\pi)$. □

For the optimal frozen policy π^i_*, the associated q^i-function is

$$\begin{aligned} q^i_*(t,x,u^i,u^{-i}) &:= \partial_t\tilde{V}^i(t,x;u^{-i}) \\ &\quad + H^i(t,x,u^i,u^{-i},\partial_x\tilde{V}^i(t,x;u^{-i}),\partial_x^2\tilde{V}^i(t,x;u^{-i})) - \beta^i\tilde{V}^i(t,x;u^{-i}). \end{aligned} \tag{35}$$

The frozen HJB (10) fixes the normalization:

$$\int_U e^{q_*^i/\gamma^i}\,du^i = e^{(\partial_t\tilde V^i-\beta^i\tilde V^i)/\gamma^i}\int_U e^{H_v^i/\gamma^i}\,du^i = 1, \qquad \pi_*^i = e^{q_*^i/\gamma^i}.$$

For the optimal pair, exponential normalization replaces the fixed-policy normalization. The next result uses this constraint to identify the optimal frozen value and response from a full-support behavior policy; cf. Jia and Zhou (2025, Theorem 9).

Theorem 5.5 *For each fixed u^{-i}, let $\hat J_*^i(\cdot,\cdot;u^{-i})\in C^{1,2}([0,T]\times\mathbb{R}^N)$ have polynomial growth, and let $\hat q_*^i$ be continuous. Assume*

$$\hat J_*^i(T,x,u^{-i}) = g^i(x), \qquad \int_U \exp\{\hat q_*^i(t,x,u^i,u^{-i})/\gamma^i\}\,du^i = 1.$$

(1) If $(\hat J_*^i,\hat q_*^i)=(\tilde V^i,q_*^i)$, then for every behavior policy $\pi\in\Pi$, every (t,x,u^{-i}), and every grid $\mathcal{G}_{t:T}$,

$$\begin{aligned} &e^{-\beta^i s}\hat J_*^i(s,X_s^{t,x,\mathcal{G},\pi,u^{-i}},u^{-i}) \\ &\quad+\int_t^s e^{-\beta^i r}\big[f^i(r,X_r^{t,x,\mathcal{G},\pi,u^{-i}},a_r^{\mathcal{G},\pi})-\hat q_*^i(r,X_r^{t,x,\mathcal{G},\pi,u^{-i}},a_r^{\mathcal{G},\pi},u^{-i})\big]dr \end{aligned} \tag{36}$$

is a martingale.

(2) Conversely, suppose the normalized candidate density $\hat\pi_*^i := \exp\{\hat q_*^i/\gamma^i\}$ belongs to Π for each fixed u^{-i}. If there exists one full-support behavior policy $\pi\in\Pi$ for which (36) is a martingale for every (t,x,u^{-i}) and every grid, then $(\hat J_*^i,\hat q_*^i)=(\tilde V^i,q_*^i)$.

Proof For a candidate value, set

$$q_{\hat J_*}^i := \partial_t\hat J_*^i + H^i(t,x,u,\partial_x\hat J_*^i,D_x^2\hat J_*^i)-\beta^i\hat J_*^i.$$

The Itô decomposition in the proof of Theorem 5.3, with $\tilde J^i$ replaced by $\hat J_*^i$, has drift $e^{-\beta^i r}(q_{\hat J_*}^i-\hat q_*^i)$ in (36). For the true optimal pair this drift vanishes by (35); equation (33) makes the remaining stochastic integral a true martingale under every admissible behavior policy. This proves (1).

For (2), the assumed martingale property makes the integral of that drift a continuous finite-variation local martingale. The continuity and full-support argument in the same proof, applied to $q_{\hat J_*}^i-\hat q_*^i$ for every initial condition and grid, gives $\hat q_*^i = q_{\hat J_*}^i$ pointwise. Substitution into the exponential normalization yields

$$1 = e^{(\partial_t\hat J_*^i-\beta^i\hat J_*^i)/\gamma^i}\int_U \exp\{H^i(t,x,u,\partial_x\hat J_*^i,D_x^2\hat J_*^i)/\gamma^i\}\,du^i.$$

Taking logarithms gives the frozen HJB (10). The equality $\hat q_*^i = q_{\hat J_*}^i$ shows that $\hat\pi_*^i$ is its Gibbs optimizer: the time-derivative and discount terms are independent of the own action and cancel in the Gibbs ratio. Its admissibility in (2), the terminal condition, and the growth hypotheses allow the verification in Proposition 2.4 with u^{-i} fixed. Thus $\hat J_*^i=\tilde V^i$, and then $\hat q_*^i = q_*^i$. □

5.3 Moment Estimators

We now turn the martingale characterizations into temporal-difference (TD) moment equations for estimation. Along each trajectory, u^{-i} is fixed and player i's action is sampled on the grid.

For candidate functions $(\hat{J}^i, \hat{q}^i)$ and a grid interval $[s_k, s_{k+1}]$, define the residual

$$\begin{aligned}\delta_k^i := e^{-\beta^i s_{k+1}} \hat{J}^i(s_{k+1}, X_{s_{k+1}}, u^{-i}) - e^{-\beta^i s_k} \hat{J}^i(s_k, X_{s_k}, u^{-i}) \\ + \int_{s_k}^{s_{k+1}} e^{-\beta^i r} \big[f^i(r, X_r, a_r^{\mathcal{G},\pi}) - \hat{q}^i(r, X_r, a_r^{\mathcal{G},\pi}, u^{-i}) \big] dr.\end{aligned} \tag{37}$$

At the true pair $(\tilde{J}^i, q^i)$, Theorem 5.3 gives $\mathbb{E}[\delta_k^i \mid \mathscr{F}_{s_k}] = 0$, with the sampled action included in $\mathscr{F}_{s_k}$. Following Jia and Zhou (2022a, Section 4.2, (22)), choose square-integrable, $\mathscr{F}_{s_k}$-measurable instruments ζ_k. Then

$$\mathbb{E}[\zeta_k \delta_k^i] = 0. \tag{38}$$

To learn the dependence on u^{-i}, choose a design distribution ν_i on U^{N-1}. At the beginning of each trajectory, draw $\bar{u}^{-i} \sim \nu_i$ independently of the Brownian motion and sampling seeds. Include this observed draw in the initial filtration and keep it fixed on the whole trajectory. Player i is sampled from a full-support behavior policy π_b at the grid points. Trajectories with different frozen profiles are combined in the same regression.

For the optimal pair, use (37) with $(\hat{J}^i, \hat{q}^i, \pi, u^{-i})$ replaced by $(\hat{J}_*^i, \hat{q}_*^i, \pi_b, \bar{u}^{-i})$, and denote this increment by $\delta_k^{i,\mathrm{off}}$. Write $\hat{J}_*^{i,\phi}$ and $\hat{q}_*^{i,\theta}$ for the value and score families, and set $z = (\phi, \theta)$. Choose a finite family of square-integrable instruments from $(s_k, X_{s_k}, a_{s_k}^i, \bar{u}^{-i})$. For n recorded transitions, define

$$\widehat{m}_n(z) = \frac{1}{n} \sum_{k=1}^{n} \zeta_k \delta_k^{i,\mathrm{off}}(z), \qquad \widehat{z}_n \in \arg\min_z \widehat{m}_n(z)^\intercal W_n \widehat{m}_n(z), \tag{39}$$

when the constrained minimum exists. Here W_n is positive definite, the residuals are square integrable, and the candidate families satisfy the growth, terminal, normalization, and admissibility requirements of Theorem 5.5.

Identification in the chosen parameter class requires a unique population root of these moments. The quality of the approximation also depends on coverage by the design law and the quadrature used for the time integrals.

The Gibbs constraint in Theorem 5.5 can be enforced by

$$\hat{q}_*^{i,\theta} = \bar{q}_*^{i,\theta} - \gamma^i \log \int_U \exp\{\bar{q}_*^{i,\theta}/\gamma^i\}\, du^i,$$

whenever the integral is finite. Quadrature gives an approximation to this exact normalization.

Algorithm 1 Off-Policy Learning of Player i's Frozen Response Map

Require: Design law ν_i; full-support behavior policy π_b; time grid; instruments ζ; positive definite weight W_n

Initialize $z = (\phi, \theta)$ in candidate families satisfying Theorem 5.5
for episode $e = 1, 2, \ldots$ **do**
  Draw $\bar{u}^{-i} \sim \nu_i$ and keep $\bar{u}^{-i}$ fixed during the episode
  for grid cells $[s_k, s_{k+1}]$ **do**
    Draw $a^i_{s_k} \sim \pi_b(\cdot; s_k, X_{s_k})$ and hold it on $[s_k, s_{k+1})$
    Evaluate ζ_k from the current time, state, action and frozen parameter
    Evolve under $(a^i_{s_k}, \bar{u}^{-i})$; store the transition and ζ_k
  end for
  Fit (ϕ^i, θ^i) using (39) on the recorded transitions
end for
return $\hat{q}^{i,\theta^i}_*$ and $\hat{J}^{i,\phi^i}_*$

6 Numerical Experiments

The LQ benchmark gives an explicit compatible joint law and conditional KL and frozen-value losses when the exploration weights are perturbed. We then give a coefficient regression with the frozen value supplied analytically.

6.1 Two-Player Stationary Discounted LQ Benchmark

Throughout this section we write $\rho > 0$ for the common discount rate of both players (i.e. $\rho = \beta^1 = \beta^2$ in the notation of Section 4); this is independent of the exploration weights $\gamma^i > 0$.

We take a shared scalar state and action space $U = \mathbb{R}$. For stationary exploratory policies $\boldsymbol{\pi} = (\pi^1, \pi^2)$, the averaged state satisfies

$$dX_t = \left(AX_t + \sum_{j=1}^{2} B^j \int_{\mathbb{R}} a\, \pi^j(a \mid X_t)\, da \right) dt + \sigma\, dW_t,$$

where $A < 0$ is the uncontrolled mean-reversion coefficient and B^i measures player i's influence on the state. The discounted exploratory performance is

$$\tilde{J}^i(x; \boldsymbol{\pi}) = \mathbb{E}_x \int_0^\infty e^{-\rho t} \int_{\mathbb{R}} \big[- Q^i X_t^2 - R^i a^2 - \gamma^i \log \pi^i(a \mid X_t) \big] \pi^i(a \mid X_t)\, da\, dt.$$

Thus $f^i(x, a) = -(Q^i x^2 + R^i a^2)$.

We use the asymmetric parameters in Table 1 throughout this section.

Table 1 Parameters of the LQ benchmark.

Parameter	Value	Role
A	-1.0	Mean-reversion
B^1, B^2	$2.0, 0.5$	Player 1 has $4\times$ stronger control
Q^1, Q^2	$1.0, 2.0$	Asymmetric state penalties
R^1, R^2	$1.0, 0.5$	Asymmetric control costs
σ	0.3	Diffusion coefficient
ρ	0.1	Discount rate
Δt	0.02	Discretization step

The continuation assessment of Section 2.4 is available in closed form. Holding the opponents at a^{-i} over the remaining horizon gives

$$\tilde{V}^i(x; a^{-i}) = -P^i x^2 + \ell^i(a^{-i})\, x + m^i(a^{-i}), \tag{40}$$

in which P^i is the positive root of the scalar Riccati equation

$$\frac{(B^i)^2}{R^i} P^2 + (\rho - 2A)P - Q^i = 0, \tag{41}$$

and

$$\ell^i(a^{-i}) = \frac{-2P^i \sum_{j\neq i} B^j a^j}{E^i}, \qquad E^i := \rho - A + \frac{(B^i)^2 P^i}{R^i}. \tag{42}$$

Put $c = a^{-i}$, write B^{-i} for the other player's coefficient, and set $\ell'^i = -2P^i B^{-i}/E^i$. Matching constants in the frozen HJB gives

$$\begin{aligned} m^i(c) &= m^i_0 + m^i_{(2)} c^2, \\ m^i_0 &= \frac{-\sigma^2 P^i + (\gamma^i/2)\log(\pi\gamma^i/R^i)}{\rho}, \qquad m^i_{(2)} = \frac{B^{-i}\ell'^i + (B^i)^2(\ell'^i)^2/(4R^i)}{\rho}. \end{aligned}$$

To verify this value on the noncompact action set, fix c and denote the quadratic candidate in (40) by v. Its frozen Hamiltonian can be written as

$$\begin{aligned} H^i(x, a, c, v', v'') &= C_i(x, c) - R^i \left(a - \frac{B^i v'(x)}{2R^i}\right)^2, \\ C_i(x, c) &= (Ax + B^{-i}c)v'(x) + \tfrac{1}{2}\sigma^2 v''(x) - Q^i x^2 + \frac{(B^i)^2 v'(x)^2}{4R^i}. \end{aligned}$$

The coefficient equations above are exactly

$$\rho v(x) = C_i(x, c) + \frac{\gamma^i}{2}\log\frac{\pi\gamma^i}{R^i} = \gamma^i \log \int_{\mathbb{R}} e^{H^i(x,a,c,v',v'')/\gamma^i} da.$$

The optimizer $\pi_*^i(\cdot \mid c; x)$ is Gaussian with mean $K^i x + B^i \ell'^i c/(2R^i)$ and variance $\gamma^i/(2R^i)$, where $K^i = -P^i B^i/R^i$.

For any admissible stationary density π, let $\mathcal{L}^{\pi,c}$ be its averaged frozen generator and set $r_\pi^i(x) = \int_{\mathbb{R}}[-Q^i x^2 - R^i a^2 - \gamma^i \log \pi(a; x)]\pi(a; x)da$. The Gibbs identity gives

$$\rho v - \mathcal{L}^{\pi,c} v - r_\pi^i = \gamma^i D_{\mathrm{KL}}(\pi(\cdot; x) \| \pi_*^i(\cdot \mid c; x)).$$

We use policies with finite discounted quadratic state and action moments, integrable discounted absolute entropy, and $\lim_{T\to\infty} e^{-\rho T}\mathbb{E}|v(X_T)| = 0$. On every finite interval, v' is affine and σ is constant, so the stochastic integral in Itô's formula is square integrable. Applying that formula, taking expectations, and then letting $T \to \infty$ gives

$$v(x) - \tilde{J}^i(x; \pi, c) = \mathbb{E}_x^{\pi,c} \int_0^\infty e^{-\rho t} \gamma^i D_{\mathrm{KL}}(\pi(\cdot; X_t) \| \pi_*^i(\cdot \mid c; X_t))dt \geq 0. \tag{43}$$

The terminal term vanishes by the stated condition, the reward integral converges absolutely, and the nonnegative KL integral passes to the limit by monotone convergence; this is the discounted verification argument of Wang et al. (2020, Appendix B). Under π_*^i, the averaged drift coefficient is $A - (B^i)^2 P^i/R^i < 0$. Its second moments are bounded for fixed c, its entropy is constant, and the transversality condition holds. Equality in (43) therefore identifies $v = \tilde{V}^i$.

These parameters give

$$P^1 = 0.3022, \quad P^2 = 0.8000, \qquad K^1 = -0.6044, \quad K^2 = -0.8000.$$

6.2 Compatibility and Approximation

For the shared scalar state, direct differentiation of the frozen Hamiltonian gives $\partial_{a^i} H_v^i = B^i[-2P^i x + \ell^i(a^{-i})] - 2R^i a^i$. Differentiating once more in the other player's action yields

$$\frac{\partial^2 H_v^i}{\partial a^j \partial a^i} = B^i \frac{\partial \ell^i}{\partial a^j} = -\frac{2P^i B^i B^j}{E^i}, \qquad j \neq i. \tag{44}$$

The compatibility gap in (22) is therefore

$$\Delta^{12} = -2B^1 B^2 \left(\frac{\varphi^1}{\gamma^1} - \frac{\varphi^2}{\gamma^2} \right), \qquad \varphi^i := \frac{P^i}{E^i}, \tag{45}$$

and does not depend on (t, x). With the parameters above, $\varphi^1 = 0.1309$ and $\varphi^2 = 0.5333$, so $\Delta^{12} = 0$ exactly when

$$\frac{\gamma^2}{\gamma^1} = \frac{\varphi^2}{\varphi^1} \approx 4.0745. \tag{46}$$

At $\gamma^1 = \gamma^2 = 1$, the gap is 0.8049. Multiplying both exploration weights by the same factor scales the gap inversely by (45). With $\gamma^1 = 1$ fixed, the gap decreases monotonically as γ^2/γ^1 increases and changes sign at (46) (Figure 1(a) and Table 2).

Table 2 Compatibility gap for the LQ benchmark with $\gamma^1 = 1$.

γ^2/γ^1	1	2	4	**4.0745**	5	8	10
Δ^{12}	0.8049	0.2715	0.0049	**0**	−0.0485	−0.1285	−0.1551

At the ratio (46), put

$$c_{12} = \frac{B^1 \ell'^1}{\gamma^1} = \frac{B^2 \ell'^2}{\gamma^2}.$$

For $s \in \mathbb{R}$, define the quadratic potential and its normalizing factor by

$$\begin{aligned}\Phi(a; x, s) = &-\frac{R^1}{\gamma^1}(a^1)^2 - \frac{R^2}{\gamma^2}(a^2)^2 + s a^1 a^2 \\ &- 2x\left(\frac{P^1 B^1}{\gamma^1} a^1 + \frac{P^2 B^2}{\gamma^2} a^2\right), \qquad Z(x, s) = \int_{\mathbb{R}^2} e^{\Phi(a;x,s)}\, da.\end{aligned}$$

The compatible joint density is $\psi(a; x) = Z(x, c_{12})^{-1} e^{\Phi(a;x,c_{12})}$, with covariance

$$\Sigma = \tfrac{1}{2}\mathsf{Q}^{-1}, \qquad \mathsf{Q} = \begin{pmatrix} R^1/\gamma^1 & -c_{12}/2 \\ -c_{12}/2 & R^2/\gamma^2 \end{pmatrix}.$$

With $\gamma^1 = 1$ and $\gamma^2 = \varphi^2/\varphi^1 \approx 4.0745$, we obtain $c_{12} = -0.2618$ and $\det \mathsf{Q} = 0.1056 > 0$. Thus $Z(x, c_{12})$ is finite, and

$$\Sigma = \begin{pmatrix} 0.5811 & -0.6199 \\ -0.6199 & 4.7357 \end{pmatrix}.$$

Its full conditionals reproduce the two Gibbs responses exactly: the conditional variances are $0.5000 = \gamma^1/(2R^1)$ and $4.0745 = \gamma^2/(2R^2)$ (Figure 1(b)).

For these parameters, compatibility requires unequal exploration weights. If $B^1 = B^2$, $Q^1 = Q^2$ and $R^1 = R^2$, then $\varphi^1 = \varphi^2$ and equal exploration weights satisfy (46).

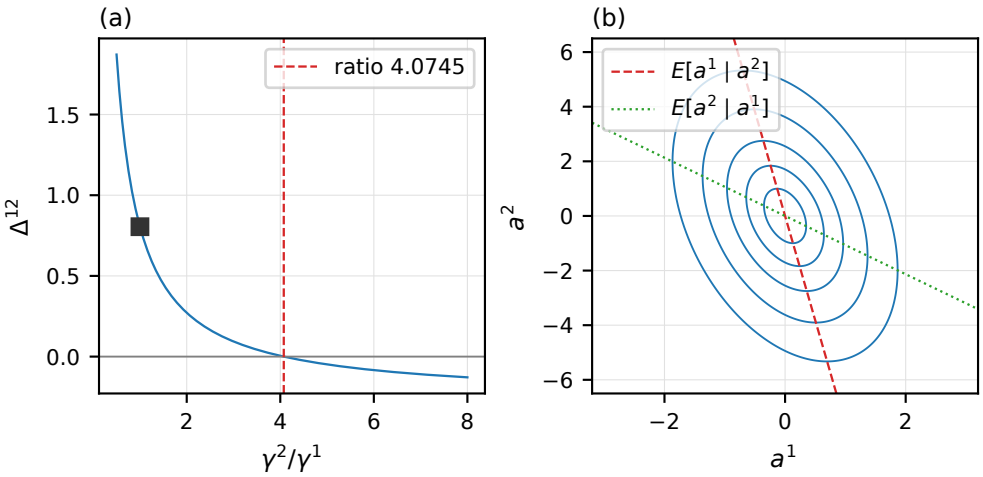


Fig. 1 Compatibility for the LQ benchmark with $\gamma^1 = 1$. *(a)* the compatibility gap Δ^{12} of (45) as a function of the exploration-weight ratio γ^2/γ^1; it is positive at equal weights and crosses zero at (46). *(b)* level sets of the joint law ψ at $x = 0$ and that ratio, with the two conditional means; the conditional variances agree with $\gamma^i/(2R^i)$ to machine precision

The coordinate-path construction gives an explicit joint approximation when the exploration weights are perturbed from (46). Put

$$d_i = \frac{B^i \ell'^i}{\gamma^i}, \qquad i = 1, 2, \qquad \Delta^{12} = d_1 - d_2.$$

Using the coordinate order (a^1, a^2) and anchor $(0, 0)$ gives

$$\widehat{\psi}(a; x) = Z(x, d_2)^{-1} e^{\Phi(a;x,d_2)},$$

which is normalizable when $d_2^2 < 4R^1R^2/(\gamma^1\gamma^2)$. Let $\widehat{\pi}^i$ be the full conditional densities of $\widehat{\psi}$. Completing the square gives $\widehat{\pi}^2 = \pi_*^2$ and

$$\widehat{\pi}^1(\cdot \mid c; x) = \mathcal{N}\left(K^1 x + \frac{\gamma^1 d_2}{2R^1} c, \frac{\gamma^1}{2R^1}\right), \qquad c = a^2.$$

The target π_*^1 has the same variance and mean $K^1x + \gamma^1 d_1 c/(2R^1)$. Hence

$$D_{\mathrm{KL}}(\widehat{\pi}^1(\cdot \mid c; x) \| \pi_*^1(\cdot \mid c; x)) = \frac{\gamma^1 (\Delta^{12})^2 c^2}{4R^1}. \tag{47}$$

For fixed c, the state drift under $\widehat{\pi}^1$ has coefficient $A + B^1K^1 < 0$. Its second moments are bounded and the discounted quadratic terminal term vanishes. Write X for the averaged exploratory state under $\widehat{\pi}^1$ with player 2 held at c, and $\tilde{J}^1(x; \widehat{\pi}^1, c)$ for its

discounted reward. Applying the verified identity (43) to $\widehat{\pi}^1$ gives

$$\begin{aligned}
&\tilde{V}^1(x;c) - \tilde{J}^1(x;\widehat{\pi}^1,c) \\
&\quad = \mathbb{E}_x \int_0^\infty e^{-\rho t}\gamma^1 D_{\mathrm{KL}}\big(\widehat{\pi}^1(\cdot \mid c;X_t) \,\|\, \pi_*^1(\cdot \mid c;X_t)\big)\, dt \\
&\quad = \frac{(\gamma^1)^2(\Delta^{12})^2 c^2}{4\rho R^1}.
\end{aligned} \tag{48}$$

The last equality uses the state-independent KL error in (47). Table 3 gives these Gaussian calculations on $U = \mathbb{R}$ at $\gamma^1 = c = 1$. Density quadrature and independent integration of the discounted Ornstein–Uhlenbeck rewards agree with these formulas to 10^{-12}.

Table 3 Conditional approximation and frozen-value loss near the compatible ratio. Here $\gamma^2/\gamma^1 = (1+\theta)\varphi^2/\varphi^1$ and $\gamma^1 = c = 1$. Player 2's conditional error is zero.

θ	γ^2/γ^1	$\lvert\Delta^{12}\rvert$	Conditional KL	Value loss
-0.2	3.2596	0.06545	0.001071	0.010708
-0.1	3.6671	0.02909	0.000212	0.002115
0.0	4.0745	0.00000	0.000000	0.000000
$+0.1$	4.4820	0.02380	0.000142	0.001416
$+0.2$	4.8894	0.04363	0.000476	0.004759

6.3 Coefficient Regression with a Known Value

With the analytic frozen value (40) supplied, we regress only the action-dependent coefficients of a one-step Euler target.

Write $a = (a^1, a^2)$ and $c = a^{-i}$. Set $s_\Delta = 1 + A\Delta t$ and $\delta_\rho = e^{-\rho\Delta t}$. With the actions held fixed over one Euler step, the transition and temporal-difference target are

$$\begin{aligned}
X' &= s_\Delta x + \Delta t(B^i a^i + B^{-i}c) + \sigma\sqrt{\Delta t}\,\xi, \qquad \xi \sim \mathcal{N}(0,1), \\
Y_\Delta^i &= f^i(x,a^i)\Delta t + \delta_\rho \tilde{V}^i(X';c) - \tilde{V}^i(x;c).
\end{aligned}$$

The conditional mean of this target is related to the stationary q-function by

$$\lim_{\Delta t \downarrow 0} \frac{\mathbb{E}[Y_\Delta^i \mid x,a]}{\Delta t} = H_v^i(x,a) - \rho\tilde{V}^i(x;c) = q_*^i(x,a).$$

At the fixed step size used in the regression, we compare the estimates with the exact Euler conditional mean computed below. Put $\chi_i = R^i\Delta t + \delta_\rho P^i (B^i)^2 \Delta t^2$. Substituting

the first two conditional moments of X' into the quadratic value and completing the square in a^i gives

$$\begin{aligned}\mathbb{E}[Y^i_\Delta \mid x, a] = &-\frac{e^{-\psi^i_3}}{2}(a^i - \psi^i_1 x - \psi^i_2)^2 + \eta^i_1 cx + \eta^i_2 c \\ &+ \eta^i_3 a^i c + \eta^i_4 c^2 + C^i_\Delta(x),\end{aligned} \tag{49}$$

with exact coefficients

$$\begin{aligned}\psi^i_1 &= -\frac{\delta_\rho P^i s_\Delta B^i \Delta t}{\chi_i}, \quad \psi^i_2 = 0, \quad e^{-\psi^i_3} = 2\chi_i, \\ \eta^i_1 &= (\delta_\rho s_\Delta - 1)\ell'^i - 2\delta_\rho P^i s_\Delta B^{-i}\Delta t, \qquad \eta^i_2 = 0, \\ \eta^i_3 &= \delta_\rho \ell'^i B^i \Delta t - 2\delta_\rho P^i B^i B^{-i}\Delta t^2, \\ \eta^i_4 &= (\delta_\rho - 1)m^i_{(2)} + \delta_\rho \ell'^i B^{-i}\Delta t - \delta_\rho P^i (B^{-i})^2 \Delta t^2.\end{aligned} \tag{50}$$

Using the exact ψ^i_1 in (50), the known state-only baseline is

$$\begin{aligned}C^i_\Delta(x) = &[P^i - Q^i\Delta t - \delta_\rho P^i s^2_\Delta + \chi_i(\psi^i_1)^2]x^2 \\ &- \delta_\rho P^i \sigma^2 \Delta t + (\delta_\rho - 1)m^i_0.\end{aligned}$$

Omitting m^i_0 from both the value and baseline leaves the centered target unchanged. We fit the seven parameters $(\psi^i_1, \psi^i_2, \psi^i_3, \eta^i_1, \ldots, \eta^i_4)$ after subtracting the baseline. The coefficient η^i_3 satisfies

$$\frac{\eta^i_3 + 2\delta_\rho P^i B^i B^{-i}\Delta t^2}{\delta_\rho \Delta t} = B^i \ell'^i.$$

Thus the fitted cross coefficient can be compared with the mixed derivative in (44) after the finite-step correction.

We use 3×10^6 Euler transitions with $\Delta t = 0.02$ and seed 20250529. At each step, the behavior actions $a^1, a^2 \sim \mathcal{N}(0, 1)$ are redrawn independently and held fixed during that step. The initial state and all fitted parameters are zero. For the three ψ coefficients we use Adam with step size 10^{-4} and componentwise gradients clipped to $[-1, 1]$. For the four η coefficients we use stochastic gradient descent with step size 4×10^{-4} during the first 66.7% of steps, then decreasing linearly to 10^{-4}. Iterates are recorded every 500 steps.

Figures 2–3 show the saved iterates against their exact finite-step targets. At the last saved iterate, the own-state slopes have relative errors 1.3% and 3.2%, and the cross coefficients have errors 2.2% and 1.9%. The small η^1_4 coefficient has absolute error 2.81×10^{-4} and relative error 75.6%.

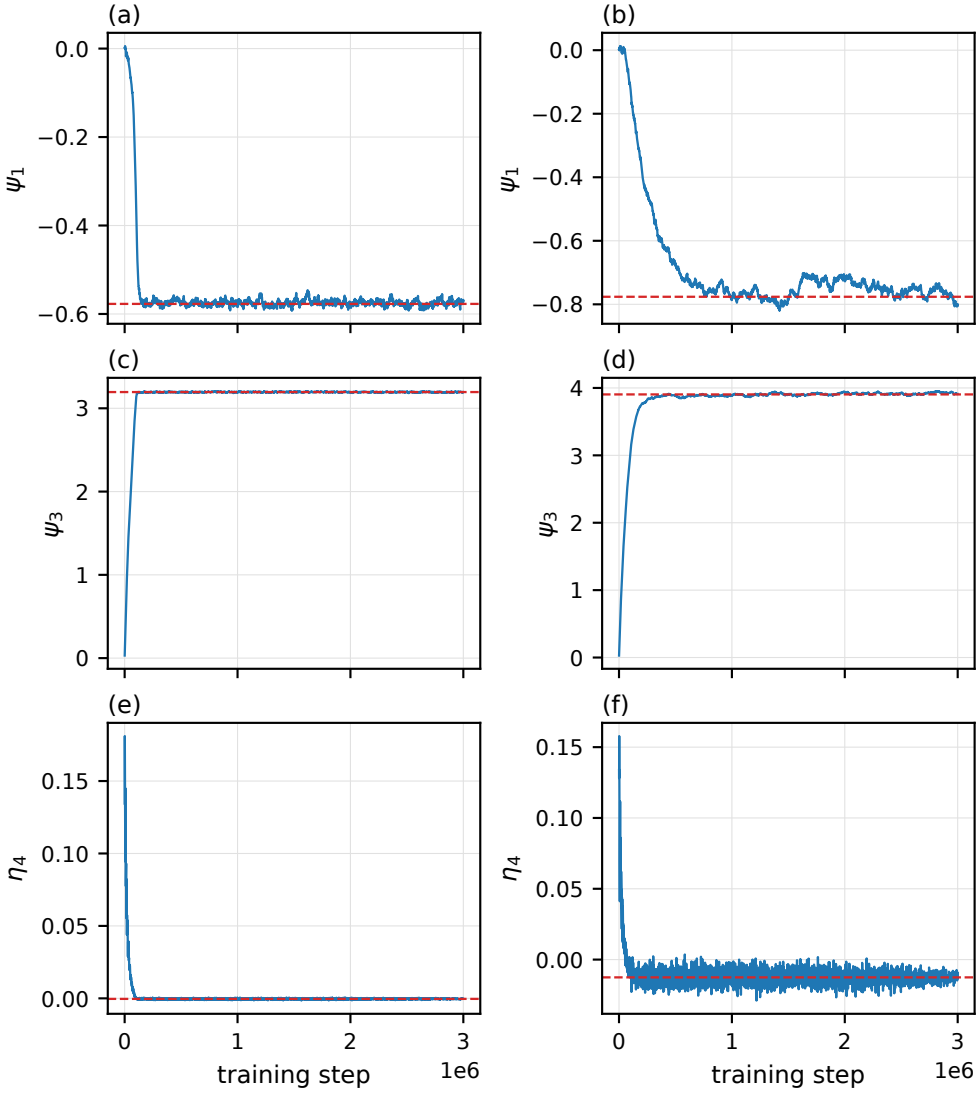


Fig. 2 Coefficient regression with a known value: rows show the own-state slope ψ_1, action-curvature parameter ψ_3, and opponent-square coefficient η_4; columns show players 1 and 2. Solid curves are estimates and dashed lines are the exact finite-step targets given in (50)

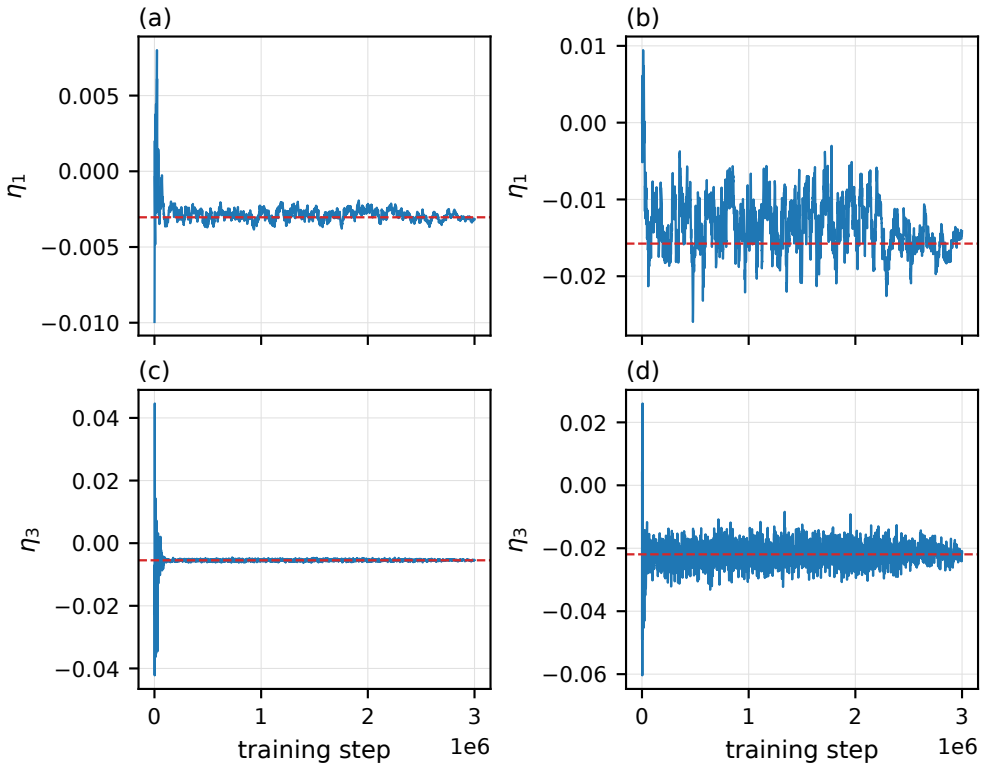


Fig. 3 Opponent-state and cross coefficients in the same run: rows show η_1 and η_3; columns show players 1 and 2. Solid curves are estimates and dashed lines are the exact finite-step targets in (50)

7 Concluding Remarks

For the prescribed current-action forecast, joint realizability is characterized by the entropy-scaled Hamiltonian cross-partial condition and a normalizable potential. A coordinate-path construction gives conditional KL and frozen-value bounds in the finite-horizon and stationary discounted settings.

The martingale characterizations provide moment conditions for estimating these response maps. In the LQ benchmark, compatibility selects an explicit ratio of exploration weights. Perturbing that ratio gives an explicit quadratic KL and frozen-value loss, and a regression with the value supplied analytically estimates the one-step coefficients.

Statements and Declarations

- **Funding.** No funding was received for this work.
- **Competing interests.** The author declares no conflict of interest.
- **Ethics approval and consent to participate.** Not applicable.

- **Consent for publication.** Not applicable.
- **Data availability.** The numerical experiments use synthetic data generated from the model in Section 6; no external dataset was analysed. The generated data are available from the author on reasonable request.
- **Code availability.** The numerical code is available from the author on reasonable request.
- **Author contributions.** The author is solely responsible for the conception of the study, the mathematical analysis, and the writing of the manuscript.
- **Use of large language models.** During the preparation of this work the author used DeepSeek and ChatGPT to assist with writing and with reasoning checks. The model, mathematical ideas, and methods presented here are the author's own. The author reviewed and verified all content and takes full responsibility for the manuscript.

References

Alós-Ferrer C, Netzer N (2010) The logit-response dynamics. Games Econ Behav 68(2):413–427. https://doi.org/10.1016/j.geb.2009.08.004

Arnold BC, Press SJ (1989) Compatible conditional distributions. J Am Stat Assoc 84(405):152–156. https://doi.org/10.1080/01621459.1989.10478750

Arnold BC, Castillo E, Sarabia JM (1992) Conditionally Specified Distributions, Lecture Notes in Statistics, vol 73. Springer, New York, https://doi.org/10.1007/978-1-4612-2912-4

Aumann RJ (1974) Subjectivity and correlation in randomized strategies. J Math Econ 1(1):67–96. https://doi.org/10.1016/0304-4068(74)90037-8

Blume LE (1993) The statistical mechanics of strategic interaction. Games Econ Behav 5(3):387–424. https://doi.org/10.1006/game.1993.1023

Blume LE, Durlauf SN (2003) Equilibrium concepts for social interaction models. Int Game Theory Rev 5(3):193–209. https://doi.org/10.1142/S021919890300101X

Candogan O, Ozdaglar A, Parrilo PA (2013) Dynamics in near-potential games. Games Econ Behav 82:66–90. https://doi.org/10.1016/j.geb.2013.07.001

Černý J, An B, Zhang AN (2022) Quantal correlated equilibrium in normal form games. In: Proceedings of the 23rd ACM Conference on Economics and Computation. ACM, pp 210–239, https://doi.org/10.1145/3490486.3538350

Chen HY (2010) Compatibility of conditionally specified models. Stat Probab Lett 80(7–8):670–677. https://doi.org/10.1016/j.spl.2009.12.025

Firoozi D, Jaimungal S (2022) Exploratory LQG mean field games with entropy regularization. Automatica 139:110177. https://doi.org/10.1016/j.automatica.2022.110177

Gilbarg D, Trudinger NS (2001) Elliptic Partial Differential Equations of Second Order, 2nd edn. Springer, Berlin, https://doi.org/10.1007/978-3-642-61798-0

Guo X, Xu R, Zariphopoulou T (2022) Entropy regularization for mean field games with learning. Math Oper Res 47(4):3239–3260. https://doi.org/10.1287/moor.2021.1238

Hommes CH, Ochea MI, Tuinstra J (2018) Evolutionary competition between adjustment processes in Cournot oligopoly: Instability and complex dynamics. Dyn Games Appl 8(4):822–843. https://doi.org/10.1007/s13235-018-0238-x

Huang Y, Jia Y, Zhou XY (2025) Sublinear regret for a class of continuous-time linear-quadratic reinforcement learning problems. SIAM J Control Optim 63(5):3452–3474. https://doi.org/10.1137/24M1695075

Jia Y, Zhou XY (2022a) Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. J Mach Learn Res 23(154):1–55. URL https://jmlr.org/papers/v23/21-0947.html

Jia Y, Zhou XY (2022b) Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. J Mach Learn Res 23(275):1–50. URL https://jmlr.org/papers/v23/21-1387.html

Jia Y, Zhou XY (2023) q-Learning in continuous time. J Mach Learn Res 24(161):1–61. URL https://jmlr.org/papers/v24/22-0755.html

Jia Y, Zhou XY (2025) Erratum to "q-learning in continuous time". J Mach Learn Res URL https://www.jmlr.org/papers/volume24/22-0755/erratum.pdf, erratum linked to volume 24, article 161 (2023)

Krylov NV (1987) Nonlinear Elliptic and Parabolic Equations of the Second Order. D. Reidel Publishing Company, Dordrecht

Li SHQ, Yu Y, Dörfler F, et al (2024) A coupled optimization framework for correlated equilibria in normal-form games. In: 2024 IEEE 63rd Conference on Decision and Control (CDC). IEEE, pp 1739–1744, https://doi.org/10.1109/CDC56724.2024.10886643

Lowd D (2012) Closed-form learning of Markov networks from dependency networks. In: Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence, Proceedings of Machine Learning Research, vol R10. PMLR, pp 531–540, URL https://proceedings.mlr.press/r10/lowd12a.html, reissued by PMLR in 2026

Mertikopoulos P, Sandholm WH (2016) Learning in games via reinforcement and regularization. Math Oper Res 41(4):1297–1324. https://doi.org/10.1287/moor.2016.0778

Monderer D, Shapley LS (1996) Potential games. Games Econ Behav 14(1):124–143. https://doi.org/10.1006/game.1996.0044

Muré J (2019) Optimal compromise between incompatible conditional probability distributions, with application to objective Bayesian kriging. ESAIM Probab Stat 23:271–309. https://doi.org/10.1051/ps/2018023

Ortiz LE, Schapire RE, Kakade SM (2007) Maximum entropy correlated equilibria. In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol 2. PMLR, pp 347–354, URL https://proceedings.mlr.press/v2/ortiz07a.html

Roberts GO, Rosenthal JS (2004) General state space Markov chains and MCMC algorithms. Probab Surv 1:20–71. https://doi.org/10.1214/154957804100000024

Savas Y, Ahmadi M, Tanaka T, et al (2019) Entropy-regularized stochastic games. In: 2019 IEEE 58th Conference on Decision and Control (CDC). IEEE, pp 5955–5962, https://doi.org/10.1109/CDC40024.2019.9029555

Sun Z, Jia G (2023) Reinforcement learning for exploratory linear-quadratic two-person zero-sum stochastic differential games. Appl Math Comput 442:127763. https://doi.org/10.1016/j.amc.2022.127763

Tang W, Zhou XY (in press) Regret of exploratory policy improvement and q-learning. J Mach Learn Res URL https://arxiv.org/abs/2411.01302v2, accepted for publication; preprint arXiv:2411.01302v2 (2026)

Tang W, Zhang YP, Zhou XY (2022) Exploratory HJB equations and their convergence. SIAM J Control Optim 60(6):3191–3216. https://doi.org/10.1137/21M1448185

Torroba Hennigen L, Kim Y (2023) Deriving language models from masked language models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, pp 1149–1159, https://doi.org/10.18653/v1/2023.acl-short.99

Wang H, Zariphopoulou T, Zhou XY (2020) Reinforcement learning in continuous time and space: A stochastic control approach. J Mach Learn Res 21(198):1–34. URL https://jmlr.org/papers/v21/19-144.html