DECENTRALIZED PREFERENCE-BASED REINFORCEMENT LEARNING UNDER BYZANTINE ATTACKS: STABILITY ANALYSIS AND RESILIENT AGGREGATION DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Preferences can specify cooperative goals that are hard to encode as numerical rewards. We study fully decentralized preference-based control, where each agent learns a private reward model from local comparisons and exchanges critic predictions to estimate the team-average reward signal. Byzantine messages can bias policy updates and hence the trajectories compared next; on unbounded state spaces, learning errors and state growth can reinforce each other. We propose PARC, an actor–critic method combining stability-constrained policy mixing with topology-preserving sender-screened clipping (TPSC). At each receiver, TPSC first uses all received predictions to bound every coordinate, then replaces suspicious senders' predictions by its own or shrinks them toward it. Because every sender keeps its slot and weight, consensus holds under the original graph conditions for any adapted screen, accurate or not, at any fixed positive retention level. For linear reward models and critics, under stability and excitation conditions, we prove almost-sure utility identification and bounded critic consensus and tracking, with tracking error controlled by aggregation bias. An average stationarity bound for the team objective separates aggregation, approximation, occupancy and reward-misspecification errors. On ten linear–quadratic plants, screening improves attacked gains when corrupted predictions fall outside the honest range, and reduced retention when they fall inside it. Under distance-constrained corruption, PARC outperforms Remove-then-Clip (RTC) under attack in LQ and has smaller clean-to-attack losses than Mean and RTC in VMAS flocking.

1 INTRODUCTION

1.1 MOTIVATION

Human preferences can specify goals that are hard to encode as numerical rewards, from robot control to language-model alignment (Christiano et al., 2017; Ouyang et al., 2022). In cooperative systems, supervisors may have different priorities, such as formation accuracy and control effort. Preference-based multi-agent reinforcement learning (MARL) uses their judgments to learn cooperative policies (Mu et al., 2026). When preference records remain local, an agent's own comparisons cannot generally determine the team-average reward signal. Agents use peer-to-peer critic communication to estimate it without a trusted coordinator. Byzantine messages can bias these estimates and misdirect policy updates (Wu et al., 2021). Updated policies generate the segments judged next. Private utilities are therefore learned from comparisons whose distribution evolves with the policies and potentially corrupted communication. On unbounded state spaces, state growth can amplify these errors and destabilize both training and the controlled process (Jiang et al., 2026). Resilient communication and trajectory stability must therefore be addressed together.

Honest critic proposals differ with private utilities and reward-estimation errors. We therefore (1) design resilient communication and stable execution, (2) trace reward-learning and aggregation errors through critic tracking to policy stationarity, and (3) evaluate aggregation components across attack geometries.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.2 Related Work

Preference-based reinforcement learning fits reward models to policy-generated comparisons (Christiano et al., 2017; Lee et al., 2021). MAGPIE combines agent-specific reward predictions during centralized training (Mu et al., 2026); here, records and reward models remain local, and critic messages influence the policies generating subsequent comparisons.

Consensus actor–critic methods estimate team returns from heterogeneous local numerical rewards (Zhang et al., 2018). Resilient methods address Byzantine communication (Wu et al., 2021), including scalar aggregation along feature directions (Ye et al., 2024); Jiang et al. (2026) combine coordinatewise W-MSR with synchronized safe-policy mixing on unbounded state spaces. With heterogeneous rewards and unknown Byzantine identities, exact honest-average evaluation cannot generally be guaranteed (Hairi et al., 2024), making target accuracy distinct from critic agreement.

Robust aggregation includes selection, coordinatewise rules, and geometric median-of-means (Blanchard et al., 2017; Yin et al., 2018; Chen et al., 2017), alongside weighted mean-subsequence-reduced (W-MSR) consensus on robust graphs (LeBlanc et al., 2013; Sundaram & Gharesifard, 2019). ClippedGossip applies receiver-centered norm clipping with base weights (He et al., 2022). Remove-then-Clip (RTC) computes a clipping radius after distance-based removal and replaces removed vectors by self (Yang & Ghaderi, 2024). TPSC instead bounds each coordinate from all received predictions before judging any sender and keeps every slot and weight, so its consensus guarantee holds for any adapted screen, accurate or not. Appendix A compares these structures.

1.3 Our Contributions

Setting. We formulate a cooperative Markov decision process with private labels on shared trajectory-segment pairs. A common comparison scale and fixed normalization make utilities comparable in the team average. Critic messages influence the policies and hence the segments evaluated by trusted local feedback (Figure 1).

Algorithm. We design Preference-based Actor–Critic with Resilient Consensus (PARC). Its aggregation rule, topology-preserving sender-screened clipping (TPSC), answers two questions separately: how far a received prediction may move the receiver, and which senders warrant intervention. Ranges computed before any sender is judged answer the first; the screen only decides whether a suspicious prediction is replaced by the receiver's own or shrunk toward it. Because no sender loses its slot or weight, the consensus certificate holds under the original graph conditions at fixed positive retention, without an accuracy assumption on the screen. PARC adopts the synchronized stability-constrained mixing to control trajectories for preference and critic learning.

Guarantees. For linear reward models and critics, we establish: (i) almost-sure private utility identification along the attacked learning trajectory (Theorem 1); (ii) critic boundedness without projection despite state-dependent update magnitudes, and consensus under admissible Byzantine messages without a small-heterogeneity condition (Theorem 2); and (iii) a step-size-weighted average stationarity bound for the discounted team objective at the limiting Byzantine physical policy (Theorem 3). Net aggregation bias controls critic tracking and enters the stationarity tolerance alongside approximation and occupancy errors. Reward misspecification adds estimation and target-shift terms.

Evidence. In linear–quadratic (LQ) control, trusted labels improve on shuffled labels, and under boundary corruption PARC keeps learned controllers stable where unscreened averaging (Mean) does not. Across ten LQ plants, screening helps against predictions outside the honest range and reduced retention against predictions inside it. Under distance-constrained corruption, PARC achieves higher attacked gain than RTC in LQ and smaller clean-to-attack losses than Mean and RTC in VMAS flocking.

2 Problem Formulation

2.1 System Model and Private Preferences

For the system in Figure 1, consider a multi-agent Markov decision process $\mathcal{M} = \langle \mathcal{G}, \mathcal{S}, \mathcal{A}, K, r^\star, \gamma \rangle$. The directed graph $\mathcal{G} = \langle \mathcal{N}, \mathcal{E} \rangle$ specifies communication. Its agents form fixed

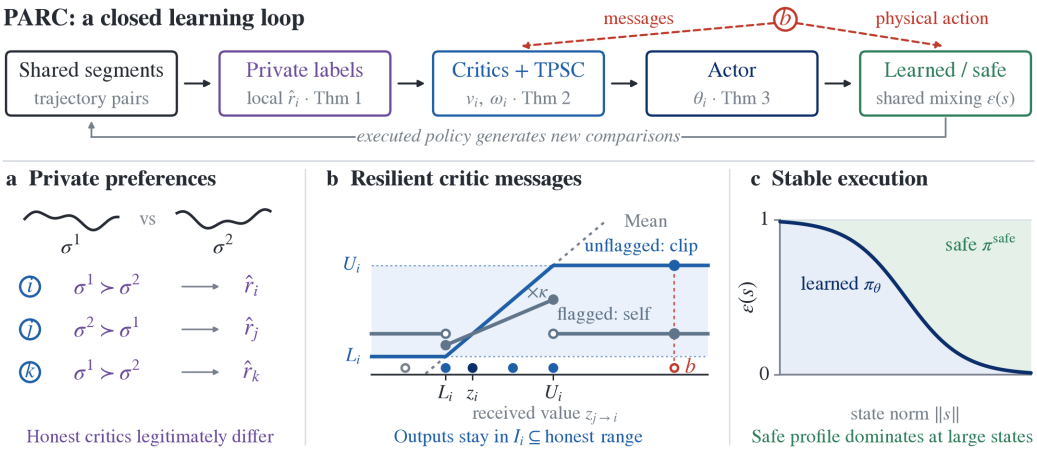


Figure 1: **PARC closes the loop from private labels to stable execution.** Top: one step at cooperative agent i; Byzantine agent b may corrupt critic messages and act physically. (a) Private labels on shared segments make honest critics differ. (b) TPSC clips unflagged values to I_i, replaces flagged outliers by self and shrinks flagged in-range values toward self by the factor κ_{keep}; every sender keeps its weight. (c) The learned policy runs with probability $\varepsilon(s)$.

cooperative and Byzantine sets $\mathcal{N}^+$ and $\mathcal{N}^-$, with $n = |\mathcal{N}^+|$. At time t, the state is $s_t \in \mathcal{S}$ and agent i takes $a_{i,t} \in \mathcal{A}_i$, with $\mathcal{A} = \prod_i \mathcal{A}_i$. Each cooperative agent has a stochastic policy $\pi_i(\cdot \mid s; \theta_i)$ on a compact convex parameter set Θ_i. The next state follows $K(\cdot \mid s_t, a_t)$, and $\gamma \in (0,1)$ is the discount. The state space may be unbounded: policies affect the states, features and update magnitudes encountered during training. Write $Z_t = (s_t, B_t, a_t, s_{t+1})$, where the shared switch B_t selects learning policies (1) or a safe profile (0; Section 3.1). Cooperative agents receive preferences rather than numeric rewards. Agent i's latent utility uses a common feature map χ,

$$r_i^\star(Z) = \chi(Z)^\top \eta_i^\star, \qquad \eta_i^\star \in \mathcal{K}_i \subseteq \mathsf{E}, \tag{1}$$

with $\mathcal{K}_i$ compact and convex and the gauge subspace E removing utility directions that leave every comparison unchanged. Features may be nonlinear and need not depend on B_t.

Comparisons use fixed blocks of (Z_t): block ℓ contains two equal-length segments separated by a fixed gap, and D_ℓ is their discounted feature difference projected onto E (Appendix B.2). All cooperative agents observe the same pair (Figure 1a); agent i's supervisor returns a private label $Y_{i,\ell} \in \{0,1\}$, with 1 favoring the first segment, under the Bradley–Terry model

$$\mathbb{P}(Y_{i,\ell} = 1 \mid \mathcal{H}_{i,\ell}) = \sigma(\beta_{\text{p}} D_\ell^\top \eta_i^\star), \qquad \sigma(u) = (1 + e^{-u})^{-1}, \tag{2}$$

where $\mathcal{H}_{i,\ell}$ contains the observed block and history preceding the fresh label. The common known inverse temperature $\beta_{\text{p}} > 0$ and gauge E fix utility scale and additive ambiguity, making utilities comparable in (3). Appendix F treats preference misspecification.

2.2 Cooperative Objective, Information Structure, and Byzantine Agents

The team utility and cooperative objective are

$$\bar{r}^\star(Z) = \frac{1}{n} \sum_{i \in \mathcal{N}^+} r_i^\star(Z), \qquad J_{\text{pref}}(\theta; \nu, \varpi) = \mathbb{E}_{\nu,\theta,\varpi}\Big[\sum_{t=0}^{\infty} \gamma^t \bar{r}^\star(Z_t)\Big], \tag{3}$$

where $\pi^+(a^+ \mid s; \theta) = \prod_{i \in \mathcal{N}^+} \pi_i(a_i \mid s; \theta_i)$ is the cooperative learning profile. The expectation uses its executed learning/safe mixture, a fixed initial law ν, and a fixed admissible state-Markov Byzantine physical policy $\pi_\varpi^-(\cdot \mid s)$.

The state, trajectories and features are common; labels, comparison records and reward models remain private. One agent's comparisons generally do not identify the other utilities, so PARC estimates the team signal through critic communication.

Cooperative agents know a bound $H \geq |\mathcal{N}^-|$ but not the Byzantine identities. The communication graph is $(2H+1)$-robust, with each cooperative receiver's in-degree at least $2H+1$ (LeBlanc et al., 2013). Byzantine agents may send arbitrary finite critic messages, adapted to the available history and differing across receivers and prediction coordinates. Their physical-policy descriptor ϖ_t is predictable, slowly varying and converges to $\varpi_\star$ (Assumption 1); Theorem 3 evaluates (3) at this limit. Each cooperative agent trusts its own updates and labels. The agents seek high team utility and a stable state process without a trusted coordinator.

3 PARC: Preference-Based Actor–Critic with Resilient Consensus

Private rewards, communicating critics and actors learn on successively slower time scales. Policy mixing controls their shared trajectory; TPSC limits corrupted critic contributions. Appendix B.1 specifies the update order.

3.1 Policy Execution and Local Learning

Following Jiang et al. (2026), PARC mixes learning policies with a known cooperative safe profile $\pi^{\mathrm{safe}} = (\pi_i^{\mathrm{safe}})_{i\in\mathcal{N}^+}$ whose kernels are geometrically ergodic uniformly over admissible Byzantine physical policies. At s_t, all cooperative agents threshold the same fresh uniform draw at $\varepsilon(s_t)$, giving $B_t \sim \mathrm{Bernoulli}(\varepsilon(s_t))$, and play

$$a_{i,t} \sim B_t\, \pi_i(\cdot \mid s_t; \theta_{i,t}) + (1 - B_t)\, \pi_i^{\mathrm{safe}}(\cdot \mid s_t), \qquad 0 \leq \varepsilon(s) \leq \min\left\{1, \frac{c}{1+\|s\|}\right\}, \tag{4}$$

with $c > 0$. The safe profile dominates at large states; the shared draw makes the joint kernel a convex mixture. With uniformly bounded one-step Lyapunov growth of learning kernels and sufficiently small c, the mixture inherits the safe profile's common drift (Appendix D.1). The LQ experiments use a stabilizing LQR reference.

After receiving the block-ℓ label, agent i takes a projected logistic-loss step:

$$\eta_{i,\ell+1} = \Pi_{\mathcal{K}_i}\left[\eta_{i,\ell} - a_{\eta,\ell}\beta_{\mathrm{p}}\big(\sigma(\beta_{\mathrm{p}} D_\ell^\top \eta_{i,\ell}) - Y_{i,\ell}\big) D_\ell\right]. \tag{5}$$

Here $a_{\eta,\ell}$ is the preference step size and $\Pi_{\mathcal{K}}$ the Euclidean projection onto a closed convex set $\mathcal{K}$. With $\ell(t)$ the latest preference-update index before transition t, the reward is $\hat{r}_{i,t} = \chi(Z_t)^\top \eta_{i,\ell(t)}$.

This reward drives the state-value critic $\phi(s)^\top v_i$ and reward critic $\psi(Z)^\top \omega_i$, which use consensus to estimate team value and team-average reward. We use linear critics with common features $\phi_t = \phi(s_t)$, $\phi_{t+1} = \phi(s_{t+1})$ and $\psi_t = \psi(Z_t)$; Section 5 also evaluates neural critics. With common $\lambda \in [0,1)$, $\mathbf{e}_0 = 0$ and step sizes $\alpha_{x,t}$, $x \in \{v, \omega\}$, TD(λ) and reward regression give

$$\begin{aligned}
\mathbf{e}_{t+1} &= \gamma\lambda \mathbf{e}_t + \phi_t, \\
\tilde{v}_{i,t+1} &= v_{i,t} + \alpha_{v,t}\big(\hat{r}_{i,t}\mathbf{e}_{t+1} - G^v_{t+1} v_{i,t}\big), \quad G^v_{t+1} = \mathbf{e}_{t+1}(\phi_t - \gamma\phi_{t+1})^\top, \\
\tilde{\omega}_{i,t+1} &= \omega_{i,t} + \alpha_{\omega,t}\big(\hat{r}_{i,t}\psi_t - G^\omega_{t+1}\omega_{i,t}\big), \quad G^\omega_{t+1} = \psi_t \psi_t^\top.
\end{aligned} \tag{6}$$

Shared observations and initialization give a common trace, carried across comparison blocks, and common sample matrices. Apart from current critic iterates, the only agent-specific input is $\hat{r}_{i,t}$, so honest proposals can differ. TPSC aggregates them into the next critics.

The actor uses the executed transition and stored time-t critics:

$$\begin{aligned}
\delta^{\mathrm{ac}}_{i,t} &= \psi_t^\top \omega_{i,t} + (\gamma\phi_{t+1} - \phi_t)^\top v_{i,t}, \quad h_{i,t} = B_t\, \delta^{\mathrm{ac}}_{i,t}\, \nabla_{\theta_i} \log \pi_i(a_{i,t} \mid s_t; \theta_{i,t}), \\
\theta_{i,t+1} &= \Pi_{\Theta_i}[\theta_{i,t} + a_{\theta,t}\, \mathsf{C}_{R_t}(h_{i,t})].
\end{aligned} \tag{7}$$

The reward term uses the consensus reward critic to estimate team-average reward rather than the private utility; B_t selects learning-policy samples. The step size is $a_{\theta,t}$. The norm safeguard

$\mathsf{C}_R(h) = h\min\{1, R/\|h\|\}$, $\mathsf{C}_R(0) = 0$, uses deterministic $R_t \uparrow \infty$ satisfying Assumption 2; under Theorem 3's conditions, it is almost surely eventually inactive.

3.2 Resilient Aggregation

Honest proposals can differ, and corrupted values may lie inside or outside their range. Figure 1b shows TPSC for one receiver and one coordinate: it builds an interval from all received values, flags suspicious senders from their complete payloads, and then clips, replaces or shrinks each value while keeping every sender's base weight.

Senders exchange predictions at common probes. For critic $x \in \{v, \omega\}$ with d_x parameters, rows of a fixed invertible $F_x \in \mathbb{R}^{d_x \times d_x}$ are probe features. Sender j broadcasts the stacked post-innovation payload $\mathbf{y}_{j,t+1} = \mathrm{col}(F_v\tilde{v}_{j,t+1}, F_\omega\tilde{\omega}_{j,t+1})$, whose coordinates are value or reward predictions. Receiver i processes the sender set $\mathcal{J}_i = \{i\} \cup \mathrm{In}(i)$. Byzantine payloads $\mathbf{y}_{j\to i}$ may differ across receivers; round indices are suppressed through (8).

For one coordinate, let $z_{j\to i}$ be the received value and z_i the self proposal. As in mean-subsequence-reduced consensus (LeBlanc et al., 2013), exclude the H largest values strictly above z_i (all if fewer exist), and symmetrically the H smallest below it. The remaining senders $\mathcal{R}_i \subseteq \mathcal{J}_i$ include i and define $I_i = [L_i, U_i]$, with $L_i = \min_{j\in\mathcal{R}_i} z_{j\to i}$ and $U_i = \max_{j\in\mathcal{R}_i} z_{j\to i}$. Trimming only constructs this interval; aggregation keeps all sender slots. With at most H Byzantine senders, I_i lies within the honest coordinate range, including self.

An adapted screen selects $\mathcal{S}_i \subseteq \mathcal{J}_i \setminus \{i\}$ from complete raw payloads, with one flag shared across critics and probes. Appendix B.3 defines the distance and MAD screens. For fixed $\kappa_{\mathrm{keep}} \in (0, 1]$ and base weights $a_{ij} \geq a_\star > 0$, $\sum_{j\in\mathcal{J}_i} a_{ij} = 1$, TPSC applies

$$\hat{z}_i = \sum_{j\in\mathcal{J}_i} a_{ij}\, T_i(z_{j\to i}), \qquad T_i(z_{j\to i}) = \begin{cases} z_i, & j \in \mathcal{S}_i,\ z_{j\to i} \notin I_i, \\ z_i + \kappa_{\mathrm{keep}}(z_{j\to i} - z_i), & j \in \mathcal{S}_i,\ z_{j\to i} \in I_i, \\ \Pi_{I_i}(z_{j\to i}), & \text{otherwise.} \end{cases} \tag{8}$$

Figure 1b plots these three cases as maps of the received value; full retention ($\kappa_{\mathrm{keep}} = 1$) leaves flagged in-range values unchanged. Because no slot is dropped or reweighted, every processed value lies in I_i and each in-range honest slot keeps coefficient at least $\kappa_{\mathrm{keep}} a_\star$; Appendix E.1 turns this into an honest-row support certificate for every admissible screen, accurate or not. Stacking outputs gives $\hat{\mathbf{y}}_{i,t+1} = \mathrm{col}(\hat{\mathbf{y}}^v_{i,t+1}, \hat{\mathbf{y}}^\omega_{i,t+1})$; reconstruct each critic as $x_{i,t+1} = F_x^{-1}\hat{\mathbf{y}}^x_{i,t+1}$.

Retention may instead target a separate adapted set $\mathcal{T}_i$, such as one uniformly random peer (RandomRetention); the certificate is unchanged (Appendix B.3).

4 Resilience Guarantees

Critic agreement and target accuracy are distinct: agents can agree on a biased team estimate. For linear reward models and critics, Theorems 1–3 successively establish utility identification, critic tracking and policy stationarity.

4.1 Assumptions

Let ϑ index execution (4) at fixed policies (θ, ϖ), with kernel P_ϑ; along the run, $\vartheta_t = \iota(\theta_t, \varpi_t)$. Stationary expectations $\mathbb{E}_\vartheta$ and $\mathbb{E}^{\mathrm{blk}}_\vartheta$ include the common trace and comparison blocks, respectively. Assumptions 1–2 control sampling and clocks, 3–4 ensure identifiability, and 5–6 govern communication and actor dynamics. Appendix D gives the full conditions.

Assumption 1 (Execution regularity and physical-policy evolution)**.** The compact executed family has a common Lyapunov envelope $V \geq 1$, drift $P_\vartheta V \leq \rho_V V + b\mathbf{1}_C$ with $\rho_V < 1$, $b < \infty$, minorization on a common small set C, and uniform weighted geometric ergodicity and Lipschitz dependence on ϑ. For some $q > 4$, the update and trace fields satisfy Appendix D.1's weighted regularity and q-/$2q$-moment conditions, including its initial and compact-critic actor bounds. The Byzantine physical policy evolves slowly: almost surely, $\varpi_t \to \varpi_\star$ with $d_\Pi(\varpi_{t+1}, \varpi_t) \leq C_\Pi a_{\theta,t}$ in a policy metric d_Π.

The drift bounds $\mathbb{E}V(s_t)$ throughout learning, while allowing update fields to grow with the state. Appendix D.1 gives a sufficient construction from the safe policy.

Assumption 2 (Learning design and clock compatibility)**.** Preference, critic and actor updates operate on successively slower time scales, with gains satisfying the Robbins–Monro, separation and block-transport conditions in Appendices D.2 and D.3. The safeguard has $R_t \uparrow \infty$ and $\sum_t R_t^{-q} < \infty$. Equation (53) gives a feasible polynomial schedule.

Assumption 3 (Calibrated preferences and bounded excitation)**.** The trusted labels obey the calibrated Bradley–Terry law and gauge of Section 2.1, with $\eta_i^\star \in \mathcal{K}_i \subseteq \mathsf{E}$. For some $B_D < \infty$, $\lambda_D > 0$ and all ϑ, $\mathbb{E}_\vartheta^{\mathrm{blk}}[DD^\top \mathbf{1}_{\{\|D\| \le B_D\}}] \succeq \lambda_D I_{\mathsf{E}}$.

The bounded comparisons retain information in every identifiable utility direction, ensuring curvature of the preference loss.

Assumption 4 (Critic feature excitation)**.** For all ϑ, $\mathbb{E}_\vartheta \phi\phi^\top \succeq c_\phi I$ and $\mathbb{E}_\vartheta \psi\psi^\top \succeq c_\psi I$, with $c_\phi, c_\psi > 0$.

Mean critic matrices are coercive: $\delta_v = (1-\gamma)c_\phi/(1-\gamma\lambda)$, $\delta_\omega = c_\psi$ (Tsitsiklis & Van Roy, 1997).

Assumption 5 (Resilient communication)**.** The static graph and Byzantine budget satisfy Section 2.2. Fixed base rows sum to one, with positive entries at least $a_\star$. Section 3.2's operator uses fixed invertible probes F_v, F_ω, fixed retention $\kappa_{\mathrm{keep}} \in (0, 1]$, and adapted screening and retention sets, each shared across critics and coordinates.

Assumption 6 (Actor dynamics and objective regularity)**.** (i) Policies are positive and C^2 on compact convex projection sets with Appendix D.6's reflection bounds; boxes suffice. The stationary actor mean is uniformly L_F-Lipschitz in the critic stack. (ii) For the discounted-objective bridge, the kernel and conditional team reward at fixed physical policy satisfy that appendix's weighted-C^1, integrability and score conditions.

Equation (124) gives a sufficient L_F from score and feature moments.

4.2 Main Results

Theorem 1 (Private utility identification)**.** *Under Assumptions 1–3, for every $i \in \mathcal{N}^+$, $\eta_{i,\ell} \to \eta_i^\star$ as $\ell \to \infty$, a.s.*

Calibration and uniform excitation identify the same utilities as policy updates change the comparison distribution. The estimation residual vanishes on critic-time windows (Lemma 12), connecting preference learning to critic and actor analysis.

Critic tracking. Assumption 4 makes the honest-team TD and reward-regression targets well defined. With $h_i^v = r_i^\star \mathbf{e}_{t+1}$ and $h_i^\omega = r_i^\star \psi_t$, they are

$$x_x^\star(\vartheta) = \bar{G}_x(\vartheta)^{-1}\bar{h}_x(\vartheta), \qquad \bar{G}_x(\vartheta) = \mathbb{E}_\vartheta G^x, \qquad \bar{h}_x(\vartheta) = \tfrac{1}{n}\textstyle\sum_{i\in\mathcal{N}^+} \mathbb{E}_\vartheta h_i^x. \tag{9}$$

Row-stochastic mixing and message processing need not preserve the honest proposal average. Their net shift, scaled by the critic step size, is

$$\Delta_{x,t} = (n\alpha_{x,t})^{-1} \textstyle\sum_{i\in\mathcal{N}^+} \big(x_{i,t+1} - \widetilde{x}_{i,t+1}\big). \tag{10}$$

Let $\mathfrak{D}_x$ bound the long-run net shift between these averages per unit of critic time (summed step sizes). Opposing shifts can cancel before the window norm is taken. Appendix C.2 specifies the bound; Lemma 11 proves finiteness.

Theorem 2 (Critic consensus and tracking)**.** *Under Assumptions 1–5, the honest critics are bounded and reach consensus a.s. For each $i \in \mathcal{N}^+$ and $x \in \{v, \omega\}$,*

$$\limsup_{t\to\infty} \|x_{i,t} - x_x^\star(\vartheta_t)\| \le \mathfrak{D}_x/\delta_x \quad \text{a.s.} \tag{11}$$

Under Assumptions 1–5, finite-block graph contraction and a positive floor on honest in-neighbor coefficients, proportional to κ_{keep}, give agreement for every admissible screen without a small-heterogeneity condition. A disagreement filter controls defect windows before coercivity establishes

critic boundedness without projection (Appendix C.2). Coercivity converts the honest-average net shift into $\mathfrak{D}_x/\delta_x$, and consensus transfers this accuracy bound to each agent. Reduced retention cuts both ways: it lowers the coefficient floor, weakening the worst-case constants (Lemma 14), and it shrinks selected in-range contributions, Byzantine or honest (Proposition 2; Section 5.2).

Under Lemma 13's additional field conditions, one may take $\mathfrak{D}_x = \mathfrak{D}_x^{\mathrm{het}} = C_x \mu_{H,x}$. Here $\mu_{H,x}$ measures feature-weighted honest utility spread, including the value trace, and the graph/probe factor C_x in (102) depends on retention; the envelope covers all admissible messages and screens.

Policy stationarity. Fix a θ-independent initial law ν with the required moments and the final physical policy $\varpi_\star$. Let $\mathcal{J}(\theta) = (1-\gamma) J_{\mathrm{pref}}(\theta; \nu, \varpi_\star)$ and $g(\theta)$ be the stationary actor mean at targets (9). Critic tracking contributes $\epsilon_{\mathrm{ac}} = L_F \sqrt{n}\, \|(\mathfrak{D}_v/\delta_v, \mathfrak{D}_\omega/\delta_\omega)\|$, where $\sqrt{n}$ comes from stacking honest critics. Approximation and sampling at those targets leave a gradient discrepancy bounded by b_{obj}:

$$\delta = \epsilon_{\mathrm{ac}} + b_{\mathrm{obj}}, \qquad b_{\mathrm{obj}} = \sup_{\theta \in \Theta} \big\{ S_\theta [E_r(\theta) + \gamma C_\lambda E_V(\theta)] + b_{\mathrm{occ}}(\theta) \big\}. \tag{12}$$

Here E_r, E_V are the best stationary L^2 reward/value errors in the spans of ψ, ϕ; S_θ is the stationary L^2 norm of the score $B \nabla_\theta \log \pi^+$, and C_λ is the TD(λ) approximation factor. The term b_{occ} measures stationary-versus-discounted occupancy mismatch (Thomas, 2014; Nota & Thomas, 2020). Lemma 19 defines these quantities and proves $\|g - \nabla \mathcal{J}\| \le b_{\mathrm{obj}}$.

For a projection step $h > 0$, define $G_h(\theta) = h^{-1}(\Pi_\Theta[\theta + h \nabla \mathcal{J}(\theta)] - \theta)$, which vanishes exactly at constrained stationary points.

Theorem 3 (Average stationarity for the team objective)**.** *Under Assumptions 1–6, for every fixed $h > 0$,*

$$\limsup_{N \to \infty} \frac{\sum_{t<N} a_{\theta,t} \|G_h(\theta_t)\|^2}{\sum_{t<N} a_{\theta,t}} \le 4\delta^2 \quad \textit{a.s.} \tag{13}$$

The expanding safeguard is eventually inactive a.s.

In particular, $\liminf_t \|G_h(\theta_t)\| \le 2\delta$. For $\delta > 0$, the cumulative fraction of actor step-size weight on iterates with $\|G_h\| \le 4\delta$ has limit inferior at least $3/4$. Under Lemma 13's conditions, identical honest utilities permit $\mathfrak{D}_x = 0$ for all admissible message attacks. If also $E_r(\theta) = E_V(\theta) = 0$ throughout Θ, one may take $\delta = \sup_{\theta \in \Theta} b_{\mathrm{occ}}(\theta)$. Appendix F adds $\epsilon_{\mathrm{pref}}^{\mathrm{ac}}$ to δ under its misspecification conditions, relative to the actual run's full latent objective.

5 Experiments

We ask whether private preferences support control (Q1), how aggregation components counter attacks (Q2), and whether protection extends to nonlinear control (Q3). Networked LQ (ten agents, unbounded 20-dimensional state) uses Lyapunov costs and spectral stability tests for Q1–Q2; VMAS flocking (six actors) addresses Q3. PARC uses full retention ($\kappa_{\mathrm{keep}} = 1$); $\mathrm{PARC}_{0.1}$ uses 0.1. Screens use distance or median absolute deviation (MAD) (Table 4). One adaptive Byzantine sender ($H = 1$) observes honest critic proposals and, with neural critics, actor scores, but not private rewards, labels or reward models. Boundary messages lie outside honest coordinate ranges; in-interval messages stay inside raw intervals. Distance-constrained messages stay closer to the receiver's prediction than its farthest honest peer. Appendix G gives protocols and statistics.

5.1 From Preference Fidelity to Control

On ten fixed-policy datasets, increasing private comparisons from 50 to 800 reduces the relative L^2 reward error about fourfold (1.225 to 0.312) (Appendix G.3).

For online control with linear function approximation (LFA), cost ratios divide the mean cost over the last five phases by the first-phase cost (lower is better). Shuffled labels leave cost nearly unchanged; trusted labels reduce the ratio by 0.0340 (clean) and 0.0320 (boundary) (Table 1, top). With matched centering and unit scale, a local numeric-reward reference yields clean/boundary ratios $0.9552/0.9568$ (trusted: $0.9664/0.9697$). Relative to shuffled labels, preferences recover about $75\%/71\%$ of the numeric improvement (Appendix G.3).

5.2 LQ Communication under Message Attacks

Main aggregation comparison. Five rules share neural learner settings, intervals and weights on ten LQ plants, two seeds each (Appendices G.1–G.2). PARC uses the distance screen; IntervalOnly disables screening. RandomScreen and RandomRetention randomize the respective targets. All five aggregation choices belong to the certified TPSC family (Appendix E.1). C1–C2 compare screened replacement and targeting; C3–C4 compare in-range retention and targeting (Table 1).

We report gain $g = (J_{0,p} - J)/J_{0,p}$ (higher is better), where J is the mean stationary honest-team cost over the final five of 20 phases and $J_{0,p}$ is the common initial-policy cost on plant p. Data collection uses the learning/safe mixture; evaluation uses learned policies without the safe fallback.

Screening counters out-of-range values. Interval clipping maps outliers to the nearest boundary; screened replacement substitutes the receiver's own prediction. Under boundary corruption, PARC exceeds IntervalOnly and RandomScreen in attacked gain by 0.0740 and 0.0608 on average, with both differences positive on all ten plants (C1–C2, Table 1). Its mean clean-gain difference from IntervalOnly is $+0.0010$. PARC preserves 94% of its clean gain under attack, versus 24% for IntervalOnly (ratios of means). Replays on the recorded messages show that the distance screen removes more Byzantine than honest exposure (weighted offset from the receiver's prediction) on 10/10 plants, whereas a random screen does the opposite in conditional mean (Appendix G.5).

Retention counters in-range values. In-range values pass clipping unchanged; reduced retention contracts flagged contributions toward self. Under in-interval corruption, $\mathrm{PARC}_{0.1}$ exceeds PARC in attacked gain by 0.0335 (10/10 plants) and RandomRetention by 0.0116 (9/10) (C3–C4); random targets thus already provide about two-thirds of the gain. In the replays, distance targets beat random ones on the same exposure margin (10/10 plants), and lowering retention to 0.1 reduces the mean realized net shift by about 9–12%. The margin itself is not positive network-wide, because receivers without a Byzantine in-neighbor must still shrink an honest peer (Appendix G.5).

Effect / condition	Contrast	Estimate	n	$+$
Preference information: cost-ratio reduction (LFA)				
Clean	Trusted vs. shuffled	+0.0340 [+0.0279, +0.0392]	10	10/10
Boundary	Trusted vs. shuffled	+0.0320 [+0.0232, +0.0406]	10	10/10
Attacked-gain difference (main neural LQ comparison, 20 phases)				
C1: Screened replacement	PARC – IntervalOnly	$+0.0740$	10	10/10
C2: Screen target	PARC – RandomScreen	$+0.0608$	10	10/10
C3: In-range retention	$\mathrm{PARC}_{0.1}$ – PARC	$+0.0335$	10	10/10
C4: Retention target	$\mathrm{PARC}_{0.1}$ – RandomRetention	$+0.0116$	10	9/10

Table 1: **Preferences inform control; targeting helps under attack.** Top: trusted-versus-shuffled cost-ratio reductions with pointwise 95% seed-bootstrap CIs. Bottom: mean paired attacked-gain differences in the main aggregation comparison (C1–C2: boundary; C3–C4: in-interval). Each plant averages two seeds; n counts seeds (top) or plants (bottom), and $+$ counts positive effects.

Attack geometry. Over 160 phases on a fixed plant with six paired seeds (Figure 2), all six Mean controllers become unstable under boundary corruption and all six PARC controllers remain finite; RTC's all-phase mean gain exceeds PARC's by 0.0183. Under distance-constrained corruption, PARC exceeds RTC by 0.3072. Nonzero offsets on actor-sensitive queries leave the injected message closer than the farthest honest peer, so it survives RTC's farthest-message removal. IntervalOnly accounts for most of the PARC–RTC loss reduction in this geometry (Appendix G.5).

Additional tests find finite PARC controllers where Mean is unstable and gains over LocalOnly (LFA, 10/10 seeds each), and smaller attack-induced cost-ratio increases than Function W-MSR (neural, 10/10 plants; Appendices G.4–G.5).

5.3 Protection Carries Over to Nonlinear Flocking

PARC, RTC and Mean share the learner, six paired seeds and 448 training phases in VMAS flocking under sparse distance-constrained corruption (Appendix G.7). PARC uses MAD; private reward

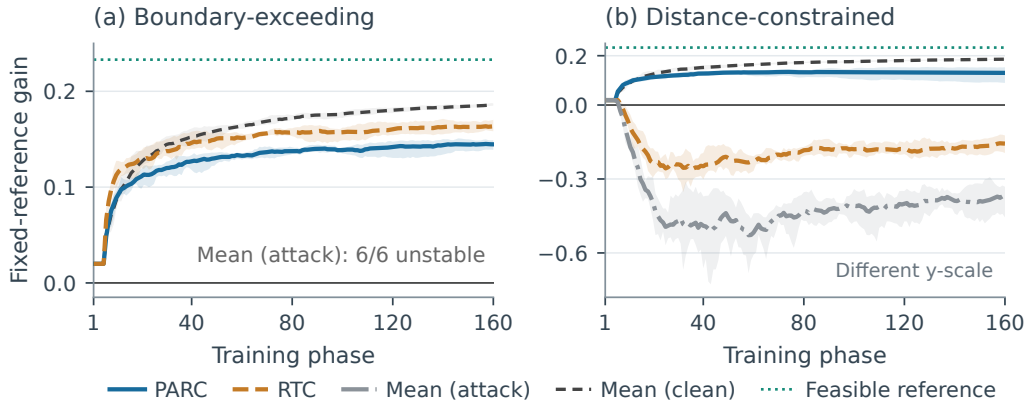


Figure 2: **PARC learns finite-cost controllers under both attacks.** Mean's controllers are unstable on 6/6 boundary runs (phases 7–8). Six paired seeds and 160 phases; PARC uses MAD (Table 4). Gains use the fixed-plant reference (Appendix G.1). Lines show means, bands span seeds, and dotted lines mark an analytic feasible controller. Both panels include clean Mean; other curves are attacked. Vertical scales differ.

models freeze after phase 32. Evaluation uses native 192-step team return. Attack-induced loss is each method's own clean minus attacked return.

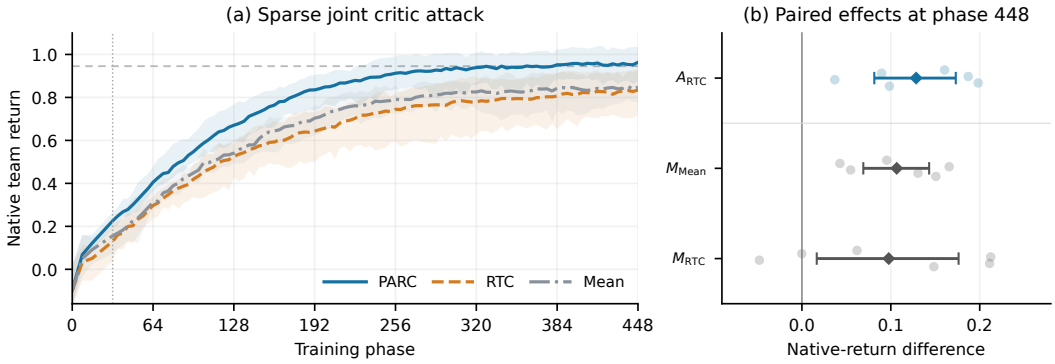


Figure 3: **Nonlinear flocking: PARC limits attack-induced loss.** Left: attacked mean returns with seed-range bands. The faint horizontal dashed line marks Mean's six-seed clean return at phase 448; the vertical line marks reward-model freezing at phase 32. Right: phase-448 attacked PARC-minus-RTC difference A and reductions M in the comparator's clean-to-attack loss, with six paired seed points and descriptive pointwise 95% bootstrap CIs.

At phase 448, PARC's attacked return exceeds RTC's by 0.1285 ($[0.0814, 0.1731]$), and its loss reductions are 0.1066 against Mean ($[0.0691, 0.1431]$; positive on 6/6 seeds) and 0.0976 against RTC ($[0.0167, 0.1762]$; positive on 4/6 seeds).

6 Conclusion

PARC couples resilient critic communication with stability-constrained mixing in decentralized preference-based control, where policies generate their own comparisons. TPSC keeps every sender's slot and weight, so consensus holds for any adapted screen. We prove utility identification, bounded critic tracking and an average stationarity bound separating aggregation, approximation, occupancy and misspecification errors. Screening and reduced retention counter out-of-range and in-range corruption; under distance-constrained attacks, PARC loses less than Mean and RTC in LQ and VMAS flocking, mostly via clipping in LQ.

AI USE STATEMENT

We used generative AI tools to assist with experimental design, manuscript writing and polishing, literature search, and figure creation. We did not use generative AI for problem formulation, modeling, or algorithm and method design. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f4b9ec30ad9f68f89b29639786cb62ef-Abstract.html.

The Viet Bui, Tien Anh Mai, and Thanh Hong Nguyen. O-MAPL: Offline multi-agent preference learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5811–5841. PMLR, 2025. URL https://proceedings.mlr.press/v267/bui25a.html.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6116–6135. PMLR, 2024. URL https://proceedings.mlr.press/v235/chakraborty24b.html.

Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):44:1–44:25, December 2017. doi: 10.1145/3154503. URL https://doi.org/10.1145/3154503.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Flint Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Cheston Tan, and Bryan Kian Hsiang Low. FedHQL: Federated heterogeneous Q-Learning. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, pp. 2810–2812. IFAAMAS, 2023. URL https://www.ifaamas.org/Proceedings/aamas2023/pdfs/p2810.pdf. Extended abstract.

Hairi, Minghong Fang, Zifan Zhang, Alvaro Velasquez, and Jia Liu. On the hardness of decentralized multi-agent policy evaluation under Byzantine attacks. In *Proceedings of the International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, pp. 285–292. IFIP, 2024. URL https://opendl.ifip-tc6.org/db/conf/wiopt2024/wiopt2024/1571066372.pdf.

Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Byzantine-robust decentralized learning via ClippedGossip, 2022. URL https://arxiv.org/abs/2202.01545. arXiv preprint arXiv:2202.01545.

Yifan Jiang, Jiasheng Pan, Mengtian Li, and Li Jin. Secure multi-agent reinforcement learning for service systems with affinity and Byzantine nodes: Stability analysis and protection design. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://gc.sjtu.edu.cn/off-the-press/2026-05-21/161822/. Accepted paper.

Philip Jordan, Florian Grötschla, Flint Xiaofeng Fan, and Roger Wattenhofer. Decentralized federated policy gradient with Byzantine fault-tolerance and provably fast convergence. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. IFAAMAS, 2024. doi: 10.5555/3635637.3662951. URL https://arxiv.org/abs/2401.03489.

Vijay R. Konda and John N. Tsitsiklis. On actor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003. doi: 10.1137/S0363012901385691. URL https://doi.org/10.1137/S0363012901385691.

Heath J. LeBlanc, Haotian Zhang, Xenofon Koutsoukos, and Shreyas Sundaram. Resilient asymptotic consensus in robust networks. *IEEE Journal on Selected Areas in Communications*, 31 (4):766–781, 2013. doi: 10.1109/JSAC.2013.130413. URL https://doi.org/10.1109/JSAC.2013.130413.

Kimin Lee, Laura M. Smith, and Pieter Abbeel. PEBBLE: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6152–6163. PMLR, 2021. URL https://proceedings.mlr.press/v139/lee21i.html.

Ni Mu, Yao Luan, Yiqin Yang, and Qing-Shan Jia. Multi-agent reinforcement learning via agent-specific preference. *IEEE Transactions on Automation Science and Engineering*, 2026. URL https://arxiv.org/abs/2608.08604. Accepted for publication; author version available as arXiv:2608.08604.

Chris Nota and Philip S. Thomas. Is the policy gradient a gradient? In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, pp. 939–947. International Foundation for Autonomous Agents and Multiagent Systems, 2020. URL https://www.ifaamas.org/Proceedings/aamas2020/pdfs/p939.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

Brahma S. Pavse, Matthew Zurek, Yudong Chen, Qiaomin Xie, and Josiah P. Hanna. Learning to stabilize online reinforcement learning in unbounded state spaces. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40015–40033. PMLR, 2024. URL https://proceedings.mlr.press/v235/pavse24a.html.

Chao Qu, Shie Mannor, Huan Xu, Yuan Qi, Le Song, and Junwu Xiong. Value propagation for decentralized networked deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8a0e1141fd37fa5b98d5bb769ba1a7cc-Abstract.html.

Devavrat Shah, Qiaomin Xie, and Zhi Xu. Stable reinforcement learning with unbounded state space. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 581. PMLR, 2020. URL https://proceedings.mlr.press/v120/shah20a.html.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7423902b5534e2b267438c85444a54b1-Abstract-Conference.html.

Shreyas Sundaram and Bahman Gharesifard. Distributed optimization under adversarial nodes. *IEEE Transactions on Automatic Control*, 64(3):1063–1076, 2019. doi: 10.1109/TAC.2018.2836919. URL https://doi.org/10.1109/TAC.2018.2836919.

Akihito Taya, Takayuki Nishio, Masahiro Morikura, and Koji Yamamoto. Decentralized and model-free federated learning: Consensus-based distillation in function space. *IEEE Transactions on Signal and Information Processing over Networks*, 8:799–814, 2022. doi: 10.1109/TSIPN.2022.3205549. URL https://arxiv.org/abs/2104.00352.

Philip Thomas. Bias in natural actor-critic algorithms. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 441–448. PMLR, 2014. URL https://proceedings.mlr.press/v32/thomas14.html.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. URL https://web.mit.edu/~jnt/www/Papers/J063-97-bvr-td.pdf.

Zhaoxian Wu, Han Shen, Tianyi Chen, and Qing Ling. Byzantine-resilient decentralized policy evaluation with linear function approximation. *IEEE Transactions on Signal Processing*, 69: 3839–3853, 2021. doi: 10.1109/TSP.2021.3090952. URL https://doi.org/10.1109/TSP.2021.3090952.

Caiyi Yang and Javad Ghaderi. Byzantine-robust decentralized learning via Remove-then-Clip aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21735–21743, 2024. doi: 10.1609/aaai.v38i19.30173. URL https://ojs.aaai.org/index.php/AAAI/article/view/30173.

Lintao Ye, Martin Figura, Yixuan Lin, Mainak Pal, Pranoy Das, Ji Liu, and Vijay Gupta. Resilient multiagent reinforcement learning with function approximation. *IEEE Transactions on Automatic Control*, 69(12):8497–8512, 2024. doi: 10.1109/TAC.2024.3409676. URL https://doi.org/10.1109/TAC.2024.3409676.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5650–5659. PMLR, 2018. URL https://proceedings.mlr.press/v80/yin18a.html.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Fully decentralized multi-agent reinforcement learning with networked agents. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5872–5881. PMLR, 2018. URL https://proceedings.mlr.press/v80/zhang18n.html.

Tianchen Zhu, Yue Qiu, Haoyi Zhou, and Jianxin Li. Decoding global preferences: Temporal and cooperative dependency modeling in multi-agent preference-based reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17202–17210, 2024. doi: 10.1609/aaai.v38i15.29666. URL https://ojs.aaai.org/index.php/AAAI/article/view/29666.

GUIDE TO THE APPENDICES

Algorithm 1 and its sampling conventions are in Appendix B; the full assumptions are in Appendix D. The main proofs use the following estimates.

Result	Proof	Supporting estimates
Theorem 1: identification	C.1	Block noise and curvature (Appendix E.3)
Theorem 2: critic tracking	C.2	Graph contraction (Appendix E.1); critic windows (Lemma 11)
Theorem 3: stationarity	C.3	Actor limits (Appendix E.7); objective bias (Lemma 19)

Shared Markov and stopping tools are in Appendix E.2, the utility-spread certificate in Appendix E.5, the fixed-input retention and screening decomposition in Appendix E.6, and preference misspecification in Appendix F. Appendix G collects the experimental protocols and results.

A EXTENDED RELATED WORK

Preference learning and heterogeneous feedback. Preference learning fits reward models to behavioral comparisons (Christiano et al., 2017; Lee et al., 2021); extensions model evaluator variation, optimize egalitarian objectives, and address global preferences and offline multi-agent learning (Siththaranjan et al., 2024; Chakraborty et al., 2024; Zhu et al., 2024; Bui et al., 2025). MAGPIE combines agent-specific reward predictions during centralized training (Mu et al., 2026). Here, a common comparison scale and fixed normalization make utilities comparable in the cooperative average; trusted labels and reward models remain local, while critic communication may be Byzantine.

Decentralized and resilient actor–critic learning. Consensus actor–critic methods estimate team returns from heterogeneous local numerical rewards (Zhang et al., 2018); value propagation extends decentralized learning to nonlinear approximation (Qu et al., 2019). Resilient variants address Byzantine policy evaluation and control (Wu et al., 2021; Ye et al., 2024); exact cooperative-average evaluation cannot generally be guaranteed with heterogeneous rewards and unknown Byzantine identities (Hairi et al., 2024), separating critic agreement from target accuracy. Decentralized federated policy gradients instead use local environments (Jordan et al., 2024), whereas PARC learns rewards from private comparisons while agents jointly control one environment.

Robust aggregation on communication graphs. Weighted mean-subsequence-reduced (W-MSR) consensus links extreme-value trimming to graph robustness (LeBlanc et al., 2013; Sundaram & Gharesifard, 2019). ClippedGossip uses receiver-centered norm clipping (He et al., 2022); RTC combines distance-based removal, self replacement and clipping, with aggregation-error and finite-time nonconvex stationarity bounds (Yang & Ghaderi, 2024). Table 2 compares their selection and bounding operations. TPSC builds intervals from the complete raw row, then applies sender flags and retention at unchanged topology weights. Positive retention preserves honest support, yielding uniform finite-block scalar contraction for each fixed retention level and every admissible adapted selection (Appendix E.1). Screens may vary in accuracy and flagged-set size within the same graph conditions. Their effect on tracking is captured by the realized defect; a smaller verified net-window bound yields a tighter tracking bound with the other constants fixed. On fixed inputs, this defect splits into an interval-mediated residual and Byzantine and honest selection terms (Appendix E.6).

Prediction-space communication. Consensus distillation in function space (Taya et al., 2022) and federated Q-learning (Fan et al., 2023) exchange predictions. Ye et al. (2024) project differences between received post-update and local parameters onto current feature directions to aggregate scalar error estimates; their neural implementation also aggregates hidden-layer parameters using trimmed means. PARC exchanges reward- and value-critic predictions at shared probes, providing a common output interface for the functions used in actor updates without requiring aligned parameter coordinates for communication. Fixed invertible probe matrices permit exact linear-critic reconstruction,

Rule	Selection or removal	Bounding construction	Treatment of retained or in-range values
W-MSR	Trim scalar extremes relative to self	Raw honest coordinate range	Retained values are unchanged and averaged with normalized weights.
Ye et al. (2024)	Trim scalar update estimates along feature directions	Honest range in the projected scalar direction	Retained scalar estimates drive the parameter update.
ClippedGossip	Clip each neighbor vector	Receiver-centered norm ball	Vectors inside the ball are unchanged.
RTC	Remove distant vectors and replace them by self	Receiver-centered ball with radius computed from retained messages	Retained vectors inside the ball are unchanged.
TPSC	Adapted sender screen	Coordinate intervals built from the complete raw row before screening	Unflagged outliers are clipped; flagged outliers use self; flagged in-range deviations from self retain fraction κ_{keep}.

Table 2: **Selection and bounding operations in robust aggregation.** For a fixed raw row, changing the TPSC screen leaves its intervals unchanged; unflagged outliers are still clipped. Positive retention preserves the honest-support certificate, whose coefficient floor depends on the retention level (Appendix E.1).

whereas neural critics fit aggregated probe targets; the convergence theory covers the linear realization.

Stability on unbounded state spaces. Lyapunov-based methods study learning stability on unbounded state spaces (Shah et al., 2020; Pavse et al., 2024). Jiang et al. (2026) combine coordinate-wise W-MSR on critic parameters with synchronized mixing between learned and stabilizing policies. PARC uses this stabilization mechanism, with the executed policies additionally determining the comparison segments from which the rewards are identified.

Analysis with policy-generated rewards. Ye et al. (2024, Assumptions 1 and 3) use uniformly bounded features and rewards; their Theorem 10 describes actor limits through a projected mean ODE evaluated at the limiting critic estimates. PARC couples private reward identification to this learning loop. Under the stated moment and excitation conditions, a shared-sample disagreement bound and Lyapunov control establish critic boundedness for state-dependent updates, while a finite-block certificate covers admissible message-dependent rows. Theorem 3 relates the actor mean to the discounted team objective, separating aggregation, critic-approximation and occupancy errors.

B Algorithm Details

B.1 PARC Update Order

Critic proposals and the actor update use the stored time-t iterates; new critics and block-end preference updates enter subsequent transitions.

Algorithm 1 PARC at cooperative agent i

1: Initialize private η_i, critics (v_i, ω_i), actor θ_i, common $\lambda \in [0, 1)$ and $\mathbf{e}_0 = 0$, fixed probes (F_v, F_ω), topology row (a_{ij}), and retention $\kappa_{\text{keep}} \in (0, 1]$.
2: **for** $t = 0, 1, \ldots$ **do**
3: At s_t, select the shared event B_t; execute (4) with θ_t and observe $Z_t = (s_t, B_t, a_t, s_{t+1})$.
4: Update the trace and form local critic proposals (6) using the latest available private reward model.
5: Send the complete probe payload $\mathbf{y}_{i,t+1} = \text{col}(F_v \tilde{v}_{i,t+1}, F_\omega \tilde{\omega}_{i,t+1})$; receive the complete row and construct each raw interval as in Section 3.2.
6: Choose the shared screening and retention sets as in Section 3.2; apply TPSC once; reconstruct each critic by $x_{i,t+1} = F_x^{-1} \hat{\mathbf{y}}^x_{i,t+1}$.
7: Update the actor by (7) using the stored time-t critics and the observed transition.
8: At a completed comparison block, obtain the private label and update (5) for subsequent rounds.
9: **end for**

B.2 Fixed Preference Blocks

The comparisons in Section 2.1 use a fixed block length $L_{\text{p}} = 2H_{\text{p}} + g_{\text{p}}$ and boundaries $\tau_\ell = \ell L_{\text{p}}$. Within block ℓ, the two segments have entries $\zeta^1_{\ell,h} = Z_{\tau_\ell + h}$ and $\zeta^2_{\ell,h} = Z_{\tau_\ell + H_{\text{p}} + g_{\text{p}} + h}$ for $0 \le h < H_{\text{p}}$. The feature difference is

$$D_\ell = P_{\mathsf{E}} \left[\sum_{h=0}^{H_{\text{p}}-1} \gamma_{\text{p}}^h \chi(Z_{\tau_\ell + h}) - \sum_{h=0}^{H_{\text{p}}-1} \gamma_{\text{p}}^h \chi(Z_{\tau_\ell + H_{\text{p}} + g_{\text{p}} + h}) \right]. \tag{14}$$

Here γ_{p} is the within-segment comparison discount, distinct from the control discount γ; the undiscounted LQ teacher uses $\gamma_{\text{p}} = 1$ (Appendix G.1). Both segments are observed before the fresh label is drawn according to (2). The block-end update of η_i is used only in subsequent transitions, as specified in Algorithm 1.

The frozen-law interface is specified in Appendix D.3.

B.3 Sender Screens

For the geometric-median/MAD screen, let m_i be the geometric median of the complete raw payload row $\{\mathbf{y}_{j\to i} : j \in \mathcal{J}_i\}$ of dimension $K = d_v + d_\omega$. For each coordinate k, set $c_{i,k} = \text{median}_j\, y_{j\to i,k}$ and $d_{i,k} = \max\{\text{median}_j\, |y_{j\to i,k} - c_{i,k}|, \epsilon_i\}$, where $\epsilon_i = 10^{-12} \max\{1, \text{median}_{j,k}\, |y_{j\to i,k}|\}$. Define

$$\begin{aligned} q_{ij} &= K^{-1/2} \left\| \text{diag}(d_i)^{-1} (\mathbf{y}_{j\to i} - m_i) \right\|_2, & \bar{q}_i &= \text{median}_{j\in\mathcal{J}_i}\, q_{ij}, \\ \sigma_{ij} &= \mathbf{1}\{q_{ij} > \bar{q}_i + \tau_{\text{MAD}}\, \text{median}_{j'\in\mathcal{J}_i}\, |q_{ij'} - \bar{q}_i| + \epsilon_i\}, & \sigma_{ii} &= 0. \end{aligned} \tag{15}$$

The multiplier τ_{MAD} is specified by study in Table 4. The flagged-sender set is $\mathcal{S}_i = \{j \in \mathcal{J}_i \setminus \{i\} : \sigma_{ij} = 1\}$. With round indices, $\sigma_{ij,t} = \mathbf{1}\{j \in \mathcal{S}_{i,t}\}$. This one flag is shared by both critics and every probe coordinate.

All medians include the receiver's own complete payload. Scalar medians average the two central order statistics for an even number of observations, and median absolute deviations are unscaled. The geometric-median calculation starts at the coordinatewise median and uses deterministic modified Weiszfeld iterations with tolerance 10^{-10} and a cap of 128 iterations. Numerical coincidence uses distance at most $10^{-10} \max\{1, \max_{j,k} |y_{j,k}|\}$; stopping uses $\|m^+ - m\| \le 10^{-10} \max\{1, \|m\|\}$.

The distance screen instead flags the nonself sender maximizing $\|\mathbf{y}_{j\to i} - \mathbf{y}_i\|_2$, breaking ties by sender ID. It uses the same complete raw row and shares its flag across both critics and all coordinates.

For interval trimming in Section 3.2, cutoff ties between nonself senders are resolved in ascending sender-identifier order; values equal to self are retained.

Separate retention targets. Starting from full-retention TPSC-processed slots u_j, choose an adapted $\mathcal{T}_i \subseteq \mathcal{J}_i \setminus \{i\}$, shared across critics and probes, and replace u_j by $z_i + \kappa_{\mathrm{keep}}(u_j - z_i)$ for $j \in \mathcal{T}_i$. Taking $\mathcal{T}_i = \mathcal{S}_i$ recovers (8). RandomRetention keeps the distance screen and draws one nonself peer uniformly for $\mathcal{T}_i$ at each receiver and round. The positive-retention argument in Appendix E.1 applies to either choice; Proposition 3 gives the conditional mean of the random choice on a fixed row.

C Proofs of the three main results

The proofs use the conditions in Appendix D and the supporting estimates in Appendix E.

C.1 Private preference identification

Proof of Theorem 1. For an honest node i, write $e_\ell = \eta_{i,\ell} - \eta_i^\star$, $a_\ell = a_{\eta,\ell}$, $g_\ell = g_i(\eta_{i,\ell}, Z_\ell^{\mathrm{blk}})$, and $\bar g_\ell = \bar g_{i,\vartheta_{\tau_\ell}}(\eta_{i,\ell})$ for the frozen mean score. Calibration gives $\bar g_{i,\vartheta}(\eta_i^\star) = 0$; Lemma 9 then yields $\langle e_\ell, \bar g_\ell\rangle \ge \lambda_{\mathrm{pref},i}\|e_\ell\|^2$. Since $\eta_i^\star \in \mathcal{K}_i$, the projected update obeys

$$\begin{aligned} \|e_{\ell+1}\|^2 &\le \|e_\ell\|^2 - 2\lambda_{\mathrm{pref},i} a_\ell \|e_\ell\|^2 + s_\ell, \\ s_\ell &= -2a_\ell\langle e_\ell, g_\ell - \bar g_\ell\rangle + a_\ell^2\|g_\ell\|^2. \end{aligned} \tag{16}$$

Lemma 8 with $v(\eta) = \eta - \eta_i^\star$, and the score second-moment bound with $\sum_\ell a_\ell^2 < \infty$, imply $\sum_\ell s_\ell$ converges a.s. Its tail $R_\ell = \sum_{k\ge\ell} s_k \to 0$ makes $\|e_\ell\|^2 + R_\ell$ nonincreasing and bounded below. Hence $\|e_\ell\|^2$ converges and $\sum_\ell a_\ell\|e_\ell\|^2 < \infty$. Since $\sum_\ell a_\ell = \infty$, this limit is zero. □

C.2 Critic consensus and tracking

For one critic with honest stack $\mathbf{x}_t$, write $L = n^{-1}(\mathbf{1}^\top \otimes I_d)$, $J = (\mathbf{1}\otimes I_d)L$, $P_\perp = I - J$ and $\bar x_t = L\mathbf{x}_t$. Write $S_t = \max_k \mathrm{sp}[(I_n \otimes F)\mathbf{x}_t]^{(k)}$ for the honest probe span. Appendix E.1 gives the honest operator $A_t = A_t^{F,\sigma,+}$ for every realized message/flag sequence.

Consensus preservation gives the proposal and center recursions

$$\begin{aligned} \widetilde{\mathbf{x}}_{t+1} &= \mathbf{x}_t + \alpha_t[\mathbf{h}_{t+1} - (I_n \otimes G_{t+1})\mathbf{x}_t], \quad \mathbf{x}_{t+1} = A_t\widetilde{\mathbf{x}}_{t+1}, \\ \bar x_{t+1} &= \bar x_t + \alpha_t[\bar h_{t+1} - G_{t+1}\bar x_t + \Delta_t], \qquad \Delta_t = \alpha_t^{-1}L(A_t - I)P_\perp\widetilde{\mathbf{x}}_{t+1}. \end{aligned} \tag{17}$$

The center defect is exactly (10), pathwise.

Net aggregation-bias bound. Let $\mathfrak{D}_x$ bound the long-run net shift per unit of critic time. Formally, it is any deterministic bound such that, for every fixed $T > 0$, almost surely,

$$\Big\|\sum_{t=m}^{k} \alpha_{x,t}\Delta_{x,t}\Big\| \le T\mathfrak{D}_x + o(1) \quad \text{whenever} \quad \sum_{t=m}^{k}\alpha_{x,t} \le T,$$

uniformly over $k \ge m$ as $m \to \infty$. Summing before taking the norm allows opposing shifts to cancel. Lemma 11 ensures a finite choice.

Proof of Theorem 2. Fix a critic channel x and use its gain α_t and coercivity δ_x.

The shared G_{t+1} cancels the center in innovation differences. Lemma 10 therefore gives $S_t \le C_Q\alpha_{t-1}\mathsf{Z}_t \to 0$ without bounded critics, using a positive filter with bounded moments. Lemma 11 bounds its weighted defect windows.

Projection and the safeguard give

$$\|\theta_{t+1}-\theta_t\| \le c_n a_{\theta,t} R_t, \qquad d(\vartheta_{t+1},\vartheta_t) = o(\alpha_{x,t}) \quad (x=v,\omega). \tag{18}$$

Indeed, the descriptor increment is at most $L_\iota a_{\theta,t}(c_n R_t + C_\Pi) = o(\alpha_{x,t})$.

Put $\mathcal{B}_t = I - \alpha_t \bar{G}(\vartheta_t)$. Coercivity contracts its products; their Abel noise bound has factor $C_\Phi(T) = 1 + M_G(T+1)$, $M_G = \sup_\vartheta \|\bar{G}(\vartheta)\|$ (Appendix E.4).

Rewrite (17) as $\bar{x}_{t+1} = \mathcal{B}_t \bar{x}_t + \alpha_t[-(G_{t+1} - \bar{G}(\vartheta_t))\bar{x}_t + \bar{h}_{t+1}(\eta_{\ell(t)}) + \Delta_t]$. Keep the possibly fast reward parameter in the uncentered forcing, bounded by the compact-family envelope K_{t+1} in (87). On critic-time blocks of mass $T + o(1)$, set $r_j = 1 + \|\bar{x}_{t_j}\|$ and stop at the first $\|\bar{x}_t\| > C_0 r_j$ or block endpoint. Lemma 7's multiplicative estimate, including the exit step, and (97) bound normalized noise by $C_\Phi(T)\varepsilon_j$, with $\varepsilon_j \to 0$. Lemmas 6 and 11 give a deterministic eventual block bound C_T for $\sum \alpha_t(K_{t+1} + \|\Delta_t\|)$. Variation of constants consequently yields

$$\sup_{t_j \le t \le \tau_j} \|\bar{x}_t\| \le r_j + C_T + C_\Phi(T)\varepsilon_j r_j, \qquad r_{j+1} \le [e^{-\delta_x T/2} + C_\Phi(T)\varepsilon_j] r_j + 1 + C_T. \tag{19}$$

Choose $C_0 > 2 + C_T$. The first inequality excludes a late stopping exit, justifying the endpoint inequality. Its coefficient is eventually below $(1 + e^{-\delta_x T/2})/2 < 1$, bounding endpoints and interiors. Lemma 10 transfers this bound to all honest agents. Hence

$$\sup_t \|\bar{x}_t\| < \infty, \qquad \sup_t \max_{i \in \mathcal{N}^+} \|x_{i,t}\| < \infty \quad \text{a.s.} \tag{20}$$

For a fixed $T > 0$, define the backward-window quantities

$$m_{x,T}(t) = \max\{s < t : \textstyle\sum_{k=s}^{t-1} \alpha_{x,k} \ge T\}, \qquad W_{x,T}(t) = \textstyle\sum_{k=m_{x,T}(t)}^{t-1} \alpha_{x,k},$$
$$\overline{\Delta}^\star_{x,T} = \limsup_{t\to\infty} \frac{\sum_{k=m_{x,T}(t)}^{t-1} \alpha_{x,k}\|\Delta_{x,k}\|}{W_{x,T}(t)}. \tag{21}$$

For large t the anchor exists and $W_{x,T}(t) \to T$. Lemma 11's one-step bound controls the endpoint overshoot, giving $\overline{\Delta}^\star_{x,T} \le \overline{D}_{\Delta,x}$.

The learned reward adds $A^x_{i,t+1}(\eta_{i,\ell(t)} - \eta^\star_i)$, with $A^v = \mathbf{e}_{t+1}\chi_t^\top$ and $A^\omega = \psi_t \chi_t^\top$. Lemma 12 and Theorem 1 give

$$\sup_k \left\| \sum_{t=m}^{k} \alpha_{x,t} A^x_{i,t+1}(\eta_{i,\ell(t)} - \eta^\star_i) \right\| \le \sup_{t \ge m} \|\eta_{i,\ell(t)} - \eta^\star_i\| \sup_k \sum_{t=m}^{k} \alpha_{x,t} \|A^x_{i,t+1}\| \longrightarrow 0, \tag{22}$$

Both suprema range over the fixed critic-time window.

For $e_t = \bar{x}_t - x^\star(\vartheta_t)$ this gives $e_{t+1} = \mathcal{B}_t e_t + \alpha_t[\epsilon_{t+1}(\bar{x}_t) + \Delta_t + r^\eta_t + r^{\mathrm{tar}}_t]$, where ϵ is the centered latent affine forcing, r^η_t is the honest mean of the learned-reward residual above, and $r^{\mathrm{tar}}_t = -[x^\star(\vartheta_{t+1}) - x^\star(\vartheta_t)]/\alpha_t \to 0$ by Lipschitz targets and slow transport. Lemma 7 on deterministic critic tubes, followed by their countable exhaustion using (20), gives vanishing weighted affine-noise tails. Together with the residual bound, the product estimate for $0 < c < \delta_x$ and (97) then give

$$\|e_t\| \le e^{-cT}\|e_{m_{x,T}(t)}\| + \left\| \sum_{k=m_{x,T}(t)}^{t-1} \Phi(t,k+1)\alpha_{x,k}\Delta_{x,k} \right\| + o(1). \tag{23}$$

For the net-window certificate of Section 4.2, put $N_k = \sum_{j=m}^{k} \alpha_j \Delta_j$ on the backward window. The one-step bound $\alpha_j\|\Delta_j\| \to 0$ controls its endpoint overshoot, so $\sup_k \|N_k\| \le T\mathfrak{D}_x + o(1)$. For sufficiently late windows, all products are nonexpansive and

$$\sum_{k=m}^{t-2} \|\Phi(t,k+1) - \Phi(t,k+2)\| \le M_G \sum_{k=m}^{t-2} \alpha_{k+1} \le M_G(T + o(1)).$$

Abel's identity (97) consequently bounds the propagated net defect by $(1 + M_G T)T\mathfrak{D}_x + o(1)$. The sharper $T + o(1)$ mass bound here, rather than the convenient $T + 1$ bound used for stopping, is needed as $T \downarrow 0$. For the finite $E^\star = \limsup_t \|e_t\|$, (23) gives

$$E^\star \le e^{-cT} E^\star + (1 + M_G T)T\mathfrak{D}_x, \qquad E^\star \le \frac{(1 + M_G T)T}{1 - e^{-\delta_x T}}\mathfrak{D}_x. \tag{24}$$

The second inequality uses $c_j = \delta_x(1 - 1/j) \uparrow \delta_x$, $j \geq 2$, on the common probability-one event. Intersect also over $T_m = 1/m$ and let $m \to \infty$ to obtain $E^\star \leq \mathfrak{D}_x/\delta_x$; consensus gives Theorem 2. Existence of a finite net-window bound follows from Lemma 11's absolute-window bound and the triangle inequality. Lemma 13 optionally supplies an absolute bound in terms of utility heterogeneity; its extra field conditions are not used in boundedness or in this net-window argument.

For the absolute-window certificate, nonexpansiveness in (23) gives $E^\star \leq e^{-cT}E^\star + T\overline{\Delta}^\star_{x,T}$. Taking the same countable limit $c \uparrow \delta_x$ and using consensus yields

$$\limsup_t \|x_{i,t} - x^\star_x(\vartheta_t)\| \leq K_x(T)\overline{\Delta}^\star_{x,T}, \qquad K_x(T) = \frac{T}{1 - e^{-\delta_x T}}. \tag{25}$$

For fixed numeric rewards, $r^\eta_t = 0$ gives the same comparison; eventual average preservation gives $\Delta_t = 0$ and exact tracking.

□

C.3 Discounted-objective stationarity

Proof of Theorem 3. Fix $h > 0$, ν and $\varpi = \varpi_\star$ on the common probability-one event; uniform family bounds permit this final-rule substitution. Write $\mathcal{J} = (1 - \gamma)J_{\mathrm{pref}}(\cdot; \nu, \varpi)$, $b_t = a_{\theta,t}$, and let $\bar{\theta}$ interpolate the actors at $\mathsf{t}_t = \sum_{k<t} b_k$. The latent actor field is

$$g_i(\theta) = \mathbb{E}_{\theta,\varpi}[B\, \delta^\star_\theta(Z)\nabla_{\theta_i} \log \pi_i(a_i \mid s; \theta_i)]\,, \tag{26}$$

with the TD signal (7) evaluated at (9) under the executed invariant law.

Let w_t stack the honest critics, $\upsilon_t = (\theta_t, \varpi_t)$, and $\overline{\mathsf{H}}$ be the stationary actor mean. Its continuity, that of the latent targets $w^\star$, and $\varpi_t \to \varpi_\star$ give $\overline{\mathsf{H}}(\upsilon_t, w^\star(\vartheta_t)) - g(\theta_t) \to 0$. The critic bound and actor-map certificate imply

$$\limsup_t \|\overline{\mathsf{H}}(\upsilon_t, w_t) - g(\theta_t)\| \leq \epsilon_{\mathrm{ac}}. \tag{27}$$

Throughout this proof, ϵ_{ac} may be any deterministic upper bound in (27), including one obtained directly from the actor map. Project this bias onto $\epsilon_{\mathrm{ac}}\mathbb{B}_2$ to obtain u_t; its remainder e_t vanishes. After the finite safeguard prefix,

$$\theta_{t+1} = \Pi_\Theta[\theta_t + a_{\theta,t}\{g(\theta_t) + u_t + e_t + \zeta_{t+1}\}], \quad \|u_t\| \leq \epsilon_{\mathrm{ac}}, \quad e_t \to 0. \tag{28}$$

Here ζ_{t+1} centers the current edge sample at the old critics. Lemma 15 gives eventual safeguard inactivity and vanishing noise windows (125). Since $\|\theta_{t+1} - \theta_t\| \leq \sqrt{n}\, b_t R_t \to 0$, Lemma 16 gives relatively compact late translates with viable limits in the field (60).

The discounted gradient identity for (132) separates the remaining errors (suppressing θ):

$$g - \nabla\mathcal{J} = \underbrace{(g - \rho F_\gamma)}_{\text{critic approximation}} + \underbrace{(\rho - d_\gamma)F_\gamma}_{\text{occupancy}}.$$

For the invariant and discounted laws ρ, d_γ, Lemma 19 bounds these terms by $S_\theta(E_r + \gamma C_\lambda E_V)$ and $b_{\mathrm{occ}}(\theta)$, with uniform sum b_{obj} in (12). Set $q = \nabla\mathcal{J}$ and $\delta = \epsilon_{\mathrm{ac}} + b_{\mathrm{obj}}$; then $\|g - q + u\| \leq \delta$ whenever $\|u\| \leq \epsilon_{\mathrm{ac}}$.

For every $\theta \in \Theta$ and $v \in \mathcal{F}^K_{\epsilon_{\mathrm{ac}}}(\theta)$, we show

$$\langle q(\theta), v\rangle \geq \tfrac{1}{4}\|G_h(\theta)\|^2 - \delta^2. \tag{29}$$

For $v = \Pi_{T_\Theta(\theta)}(q + e)$, $\|e\| \leq \delta$ and $x = \|v\|$, Moreau orthogonality and nonexpansiveness give $q \cdot v = \|v\|^2 - e \cdot v \geq x^2 - \delta x$ and $\|\Pi_{T_\Theta(\theta)}q\| \leq x + \delta$. The tangent/normal decomposition $q = p + n$ gives $\Pi_\Theta(\theta + hn) = \theta$ and $\|G_h(\theta)\| \leq \|p\| \leq x + \delta$. Since $x^2 - \delta x - \frac{1}{4}(x+\delta)^2 + \delta^2 = \frac{3}{4}(x - \delta)^2 \geq 0$, (29) follows. This covers all raw velocities of $g + \epsilon_{\mathrm{ac}}\mathbb{B}_2$. Continuity of q, G_h and bounded velocities extend it through nearby-point limits, convex combinations and closure to the full Krasovskii field, including the boundary.

Integrating (29) along limit arcs, then using uniform continuity of $\mathcal{J}, G_h$ and translated-path compactness, gives on every late window of fixed length L:

$$\int_r^{r+L} \|G_h(\bar{\theta}(s))\|^2 \, ds \le 4[\mathcal{J}(\bar{\theta}(r+L)) - \mathcal{J}(\bar{\theta}(r))] + 4\delta^2 L + \varepsilon_L(r), \qquad \varepsilon_L(r) \longrightarrow 0. \quad (30)$$

Otherwise a convergent subsequence of violating windows contradicts the limit-arc bound. Sum consecutive windows and divide by actor time: the bounded objective telescopes, remainders average to zero, and initial and partial windows have vanishing weight. Thus

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T \|G_h(\bar{\theta}(s))\|^2 \, ds \le 4\delta^2. \quad (31)$$

For a modulus of continuity m_f of $f = \|G_h\|^2$,

$$\left| \int_{\mathfrak{t}_t}^{\mathfrak{t}_{t+1}} f(\bar{\theta}(s)) \, ds - b_t f(\theta_t) \right| \le b_t m_f(\|\theta_{t+1} - \theta_t\|).$$

Summing and dividing by $B_N \to \infty$ makes this discrepancy vanish, proving (13). For every $\xi > 0$, the pathwise inequality

$$\mathbf{1}_{\{\|G_h(\theta_t)\| > \xi\}} \le \xi^{-2} \|G_h(\theta_t)\|^2$$

gives, on the same probability-one event,

$$\limsup_{N\to\infty} \frac{\sum_{t<N} a_{\theta,t} \mathbf{1}_{\{\|G_h(\theta_t)\| > \xi\}}}{\sum_{t<N} a_{\theta,t}} \le \frac{4\delta^2}{\xi^2}. \quad (32)$$

If $\delta > 0$, take $\xi = 4\delta$: the liminf of the normalized actor-gain weight inside $\{\|G_h\| \le 4\delta\}$ is at least $3/4$. If $\delta = 0$, the outside weight tends to zero for every fixed $\xi > 0$. The average bound also implies $\liminf_t \|G_h(\theta_t)\| \le 2\delta$, since a larger liminf would contradict $\sum_t b_t = \infty$. The corresponding unnormalized bounds follow from

$$G_h^{\mathcal{J}} = (1-\gamma) G_{h(1-\gamma)}^{J_{\mathrm{pref}}}. \quad (33)$$

The norm threshold is $2\delta/(1-\gamma)$ at step $h(1-\gamma)$, and the squared-average bound is $4\delta^2/(1-\gamma)^2$. □

D Conditions for the six main assumptions

All bounds are uniform on the compact safe family $\mathcal{Z}_{\mathrm{safe}}$. The pre-action history $\mathcal{F}_t$ contains s_t, actors, critics, trace, available reward estimates and completed rounds. For the predictable full-law descriptor $\vartheta_t = \iota(\theta_t, \varpi_t)$,

$$Z_t = (s_t, B_t, a_t, s_{t+1}), \qquad \mathcal{L}(Z_t \mid \mathcal{F}_t) = Q_{\vartheta_t}(s_t, dz), \qquad P_\vartheta(s, D) = \int \mathbf{1}_{\{s' \in D\}} Q_\vartheta(s, dz). \quad (34)$$

The switch, actions and next state are drawn afterward; updates and block-end labels enter $\mathcal{F}_{t+1}$. Thus $\phi_t = \phi(s_t)$, $\psi_t = \psi(Z_t)$, $\chi_t = \chi(Z_t)$, and the actor multiplies the current edge by the old critics. Write ρ_ϑ, $\mu_\vartheta = \rho_\vartheta Q_\vartheta$, and $\nu_{\vartheta,\lambda}$ for the invariant state, edge and stationary-past trace laws. Expectations use the relevant law; drift and geometric bounds concern P_ϑ.

D.1 Execution regularity and physical-policy evolution

Executed descriptors and physical policies. The executed-descriptor map $\vartheta = \iota(\theta, \varpi)$ satisfies $d(\iota(\theta', \varpi'), \iota(\theta, \varpi)) \le L_\iota(\|\theta' - \theta\| + d_\Pi(\varpi', \varpi))$. The adapted physical policy satisfies $\varpi_t \to \varpi_\star$ and $d_\Pi(\varpi_{t+1}, \varpi_t) \le C_\Pi a_{\theta,t}$ almost surely in d_Π, with $C_\Pi < \infty$.

Drift and weighted regularity. The frozen chains are ψ-irreducible and strongly aperiodic. For a common small set C, a Lyapunov envelope $V \ge 1$, $0 < \rho_V < 1$, $0 < b < \infty$, $\epsilon_C > 0$, and probability measure ν_C,

$$P_\vartheta V \le \rho_V V + b\mathbf{1}_C, \qquad P_\vartheta(s, \cdot) \ge \epsilon_C \nu_C(\cdot) \quad (s \in C). \quad (35)$$

Assume $\mathbb{E}V(s_0) < \infty$. For an observable weight $W \geq 1$, write $\|f\|_{W^j} = \sup_s \|f(s)\|/W(s)^j$. For $j = 1, 2$, require

$$\|P_{\vartheta'} - P_\vartheta\|_{W^j \to W^j} \leq L_P d(\vartheta', \vartheta), \tag{36}$$

$$\sup_\vartheta \|P_\vartheta^m - \Pi_\vartheta\|_{W^j \to W^j} \leq C_P \rho_P^m, \qquad \Pi_\vartheta f = \rho_\vartheta(f)\mathbf{1}, \quad 0 < \rho_P < 1. \tag{37}$$

The drift gives, including adapted descriptors,

$$\sup_t \mathbb{E}V(s_t) \leq \mathbb{E}V(s_0) + \frac{b}{1-\rho_V}, \qquad \sup_\vartheta \rho_\vartheta(V) \leq \frac{b}{1-\rho_V}. \tag{38}$$

Iterate the conditional drift; then use frozen convergence on $V \wedge M$ and $M \uparrow \infty$ for the stationary bound.

Safe-policy construction of the drift. Following the mixture argument of Jiang et al. (2026, Proposition 3.9 and Appendix C.1), we give a sufficient drift condition for the present execution family. Let P_ϖ^{safe} and $P_{\theta,\varpi}^{\mathrm{learn}}$ denote the safe and learning state kernels. Suppose the safe kernels have the common minorization in (35), $P_\varpi^{\mathrm{safe}} V \leq \rho_{\mathrm{s}} V + b\mathbf{1}_C$ with $0 < \rho_{\mathrm{s}} < 1$, and $P_{\theta,\varpi}^{\mathrm{learn}} V \leq KV$ with $K \geq 1$, uniformly over $\theta \in \Theta$ and admissible ϖ. Choose $c > 0$ such that $\rho_V := \rho_{\mathrm{s}} + (K - \rho_{\mathrm{s}})c < 1$; then $c < 1$ and (4) gives

$$\begin{aligned} P_\vartheta V &\leq [\rho_{\mathrm{s}} + (K - \rho_{\mathrm{s}})\varepsilon(s)]V + b\mathbf{1}_C \leq \rho_V V + b\mathbf{1}_C, \\ P_\vartheta(s, \cdot) &\geq (1-c)\epsilon_C \nu_C(\cdot), \qquad s \in C. \end{aligned} \tag{39}$$

Thus the mixture preserves common drift and minorization, supplying (38) even along adapted executions. This is a sufficient check for the state-process conditions; the weighted regularity and update-field moments are imposed separately, as in general-state actor–critic analysis (Konda & Tsitsiklis, 2003).

Update fields and conditional means. Choose one moment order $q_{\mathrm{cat}} = q > 4$, as in Assumption 1. The base catalogue comprises:

(i) Shared samples and reward regression: $\phi_t, \phi_{t+1}, \psi_t, \chi_t, G_\omega, h_i^\omega$, uniformly over $\eta_i \in \mathcal{K}_i$.

(ii) Reward-critic envelopes: $H_\omega^{\mathcal{K}}, U_\omega^{\mathcal{K}}, K_\omega^{\mathcal{K}}, g_\omega$ from (87).

(iii) Reward-model residuals: $A^\omega = \psi\chi^\top$ and its operator norm.

(iv) Comparisons: $D, DD^\top, \|D\|^2$, the BT score and its parameter-Lipschitz envelope.

(v) Actor updates: the primitive coefficient K^{ac} in (47) and, on each compact critic tube, the actor sample, frozen mean and $(\theta, \varpi, v, \omega)$-Lipschitz envelope.

For each centered edge field $f_{\vartheta,p}$, including deterministic critic tubes, require

$$\bar{f}_{\vartheta,p}(s) = \int f_{\vartheta,p}(z) Q_\vartheta(s, dz), \qquad \sup_{\vartheta,p} \|\bar{f}_{\vartheta,p}\|_W < \infty, \tag{40}$$

and Lipschitzness in (ϑ, p) in $\mathcal{B}_W$ on compact parameter sets, before trace multiplication. Include $\|\chi\|$ and $\|F_\omega\psi\|_\infty\|\chi\|$. Comparison fields use W_B below; envelopes cover adapted choices.

A sufficient primitive check for an edge weight $W_E(z) = W(s) + W(s')$ is

$$\begin{aligned} \int W_E(z) Q_\vartheta(s, dz) &\leq CW(s), \\ \int W_E(z) |Q_{\vartheta'} - Q_\vartheta|(s, dz) &\leq L d(\vartheta', \vartheta) W(s), \end{aligned} \tag{41}$$

with raw size and parameter-Lipschitz envelopes CW_E: add and subtract $Q_\vartheta f_{\vartheta',p'}$ to obtain (40). Products and blocks may use compatible weights. Marginal P-regularity alone does not imply these marked-law bounds.

Moment orders and weighted products. All listed fields and their products with either adjacent W have uniform q_{cat}-moments, including $W^2, WU_\omega^{\mathcal{K}}, WK_\omega^{\mathcal{K}}, WK^{\mathrm{ac}}$. Block counterparts use either boundary weight W_B, including W_BK^{ac}; bare W_B needs only second moments. Bounds cover the controlled path, all frozen edge/block laws and the initial critic stack. In particular, the initial honest critic stack is q_{cat}-integrable.

Eligibility traces and derived products. The value critic additionally requires $\phi \in \mathcal{B}_W$ and

$$\sup_t \mathbb{E}W(s_t)^{2q_{\mathrm{cat}}} < \infty, \qquad \sup_\vartheta \rho_\vartheta(W^{2q_{\mathrm{cat}}}) < \infty. \tag{42}$$

A sufficient check is $W^{2q_{\mathrm{cat}}} \le C_W V$, for example $W = V^\beta$ with $2\beta q_{\mathrm{cat}} \le 1$. State trace generators are coordinates of ϕ and $\|\phi\|$; raw edge multipliers are coordinates of $\chi(Z_t)$, $\|\chi(Z_t)\|$, $\phi_b(s_t) - \gamma\phi_b(s_{t+1})$, and the fixed latent utility multipliers. For each such multiplier $k_\vartheta^{\mathrm{edge}}$, require either

$$|k_\vartheta^{\mathrm{edge}}(z)| \le C[W(s) + W(s')], \quad \text{or} \quad \mathbb{E}[|k_{\vartheta_t}^{\mathrm{edge}}(Z_t)|^{2q_{\mathrm{cat}}} \mid \mathcal{F}_t] \le CW(s_t)^{2q_{\mathrm{cat}}}, \tag{43}$$

with the frozen counterpart and separate conditional-mean regularity (40). Hölder gives (78) and $q_{\mathrm{cat}}/2$ moments for pair-products, without moments of unbounded-critic actor samples or random eventual critic bounds.

For $x = \omega$, Lemma 10's deterministic-start filter inherits the primitive moments; Hölder gives

$$\sup_t \mathbb{E}[W_t^2 + W_tU_{t+1} + W_tK_{t+1}^{\mathcal{K}} + W_t\mathsf{Z}_t + g_{t+1}\mathsf{Z}_t + W_tg_{t+1}\mathsf{Z}_t] < \infty, \tag{44}$$

also with W_{t+1} in place of $W_t = W(s_t)$. For $x = v$, use (78). Finite random filter transients are handled pathwise.

D.2 Learning design and time-scale compatibility

All gains are positive, deterministic, eventually nonincreasing, and satisfy $\sum a = \infty, \sum a^2 < \infty$. In addition, $\alpha_{x,t+1}/\alpha_{x,t} \to 1$. Numeric-reward tracking requires

$$d(\vartheta_{t+1}, \vartheta_t) \le \alpha_{x,t}c_{x,t}, \qquad c_{x,t} \to 0 \quad \text{a.s.} \tag{45}$$

Under the descriptor and physical-policy conditions of Appendix D.1, the coupled safeguard implies this when

$$R_t \uparrow \infty, \qquad \sum_t R_t^{-q_{\mathrm{cat}}} < \infty, \qquad \max_x a_{\theta,t}R_t/\alpha_{x,t} \to 0. \tag{46}$$

The primitive coefficient K^{ac} below has a uniform q_{cat}-moment, so $\sum_t \Pr(K_{t+1}^{\mathrm{ac}} > cR_t) < \infty$ for every $c > 0$. Define the safeguard coefficient

$$K_{t+1}^{\mathrm{ac}} = B_t\Big(\sum_{i\in\mathcal{N}^+} \|\nabla_{\theta_i} \log \pi_i(a_{i,t} \mid s_t; \theta_{i,t})\|^2\Big)^{1/2}[\|\psi_t\| + \|\phi_t\| + \gamma\|\phi_{t+1}\|]. \tag{47}$$

D.3 Calibrated preferences and block transport

Use the calibrated label law, gauge and compact convex $\mathcal{K}_i$ of Assumption 3 and the blocks of Appendix B.2. Write $X_\ell = s_{\tau_\ell}$ and $X_{\ell+1} = s_{\tau_\ell+L_{\mathrm{p}}}$ for their boundary states, and Z_ℓ^{blk} for a whole block and label. The frozen marked law $\mathcal{Q}_\vartheta^{\mathrm{blk}}$ composes L_{p} copies of Q_ϑ and the trusted label kernel; its boundary marginal is $K_\vartheta = P_\vartheta^{L_{\mathrm{p}}}$. For its stationary law $\nu_\vartheta^{\mathrm{blk}}$, require, for $B_D < \infty$, $\lambda_D > 0$, uniform bounded information:

$$\mathbb{E}_{\nu_\vartheta^{\mathrm{blk}}}[DD^\top \mathbf{1}_{\{\|D\| \le B_D\}}] \succeq \lambda_D I_{\mathsf{E}}. \tag{48}$$

Online block transport. Choose $W_B \ge 1$. For the block score $g_i(\eta, Z^{\mathrm{blk}}) = \beta_{\mathrm{p}}[\sigma(\beta_{\mathrm{p}}D^\top\eta) - Y_i]D$, the conditional field $f_{i,\vartheta,\eta}(x) = \mathbb{E}_\vartheta[g_i(\eta, Z^{\mathrm{blk}}) \mid X_\ell = x]$, its η-derivative, and its scalarizations by bounded Lipschitz functions of η obey the catalogue in W_B-norm. The boundary kernel has the corresponding weighted geometric bound and the operator regularity

$$\|K_{\vartheta'} - K_\vartheta\|_{W_B \to W_B} \le L_B d(\vartheta', \vartheta).$$

The actual kernels $\widetilde{\mathcal{Q}}_\ell^{\mathrm{blk}}$, $\widetilde{K}_\ell$ are conditional on $\mathcal{F}_{\tau_\ell}$ and include within-block actor and physical-policy updates. With $d_\ell^{\mathrm{blk}} = d(\vartheta_{\tau_{\ell+1}}, \vartheta_{\tau_\ell})$, require $\delta_\ell^{\mathrm{blk}} \ge 0$ such that the actual score $\widetilde{f}_\ell$ and each normalized block Poisson solution $\widehat{f}$ satisfy

$$\|(\widetilde{f}_\ell - f_{i,\vartheta_{\tau_\ell},\eta})(X_\ell)\| + \|[(\widetilde{K}_\ell - K_{\vartheta_{\tau_\ell}})\widehat{f}](X_\ell)\| \le C\delta_\ell^{\mathrm{blk}} W_B(X_\ell). \tag{49}$$

Require

$$\sum_\ell a_{\eta,\ell}[\delta_\ell^{\mathrm{blk}} W_B(X_\ell) + d_\ell^{\mathrm{blk}} W_B(X_{\ell+1})] < \infty \quad \text{a.s.} \tag{50}$$

A sufficient deterministic clipped-path schedule is

$$\sum_\ell a_{\eta,\ell} \sum_{t=\tau_\ell}^{\tau_{\ell+1}} a_{\theta,t}(R_t + 1) < \infty. \tag{51}$$

Under marked-kernel regularity at compatible block weights (41), telescope conditional law deviations. The safeguard and physical-policy rate bound every within-block history by the inner sum; boundary moments and Tonelli give (50). Also require

$$\bar{\alpha}_{x,\ell}/a_{\eta,\ell} \to 0, \qquad \bar{a}_{\theta,\ell}/\bar{\alpha}_{x,\ell} \to 0, \tag{52}$$

where bars sum gains over one block, separately from transport.

For fixed block length and $q_{\mathrm{cat}} > 4$, a polynomial schedule satisfying the gain, safeguard and separation requirements and (51) is

$$a_{\eta,\ell} = (\ell+1)^{-.55}, \quad \alpha_{v,t} = \alpha_{\omega,t} = (t+1)^{-.65}, \quad a_{\theta,t} = (t+1)^{-1}, \quad R_t = (t+1)^{.30}. \tag{53}$$

Here $a_{\theta,t}R_t/\alpha_{x,t} = O(t^{-.05})$ and the summand in (51) is $O(\ell^{-1.25})$.

D.4 Critic feature excitation

The proposals (6) have affine innovations $h_{i,t+1}^x(\eta_i) - G_{t+1}^x x_i$, where

$$\begin{aligned} G_{t+1}^v &= \mathbf{e}_{t+1}(\phi_t - \gamma\phi_{t+1})^\top, & h_{i,t+1}^v(\eta_i) &= \mathbf{e}_{t+1}\chi_t^\top \eta_i, \\ G_{t+1}^\omega &= \psi_t\psi_t^\top, & h_{i,t+1}^\omega(\eta_i) &= \psi_t\chi_t^\top \eta_i. \end{aligned} \tag{54}$$

Shared trajectories, features and trace initialization make G^x common to all honest nodes. Require

$$C_\phi(\vartheta) = \mathbb{E}_\vartheta[\phi\phi^\top] \succeq c_\phi I, \qquad C_\psi(\vartheta) = \mathbb{E}_\vartheta[\psi\psi^\top] \succeq c_\psi I. \tag{55}$$

Lemmas 4–5 give bounded Lipschitz means, and the Grams give

$$\tfrac{1}{2}(\bar{G}_x + \bar{G}_x^\top) \succeq \delta_x I, \qquad \|\bar{G}_x\| \le M_x, \qquad \|\bar{h}_x\| \le H_x. \tag{56}$$

Thus $x_x^\star = \bar{G}_x^{-1}\bar{h}_x$ is bounded and Lipschitz, with $L_{x,\star} \le L_{h,x}/\delta_x + H_x L_{G,x}/\delta_x^2$. For the reward critic, $\bar{G}_\omega = C_\psi$. For the value critic, with $\Phi u = \phi^\top u$ in $L^2(\rho_\vartheta)$,

$$\bar{G}_v = \Phi^*(I - \gamma P_\vartheta)(I - \gamma\lambda P_\vartheta)^{-1}\Phi, \qquad \operatorname{sym}\bar{G}_v \succeq \frac{1-\gamma}{1-\gamma\lambda} C_\phi. \tag{57}$$

Expand the stationary trace and use $\|P_\vartheta\|_{2\to 2} \le 1$ on $I - \gamma(1-\lambda)P_\vartheta(I - \gamma\lambda P_\vartheta)^{-1}$. Thus $\delta_v = (1-\gamma)c_\phi/(1-\gamma\lambda)$ and $\delta_\omega = c_\psi$.

D.5 Resilient communication

The fixed Byzantine set is H-total; the static digraph is $(2H+1)$-robust with honest in-degrees at least $2H+1$. Each fixed base row sums to one and its positive coefficients satisfy $a_{ij} \ge a_\star > 0$. The d_x fixed probes form an invertible F_x with fixed inverse norm. The rule is the TPSC family of Section 3.2, with fixed $\kappa_{\mathrm{keep}} \in (0,1]$ and unchanged topology weights. Screening and retention sets are adapted, exclude self, and are each shared across critics and probes; they may be selected separately. Messages are finite, adapted, and may equivocate by receiver and probe; sender selection has no accuracy or cardinality premise.

D.6 Actor dynamics and objective regularity

Projected dynamics and critic sensitivity. Policies are positive and twice continuously differentiable on compact convex Θ_i. On $\Theta = \prod_i \Theta_i$, the Euclidean normal-reflection map for step inputs, including their initial values, obeys

$$\begin{aligned} \|\mathscr{R}_\Theta(z) - \mathscr{R}_\Theta(\widetilde{z})\|_{\infty,[0,T]} &\le C_\Theta \|z - \widetilde{z}\|_{\infty,[0,T]}, \\ \operatorname{osc}(\mathscr{R}_\Theta(z); [s,t]) &\le C_\Theta \operatorname{osc}(z; [s,t]). \end{aligned} \tag{58}$$

The same bounds extend to continuous and absolutely continuous inputs; Cartesian boxes suffice. The synchronized mixture belongs to the safe family, with $\varepsilon(s) \le c/(1 + \|s\|)$. Its learning-event law at the final physical policy is

$$\mu^{\mathrm{L}}_{\theta,\varpi_\star}(ds, da) = \varepsilon(s)\rho_{\theta,\varpi_\star}(ds)\pi^+_\theta(da^+ \mid s)\pi^-_{\varpi_\star}(da^- \mid s). \tag{59}$$

The frozen actor field is L_F-Lipschitz in the critic stack. For $\epsilon \ge 0$, define

$$\mathcal{F}^K_\epsilon(\theta) = \bigcap_{\delta>0} \overline{\mathrm{co}} \bigcup_{\substack{\theta' \in \Theta \\ \|\theta' - \theta\| < \delta}} \{\Pi_{T_\Theta(\theta')}[g(\theta') + b] : \|b\| \le \epsilon\}. \tag{60}$$

Differentiation of the frozen discounted objective. Fix an admissible final rule ϖ; use subscripts (θ, ϖ) for descriptor $\iota(\theta, \varpi)$. For $r = n^{-1}\sum_{i \in \mathcal{N}^+} r^\star_i$ and $r_{\theta,\varpi}(s) = (Q_{\theta,\varpi} r)(s)$, require honest-policy and kernel extensions to an open $U \supset \Theta$ for which

$$\theta \mapsto P_{\theta,\varpi} \in \mathcal{L}(\mathcal{B}_W), \qquad \theta \mapsto r_{\theta,\varpi} \in \mathcal{B}_W \tag{61}$$

are continuously differentiable in the displayed norms with uniformly bounded derivatives. With the synchronized score $\ell_\theta = B(\nabla_{\theta_i} \log \pi_i(a_i \mid s; \theta_i))_i$, for every $f \in \mathcal{B}_W$ require

$$\begin{aligned} \nabla r_{\theta,\varpi}(s) &= \mathbb{E}_{Q_{\theta,\varpi}(s)}[\ell_\theta r], \\ [(\nabla P_{\theta,\varpi}) f](s) &= \mathbb{E}_{Q_{\theta,\varpi}(s)}[\ell_\theta f(s')], \qquad \mathbb{E}_{Q_{\theta,\varpi}(s)} \ell_\theta = 0. \end{aligned} \tag{62}$$

Dominated differentiable action densities suffice if their derivative integrals against $|r| + W(s')$ are uniformly integrable and weighted-norm continuous. The transition, safe law, switch, utility and physical rule have no other θ-dependence. For finite constants R, S, A_r, A_P, uniformly on the admitted family,

$$\begin{aligned} \|r_{\theta,\varpi}\|_W &\le R, & \mathbb{E}_{\mu_{\theta,\varpi}} \|\ell_\theta\|^2 &\le S^2, \\ \|\mathbb{E}[\ell_\theta r \mid s]\| &\le A_r W(s), & \|\mathbb{E}[\ell_\theta f(s') \mid s]\| &\le A_P \|f\|_W W(s). \end{aligned} \tag{63}$$

The θ-independent initial law satisfies $\nu W < \infty$ and the Lyapunov moment if used online. Bounds are uniform in final rules, including continuity bounds; the extension defines boundary gradients with any random $\varpi_\star$ held fixed pathwise.

E Supporting derivations

E.1 Scalar communication and fixed-probe contraction

Fix a receiver i, critic, probe, and round. Write $z_{j\to i}$ for a raw scalar, z_i for the honest self proposal, $\mathcal{R}_i$ for the slots retained by the strict self-centered trim, and $I_i = [L_i, U_i]$ for their range. Write $T_i(z_{j\to i}; \sigma_{ij})$ for the map (8) with its sender flag explicit. Condition on the raw row and shared flags; $d_i^- = |\mathrm{In}(i)|$ is the in-degree. The next lemma and proposition first treat full-retention TPSC ($\kappa_{\mathrm{keep}} = 1$).

Lemma 1 (Honest interval and unchanged survivors)**.** *Let $m_i = \min_{h \in \mathcal{J}_i \cap \mathcal{N}^+} z_h$ and $M_i = \max_{h \in \mathcal{J}_i \cap \mathcal{N}^+} z_h$. Then $I_i \subseteq [m_i, M_i]$, every transformed scalar lies in I_i, and $T_i(z_{j\to i}; \sigma_{ij}) = z_{j\to i}$ for $j \in \mathcal{R}_i$.*

Proof. At most H slots lie above M_i; they are Byzantine, strictly above z_i, and among the H largest upper values. Trimming removes them, also when fewer than H upper slots exist. The lower side is symmetric. Self and ties remain, so $I_i \subseteq [m_i, M_i]$. Transformation preserves retained values; outsiders are projected into I_i or replaced by $z_i \in I_i$. □

Proposition 1 (Honest-row support certificate)**.** *For full-retention TPSC, every realized row admits a nonnegative honest row c_i such that*

$$\widehat{z}_i = \sum_{h\in\mathcal{N}^+} c_{ih} z_h, \qquad \sum_h c_{ih} = 1, \qquad c_{ii} \geq a_\star, \\ \#\{h \in \mathrm{In}(i) \cap \mathcal{N}^+ : c_{ih} \geq a_\star/2\} \geq d_i^- - 2H. \tag{64}$$

The row is supported on self and original honest in-neighbors.

Proof. Let $s_i = \sum_{j\in\mathcal{R}_i} a_{ij} > 0$ and $w_i = \sum_{j\in\mathcal{R}_i}(a_{ij}/s_i) z_{j\to i}$. The normalized survivor coefficients have lower bound $\lambda_i^{\mathrm{surv}} \geq a_\star/s_i$. The scalar survivor representation of Sundaram & Gharesifard (2019, Proposition 5.1), applied with local bound H and $d_i^- \geq 2H+1$, supplies an honest row q_i with $q_{ii} \geq \lambda_i^{\mathrm{surv}}$ and at least $d_i^- - 2H$ other honest coefficients at least $\lambda_i^{\mathrm{surv}}/2$. For an excluded transformed value $u_j \in [m_i, M_i]$, when $m_i < M_i$,

$$u_j = \lambda(u_j) m_i + [1-\lambda(u_j)] M_i, \qquad \lambda(u) = \frac{M_i - u}{M_i - m_i}. \tag{65}$$

Both endpoints are honest raw values. If they coincide, use one unit mass at that value. Denote the resulting honest probability row by $p^{[j]}$. By Lemma 1,

$$\widehat{z}_i = s_i w_i + \sum_{j\notin\mathcal{R}_i} a_{ij} u_j, \qquad c_{ih} = s_i q_{ih} + \sum_{j\notin\mathcal{R}_i} a_{ij} p_h^{[j]}. \tag{66}$$

The coefficients sum to one; scaling the survivor floors by s_i proves (64). Ordering auxiliary choices by sender identity gives a measurable selection. This representation uses a normalized survivor mean; the executed weights remain a_{ij} for every admitted flag sequence. □

Positive retention. For fixed $\kappa_{\mathrm{keep}} \in (0,1]$, let C be the full-retention honest row above. For an adapted retention set $\mathcal{T}_i$ as in Appendix B.3, full self-substitution of its processed slots also admits an honest stochastic row P, since every slot remains in I_i. Assigning the original self weight to i gives $p_{ii} \geq a_\star$. The retained update has

$$C_{\kappa_{\mathrm{keep}}} = \kappa_{\mathrm{keep}} C + (1-\kappa_{\mathrm{keep}}) P \geq \kappa_{\mathrm{keep}} C.$$

Thus the self floor stays $a_\star$, the other certified floors become $\kappa_{\mathrm{keep}} a_\star/2$, and all slots remain in I_i. Taking $\mathcal{T}_i = \mathcal{S}_i$ gives (8); the same argument allows $\mathcal{T}_i$ to be selected independently of the screen. In operator superscripts below, σ records both selections when they differ.

Lemma 2 (Rooted support and arbitrary products)**.** *Set $\mu_\star = \kappa_{\mathrm{keep}} a_\star/2$, $B_G = n-1$, and $\rho_G = 1-\mu_\star^{n-1}$. Every admissible scalar row sequence satisfies*

$$\mathrm{sp}(C_{t+B_G-1}\cdots C_t z) \leq \rho_G \,\mathrm{sp}(z). \tag{67}$$

Proof. Select self and exactly $d_i^- - 2H$ certified other supports in each row. If receiver i has $b_i \leq H$ Byzantine in-neighbors, the selection omits $(d_i^- - b_i) - (d_i^- - 2H) = 2H - b_i$ honest incoming edges. Including the b_i Byzantine edges, at most $2H$ original incoming edges are omitted. For two disjoint nonempty honest vertex sets, $(2H+1)$-robustness supplies a vertex in one set with at least $2H+1$ outside incoming neighbors. At least one such edge survives, so each realized honest support graph is 1-robust and therefore rooted.

Orient a support edge $h \to i$ when $c_{ih} \geq \mu_\star$ and write $\mathrm{Pred}_G(A) = A \cup \{h : \exists i \in A, h \to i\}$. Starting backward from distinct terminal rows i, j, set $A_0 = \{i\}$, $B_0 = \{j\}$. Self-loops make their predecessor sets monotone. Until the sets intersect, their union grows at every step: otherwise both would be in-closed, contradicting 1-robustness. Within $n-1$ steps a common ancestor therefore reaches both terminal rows, with self-loops padding shorter paths. Each path product is at least $\mu_\star^{n-1}$. For the product P,

$$\tau(P) = 1 - \min_{i,j} \sum_h \min(P_{ih}, P_{jh}) \leq 1 - \mu_\star^{n-1}. \tag{68}$$

The scalar span inequality follows from the Dobrushin coefficient. □

Lemma 3 (Exact fixed-probe reconstruction). *Let $f_{x,k}^\top$ be row k of F_x, and put*

$$\Gamma_{x,k}^F = (F_x^{-1}e_k)f_{x,k}^\top, \qquad A_{x,t}^{F,\sigma,+} = \sum_{k=1}^{d_x} C_{x,t}^{(k),\sigma,+} \otimes \Gamma_{x,k}^F. \tag{69}$$

This operator gives the simultaneous parameter update, preserves the honest consensus subspace, and is similar, after the probe-major permutation, to $\operatorname{diag}_k C_{x,t}^{(k),\sigma,+}$.

Proof. Writing $g_k = F_x^{-1}e_k$ gives $f_{x,k}^\top g_\ell = \delta_{k\ell}$, so $\Gamma_k^F\Gamma_\ell^F = \delta_{k\ell}\Gamma_k^F$ and $\sum_k \Gamma_k^F = I$. Reconstructing each scalar honest-row representation gives (69). Further, $F_x\Gamma_k^F F_x^{-1} = e_k e_k^\top$, which proves the similarity. Each scalar row preserves constants, hence the lifted operator preserves $\mathbf{1}\otimes z$. Fixed norms of F_x and F_x^{-1} transfer scalar span estimates into parameter disagreement estimates. Different probes may have different realized rows. □

For fixed κ_{keep}, the honest-hull, block-contraction and reconstruction properties above are the communication inputs to the critic and actor proofs. Lemmas 10, 11 and 13 use B_G, ρ_G; their constants inherit the retention dependence through ρ_G. Theorems 2 and 3 therefore retain their stated form. Appendix E.6 gives a complementary fixed-input net-shift decomposition and quantifies the coefficient-floor filter gain.

E.2 Shared Markov and stopping tools

Lemma 4 (Uniform Poisson solutions). *For $j \in \{1,2\}$ and every bounded, weighted-Lipschitz family $f_\vartheta \in \mathcal{B}_{W^j}$,*

$$\widehat{f}_\vartheta = \sum_{m\ge 0}(P_\vartheta^m - \Pi_\vartheta)f_\vartheta, \qquad (I-P_\vartheta)\widehat{f}_\vartheta = f_\vartheta - \rho_\vartheta f_\vartheta \tag{70}$$

converges uniformly in the weighted norm; the mean, solution and $P_\vartheta\widehat{f}_\vartheta$ are Lipschitz, including in additional compact field parameters.

Proof. Equation (37) gives convergence. With $Z_\vartheta = (I - P_\vartheta + \Pi_\vartheta)^{-1}$, $\Pi_{\vartheta'} - \Pi_\vartheta = \Pi_{\vartheta'}(P_{\vartheta'} - P_\vartheta)(Z_\vartheta - \Pi_\vartheta)$ and $Z_{\vartheta'} - Z_\vartheta = Z_{\vartheta'}[(P_{\vartheta'} - P_\vartheta) - (\Pi_{\vartheta'} - \Pi_\vartheta)]Z_\vartheta$ give Lipschitzness by the uniform operator bounds. □

Lemma 5 (Trace moments and Poisson solutions). *Put $a = \gamma\lambda$, $E_0 = 0$, $E_{t+1} = aE_t + \|\phi_t\|$, and $J_t = 1 + W_t^2 + E_tW_t$. Then $\|\mathrm{e}_t\| \le E_t$ and $\sup_t \mathbb{E}E_t^{2q_{\mathrm{cat}}} < \infty$. For a scalar state generator c and conditional edge multiplier $k_\vartheta(s) = \int k_\vartheta^{\mathrm{edge}}(z)Q_\vartheta(s,dz)$ from Appendix D.1, let $e_{t+1} = ae_t + c(s_t)$. The trace kernel is $\mathscr{Q}_\vartheta : (x,e) \mapsto (X', ae + c(x))$, where $X' \sim P_\vartheta(x,\cdot)$. With the Poisson notation of Lemma 4, define*

$$\begin{aligned} r_\vartheta &= (I - aP_\vartheta)^{-1}k_\vartheta, & m_\vartheta &= \rho_\vartheta(cr_\vartheta), \\ B_\vartheta &= \widehat{cr}_\vartheta, & H_\vartheta(x,e) &= ae\, r_\vartheta(x) + B_\vartheta(x). \end{aligned} \tag{71}$$

Under the stationary trace law the mean is m_ϑ, and

$$(I - \mathscr{Q}_\vartheta)H_\vartheta(x,e) = (ae + c(x))k_\vartheta(x) - m_\vartheta. \tag{72}$$

The means are bounded and Lipschitz; H_ϑ and its descriptor Lipschitz envelope are bounded by $C[W(x)^2 + |e|W(x)]$.

Proof. Minkowski gives $\|E_{t+1}\|_{L^q} \le \sum_{j=0}^t a^j\|\phi_{t-j}\|_{L^q} \le C_q/(1-a)$ for $q \le 2q_{\mathrm{cat}}$. The stationary series $e_t = \sum_{j\ge 0} a^j c(s_{t-1-j})$ converges in these norms; conditioning gives $\mathbb{E}[e_{t+1}k_\vartheta(s_t)] = m_\vartheta = \rho_\vartheta[c(I - aP_\vartheta)^{-1}k_\vartheta]$. For $R_a = (I - aP)^{-1}$, the resolvent identity $R_{a,\vartheta'} - R_{a,\vartheta} = aR_{a,\vartheta'}(P_{\vartheta'} - P_\vartheta)R_{a,\vartheta}$ and Lemma 4 for $cr \in \mathcal{B}_{W^2}$ give the bounds. Direct substitution yields $H - \mathscr{Q}H = ae(r - aPr) + (I - P)B - acPr = (ae + c)k - m$. □

Lemma 6 (Current-edge noise and weighted tails)**.** *Let $a_t \downarrow 0$, $\sum_t a_t^2 < \infty$, and let the field parameter p_t be predictable in a deterministic compact set. For a catalogued edge field, put $m(\vartheta, p) = \rho_\vartheta \bar{f}_{\vartheta,p}$. If its Poisson envelope satisfies*

$$\sum_t a_t \mathcal{W}_{t+1}[d(\vartheta_{t+1}, \vartheta_t) + \|p_{t+1} - p_t\|] < \infty \quad a.s., \tag{73}$$

then

$$\lim_{m\to\infty} \sup_{\ell \geq m} \left\| \sum_{t=m}^{\ell} a_t [f_{\vartheta_t,p_t}(Z_t) - m(\vartheta_t, p_t)] \right\| = 0 \quad a.s. \tag{74}$$

Use $\mathcal{W}_t = W(s_t)$ for edge fields. The same conclusion holds for $S_t = e_{t+1} k^{\mathrm{edge}}_{\vartheta_t}(Z_t)$, with stationary mean m_{ϑ_t} and $\mathcal{W}_t = J_t$. For a fixed field parameter, $d(\vartheta_{t+1}, \vartheta_t) \leq a_t c_t$, $c_t \to 0$ a.s., suffices for (73).

Proof. Let $h_{\vartheta,p}$ solve the physical-state Poisson equation for $\bar{f}_{\vartheta,p}$ (Lemma 4). Write $h_t = h_{\vartheta_t,p_t}$ and

$$\begin{aligned} M_{t+1} &= f_{\vartheta_t,p_t}(Z_t) - \bar{f}_{\vartheta_t,p_t}(s_t) + h_t(s_{t+1}) - (P_{\vartheta_t} h_t)(s_t), \\ R_t &= h_t(s_t), \qquad D_{t+1} = (h_{t+1} - h_t)(s_{t+1}). \end{aligned} \tag{75}$$

Both differences in M_{t+1} are $\mathcal{F}_t$-centered, and

$$f_{\vartheta_t,p_t}(Z_t) - m(\vartheta_t, p_t) = M_{t+1} + R_t - R_{t+1} + D_{t+1}. \tag{76}$$

The Lipschitz bound on D_{t+1} is the increment in (73) times $CW(s_{t+1})$.

For trace fields, use $k_\vartheta = Q_\vartheta k^{\mathrm{edge}}_\vartheta$ in Lemma 5. Put $Y_t = (s_t, e_t)$, $f_\vartheta(s, e) = (ae + c(s))k_\vartheta(s)$, and

$$\begin{aligned} M_{t+1} &= S_t - f_{\vartheta_t}(Y_t) + H_{\vartheta_t}(Y_{t+1}) - (\mathscr{Q}_{\vartheta_t} H_{\vartheta_t})(Y_t), \\ R_t &= H_{\vartheta_t}(Y_t), \qquad D_{t+1} = (H_{\vartheta_{t+1}} - H_{\vartheta_t})(Y_{t+1}). \end{aligned}$$

Predictability of e_{t+1} gives conditional centering. Equation (72) gives (76) with left side $S_t - m_{\vartheta_t}$ and $|D_{t+1}| \leq C J_{t+1} d(\vartheta_{t+1}, \vartheta_t)$. Equation (43), trace bounds and Hölder put S, J, R, M in $L^{q_{\mathrm{cat}}}$. Square-summable gains give martingale convergence and, by Borel–Cantelli, $a_t R_{t+1} \to 0$. Monotonicity and Tonelli give $\sum_t |a_t - a_{t-1}||R_t| < \infty$; (73) controls D. Abel summation proves (74) for both cases, including zero-start traces. For fixed p, the sufficient condition follows from eventual $c_t \leq 1$ and $\sum_t a_t^2 \mathcal{W}_{t+1} < \infty$ a.s. □

Lemma 7 (Stopped affine estimate)**.** *For a fixed numeric-reward family, let $\epsilon_{t+1}(z) = [\bar{h}_{t+1} - G_{t+1} z] - [\bar{h}(\vartheta_t) - \bar{G}(\vartheta_t) z]$. For fixed $C_0 > 1$, on disjoint critic-time blocks $[t_j, t_{j+1})$ of mass $T + o(1)$, put $r_j = 1 + \|\bar{x}_{t_j}\|$ and $\tau_j = t_{j+1} \wedge \inf\{t \geq t_j : \|\bar{x}_t\| > C_0 r_j\}$. Under (45),*

$$\max_{t_j \leq k \leq \tau_j} \left\| r_j^{-1} \sum_{t=t_j}^{k-1} \alpha_t \epsilon_{t+1}(\bar{x}_t) \right\| \longrightarrow 0 \quad a.s. \tag{77}$$

The same estimate holds for the multiplicative part alone, leaving an arbitrary compact-Preference forcing uncentered.

Proof. Apply the decompositions to h, G, with $\mathcal{W} = W$ for ω and J for v. The catalogue, trace and deterministic-filter moments give, by Hölder,

$$\sup_t \mathbb{E}[J_{t+1}^2 + J_{t+1}(U_{t+1}^+ + K_{t+1}^+ + \mathsf{Z}_t) + g_{t+1}^+ \mathsf{Z}_t + J_{t+1} g_{t+1}^+ \mathsf{Z}_t] < \infty. \tag{78}$$

The analogous J_t, W_{t+1} bounds complement (44). For $z_t = \bar{x}_t$,

$$\begin{aligned} \epsilon_{t+1}(z_t) = M_{t+1}^h - M_{t+1}^G z_t + (R_t^h - R_t^G z_t) - (R_{t+1}^h - R_{t+1}^G z_{t+1}) \\ + D_{t+1}^h - D_{t+1}^G z_t - R_{t+1}^G (z_{t+1} - z_t). \end{aligned} \tag{79}$$

The center recursion, norm equivalence and (93) give, for $t_j \leq t < \tau_j$,

$$\frac{\|\bar{x}_{t+1} - \bar{x}_t\|}{r_j} \leq C C_0 \alpha_t g_{t+1} + \alpha_t K_{t+1}^{\mathcal{K}} + C_\Delta[\alpha_t \mathsf{Z}_t + \alpha_t U_{t+1}^{\mathcal{K}} + \alpha_t^2 g_{t+1} \mathsf{Z}_t] =: b_t(C_0). \tag{80}$$

Primitive moments and Lemma 10 give $\alpha_t g_{t+1}, \alpha_t K^{\mathcal{K}}_{t+1}, \alpha_t U^{\mathcal{K}}_{t+1}, \alpha_t \mathsf{Z}_t \to 0$. The product bounds give $\sum_t \alpha_t^2 g_{t+1} \mathsf{Z}_t < \infty$, hence

$$\beta_j := \sup_{t \ge t_j} b_t(C_0) \to 0, \qquad \|\bar{x}_{\tau_j}\|/r_j \le C_0 + \beta_j. \tag{81}$$

The scale r_j is $\mathcal{F}_{t_j}$-measurable and $\mathbf{1}_{\{t<\tau_j\}}\bar{x}_t/r_j$ is predictable, bounded by C_0. Doob's block bound $C\sum_{t=t_j}^{t_{j+1}-1}\alpha_t^2$, summed over disjoint blocks, and Borel–Cantelli eliminate the normalized martingale tails. The coboundary endpoints vanish since $\alpha_{\tau_j-1}\mathcal{W}_{\tau_j}(1 + C_0 + \beta_j) \to 0$; gain variation is summable. The remaining weighted terms are bounded by

$$C\alpha_t \mathcal{W}_{t+1}\left[(1+C_0)d(\vartheta_{t+1},\vartheta_t) + \|\bar{x}_{t+1} - \bar{x}_t\|/r_j\right].$$

Slow transport and (80) reduce these to the summable products (44) and (78). Omitting the h-terms proves the multiplicative claim, with $K^{\mathcal{K}}_{t+1}$ bounding the uncentered forcing. □

E.3 Preference noise and curvature

Lemma 8 (Boundary-chain score noise). *For any bounded Lipschitz $v(\eta)$, the conditional score and kernel of Appendix D.3 have a uniformly W_B-bounded, weighted-Lipschitz centered scalar Poisson solution in (ϑ, η). Under (50),*

$$\sum_\ell a_{\eta,\ell}\langle v(\eta_{i,\ell}), g_i(\eta_{i,\ell}, Z_\ell^{\mathrm{blk}}) - \bar{g}_{i,\vartheta_{\tau_\ell}}(\eta_{i,\ell})\rangle \quad \textit{converges a.s.}, \tag{82}$$

where $\bar{g}_{i,\vartheta} = \mathbb{E}_{\nu_\vartheta^{\mathrm{blk}}} g_i$. This includes $v_r(e) = 2(1 - r/\|e\|)_+ e$, with $v_r(0) = 0$, $e = \eta - \eta_i^{\mathrm{ref}}$ and $r \ge 0$. Misspecified labels require Appendix F's conditions.

Proof. Apply Lemma 4 to $v(\eta)^\top f_{i,\vartheta,\eta}$ before evaluating the adapted parameter. At $X_\ell = s_{\tau_\ell}$, write $q_\ell = v(\eta_{i,\ell})^\top g_i(\eta_{i,\ell}, Z_\ell^{\mathrm{blk}})$, $f_\ell = v(\eta_{i,\ell})^\top f_{i,\vartheta_{\tau_\ell},\eta_{i,\ell}}$, and $\bar{q}_\ell = \rho_{\vartheta_{\tau_\ell}} f_\ell$. Let $(I - K_{\vartheta_{\tau_\ell}})h_\ell = f_\ell - \bar{q}_\ell$; $\widetilde{f}_\ell, \widetilde{K}_\ell$ denote the actual conditional scalar score and boundary law. Exactly,

$$\begin{aligned} M_{\ell+1} &= q_\ell - \widetilde{f}_\ell(X_\ell) + h_\ell(X_{\ell+1}) - (\widetilde{K}_\ell h_\ell)(X_\ell), \\ A_\ell &= \widetilde{f}_\ell(X_\ell) - f_\ell(X_\ell) + [(\widetilde{K}_\ell - K_{\vartheta_{\tau_\ell}})h_\ell](X_\ell), \\ R_\ell &= h_\ell(X_\ell), \qquad D_{\ell+1} = (h_{\ell+1} - h_\ell)(X_{\ell+1}). \end{aligned} \tag{83}$$

Both differences in $M_{\ell+1}$ are $\mathcal{F}_{\tau_\ell}$-centered; the marked-block interface gives

$$q_\ell - \bar{q}_\ell = M_{\ell+1} + R_\ell - R_{\ell+1} + D_{\ell+1} + A_\ell, \qquad |A_\ell| \le C\delta_\ell^{\mathrm{blk}} W_B(X_\ell). \tag{84}$$

Square-summable gains and score/bare-W_B second moments control martingales; Abel summation controls boundaries and gain variation. Projection gives $\|\eta_{i,\ell+1} - \eta_{i,\ell}\| \le a_{\eta,\ell}\|g_i(\eta_{i,\ell}, Z_\ell^{\mathrm{blk}})\|$; adjacent-$W_B$ score-product moments make its Poisson remainder summable. Condition (50) makes descriptor and actual/frozen transport summable, proving (82). □

Lemma 9 (Bounded excitation gives curvature). *Uniformly in ϑ and $\eta \in \mathcal{K}_i$,*

$$\nabla^2 \mathcal{L}_{i,\vartheta}(\eta) \succeq \lambda_{\mathrm{pref},i} I_{\mathsf{E}}, \qquad \lambda_{\mathrm{pref},i} = \beta_{\mathrm{p}}^2 m_{B,i}\lambda_D > 0, \tag{85}$$

where $R_{\eta,i} = \sup_{\eta\in\mathcal{K}_i}\|\eta\|$, and $m_{B,i} = \min_{|u|\le\beta_{\mathrm{p}}B_D R_{\eta,i}} \sigma'(u) > 0$. For $\eta_i^0, \eta_i^0 + e \in \mathcal{K}_i$ with $\|\bar{g}_{i,\vartheta}(\eta_i^0)\| \le \epsilon_{i,\mathrm{score}}$,

$$\langle e, \bar{g}_{i,\vartheta}(\eta_i^0 + e)\rangle \ge \lambda_{\mathrm{pref},i}\|e\|^2 - \epsilon_{i,\mathrm{score}}\|e\|. \tag{86}$$

Proof. The Hessian is $\beta_{\mathrm{p}}^2 \mathbb{E}_{\nu_\vartheta^{\mathrm{blk}}}[\sigma'(\beta_{\mathrm{p}} D^\top \eta) DD^\top]$. On $\{\|D\| \le B_D\}$, $\sigma' \ge m_{B,i}$. Discarding the positive remainder and using (48) proves (85). Integration along the segment from η_i^0 to $\eta_i^0 + e$ and Cauchy–Schwarz give (86). □

E.4 Disagreement filters and critic-time estimates

Use the one-channel notation and actual defect of Appendix C.2.

For projected reward parameters, define

$$\begin{aligned} H_{t+1} &= \sup_{\eta\in\prod_i \mathcal{K}_i} \max_k \operatorname{sp}[(I_n\otimes F)\mathbf{h}_{t+1}(\eta)]^{(k)}, \quad U_{t+1} \ge 1 + H_{t+1}, \\ K_{t+1} &\ge \sup_{\eta\in\prod_i \mathcal{K}_i} \|n^{-1}\textstyle\sum_i h_{i,t+1}(\eta_i)\|, \qquad\qquad g_{t+1} \ge \|FG_{t+1}F^{-1}\|_\infty. \end{aligned} \tag{87}$$

Use equality for $x=\omega$. For $x=v$, the positive trace of Lemma 5 gives

$$\begin{aligned} U^+_{t+1} &= 1 + c_H E_{t+1}\|\chi_t\|, \quad K^+_{t+1} = R_\eta E_{t+1}\|\chi_t\|, \\ A^+_{t+1} &= E_{t+1}\|\chi_t\|, \qquad\qquad g^+_{t+1} = c_g E_{t+1}(\|\phi_t\| + \gamma\|\phi_{t+1}\|), \end{aligned} \tag{88}$$

where $R_\eta = \max_i \sup_{\eta\in\mathcal{K}_i}\|\eta\|$, $c_H = 2R_\eta\|F_v\|_{2\to\infty}$, and $c_g = \sqrt{d_v}\|F_v\|_{2\to\infty}\|F_v^{-1}\|_2$. Set $(U,K,g) = (U^+,K^+,g^+)$; then $\|\mathbf{e}_{t+1}\chi_t^\top\| \le A^+_{t+1}$. Lemma 6 controls the nonconstant parts of U^+, K^+, A^+ with generator $\|\phi\|$ and multiplier $\|\chi\|$.

Cancellation of the common G_{t+1} and network contraction give

$$\begin{aligned} S_{t+1} &\le (1+\alpha_t g_{t+1})S_t + \alpha_t U_{t+1}, \\ S_{t+B_G} &\le \rho_G S_t + \sum_{s=t}^{t+B_G-1} \alpha_s(H_{s+1} + g_{s+1}S_s). \end{aligned} \tag{89}$$

Lemma 10 (Disagreement filter). *Choose $\rho_G < \rho_L < 1$ and $\rho_L^{1/B_G} < \rho_2 < 1$. The adapted filter $\mathsf{Z}_0 = 1$, $\mathsf{Z}_t = U_t + \rho_2\mathsf{Z}_{t-1}$ satisfies, for a deterministic $C_Q < \infty$,*

$$S_t \le C_Q\alpha_{t-1}\mathsf{Z}_t \quad \textit{eventually}, \qquad \sup_t \mathbb{E}\mathsf{Z}_t^2 < \infty, \qquad \alpha_t\mathsf{Z}_t \to 0, \quad S_t \to 0 \quad \textit{a.s.} \tag{90}$$

Proof. Put $A_t^B = \sum_{s=t}^{t+B_G-1}\alpha_s g_{s+1}$ and $B_t^B = \sum_{s=t}^{t+B_G-1}\alpha_s U_{s+1}$. Discrete Gronwall bounds the within-block span by $e^{A_t^B}(S_t + B_t^B)$. Inserting this in (89) yields $S_{t+B_G} \le (\rho_G + A_t^B e^{A_t^B})S_t + (1 + A_t^B e^{A_t^B})B_t^B$. The catalogue moments and Borel–Cantelli imply $A_t^B \to 0$. With $Q_t = S_t/\alpha_{t-1}$, the gain-ratio condition then gives eventually

$$Q_{t+B_G} \le \rho_L Q_t + 4\sum_{s=t}^{t+B_G-1} U_{s+1}. \tag{91}$$

Unrolling from the finite pathwise starting time N on the B_G residue classes gives, for a finite random R,

$$Q_t \le R\rho_L^{\lfloor (t-N)/B_G\rfloor} + 4\rho_2^{-(B_G-1)}\mathsf{Z}_t, \qquad C_Q = 1 + 4\rho_2^{-(B_G-1)}. \tag{92}$$

The convolution bound uses $\rho_L^j \le \rho_2^{jB_G}$ and at most $B_G - 1$ remaining positions per block. Since $\mathsf{Z}_t \ge 1$, the transient is absorbed pathwise. Minkowski gives $\sup_t\|\mathsf{Z}_t\|_{L^q} \le 1 + \sup_t\|U_t\|_{L^q}/(1-\rho_2)$ for each available catalogue order q. Finally, $\sum_t \Pr(\alpha_t\mathsf{Z}_t > \varepsilon) \le \varepsilon^{-2}\sup_t\mathbb{E}\mathsf{Z}_t^2\sum_t\alpha_t^2 < \infty$. □

Lemma 11 (Aggregation defect: envelope and intensity). *For $c_F = \sqrt{d}\|F^{-1}\|_2$ and $C_\Delta = c_F\max\{2C_Q, 1\}$, eventually*

$$\|\Delta_t\| \le C_\Delta(\mathsf{Z}_t + U_{t+1} + \alpha_t g_{t+1}\mathsf{Z}_t). \tag{93}$$

If $d(\vartheta_{t+1},\vartheta_t) = o(\alpha_t)$, every fixed $T > 0$ satisfies

$$\begin{aligned} \limsup_{m\to\infty} \sup_{\ell:\,\sum_{t=m}^{\ell}\alpha_t\le T} \sum_{t=m}^{\ell}\alpha_t\|\Delta_t\| &\le T\overline{D}_\Delta, \\ \overline{D}_\Delta = C_{\mathrm{net}}\mu_U, \qquad C_{\mathrm{net}} &= C_\Delta\left(1 + \frac{1}{1-\rho_2}\right), \end{aligned} \tag{94}$$

where $\mu_U = \sup_\vartheta \mathbb{E}_\vartheta U < \infty$; for $x = v$, this is $\sup_\vartheta \mathbb{E}_{\nu_{\vartheta,\lambda}}U^+$. Also $\alpha_t\|\Delta_t\| \to 0$ almost surely.

Proof. Both the pre- and post-aggregation means of each probe lie in its honest proposal span. Reconstruction thus gives $\|\Delta_t\| \le c_F[S_t/\alpha_t + H_{t+1} + g_{t+1}S_t]$. Apply Lemma 10 and $\alpha_{t-1}/\alpha_t \le 2$ eventually.

Lemma 6, using its trace version for $x = v$, bounds the α-weighted U sums by $\mu_U(T+o(1))+o(1)$. Eventual monotonicity and the filter identity give

$$(1-\rho_2)\sum_{t=m}^{\ell}\alpha_t \mathsf{Z}_t \le \sum_{t=m}^{\ell}\alpha_t U_t + \rho_2\alpha_m \mathsf{Z}_{m-1}. \tag{95}$$

The boundary vanishes; bounded gain variation and moments control the shift from U_{t+1} to U_t. Also $\sum_t \alpha_t^2 g_{t+1}\mathsf{Z}_t < \infty$ by (44) and (78). These estimates prove (94) and the one-step claim. □

For the mean recursion, write $\delta = \delta_x$. For $\mathcal{B}_t = I - \alpha_t \bar{G}(\vartheta_t)$ and $\Phi(t,s) = \mathcal{B}_{t-1}\cdots\mathcal{B}_s$, coercivity implies eventually

$$\|\Phi(t,s)\| \le e^{-(\delta/2)\sum_{k=s}^{t-1}\alpha_k}. \tag{96}$$

Indeed $\|I - \alpha_t\bar{G}\|^2 \le 1 - 2\delta\alpha_t + M_G^2\alpha_t^2$ is eventually at most $e^{-2c\alpha_t}$ for each $0 < c < \delta$. Multiplication gives (96), also with c in place of $\delta/2$. Abel summation, with $N_\ell = \sum_{k=m}^{\ell}\alpha_k\zeta_{k+1}$, gives

$$\sum_{k=m}^{t-1}\Phi(t,k+1)\alpha_k\zeta_{k+1} = N_{t-1} + \sum_{k=m}^{t-2}[\Phi(t,k+1) - \Phi(t,k+2)]N_k. \tag{97}$$

Thus transition weighting increases a maximal partial-sum bound by at most $C_\Phi(T) = 1 + M_G(T+1)$ on a critic-time T-window.

For large m, let $\tau_x(m,T) = \max\{k \ge m : \sum_{t=m}^{k}\alpha_{x,t} \le T\}$.

Lemma 12 (Reward-residual windows)**.** *Let $A^v_{i,t+1} = \mathbf{e}_{t+1}\chi_t^\top$, $A^\omega_{i,t+1} = \psi_t\chi_t^\top$, $\mathcal{A}_v = A^+$ and $\mathcal{A}_\omega = \|\psi_t\chi_t^\top\|$. With $\mu_{A,x} = \sup_\vartheta \mathbb{E}_\vartheta\mathcal{A}_x$, the catalogue and slow-path conditions give, for every fixed $T > 0$, almost surely,*

$$\limsup_{m\to\infty}\sup_{m\le k\le\tau_x(m,T)}\sum_{t=m}^{k}\alpha_{x,t}\frac{1}{n}\sum_i\|A^x_{i,t+1}\| \le \mu_{A,x}T. \tag{98}$$

Proof. Apply Lemma 6 to $\mathcal{A}_x$, using the positive trace for v. Centered tails vanish and the window mass tends to T. □

E.5 An optional worst-case certificate from utility heterogeneity

For $x \in \{v,\omega\}$, the shared-sample proposal is

$$\widetilde{x}_{i,t+1} = (I - \alpha_t G_{t+1})x_{i,t} + \alpha_t\zeta_{t+1}\chi(Z_t)^\top\eta_{i,\ell(t)}, \tag{99}$$

where $\zeta = \psi(Z_t)$ for the reward critic and $\zeta = \mathbf{e}_{t+1}$ for the value critic. Define $k(Z) = \mathrm{osc}_{i\in\mathcal{N}^+} r_i^\star(Z)$, with $\mathrm{osc}_i\, u_i = \max_i u_i - \min_i u_i$, and the positive trace $E_0 = 0$, $E_{t+1} = \gamma\lambda E_t + \|\phi(s_t)\|$. Set

$$\begin{aligned} H^\star_{\omega,t+1} &= \|F_\omega\psi(Z_t)\|_\infty k(Z_t),\\ H^+_{v,t+1} &= \|F_v\|_{2\to\infty}E_{t+1}k(Z_t),\\ \mu_{H,\omega} &= \sup_\vartheta \mathbb{E}_\vartheta H^\star_\omega, \qquad \mu_{H,v} := \mu^+_{H,v} = \sup_\vartheta \mathbb{E}_\vartheta H^+_v. \end{aligned} \tag{100}$$

Here $\mu_{H,x}$ measures honest utility-span intensity. The value expectation includes the stationary positive trace. Require conditional weighted bounds, descriptor-Lipschitzness and catalogue moments for

$$k(Z), \quad \|F_\omega\psi(Z)\|_\infty k(Z), \quad \|F_\omega\psi(Z)\|_\infty\|\chi(Z)\|, \quad \|\chi(Z)\|, \tag{101}$$

The value fields have generator $\|\phi\|$ and multipliers $k, \|\chi\|$; require the raw-multiplier moments and conditional regularity of (43) for k as well. The conditions on k and its products belong to this optional certificate, not to Assumption 1. The general critic proof uses the compact reward-family envelopes of (87) instead. By the triangle inequality, an absolute-window certificate below is also a valid net-window certificate in Section 4.2.

Lemma 13 (True-heterogeneity intensity). *Suppose Assumptions 1–5 and (101) hold, and the calibrated local estimates converge to $\eta_i^\star$. Let B_G, ρ_G be any valid scalar honest-row contraction certificate from Appendix E.1. Choose $\rho_G < \rho_L < 1$ and $\rho_L^{1/B_G} < \rho_2 < 1$, and put*

$$C_H = 4\rho_2^{-(B_G-1)}, \qquad c_{F,x} = \sqrt{d_x}\|F_x^{-1}\|_2, \qquad \mathfrak{D}_x^{\mathrm{het}} = c_{F,x}\left(1 + \frac{2C_H}{1-\rho_2}\right)\mu_{H,x}. \tag{102}$$

For every fixed $T > 0$, almost surely,

$$\limsup_{m\to\infty} \sup_{\ell\geq m:\, \sum_{t=m}^{\ell}\alpha_t \leq T} \sum_{t=m}^{\ell} \alpha_t\|\Delta_{x,t}\| \leq T\mathfrak{D}_x^{\mathrm{het}}, \qquad \alpha_t\|\Delta_{x,t}\| \longrightarrow 0. \tag{103}$$

The intensity is zero when the true honest utilities coincide pointwise.

Proof. Suppress the channel index. For the honest probe span S_t, use the actual learned forcing

$$H_{t+1} = \|F\zeta_{t+1}\|_\infty \operatorname{osc}_i[\chi(Z_t)^\top \eta_{i,\ell(t)}], \qquad g_{t+1} = \|FG_{t+1}F^{-1}\|_\infty.$$

Common G_{t+1} and scalar nonexpansion give $S_{t+1} \leq (1+\alpha_t g_{t+1})S_t + \alpha_t H_{t+1}$. The block argument of (91), using this realized forcing, gives

$$Q_{t+B_G} \leq \rho_L Q_t + 4\sum_{s=t}^{t+B_G-1} H_{s+1} \quad \text{eventually}, \qquad Q_t = S_t/\alpha_{t-1}. \tag{104}$$

Since $\alpha_t g_{t+1} \to 0$, for $\mathsf{Z}_0 = 0$, $\mathsf{Z}_t = H_t + \rho_2 \mathsf{Z}_{t-1}$, unrolling the B_G residue classes gives

$$Q_t \leq r_t + C_H \mathsf{Z}_t, \qquad r_t = R\rho_L^{\lfloor (t-N)/B_G \rfloor} \longrightarrow 0, \tag{105}$$

for pathwise finite R, N, since $\rho_L^j \leq \rho_2^{jB_G + r - (B_G-1)}$ for $0 \leq r < B_G$. No moments of R, N are used. Compact-forcing domination gives $\alpha_t \mathsf{Z}_t \to 0$ and $\sum_t \alpha_t^2 g_{t+1}\mathsf{Z}_t < \infty$ by the catalogue estimates.

Lemma 11 and $\alpha_{t-1}/\alpha_t \leq 2$ eventually give

$$\|\Delta_t\| \leq c_F\big[H_{t+1} + 2C_H\mathsf{Z}_t + 2r_t + 2C_H\alpha_t g_{t+1}\mathsf{Z}_t + 2\alpha_t g_{t+1} r_t\big]. \tag{106}$$

The last three terms vanish on weighted windows by summability and $r_t, \alpha_t g_{t+1} \to 0$.

Let $e_m = \sup_{t\geq m} \max_i \|\eta_{i,\ell(t)} - \eta_i^\star\| \to 0$ almost surely. The oscillation inequality gives, on $t \geq m$,

$$\begin{aligned} H_{\omega,t+1} &\leq H^\star_{\omega,t+1} + 2e_m\|F_\omega\psi(Z_t)\|_\infty\|\chi(Z_t)\|, \\ H_{v,t+1} &\leq H^+_{v,t+1} + 2e_m\|F_v\|_{2\to\infty}E_{t+1}\|\chi(Z_t)\|. \end{aligned} \tag{107}$$

The Poisson estimates for (101) bound the multiplier windows and center the latent fields, so

$$\limsup_{m\to\infty} \sup_{\ell:\, \sum_{t=m}^{\ell}\alpha_t \leq T} \sum_{t=m}^{\ell} \alpha_t H_{t+1} \leq T\mu_H. \tag{108}$$

The filter identity (95), with H replacing U, bounds the Z windows by $T\mu_H/(1-\rho_2)$. Its boundary vanishes, and the shift from H_{t+1} to H_t costs a vanishing endpoint and a series of absolute expectation at most $C\sum_t|\alpha_t - \alpha_{t-1}| < \infty$. Substitution in (106) proves the intensity bound. Compact moments give $\alpha_t H_{t+1} \to 0$, proving the one-step claim with the filter and transient bounds. □

For $r_i^\star = r_{\mathrm{common}} + \tau d_i$ and fixed difference envelope, graph, features, policy family and (γ, λ), the certificate is $O(\tau)$ under exact preference identification.

E.6 Fixed-input decomposition of retention and screening effects

Theorem 2 links critic accuracy to the realized net shift (10). Holding the raw messages and first-stage screen fixed, we split that shift into residual, missed Byzantine, targeted Byzantine and targeted honest parts, isolating retention level and target selection.

Record, slots and offsets. Fix a critic channel x and write $F = F_x$, $\alpha_t = \alpha_{x,t}$ and $c_F = \sqrt{d_x}\|F^{-1}\|_2$. A record $\Xi = (\Xi_t)_{t\ge 0}$ lists the step sizes and, for each round and honest receiver i, the complete joint-critic raw row, the self proposal $\mathbf{z}_{i,t} = F\widetilde{x}_{i,t+1}$, the raw intervals, the base weights, the first-stage screen $\mathcal{S}_{i,t}$ and the retention targets $\mathcal{T}_{i,t}$. Let $\mathbf{u}_{ij,t}$ be sender j's full-retention TPSC-processed probe vector (Appendix B.3), determined by Ξ_t. Retention $\kappa \in (0,1]$ on Ξ gives the slots

$$\mathbf{u}_{ij,t}^{(\kappa)} = \mathbf{z}_{i,t} + \big[1-(1-\kappa)\mathbf{1}\{j\in\mathcal{T}_{i,t}\}\big](\mathbf{u}_{ij,t} - \mathbf{z}_{i,t}), \tag{109}$$

which reduce to (8) when $\mathcal{T}_{i,t} = \mathcal{S}_{i,t}$. Write $\mathrm{In}^{\pm}(i) = \mathrm{In}(i)\cap\mathcal{N}^{\pm}$ and $o_{ij,t} = F^{-1}(\mathbf{u}_{ij,t} - \mathbf{z}_{i,t})$; sums over i range over $\mathcal{N}^{+}$. Define the net-shift components

$$\begin{aligned}
\Delta_t^{\mathrm{res}} &= \frac{1}{n\alpha_t}\sum_i \sum_{h\in\mathrm{In}^+(i)} a_{ih}o_{ih,t}, & \Delta_t^{\mathrm{hon}} &= \frac{1}{n\alpha_t}\sum_i \sum_{h\in\mathrm{In}^+(i)\cap\mathcal{T}_{i,t}} a_{ih}o_{ih,t},\\
\Delta_t^{\mathrm{miss}} &= \frac{1}{n\alpha_t}\sum_i \sum_{b\in\mathrm{In}^-(i)\setminus\mathcal{T}_{i,t}} a_{ib}o_{ib,t}, & \Delta_t^{\mathrm{hit}} &= \frac{1}{n\alpha_t}\sum_i \sum_{b\in\mathrm{In}^-(i)\cap\mathcal{T}_{i,t}} a_{ib}o_{ib,t},
\end{aligned} \tag{110}$$

and the exposures $\mathsf{e}_t^{\mathrm{hon}}$, $\mathsf{e}_t^{\mathrm{miss}}$ and $\mathsf{e}_t^{\mathrm{hit}}$ obtained by replacing each o with $\|o\|$ in the last three components. The residual Δ_t^{res} replaces every Byzantine processed slot by self but keeps the intervals and screen of Ξ, hence their influence on honest slots. For a finite window W of rounds with $|W|_\alpha = \sum_{t\in W}\alpha_t > 0$, put

$$\bar{d}_W^{\mathrm{res}} = \frac{1}{|W|_\alpha}\Big\|\sum_{t\in W}\alpha_t\Delta_t^{\mathrm{res}}\Big\|, \qquad \bar{\mathsf{e}}_W^{\bullet} = \frac{1}{|W|_\alpha}\sum_{t\in W}\alpha_t\mathsf{e}_t^{\bullet}, \quad \bullet\in\{\mathrm{hon},\mathrm{miss},\mathrm{hit}\}. \tag{111}$$

Proposition 2 (Fixed-input decomposition and retention envelope)**.** *Let $\Delta_t(\kappa \mid \Xi)$ be the net shift (10) produced by the slots (109). For every $\kappa\in(0,1]$ and every finite window W,*

$$\Delta_t(\kappa\mid\Xi) = \Delta_t^{\mathrm{res}} + \Delta_t^{\mathrm{miss}} + \kappa\Delta_t^{\mathrm{hit}} - (1-\kappa)\Delta_t^{\mathrm{hon}}, \tag{112}$$

$$\frac{1}{|W|_\alpha}\Big\|\sum_{t\in W}\alpha_t\Delta_t(\kappa\mid\Xi)\Big\| \le \mathscr{E}_W(\kappa\mid\Xi) := \bar{d}_W^{\mathrm{res}} + \bar{\mathsf{e}}_W^{\mathrm{miss}} + \kappa\bar{\mathsf{e}}_W^{\mathrm{hit}} + (1-\kappa)\bar{\mathsf{e}}_W^{\mathrm{hon}}, \tag{113}$$

$$\mathscr{E}_W(1\mid\Xi) - \mathscr{E}_W(\kappa\mid\Xi) = (1-\kappa)\mathsf{M}_W, \qquad \mathsf{M}_W = \bar{\mathsf{e}}_W^{\mathrm{hit}} - \bar{\mathsf{e}}_W^{\mathrm{hon}}. \tag{114}$$

If Ξ is the record of a run with retention κ_{keep}, then $\Delta_t(\kappa_{\mathrm{keep}}\mid\Xi) = \Delta_{x,t}$.

Proof. Unit row sums and $\mathbf{u}_{ii,t} = \mathbf{z}_{i,t}$ give $x_{i,t+1} - \widetilde{x}_{i,t+1} = \sum_{j\in\mathrm{In}(i)} a_{ij}F^{-1}(\mathbf{u}_{ij,t}^{(\kappa)} - \mathbf{z}_{i,t})$. By (109), a targeted slot keeps the fraction κ of its offset and every other slot keeps all of it. Splitting $\mathrm{In}(i)$ into honest and Byzantine senders, targeted or not, gives (112). Summing it with weights α_t over W before taking any norm, and applying the triangle inequality only to the three selection sums, gives (113); subtraction gives (114). With $\kappa = \kappa_{\mathrm{keep}}$ the slots are the executed ones, so the last claim is (10). □

The residual keeps cancellation across receivers and time, whereas exposures are computed slot by slot. For $0 < \kappa < 1$ the envelope tightens exactly when $\mathsf{M}_W > 0$, for arbitrary Byzantine in-neighbor sets and message coordinates; for a wholly in-interval Byzantine message, $o_{ib,t}$ is its raw offset. The envelope and the realized shift need not move together. On the complete graph with four honest nodes and one Byzantine node b ($H = 1$), weights $1/5$, $F = 1$ and one round with $\alpha_t = 1$, let the honest proposals be $(0,-2,-1,2)$ and let b send $(-2,-1,2,2)$, with $\mathcal{T}_i = \mathcal{S}_i = \{b\}$ at every receiver. For any tie-breaking the raw intervals are $[-2,0],[-2,0],[-1,2],[-1,2]$, each containing the Byzantine value, and $\Delta(\kappa\mid\Xi) = (\kappa-1)/10$ while $\mathscr{E}(\kappa\mid\Xi) = (1+3\kappa)/10$: lower retention tightens the envelope but removes cancellation from the realized shift.

Corollary 1 (Interface with Theorem 2)**.** *Under Assumptions 1–5, consider a run with fixed retention κ_{keep} and record Ξ. Let $\mathfrak{D}_x^{\mathrm{res}}$, $\mathfrak{E}_x^{\mathrm{miss}}$, $\mathfrak{E}_x^{\mathrm{hit}}$ and $\mathfrak{E}_x^{\mathrm{hon}}$ be deterministic bounds, in the window sense of Appendix C.2, on the net residual windows $\|\sum_t\alpha_t\Delta_t^{\mathrm{res}}\|$ and the exposure windows $\sum_t\alpha_t\mathsf{e}_t^{\bullet}$ of this run. Then*

$$\mathfrak{D}_x^{\mathrm{ret}}(\kappa_{\mathrm{keep}}) = \mathfrak{D}_x^{\mathrm{res}} + \mathfrak{E}_x^{\mathrm{miss}} + \kappa_{\mathrm{keep}}\mathfrak{E}_x^{\mathrm{hit}} + (1-\kappa_{\mathrm{keep}})\mathfrak{E}_x^{\mathrm{hon}} \tag{115}$$

is an admissible $\mathfrak{D}_x$, and (11) holds with it. Such bounds exist: each may be taken as $\overline{D}_\Delta$ in (94).

Proof. Summing (112) at $\kappa = \kappa_{\mathrm{keep}}$ over a window and applying the triangle inequality to the three selection sums bounds the net window of $\Delta_{x,t}$ by $T\mathfrak{D}_x^{\mathrm{ret}} + o(1)$, so Theorem 2 applies. For existence, every coordinate of $\mathbf{u}_{ij,t}$ and $\mathbf{z}_{i,t}$ lies in its raw interval, hence in the honest post-innovation range (Lemma 1). With $\widetilde{S}_t$ the honest post-innovation probe span, $\|o_{ij,t}\| \le c_F \widetilde{S}_t$, and unit row sums give

$$\|\Delta_t^{\mathrm{res}}\|,\ \mathsf{e}_t^{\mathrm{miss}},\ \mathsf{e}_t^{\mathrm{hit}},\ \mathsf{e}_t^{\mathrm{hon}} \le c_F \widetilde{S}_t/\alpha_t \le c_F\big[S_t/\alpha_t + H_{t+1} + g_{t+1}S_t\big],$$

the bound used in the proof of Lemma 11. That proof then gives (93) and (94) for each component, together with the one-step bound that controls endpoint overshoot. □

Over a finite collection of fixed-retention runs, componentwise maxima give common finite bounds in (115).

Lemma 14 (Retention dependence of the generic filter gain)**.** *For Lemma 2's certificate $B_G = n-1$, $\rho_G = 1 - \mu_\star^{B_G}$, every choice $\rho_G < \rho_L < 1$, $\rho_L^{1/B_G} < \rho_2 < 1$ in Lemma 10 satisfies*

$$\frac{1}{1-\rho_2} > B_G\, \rho_G^{1-1/B_G}\, \mu_\star^{-B_G}, \qquad \mu_\star = \frac{\kappa_{\mathrm{keep}} a_\star}{2}. \tag{116}$$

Hence, for fixed $n \ge 2$, $a_\star$ and probes, the generic constant C_{net} in (94) and the multiplier $c_{F,x}[1 + 2C_H/(1-\rho_2)]$ in (102), built from this certificate, have lower bounds of order $\kappa_{\mathrm{keep}}^{-(n-1)}$ as $\kappa_{\mathrm{keep}} \downarrow 0$.

Proof. Since $\rho_2 > \rho_L^{1/B_G} > \rho_G^{1/B_G}$, the mean-value theorem for $y \mapsto y^{1/B_G}$ gives $1 - \rho_2 < (1-\rho_G)/(B_G \rho_G^{1-1/B_G})$, with $1 - \rho_G = \mu_\star^{B_G}$. Both constants contain $(1-\rho_2)^{-1}$ times factors bounded below uniformly in κ_{keep}. In particular, $C_Q = 1 + 4\rho_2^{-(B_G-1)} \ge 5$ is nonincreasing in ρ_2 and tends to 5 as $\rho_2 \uparrow 1$. □

These bounds are finite at each fixed positive retention level but cannot favor reduced retention; Proposition 2 compares retention levels on fixed inputs instead.

Proposition 3 (Random retention targets on a fixed record)**.** *Fix a record Ξ° of raw rows and first-stage screens, with retention targets left unspecified. Conditional on this record, draw fresh $J_{i,t}$ uniformly on $\mathrm{In}(i)$, independently across receivers and rounds, and put $\mathcal{T}_{i,t}^{\mathrm{rand}} = \{J_{i,t}\}$. At one round, with $\mathbf{v}_{ij} = a_{ij}(\mathbf{u}_{ij} - \mathbf{z}_i)$ and $\mathbf{v}_i = \sum_{j \in \mathrm{In}(i)} \mathbf{v}_{ij}$,*

$$\mathbb{E}\big[\widehat{\mathbf{y}}_i^{(\kappa),\mathrm{rand}} \,\big|\, \Xi^\circ\big] = \mathbf{z}_i + \Big(1 - \frac{1-\kappa}{d_i^-}\Big)\mathbf{v}_i. \tag{117}$$

For a window W, $\mathbb{E}[\mathsf{M}_W^{\mathrm{rand}} \mid \Xi^\circ]$ is the α-weighted window average of

$$\frac{1}{n\alpha_t} \sum_i \frac{1}{d_i^-}\Big[\sum_{b \in \mathrm{In}^-(i)} a_{ib}\|o_{ib,t}\| - \sum_{h \in \mathrm{In}^+(i)} a_{ih}\|o_{ih,t}\|\Big].$$

If distance targets $\mathcal{T}^{\mathrm{dist}}$ are computed from Ξ° with the same first-stage screen, then

$$\mathbb{E}\big[\mathscr{E}_W^{\mathrm{rand}}(\kappa) \,\big|\, \Xi^\circ\big] - \mathscr{E}_W^{\mathrm{dist}}(\kappa) = (1-\kappa)\big(\mathsf{M}_W^{\mathrm{dist}} - \mathbb{E}[\mathsf{M}_W^{\mathrm{rand}} \mid \Xi^\circ]\big). \tag{118}$$

Proof. At one round, $\widehat{\mathbf{y}}_i^{(\kappa),\mathrm{rand}} = \mathbf{z}_i + \mathbf{v}_i - (1-\kappa)\mathbf{v}_{iJ_i}$, and $\Pr(J_i = j \mid \Xi^\circ) = 1/d_i^-$ gives (117) and the margin by linearity. Both targets share $\mathbf{u}$, hence $\bar{d}_W^{\mathrm{res}}$ and $\bar{\mathsf{e}}_W^{\mathrm{byz}} = \bar{\mathsf{e}}_W^{\mathrm{hit}} + \bar{\mathsf{e}}_W^{\mathrm{miss}}$; for either target, (113) reads $\mathscr{E}_W(\kappa) = \bar{d}_W^{\mathrm{res}} + \bar{\mathsf{e}}_W^{\mathrm{byz}} - (1-\kappa)\mathsf{M}_W$. Taking conditional expectations gives (118). □

In the LQ graph $d_i^- = 6$, so $\mathbb{E}[\mathsf{M}_W^{\mathrm{rand}} \mid \Xi^\circ] = (\bar{\mathsf{e}}_W^{\mathrm{byz}} - \bar{\mathsf{e}}_W^{\mathrm{peer}})/6$, with $\bar{\mathsf{e}}_W^{\mathrm{peer}}$ the honest exposure when every honest peer is targeted, including at receivers without a Byzantine in-neighbor. At $\kappa = 0.1$, (117) multiplies every peer offset by 0.85 in conditional mean, whereas a selected Byzantine peer keeps 0.1 of its offset. Every singleton selection adds the same self mass $(1-0.1)\cdot 5/42 = 3/28$; the envelope comparison depends on the margin in (118). Along an executed run, the identity conditions on the history and raw row before the current draw; realized offsets remain selection-dependent.

Receivers without a Byzantine in-neighbor. If $\mathrm{In}^-(i) = \varnothing$, receiver i contributes $-(n\alpha_t)^{-1}\sum_{h\in\mathrm{In}^+(i)\cap\mathcal{T}_{i,t}} a_{ih}\|o_{ih,t}\| \le 0$ to the margin, with equality for an empty target set. The LQ graph has six exposed and three unexposed honest receivers, so its single-peer distance target needs hit exposure at the six exposed receivers to outweigh targeted honest exposure at all nine; Appendix G.5 quantifies this cost.

Screening counterpart. The same bookkeeping compares screening with interval clipping at $\kappa = 1$. For sender j at receiver i, let $\mathbf{c}_{ij,t} = \Pi_{\mathbf{I}_i}(\mathbf{y}_{j\to i}) - \mathbf{z}_{i,t}$ be its IntervalOnly offset and $O_{ij,t}$ the diagonal indicator of its coordinates outside the raw intervals. By (8), flagging j at full retention changes this offset to $(I - O_{ij,t})\mathbf{c}_{ij,t}$, so on a fixed record

$$\Delta_t^{\mathrm{screen}} = \Delta_t^{\mathrm{IO}} - \frac{1}{n\alpha_t}\sum_i \sum_{j\in\mathcal{S}_{i,t}} a_{ij} F^{-1} O_{ij,t}\mathbf{c}_{ij,t}. \tag{119}$$

Suppose every coordinate of each Byzantine message lies outside its interval, so a flagged Byzantine slot contributes zero. Let $\bar{\mathsf{e}}_W^{\mathrm{out,hit}}$ and $\bar{\mathsf{e}}_W^{\mathrm{out,hon}}$ be the removed Byzantine and honest exposures defined below. With the honest IntervalOnly residual $\bar{d}_W^{\mathrm{IO,res}}$ and the total Byzantine IntervalOnly exposure $\bar{\mathsf{e}}_W^{\mathrm{IO,byz}}$, the triangle inequality gives the envelopes $\mathscr{E}_W^{\mathrm{IO}} = \bar{d}_W^{\mathrm{IO,res}} + \bar{\mathsf{e}}_W^{\mathrm{IO,byz}}$ and $\mathscr{E}_W^{\mathrm{screen}} = \mathscr{E}_W^{\mathrm{IO}} - \bar{\mathsf{e}}_W^{\mathrm{out,hit}} + \bar{\mathsf{e}}_W^{\mathrm{out,hon}}$, so

$$\mathscr{E}_W^{\mathrm{IO}} - \mathscr{E}_W^{\mathrm{screen}} = \bar{\mathsf{e}}_W^{\mathrm{out,hit}} - \bar{\mathsf{e}}_W^{\mathrm{out,hon}}, \tag{120}$$

where $\bar{\mathsf{e}}_W^{\mathrm{out,hit}}$ and $\bar{\mathsf{e}}_W^{\mathrm{out,hon}}$ are the window averages of $(n\alpha_t)^{-1}\sum_i\sum_{b\in\mathrm{In}^-(i)\cap\mathcal{S}_{i,t}} a_{ib}\|F^{-1}\mathbf{c}_{ib,t}\|$ and $(n\alpha_t)^{-1}\sum_i\sum_{h\in\mathrm{In}^+(i)\cap\mathcal{S}_{i,t}} a_{ih}\|F^{-1}O_{ih,t}\mathbf{c}_{ih,t}\|$. Screening thus tightens the clipping envelope when it removes more out-of-range Byzantine than honest exposure; reduced retention faces the same trade-off on selected slots, scaled by $1-\kappa$. For partially in-range messages, (119) keeps both components explicit. A uniformly random one-peer screen, drawn afresh on the same raw rows, flags each peer with conditional probability $1/d_i^-$; both screens share the raw intervals and hence the honest IntervalOnly residual, so the linearity argument of Proposition 3 gives

$$\begin{aligned}\mathbb{E}[\mathscr{E}_W^{\mathrm{RS}} \mid \Xi^\circ] - \mathscr{E}_W^{\mathrm{screen}} &= \mathsf{M}_W^{\mathrm{out}} - \mathbb{E}[\mathsf{M}_W^{\mathrm{out,RS}} \mid \Xi^\circ],\\ \mathsf{M}_W^{\mathrm{out}} &= \bar{\mathsf{e}}_W^{\mathrm{out,hit}} - \bar{\mathsf{e}}_W^{\mathrm{out,hon}}.\end{aligned} \tag{121}$$

where $\mathbb{E}[\mathsf{M}_W^{\mathrm{out,RS}} \mid \Xi^\circ]$ is the random-target margin of Proposition 3 with $\|o_{ib,t}\|$ and $\|o_{ih,t}\|$ replaced by $\|F^{-1}\mathbf{c}_{ib,t}\|$ and $\|F^{-1}O_{ih,t}\mathbf{c}_{ih,t}\|$.

Neural realization and message-space replay. Replacing F^{-1} by the identity gives the same per-round algebra for aggregated prediction targets. In one channel, write $\mathbf{o}_{ij,t}^{\mathrm{msg}} = \mathbf{u}_{ij,t} - \mathbf{z}_{i,t}$ and use the query-RMS norm $\|v\|_{\mathrm{q},t} = \|v\|_2/\sqrt{K_{x,t}}$, where $K_{x,t}$ is the number of queries. Define the message exposures $\widehat{\mathsf{e}}_t^{\mathrm{hon}}$, $\widehat{\mathsf{e}}_t^{\mathrm{miss}}$ and $\widehat{\mathsf{e}}_t^{\mathrm{hit}}$ by replacing $\|o_{ij,t}\|$ with $\|\mathbf{o}_{ij,t}^{\mathrm{msg}}\|_{\mathrm{q},t}$ in the exposure sums and omitting α_t^{-1}. The residual and realized shift are

$$\begin{aligned}\widehat{d}_t^{\mathrm{res}} &= \left\| \frac{1}{n}\sum_i \sum_{h\in\mathrm{In}^+(i)} a_{ih}\mathbf{o}_{ih,t}^{\mathrm{msg}} \right\|_{\mathrm{q},t},\\ \widehat{s}_t(\kappa) &= \left\| \frac{1}{n}\sum_i (\widehat{\mathbf{y}}_{i,t}^{(\kappa)} - \mathbf{z}_{i,t}) \right\|_{\mathrm{q},t}.\end{aligned} \tag{122}$$

For $\langle f_t\rangle_W = |W|^{-1}\sum_{t\in W} f_t$, the reported summaries obey

$$\begin{aligned}\langle \widehat{s}_t(\kappa)\rangle_W \le \widehat{\mathscr{E}}_W(\kappa) &:= \langle \widehat{d}_t^{\mathrm{res}} + \widehat{\mathsf{e}}_t^{\mathrm{miss}} + \kappa\widehat{\mathsf{e}}_t^{\mathrm{hit}} + (1-\kappa)\widehat{\mathsf{e}}_t^{\mathrm{hon}}\rangle_W,\\ \widehat{\mathsf{M}}_W &:= \langle \widehat{\mathsf{e}}_t^{\mathrm{hit}} - \widehat{\mathsf{e}}_t^{\mathrm{hon}}\rangle_W.\end{aligned} \tag{123}$$

The per-round triangle inequality proves the bound, and (114) and (118) hold with these hatted quantities; the IntervalOnly residual and removed offsets in the same norm give the message-space versions of (120) and (121). Norms are taken within rounds, accommodating query sets that change during training; targets are shared across critics, while exposures and margins are channel-specific. These equal-round diagnostics describe target formation, whereas Corollary 1 also uses F^{-1} and asymptotic critic-time windows. Appendix G.5 reports the replays.

E.7 Actor noise and projected limits

For $w_t = (\mathbf{v}_t, \boldsymbol{\omega}_t)$, $\upsilon_t = (\theta_t, \varpi_t)$ and $\vartheta_t = \iota(\upsilon_t)$, write the actor sample as $\mathsf{H}_{\upsilon_t, w_t}(Z_t)$.

For $D\phi = \gamma\phi(s') - \phi(s)$, one sufficient constant in Assumption 6 is

$$L_F = \sup_{\vartheta} \mathbb{E}_{\mu_\vartheta}\left[\|\ell_\theta\|(\|\psi\|^2 + \|D\phi\|^2)^{1/2}\right] < \infty. \tag{124}$$

The catalogue ensures finiteness; block-diagonal Cauchy–Schwarz gives $\|\overline{\mathsf{H}}(\upsilon, w) - \overline{\mathsf{H}}(\upsilon, \widetilde{w})\| \le L_F\|w - \widetilde{w}\|$.

Lemma 15 (Weighted critic variation and actor noise)**.** *For $x \in \{v, \omega\}$ and each fixed $T < \infty$, almost surely,*

$$\begin{aligned} &\sum_t a_{\theta,t} W(s_{t+1})\|\mathbf{x}_{t+1} - \mathbf{x}_t\| < \infty, \\ &\sup_{m \le k \le \tau_\theta(m,T)} \left\| \sum_{t=m}^{k} a_{\theta,t}\zeta_{t+1} \right\| \longrightarrow 0, \end{aligned} \tag{125}$$

where $\tau_\theta(m, T)$ ends the forward actor-time window of mass T and $\zeta_{t+1} = \mathsf{H}_{\upsilon_t, w_t}(Z_t) - \overline{\mathsf{H}}(\upsilon_t, w_t)$. The safeguard is eventually inactive almost surely.

Proof. Stop at $\tau_R = \inf\{t : \|w_t\| > R\}$ for integer R. The nonexpansive extension $w_t^R = \Pi_{\overline{B}_R} w_t$ equals w_t before exit and controls the Poisson endpoint at exit. Lemmas 10–11, the deterministic-start compact-forcing filter and (78) give, after a finite pathwise transient,

$$\|\mathbf{x}_{t+1} - \mathbf{x}_t\|\mathbf{1}_{\{t<\tau_R\}} \le C_R \alpha_{x,t} K_{x,t+1}^{(R)}, \qquad \sup_t \mathbb{E}[W(s_{t+1})K_{x,t+1}^{(R)}] < \infty.$$

Since eventually $a_{\theta,t} \le \alpha_{x,t}$, Tonelli proves weighted variation; the transient is handled pathwise.

The primitive actor bound gives $\Pr(t < \tau_R, \|\mathsf{H}_t\| > R_t) \le \Pr(K_{t+1}^{\mathrm{ac}} > R_t/C_R) \le C_R^{q_{\mathrm{cat}}} R_t^{-q_{\mathrm{cat}}} \sup_u \mathbb{E}[(K_{u+1}^{\mathrm{ac}})^{q_{\mathrm{cat}}}]$, a summable sequence by (46). Thus clipping on this tube is eventually inactive and $\|\theta_{t+1} - \theta_t\| \le C_R a_{\theta,t} K_{t+1}^{\mathrm{ac}}$ there. The adjacent-weight moments, physical-policy rate and weighted variation yield

$$\sum_t a_{\theta,t} W(s_{t+1})\big(\|\theta_{t+1} - \theta_t\| + d_\Pi(\varpi_{t+1}, \varpi_t) + \|w_{t+1} - w_t\|\big)\mathbf{1}_{\{t<\tau_R\}} < \infty. \tag{126}$$

Apply Lemma 6 to the actor edge field, with physical conditional mean $Q_{\iota(\upsilon)}\mathsf{H}_{\upsilon,w}$, using w^R and predictable stopping indicators on the deterministic tube. Its martingale series converge and Abel endpoints vanish; (126) controls the parameter remainder. The centered noise windows therefore vanish on $\{\tau_R = \infty\}$. By boundedness (Appendix C.2), $\bigcup_R\{\tau_R = \infty\}$ is almost sure. Countable intersection removes stopping under the original law, without conditioning on future boundedness. □

Lemma 16 (Viable projected limits)**.** *For the affine actor-clock interpolation $\bar{\theta}$ and each fixed $T < \infty$, almost surely,*

$$\lim_{r\to\infty} \inf_{\phi \in \mathrm{Sol}_\Theta(\mathcal{F}_{\epsilon_{\mathrm{ac}}}^K)} \sup_{0\le s\le T} \|\bar{\theta}(r+s) - \phi(s)\| = 0, \tag{127}$$

where Sol_Θ denotes viable arcs of (60) in Θ.

Proof. Put $M = \sup_\Theta \|g\| + \epsilon_{\mathrm{ac}}$. By (125), $e_t \to 0$ and the vanishing gain mesh, the unconstrained input in (28) has oscillation at most $M\delta + o(1)$ on late intervals of length δ, uniformly over each fixed horizon. Reflection (58) and the vanishing endpoint error $\|\theta_{t+1} - \theta_t\| \le c_n a_{\theta,t} R_t \to 0$ give equicontinuous translated paths.

Take a path subsequence and a weak-* control limit u. Inputs converge uniformly to $z(s) = \theta(0) + \int_0^s [g(\theta(q)) + u(q)]\,dq$, with $\|u(q)\| \le \epsilon_{\mathrm{ac}}$ almost everywhere, and reflection continuity gives $\theta = \mathscr{R}_\Theta(z)$. Passing the projection inequality to step-function tests yields $\dot\theta = g(\theta) + u - p$, $p \in N_\Theta(\theta)$, $\dot\theta \in T_\Theta(\theta)$ and $\langle p, \dot\theta\rangle = 0$ almost everywhere. Moreau decomposition then gives $\dot\theta = \Pi_{T_\Theta(\theta)}[g(\theta) + u] \in \mathcal{F}_{\epsilon_{\mathrm{ac}}}^K(\theta)$. Every translation sequence admits this limit, proving (127). □

E.8 Critic approximation and the discounted objective

Hold ϖ and ν fixed, with $\nu W < \infty$; suppress ϖ in $P_\theta = P_{\iota(\theta,\varpi)}$, ρ_θ and $\mu_\theta = \rho_\theta Q_\theta$. For latent utility $r(Z) = n^{-1}\sum_i r_i^\star(Z)$ and $r_\theta(s) = \mathbb{E}[r(Z) \mid s]$, set

$$V_\theta = (I - \gamma P_\theta)^{-1} r_\theta, \qquad d_{\gamma,\theta} = (1-\gamma)\sum_{k\geq 0}\gamma^k \nu P_\theta^k, \tag{128}$$
$$\mathcal{J}(\theta) = (1-\gamma)\nu V_\theta = (1-\gamma) J_{\mathrm{pref}}(\theta;\nu,\varpi).$$

The synchronized score is $\ell_\theta(Z) = B\,\mathrm{col}_{i\in\mathcal{N}^+}\nabla_{\theta_i}\log\pi_i(a_i \mid s;\theta_i)$.

Let $\Phi v = \phi^\top v$, with adjoint Φ^* in $L^2(\rho_\theta)$, and write $\widehat{V}_\theta = \Phi v^\star(\theta)$ and $\widehat{r}_\theta = \psi^\top\omega^\star(\theta)$. Define

$$E_r(\theta) = \inf_\omega \|\psi^\top\omega - r\|_{L^2(\mu_\theta)}, \quad E_V(\theta) = \inf_v \|\phi^\top v - V_\theta\|_{L^2(\rho_\theta)},$$
$$k_\lambda = \frac{\gamma(1-\lambda)}{1-\gamma\lambda}, \qquad C_\lambda = (1-k_\lambda^2)^{-1/2}. \tag{129}$$

Lemma 17 (Reward projection and TD approximation)**.** *Under the stated moment and excitation conditions,*

$$\|\widehat{r}_\theta - r\|_{L^2(\mu_\theta)} = E_r(\theta), \qquad \|\widehat{V}_\theta - V_\theta\|_{L^2(\rho_\theta)} \leq C_\lambda E_V(\theta). \tag{130}$$

Proof. Reward regression is orthogonal projection. For $a = \gamma\lambda$, stationary trace expansion gives

$$\bar{h}_v = \Phi^*(I - aP_\theta)^{-1} r_\theta. \tag{131}$$

Here $\mathbb{E}[\phi(s_{t-j})r(Z_t)] = \rho_\theta[\phi P_\theta^j r_\theta]$; Cauchy–Schwarz and $a<1$ justify summation. With the $\bar{G}_v$ identity in (57) and the orthogonal feature projection Π_ϕ, the critic normal equations become

$$\widehat{V}_\theta = \Pi_\phi T_\lambda \widehat{V}_\theta, \qquad T_\lambda f = (I - aP_\theta)^{-1}[r_\theta + \gamma(1-\lambda)P_\theta f].$$

Since $\|P_\theta\|_{2\to 2} \leq 1$, contraction by k_λ, the fixed point V_θ and orthogonality give

$$\|V_\theta - \widehat{V}_\theta\|_2^2 = E_V(\theta)^2 + \|\Pi_\phi V_\theta - \widehat{V}_\theta\|_2^2 \leq E_V(\theta)^2 + k_\lambda^2\|V_\theta - \widehat{V}_\theta\|_2^2. \qquad \square$$

Use the score identities (62) under the weighted C^1 conditions of Appendix D.6. Define

$$F_{\gamma,\theta}(s) = \mathbb{E}[\ell_\theta\{r(Z) + \gamma V_\theta(s') - V_\theta(s)\} \mid s], \qquad S_\theta = \left(\mathbb{E}_{\mu_\theta}\|\ell_\theta\|^2\right)^{1/2}. \tag{132}$$

Lemma 18 (Discounted policy-gradient identity)**.** *Under Appendix D.6's conditions,* $\nabla\mathcal{J}(\theta) = d_{\gamma,\theta}F_{\gamma,\theta}$.

Proof. The state-value baseline cancels by (62). Splitting $P_\theta^k = \mathbf{1}\rho_\theta + (P_\theta^k - \mathbf{1}\rho_\theta)$ bounds the resolvent on $\mathcal{B}_W$ by weighted ergodicity. Inverse differentiation gives

$$(I - \gamma P_\theta)\nabla V_\theta = \nabla r_\theta + \gamma(\nabla P_\theta)V_\theta = F_{\gamma,\theta}. \tag{133}$$

Multiply by $(1-\gamma)\nu(I-\gamma P_\theta)^{-1}$ to conclude. $\square$

Lemma 19 (Derived objective discrepancy)**.** *Put* $b_{\mathrm{occ}}(\theta) = \|(\rho_\theta - d_{\gamma,\theta})F_{\gamma,\theta}\|$. *For the latent actor field g in (26),*

$$\|g(\theta) - \nabla\mathcal{J}(\theta)\| \leq S_\theta[E_r(\theta) + \gamma C_\lambda E_V(\theta)] + \|(\rho_\theta - d_{\gamma,\theta})F_{\gamma,\theta}\|. \tag{134}$$

For R, A_r, A_P from Appendix D.6 and C_P, ρ_P from (37),

$$\|(\rho_\theta - d_{\gamma,\theta})F_{\gamma,\theta}\| \leq \frac{C_P\nu W(1-\gamma)}{1-\gamma\rho_P}\left(A_r + \frac{\gamma A_P C_P R}{1-\gamma\rho_P}\right). \tag{135}$$

Proof. The current value-error baseline cancels by (62). Cauchy–Schwarz, stationarity of s', and Lemma 17 give

$$\|g - \rho_\theta F_{\gamma,\theta}\| \leq S_\theta\big(\|\widehat{r}_\theta - r\|_{L^2(\mu_\theta)} + \gamma\|\widehat{V}_\theta - V_\theta\|_{L^2(\rho_\theta)}\big).$$

Lemma 18 proves (134). Weighted mixing and $\nu\mathbf{1} = 1$ give $\|d_{\gamma,\theta} - \rho_\theta\|_W \leq C_P\nu W(1-\gamma)/(1-\gamma\rho_P)$. For $j_\theta = \rho_\theta r_\theta$, center the resolvent:

$$V_\theta = \frac{j_\theta}{1-\gamma} + H_\theta, \qquad H_\theta = \sum_{k\geq 0}\gamma^k(P_\theta^k - \mathbf{1}\rho_\theta)r_\theta, \qquad \|H_\theta\|_W \leq \frac{C_P R}{1-\gamma\rho_P}. \tag{136}$$

The constant part has zero score mean, so $F_{\gamma,\theta} = \mathbb{E}[\ell_\theta(r + \gamma H_\theta(s')) \mid s]$ and $\|F_{\gamma,\theta}\|_W \leq A_r + \gamma A_P C_P R/(1-\gamma\rho_P)$. The dual-norm bound proves (135). $\square$

F REWARD-MODEL EXTENSIONS

This branch replaces the calibrated label clause of Assumption 3, retaining bounded excitation and the other conditions for the critic and actor conclusions. All field, transport and objective bounds below concern this branch's executed law and full latent utility. Let the true label probability be $p_{i,\vartheta}(\mathcal{T}^{\mathrm{blk}})$, where $\mathcal{T}^{\mathrm{blk}}$ contains the observed block and pre-label history but excludes the fresh label. Require (49)–(50) for this label kernel, with the moment/Poisson catalogue covering $p_{i,\vartheta}(\mathcal{T}^{\mathrm{blk}})D$, its weighted products and increments. For fixed $\eta_i^0 \in \mathcal{K}_i$, define

$$\epsilon_{i,\mathrm{score}} = \sup_{\vartheta} \left\| \beta_{\mathrm{p}} \mathbb{E}_{\nu_\vartheta^{\mathrm{blk}}} [(\sigma(\beta_{\mathrm{p}} D^\top \eta_i^0) - p_{i,\vartheta}(\mathcal{T}^{\mathrm{blk}}))D] \right\|. \tag{137}$$

Put $r = \epsilon_{\eta,i} = \epsilon_{i,\mathrm{score}}/\lambda_{\mathrm{pref},i}$. Strong monotonicity (86) gives $\langle \nabla V_r(e), \bar{g}_i(\eta_i^0+e)\rangle \geq 2\lambda_{\mathrm{pref},i} V_r(e)$ for $V_r(e) = (\|e\| - r)_+^2$, whose gradient is Lipschitz. Projection about η_i^0 cannot increase V_r. Lemma 8 with multiplier ∇V_r, the summable quadratic term and the exact proof's tail correction give convergence of V_r and $\sum_\ell a_{\eta,\ell} V_r < \infty$. Hence, almost surely,

$$\limsup_{\ell} \|\eta_{i,\ell} - \eta_i^0\| \leq \epsilon_{\eta,i}. \tag{138}$$

Corollary 2 (Preference bias through the critics)**.** *Under these additional conditions, for each fixed $T > 0$, almost surely,*

$$\limsup_{t} \|x_{i,t} - x_x^\star(\vartheta_t; \eta^0)\| \leq K_x(T) \left[\overline{\Delta}_{x,T}^\star + \mu_{A,x} \max_i \epsilon_{\eta,i} \right]. \tag{139}$$

Proof. Equations (98) and (138) bound the window intensity of $A_{i,t+1}^x(\eta_{i,\ell(t)} - \eta_i^0)$ by $\mu_{A,x} \max_i \epsilon_{\eta,i}$. Adding this forcing to the comparison yielding (25) gives (139). Adding the same residual bound to the net-window comparison (24) gives $E^\star \leq e^{-cT} E^\star + (1 + M_G T) T \mathfrak{D}_x + T \mu_{A,x} \max_i \epsilon_{\eta,i}$. Taking $c \uparrow \delta_x$ and $T_m = 1/m \downarrow 0$ gives $(\mathfrak{D}_x + \mu_{A,x} \max_i \epsilon_{\eta,i})/\delta_x$. □

Latent utility targets. For utility $\chi^\top \eta_i^\star$, coercivity and $L_{x,\eta} = \mu_{A,x}$ give the target shift $b_x^{\mathrm{tar}} = (L_{x,\eta}/\delta_x)\epsilon_\eta^{\mathrm{tar}}$, where $\epsilon_\eta^{\mathrm{tar}} = \max_i \|\eta_i^0 - \eta_i^\star\|$. For out-of-span utility $r_i^\star = \chi^\top \eta_i^0 + e_i$, use instead $b_x^{\mathrm{tar}} = \epsilon_x^{\mathrm{app}}$ with

$$\epsilon_x^{\mathrm{app}} = \delta_x^{-1} \sup_{\vartheta} \left\| \frac{1}{n} \sum_i \mathbb{E}_\vartheta[\zeta_x e_i] \right\|, \qquad \zeta_v = \mathbf{e}_{t+1}, \quad \zeta_\omega = \psi_t. \tag{140}$$

For full-latent-utility BT labels, $\sup \sigma' = 1/4$ bounds the score bias by $(\beta_{\mathrm{p}}^2/4) \sup_\vartheta \mathbb{E}_\vartheta[|E_i^{\mathrm{pair}}| \|D\|]$, where E_i^{pair} is the discounted segment-pair residual. For v, require the residual's conditional mean to be bounded and Lipschitz in $\mathcal{B}_W$, and $|e_i(Z)| \leq C[W(s) + W(s')]$; Lemma 5 controls its forcing. The residual-label and reward-critic catalogues also remain required. The resulting latent-target radius is

$$R_x^{\mathrm{lat}} = \frac{\mathfrak{D}_x + \mu_{A,x} \max_i \epsilon_{\eta,i}}{\delta_x} + b_x^{\mathrm{tar}}. \tag{141}$$

Replacing $\mathfrak{D}_x/\delta_x$ by R_x^{lat} in the actor-error bound of Section 4.2 extends Theorem 3, provided the weighted objective conditions and approximation/occupancy bounds hold for the full latent utility. The defect $\mathfrak{D}_x$ remains that of the actual learned run; Lemma 13 requires exact identification.

Additive actor-tolerance bound. Write $\epsilon_\eta = \max_i \epsilon_{\eta,i}$ and $e_x^{\mathrm{pref}} = \mu_{A,x}\epsilon_\eta/\delta_x + b_x^{\mathrm{tar}}$. For this run define the norm-based communication transfer bound

$$\epsilon_{\mathrm{ac}} = L_F \sqrt{n} \left\| (\mathfrak{D}_x/\delta_x)_{x \in \{v,\omega\}} \right\|, \qquad \epsilon_{\mathrm{pref}}^{\mathrm{ac}} = L_F \sqrt{n} \left\| (e_x^{\mathrm{pref}})_{x \in \{v,\omega\}} \right\|.$$

The latent-target radii in (141) give

$$\delta_{\mathrm{ext}} := b_{\mathrm{obj}} + L_F \sqrt{n} \left\| (R_x^{\mathrm{lat}})_{x \in \{v,\omega\}} \right\| \leq b_{\mathrm{obj}} + \epsilon_{\mathrm{ac}} + \epsilon_{\mathrm{pref}}^{\mathrm{ac}} \tag{142}$$

by the triangle inequality. The squared-average and weighted-density conclusions use δ_{ext} in place of δ. The term b_{obj} is evaluated for the full latent objective under the final physical policy of this run. In the calibrated model choose $\eta_i^0 = \eta_i^\star$, giving $\epsilon_{\eta,i} = b_x^{\mathrm{tar}} = \epsilon_{\mathrm{pref}}^{\mathrm{ac}} = 0$. Without exact identification, neither Lemma 13's utility-span envelope nor its zero-span reduction follows.

G Experimental Details

Tables 3–4 collect the learner settings and evaluation protocols. Section G.3 supports the reward-source comparison, Section G.5 the LQ aggregation and attack-geometry results, and Section G.7 flocking; the remaining studies test linear critics and retention sensitivity.

G.1 Tasks, Learners and Evaluation

Networked LQ. Ten agents have joint state $s \in \mathbb{R}^{20}$, local states $s^i \in \mathbb{R}^2$ and scalar actions, with

$$s_{t+1} = As_t + Ba_t + w_t, \quad w_t \sim \mathcal{N}(0, 0.1^2 I), \qquad r_i(s,a) = -(s^i)^\top Q_i s^i - R_i a_i^2.$$

The dynamics couple neighboring two-dimensional blocks on a ring and have spectral radius 1.10; positive quadratic costs differ across agents. The main aggregation comparison uses ten LQ plants. Fixed-plant studies use the default system; the Function W-MSR comparison uses a distinct ten-plant suite. Policies are $a_i = \theta_i^\top s + \xi_i$, with $\xi_i \sim \mathcal{N}(0, 0.2^2)$. Node 9 corrupts messages while retaining a stabilizing LQR actor row. Evaluation uses the nine honest agents' stationary mean cost under the learned rows, computed from the closed-loop Lyapunov equation. A learned-controller spectral radius of at least one gives nonfinite stationary cost; reported controller failures are determined by this criterion.

Learning updates. Each LQ phase collects 2,000 transitions to fit local rewards, followed by 1,000 actor–critic transitions. Collection chooses the learned policy with probability $1/(1+\|s\|)$ and otherwise a stabilizing LQR reference; evaluation uses learned rows alone. Actor updates begin after 5,000 actor–critic transitions. The stepsize is $10^{-3}/\sqrt{1+u/5000}$ at update u, multiplied by three in the 160-phase study. Local Bradley–Terry teachers use undiscounted utilities over length-ten segments and inverse temperature one; each agent fits its own quadratic reward model. Pair budgets include a one-fifth held-out split. Critics communicate every ten transitions. Tables 3–4 give the implementations and study budgets. In the aggregation comparison, all methods use successive actor projections onto the origin-centered ball and $\{\theta : \|\theta - \theta_{i,0}\|_2 \le 0.25\|\theta_{i,0}\|_2\}$, with $\theta_{i,0}$ the initial half-gain LQR row.

Evaluation metrics. The main aggregation comparison (Tables 1, bottom, and 7) uses gain $g = (J_{0,p} - J)/J_{0,p}$, with J the final-five-phase mean cost and $J_{0,p}$ the plant's initial half-gain policy cost, shared by methods and seeds. Reward-source and Function W-MSR comparisons (Tables 5 and 10) use final-five/first-phase cost ratios. Fixed-plant neural comparisons (Figure 2 and the clipping, reward-budget and retention studies) use gain $(J_0 - J)/J_0$ with $J_0 \approx 0.2374$, the default plant's all-agent half-gain LQR cost. The linear-critic studies (Table 6) use $J_0 \approx 0.2314$. Flocking uses native team return. For a higher-is-better measure y, define

$$D_m = y_m^{\text{clean}} - y_m^{\text{attack}}, \qquad M_{m,b} = D_b - D_m, \qquad A_{m,b} = y_m^{\text{attack}} - y_b^{\text{attack}}.$$

Thus $M_{m,b}$ is own-clean loss reduction and $A_{m,b}$ is the direct advantage of the first method over the second under attack. We omit the first index when it is PARC.

Implementation	Linear LQ (LFA)	Neural LQ	Flocking
Private reward	Local quadratic fit	Local quadratic fit	Potential/residual kernel
Value / reward critic	231 / 496 coefficients	Two 64-unit ReLU layers	Two 128-unit tanh layers
Joint message	$231 + 496$ predictions	$64 + 64$ predictions	$1664 + 1664$ per fold
Realization	Full-rank linear inverse	Conditioned Adam fit	Persistent neural fit
Actor update	Buffered omega return	One-step critic TD	PPO with GAE(0.95)
Discount	0.95	0.95	1

Table 3: **Learning implementations.** Local reward models and communicated reward critics are distinct. Section G.7 specifies Flocking's action schedule.

G.2 Message Attacks, Baselines and Ablations

Topology and weights. LQ receiver i takes senders $i, i \pm 1, i \pm 2, i \pm 3$ modulo ten; Flocking uses the complete six-node graph, both with $H = 1$. LQ rows use $1/7$, except that the main aggregation

Comparison	Attacks	Phases	Pairs	Screen	Units
Main-text comparisons					
Reward sources (LFA)	B	20	200 / 0	3	10 paired seeds
LQ aggregation	B, I	20	800	Distance	10 plants
LQ attack geometry	B, D	160	800	6	6 paired seeds
Flocking	D	448	256 until 32	3	6 paired seeds
Additional studies					
Linear critics	B	20	200	3	10 seeds/test
Clipping and screening	B, D	20	50 / 800	6	6 paired seeds
Reward-budget sensitivity	B	20	50 / 200 / 800	3	6 paired seeds
Function W-MSR	B	20	200	3	10 plants
Retention sensitivity	I	20	800	Distance	10 paired seeds

Table 4: **Evaluation protocols.** B/D/I denote boundary, distance-constrained/in-interval attacks; each comparison includes clean runs. Pairs are new comparisons/node/phase (zero for numeric rewards); screen gives PARC's MAD multiplier or distance selection. Each plant averages two training seeds. LQ endpoints average the final five phases; Figure 2 also reports all-phase averages.

comparison and retention sensitivity use self/peer weights $2/7, 5/42$; Flocking PARC and Mean use $1/3, 2/15$. RTC uses uniform weights ($1/7$ or $1/6$); replacing one peer by self matches the weighted TPSC row's explicit self mass before retention.

Baselines and ablations. Mean averages raw proposals. IntervalOnly disables the sender screen while keeping TPSC's coordinate intervals and clipping. Function W-MSR trims coordinate extremes and renormalizes retained mass. Directional W-MSR adapts Ye et al. (2024) to LFA by trimming in the current feature direction. LocalOnly uses the same local learner without peer messages.

RTC adaptation. The message-space RTC baseline applies weighted distance-based removal followed by receiver-centered norm clipping (Yang & Ghaderi, 2024). Uniform weights and Byzantine weight bounds are $1/7$ for LQ and $1/6$ for flocking. The farthest nonself message is replaced by self at its original weight; the second-moment clipping radius is evaluated from the current retained message row. With retained set R, distances d_j from self and Byzantine weight bound $\delta_{\max}$, the radius is $\tau^2 = \sum_{j\in R} w_j d_j^2/\delta_{\max}$; retained offsets are scaled by $\min\{1, \tau/d_j\}$ for $d_j > 0$.

Screen and retention ablations. PARC flags the farthest of six nonself senders in the raw 128-dimensional joint row, with sender-ID tie breaks. RandomScreen draws one peer uniformly per receiver and round from an independent stream. $\mathrm{PARC}_{0.1}$ uses PARC's flag in both stages with $\kappa_{\mathrm{keep}} = 0.1$; RandomRetention preserves the first stage and draws only the retention target uniformly. Both add self mass $(1 - 0.1)5/42 = 3/28$, giving self weight $11/28$ relative to processed slots; changed coordinates depend on the target. Complete raw rows determine intervals, whose values can differ across the methods' learned trajectories.

Message access and attack geometry. Attacks act after local innovation and before aggregation, with receiver-specific messages. The attacker observes honest critic proposals; neural LQ additionally uses current actor scores, and VMAS's reward-critic attack uses lag-one actor-score sensitivity. Private rewards, labels and reward models are excluded. Within each condition, methods share the attack-generation rule, access and budget; their different trajectories can produce different malicious messages.

Boundary-exceeding corruption. Predictions are placed outside the honest coordinate range. LFA uses an eight-MAD overshoot; the linear-critic robustness study adds an alternating four-MAD perturbation on 200 joint state–action reward probes. Neural LQ selects opposite value/reward extremes using the actor-score direction toward $-\theta_i$.

Distance-constrained corruption. Neural LQ selects eight query indices by actor-score magnitude divided by honest prediction scale. It sends self plus a sparse joint value/reward offset with norm 0.9 times the maximum honest distance from self; other coordinates copy self. This combines

sparse, actor-sensitive corruption with a receiver-relative norm budget. VMAS preserves the sender background and imposes additional displacement and channel bounds (Section G.7).

In-interval corruption. Actor-score signs select the 0.05 or 0.95 interpolation point of each honest coordinate range, in opposite directions for value and reward critics.

G.3 Preference Information and Reward Accuracy

The reward-source comparison uses ten paired training seeds on one fixed LQ plant. Numeric utilities and preference predictions use the same local collection-mean centering rule, frozen within each phase, and common scale one. Interaction budgets, critic and MC actor updates, replay, communication, attacks and evaluation are otherwise unchanged. Numeric runs use no preference queries. Shuffling permutes each node's new labels within phase while keeping query pairs and held-out labels. All sources have identical first-phase costs within each seed.

Reward source	Critic communication	Clean	Boundary corruption
Matched numeric	TPSC	0.9552	0.9568
Trusted labels	TPSC	0.9664	0.9697
Shuffled labels	TPSC	1.0004	1.0017
Trusted labels	LocalOnly	0.9901	–

Table 5: **Reward information and control: cost ratio ↓.** Mean last-five-phase cost divided by first-phase cost, averaged over ten paired seeds on one fixed plant. Numeric and preference sources use matched centering and unit scale. LocalOnly uses the same reward-source protocol; a dash marks a condition not included in this design.

Trusted labels improve on shuffled labels by 0.0340 clean and 0.0320 under boundary corruption (Table 1). Local numeric rewards provide an information-rich reference: matched numeric runs are all finite and have lower ratios than trusted preferences by 0.0112 ($[0.0093, 0.0130]$) clean and 0.0129 ($[0.0109, 0.0149]$) under attack, positive on all ten paired seeds per condition; intervals are pointwise 95% bootstrap CIs. Relative to shuffled labels, trusted preferences recover about 75% of the numeric-reward improvement without attack and 71% under boundary corruption. These percentages use the mean cost ratios in Table 5.

Preprocessing sensitivity. Without centering, the same numeric utilities yield ratios $1.0499/1.0455$ (clean/boundary). Centering reduces them by $0.0947/0.0887$, with improvement on all ten seeds in each condition. This comparison varies preprocessing rather than reward information.

The ten fixed-policy datasets use nested training prefixes and common independent held-out data (Figure 4). At 50/200/800 comparisons, held-out NLL is $0.659/0.610/0.598$ and AUC is $0.693/0.724/0.737$. Reward error centers predictions and true rewards on the same held-out trajectory.

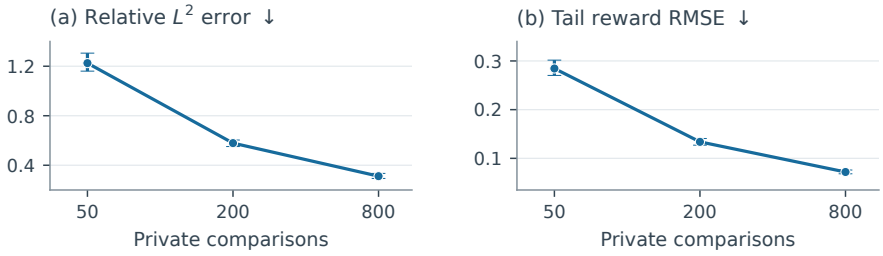


Figure 4: **Private reward accuracy.** Means and 95% bootstrap intervals over ten datasets, with node averages formed first. Predictions and true rewards are centered on the full held-out set; tail RMSE uses its top 10% of state norms with the same centering. Both errors decrease with budget.

G.4 Robust Control with Linear Critics

Under structured boundary corruption, Mean's learned controller becomes unstable on all ten seeds, whereas PARC and Directional W-MSR remain finite (Table 6); PARC attains mean attacked gain 0.0373. In the ten-seed LFA screening comparison, PARC's median loss reduction relative to IntervalOnly is small: 7.7×10^{-5} (95% CI $[-1.91, 1.14] \times 10^{-3}$).

Comparison	Method	Clean	Structured corruption
Protection	Mean	+0.0230	0/10 finite
Protection	Directional W-MSR	+0.0114	+0.0129
Protection	PARC	+0.0343	+0.0373
Screening	IntervalOnly	+0.0213	+0.0265
Screening	PARC	+0.0230	+0.0275

Table 6: **Linear critics: PARC learns finite-cost controllers under attack.** Under boundary corruption, Mean's learned controller is unstable on all ten seeds, whereas PARC and Directional W-MSR remain finite. Entries are mean last-five-phase gains (↑) relative to the fixed initial-policy cost; the protection and screening comparisons each use ten paired seeds and matched clean runs. Nonfinite groups report finite counts.

Critic communication improves control: in the ten-seed LocalOnly comparison, PARC exceeds the matched LocalOnly learner's fixed-reference gain by 0.0354 (95% CI $[0.0165, 0.0625]$), with positive differences on all ten seeds.

G.5 Neural Control: Screening and Retention

Main aggregation comparison. All 240 runs of the five-rule comparison have finite stationary costs and positive final-five-phase gains. The comparison uses ten LQ plants and two seeds each under the shared learner settings of Section G.1, covering five clean, three boundary and four in-interval cells; all clean, attacked and own-clean contrasts use all ten plants (Table 7).

Rule	Clean	Boundary	In-interval
IntervalOnly	0.1046 (0.0103)	0.0249 (0.0138)	0.0417 (0.0142)
PARC	0.1056 (0.0107)	0.0988 (0.0106)	0.0533 (0.0116)
RandomScreen	0.1056 (0.0094)	0.0381 (0.0120)	—
$\text{PARC}_{0.1}$	0.1061 (0.0099)	—	0.0868 (0.0129)
RandomRetention	0.1072 (0.0104)	—	0.0752 (0.0123)

Table 7: **Main aggregation comparison: all 240 runs finite with positive gains.** Entries are mean gain (↑) with sample SD across ten plant means; each plant averages two seeds. Each populated cell has 20/20 finite runs and 20/20 positive endpoint gains; dashes mark conditions not included in this design.

Table 8 reports clean differences and clean-to-attack loss reductions for C1–C4. The in-interval attacked-to-clean ratios of mean gains are 82% for $\text{PARC}_{0.1}$, 50% for PARC and 70% for RandomRetention. RandomRetention exceeds PARC by approximately 0.022 in in-interval gain, positive on all ten plants and accounting for about two-thirds of C3; its clean-to-attack loss is 0.0204 smaller. By Proposition 3, a uniformly random singleton target multiplies every peer offset by $1 - (1 - \kappa)/d_i^- = 0.85$ in conditional mean; the additional targeting gain is C4.

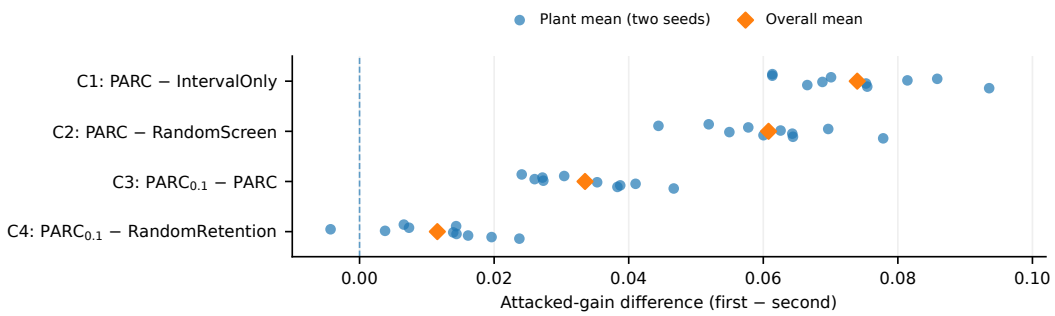


Figure 5: **Intervention and selection effects across the ten LQ plants.** Points are two-seed plant means for C1–C4 (Table 1); diamonds show overall means. All ten plants contribute to every paired contrast.

Condition	Contrast	Clean gain difference	Loss reduction	n
Boundary	C1	+0.0010	+0.0729	10
Boundary	C2	+0.0000	+0.0608	10
In-interval	C3	+0.0005	+0.0330	10
In-interval	C4	−0.0010	+0.0126	10

Table 8: **Clean differences and own-clean loss reductions.** Mean paired differences over all ten plants. C1–C4 follow Table 1. Clean difference is first minus second; loss reduction is second minus first, using each rule's own clean run.

Same-input exposure replay. We replay aggregation on the messages from 20 boundary PARC runs and 20 in-interval $\mathrm{PARC}_{0.1}$ runs (ten plants, two seeds per cell). A read-only observer regenerates each trajectory from its original configuration, reproducing phase tables, model states and random-number states exactly. It records all 2,000 communication rounds and all nine honest receivers, using public messages and the simulator's sender identities rather than private labels, utilities or reward models. Retention replays fix raw intervals and first-stage processing while varying retention levels and targets. Screening replays fix raw intervals and vary flags. Alternative outputs are evaluated without advancing the learner. Random selections use the analytic conditional expectations in (118) and (121).

Exposures, margins, envelopes and realized shifts are the query-RMS message-space quantities in (122)–(123). The two channels share a target selected from the joint 128-dimensional row. Rounds have equal weight; two seeds are averaged within each plant. Reported intervals are descriptive, pointwise 95% plant-bootstrap intervals from 20,000 resamples. The primary window uses all 2,000 rounds; the last 500 form an auxiliary window.

Retention targets: in-interval $\mathrm{PARC}_{0.1}$ *rows*

Channel	Hit	Honest	Missed	Residual	Dist. margin	Random	Dist. − Random
Value	0.0133	0.0135	0.0132	0.0168	−0.0002 [−0.0024, 0.0019]	−0.0189	+0.0187 [0.0163, 0.0211]
Reward	0.0038	0.0062	0.0049	0.0094	−0.0024 [−0.0032, −0.0017]	−0.0091	+0.0067 [0.0059, 0.0075]

First-stage screening: boundary PARC rows

Channel	Removed Byzantine	Removed honest	Dist. margin	RandomScreen	All outside
Value	0.0276	0.0060	+0.0217 [0.0203, 0.0232]	−0.0041	100%
Reward	0.0091	0.0027	+0.0064 [0.0060, 0.0069]	−0.0033	100%

Table 9: **Same-input exposure replay.** Message-space averages over all nine honest receivers and 2,000 rounds; rounded ten-plant means (two seeds each) with descriptive pointwise 95% plant-bootstrap intervals in brackets. Hit, Honest, Missed and Residual average $\widehat{\mathrm{e}}^{\mathrm{hit}}$, $\widehat{\mathrm{e}}^{\mathrm{hon}}$, $\widehat{\mathrm{e}}^{\mathrm{miss}}$ and $\widehat{d}^{\mathrm{res}}$; Random columns are conditional-mean margins. Positive plants: retention distance margin 5/10 (value) and 0/10 (reward); screening margins and retention Dist.−Random 10/10 in both channels. "All outside" is the fraction of Byzantine payloads with every coordinate outside its raw interval.

Screening margins. Every recorded boundary Byzantine payload lies outside all raw intervals, so (120) applies throughout. Distance screening removes more Byzantine than honest exposure in both channels on all ten plants, whereas RandomScreen's conditional-mean margins are negative on the same inputs (Table 9, Figure 6a,b), supporting C1 and C2.

Targeting advantage and network-level retention cost. Distance targets exceed random targets in margin by $+0.0187$ (value) and $+0.0067$ (reward) on all ten plants (Figure 6c,d); greater hit exposure and lower targeted honest exposure both contribute. For in-interval messages the distance target captures about half of the Byzantine exposure in value (hit 0.0133, missed 0.0132) and about 44% in reward (0.0038, 0.0049). The distance margin itself is -0.0002 in value (interval spanning zero) and -0.0024 in reward (negative on every plant). It is $+0.0099$ and $+0.0010$ at the six receivers with a Byzantine in-neighbor and -0.0202 and -0.0092 at the other three, where a singleton target can only shrink an honest peer (Appendix E.6); the network margin weights these within-group averages by $6/9$ and $3/9$. Emptying only those three receivers' targets on the same records would leave approximately $+0.0066$ and $+0.0007$ ($6/9$ of the exposed-subset margins); this diagnostic uses sender identities that receivers do not observe. Reducing retention from 1 to 0.1 raises the network envelopes by about 0.4% (value) and 12.0% (reward), with residual shares 39% and 46% at 0.1, yet mean realized shifts fall from 0.0180 to 0.0158 and from 0.0087 to 0.0079 (about 12% and 9%; ratios of means). These norms retain vector cancellation that the sum-of-norms envelope discards. The last 500 rounds give margins $-0.0011/-0.0027$ and targeting advantages $+0.0167/+0.0064$ (value/reward). Exposure refers to direct in-neighbors.

Clipping and screening ablations. On the default plant, six paired seeds average budgets 50/800 using the MAD screen (Table 4); gains use the fixed-plant reference in Section G.1. The ablations show complementary contributions from screening and clipping. Under boundary corruption, screening adds to interval clipping: against IntervalOnly, PARC's attacked-gain difference is 0.1108 ($[0.1027, 0.1190]$) and its loss reduction is 0.1084 ($[0.1016, 0.1145]$). Under distance-constrained corruption, IntervalOnly accounts for most of PARC's advantage: against RTC, PARC's loss reduction is 0.1415 ($[0.1295, 0.1529]$) and its attacked-gain advantage is 0.1244 ($[0.1130, 0.1354]$), while IntervalOnly alone attains a loss reduction of 0.1397 ($[0.1278, 0.1507]$), leaving a screening increment of 1.77×10^{-3} ($[-3.10, 5.05] \times 10^{-3}$). RTC's mean clean gain is higher by 0.0170 ($[0.0146, 0.0196]$).

Comparison with Function W-MSR. On a distinct ten-plant suite (200 pairs, uniform weights), PARC reduces attack-induced cost-ratio increase by 0.1657 ($[0.1374, 0.2001]$) versus IntervalOnly and 0.0684 ($[0.0534, 0.0821]$) versus Function W-MSR, both positive on all ten plants (Table 10). Its cost-ratio metric is defined in Section G.1.

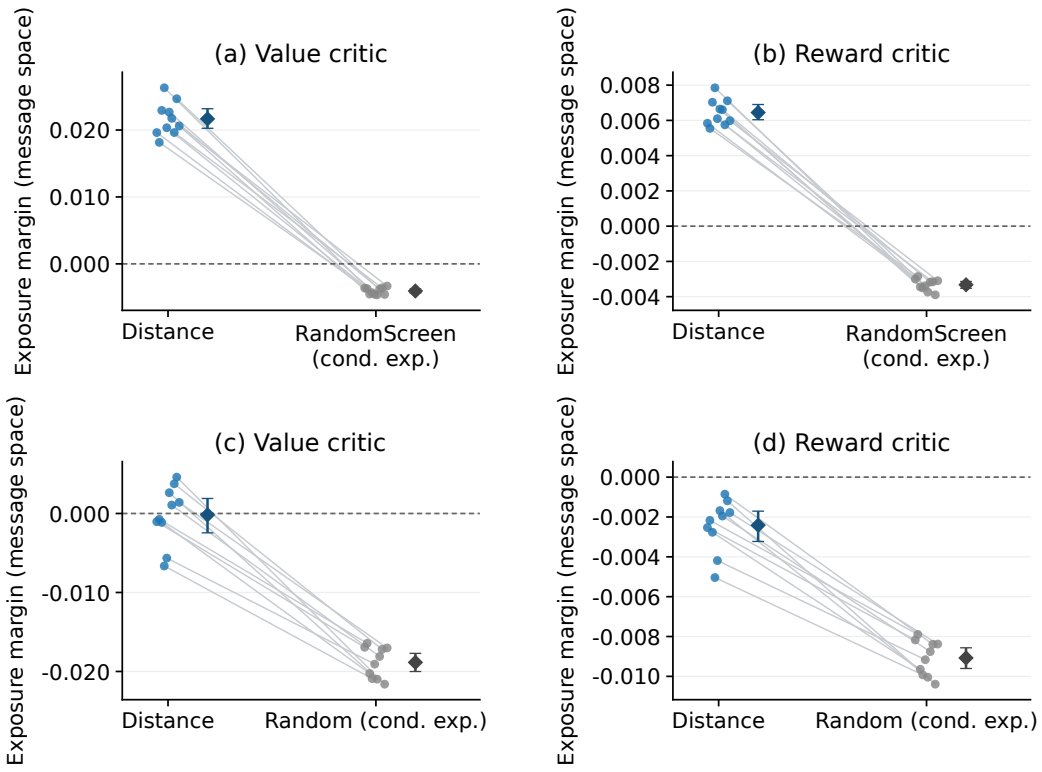


Figure 6: **Same-input exposure margins.** Message-space margins; points are plant means over two seeds, lines pair plants, and diamonds show means with descriptive pointwise 95% plant-bootstrap intervals. Top (a, b): screening on boundary PARC rows, distance versus RandomScreen. Bottom (c, d): retention targets on in-interval $\mathrm{PARC}_{0.1}$ rows at all nine honest receivers, distance versus uniform one-peer targeting. Random entries are analytic conditional expectations ((121) and Proposition 3). Vertical scales differ.

Method	Clean cost ratio	Attack cost ratio	Clean finite	Attack finite
Mean	0.9158	–	20/20	0/20
IntervalOnly	0.9082	1.0882	20/20	20/20
Function W-MSR	0.9056	0.9882	20/20	20/20
PARC	0.9051	0.9194	20/20	20/20

Table 10: **Comparison with Function W-MSR (cost ratio ↓).** PARC's attack-induced increase in cost ratio is smaller than IntervalOnly's and Function W-MSR's on all ten plants. Last-five/first-phase cost over ten paired plants (two seeds each). A dash denotes an entirely nonfinite group.

Attack geometry. Over 160 phases, PARC-minus-RTC all-phase mean gain is -0.0183 ($[-0.0206, -0.0154]$) under boundary corruption and $+0.3072$ ($[0.2943, 0.3190]$) under distance-constrained corruption (Figure 2). The analytic feasible controller uses the true plant, honest-team objective and matched actor-row constraints. As a supplementary two-attack summary, we take each method's lower all-phase mean gain for each paired seed; the PARC-minus-RTC difference averages 0.3043 (95% CI $[0.2928, 0.3149]$).

Reward-budget sensitivity. The 20-phase reward-source study gives boundary-screening loss reductions $0.1263/0.1048/0.1051$ at 50/200/800 preference pairs and 0.1061 with numeric rewards. Within-seed averaging of preference budgets gives 0.1121 ($[0.1096, 0.1151]$).

G.6 In-interval Attacks and Self-retention

All 100 runs of the retention-sensitivity study are finite. The study uses the default plant, ten paired training seeds and the distance screen (Table 4); gains use the fixed-plant reference in Section G.1.

Retention level. Each level applies (8) with the distance screen and $\mathcal{T}_i = \mathcal{S}_i$ (Appendix B.3); raw intervals and topology weights are unchanged, and $\kappa_{\text{keep}} = 1$ is PARC. Table 11 reports all levels with their own clean runs.

κ_{keep} / rule	Clean g	Attack g	Loss D [95% CI]
1 (PARC)	0.11999	0.07140	0.04859 $[0.04601, 0.05141]$
0.5	0.12363	0.09348	0.03014 $[0.02858, 0.03166]$
0.25	0.12509	0.10110	0.02399 $[0.02259, 0.02550]$
0.1	0.12564	0.10381	0.02183 $[0.01948, 0.02422]$
RTC	0.11456	0.09949	0.01507 $[0.01276, 0.01718]$

Table 11: **Reduced retention ($\kappa_{\text{keep}} = 0.1$) improves attacked and clean gain over full retention.** Distance screen, ten paired seeds; mean last-five-phase fixed-reference gains (↑) and own-clean losses (↓) with pointwise 95% bootstrap CIs. RTC is a separately trained reference.

Reduced retention improves both clean and attacked mean gain in this comparison. Relative to $\kappa_{\text{keep}} = 1$, $\kappa_{\text{keep}} = 0.1$ raises attacked gain by 0.0324 ($[0.0303, 0.0346]$), lowers own-clean loss by 0.0268 ($[0.0244, 0.0297]$) and raises clean gain by 0.0056 ($[0.0037, 0.0076]$); these intervals adjust for the nine retention contrasts (Section G.8). Against the separately trained RTC reference, attacked gain is higher by 0.0043 ($[0.0015, 0.0070]$), while own-clean loss is larger by 0.0068 ($[0.0031, 0.0108]$), using the twelve-comparison adjustment.

Certificate. Every level in Table 11 is covered by the positive-retention certificate of Appendix E.1, with coefficient floor $\kappa_{\text{keep}} a_\star / 2$; Appendix E.6 gives the complementary fixed-input comparison. The replays in Appendix G.5 use the main comparison, not this single-plant study.

G.7 VMAS Flocking

Task and learner. VMAS 1.5.2 `flocking` uses six independent agents, a scripted target and five obstacles; one agent corrupts critic messages only. Native utility rewards progress toward the desired 0.1 agent/target distance and penalizes collisions. All methods train for 448 phases. Evaluation averages native 192-step returns over six agents and twelve episodes per seed (three environment seeds, four replicas each).

Each phase collects 24 rollouts, with three cross-fit folds using sixteen replicas for critic fitting and eight for actor updates. Actors use local observations, target and nearest-peer features; critics use global state, plus joint actions and pairwise geometry for the reward critic. Actors have two 64-unit tanh layers; persistent critics have two 128-unit tanh layers. Messages contain 128 realization queries and 1,536 actor-facing predictions. Post-consensus rewards and realized value baselines drive GAE($\lambda = 0.95$, $\gamma = 1$) and PPO.

Local 67D potential/residual kernel reward models use private comparisons: 64/64/128 new pairs of lengths 32/64/192 per phase, split into 192 training and 64 validation pairs. Predictions enter one phase later. After phase 32, the last actor-used model is frozen and labeling stops. All methods use eight unguarded phases, then a fixed local separation guard, also applied to the deterministic evaluation policy.

Distance-constrained attack. Let R_i be receiver i's maximum honest distance from self on the complete joint payload. The attack modifies 16 value realization and 16 actor-facing reward queries, preserving all other predictions. Dynamic supports use sender-minus-honest-median discrepancy for value and lag-one actor-score sensitivity for reward, scaled by channel-wise honest MAD. Value perturbations reverse the centered discrepancy; reward perturbations follow sensitivity toward actions away from the target and nearest peer. The faulty payload obeys

$$\|x'_b - x_i\|_2 \le 0.9 R_i, \qquad \|x'_b - x_b\|_2 \le 0.1 R_i, \qquad \|(x'_b - x_b)^\omega\|_2 \le 0.01 R_i.$$

Value displacement is also bounded by 0.45 times the honest radius on its selected support. Infeasible constructions leave the message unchanged.

Phase	Rule	Clean return	Attack return	Loss D [95% CI]
448	PARC	0.95416	0.96330	$-0.00914\ [-0.03436, 0.01512]$
448	RTC	0.92331	0.83481	$0.08850\ [0.00795, 0.16612]$
448	Mean	0.94541	0.84792	$0.09749\ [0.05812, 0.14067]$
Phase	Effect	Mean [95% CI]		Positive seeds
448	A_{RTC}	$+0.12849\ [0.08143, 0.17308]$		6/6
448	A_{Mean}	$+0.11538\ [0.09121, 0.14040]$		6/6
448	M_{Mean}	$+0.10663\ [0.06906, 0.14311]$		6/6
448	M_{RTC}	$+0.09764\ [0.01667, 0.17623]$		4/6

Table 12: **Flocking at 448 phases: PARC has higher attacked return than RTC and smaller attack-induced loss than Mean and RTC.** Native team returns (↑) and paired effects with descriptive pointwise 95% seed-bootstrap intervals. A is attacked PARC minus comparator return; M is comparator own-clean loss minus PARC loss.

G.8 Statistical Analysis

Bootstrap intervals resample method/condition-paired units in Table 4 after agent/replica averaging. Effects are means (LFA screening: median); nonfinite runs retain failure counts. Reward-fit intervals resample fixed-policy datasets.

C1–C4 average two seeds per plant and report mean paired differences and positive-plant counts; Table 7 gives between-plant sample SD. Matched numeric and preference runs use ten paired seeds and pointwise 95% seed-bootstrap intervals.

Clipping/screening sign-flip and attack-geometry sign tests use Holm correction (adjusted $p = 0.03125$); sign-flip inference assumes symmetry and exchangeability. The two-attack summary is $\Delta_s = \min_a \bar{g}_{\mathrm{PARC},s,a} - \min_a \bar{g}_{\mathrm{RTC},s,a}$, where $\bar{g}$ is the 160-phase mean and a indexes the two prespecified attacks. Its mean and pointwise bootstrap interval use six paired seeds; sign tests rank failures below finite performance and omit ties. Retention sensitivity adjusts nine contrasts against $\kappa_{\mathrm{keep}} = 1$ and twelve against RTC. Flocking pairs six seeds with faulty senders 0–5; phase-448 descriptive, pointwise 95% intervals use 100,000 paired seed resamples.