Quantum Mean-Field Games: Numerical Nash Equilibria via PDE-form HJB Learning

Nevroz Sen [1]

Abstract

Quantum mean-field games (QMFGs) extend classical mean-field games to populations of indirectly observed quantum systems governed by Belavkin stochastic equations. Prior QMFG literature provides theoretical foundations but, to our knowledge, no computed Hamilton–Jacobi–Bellman (HJB) optimal Nash equilibria. We present two contributions. First, we identify a structural failure mode of trajectory-based discrete Bellman / deep-BSDE training on Belavkin SDEs: when the measurement diffusion direction is aligned with the value-function gradient, the Monte Carlo Bellman residual is dominated by innovation noise, and an asymptotic signal-to-noise analysis shows the Bellman-loss minimizer is limited by an alignment-dependent noise floor that does not vanish as the timestep is refined. We confirm the geometric prediction on two structurally distinct quantum benchmarks: under aligned noise the trained policy matches zero control, while in a near-orthogonal benchmark Bellman training recovers PDE-form performance. Second, using PDE-form HJB learning inside a fictitious-play loop, we compute, to our knowledge, the first numerical Nash equilibrium for a multi-type Belavkin-filtered QMFG, exhibiting equilibria for priors $p \in \{0.25, 0.5, 0.75\}$ and proving the symmetric Nash equilibrium at $p = 1/2$ by reflection symmetry.

1. Introduction

Mean-field game (MFG) theory (Lasry and Lions, 2006; Huang et al., 2006) analyzes symmetric large-population games in the limit where each agent becomes negligible, replacing explicit multi-agent interactions with a coupled HJB–Fokker–Planck system characterizing the self-consistent equilibrium. *Quantum* MFG (QMFG) (Kolokoltsov, 2022; 2021), introduced by Kolokoltsov, apply the same reduction to populations of continuously-monitored quantum systems governed by Belavkin SDEs (Belavkin, 1987; Bouten et al., 2007), opening a control-theoretic path to systems such as arrays of measurement-feedback-stabilized qubits. Numerical methods for QMFG lag significantly behind their classical counterparts. The recent result of Chalal, Amini, Amini, and Laurière (Chalal et al., 2026) establishes a dynamic programming principle on Hilbert–Schmidt space and proves exponential convergence of a Lyapunov-feedback controller for a qubit Ising benchmark. Kolokoltsov (Kolokoltsov, 2021) proves that solutions to the limiting MFG system induce an ϵ-Nash equilibrium in the corresponding N-agent game, but leaves existence of such solutions open. Adapting the standard classical-MFG numerical machinery (trajectory-based Bellman/deep-BSDE losses (Han et al., 2018; Carmona and Laurière, 2021; Ruthotto et al., 2020) and fictitious play (Perrin et al., 2020; Hu, 2020)) to this setting reveals a subtlety specific to quantum filtering that, to our knowledge, has not been identified. We herein show that the choice of HJB loss matters that it can be missed on classical benchmarks. The stochastic increment in the Belavkin filter is $\sigma dW = O(\sqrt{\Delta t})$, while the Bellman signal is only $O(\Delta t)$, so the Monte Carlo Bellman residual can be dominated by innovation noise and can admit spurious minimizers. Training under the Bellman residual produces a policy no better than zero control despite reaching a low training loss; training the same architecture under the PDE-form residual produces a useful policy, competitive with the Chalal Lyapunov feedback (Chalal et al., 2026) across three problem variants in the Chalal benchmark family. Section 3 gives the asymptotic argument and the experimental comparison.

Hence, using the PDE-form solver inside a fictitious-play outer loop, we compute a numerical Nash equilibrium for a two-type QMFG. At the symmetric prior $p = 1/2$, a reflection-symmetry argument proves self-consistency of the symmetric Nash equilibrium. Across $p \in \{0.25, 0.5, 0.75\}$, the population-weighted Nash cost varies moderately, by $\approx 13\%$ across the tested priors, with no strong monotone dependence on p. This is consistent with the analytical no-linear-correction result of Proposition 4; whether the varia-

[1]C3 AI, Redwood City, CA, USA. Correspondence to: Nevroz Sen <nevroz.sen@mail.mcgill.ca>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

tion tends to zero in the limit of optimization error tending to zero. This paper makes two contributions. *Methodological* (Section 3): on Belavkin SDEs with measurement aligned with the value-function gradient, the discrete Bellman residual fails, i.e., its minimizer is bounded away from any HJB solution by an innovation-noise floor (Proposition 1), while the PDE-form (PINN) residual succeeds. Rotating the measurement axis breaks the alignment and Bellman recovers, confirming the failure is geometric rather than universal to Belavkin filtering. On the single-type Chalal et al. benchmark, PINN-HJB inside fictitious play attains the best two-parameter Lyapunov controller of (Chalal et al., 2026). *Application* (Section 5): we compute, to our knowledge, the first numerical Nash equilibrium for a multi-type Belavkin-filtered QMFG, a setting with no known Lyapunov closed form due to the shared-mean-field coupling between per-type HJBs. Existence results for limiting QMFG systems appear in (Kolokoltsov, 2021; 2022) but no Nash solutions are computed there. We then prove self-consistency of the symmetric Nash equilibrium at $p = 1/2$ via a reflection-symmetry argument (Section 6), and exhibit numerical Nash equilibria across $p \in \{0.25, 0.5, 0.75\}$. We also remark few related work beyond that cited above. First, the QMFGs are inherently connected to partially observed MFGs in the non-linear settings (Şen and Caines, 2016). Also, reinforcement learning for single-system quantum feedback (Fösel et al., 2018) solves a related problem (single system, no mean-field coupling) and deep RL has more recently been applied to feedback control of continuously-monitored qubits (Sivak et al., 2022; Porotti et al., 2022; Borah et al., 2021), again without mean-field coupling, and a per-type model-free RL solver would not consider the analytical $\alpha^\star$ in (3) and the consistency structure we exploit in Section 4. Mean-field limits of Belavkin filtering have been studied at the level of the filtering equations (Kolokoltsov, 2023), and multi-population structure is standard in classical MFG (Cirant, 2015; Bensoussan et al., 2013). Classical-MFG trajectory-based deep-BSDE methods (Han et al., 2018; Carmona and Laurière, 2021; Ruthotto et al., 2020; Sirignano and Spiliopoulos, 2018; Hu, 2020) do not address the Belavkin-filter noise scaling. The broader observation that minimizing the Bellman/temporal-difference residual is not equivalent to minimizing value error originates in the double-sampling analysis of (Baird, 1995) and has been documented in deep RL (Geist et al., 2017; Fujimoto et al., 2022) and is the subject of variance-reduction work in deep-BSDE methods (Yang and Li, 2024). Our contribution complements that line: rather than proposing a variance-reduction trick or a tabular cancellation argument, we identify a structural geometric failure mode for continuous-time SDEs, characterize it by a closed-form alignment factor $|\nabla V \cdot \sigma|$, and verify the prediction by rotating that factor on a controlled quantum benchmark. **Notation.** We use $\gamma \in \mathbb{C}^{2\times 2}$ for a qubit density matrix, parameterized by Bloch coordinates $(x, y, z) \in \mathbb{R}^3$ with $\gamma = \frac{1}{2}(I + x\sigma_x + y\sigma_y + z\sigma_z)$. The Pauli matrices are $\sigma_x, \sigma_y, \sigma_z$. For the two-type setup, $\theta \in \{A, B\}$ indexes the type with prior probability p and $1 - p$ respectively, with targets ρ_g^A ($z = -1$) and ρ_g^B ($z = +1$). The shared mean-field trajectory is $\bar{z}_t \in \mathbb{R}$; per-type trajectories are $\bar{z}_t^\theta$. The control is $\alpha \in \mathbb{R}$; the cost is $\mathcal{J}$; the value network is V_ϑ (subscript ϑ denotes network parameters, not type). We write $\mathbb{E}$ for expectation, Δt for timestep, dW for the Belavkin innovation.

2. Problem setup

We adopt the qubit Ising benchmark of (Chalal et al., 2026). In the mean-field limit, a typical agent carries state $\gamma_t \in \mathbb{C}^{2\times 2}$ (a qubit density matrix), evolving under continuous σ_z-measurement with control $\alpha_t \in \mathbb{R}$ entering the Hamiltonian as $H = \alpha\sigma_x + J\,\bar{z}_t\,\sigma_z$ (we take coupling $J = 1$ throughout), where $\bar{z}_t = \mathrm{tr}(\sigma_z\gamma_t)$ is the population-average z-coordinate and $\sigma_x, \sigma_y, \sigma_z$ are the Pauli matrices. The Belavkin SDE in Bloch coordinates $r = (x, y, z) \in \mathbb{R}^3$, defined by $\gamma = \frac{1}{2}(I + x\sigma_x + y\sigma_y + z\sigma_z)$, reads

$$\begin{aligned} dx &= (-2\bar{z}\,y - 2x)\,dt - 2xz\,dW, \\ dy &= (2\bar{z}\,x - 2y - 2\alpha z)\,dt - 2yz\,dW, \\ dz &= 2\alpha y\,dt + 2(1 - z^2)\,dW, \end{aligned} \tag{1}$$

with dW the Belavkin innovation. The state space is the closed Bloch ball $\overline{B}_1 = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq 1\}$, reflecting positivity and unit trace of γ; the dynamics (1) preserve $\overline{B}_1$ in continuous time. The target is the computational-basis state $\rho_g = \frac{1}{2}(I - \sigma_z)$ (i.e. $z = -1$); the infidelity/running state cost is $\ell(\gamma) = 1 - \mathrm{tr}(\rho_g\gamma) = (1 + z)/2$. The control problem minimizes, over $\alpha(\cdot)$,

$$\mathcal{J}(\alpha) = \mathbb{E}\left[\int_0^T \left(\ell(\gamma_t) + \tfrac{\lambda_u}{2}\alpha_t^2\right) dt + \ell(\gamma_T)\right], \tag{2}$$

with $\lambda_u = 0.5$, $T = 5$ throughout this paper. We take admissible controls to be measurable feedbacks $\alpha : [0, T] \times \overline{B}_1 \to \mathbb{R}$, adapted to the filtration generated by $(\gamma_s)_{s\leq t}$, with $\mathbb{E}\int_0^T \alpha_t^2\,dt < \infty$. The value function $V(t, \gamma, \bar{z})$ satisfies an HJB equation. From (1), the α-dependent part of the generator is $2\alpha(y\,\partial_z V - z\,\partial_y V)$, so the α-dependent part of the HJB objective is $2\alpha(y\,\partial_z V - z\,\partial_y V) + \frac{\lambda_u}{2}\alpha^2$. Setting the derivative in α to zero yields $\lambda_u\,\alpha + 2(y\,\partial_z V - z\,\partial_y V) = 0$:

$$\alpha^\star(t, \gamma, \bar{z}) = \frac{2}{\lambda_u}(z\,\partial_y V - y\,\partial_z V). \tag{3}$$

The mean-field consistency condition is $\bar{z}_t = \mathbb{E}[z_t]$ under the feedback $\alpha^\star$ itself: a fixed-point requirement that our fictitious-play loop (Section 4) solves numerically. For the two-type extension developed in Section 5, each agent has a type $\theta \in \{A, B\}$ with prior probability $(p, 1 - p)$. Type

A targets $\rho_g^A = \frac{1}{2}(I - \sigma_z)$ $(z = -1)$; type B targets $\rho_g^B = \frac{1}{2}(I + \sigma_z)$ $(z = +1)$. Both types see the same *combined* mean field $\bar{z}_t = p\,\bar{z}_t^A + (1-p)\,\bar{z}_t^B$. A Nash equilibrium is a pair $(\alpha^{A,\star}, \alpha^{B,\star})$ of feedbacks such that each $\alpha^{\theta,\star}$ is optimal against the combined mean field generated by the two populations applying these feedbacks: per-type HJB solves coupled through a shared $\bar{z}_t$. Both Bellman and PINN training use Euler–Maruyama integration with a Bloch-ball projection step (radius 0.999) after each Euler update. The same projection is used during evaluation rollouts. For the Bellman residual (4) the collocation distribution is restricted to $t \in [0, T - \Delta t]$ to avoid evaluating $V_\vartheta(t + \Delta t, \cdot)$; the PINN residual is pointwise and uses $t \in [0, T]$.

3. HJB residual choice for Belavkin SDEs

We parameterize $V_\vartheta(t, x, y, z) : [0, T] \times \mathbb{R}^3 \to \mathbb{R}$ as an MLP with two hidden layers of width 64 and tanh activations, softly constrained to the Bloch ball ($|r| \le 1$, where $r = (x, y, z)$). Optimal control of a single continuously-monitored qubit via the Belavkin HJB equation goes back to (van Handel et al., 2005); physics-informed neural networks have recently been applied to quantum control problems (Norambuena et al., 2024). Two loss formulations for the HJB equation are common, corresponding to the two broad families of deep PDE solvers: BSDE-type methods (Han et al., 2018; Carmona and Laurière, 2021) that work with one-step Bellman backups along simulated trajectories, and PINN/DGM-type methods (Sirignano and Spiliopoulos, 2018) that enforce the PDE pointwise via automatic differentiation.

3.1. The two residuals

The *discrete Bellman residual* at (t, γ) is estimated by

$$\widehat{\mathcal{R}}_{\mathrm{B},n}[V_\vartheta] = V_\vartheta(t,\gamma) - \frac{1}{n}\sum_{i=1}^{n} V_\vartheta(t+\Delta t,\, \gamma_i') - C(\gamma, \alpha^\star)\Delta t, \tag{4}$$

where $C(\gamma, \alpha) := \ell(\gamma) + \frac{\lambda_u}{2}\alpha^2$ is the running cost integrand from (2), and each γ_i' is one Euler–Maruyama step of (1) under the current $\alpha^\star = \alpha^\star(V_\vartheta)$ and an independent innovation dW_i. This is the random n-sample Monte Carlo estimator of the exact one-step Bellman residual. The *PDE-form (PINN) residual* is the pointwise HJB PDE:

$$\mathcal{R}_{\mathrm{P}}[V_\vartheta] = \partial_t V_\vartheta + \mathcal{L}^{\alpha^\star} V_\vartheta + C(\gamma, \alpha^\star), \tag{5}$$

with infinitesimal generator $\mathcal{L}^\alpha V = \mu \cdot \nabla V + \frac{1}{2}\sigma^\top(\nabla^2 V)\sigma$ of (1), where $\mu(\gamma, \alpha, \bar{z}) \in \mathbb{R}^3$ and $\sigma(\gamma) \in \mathbb{R}^3$ collect the drift and diffusion vectors of (1) (so $dr = \mu\, dt + \sigma\, dW$ in compact form), evaluated through V_ϑ by automatic differentiation. Training minimizes the mean-square residual over collocation points $(t, \gamma) \in [0, T] \times$ Bloch ball, with terminal-condition loss $\|V_\vartheta(T, \gamma) - \ell(\gamma)\|^2$ and a soft positivity penalty. The penalty term $\lambda_+ \max(0, -V_\vartheta)^2$ in line 7 enforces nonnegativity of the *value* V_ϑ, not of any Bloch-ball coordinate: the cost-to-go is an integral of the nonnegative running cost $\ell \ge 0$ plus the nonnegative terminal cost, so $V^\star \ge 0$ pointwise, and the soft penalty discourages the optimizer from settling in spurious negative-value regions far from the terminal condition. It is a mild regularizer, active only where $V_\vartheta < 0$; the trained networks satisfy it with negligible penalty contribution at convergence.

Algorithm 1 PDE-form HJB training loop (one outer FP)

Require: Mean-field trajectory $\bar{z}(\cdot)$, training steps S, batch size B, learning rate η

1: Initialize MLP parameters ϑ_0
2: **for** $s = 0, 1, \ldots, S-1$ **do**
3: Sample collocation $\{(t_b, \gamma_b)\}_{b=1}^B$ uniformly in $[0, T] \times$ Bloch ball
4: Sample terminal points $\{(\gamma_b^T)\}_{b=1}^B$ uniformly in Bloch ball
5: Compute $\partial_t V_\vartheta, \nabla V_\vartheta, \nabla^2 V_\vartheta$ at each (t_b, γ_b) via autodiff
6: Form $\alpha_b^\star$ from (3) and the PDE residual $\mathcal{R}_{\mathrm{P}}[V_\vartheta](t_b, \gamma_b)$ from (5)
7: Loss $\mathcal{L}(\vartheta_s) = \frac{1}{B}\sum_b \mathcal{R}_{\mathrm{P}}^2 + \frac{1}{B}\sum_b (V_\vartheta(T, \gamma_b^T) - \ell(\gamma_b^T))^2 + \lambda_+ \max(0, -V_\vartheta)^2$ (with $\lambda_+ > 0$ a fixed positivity-penalty weight)
8: $\vartheta_{s+1} \leftarrow \mathrm{Adam}(\vartheta_s, \nabla_\vartheta \mathcal{L}, \eta)$
9: **end for**
10: **return** V_{ϑ_S} and policy $\alpha^\star(\cdot; V_{\vartheta_S})$ via (3).

3.2. Why $\mathcal{R}_{\mathrm{B}}$ fails for Belavkin dynamics

The stochastic term σdW in the Belavkin SDE (1) is $O(\sqrt{\Delta t})$, while the deterministic HJB signal in one Bellman step is $O(\Delta t)$, i.e., the noise can swamp the signal. This is the classical double-sampling problem (Baird, 1995), made quantitative for HJB losses on classical SDEs by (Nüsken and Richter, 2021; Richter et al., 2022): variance scales as $|\nabla V \cdot \sigma|^2$. Proposition 1 specializes this to Belavkin filtering; the new ingredient is the closed-form *alignment factor* $|\nabla V \cdot \sigma|$, which predicts *when* Bellman fails and is verified by rotating the measurement axis (Table 4).

Proposition 1 (Alignment-controlled SNR of the Bellman residual)**.** *Let V be smooth and let $\widehat{\mathcal{R}}_{B,n}[V]$ be the n-sample Monte Carlo Bellman residual* (4) *for the Belavkin SDE* (1) *with scalar Wiener innovation. Assume sufficient smoothness of V and the SDE coefficients on the collocation domain so that the standard one-step weak Itô–Taylor expansion has remainder $O(\Delta t^2)$. Then, with σ the state-*

dependent diffusion vector,

$$\mathbb{E}[\widehat{\mathcal{R}}_{B,n}[V]] = -\mathcal{R}_P[V]\,\Delta t + O(\Delta t^2), \tag{6}$$

$$\mathrm{Var}[\widehat{\mathcal{R}}_{B,n}[V]] = \frac{(\nabla V \cdot \sigma)^2}{n}\,\Delta t + O(\Delta t^2), \tag{7}$$

yielding signal-to-noise

$$\mathrm{SNR}(\widehat{\mathcal{R}}_{B,n}[V]) = \frac{|\mathcal{R}_P[V]|}{|\nabla V \cdot \sigma|}\sqrt{n\,\Delta t}. \tag{8}$$

Here $\nabla V \cdot \sigma = \sum_j (\partial_j V)\,\sigma_j \in \mathbb{R}$ is the scalar inner product of the value gradient with the diffusion vector $\sigma \in \mathbb{R}^3$ of (1)*, and $|\nabla V \cdot \sigma|$ denotes its absolute value (equivalently, $(\nabla V \cdot \sigma)^2$ in* (7) *is the squared scalar projection of ∇V onto σ). Treating the induced feedback consistently and assuming the true value function $V^\star$ lies in the hypothesis class, any population-loss minimizer $\hat{V}$ of $\mathbb{E}[\widehat{\mathcal{R}}^2_{B,n}]$ satisfies the asymptotic noise-floor bound*

$$\mathbb{E}[\mathcal{R}_P[\hat{V}]^2] \le \frac{1}{n\,\Delta t}\,\mathbb{E}[(\nabla V^\star \cdot \sigma)^2] + O(\Delta t). \tag{9}$$

For fixed n, the leading bound worsens as $\Delta t \to 0$: shrinking the timestep dilutes the deterministic residual signal relative to the innovation noise floor.

Sketch. Let $\delta = \mu\,\Delta t + \sigma\,dW$ with $dW \sim \mathcal{N}(0, \Delta t)$. The standard weak Itô–Taylor expansion (under the stated smoothness) gives $V(t{+}\Delta t, \gamma{+}\delta) = V + \partial_t V \Delta t + \nabla V \cdot \delta + \frac{1}{2}\delta^\top \nabla^2 V \delta + R$, with $\mathbb{E}[R] = O(\Delta t^2)$. Taking expectation, using $\mathbb{E}[dW] = 0$ and $\mathbb{E}[\delta_i \delta_j] = \sigma_i \sigma_j \Delta t + O(\Delta t^2)$,

$$\begin{aligned}\mathbb{E}[V(t+\Delta t, \gamma')] = V + (&\partial_t V + \mu \cdot \nabla V \\ &+ \tfrac{1}{2}\sigma^\top \nabla^2 V \sigma)\,\Delta t + O(\Delta t^2).\end{aligned}$$

Substituting into (4) and noting $\partial_t V + \mu \cdot \nabla V + \frac{1}{2}\sigma^\top \nabla^2 V \sigma + C = \mathcal{R}_{\mathrm{P}}[V]$ gives (6). Now, decompose $\widehat{\mathcal{R}}_{\mathrm{B},n}[V] - \mathbb{E}[\widehat{\mathcal{R}}_{\mathrm{B},n}[V]] = -\frac{1}{n}\sum_{k=1}^n (\nabla V \cdot \sigma) dW_k + \mathcal{O}$, where $\mathcal{O}$ collects centered higher-order Itô–Taylor terms. The leading term has variance $(\nabla V \cdot \sigma)^2 \Delta t / n$; centered quadratic terms in $dW^2 - \Delta t$ contribute $O(\Delta t^2/n)$. This proves (7), and (8) follows by taking the ratio. Moreover, at $V^\star$ we have $\mathcal{R}_{\mathrm{P}}[V^\star] = 0$, hence $\mathbb{E}[\widehat{\mathcal{R}}_{\mathrm{B},n}[V^\star]^2] = \frac{\Delta t}{n}\mathbb{E}[(\nabla V^\star \cdot \sigma)^2] + O(\Delta t^2)$. A population-loss minimizer $\hat{V}$ has loss bounded by this quantity. On the other hand, Jensen's inequality and the bias expansion (6) give $\mathbb{E}[\widehat{\mathcal{R}}_{\mathrm{B},n}[\hat{V}]^2] \ge \Delta t^2\,\mathbb{E}[\mathcal{R}_{\mathrm{P}}[\hat{V}]^2] + O(\Delta t^3)$. Combining and dividing by Δt^2 yields (9). □

Corollary 1 (Bellman–PINN consistency under small alignment)**.** *If $\mathbb{E}[(\nabla V^\star \cdot \sigma)^2] \le \varepsilon^2 \cdot n\Delta t$ on the collocation distribution, then any minimizer $\hat{V}$ of $\mathbb{E}[\widehat{\mathcal{R}}^2_{B,n}]$ satisfies $\mathbb{E}[\mathcal{R}_P[\hat{V}]^2] \le \varepsilon^2 + O(\Delta t)$. In the orthogonal limit $\nabla V^\star \perp \sigma$ (a.e. on the collocation distribution), the Bellman minimizer is also a PINN minimizer to leading order.*

Proof. Substitute the hypothesis into (9). □

We note that the bound (9) is the central object: the achievable PDE-residual of the Bellman minimizer is bounded by $\mathbb{E}[(\nabla V^\star \cdot \sigma)^2]/(n\Delta t)$ to leading order. Where this alignment is small, $\hat{V}$ approximately satisfies the HJB; where it is large, $\hat{V}$ can be a *spurious minimizer*, i.e., far from any HJB solution while still minimizing $\mathbb{E}[\widehat{\mathcal{R}}^2_{\mathrm{B},n}]$. Decreasing Δt *worsens* the floor (the $1/(n\Delta t)$ factor grows as $\Delta t \to 0$), so Δt-shrinkage cannot rescue Bellman; we verify this across $\Delta t \in \{0.1, 0.05, 0.02, 0.005\}$. We further tested standard variance-reduction and higher-order integration variants (Table 1): antithetic sampling, multi-step ($K{=}5$) backups, and a Milstein discretization all leave the trained policy within $\approx 5\%$ of zero control. This is consistent with Proposition 1: these variants reduce variance or discretization error but do not remove the $(\nabla V \cdot \sigma)^2$ term that sets the floor. A control-variate subtracting the $\sigma^\top \nabla V$ innovation (as in the martingale / Z-process deep-BSDE formulation (Han et al., 2018; Yang and Li, 2024)) targets that term directly and would be expected to recover PINN performance. We leave this baseline to future work. In our benchmark (σ_z measurement, target $z = -1$), σ has a strong z-component throughout the bulk of the Bloch ball, parallel to ∇V, placing us in a structurally adverse regime for Proposition 1. We confirm the geometric prediction empirically: on a structurally distinct benchmark with σ_x measurement (so the diffusion is dominantly in x, while ∇V still points along z), the alignment factor $|\nabla V \cdot \sigma|$ is much smaller in the bulk and Bellman recovers, training a policy nearly matching PINN's. The PDE-form residual avoids this particular one-step innovation Monte Carlo trade-off: $\mathcal{R}_{\mathrm{P}}$ is computed through automatic differentiation, although the method still has collocation-sampling, approximation, rollout, and optimization error.

Table 1. One-step Bellman variants on the aligned baseline ($\Delta t = 0.05$). Reference: zero control $= 2.96$, PINN $= 1.32$. Antithetic, multi-step, and Milstein variants reduce variance or discretization error but remain within $\approx 5\%$ of zero control: none escape the alignment-controlled noise floor of Proposition 1.

Bellman variant	$\mathbb{E}[\mathcal{J}]$	gap vs. zero
Vanilla one-step	3.00	+0.04
Antithetic variate	2.87	−0.09
Multi-step ($K{=}5$)	3.11	+0.16
Milstein scheme	3.00	+0.04
PINN (reference)	1.32	−1.64

Remark 1 (Proposition 1 and relevance to Belavkin filtering)**.** *The SNR and noise-floor results apply to any Itô SDE with $\sigma\,dW = O(\sqrt{\Delta t})$; nothing quantum is required, and the operative quantity is $|\nabla V \cdot \sigma|$. Discrete Bellman/deep-BSDE training has succeeded on classical benchmarks (Han et al., 2018; Carmona and Laurière, 2021; Ruthotto et al.,*

2020; Sirignano and Spiliopoulos, 2018) consistent with small alignment there (we do not claim this is universal). Belavkin-filtered stabilization toward an eigenstate is structurally adverse: σ encodes measurement back-action along the measured observable and the cost depends on that same observable, so $|\nabla V \cdot \sigma|$ is large in the bulk. Table 4 isolates this mechanism: rotating the measurement axis to σ_x while keeping the cost on z breaks the alignment, and Bellman recovers. Result is confined only to the single-qubit Ising mean-field family.

3.3. Experimental demonstration

We train the same MLP under both residuals with identical hyperparameters (Adam, lr 10^{-3}, 1500 steps, batch 128, same initialization) across three problem variants: (a) baseline with $T = 5$, initial state $\gamma_0 = I/2$ ($z_0 = 0$); (b) *hard start* from $z_0 = 0.8$, far from target; (c) *short horizon* with $T = 3$. The Bellman variant uses $\Delta t = 0.1$, $n_{\text{noise}} = 2$. Rolled out over $K = 2048$ trajectories in each case, with baselines of zero control ($\alpha \equiv 0$) and the Lyapunov feedback. Figure 1 shows the qualitative outcome and Table 2 gives the multi-seed quantitative result. In every variant, the Bellman policy lies within $\approx 3\%$ of zero control (mean across five seeds): 2.92 vs 2.85 on baseline, 5.32 vs 5.33 on hard-start, 1.96 vs 1.90 on short-horizon. The PINN policy is better in every case (1.10, 1.56, 0.97 respectively), competitive with or slightly above the single-Lyapunov baseline (1.33, 2.24, 1.21). The hard-start variant exhibits the spurious-minimizer mechanism: Bellman fits the discrete backup but the resulting policy is no better than zero control, indicating V_ϑ satisfies the discrete equation while not encoding an admissible cost-to-go. To rule out a fortunate-seed explanation, Table 2 reports rollout cost as mean $\pm$ standard deviation across five independent training seeds (each seed re-initializing both the network and the optimizer key) on the same three variants. Bellman variance across seeds is ≤ 0.06 in every variant. PINN variance is 0.05–0.08, reflecting a well-converged regime.

Table 2. Mean $\pm$ standard deviation of $\mathbb{E}[\mathcal{J}]$ over five seeds. Zero control and Lyapunov are deterministic baselines.

Variant	Zero	Lyap	Bellman	PINN
baseline	2.85	1.33	2.92 ± 0.05	1.10 ± 0.08
hard_start	5.33	2.24	5.31 ± 0.03	1.56 ± 0.08
short_horizon	1.90	1.21	1.96 ± 0.02	0.97 ± 0.05

Proposition 1 predicts that the noise floor on $\mathbb{E}[\mathcal{R}_{\text{P}}^2]$ scales as $1/(n\,\Delta t)$, so increasing n_{noise} should help and decreasing Δt should hurt. Table 3 sweeps n_{noise}, Δt, and collocation batch size on the baseline variant. Across all 15 conditions, Bellman's rollout cost stays in $[2.83, 3.20]$, against zero control 2.85 and PINN 1.14. Larger n does not bring Bellman to PINN performance because the bound is $\propto 1/(n\Delta t)$ at fixed Δt and we are far from the regime where that quotient becomes small enough to matter; smaller Δt does not help either, consistent with the bound *worsening* as $\Delta t \to 0$. Batch size has no effect, since it controls Monte Carlo error of the empirical loss but not the per-point noise floor.

Table 3. Bellman ablations on the baseline variant. Reference: zero $= 2.85$, PINN $= 1.14$. Default cell unless varied: $\Delta t = 0.1$, $n_{\text{noise}} = 2$, batch $= 128$.

(a) n_{noise} sweep ($\Delta t = 0.1$, batch $= 128$)					
n_{noise}	2	4	8	16	32
$\mathbb{E}[\mathcal{J}]$	2.92	2.94	2.94	2.94	2.83
(b) Δt sweep ($n_{\text{noise}} = 2$, batch $= 128$)					
Δt	0.20	0.10	0.05	0.02	0.01
$\mathbb{E}[\mathcal{J}]$	3.00	2.92	3.20	3.05	2.92
(c) batch size sweep ($\Delta t = 0.1$, $n_{\text{noise}} = 2$)					
batch B	32	64	128	256	512
$\mathbb{E}[\mathcal{J}]$	2.92	2.92	2.92	2.94	2.91

Scaling up the PINN solver (width 128, 5000 steps, batch 256) on the baseline variant gives $\mathbb{E}[\mathcal{J}_{\text{PDE, fixed-}\bar{z}}] \approx 1.06$. A grid search over the two-parameter Lyapunov family $\alpha = \kappa_1(1+z)/2 + \kappa_2 x$ on $\{1, 2, 4, 8\} \times \{0.5, 1, 2\}$ identifies $(\kappa_1, \kappa_2) = (2.0, 0.5)$ as the best, with $\mathbb{E}[\mathcal{J}] \approx 0.97$. Fictitious play (Section 4) brings the PINN-HJB+FP cost to ≈ 0.99. The difference from the best hand-tuned Lyapunov controller (0.99 vs 0.97) is within the across-seed standard deviation reported in Table 2 (≤ 0.08 for PINN), so we do not claim the learned policy *beats* the Lyapunov baseline; the appropriate reading is that PINN-HJB+FP *attains the analytical baseline without manual controller design*. Its value over Lyapunov in this single-type case is therefore not absolute cost reduction but (i) yielding a value function, which the multi-type Nash extension below requires, and (ii) not requiring problem-specific manual tuning of a control ansatz. To test the geometric prediction of Proposition 1 directly, we construct a benchmark with σ_x measurement and σ_y control (target $z = -1$, same cost (2)). Here the diffusion vector σ is dominantly along x while ∇V remains along z, so $|\nabla V \cdot \sigma|$ is small in the bulk of the Bloch ball. Proposition 1/Corollary 1 predict that Bellman should recover on this variant. Table 4 confirms this.

Table 4. Bellman vs PINN on two benchmarks. **Original**: σ_z measurement, target $z = -1$; σ aligned with ∇V in the bulk (large $|\nabla V \cdot \sigma|$). **Cross**: σ_x measurement, σ_y control, target $z = -1$; σ approximately orthogonal to ∇V (small $|\nabla V \cdot \sigma|$). Other hyperparameters identical. The Bellman–PINN gap shrinks by an order of magnitude, consistent with Proposition 1; a direct evaluation of $\mathbb{E}[(\nabla V^\star \cdot \sigma)^2]$ is left to future work.

Benchmark	Zero	Bellman	PINN	Gap
Original (σ_z meas.)	2.85	2.92	1.10	1.82
Cross (σ_x meas.)	3.00	2.30	2.22	0.08

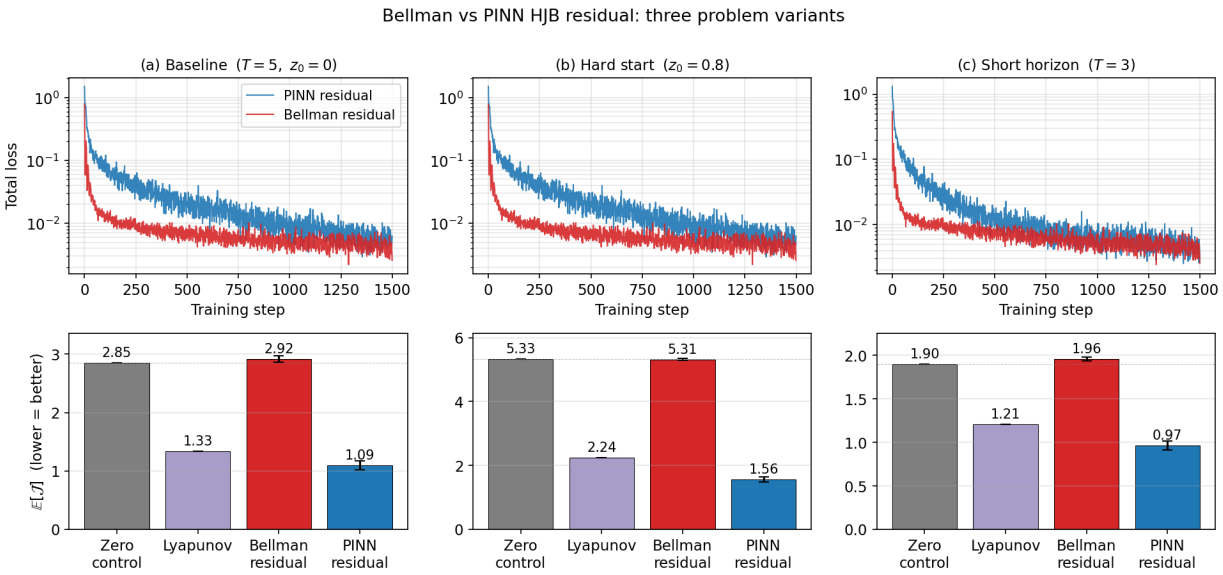

Figure 1. Bellman vs PDE-form HJB residual across three problem variants. **Top row:** both losses decay to comparable floors $\approx 10^{-3}$. **Bottom row:** rollout validation; bars show mean over five training seeds with error bars showing one standard deviation (Table 2 numbers). In all three variants, the Bellman policy matches zero control while the PINN policy is competitive with the Chalal Lyapunov baseline (Chalal et al., 2026). Same architecture, same hyperparameters, same initialization throughout.

Algorithm 2 Fictitious play for single-type QMFG

1: Initialize $\bar{z}^{(0)}(\cdot) \equiv 0$
2: **for** $k = 0, 1, \ldots$ **do**
3: Train $V_\vartheta^{(k)}$ with the PDE-form residual (5) using mean field $\bar{z}^{(k)}(\cdot)$
4: Roll out K trajectories under $\alpha^\star$ from $V_\vartheta^{(k)}$; compute $m^{(k)}(t) = \hat{\mathbb{E}}[z_t]$
5: Update $\bar{z}^{(k+1)}(t) = (1 - \eta_k)\, \bar{z}^{(k)}(t) + \eta_k\, m^{(k)}(t)$
6: Check $\|\bar{z}^{(k+1)} - \bar{z}^{(k)}\|_\infty < \epsilon_{\text{tol}}$
7: **end for**

Algorithm 3 Fictitious play for two-type QMFG

1: Initialize $\bar{z}^{(0)}(\cdot) \equiv 0$
2: **for** $k = 0, 1, \ldots$ **do**
3: Train $V_\vartheta^{A,(k)}$, $V_\vartheta^{B,(k)}$ independently against $\bar{z}^{(k)}(\cdot)$
4: Roll out $K_A = pK$ type-A and $K_B = (1-p)K$ type-B trajectories
5: $m^{(k)}(t) = p\, \hat{\mathbb{E}}_A[z_t] + (1-p)\, \hat{\mathbb{E}}_B[z_t]$
6: $\bar{z}^{(k+1)} = (1 - \eta_k)\bar{z}^{(k)} + \eta_k m^{(k)}$; check tolerance.
7: **end for**

On the cross-benchmark, the Bellman policy moves below zero control (2.30 vs 3.00, a 23% reduction) and lies within 4% of the PINN policy (2.22). The Bellman–PINN gap is roughly 23× smaller than on the original benchmark. This is a direct empirical confirmation of the geometric prediction: when alignment is small, Bellman recovers PINN performance to leading order, as Corollary 1 states.

4. Fictitious play for mean-field consistency

The single-type PDE-form HJB solve above assumes a fixed $\bar{z}$. The true MFG equilibrium requires $\bar{z}_t = \mathbb{E}[z_t]$ under the feedback derived from V itself: a fixed-point condition not satisfied by the fixed-$\bar{z}$ solution. We close this gap by fictitious play (FP): With $\eta_k = 1/(k+1)$ (classical FP averaging) and $\epsilon_{\text{tol}} = 10^{-2}$, Algorithm 2 converges within 5–6 iterations on our qubit Ising benchmark. Figure 2 shows the mean-field trajectory $\bar{z}^{(k)}(t)$ stabilizing and the achieved cost fluctuating but settling near ≈ 0.99 over the 5–6 FP iterations, starting from ≈ 1.14 at the fixed-$\bar{z}$ initialization. The converged value matches the best two-parameter Lyapunov controller within numerical noise (Section 3); the methodological value of FP here is that it produces a value function V_ϑ which the multi-type extension below requires.

5. Two-type Nash equilibrium

We now compute a Nash equilibrium for the two-type QMFG defined in Section 2. Algorithm 3 iterates per-type PDE-form HJB training against a shared mean field, each per-type policy best-responding to the current estimate of $\bar{z}$.

5.1. Symmetric case $p = 0.5$

As a sanity check, $p = 0.5$ should produce a Nash equilibrium by symmetry: since both types are equally popu-

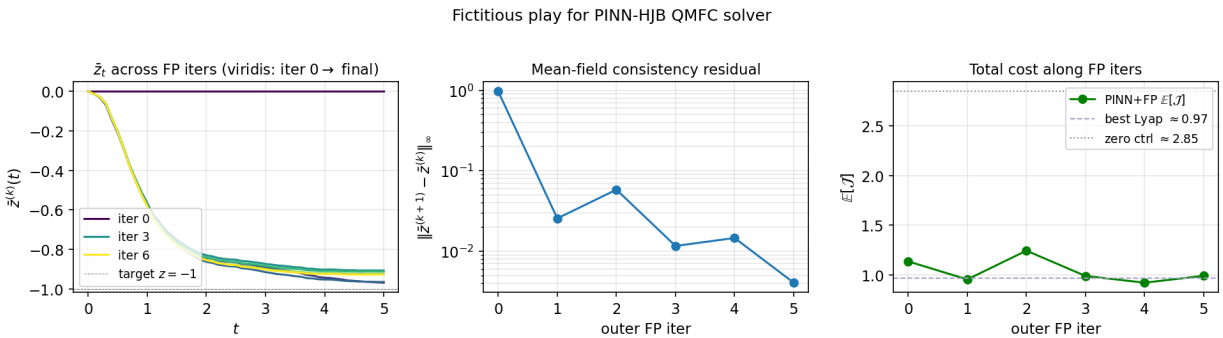

Figure 2. Fictitious play convergence for the single-type QMFG. **Left:** mean-field trajectory $\bar{z}^{(k)}(t)$ across FP iterations, converging to a consistent limit near the target $z = -1$. **Middle:** mean-field consistency residual $\|\bar{z}^{(k+1)} - \bar{z}^{(k)}\|_\infty$ per FP iteration, decaying roughly exponentially. **Right:** expected total cost $\mathbb{E}[\mathcal{J}]$ per FP iteration, settling near the best two-parameter Lyapunov baseline of ≈ 0.97.

lous and pull toward opposite targets, the combined mean field $\bar{z}$ should remain identically zero, and each type should solve its own single-type QMFG independently with mean field $\bar{z} \equiv 0$. Figure 3 confirms this: the combined $\bar{z}$ stays within ≈ 0.04 of zero through FP, and each per-type policy replicates the single-type optimal feedback. The achieved per-type costs are $\mathbb{E}[\mathcal{J}^A] \approx \mathbb{E}[\mathcal{J}^B] \approx 1.07$, comparable to the single-type FP-converged value of Section 4. The numerical exchange symmetry $\mathbb{E}[\mathcal{J}^A] \approx \mathbb{E}[\mathcal{J}^B]$ at $p = 1/2$ is the empirical signature of Proposition 2 below.

5.2. Asymmetric case: prior scan

Algorithm 3 runs at $p \in \{0.25, 0.50, 0.75\}$. Results are in Table 5. The qualitative pattern of Table 5 is consistent with our analytical results below: per-type costs are roughly comparable, with the minority type tending to pay slightly more in absolute terms, and the population-weighted total varies moderately ($\approx 13\%$) without a strong monotone dependence on p. Whether the population-weighted Nash cost is exactly p-invariant in the limit of optimization error tending to zero, rollout batch tending to infinity, and an FP-converged equilibrium remains an open empirical question (see Section 7). Figure 4 visualizes the prior scan.

Table 5. Two-type QMFG Nash equilibrium across priors. Final values from 4 FP iterations with population-proportional batches $K_A = pK$, $K_B = (1-p)K$, $K = 2048$. Inter-iteration oscillations ± 0.1 persist. Population-weighted total varies $\approx 13\%$. Let $\mathcal{J}^{A,B} := p\,\mathcal{J}^A + (1-p)\,\mathcal{J}^B$

p (prior A)	$\mathbb{E}[\mathcal{J}^A]$	$\mathbb{E}[\mathcal{J}^B]$	$\mathcal{J}^{A,B}$	$\max_t \lvert\bar{z}_t^\star\rvert$
0.25	0.99	0.96	0.97	0.49
0.50	1.12	1.01	1.07	0.04
0.75	1.15	0.92	1.10	0.48

6. Analytical results: summary

We close the empirical part of the paper with two analytical observations that anchor the prior-scan results, deferring full proofs to the appendix.

First, the symmetric Nash equilibrium at $p = 1/2$ is provable: an involution $\mathcal{R} : (x, y, z) \mapsto (x, -y, -z)$ exchanges types $A \leftrightarrow B$ while flipping the sign of the combined mean field, so $\bar{z}^\star \equiv 0$ is self-consistent (Lemma 1, Proposition 2 in Appendix A). Figure 3 confirms this numerically, with residual $|\bar{z}^\star| \lesssim 0.04$ attributable to finite rollout batch and FP tolerance. Second, the planner-versus-Nash gap (price of anarchy (Carmona et al., 2019)) admits a partial characterization. The planner value $J_{\text{MFC}}(p) \leq p\mathcal{J}^{A,\star}(p) + (1-p)\mathcal{J}^{B,\star}(p)$ trivially (the Nash feedback is one feasible joint control). At $p = 1/2$, symmetry implies stationarity of the reduced planner objective at $\bar{z} \equiv 0$, and a smoothness assumption forces the Taylor expansion of $J_{\text{MFC}}(p)$ around $p = 1/2$ to contain only even powers of $p - 1/2$ (Propositions 3–4 in Appendix A). The moderate variation of the Nash population-weighted cost across $p \in \{0.25, 0.5, 0.75\}$ ($\approx 13\%$) is consistent with this analytical structure; whether the second-order coefficient is exactly zero remains an open empirical question.

7. Discussion

We have shown that for Belavkin-filtered QMFG the one-step stochastic displacement $\sigma dW = O(\sqrt{\Delta t})$ can make the discrete Bellman HJB residual fail to train a useful value function (Tables 2, 3, Fig. 1); we proved the underlying SNR / noise-floor structure (Proposition 1) and confirmed its geometric prediction on a structurally distinct σ_x-measurement benchmark (Table 4). Using the PDE-form residual inside a fictitious-play loop yields a value function that, on the single-type Chalal et al. benchmark, matches the best two-parameter Lyapunov controller within numerical noise. The framework then extends to the two-type case where no Lya-

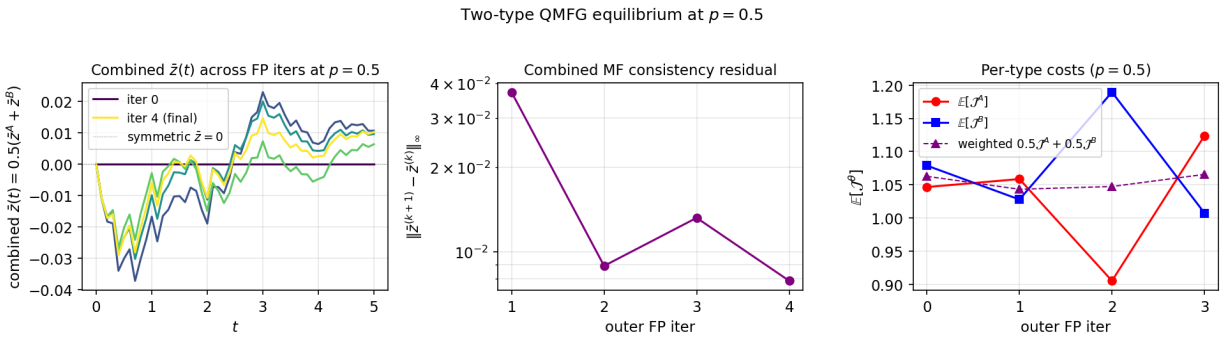

Figure 3. Two-type QMFG equilibrium at $p = 0.5$. **Left:** combined mean-field trajectory $\bar{z}(t) = 0.5(\bar{z}^A + \bar{z}^B)$ across FP iterations (viridis: iter 0 to final), remaining within ≈ 0.04 of zero throughout, confirming the symmetric Nash equilibrium. **Middle:** mean-field consistency residual $\|\bar{z}^{(k+1)} - \bar{z}^{(k)}\|_\infty$ per FP iteration. **Right:** per-type costs $\mathbb{E}[\mathcal{J}^A]$, $\mathbb{E}[\mathcal{J}^B]$ and the population-weighted total per FP iteration; type-symmetric values fluctuate around ≈ 1.05.

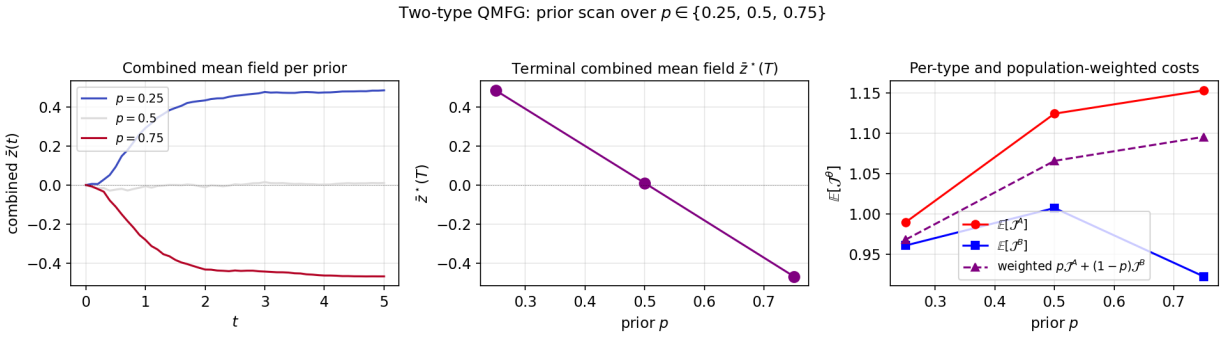

Figure 4. Two-type QMFG across prior $p \in \{0.25, 0.5, 0.75\}$. **Left:** combined mean-field trajectory $\bar{z}(t)$ at each p. **Middle:** terminal combined mean field $\bar{z}^\star(T)$ vs prior, passing approximately through zero at $p = 1/2$ as predicted by reflection symmetry. **Right:** per-type and population-weighted costs across priors (numerical values match Table 5).

punov closed form is known, exhibiting numerical Nash equilibria across $p \in \{0.25, 0.5, 0.75\}$ (Table 5). We prove the symmetric Nash equilibrium at $p = 1/2$ (Proposition 2), prove stationarity of the symmetric *planner* (mean-field-control) objective (Corollary 2), and show that the *planner* value $J_{\text{MFC}}(p)$ has no linear correction around $p = 1/2$ (Proposition 4). We stress that Proposition 4 concerns the planner value, not the Nash value: the empirical observation that the population-weighted *Nash* cost is flat to within $\approx 13\%$ across $p \in \{0.25, 0.5, 0.75\}$ is a separate, stronger statement that our symmetry argument does not establish. As limitations and directions for future work, we note that the benchmark is single-qubit Ising mean-field with σ_z measurement and σ_x control; the innovation-noise SNR argument is asymptotic, and fictitious-play convergence is observed empirically without a theorem. Companion work (Şen, 2026) develops a permutation-invariant DeepSet architecture for finite-N QMFG using the same HJB-learning machinery; the PDE-form residual is similarly preferred there for the aligned-measurement regime. Extensions to $K > 2$ types, asymmetric Nash, and qutrit QMFG would test whether the observed zero planner gap survives beyond the symmetric benchmark.

Two open questions remain. *(i) Magnitude of β on this benchmark.* Second-order HJB expansion gives $\beta = \mathrm{I}_1 + \mathrm{I}_2$ with $\mathrm{I}_2 = \frac{1}{2\lambda_u}\int_0^T \mathbb{E}[(z\partial_y V_1 - y\partial_z V_1)^2]\,dt \geq 0$ and I_1 unsigned. Whether β is small (consistent with the relative flatness of $p\,\mathcal{J}^A + (1-p)\mathcal{J}^B$ in Table 5) or whether the observed flatness is a finite-resolution artifact of FP convergence and rollout sampling remains open. *(ii) Robustness beyond symmetry.* Do the qualitative features of Table 5 survive asymmetric costs or Hamiltonians? The reflection symmetry in Lemma 1 breaks in those settings, so neither Proposition 2 nor Proposition 4 would apply.

Acknowledgments

The author thanks C3 AI for institutional support. Anthropic Claude was used to assist with LaTeX, code refactoring, and proof drafting; the author verified all results and takes full responsibility for the content.

References

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*, pages 30–37, 1995.

Viacheslav P. Belavkin. Nondemolition measurement and control in quantum dynamical systems. In *Information Complexity and Control in Quantum Physics*, CISM Courses and Lectures. Springer, 1987.

Alain Bensoussan, Jens Frehse, and Phillip Yam. *Mean Field Games and Mean Field Type Control Theory*. Springer, 2013.

Sangkha Borah, Bijita Sarma, Michael Kewming, Gerard J. Milburn, and Jason Twamley. Measurement-based feedback quantum control with deep reinforcement learning for a double-well nonlinear potential. *Physical Review Letters*, 127(19):190403, 2021.

Luc Bouten, Ramon van Handel, and Matthew R. James. An introduction to quantum filtering. *SIAM Journal on Control and Optimization*, 46(6):2199–2241, 2007. doi: 10.1137/060651239.

René Carmona and Mathieu Laurière. Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games I: The ergodic case. *SIAM Journal on Numerical Analysis*, 59(3):1455–1485, 2021. doi: 10.1137/19M1274377.

René Carmona, Christy V. Graves, and Zongjun Tan. Price of anarchy for mean field games. *ESAIM: Proceedings and Surveys*, 65:349–383, 2019. doi: 10.1051/proc/201965349. arXiv:1802.04644.

Sofiane Chalal, Nina H. Amini, Hamed Amini, and Mathieu Laurière. Dynamic programming principle and stabilization for mean-field quantum filtering systems. *arXiv preprint arXiv:2602.12472*, 2026. Preprint, Feb. 2026.

Marco Cirant. Multi-population mean field games systems with Neumann boundary conditions. *Journal de Mathématiques Pures et Appliquées*, 103(5):1294–1315, 2015.

Nevroz Şen. Function-space value networks for finite-N quantum mean-field games. Under review (anonymized for double-blind submission), 2026.

Nevroz Şen and Peter E. Caines. Mean field game theory with a partially observed major agent. *SIAM Journal on Control and Optimization*, 54(6):3174–3224, 2016. doi: 10.1137/16M1063010. URL `https://doi.org/10.1137/16M1063010`.

Thomas Fösel, Petru Tighineanu, Talitha Weiss, and Florian Marquardt. Reinforcement learning with neural networks for quantum feedback. *Physical Review X*, 8(3):031084, 2018. doi: 10.1103/PhysRevX.8.031084.

Scott Fujimoto, David Meger, Doina Precup, Ofir Nachum, and Shixiang Shane Gu. Why should I trust you, Bellman? The Bellman error is a poor replacement for value error. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 6918–6943, 2022.

Matthieu Geist, Bilal Piot, and Olivier Pietquin. Is the Bellman residual a bad proxy? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Jiequn Han, Arnulf Jentzen, and Weinan E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018. doi: 10.1073/pnas.1718942115.

Ruimeng Hu. Deep fictitious play for stochastic differential games. In *Proceedings of Machine Learning Research*, volume 107, pages 221–245, 2020.

Minyi Huang, Roland P. Malhamé, and Peter E. Caines. Large population stochastic dynamic games: closed-loop McKean–Vlasov systems and the Nash certainty equivalence principle. *Communications in Information and Systems*, 6(3):221–252, 2006.

Vassili N. Kolokoltsov. Quantum mean-field games with the observations of counting type. *Games*, 12(1):7, 2021. doi: 10.3390/g12010007.

Vassili N. Kolokoltsov. Quantum mean-field games. *The Annals of Applied Probability*, 32(3):2254–2288, 2022. doi: 10.1214/21-AAP1733.

Vassili N. Kolokoltsov. Mean-field limits of quantum filtering and control. *arXiv preprint arXiv:2303.09667*, 2023.

Jean-Michel Lasry and Pierre-Louis Lions. Jeux à champ moyen. I – Le cas stationnaire. *Comptes Rendus Mathématique. Académie des Sciences. Paris*, 343(9):619–625, 2006. doi: 10.1016/j.crma.2006.09.019.

Ariel Norambuena, Marios Mattheakis, Francisco A. González, and Raúl Coto. Physics-informed neural networks for quantum control. *Physical Review Letters*, 132(1):010801, 2024.

Nikolas Nüsken and Lorenz Richter. Solving high-dimensional Hamilton–Jacobi–Bellman PDEs using neural networks: perspectives from the theory of controlled

diffusions and measures on path space. *Partial Differential Equations and Applications*, 2(4):48, 2021. doi: 10.1007/s42985-021-00102-x.

Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Élie, and Olivier Pietquin. Fictitious play for mean field games: continuous time analysis and applications. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 13199–13213, 2020.

Riccardo Porotti, Antoine Essig, Benjamin Huard, and Florian Marquardt. Deep reinforcement learning for quantum state preparation with weak nonlinear measurements. *Quantum*, 6:747, 2022.

Lorenz Richter, Leon Sallandt, and Nikolas Nüsken. Robust SDE-based variational formulations for solving linear PDEs via deep learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 18649–18666, 2022.

Lars Ruthotto, Stanley J. Osher, Wuchen Li, Levon Nurbekyan, and Samy Wu Fung. A machine learning framework for solving high-dimensional mean field game and mean field control problems. *Proceedings of the National Academy of Sciences*, 117(17):9183–9193, 2020. doi: 10.1073/pnas.1922204117.

Justin Sirignano and Konstantinos Spiliopoulos. DGM: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339–1364, 2018. doi: 10.1016/j.jcp.2018.08.029.

V. V. Sivak, A. Eickbusch, H. Liu, B. Royer, I. Tsioutsios, and M. H. Devoret. Model-free quantum control with reinforcement learning. *Physical Review X*, 12(1):011059, 2022.

Ramon van Handel, John K. Stockton, and Hideo Mabuchi. Feedback control of quantum state reduction. *IEEE Transactions on Automatic Control*, 50(6):768–780, 2005.

Jiefei Yang and Guanglian Li. A deep primal-dual BSDE method for optimal stopping problems. *arXiv preprint arXiv:2409.06937*, 2024.

A. Proofs of analytical results

This appendix gives the formal statements and proofs of the analytical results summarized in Section 6 of the main text.

Symmetry and Nash equilibrium at $p = 1/2$

Consider the involution $\mathcal{R}$ acting on density matrices by $\mathcal{R}\gamma = \sigma_x \gamma \sigma_x$. On Bloch coordinates it acts as

$$\mathcal{R} : (x, y, z) \mapsto (x, -y, -z). \tag{10}$$

Lemma 1. *Under $\mathcal{R}$, the two-type QMFG of Section 2 is invariant under the combined transformation: (type $A \leftrightarrow$ type B, $\gamma \mapsto \mathcal{R}\gamma$, $\bar{z} \mapsto -\bar{z}$).*

Proof. Write $(x', y', z') = (x, -y, -z)$ for the $\mathcal{R}$-image of the Bloch coordinates and $\bar{z}' = -\bar{z}$ for the sign-flipped mean field. We verify invariance of each component of the QMFG specification: target states, costs, Hamiltonian, and Belavkin SDE.

Targets and running cost. The targets satisfy $\mathcal{R}\rho_g^A = \mathcal{R}\,\frac{1}{2}(I - \sigma_z) = \frac{1}{2}(I - \mathcal{R}\sigma_z) = \frac{1}{2}(I + \sigma_z) = \rho_g^B$, since $\sigma_x \sigma_z \sigma_x = -\sigma_z$. Hence $\ell^A(\gamma) = 1 - \mathrm{tr}(\rho_g^A \gamma) = (1 + z)/2$ and $\ell^A(\mathcal{R}\gamma) = (1 + z')/2 = (1 - z)/2 = 1 - \mathrm{tr}(\rho_g^B \gamma) = \ell^B(\gamma)$. The control penalty $\frac{\lambda_u}{2}\alpha^2$ is invariant since $\alpha \in \mathbb{R}$ is unchanged.

Hamiltonian. The control Hamiltonian is $H_{\bar{z}} = \alpha\sigma_x + J\bar{z}\,\sigma_z$. Conjugation by σ_x gives $\mathcal{R}\sigma_x = \sigma_x\sigma_x\sigma_x = \sigma_x$ and $\mathcal{R}\sigma_z = \sigma_x\sigma_z\sigma_x = -\sigma_z$. Therefore

$$\mathcal{R}H_{\bar{z}}\mathcal{R}^\dagger = \alpha\sigma_x - J\bar{z}\,\sigma_z = H_{-\bar{z}} = H_{\bar{z}'},$$

so the Hamiltonian is invariant under the combined transformation $(\gamma, \bar{z}) \mapsto (\mathcal{R}\gamma, -\bar{z})$.

Belavkin SDE. We verify directly that (1) transforms covariantly. Substituting $x = x'$, $y = -y'$, $z = -z'$, $\bar{z} = -\bar{z}'$ into (1) and using $dx' = dx$, $dy' = -dy$, $dz' = -dz$, we obtain:

$$\begin{aligned} dx' &= (-2\bar{z}'y' - 2x')\,dt + 2x'z'\,dW, \\ dy' &= (2\bar{z}'x' - 2y' - 2\alpha z')\,dt + 2y'z'\,dW, \\ dz' &= 2\alpha y'\,dt - 2(1 - z'^2)\,dW. \end{aligned}$$

The drifts coincide with the canonical form of (1) expressed in primed coordinates against mean field $\bar{z}'$, while every diffusion coefficient has the opposite sign. Setting $W'_t := -W_t$, the process W' is again a standard Wiener process and $dW = -dW'$. Substituting $dW = -dW'$ into the displays above produces exactly (1) in primed coordinates against $\bar{z}'$. Since the law of the SDE depends on W only through its distribution, and $W \overset{d}{=} W'$, the joint law of the trajectory $(\gamma'_t, \bar{z}'_t)$ under the transformed dynamics equals the joint law of $(\gamma_t, \bar{z}_t)$ under the original dynamics.

Cost exchange. Combining the running-cost identity $\ell^A(\mathcal{R}\gamma) = \ell^B(\gamma)$, terminal-cost identity $\ell^A(\mathcal{R}\gamma_T) = \ell^B(\gamma_T)$, control-penalty invariance, and the equality in law of trajectories, we obtain

$$\mathcal{J}^A[\alpha; \bar{z}](\mathcal{R}\gamma_0) = \mathcal{J}^B[\alpha; -\bar{z}](\gamma_0). \qquad \square$$

Proposition 2. *At prior $p = 1/2$, the profile in which both types play their single-type mean-field-$\bar{z} \equiv 0$ optimal feedback is a Nash equilibrium of the two-type QMFG, with self-consistent combined mean field $\bar{z}_t^\star \equiv 0$.*

Proof. Let $\alpha^{\theta,\star}$ be type θ's best response against $\bar{z} \equiv 0$. By Lemma 1, $\alpha^{B,\star}(\gamma, t) = \alpha^{A,\star}(\mathcal{R}\gamma, t)$. Both types start from the maximally mixed state $\gamma_0 = I/2$, which is $\mathcal{R}$-invariant ($\mathcal{R}\gamma_0 = \gamma_0$). The trajectory under type-B dynamics and optimal feedback is therefore $\mathcal{R}$-conjugate in distribution to the type-A trajectory driven by independent noise with the same law; hence $\mathbb{E}[z_t^B] = -\mathbb{E}[z_t^A]$ for all t, giving

$$\bar{z}_t^\star = \tfrac{1}{2}\mathbb{E}[z_t^A] + \tfrac{1}{2}\mathbb{E}[z_t^B] = 0.$$

The profile is self-consistent: each type's response is optimal against $\bar{z} \equiv 0$, which is indeed the combined mean field produced. $\square$

Figure 3 confirms this proposition numerically, with the residual $\bar{z}^\star \lesssim 0.03$ attributable to the finite rollout batch ($K = 2048$) and finite FP tolerance.

Mean-field-control lower bound and symmetry

Let $J_{\text{MFC}}(p)$ denote the cost achieved by a central planner who chooses (α^A, α^B) jointly to minimize $p\,\mathcal{J}^A + (1-p)\,\mathcal{J}^B$ under the same self-consistent mean-field constraint. The gap between the Nash population cost and this planner value is the price of anarchy, introduced to MFG by Carmona, Graves, and Tan (Carmona et al., 2019).

Proposition 3 (Planner lower bound)**.** *For the two-type QMFG,*

$$J_{MFC}(p) \leq p\,\mathcal{J}^{A,\star}(p) + (1-p)\,\mathcal{J}^{B,\star}(p), \tag{11}$$

where the right-hand side is the population-weighted Nash cost.

Proof. The central planner minimizes over joint controls. The Nash feedback pair is one feasible joint-control pair for that optimization problem, so the planner value cannot exceed the population-weighted Nash cost. □

Equality in (11) requires more than Nash self-consistency: the Nash feedback pair must also solve the planner's joint optimization problem. A first-order route to this condition is to view the planner objective as a functional of the shared trajectory $\bar{z}(\cdot)$ after each type has optimized against that trajectory. Symmetry at $p = 1/2$ implies stationarity of this reduced planner objective at $\bar{z} \equiv 0$, but stationarity alone does not imply global optimality without an additional convexity or second-order argument.

Corollary 2 (Symmetric planner stationarity)**.** *At $p = 1/2$, the symmetric Nash mean field $\bar{z}^\star \equiv 0$ is a stationary point of the population-weighted planner objective. Consequently, zero price of anarchy at $p = 1/2$ follows if this stationary point is the global planner minimizer; our numerical MFC comparison supports this condition in the benchmark.*

Proof. Let $\mathcal{W}(\bar{z})$ denote the weighted optimized cost after both types best respond to the imposed shared mean-field trajectory $\bar{z}(\cdot)$. Consider a variation $\bar{z} \mapsto \epsilon\eta(t)$ for arbitrary $\eta \in L^2([0,T])$. By Lemma 1, applying the involution $\mathcal{R}$ exchanges the two types and maps $\bar{z}$ to $-\bar{z}$. At $p = 1/2$, the population weights are equal, hence

$$\mathcal{W}(\epsilon\eta) = \mathcal{W}(-\epsilon\eta).$$

Thus $\mathcal{W}$ is even in ϵ, and $d\mathcal{W}/d\epsilon|_{\epsilon=0} = 0$ for every perturbation direction η. This proves the stationarity statement. □

For $p \neq 1/2$, the equal-weight cancellation no longer applies. The planner value is nevertheless symmetric under swapping the labels of the two types.

Proposition 4 (No linear planner correction)**.** *Assume the planner value is smooth in p near $p = 1/2$. Then*

$$J_{MFC}(p) = J_{MFC}(1/2) + \beta(p - 1/2)^2 + O((p - 1/2)^4), \tag{12}$$

for some second-order coefficient β.

Proof. By Lemma 1, exchanging type labels and applying $\mathcal{R}$ maps the planner problem with prior p to the planner problem with prior $1-p$ without changing the optimal value. Hence $J_{\text{MFC}}(p) = J_{\text{MFC}}(1-p)$. Smoothness around $p = 1/2$ then forces the Taylor expansion to contain only even powers of $p - 1/2$, which removes the linear term. □

By Proposition 4, the planner value cannot have a linear departure from its symmetric-prior value. The observed flatness of the Nash population-weighted cost across $p \in \{0.25, 0.5, 0.75\}$ is stronger than this result: it suggests either a very small quadratic coefficient or an additional cancellation not captured by the symmetry proof. Whether the exact coefficient is zero remains open; see Open Question (i).