# WHEN DOES LOCAL CREDIT ASSIGNMENT IMPROVE SHARED POLICY GRADIENTS IN MULTI-AGENT REINFORCEMENT LEARNING?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Credit assignment in multi-agent reinforcement learning is justified agent by agent: a counterfactual or teammate-marginalized critic can reduce the variance of a policy-gradient term. With parameter sharing, the optimizer updates the sum of agent–time contributions, so local variance reduction need not improve the shared gradient. This gap is fundamental. For every team size, we construct bounded games with an exact critic and a nonzero shared gradient in which every locally variance-optimal correction reduces local variance but increases shared-gradient variance by an arbitrarily large factor. For two agents sharing a Bernoulli policy with the value baseline, we identify a sharp regime in which exact teammate marginalization is guaranteed not to increase shared-gradient covariance, and extend it to an exact criterion for $n$ identical agents. These results motivate PLACID, which separates *admission*—whether and at what strength a correction improves the shared-gradient estimate—from *realization*—how to compute it with less Monte Carlo noise. On 16 fixed tabular games, empirical admission removes up to $9.21\%$ of sparse population mean-squared error. On the two fitted critic families, it exceeds unit marginalization at every tested batch budget. Its certified variant selects 30,101 corrections with none increasing population shared-gradient variance. In 1,566 enumerated games, even population-moment local screens can select corrections that harm the shared gradient. In shared-actor training, a lagged empirical admission variant increases normalized return AUC on coverage games by $22.54 \times 10^{-3}$ over the sparse update and exceeds unit marginalization on team and coverage games with one teammate proposal.

## 1 INTRODUCTION

Cooperative multi-agent reinforcement learning (MARL) often learns decentralized policies from a shared team outcome that may arrive only at the end of an episode. Centralized critics, reward redistribution, and counterfactual credit methods address this sparse-feedback problem by assigning more informative credit to individual agent decisions (Foerster et al., 2018; Arjona-Medina et al., 2019; She et al., 2022; Chen et al., 2024; Bui et al., 2025b). A common justification for such methods is local: hold one agent's action fixed, average over its teammates, and show that the resulting credit signal reduces the variance of that agent's policy-gradient contribution (Kuba et al., 2021; Jafferjee et al., 2025). With parameter sharing, however, the optimizer updates a common policy using the sum of many agent–time contributions. The quality of this shared update depends not only on individual variances, but also on how those contributions covary. Thus, improving every agent locally need not improve the gradient that actually updates the shared policy. This raises our central question: *When does a local credit guarantee survive aggregation into the gradient that updates a shared policy?*

We give a negative and an exact positive answer. For every team size, bounded games exist in which exact, locally optimal credit improves every agent's term but makes the shared gradient arbitrarily noisier. In the setting of two agents sharing a Bernoulli policy, a common value baseline, and exact teammate marginalization, we identify the sharp interval around the balanced policy in which the shared gradient does not worsen for every bounded payoff. Outside settings with a comparable structural guarantee, the correction must be evaluated after agent contributions are aggregated.



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.

PLACID follows this logic: *admission* selects a correction from the complete shared gradient; *prospective recycling* retains the selecting prefix as raw gradients and corrects only unseen trajectories; *realization* reduces teammate-sampling noise while preserving proposal dependencies. Our main contributions are:

- **Local-to-shared theory.** We identify omitted cross-agent and cross-time covariance terms, prove an arbitrarily large reversal for every team size, and derive a sharp two-agent Bernoulli regime where local marginalization is uniformly safe, with an exact criterion for a shared Bernoulli policy of any team size. We also construct identical agent-wise gradient laws and prediction errors with opposite effects on the shared gradient.
- **Complete-block admission with prospective reuse.** PLACID decides whether and how strongly to use learned credit from the shared gradient, applies it only to unseen trajectories, and retains the selecting prefix; abstention recovers the all-data sparse update.
- **Lower-noise realization.** Dependency-preserving stratification preserves the admitted correction's mean and cannot increase its conditional covariance relative to iid teammate sampling. Experiments separately measure complete-estimator and proposal-sampling error, and track local-to-shared reversal and admission during shared-actor training.

## 2 Related Work

**Multi-agent credit and shared gradients.** COMA, ROLA, ASAE, PERLA, and GPAE construct focal or teammate-marginal actor signals (Foerster et al., 2018; Xiao et al., 2021; Wan et al., 2021; Jafferjee et al., 2025; Kim et al., 2026). Their local variance analyses motivate our question (Kuba et al., 2021; Jafferjee et al., 2025). Under score orthogonality, local and total baseline optimality can coincide (Wu et al., 2018); HyperMARL instead changes the actor architecture to address cross-agent covariance (Tessera et al., 2025). We keep the shared actor fixed and evaluate the covariance of its complete update. Appendix G gives the detailed comparison.

**Control variates, reuse, and realization.** Outcome-retaining control variates and full-vector weight selection provide related estimator tools (Gu et al., 2017; Huang and Jiang, 2020; Cheng et al., 2020; Geffner and Domke, 2020). Sample splitting, adaptive control variates, and cross-fitting separate selection from evaluation (Kim and Henderson, 2007; Lapeyre and Lelong, 2011; Kota et al., 2017; Chernozhukov et al., 2018). PLACID retains its selecting prefix as raw actor data and applies one frozen correction to an unseen suffix. After admission, it uses stratification to reduce proposal noise without breaking existing proposal dependencies (Etore et al., 2011; Jourdain et al., 2011). Further distinctions and sampling-direction precedents appear in Appendix G.

## 3 Problem Setup: The Shared Gradient

**Cooperative MARL and parameter sharing.** Consider a cooperative partially observed Markov game with agents $\mathcal{N} = \{1, \dots, n\}$, horizon $T$, local records $s_{i,t}$, joint action $\mathbf{a}_t$, and shared return

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\pi}_\theta}}[Y], \qquad Y = \sum\nolimits_{t=0}^{T-1} \gamma^t r_t. \tag{1}$$

We focus on sparse feedback, where $Y$ identifies team performance but not the responsible agent–time decisions. Under centralized training with decentralized execution, the policy factorizes as

$$\boldsymbol{\pi}_\theta(\mathbf{a}_t \mid \mathbf{s}_t) = \prod\nolimits_{i=1}^{n} \pi_{i,\theta}(a_{i,t} \mid s_{i,t}), \tag{2}$$

while the critic may use joint information. Partition shared actor parameters into blocks $\{\theta_b : b \in \mathcal{B}\}$ and let $\mathcal{I}_b$ contain the agents depending on block $\theta_b$.

**The shared policy-gradient object.** For agent $i \in \mathcal{I}_b$, define its *policy score* for block $\theta_b$ as $S_{i,t}^{(b)} = \nabla_{\theta_b} \log \pi_{i,\theta}(a_{i,t} \mid s_{i,t})$. Let the baseline rule be fixed before the trajectory, and let $B_t$ denote its action-independent value at time $t$. The contribution of one trajectory to block $b$ is determined as:

$$X_b = \sum\nolimits_{t=0}^{T-1} \sum\nolimits_{i \in \mathcal{I}_b} S_{i,t}^{(b)} (Y - B_t), \qquad \mu_b = \mathbb{E}_{\tau \sim p_{\boldsymbol{\pi}_\theta}}[X_b]. \tag{3}$$

These are vectors in block-$b$ coordinates; under standard score-function conditions, $\mu_b$ is the policy gradient at the collecting policy. Because shared-agent contributions are summed in the same coordinates, we study the covariance of $X_b$ rather than separate agent-wise variances.

**Teammate-centered credit.** Let $Q(\mathbf{s}_t, \mathbf{a}_t)$ be an auxiliary score fitted from data and frozen before the trajectory on which it is used. It may be an exact critic, but exactness is not required. For focal agent $i$, let $\boldsymbol{\pi}_{-i,\theta}$ denote the joint policy of all other agents and $\widetilde{\mathbf{a}}_{-i,t}$ denote an alternative joint action of agent $i$'s teammates. Holding $a_{i,t}$ fixed, define the teammate-centered correction

$$d_{i,t}[Q] = Q(\mathbf{s}_t, \mathbf{a}_t) - \mathbb{E}_{\widetilde{\mathbf{a}}_{-i,t} \sim \boldsymbol{\pi}_{-i,\theta}}[Q(\mathbf{s}_t, a_{i,t}, \widetilde{\mathbf{a}}_{-i,t})], \tag{4}$$

Thus $d_{i,t}[Q]$ compares the score under the realized teammate actions with its average over alternative ones while keeping the focal action fixed. In practice, this expectation is approximated using $K$ teammate actions $\widetilde{\mathbf{a}}^{(1)}_{-i,t}, \ldots, \widetilde{\mathbf{a}}^{(K)}_{-i,t}$ drawn from the frozen teammate policy: $d^{(K)}_{i,t}[Q] = Q(\mathbf{s}_t, \mathbf{a}_t) - \frac{1}{K}\sum_{k=1}^{K} Q(\mathbf{s}_t, a_{i,t}, \widetilde{\mathbf{a}}^{(k)}_{-i,t})$. For parameter block $b$, we define the corresponding complete correction:

$$H_{K,b}(Q) = \sum\nolimits_{t=0}^{T-1} \sum\nolimits_{i \in \mathcal{I}_b} S^{(b)}_{i,t} d^{(K)}_{i,t}[Q]. \tag{5}$$

For a fixed coefficient $\lambda_b$, the corrected block gradient is defined as follows:

$$G^{\lambda_b}_b = X_b - \lambda_b H_{K,b}(Q). \tag{6}$$

The observed outcome $Y$ remains in the estimator; $Q$ supplies only a control-variate correction.

**Target preservation.** Conditional on $(\mathbf{s}_t, a_{i,t})$, the realized teammate action and each independent teammate proposal follow the same policy distribution. Hence $\mathbb{E}\big[S^{(b)}_{i,t} d^{(K)}_{i,t}[Q]\big] = 0$, It follows that $\mathbb{E}[H_{K,b}(Q)] = 0,\ \mathbb{E}[G^{\lambda_b}_b] = \mathbb{E}[X_b] = \mu_b$, provided that $Q$, the baseline rule, and $\lambda_b$ are fixed before the corrected trajectory and the required moments exist. Theorem 9 gives the full statement and proof. Thus critic misspecification may make a correction ineffective or increase its variance, but it does not itself change the expected policy gradient. The remaining question is whether a correction that improves an individual agent's gradient also improves the *complete shared gradient*.

## 4 When Do Local Guarantees Survive Sharing?

A covariance expansion identifies the cross terms omitted by agent-wise analysis. Theorem 1 shows that they can dominate for any team size, while Theorem 2 gives a sharp payoff-uniform safe interval for a two-agent Bernoulli policy. Outside this regime, the complete correction must be tested.

**Where the local argument loses information.** Index the agent–time terms that update one shared block by $\ell$, and write $X_b = \sum_\ell x_\ell$ and $H_b = \sum_\ell h_\ell$. For a common coefficient $\lambda$, we define the *covariance-trace reduction* $\mathcal{D}_b(\lambda) := \operatorname{tr}\operatorname{Cov}(X_b) - \operatorname{tr}\operatorname{Cov}(X_b - \lambda H_b)$. Thus $\mathcal{D}_b(\lambda) > 0$ means that the correction reduces the total coordinate-wise variance of the shared-block gradient, while $\mathcal{D}_b(\lambda) < 0$ means that it increases it. Since $\mathbb{E}[H_b] = 0$, expanding $X_b$ and $H_b$ gives

$$\mathcal{D}_b(\lambda) = \sum\nolimits_\ell \Big(2\lambda\,\mathbb{E}[x_\ell^\mathsf{T} h_\ell] - \lambda^2 \mathbb{E}\|h_\ell\|^2\Big) + \sum\nolimits_{\ell \neq r} \Big(2\lambda\,\mathbb{E}[x_\ell^\mathsf{T} h_r] - \lambda^2 \mathbb{E}[h_\ell^\mathsf{T} h_r]\Big). \tag{7}$$

Appendix F.1 shows Equation (7). The first sum contains local terms seen by agent-wise analyses; the second has cross-agent and cross-time interactions from aggregation. Disjoint parameter supports remove cross-agent products, but parameter sharing retains them with signs not determined locally.

**One-step shared-Bernoulli setting.** Consider a one-step cooperative game with $n$ agents sharing one logit parameter. Each agent chooses a binary action $a_i \sim \mathrm{Bernoulli}(p)$, so $\Pr(a_i = 1) = p$, and the joint action is $\mathbf{a} = (a_1, \ldots, a_n)$. The corresponding policy score is $S_i = a_i - p$. For a shared outcome $Y$, an action-independent baseline $B$, and an exact critic $Q(\mathbf{a}) = \mathbb{E}[Y \mid \mathbf{a}]$, define agent $i$'s raw gradient contribution and teammate-centered correction by $x_i = S_i(Y - B),\ h_{i,K} = S_i d^{(K)}_i[Q]$. Here $d^{(K)}_i[Q]$ uses $K$ iid joint teammate-action proposals with $a_i$ held fixed; $K = \infty$ denotes exact teammate marginalization. The complete shared gradient and correction are $X = \sum_{i=1}^n x_i,\ H_K = \sum_{i=1}^n h_{i,K}$. Proposal sets corresponding to different focal agents may be arbitrarily coupled.

**Theorem 1** (An obstruction for every team size)**.** *For every $n \geq 2$ and $0 < \sigma \leq 1$, there exists a bounded game in the setting above with $Y \in [0,1]$ and baseline $B = 1/2$ such that $\mathbb{E}[x_i] = \frac{1}{8n}$ for every $i$, $\mathbb{E}[X] = \frac{1}{8}$. Moreover, for every $K \in \mathbb{N} \cup \{\infty\}$, each local variance $\mathrm{Var}(x_i - \lambda h_{i,K})$ is uniquely minimized at $\lambda_K = \frac{K}{K+1}$, $\lambda_\infty = 1$, and strictly decreases there. Nevertheless,*

$$\frac{\mathrm{Var}(X - \lambda_K H_K)}{\mathrm{Var}(X)} > 1 + \frac{1}{16\sigma^2}. \tag{8}$$

*Indeed, every nonzero common coefficient increases the shared variance. The baseline $B = 1/2$ is also the variance-optimal constant baseline for the raw shared gradient.*

The critic is exact and the shared mean nonzero. Conditional expectation reduces variance for each agent, but each conditions on its own action. These different conditioning choices disrupt cancellation in the raw sum. The reversal holds even with exact marginalization and bounded rewards. Appendix A strengthens the result by jointly optimizing local baselines and finite-$K$ coefficients.

**Two-agent shared-Bernoulli setting.** Consider two agents with independent binary actions $a_i \sim \mathrm{Bernoulli}(p)$, where $\Pr(a_i = 1) = p$ and $0 < p < 1$. Both agents use the same logit $\eta(\theta)$, so, writing $j_\eta = \nabla_\theta \eta(\theta) \neq 0$, their policy scores are $S_i = (a_i - p) j_\eta$. Let $Q = \mathbb{E}[Y \mid a_1, a_2]$ be the exact critic and use the common value baseline $B = \mathbb{E}[Q]$. Define the raw local gradients $x_i = S_i(Y - B)$, their exact teammate-centered corrections $h_{i,\infty} = S_i d_i[Q]$, and $X = x_1 + x_2$, $H_\infty = h_{1,\infty} + h_{2,\infty}$.

**Theorem 2** (A sharp payoff-uniform transfer window)**.** *Exact teammate marginalization with coefficient $\lambda = 1$ satisfies $\mathrm{Cov}(X - H_\infty) \preceq \mathrm{Cov}(X)$ for every bounded two-agent payoff $Q$ iff*

$$(1-2p)^2 \leq \frac{1}{11}, \qquad \textit{equivalently} \qquad p \in \left[\frac{1 - 1/\sqrt{11}}{2}, \frac{1 + 1/\sqrt{11}}{2}\right]. \tag{9}$$

*Here $\preceq$ denotes the positive-semidefinite (Loewner) order. Outside this interval, some bounded payoff has lower local variances after exact marginalization but higher shared-gradient covariance.*

Here *payoff-uniform* means the guarantee holds for every bounded joint-action payoff under the stated policy and baseline. With $z_i = a_i - p$, expanding any payoff in the basis $1, z_1, z_2, z_1 z_2$ reduces the covariance difference to a quadratic form that is nonnegative for every payoff exactly when $(1-2p)^2 \leq 1/11$; Appendix B gives the derivation. Figure 1 visualizes both the all-team obstruction and this sharp window. For $n$ identical agents, two binomial-moment inequalities give the exact payoff-uniform criterion (Proposition 8); sampling uniform $p$ estimates an unsafe share above one half for each $n = 2, \ldots, 128$. The safe window assumes exact marginalization; finite teammate sampling can reverse it even inside the window (Appendix B). A harmful local coefficient also need not make the correction direction useless: Appendix C shows it can help at a smaller complete-block coefficient.

**Proposition 3** (Agent-wise information is insufficient)**.** *There are two fixed credit functions on the same bounded cooperative game and factorized shared policy such that, for every agent $i$, the pair $(x_i, h_i)$ has the same distribution under the two credit functions. The functions also have the same outcome-prediction mean-squared error (MSE). Nevertheless, at $\lambda = 1$, one has positive complete-block covariance reduction and the other has negative reduction. Thus agent-wise laws and prediction error do not determine whether a correction improves the complete shared gradient.*

The proof (Appendix F.1) changes only cross-agent coupling. Hence, admission must test the aggregated correction; teammate-expectation approximation is a separate sampling problem. In 1,566 enumerated games, an agent-diagonal population screen deploys 35 harmful corrections under full sharing (up to $17.56\%$ more shared covariance); the complete-block screen deploys none (Appendix H.1).

## 5 PLACID: Admit First, Realize Second

PLACID selects a correction from the complete shared gradient, retains the selecting prefix, and refines its teammate sampling.

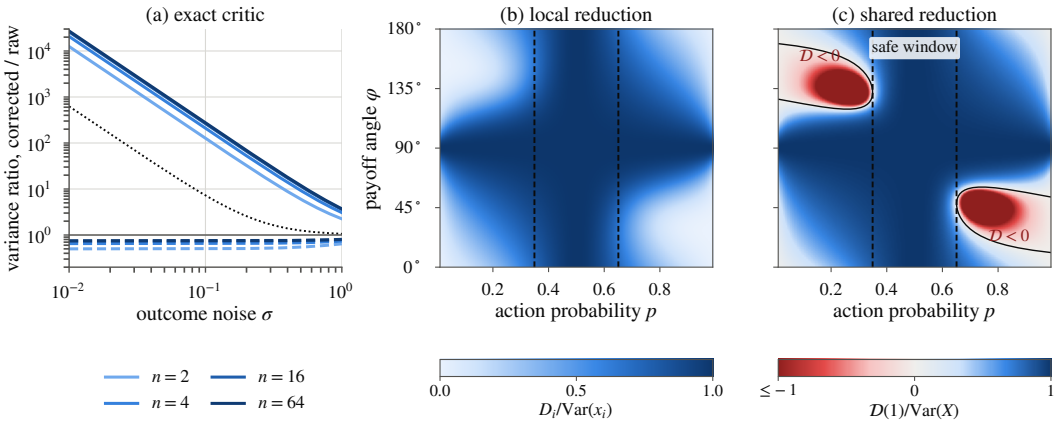


Figure 1: Local guarantees versus the shared gradient. (a) In Theorem 1 with exact critic and $\lambda_\infty = 1$, each agent's term (dashed) improves, while the shared gradient (solid) becomes arbitrarily noisier as $\sigma$ decreases; curves for $n \in \{2, 4, 16, 64\}$ nearly coincide and remain above the dotted bound in Equation (8). (b, c) For Theorem 2, let $Q - B = \frac{s}{2}(z_1 + z_2) + tz_1z_2$ with $(s, t) = (\cos\varphi, \sin\varphi)$. The agent-wise reduction in (b) is positive everywhere, whereas the shared reduction in (c) changes sign (black contour) exactly outside $(1 - 2p)^2 \le 1/11$ (dashed lines); negative values are clipped at $-1$.

### 5.1 Complete-block admission criterion

Suppress the block index, let $X$ be its raw gradient, and let $H_K = H_K(Q)$ be a candidate correction from a fitted, held-fixed credit function $Q$. For a candidate coefficient $\lambda$, define the population reduction in covariance trace by

$$\mathcal{D}(\lambda) = \operatorname{tr}\operatorname{Cov}(X) - \operatorname{tr}\operatorname{Cov}(X - \lambda H_K) = 2\lambda\,\mathbb{E}[X^\top H_K] - \lambda^2\mathbb{E}\|H_K\|^2, \tag{10}$$

where the second equality uses $\mathbb{E}[H_K] = 0$. Equation (7) separates the local and cross terms in this criterion.

Writing $\chi = \mathbb{E}[X^\top H_K]$, $\omega = \mathbb{E}\|H_K\|^2$, gives $\mathcal{D}(\lambda) = 2\lambda\chi - \lambda^2\omega$. When $\omega > 0$, the unconstrained coefficient that maximizes the reduction is $\lambda^* = \frac{\chi}{\omega}$. Figure 2(a) illustrates how the useful shared coefficient can differ in size and sign from the local optimum (Appendix C). The population quantity $\mathcal{D}(\lambda)$ is unknown. For an observed data cluster $j$, let $X_j$ and $H_j$ be the corresponding complete raw gradient and correction, and define

$$Z_j(\lambda) = 2\lambda X_j^\top H_j - \lambda^2\|H_j\|^2 = \|X_j\|^2 - \|X_j - \lambda H_j\|^2. \tag{11}$$

If $Q$ and $\lambda$ are held fixed independently of cluster $j$, then $\mathbb{E}[Z_j(\lambda)] = \mathcal{D}(\lambda)$. The gate therefore estimates $\mathcal{D}(\lambda)$ from complete-block observations $Z_j(\lambda)$. Proposition 3 shows why agent-wise statistics and prediction error do not suffice.

### 5.2 Prospective recycling

The admission criterion is estimated from data, but reusing the same observations for selection and correction would mix selection with evaluation; discarding them would waste gradient information. PLACID instead selects from an initial prefix, applies the correction only to unseen data, and retains the prefix as ordinary uncorrected gradients. Dependency clusters serve as independent collection units under pre-batch conditioning (Appendix I). For $N$ ordered clusters, the first $m_0$ fit the held-fixed $Q$; later prefix clusters provide $Z_j(\lambda)$.

The admission procedure may stop at a data-dependent but bounded index $M$. At that point, the selected coefficient $\lambda_M$ is held fixed and is not changed using the remaining clusters. Clusters $j \le M$ retain their ordinary raw contributions $X_j$, while only the previously unseen suffix clusters $j > M$ receive the selected correction. The resulting *prospective recycled estimator* is

$$\widehat{G}^{\mathrm{R}} = \frac{1}{N}\Big\{\sum\nolimits_{j\le M} X_j + \sum\nolimits_{j>M}\big(X_j - \lambda_M H_j(Q_M)\big)\Big\}. \tag{12}$$

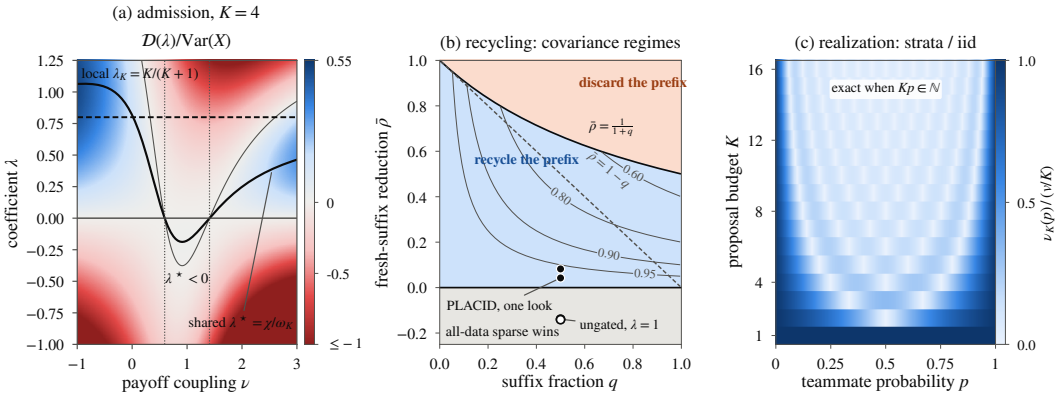


Figure 2: The three stages of PLACID in exactly solvable cases. (a) Admission: complete-block variance reduction for the Gaussian family $Q_\nu$ with $K = 4$. The locally optimal coefficient $K/(K+1)$ (dashed) is often harmful, whereas the complete-block optimum $\lambda^\star = \chi/\omega_K$ (solid) reduces variance whenever $\chi \neq 0$ and is negative for $|\nu - 1| < 1/\sqrt{6}$. (b) Prospective recycling: the lowest-covariance estimator as a function of corrected fraction $q$ and fresh-suffix reduction $\bar{\rho}$; contours show the recycling ratio $1 - q\bar{\rho}$, with confirmation operating points at $q = 1/2$. (c) Realization: variance of equal-mass stratification relative to $K$ iid samples for one Bernoulli$(p)$ teammate, $\nu_K(p)/(v/K)$, where $v = p(1-p)$; it never exceeds one and vanishes when $Kp$ is an integer.

Here $Q_M = Q$ is fitted on the model prefix and held fixed. If no correction is selected, $\lambda_M = 0$ and $\widehat{G}^{\mathrm{R}} = \widehat{G}^0 := \frac{1}{N}\sum_{j=1}^{N} X_j$, so abstention recovers the ordinary all-data sparse estimator exactly. To quantify the effect of recycling, let $\mathcal{D}_M(\lambda)$ be the population covariance-trace reduction that the selected credit function would achieve, with coefficient $\lambda$, on one fresh suffix cluster.

**Proposition 4** (Exact prospective accounting)**.** *Condition on all pre-batch choices, including the behavior policy, baseline, parameter blocks, and admission procedure. Suppose the $N$ clusters are iid, $\mathrm{tr}\,\mathrm{Cov}(X) > 0$, and raw gradients have finite second moments. Let $M$ be a bounded prefix stopping time, with $Q_M$ and $\lambda_M$ determined from the stopped prefix and fixed on the unseen suffix. Generate fresh proposal streams independently across suffix clusters, and assume $\mathbb{E}\|\lambda_M H_j(Q_M)\|^2 < \infty$, $\mathbb{E}[(N-M)|\mathcal{D}_M(\lambda_M)|] < \infty$. Then Equation (12) is unbiased and*

$$\frac{\mathrm{tr}\,\mathrm{Cov}(\widehat{G}^{\mathrm{R}})}{\mathrm{tr}\,\mathrm{Cov}(\widehat{G}^0)} = 1 - \mathbb{E}\left[\left(1 - \frac{M}{N}\right)\rho_M\right], \qquad \rho_M = \frac{\mathcal{D}_M(\lambda_M)}{\mathrm{tr}\,\mathrm{Cov}(X)}. \tag{13}$$

Eq. (13) separates correction quality from exposure: $\rho_M$ is the fresh-suffix reduction, and $1 - M/N$ the corrected fraction. For fixed $q = (N-M)/N$, recycling improves on the sparse estimator whenever the average suffix reduction is positive; discarding the prefix requires it to exceed $1 - q$. Appendix F.4 derives these ratios, and Figure 2(b) shows the regimes. The confirmation study uses $q = 1/2$ in the recycling-favorable region. We next describe PLACID's finite-prefix selection rule.

### 5.3 Finite-sample admission and actor update

Before screening, PLACID fixes a finite coefficient grid $\Lambda$, a finite look schedule $\mathcal{T}$, a per-refresh error level $\delta_r$, and valid support bounds for every $Z_j(\lambda)$. At each scheduled index $m$, it forms simultaneous empirical-Bernstein (EB) lower confidence bounds $L_m^{\mathrm{EB}}(\lambda)$ from the screening-cluster means, variances, and fixed support ranges. The multiplicity penalty covers every pair in $\mathcal{T} \times \Lambda$. The exact bound is given in Appendix F.5.

PLACID stops at the first look with $\max_{\lambda\in\Lambda} L_m^{\mathrm{EB}}(\lambda) > 0$, sets $M = m$, and freezes the maximizing $\lambda_M$ on the unseen suffix. On the simultaneous coverage event, the selected coefficient has positive complete-block reduction. If no bound crosses zero, PLACID sets $M = N$ and $\lambda_M = 0$, recovering the all-data sparse estimator exactly. Appendix F.5 gives the formal guarantee and a sub-gamma alternative. The overall estimator effect is governed by $\mathbb{E}[(1 - M/N)\rho_M]$ in Proposition 4.

**Two admission rules.** We call the certified rule above PLACID-C. A second rule, PLACID-E, reads the same prefix statistics at a prespecified index $m < N$ but replaces the confidence bound by the sample mean. It selects the $\lambda \in \Lambda \cup \{0\}$ that maximizes $\bar{Z}_m(\lambda)$ over the screening prefix, breaking ties by the smallest coefficient, and sets $M = m$ if the maximum is positive. Otherwise it sets $M = N$ and $\lambda_M = 0$. The experiments use $m = N/2$. PLACID-C controls the probability of harmful admission at level $\delta_r$; PLACID-E selects by empirical reduction. Both rules follow the information contract of Section 5.2: the prefix is retained as raw data, the selected coefficient is frozen on the unseen suffix, abstention recovers the sparse estimator exactly, and Proposition 4 applies to both. Either rule accepts any prespecified finite grid. Because the complete-block optimum $\lambda^\star = \chi/\omega$ can be negative (Appendix C), we also use the mirrored grid $\Lambda \cup (-\Lambda)$. The support construction in Appendix F.5 applies to either sign. The simultaneous coverage level remains $1 - \delta_r$, with the multiplicity penalty recomputed for all ten nonzero coefficients in the mirrored grid.

**Actor update.** Eq. (12) is used for the full-batch actor step at the behavior policy. The baseline, credit function, coefficient, and teammate samples are detached; only the actor likelihood is differentiated. A clipped PPO loss has the same derivative at likelihood ratio one. Later PPO epochs use the sparse objective because the actor has moved from the policy where the guarantee holds. Appendix I gives the surrogate, preconditioned extension, and full algorithm.

### 5.4 Realization: reducing teammate-sampling noise

With admission fixed, finite-$K$ teammate sampling still adds noise to $H_K(Q)$. Realization reduces this noise at the same query budget while preserving the selected correction's mean.

**Which correction terms must be sampled together?** Each summand $S_{i,t}^{(b)} d_{i,t}^{(K)}[Q]$ in Equation (5) is a correction term. Terms sharing primitive teammate-proposal randomness form one *proposal component*; only components with independent inputs may be sampled separately. This grouping follows sampling dependence, not agent identity: splitting a shared draw can destroy useful covariance cancellation. Under common-joint proposals, the two focal terms for two agents use disjoint teammate coordinates, whereas focal terms for larger teams generally overlap. Appendix D specifies the dependency graph and gives a cancellation example.

**How is teammate sampling stratified?** Let $\mathcal{R}$ contain the recorded trajectories, fitted score, admitted coefficient, coupling rule, and partition, all fixed before production proposals. For component $g$, write $F_g(U_g)$ for its vector contribution and let iid realization average $K_g$ independent inputs. Stratification partitions the conditional law of $U_g$ into $K_g$ disjoint cells, each of probability $1/K_g$ given $\mathcal{R}$, draws once independently from each cell, and averages the $K_g$ contributions. Thus query count and the component's mean are unchanged; only the allocation of proposal draws changes.

**Theorem 5** (Conditional refinements compose under sharing)**.** *Condition on $\mathcal{R}$ and let the realized gradient be an affine combination of component averages with $\mathcal{R}$-measurable weights. Assume independent component streams and finite second moments. Replacing iid draws by the equal-probability stratification above, while preserving each component's joint input, preserves the conditional mean and gives* $\mathrm{Cov}(G_{\mathrm{strat}} \mid \mathcal{R}) \preceq \mathrm{Cov}(G_{\mathrm{iid}} \mid \mathcal{R})$. *The same order holds unconditionally and for the second error-moment matrix around any $\mathcal{R}$-measurable target with finite second moments.*

This compares two realizations of the *same admitted correction*, not a new admission test; Appendix D gives the proof. For neural actors, PLACID stratifies a one-dimensional latent projection chosen by the complete correction's sensitivity (Appendices E, N). Figure 2(c) shows that for one Bernoulli teammate, equal-mass strata never increase iid proposal variance and remove it when $Kp$ is an integer (Proposition 7).

## 6 Experiments

We separate the two effects targeted by PLACID. First, we ask whether *admission and prospective reuse* improve the complete shared-gradient estimator. Second, after a correction is fixed, we ask whether *realization* reduces the Monte Carlo noise from using finitely many teammate-action samples. Third, we train shared actors and measure reversal and learning.

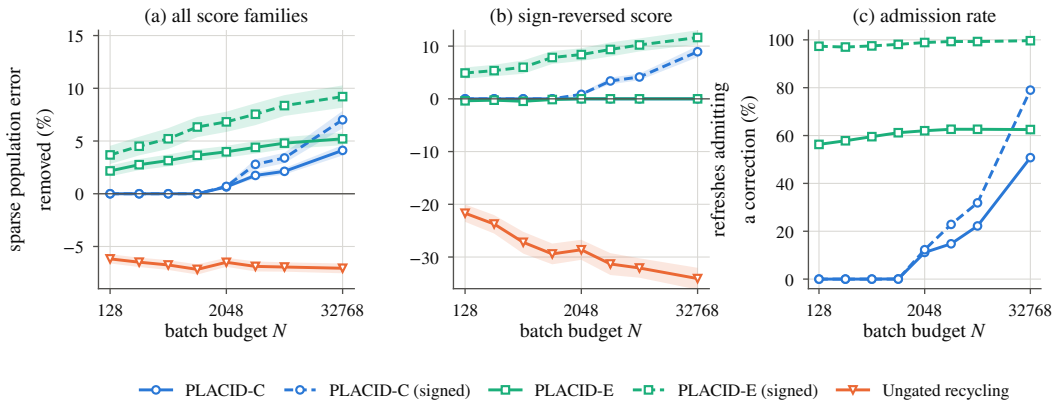


Figure 3: Complete-estimator admission on the 16-game confirmation suite (three credit-function families, 256 repetitions per budget). Positive values in (a,b) indicate lower shared-gradient MSE than sparse. Solid curves use the positive coefficient grid; dashed curves allow both signs. (a) Sparse population error removed across batch sizes. (b) For the sign-reversed credit family, positive-only PLACID-C falls back to sparse, whereas signed rules use the reversed direction. (c) Admission frequency. Across both grids, PLACID-C admits 30,101 corrections. Shading gives paired pointwise 95% bootstrap intervals over batches within the fixed games.

### 6.1 Complete-estimator accuracy

We use 16 exactly enumerable tabular games with two or three agents, so the true population shared gradient and the population effect of every selected correction are known. Each replicate draws a fresh batch, fits a credit function on the model prefix, selects a coefficient on the screening prefix, and applies the frozen correction only to the unseen suffix; the prefix remains in the estimator through its ordinary sparse-gradient contributions. All methods use the same trajectories, actors, baselines, and credit-function families. The families are a fitted joint table, an additive fit, and the sign-reversed joint table. Exact teammate marginalization removes finite-$K$ proposal noise, so this experiment isolates admission and data reuse. For an estimate $\widehat{G}$ of the exact population gradient $\mu$, we measure $\|\widehat{G} - \mu\|^2$ and normalize by $\mathcal{V}_{\mathrm{raw}}/N$, the population MSE of the all-data sparse estimator. We report *sparse population error removed*: positive values mean lower complete-gradient MSE than sparse, and negative values mean higher MSE. PLACID-C uses the confidence-bound admission rule; PLACID-E selects by empirical reduction. The positive grid allows $\lambda \geq 0$, whereas the signed grid also allows negative coefficients.

At the locked budgets $N = 8192$ and 32768, positive-grid PLACID-C removes $2.12\%$ and $4.11\%$ of sparse population error, while PLACID-E removes $4.80\%$ and $5.21\%$ on the same batches. Across $N = 128$–$32{,}768$, signed-grid PLACID-E improves from $3.68\%$ to $9.21\%$, and signed-grid PLACID-C reaches $7.01\%$ at the largest budget. Across both grids, PLACID-C makes 30,101 admissions with no negative population reduction. With the two fitted critics, ungated recycling removes $1.58$–$6.45\%$, and signed PLACID-E removes $1.30$–$1.95$ percentage points more at every budget (paired 95% intervals exclude zero, Table 10); on the sign-reversed control, ungated correction increases error by $21.69$–$34.09\%$. The unbounded plug-in coefficient $\hat{\chi}/\hat{\omega}$ reaches $9.30\%$ at $N = 32{,}768$ but increases error by $99.25\%$ at $N = 128$. Signed one-standard-error shrinkage removes $9.17\%$ at the larger budget and increases error by $1.22\%$ at the smaller one. Empirical grid selection reduces error at every tested budget. Appendix J.2 reports intervals, cross-fitting and plug-in controls, all budgets, and costs.

### 6.2 Conditional proposal-sampling error

We freeze the actor, credit function, and trajectory and resample only the teammate actions, measuring the proposal noise of Theorem 5 on 12 locked native trajectories of SMAC 8m (eight masked discrete-action agents) and MA-MuJoCo CoupledHalfCheetah (two continuous-action agents) (Samvelyan

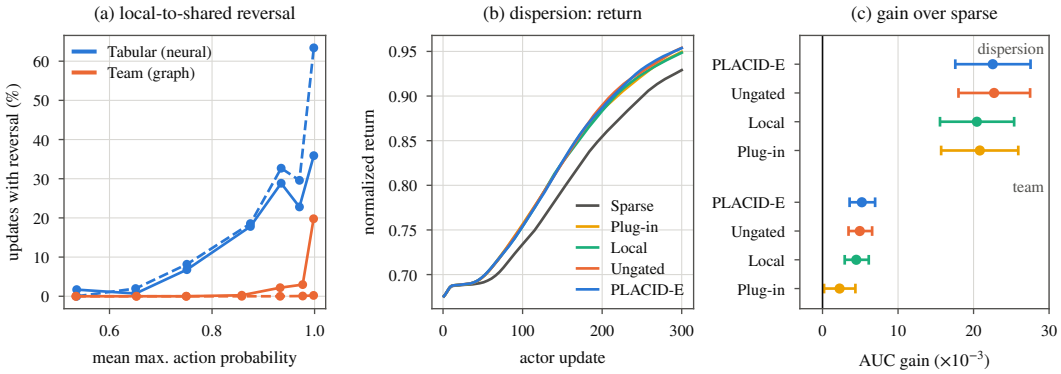


Figure 4: Shared-actor training. (a) Fraction of sparse-training updates whose locally optimal corrections increase shared-gradient covariance trace, by mean maximum action probability; solid/dashed lines use fitted/exact critics. (b) Mean normalized return on dispersion games (384 runs per method). (c) Return-AUC gains over sparse with paired pointwise 95% bootstrap intervals over complete runs within fixed games. Panels (b,c) use graph critics and lagged selection.

et al., 2019; de Witt et al., 2020), with $K = 4$ proposals and unit coefficient. We report conditional proposal covariance removed relative to joint iid with $K = 4$.

At $K = 4$, focal stratification removes $82.43\%$ $[81.76, 83.09]$ of proposal variance relative to joint iid on CoupledHalfCheetah, and $25.09\%$ $[14.58, 34.58]$ on the corrected-observation SMAC panel (Appendix N, Figure 7).

### 6.3 Local-to-shared reversal during training

We train a shared two-layer tanh actor with agent-identity inputs and fit a joint critic from past data. We use the 16 tabular games of Section 6.1, 12 one-step team games with $n \in \{4, 8, 12\}$, and 12 dispersion games in which $n \in \{3, \ldots, 6\}$ agents are paid for weighted coverage of three or four resources. In this lagged PLACID-E variant, the previous $N = 64$ batch selects a coefficient that corrects every episode of the fresh current batch, using exact teammate marginals. We compare it with unit marginalization (ungated), per-agent grid admission (local), and plug-in $\hat{\chi}/\hat{\omega}$ over 32 paired runs per game (Appendix K).

Along 153,600 sparse-training updates on the tabular suite, locally optimal credit increases shared-gradient covariance trace in $26.04\%$ of updates with the neural critic and $38.06\%$ with the exact critic (Figure 4a). For the neural critic, the 99th percentile of these increases is $9.80\%$ of raw covariance trace, and $1.43\%$ of all updates exceed $1\%$ (Table 16). With the graph critic, lagged PLACID-E increases normalized return AUC over sparse by $22.54 \times 10^{-3}$ $[17.58, 27.55]$ on dispersion games and $5.20 \times 10^{-3}$ $[3.59, 6.96]$ on team games (Figure 4b,c). Table 13 gives all paired comparisons with ungated, local and plug-in corrections for both critic architectures. With one sampled teammate proposal, PLACID-E exceeds ungated correction by $9.08$ $[4.73, 13.38]$ and $1.66$ $[0.35, 3.00]$ $(\times 10^{-3})$ on dispersion and team games (Appendix K.1).

## 7 Conclusion

Locally optimal credit can make a shared gradient arbitrarily noisier for every team size, even with bounded rewards and an exact critic, because reducing each term can remove cancellation in their sum. The sharp two-agent Bernoulli window identifies when exact marginalization remains safe, and an exact criterion extends it to $n$ agents. PLACID therefore evaluates the complete shared gradient before choosing a correction and recovers the sparse estimator exactly on abstention. It removes up to $9.21\%$ of sparse population error on the fixed tabular suite and records no harmful selection among 30,101 certified admissions. Population-moment comparisons isolate the cross terms missed by local screens; training improves over sparse and, with one teammate proposal, over unit marginalization on team and coverage games.

## STATEMENT OF USE OF AI

The research questions, the main ideas, the methodology, and the experimental protocols of this paper were chosen by the authors, who divided the work among themselves, supervised every stage of the project, and made all decisions on claims and conclusions. Under the authors' direction, we used generative AI tools as assistants for the following tasks with required disclosure: refining the conceptual framing, discussing candidate mathematical claims, designing experiments and ablations, and implementing the methods and experiments. For the theoretical component, generative AI was also used in part as a "theory attacker": it assisted in constructing and computationally testing large collections of examples with the aim of identifying potential counterexamples, edge cases, or weaknesses in candidate theorem statements and proof arguments. These searches were used as a stress-testing mechanism to help the authors examine and refine the theoretical claims. The absence of a counterexample in such computational searches does not constitute a proof, and no theorem in this paper is considered established on the basis of AI-assisted simulation, search, or testing.

The authors independently verified the theorem statements and proofs, citations, experimental specifications, and AI-assisted text against their research artifacts. No experimental data or results were produced by generative AI tools themselves: every reported number comes from experiments that were actually executed, and translation and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools for literature discovery and for drafting and editing the manuscript, including checks of notation and of the consistency of reported numbers. The authors take full responsibility for every claim, proof, citation, and experimental result in this paper; no AI system is an author.

## REPRODUCIBILITY STATEMENT

Appendices I and J specify the algorithm, its implementation, and the protocols of all experiments. The supplementary material provides the source code and configurations for all experiments, together with summaries of the recorded outputs.

## REFERENCES


Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

The Viet Bui, Tien Mai, and Thanh Nguyen. Inverse factorized soft Q-learning for cooperative multi-agent imitation learning. *Advances in Neural Information Processing Systems*, 37:27178–27206, 2024.

The Viet Bui, Thanh Nguyen, and Tien Mai. ComaDICE: Offline cooperative multi-agent reinforcement learning with stationary distribution shift regularization. *International Conference on Learning Representations*, 2025.

The Viet Bui, Tien Mai, and Hong Thanh Nguyen. Preference-guided learning for sparse-reward multi-agent reinforcement learning. arXiv:2509.21828, 2025.

The Viet Bui, Tien Anh Mai, and Thanh Hong Nguyen. O-MAPL: Offline multi-agent preference learning. *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267:5811–5841, 2025.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

Sirui Chen, Zhaowei Zhang, Yaodong Yang, and Yali Du. STAS: Spatial-temporal return decomposition for solving sparse rewards problems in multi-agent reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:17337–17345, 2024.

Ching-An Cheng, Xinyan Yan, and Byron Boots. Trajectory-wise control variates for variance reduction in policy gradient methods. *Proceedings of the Conference on Robot Learning*, PMLR 100:1379–1394, 2020.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Thomas E. Coons, Aniket Jivani, and Xun Huan. Bayesian covariance uncertainty for adaptive pilot-sampling termination in multi-fidelity uncertainty quantification. arXiv:2508.18490, 2025.

Christian Schroeder de Witt, Bei Peng, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. Deep multi-agent reinforcement learning for decentralized continuous cooperative control. arXiv:2003.06709, 2020.

Pierre Etore, Gersende Fort, Benjamin Jourdain, and Eric Moulines. On adaptive stratification. *Annals of Operations Research*, 189(1):127–154, 2011.

Chengjun Pan, Shichun Liu, Jiahang Lin, Dingwei Zhu, Jiazheng Zhang, Shihan Dou, Songyang Gao, Zhenhua Han, Binghai Wang, Rui Zheng, Xuanjing Huang, Tao Gui, and Yansong Feng. EVPO: Explained variance policy optimization for adaptive critic utilization in LLM post-training. arXiv:2604.19485, 2026.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, 2018.

Tomas Geffner and Justin Domke. A rule for gradient estimator selection, with an application to variational inference. *Proceedings of AISTATS*, PMLR 108:1803–1812, 2020.

Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E. Turner, and Sergey Levine. Q-Prop: Sample-efficient policy gradient with an off-policy critic. *International Conference on Learning Representations*, 2017.

Jiawei Huang and Nan Jiang. From importance sampling to doubly robust policy gradient. *Proceedings of ICML*, PMLR 119:4434–4443, 2020.

Siyi Hu, Chuanlong Xie, Xiaodan Liang, and Xiaojun Chang. Policy diagnosis via measuring role diversity in cooperative multi-agent RL. *Proceedings of ICML*, PMLR 162:9041–9071, 2022.

Taher Jafferjee, Juliusz Ziomek, Tianpei Yang, Zipeng Dai, Jianhong Wang, Matthew E. Taylor, Kun Shao, Jun Wang, and David Mguni. Taming multi-agent reinforcement learning with estimator variance reduction. *Proceedings of AAMAS*, 1042–1050, 2025.

Martin Jankowiak and Theofanis Karaletsos. Pathwise derivatives for multivariate distributions. *Proceedings of AISTATS*, PMLR 89:333–342, 2019.

Benjamin Jourdain, Bernard Lapeyre, and Piergiacomo Sabino. Convenient multiple directions of stratification. *International Journal of Theoretical and Applied Finance*, 14(6):867–897, 2011.

Nathan Kallus and Masatoshi Uehara. Statistically efficient off-policy policy gradients. *Proceedings of ICML*, PMLR 119:5089–5100, 2020.

Sujin Kim and Shane G. Henderson. Adaptive control variates for finite-horizon simulation. *Mathematics of Operations Research*, 32(3):508–527, 2007.

Seongmin Kim, Giseung Park, Woojun Kim, Jiwon Jeon, Seungyul Han, and Youngchul Sung. Generalized per-agent advantage estimation for multi-agent policy optimization. *Proceedings of AAMAS*, 2895–2903, 2026.

Nithyanand Kota, Abhishek Mishra, Sunil Srinivasa, Xi Chen, and Pieter Abbeel. A K-fold method for baseline estimation in policy gradient algorithms. arXiv:1701.00867, 2017.

Jakub Grudzien Kuba, Muning Wen, Linghui Meng, Shangding Gu, Haifeng Zhang, David Mguni, Jun Wang, and Yaodong Yang. Settling the variance of multi-agent policy gradients. *Advances in Neural Information Processing Systems*, 34, 2021.

Bernard Lapeyre and Jérôme Lelong. A framework for adaptive Monte Carlo procedures. *Monte Carlo Methods and Applications*, 17(1):77–98, 2011.

Yueheng Li, Guangming Xie, and Zongqing Lu. Difference advantage estimation for multi-agent policy gradients. *Proceedings of ICML*, PMLR 162:13066–13085, 2022.

Iou-Jen Liu, Raymond A. Yeh, and Alexander G. Schwing. PIC: Permutation invariant critic for multi-agent deep reinforcement learning. *Proceedings of the Conference on Robot Learning*, PMLR 100:590–602, 2020.

Xinjie Liu, Cyrus Neary, Kushagra Gupta, Wesley A. Suttle, Christian Ellis, Ufuk Topcu, and David Fridovich-Keil. A multi-fidelity control variate approach for policy gradient estimation. *Transactions on Machine Learning Research*, 2026.

Xueguang Lyu, Andrea Baisero, Yuchen Xiao, Brett Daley, and Christopher Amato. On centralized critics in multi-agent reinforcement learning. *Journal of Artificial Intelligence Research*, 77:295–354, 2023.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. *Proceedings of COLT*, 2009.

Roman Novak, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Fast finite width neural tangent kernel. *Proceedings of ICML*, PMLR 162:17018–17044, 2022.

Chris J. Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B*, 79(3):695–718, 2017.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. The StarCraft multi-agent challenge. *Proceedings of AAMAS*, 2186–2188, 2019.

Jennifer She, Jayesh K. Gupta, and Mykel J. Kochenderfer. Agent-time attention for sparse rewards multi-agent reinforcement learning. *Proceedings of AAMAS*, 2022.

Kale-ab Tessera, Muhammad Arrasy Rahman, Amos Storkey, and Stefano Albrecht. HyperMARL: Adaptive hypernetworks for multi-agent RL. *Advances in Neural Information Processing Systems*, 38:114352–114398, 2025.

Brendan D. Tracey and David H. Wolpert. Reducing the error of Monte Carlo algorithms by learning control variates. arXiv:1606.02261, 2016.

George Tucker, Surya Bhupatiraju, Shixiang Gu, Richard E. Turner, Zoubin Ghahramani, and Sergey Levine. The mirage of action-dependent baselines in reinforcement learning. *Proceedings of ICML*, PMLR 80:5015–5024, 2018.

Lipeng Wan, Xuwei Song, Xuguang Lan, and Nanning Zheng. Multi-agent policy optimization with approximatively synchronous advantage estimation. arXiv:2012.03488, 2021.

Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre M. Bayen, Sham Kakade, Igor Mordatch, and Pieter Abbeel. Variance reduction for policy gradient with action-dependent factorized baselines. *International Conference on Learning Representations*, 2018.

Jiezhong Wu and Reiichiro Kawai. Stopping rules for Monte Carlo methods of martingale difference type. *SIAM Journal on Scientific Computing*, 48(4):A1979–A2001, 2026.

Yuchen Xiao, Xueguang Lyu, and Christopher Amato. Local advantage actor-critic for robust multi-agent deep reinforcement learning. *International Symposium on Multi-Robot and Multi-Agent Systems*, 155–163, 2021.

Olivier Zahm, Paul G. Constantine, Clémentine Prieur, and Youssef M. Marzouk. Gradient-based dimension reduction of multivariate vector-valued functions. *SIAM Journal on Scientific Computing*, 42(1):A534–A558, 2020.

Zijian Zhao and Sen Li. Aggregate in the advantage, not the ratio: A canonical-form analysis of cooperative multi-agent policy optimization. arXiv:2607.17924, 2026.

## APPENDIX ROADMAP

Appendices A–C prove the local-to-shared results and give the Gaussian coefficient example. Appendices D–F establish the sampling and estimator guarantees, including target preservation, prospective recycling, and the scheduled-look gate. Appendices G–I provide detailed positioning, exact checks, and the implementation contract. The remaining appendices specify the primary complete-estimator and proposal-realization experiments and summarize the secondary scope checks.

## A A BOUNDED OBSTRUCTION FOR EVERY TEAM SIZE

**Purpose and proof idea.** This appendix proves Theorem 1. The construction is designed so that the *raw shared gradient* contains a cancellation that keeps its variance small, while each agent's locally optimal teammate-marginal correction removes a different piece of that cancellation. The proof therefore has two tasks: keep the reward and policy-gradient target well behaved for every team size, and show that local variance reduction coexists with arbitrarily worse shared variance. We first build the bounded reciprocal-payoff family, then compute the local optima and the resulting complete-block variance. A final proposition strengthens the message by allowing each agent to jointly optimize its own constant baseline and finite-$K$ coefficient.

*Proof of Theorem 1.* Fix an integer $n \geq 2$. Initially let $a_1, \dots, a_n$ be independent $\mathrm{Bernoulli}(p)$ variables, and write

$$C = \sum_i a_i, \qquad z_i = a_i - p, \qquad Z = C - np, \qquad v = p(1-p), \qquad r = 1 - 2p.$$

The shared logit score of agent $i$ is $z_i$. All variables in this proof are scalar; they therefore form one complete shared parameter block.

**A root with uniformly bounded reciprocal payoff.** For $p \notin \{0, 1/n, \dots, 1\}$, define

$$f_n(p) = \sum_{k=0}^{n} \binom{n}{k} \frac{p^k (1-p)^{n-k}}{k - np}.$$

Differentiating this finite sum, using $\partial_p \Pr_p(C = k) = \Pr_p(C = k)(k - np)/v$, gives

$$f_n'(p) = \frac{1}{v} + n\mathbb{E}_p\left[\frac{1}{(C - np)^2}\right] > 0. \tag{14}$$

On each interval $(k/n, (k+1)/n)$, the pole at the left endpoint has negative sign and the pole at the right endpoint has positive sign. Their binomial masses are nonzero at interior endpoints. At 0 and 1, the corresponding endpoint mass tends to one. Thus $f_n$ has exactly one root in each such interval.

Let $p_n$ be its first root and put $c_n = np_n$. At $np = 1/2$, $P_0 = \Pr(C = 0) = (1 - 1/(2n))^n > 1/2$, by the strict Bernoulli inequality. Every positive reciprocal term is at most 2, so $f_n(1/(2n)) \leq -2P_0 + 2(1 - P_0) < 0$. At $np = 2/3$, retaining only the positive $C = 1$ term gives

$$f_n(2/(3n)) \geq -\frac{3}{2}P_0 + 3P_1 > 0, \qquad \frac{P_1}{P_0} = \frac{2/3}{1 - 2/(3n)} > \frac{2}{3}.$$

Monotonicity therefore yields

$$\frac{1}{2} < c_n < \frac{2}{3}. \tag{15}$$

**Fixed environment and nonzero target.** Now fix the behavior probability at $p = p_n$, and define the environment payoff and outcome by

$$Q(\mathbf{a}) = \frac{1}{\sum_i a_i - np_n}, \qquad Y = Q(\mathbf{a}) + \varepsilon, \qquad \Pr(\varepsilon = \sigma) = \Pr(\varepsilon = -\sigma) = \frac{1}{2}, \quad 0 < \sigma \le 1, \tag{16}$$

with $\varepsilon$ independent of the actions. The number $np_n$ in this environment is fixed when differentiating the policy objective. By Equation (15), $|Q| < 3$ and $|Y| < 4$, uniformly in $n$ and $\sigma$. Moreover,

$$\mathbb{E}[Q] = 0, \qquad ZQ = 1, \qquad X := Z(Y - B) = 1 + Z\varepsilon \quad \text{at } B = 0.$$

Consequently

$$\mathbb{E}[X] = 1, \qquad \mathrm{Var}(X) = nv\sigma^2. \tag{17}$$

The baseline is both the outcome mean and the unique optimal constant baseline for the *raw shared* estimator: since $\mathbb{E}[Z] = 0$,

$$B^*_{\text{shared,raw}} = \frac{\mathbb{E}[Z^2 Y]}{\mathbb{E}[Z^2]} = \frac{\mathbb{E}[Z]}{nv} = 0.$$

Exchangeability and $ZQ = 1$ imply $\mathbb{E}[z_i Q] = 1/n$, so every agent has a nonzero local gradient. As a function of one binary action, $q_i(a_i) := \mathbb{E}[Q \mid a_i]$ has mean zero and is proportional to $z_i$. Its coefficient is therefore

$$q_i(a_i) = \frac{z_i}{nv}. \tag{18}$$

**Local optima and the complete correction.** For finite $K$, let $\overline{Q}_i$ average $K$ fresh independent teammate-joint proposals, keeping the observed $a_i$ fixed. Conditional on the complete actual record, each focal stream consists of $K$ iid draws from the teammate behavior law. Different focal streams may be coupled, including conditionally on that record. Set

$$x_i = z_i Y, \qquad h_i^{(K)} = z_i(Q - \overline{Q}_i), \qquad D = \mathbb{E}\big[z_i^2 \,\mathrm{Var}(Q \mid a_i)\big].$$

The quantity $D$ is positive: $z_i \neq 0$, every teammate action has positive probability, and changing a teammate action changes the reciprocal payoff. Conditional independence of the actual and proposed teammate actions, and the conditional mean-zero outcome noise, give

$$\mathrm{Cov}(x_i, h_i^{(K)}) = D, \qquad \mathrm{Var}(h_i^{(K)}) = (1 + 1/K)D. \tag{19}$$

Thus the unique scalar variance-minimizing coefficient at this baseline is $\lambda_K = K/(K+1)$, and each agent's variance drops by $\lambda_K D > 0$. For exact marginalization, take $1/K = 0$ and $\lambda_\infty = 1$.

The binary identity $z_i^2 = v + rz_i$, together with Equation (18), gives the exact shared correction

$$H_\infty = \sum_i z_i(Q - q_i) = 1 - \frac{\sum_i z_i^2}{nv} = -\frac{rZ}{nv}, \qquad \mathrm{Var}(H_\infty) = \frac{r^2}{nv}. \tag{20}$$

For $H_K = \sum_i h_i^{(K)}$, conditional unbiasedness of each proposal average implies

$$\mathbb{E}[H_K \mid \mathbf{a}, \varepsilon] = H_\infty, \qquad \mathrm{Cov}(X, H_K) = 0, \qquad \mathrm{Var}(H_K) \ge \mathrm{Var}(H_\infty).$$

The last inequality is total variance and does not require independence between focal streams. Hence

$$\frac{\mathrm{Var}(X - \lambda_K H_K)}{\mathrm{Var}(X)} \ge 1 + \frac{\lambda_K^2 r^2}{\sigma^2 n^2 v^2} > 1 + \frac{1}{16\sigma^2}. \tag{21}$$

For the final inequality, Equation (15) gives $r > 1/3$ and $nv < 2/3$, while $\lambda_K \ge 1/2$.

Finally, replace the outcome, critic and baseline by $\widetilde{Y} = (Y+4)/8$, $\widetilde{Q} = (Q+4)/8$, and $\widetilde{B} = 1/2$. Then $\widetilde{Y} \in (0,1)$, both gradient estimators scale by $1/8$, and the shared task gradient is $1/8$ for every team size. The coefficients and variance ratios are unchanged. Letting $\sigma \downarrow 0$ establishes the claimed unbounded ratio. □

**Baseline and outcome conventions.** The construction uses bounded scalar terminal rewards. At the specified shared-optimal baseline, every correction coefficient is individually optimal. The raw-local optimal constant for the unnormalized first-root family is instead $r/(nv)$. Optimizing separate local baselines, or reoptimizing a shared baseline after choosing a correction, defines a different estimator. The divergence in Equation (21) is a variance *ratio* as the raw variance decreases; all displayed absolute moments remain finite.

**Proposition 6** (Jointly optimal local baselines at finite proposal budgets)**.** *For every odd $n \geq 3$ and every finite $K \geq 1$, there is a bounded cooperative one-step game in which jointly optimizing each agent's constant baseline and correction coefficient strictly reduces every local variance but increases the shared variance by an arbitrarily large factor.*

*Proof.* Take $p = 1/2$, $Z = \sum_i (a_i - 1/2)$, $Q = 1/Z$, and $Y = Q + \varepsilon$ with the independent noise in Equation (16). Odd $n$ ensures $Z \neq 0$; symmetry gives $\mathbb{E}[Q] = 0$, and $|Q| \leq 2$. The same exchangeability argument gives $q_i = z_i/(nv)$, where $v = 1/4$. Thus

$$X = 1 + Z\varepsilon, \quad \mathrm{Var}(X) = n\sigma^2/4, \quad H_\infty = 0, \quad D = \frac{1}{4}\mathbb{E}[Z^{-2}] - \frac{1}{n^2} > 0.$$

Because $z_i^2 = 1/4$, $\mathrm{Cov}(x_i, z_i) = \mathbb{E}[z_i^2 Y] = 0$. Also $\mathbb{E}[h_i^{(K)} \mid a_i] = 0$, so $\mathrm{Cov}(z_i, h_i^{(K)}) = 0$. The local quadratic minimization in baseline and coefficient separates, with the unique optimum $B_i = 0$, $\lambda_i = \lambda_K = K/(K+1)$. With conditionally independent focal proposal streams, total variance gives $\mathrm{Var}(H_K) = nD/K$ and $\mathrm{Cov}(X, H_K) = 0$. Therefore

$$\frac{\mathrm{Var}(X - \lambda_K H_K)}{\mathrm{Var}(X)} = 1 + \frac{4\lambda_K^2 D}{K\sigma^2}.$$

Each local variance drops by $\lambda_K D$, while the shared ratio diverges as $\sigma \downarrow 0$. The preceding affine normalization puts the terminal reward in $[0, 1]$ and fixes the shared gradient at $1/8$. This strengthening concerns finite $K$: at exact marginalization, $H_\infty = 0$ and the shared estimator is unchanged. □

## B THE SHARP TWO-AGENT TRANSFER WINDOW

**Purpose and proof idea.** Theorem 2 asks the complementary question: when is exact local marginalization guaranteed to be safe after sharing? For two Bernoulli agents, every payoff can be expanded in the orthogonal basis $1, z_1, z_2, z_1 z_2$. This reduces the shared covariance difference to a quadratic form whose sign is controlled by the policy imbalance $r = 1 - 2p$. The proof identifies exactly when that quadratic form is nonnegative for every payoff and then constructs a violating payoff outside the window. The finite-$K$ proposition at the end isolates proposal noise from this exact-marginalization geometry.

*Proof of Theorem 2.* Let $a_1, a_2$ be independent $\mathrm{Bernoulli}(p)$, $z_i = a_i - p$, $v = p(1-p)$, and $r = 1 - 2p$, with scores $S_i = z_i j_\eta$ for a fixed nonzero vector $j_\eta$. Take the exact critic $Q = \mathbb{E}[Y \mid a_1, a_2]$ and the common baseline $B = \mathbb{E}[Q]$. The orthogonal basis $1, z_1, z_2, z_1 z_2$ spans every four-entry payoff table, so uniquely

$$Q - B = c_1 z_1 + c_2 z_2 + t z_1 z_2, \qquad s = c_1 + c_2.$$

First take $Y = Q$. Exact teammate marginalization at coefficient one leaves the scalar shared estimator

$$W = c_1 z_1^2 + c_2 z_2^2 = vs + r(c_1 z_1 + c_2 z_2).$$

Writing the raw scalar estimator as $X = W + H$, the same binary identity $z_i^2 = v + r z_i$ gives

$$H = (s + 2rt) z_1 z_2 + vt(z_1 + z_2), \qquad \mathbb{E}[H] = 0.$$

Independence and centering yield

$$\mathrm{Var}(H) = v^2(s + 2rt)^2 + 2v^3 t^2, \qquad \mathrm{Cov}(W, H) = rv^2 ts.$$

It follows that

$$\begin{aligned} \mathrm{Cov}(j_\eta X) - \mathrm{Cov}(j_\eta W) &= v^2\big[(s + 2rt)^2 + 2vt^2 + 2rts\big] j_\eta j_\eta^{\mathsf{T}} \\ &= v^2\big[(s + 3rt)^2 + (2v - 5r^2)t^2\big] j_\eta j_\eta^{\mathsf{T}}. \end{aligned} \tag{22}$$

Since $2v = (1 - r^2)/2$, this matrix is positive semidefinite for every payoff table if $r^2 \leq 1/11$. Conversely, outside that interval choose $t = 1$ and $c_1 = c_2 = -3r/2$. The squared term then vanishes and the remaining coefficient is strictly negative. For either focal agent, $\mathrm{Var}(Q \mid a_i) = v(c_{-i} + tz_i)^2$; its expectation weighted by $z_i^2$ is positive because $t = 1$ and the two possible own actions differ. Thus both local variance reductions are strict in this witness. At the boundary the shared difference may be zero, which is why the classification concerns non-increase:

$$p \in \left[\frac{1 - 1/\sqrt{11}}{2}, \frac{1 + 1/\sqrt{11}}{2}\right].$$

Finally, for $Y = Q + \eta$ with $\mathbb{E}[\eta \mid a_1, a_2] = 0$, the same term $(z_1 + z_2)\eta$ is retained in the raw and corrected scalar estimators. It is uncorrelated with every action-measurable term. Its covariance therefore cancels from their difference, leaving Equation (22) unchanged. □

**Why the baseline is part of the classification.** At the stated baseline, exact teammate marginalization is locally variance-optimal: conditional on the own action it replaces the action-measurable part by its conditional expectation. Equivalently, its local covariance and control variance both equal $D_i = \mathbb{E}[z_i^2 \,\mathrm{Var}(Q \mid a_i)]$. If separate local constants are optimized instead, their optima are $B_i^* = \mathbb{E}[Q] + rc_i$. Their exact corrected action terms become $z_i(c_i z_i - rc_i) = vc_i$, constants. Thus the displayed window is the payoff-uniform classification at the specified common outcome-mean baseline, not a classification of joint local baseline optimization.

**Proposition 7** (Finite iid proposals and aligned Bernoulli strata)**.** *Suppose the joint collection of teammate proposals is independent of the actual actions. For every $0 < p < 1$ and every finite iid proposal budget $K$, there is an exact two-agent critic whose locally optimal corrections strictly increase shared variance at the common outcome-mean baseline. If the single teammate's inverse cumulative distribution function (CDF) uniform is instead stratified into $K$ equal-width cells and $Kp$ is an integer, its marginal is evaluated exactly, restoring the classification in Theorem 2 for every payoff table.*

*Proof.* For the first assertion, take $Q = z_1 - z_2$ and $B = 0$. Let $\bar{u}_2, \bar{u}_1$ be the centered teammate proposal averages for focal agents 1 and 2. Each averages $K$ iid centered Bernoulli draws; the joint collection is independent of the actual actions. Then

$$X = z_1^2 - z_2^2 = r(z_1 - z_2), \qquad H_K = z_1\bar{u}_2 - z_2\bar{u}_1, \qquad \mathbb{E}[H_K \mid a_1, a_2] = 0.$$

The cross term in $\mathbb{E}[H_K^2]$ vanishes because it factors through $\mathbb{E}[z_1 z_2] = 0$, even if the two focal proposal collections are coupled. Consequently

$$\mathrm{Var}(H_K) = \frac{2v^2}{K}, \qquad \mathrm{Cov}(X, H_K) = 0.$$

For each agent, the local covariance is $v^2$, the control variance is $(1 + 1/K)v^2$, and the unique optimal coefficient is $\lambda_K = K/(K + 1)$. Both local variances decrease, whereas

$$\mathrm{Var}(X - \lambda_K H_K) - \mathrm{Var}(X) = \frac{2\lambda_K^2 v^2}{K} > 0. \tag{23}$$

This example has zero shared target; Equation (23) is an absolute increase. The nonzero-target, bounded-ratio construction is given separately in Theorem 1.

For the second assertion, draw one independent uniform from each interval $((k-1)/K, k/K)$ and apply the Bernoulli inverse CDF. Write $Kp = q + f$, with $q = \lfloor Kp \rfloor$ and $0 \leq f < 1$. Exactly $q$ strata give deterministic successes and at most one boundary stratum gives a $\mathrm{Bernoulli}(f)$ success. The variance of the empirical mean is therefore

$$\nu_K(p) = \frac{f(1 - f)}{K^2} \leq \frac{v}{K}. \tag{24}$$

The inequality also follows directly from total variance for the equal-mass partition. With one binary teammate, the conditional payoff is affine in that teammate's action. Hence, for every payoff table,

$$\mathrm{Cov}(x_i, h_i^{\mathrm{strat}}) = D_i, \quad \mathrm{Var}(h_i^{\mathrm{strat}}) = (1 + \nu_K/v)D_i, \quad \lambda_i^* = \frac{v}{v + \nu_K} \quad (D_i > 0).$$

When $Kp$ is integral, $\nu_K = 0$: each equally weighted focal average is the exact conditional expectation, using $K$ critic queries. Then $\lambda_i^* = 1$ and Theorem 2 applies. For example, $p = 1/2$ is aligned for every even $K$. □

### B.1 THE PAYOFF-UNIFORM WINDOW FOR $n$ IDENTICAL AGENTS

Let $n \geq 2$ agents act independently with a common Bernoulli$(p)$ policy, $0 < p < 1$, whose shared score is $S_i = z_i J$, $z_i = a_i - p$, for a fixed nonzero vector $J$. Keep the exact critic $Q$, the value baseline $B = \mathbb{E}[Q]$ and $Y = Q + \eta$ with $\mathbb{E}[\eta \mid \mathbf{a}] = 0$ and $\mathbb{E}[\eta^2] < \infty$. Write $C = \sum_i a_i$, $T = C - np$, $v = p(1-p)$ and $r = 1 - 2p$; when $np \notin \mathbb{Z}$, put $\phi_1 = \mathbb{E}[T^{-1}]$ and $\phi_2 = \mathbb{E}[T^{-2}]$.

**Proposition 8** (Exact $n$-agent safety criterion)**.** *Exact teammate marginalization satisfies* $\mathrm{Cov}(X - H_\infty) \preceq \mathrm{Cov}(X)$ *for every bounded payoff iff $p$ is safe in the following sense. The balanced policy $p = 1/2$ is safe; every other $p$ with $np \in \mathbb{Z}$ is unsafe. For $p \neq 1/2$ and $np \notin \mathbb{Z}$, $p$ is safe iff*

$$\textit{(i)}\ \ r^2\,\mathbb{E}\left[\frac{C(n-C)}{T^2}\right] \leq v\,n(n-1) \qquad \textit{and} \qquad \textit{(ii)}\ \ r^2\phi_2 \leq (nv + r^2)\,\phi_1^2. \tag{25}$$

*For $n = 2$, (i) holds with equality and (ii) is $22p^2 - 22p + 5 \leq 0$, which is exactly Equation (9).*

*Proof.* Since $X = J\,T(Y - B)$ and $X - H_\infty = J(W + T\eta)$ with $W = \sum_i z_i(\mathbb{E}[Q \mid a_i] - B)$, and $T\eta$ is uncorrelated with every action-measurable variable, $\mathrm{Cov}(X) - \mathrm{Cov}(X - H_\infty) = JJ^{\mathsf{T}}\{\mathrm{Var}(TF) - \mathrm{Var}(W)\}$ with $F = Q - \mathbb{E}Q$. Work in $\mathcal{H}_0 = \{F : \mathbb{E}F = 0\}$ with $\langle F, G\rangle = \mathbb{E}[FG]$. The $\psi_i = z_i/\sqrt{v}$ are orthonormal; let $P_1$ project onto their span. Since $a_i$ is binary, $\mathbb{E}[F \mid a_i] = c_i z_i$ with $c_i = \mathbb{E}[F z_i]/v$, and $z_i^2 = v + r z_i$ gives $W = v\sum_i c_i + r\sum_i c_i z_i$, hence $\mathrm{Var}(W) = r^2\|P_1 F\|^2$. In particular $r = 0$ is safe. Also $\mathrm{Var}(TF) = \|LF\|^2$ for the self-adjoint map $LF = TF - \mathbb{E}[TF]$ on $\mathcal{H}_0$.

If $np \notin \mathbb{Z}$ and $\phi_1 \neq 0$, then $L$ is invertible with $L^{-1}h = (h - \mathbb{E}[h/T]/\phi_1)/T$: this function has mean zero and $LL^{-1}h = h$. Setting $h = LF$, safety for every payoff is $\|rP_1L^{-1}\|_{\mathrm{op}} \leq 1$, and $\|P_1L^{-1}\|_{\mathrm{op}}^2$ is the largest eigenvalue of the Gram matrix of $u_i = L^{-1}\psi_i$. Because $\sum_i z_i = T$, exchangeability gives $\mathbb{E}[z_i/T] = 1/n$ and $u_i = (z_i - 1/(n\phi_1))/(\sqrt{v}\,T)$. The Gram matrix is permutation invariant, so its eigenvalues belong to the common direction and to the contrast space. On the common direction, $n^{-1/2}\sum_i u_i = (1 - 1/(\phi_1 T))/\sqrt{nv}$ has squared norm $(\phi_2/\phi_1^2 - 1)/(nv)$. For a unit contrast $b$ ($\sum_i b_i = 0$), $\sum_i b_i u_i = \sum_i b_i z_i/(\sqrt{v}\,T)$, and sampling without replacement on the count slice gives $\mathbb{E}[(\sum_i b_i z_i)^2 \mid C] = C(n-C)/\{n(n-1)\}$, so its eigenvalue is $\mathbb{E}[C(n-C)/T^2]/\{vn(n-1)\}$. Multiplying both eigenvalues by $r^2$ and requiring them to be at most one gives (ii) and (i).

If $np \notin \mathbb{Z}$ but $\phi_1 = 0$, the payoff $F = 1/T$ has $\mathrm{Var}(TF) = 0$ and $c_i = 1/(nv)$, so $\mathrm{Var}(W) = r^2/(nv) > 0$; condition (ii) also fails. If $np = k \in \{1, \ldots, n-1\}$ and $r \neq 0$, the payoff $F = (a_1 - a_2)\mathbf{1}\{C = k\}$ has mean zero by symmetry and $TF = 0$, while $\mathbb{E}[F(z_1 - z_2)] = \mathbb{E}[(a_1 - a_2)^2\mathbf{1}\{C = k\}] > 0$ makes $\mathrm{Var}(W) > 0$. For $n = 2$, direct evaluation over $C \in \{0, 1, 2\}$ gives the stated forms of (i) and (ii). □

**How much of the policy interval is safe?** The criterion reduces the payoff-uniform comparison over all $2^n$ payoff entries to three binomial moments, so the unsafe share of $(0, 1)$ can be evaluated for every team size. We draw 131,072 uniform policies per $n$ and evaluate Equation (25) with exact count sums in float64 and a relative rounding guard of $10^{-9}$. For $n = 2$ the classification agrees with Equation (9) on every sampled policy; an independent CUDA float64 evaluation agrees within sampling error at every shared team size, and 430 full payoff-matrix eigenvalue tests for $n = 2, \ldots, 8$ agree with the criterion without exception. The estimated unsafe share is 0.70 for $n = 2$ (exactly $1 - 1/\sqrt{11}$), 0.78 for $n = 3$, 0.75 for $n = 4$, 0.74 for $n = 8$, 0.71 for $n = 16$, 0.66 for $n = 32$, 0.61 for $n = 64$ and 0.55 for $n = 128$, and it is at least 0.55 for every $n \in \{2, \ldots, 128\}$.

## C AN EXACT GAUSSIAN COORDINATION EXAMPLE

**Why this example is useful.** The all-team obstruction shows that locally optimal credit can be harmful, but it does not by itself tell us whether the underlying credit direction should always be rejected. This Gaussian example separates those questions. Exact marginalization at the local coefficient increases shared variance, while a smaller complete-block coefficient strictly decreases it. The calculations therefore motivate PLACID's coefficient selection rather than a binary "use local credit or abstain" rule. A second quadratic family shows that the useful shared coefficient can even have the opposite sign.

Consider two independent actions $a_i \sim \mathcal{N}(\theta, 1)$ with a shared scalar mean, evaluated at $\theta = 0$, so $S_i = a_i$. Let the deterministic terminal payoff be

$$Q(a_1, a_2) = \frac{a_1 + a_2}{2} - (a_1 - a_2)^2 + 2, \qquad Y = Q, \qquad B = \mathbb{E}[Q] = 0. \tag{26}$$

The payoff rewards aggregate progress and penalizes disagreement. For every $\theta$, $\mathbb{E}_\theta[Q] = \theta$, so its true shared gradient is one. This polynomial payoff has all required Gaussian moments.

Set $S = a_1 + a_2$ and $D = a_1 - a_2$. At the behavior point they are independent $\mathcal{N}(0, 2)$ variables, and

$$X = SQ = \frac{S^2}{2} + S(2 - D^2), \qquad \mathrm{Var}(X) = \frac{\mathrm{Var}(S^2)}{4} + \mathbb{E}[S^2]\,\mathrm{Var}(D^2) = 2 + 16 = 18.$$

Also $\mathbb{E}[S^2 Q] = 0$, so $B = 0$ minimizes variance among constant baselines for the raw shared gradient. The exact own-action conditional critic and corrected local gradient are

$$q_i(a_i) = \frac{a_i}{2} - a_i^2 + 1, \qquad u_i = a_i q_i = \frac{a_i^2}{2} - a_i^3 + a_i.$$

Using $\mathbb{E}[a_i^2] = 1$, $\mathbb{E}[a_i^4] = 3$, and $\mathbb{E}[a_i^6] = 15$,

$$\mathbb{E}[u_i] = \frac{1}{2}, \qquad \mathrm{Var}(u_i) = \frac{1}{2} + (15 - 6 + 1) = \frac{21}{2}.$$

For $x_i = a_i Q$, symmetry and the independence of $S, D$ give

$$\mathbb{E}[x_i^2] = \frac{\mathbb{E}[S^2 Q^2] + \mathbb{E}[D^2 Q^2]}{4} = \frac{19 + 81}{4} = 25, \qquad \mathrm{Var}(x_i) = \frac{99}{4}.$$

Thus each exact local correction removes $57/4$ variance units, yet $U = u_1 + u_2$ has variance 21, exceeding 18. Writing $H_\infty = X - U$, direct expansion gives

$$H_\infty = a_1 a_2 + (a_1 a_2 + 1)(a_1 + a_2), \qquad \mathrm{Var}(H_\infty) = 13, \qquad \mathrm{Cov}(X, H_\infty) = 5.$$

For example, the even part of $H_\infty$ has variance one, its odd part has variance twelve, and their covariance vanishes. Furthermore $\mathrm{Cov}(U, H_\infty) = -8$, giving the stated covariance with $X = U + H_\infty$.

For $K$ iid teammate proposals with conditionally independent focal streams, the local residual variance is $D_i = 57/4$. The locally optimal coefficient is $\beta_K = K/(K + 1)$, whereas

$$\mathrm{Var}(H_K) = 13 + \frac{57}{2K}, \qquad \mathrm{Cov}(X, H_K) = 5.$$

This independence holds also for common joint Gaussian proposals with two agents, since the two focal integrands use disjoint independent teammate coordinates. The shared variance reduction at the local optimum is

$$10\beta_K - \beta_K^2 \left(13 + \frac{57}{2K}\right) = -\frac{K(6K + 37)}{2(K + 1)^2} < 0. \tag{27}$$

In contrast, complete-block coefficient optimization yields

$$\lambda^*_{\mathrm{shared},K} = \frac{5}{13 + 57/(2K)}, \qquad \Delta^*_{\mathrm{shared},K} = \frac{25}{13 + 57/(2K)} > 0. \tag{28}$$

The exact-marginal values follow by $1/K = 0$. Thus the same exact critic supports a beneficial nonzero shared correction even when its locally optimal coefficient is harmful. Here the raw-local optimal constant baseline is $-2$; the comparison fixes the outcome-mean, raw-shared-optimal baseline $B = 0$.

**Orientation reversal on an open quadratic family.** Keep the shared Gaussian actor and fix

$$Q_\nu = Y_\nu = \tfrac{1}{2}(a_1 + a_2) - (a_1^2 + a_2^2) + \nu a_1 a_2 + 2.$$

At $\theta = 0$, $B = \mathbb{E}[Y_\nu] = 0$ and the shared mean is one; the objective is $J(\theta) = \theta + (\nu - 2)\theta^2$. The reward is strictly concave in the joint action for $|\nu| < 2$. Its own-action conditional critics and $U = u_1 + u_2$ are unchanged. Writing $P = a_1 a_2$ and $S = a_1 + a_2$ gives

$$H_\infty = P + \{1 + (\nu - 1)P\}S, \qquad D_i = \mathbb{E}[a_i^2\,\mathrm{Var}(Q_\nu \mid a_i)] = \tfrac{9}{4} + 3\nu^2.$$

The local cross moment is $D_i > 0$, hence the local optimum remains $\lambda_K = K/(K+1)$. Using $\mathbb{E}[P^2] = 1$, $\mathbb{E}[S^2] = 2$, $\mathbb{E}[S^2P] = 2$ and $\mathbb{E}[S^2P^2] = 6$, together with $\mathrm{Cov}(U, H_\infty) = -4\nu$, yields

$$\mathrm{Var}(X) = 6\nu^2 - 16\nu + 26,$$
$$\chi = \mathrm{Cov}(X, H_K) = 6(\nu-1)^2 - 1,$$
$$\omega_K = \mathrm{Var}(H_K) = 6\nu^2 - 8\nu + 5 + \frac{2(9/4 + 3\nu^2)}{K}.$$

The finite-$K$ formula uses independent focal streams, also realized by common joint proposals for these two agents. Thus throughout $|\nu - 1| < 1/\sqrt{6}$, every positive coefficient increases shared variance, whereas $\lambda^*_{\mathrm{shared,K}} = \chi/\omega_K < 0$ removes $\chi^2/\omega_K > 0$. At $\nu = 1$, exact local correction reduces each variance from $63/4$ to $21/2$ but raises shared variance from 16 to 21; the shared optimum is $-1/3$ and removes $1/3$. This sign-reversal statement fixes the value baseline: the raw shared-optimal constant is instead $B^*_{\mathrm{shared}} = \nu - 2$, at which the cross moment becomes $4(\nu-1)^2 + 1 > 0$. Theorem 1, separately, establishes an obstruction at the raw shared-optimal baseline itself.

## D Conditional Sampling Refinement and Composition

**Role of this appendix.** Admission decides which correction should reach the actor. Here that decision is held fixed, and we compare only different Monte Carlo realizations of the same correction. Conditioning on the recorded data and all admitted choices removes the task-gradient uncertainty from the comparison. The remaining question is purely a sampling question: can stratifying each independent proposal component reduce covariance without changing its mean? Theorem 5 answers yes, provided terms that share proposal randomness remain in the same component.

### D.1 The conditional comparison

The stratification identity used here is the usual total-covariance identity. We state its conditioning and composition explicitly to identify which sampling replacements preserve a shared-gradient comparison.

*Proof of Theorem 5.* Let $\mathcal{R}$ contain the recorded actual data and all choices held fixed between the two realizations, including fitted models, actor scores, baselines, coefficients, routes, proposal budgets, directions, and any upstream pilot or gate randomness. The fresh production randomness being compared is generated after this information is fixed. Conditional on $\mathcal{R}$, let component $g$ have proposal law $\nu_g$ and vector integrand $F_g$. Partition $\nu_g$ into $K_g$ measurable cells $C_{gk}$, each of conditional probability $1/K_g$. These objects, including the finite number of components and their positive integer budgets, are $\mathcal{R}$-measurable. Suppose all weighted second moments below are integrable.

Write

$$m_g = \mathbb{E}_{\nu_g}[F_g \mid \mathcal{R}], \qquad m_{gk} = \mathbb{E}_{\nu_g}[F_g \mid \mathcal{R}, U_g \in C_{gk}],$$

and denote the corresponding conditional covariance matrices by $V_g, V_{gk}$. The iid realization averages $K_g$ independent base-law draws. The stratified realization draws independently once from each conditional cell law and averages the $K_g$ results. Equal-mass mixture recovery gives the same conditional mean $m_g$. Independence within either realization gives

$$\mathrm{Cov}(\overline{F}_g^{\mathrm{iid}} \mid \mathcal{R}) = \frac{V_g}{K_g}, \qquad \mathrm{Cov}(\overline{F}_g^{\mathrm{strat}} \mid \mathcal{R}) = \frac{1}{K_g^2}\sum_{k=1}^{K_g} V_{gk}.$$

Conditional total covariance therefore yields

$$\Gamma_g := \mathrm{Cov}(\overline{F}_g^{\mathrm{iid}} \mid \mathcal{R}) - \mathrm{Cov}(\overline{F}_g^{\mathrm{strat}} \mid \mathcal{R}) = \frac{1}{K_g^2}\sum_{k=1}^{K_g}(m_{gk} - m_g)(m_{gk} - m_g)^{\mathsf{T}} \succeq 0. \tag{29}$$

Now let $C(\mathcal{R})$ be any vector and $W_g(\mathcal{R})$ compatible matrices, and form

$$G^r = C(\mathcal{R}) + \sum_g W_g \overline{F}_g^r, \qquad r \in \{\mathrm{iid}, \mathrm{strat}\}.$$

Suppose different component production streams are conditionally independent within each scheme. The two schemes may be coupled to each other for paired evaluation. They have identical conditional mean $m_G = C + \sum_g W_g m_g$, while

$$\Delta(\mathcal{R}) := \mathrm{Cov}(G^{\mathrm{iid}} \mid \mathcal{R}) - \mathrm{Cov}(G^{\mathrm{strat}} \mid \mathcal{R}) = \sum_g W_g \Gamma_g W_g^{\mathsf{T}} \succeq 0. \tag{30}$$

In total covariance the two $\mathrm{Cov}(m_G)$ terms cancel, proving

$$\mathrm{Cov}(G^{\mathrm{iid}}) - \mathrm{Cov}(G^{\mathrm{strat}}) = \mathbb{E}[\Delta(\mathcal{R})] \succeq 0. \tag{31}$$

More generally, for any square-integrable $\mathcal{R}$-measurable target $T$, each conditional error second moment is its conditional covariance plus $(m_G - T)(m_G - T)^{\mathsf{T}}$. Hence

$$\begin{aligned} &\mathbb{E}[(G^{\mathrm{iid}} - T)(G^{\mathrm{iid}} - T)^{\mathsf{T}} \mid \mathcal{R}] \\ &\quad - \mathbb{E}[(G^{\mathrm{strat}} - T)(G^{\mathrm{strat}} - T)^{\mathsf{T}} \mid \mathcal{R}] = \Delta(\mathcal{R}). \end{aligned} \tag{32}$$

This also proves domination for any integrable quadratic loss with a $\mathcal{R}$-measurable positive-semidefinite weight matrix, by taking the trace after multiplication by that matrix. □

**Application to frozen updates.** For prospective recycling, put every raw-prefix and actual-action term in $C$. A suffix proposal contribution receives the corresponding frozen weight $W_g = \lambda_M I/N$, with its sign fixed by the integrand convention. Its covariance drop is weighted by $\lambda_M^2/N^2$. Reciprocal folds and frozen linear preconditioners are handled by their own $W_g$. The actual data and frozen choices may be dependent: their contribution to total covariance cancels because they are identical in the two realizations. Predictable-prefix conditions separately establish the task-target claim; no iid-data or environment-tail condition is needed for this relative production-sampling comparison. If all correction weights vanish, both realizations reduce to the same raw estimator samplewise.

The conditioning excludes the production draws being replaced and later training data influenced by an update using those draws. In particular, using those production draws to select a new coefficient, partition, or sample count is not the frozen comparison above. A frozen upstream selection followed by fresh production draws satisfies the stated order.

### D.2 Which proposal terms form one independent component?

Condition on $\mathcal{R}$. Let $U_1, \ldots, U_m$ be independent primitive blocks under the declared proposal law, and let focal vector integrand $F_i$ depend on the declared subset $D_i$ of these blocks. The connected components of the bipartite graph joining each integrand to the primitive blocks it uses yield a decomposition

$$F(U) = F_0(\mathcal{R}) + \sum_g F_g(U_{C_g}; \mathcal{R}) \tag{33}$$

with disjoint independent $U_{C_g}$. A valid functional identity or architectural restriction may split an integrand further before this graph is built. A zero derivative at one expansion point does not establish functional independence.

Apply Equation (29) independently to each component in Equation (33). This preserves the iid comparator's component factorization. The proof uses each component's equal-mass mixture separately; it does not require that joint tuples reconstructed by aligning component stratum indices themselves have the original product-mixture law.

For common joint proposals with independent agent blocks $U_j$, the focal integrand is

$$F_i = S_i Q(a_i, T_{-i}(U_{-i})).$$

With two agents, $F_1$ uses only $U_2$ and $F_2$ only $U_1$. They are two independent components even when stored in one joint tensor. Independent focal stratification therefore refines the same iid complete-integrand law, using the same $2K$ focal critic evaluations. For generic $n \geq 3$ critics depending on every teammate block, the overlapping dependencies connect all focal terms into one component. If the comparator instead declares disjoint proposal variables for each focal agent, its focal components are independent for every $n$. For $n \geq 3$, that independent-focal iid comparator generally differs from the common-joint iid comparator.

For instance, with three recorded scalar scores $(1, -1, 1)$ and $Q(a) = La_3$, common proposals give $F_1 = LU_3$, $F_2 = -LU_3$, and the constant $F_3 = La_3$. Their proposal fluctuations cancel exactly. Replacing the first two uses of $U_3$ by independent copies destroys that cancellation, even if both copies are separately stratified. Keeping the coupled integrands in one component retains the cancellation and makes Theorem 5 applicable.

### D.3 An affine comparison between fine and coarse partitions

For a standard normal scalar $Z$, let $q_k = \Phi^{-1}(k/K)$, $\mu_k = \mathbb{E}[Z \mid q_{k-1} < Z \le q_k]$, and

$$\alpha_K = \frac{1}{K}\sum_{k=1}^{K} \mu_k^2, \qquad 0 < \alpha_K < 1 \quad (K \ge 2). \tag{34}$$

For $U \sim \mathcal{N}(0, I)$, an affine vector integrand $c + BU$ has stratum means $c + Bv\mu_k$ when partitioned along a unit vector $v$. Thus Equation (29) gives the classical covariance drop

$$\Gamma(v) = \frac{\alpha_K}{K} Bvv^\mathsf{T}B^\mathsf{T}. \tag{35}$$

Suppose the independent focal components instead have the form $F_i(U_i) = S_i(b_i + A_iU_i)$, where $A_i$ is a row vector and $S_i$ a recorded score vector. Concatenating the matrices $S_iA_i$ gives $B$, with iid-average covariance $BB^\mathsf{T}/K$. Within each component choose $A_i^\mathsf{T}/\|A_i\|$, arbitrarily if $A_i = 0$. Summing its drops gives

$$\Gamma_{\text{fine}} = \frac{\alpha_K}{K} BB^\mathsf{T}, \qquad \Gamma_{\text{fine}} - \Gamma(v) = \frac{\alpha_K}{K} B(I - vv^\mathsf{T})B^\mathsf{T} \succeq 0.$$

Hence these component-aligned strata dominate every single global direction for this affine scalar-critic-times-score class, at the same number of focal critic evaluations. For nonlinear integrands, each frozen partition still satisfies Theorem 5; the ordering between two different partitions is not implied by this affine calculation. Evaluation-count matching and timing comparisons are separate questions.

## E Shared Geometry Versus Agent-Diagonal Directions

**Purpose.** Theorem 5 guarantees non-increase for any frozen stratification, but it does not say which direction is most useful. This appendix gives a simple affine example showing why the direction should be chosen from the complete shared-gradient geometry rather than by summing agent-wise objectives. Large local sensitivities can cancel after sharing, so an agent-diagonal direction can focus on variation that is almost irrelevant to the final update.

The following conditional construction isolates which score geometry controls useful stratification of a coupled proposal component. It uses the classical identity Equation (35).

Fix three nonzero recorded scalar scores $s = (1, -1, 1)$, a common joint proposal $U \sim \mathcal{N}(0, I_3)$, and the affine critic

$$Q(a_1, a_2, a_3) = a_1 + a_2 + La_3, \qquad L > 0.$$

Holding the focal action fixed leaves sensitivity rows

$$A_1 = (0, 1, L), \qquad A_2 = (1, 0, L), \qquad A_3 = (1, 1, 0).$$

The complete proposal integrand is $F(U) = c + b^\mathsf{T}U$, with $b = \sum_i s_iA_i = (0, 2, 0)$. Its exactly optimal direction is $e_2$, giving variance drop $4\alpha_K/K$. For a general frozen affine component, the trace-drop objective is $v^\mathsf{T}A^\mathsf{T}GAv$, where $G_{ij} = S_i^\mathsf{T}S_j$. Here its full and agent-diagonal matrices are respectively

$$A^\mathsf{T}ss^\mathsf{T}A = \operatorname{diag}(0, 4, 0), \qquad A^\mathsf{T}\operatorname{diag}(s_i^2)A = \begin{pmatrix} 2 & 1 & L \\ 1 & 2 & L \\ L & L & 2L^2 \end{pmatrix}. \tag{36}$$

The latter exactly maximizes the *sum of individual* variance drops over single directions; this is not a linearization claim for the affine example.

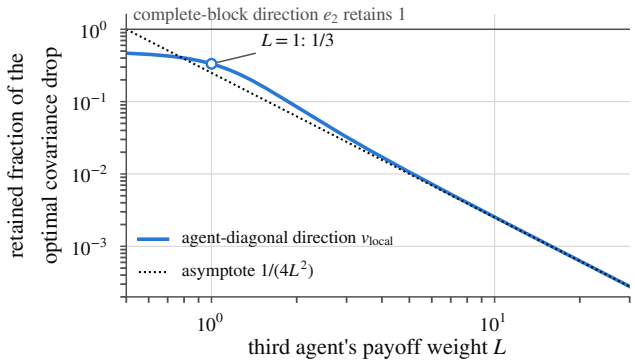


Figure 5: Retained fraction Equation (37) of the optimal complete-block covariance drop when the stratification direction maximizes the agent-diagonal objective in the three-agent affine example. The complete-block direction $e_2$ retains the full drop for every $L$; the agent-diagonal direction retains $1/3$ at $L = 1$ and decays as $1/(4L^2)$.

The vector $(1, -1, 0)/\sqrt{2}$ is an eigenvector of the diagonal geometry with eigenvalue one. On the span of $u = (1, 1, 0)/\sqrt{2}$ and $e_3$, the matrix is

$$\begin{pmatrix} 3 & \sqrt{2}L \\ \sqrt{2}L & 2L^2 \end{pmatrix}.$$

Its larger eigenvalue is

$$\lambda_+ = \frac{3 + 2L^2 + \sqrt{4L^4 - 4L^2 + 9}}{2} > 3,$$

so its leading unit direction $v_{\text{local}}$ is unique up to sign. The squared $u$-coordinate of this direction equals $\frac{1}{2}[1 + (3 - 2L^2)/\sqrt{4L^4 - 4L^2 + 9}]$. Therefore its retained fraction of the optimal complete-block drop is

$$\frac{\Gamma_{\text{scalar}}(v_{\text{local}})}{\Gamma_{\text{scalar}}(e_2)} = (v_{\text{local}})_2^2 = \frac{1}{4}\left(1 + \frac{3 - 2L^2}{\sqrt{4L^4 - 4L^2 + 9}}\right) = \frac{1}{4L^2} + O(L^{-4}) \longrightarrow 0. \tag{37}$$

Every frozen direction remains non-increasing relative to iid proposal variance. The separation is in the fraction of attainable reduction: the large $L$-coordinate cancels in the shared integrand but dominates the local objective. Figure 5 plots this fraction.

**A positive-probability policy realization.** The displayed score event is realized exactly by the factorized shared policy

$$\pi_\theta(a) = \frac{\phi(a)\exp(\theta\,\text{sign}(a))}{\cosh\theta}, \qquad \nabla_\theta \log \pi_\theta(a) = \text{sign}(a) - \tanh\theta.$$

At $\theta = 0$, its action law is standard Gaussian and its scores are independent signs. The pattern $(1, -1, 1)$ has probability $1/8$; its global sign reversal has the same covariance drops, giving a combined probability $1/4$. The construction is conditional on these recorded-score events, for which all three agents are active. It does not assert a vanishing ratio after averaging over all possible score patterns.

Finally, the three patterns $(-1, 1, 1)$, $(1, -1, 1)$, and $(1, 1, -1)$ have identical agent-diagonal matrices but complete coefficients $(2, 0, 0)$, $(0, 2, 0)$, and $(0, 0, 2L)$. Their shared optima are the three orthogonal coordinate axes. Consequently a direction depending only on this diagonal geometry cannot be optimal on all three; for any fixed unit direction its smallest retained fraction is at most $1/3$. Equation (37) gives the stronger vanishing fraction for the particular direction that optimizes the local objective.

## F Additional Estimator Results and Proofs

**How to read this appendix.** The subsections here show why local summaries do not determine complete-block admission, prove target preservation and prospective recycling, and justify the

scheduled-look gate. These are one-refresh estimator guarantees, not claims about long-run training return.

**Setup and notation.** All expectations condition on the information frozen before the evaluated cluster unless stated otherwise. Before cluster 1 this includes the behavior policy, baseline, actor-block structure, and prespecified protocol; for a corrected suffix cluster it additionally includes the stopped prefix and the selected $Q_M, \boldsymbol{\lambda}_M$. A score fitted during the refresh is therefore pre-deployment, not literally pre-refresh. A dependency cluster is the smallest unit independent of the others under this conditioning. Bold symbols concatenate the complete shared-actor blocks.

### F.1 Local Information Does Not Determine Shared Admission

Proposition 3 is stated in Section 4; we prove it here.

*Proof of Proposition 3.* Let $a_1, a_2$ be independent uniform Rademacher actions under a shared Bernoulli logit, so $S_i = a_i/2$. Set $Y = 5a_1a_2/4$, $B = 0$, and

$$Q_+ = a_1 + a_2 + a_1a_2, \qquad Q_- = a_1 - a_2 + a_1a_2.$$

Both prediction MSEs equal $1 + 1 + (1 - 5/4)^2 = 33/16$. With exact teammate marginalization,

$$x_1 = \frac{5}{8}a_2, \qquad x_2 = \frac{5}{8}a_1,$$
$$h_{1,+} = \frac{a_1a_2 + a_2}{2}, \qquad h_{1,-} = \frac{-a_1a_2 + a_2}{2},$$
$$h_{2,+} = h_{2,-} = \frac{a_1a_2 + a_1}{2}.$$

For either score and either agent, $(x_i, h_i)$ is uniform over

$$\{(-5/8, -1),\ (-5/8, 0),\ (5/8, 0),\ (5/8, 1)\}.$$

Every local reduction is $1/8$. For the shared update, however,

$$X = \frac{5}{8}(a_1 + a_2), \qquad H_+ = a_1a_2 + \frac{a_1 + a_2}{2}, \qquad H_- = \frac{a_1 + a_2}{2},$$

so $\mathrm{Var}\, X = 25/32$, $\mathcal{D}_+(1) = -1/4$, and $\mathcal{D}_-(1) = 3/4$. The allowed summaries are identical while the correct signs differ, proving the claim. A randomized binary classifier has average error $1/2$ under an equal mixture of the two cases. A rule may abstain on both, in which case it misses the beneficial score. Paired joint observations are not restricted to these local laws: they reveal the distinguishing cross-agent dependence. □

**Derivation of the local–shared decomposition.** Index the agent–time terms in one shared block by $\ell$, with $X_b = \sum_\ell x_\ell$ and $H_b = \sum_\ell h_\ell$. By target preservation, $\mathbb{E}[H_b] = 0$, so $X_b$ and $X_b - \lambda H_b$ have the same mean. Therefore the covariance-trace reduction satisfies

$$\begin{aligned}
\mathcal{D}_b(\lambda) &= \operatorname{tr}\mathrm{Cov}(X_b) - \operatorname{tr}\mathrm{Cov}(X_b - \lambda H_b) \\
&= \mathbb{E}\|X_b - \mathbb{E}[X_b]\|^2 - \mathbb{E}\|X_b - \mathbb{E}[X_b] - \lambda H_b\|^2 \\
&= 2\lambda\, \mathbb{E}\big[(X_b - \mathbb{E}[X_b])^\mathsf{T} H_b\big] - \lambda^2 \mathbb{E}\|H_b\|^2 \\
&= 2\lambda\, \mathbb{E}[X_b^\mathsf{T} H_b] - \lambda^2 \mathbb{E}\|H_b\|^2,
\end{aligned} \tag{38}$$

where the last equality again uses $\mathbb{E}[H_b] = 0$. Expanding the two inner products gives

$$\mathbb{E}[X_b^\mathsf{T} H_b] = \sum_\ell \mathbb{E}[x_\ell^\mathsf{T} h_\ell] + \sum_{\ell \neq r} \mathbb{E}[x_\ell^\mathsf{T} h_r],$$

and

$$\mathbb{E}\|H_b\|^2 = \sum_\ell \mathbb{E}\|h_\ell\|^2 + \sum_{\ell \neq r} \mathbb{E}[h_\ell^\mathsf{T} h_r].$$

Substituting these identities into Equation (38) yields

$$\mathcal{D}_b(\lambda) = \sum_\ell \{2\lambda \mathbb{E}[x_\ell^\mathsf{T} h_\ell] - \lambda^2 \mathbb{E}\|h_\ell\|^2\} + \sum_{\ell \neq r} \{2\lambda \mathbb{E}[x_\ell^\mathsf{T} h_r] - \lambda^2 \mathbb{E}[h_\ell^\mathsf{T} h_r]\}, \tag{39}$$

which is Equation (7). The first sum contains the local terms, whereas the second contains the cross-agent and cross-time terms created by aggregation.

**What information composes.** Disjoint parameter supports make the cross-agent inner products vanish pointwise, so complete agent-trajectory reductions add. Likewise, if all agents use the same conditioning sigma-field $\mathcal{A}$, then $\sum_i \mathbb{E}[x_i \mid \mathcal{A}] = \mathbb{E}[X \mid \mathcal{A}]$ and ordinary conditional-expectation variance reduction applies to the sum. Teammate marginalization instead conditions each agent on its own action; Theorem 1 shows the consequence of those different information sets. PLACID retains the joint cross terms in Equation (39) when deciding whether a correction is admissible.

### F.2 Centering and Target Preservation

**Theorem 9** (Target preservation)**.** *Under the factorized behavior policy, fix the score, action-independent baseline rule, and coefficient before a corrected cluster. Suppose the $K$ teammate proposals follow the conditional teammate policy, are independent of the realized teammate actions given the state and focal action, and have the required integrable moments. Then for every shared actor block,*

$$\mathbb{E}[H_{j,b}(Q)] = 0, \qquad \mathbb{E}[X_{j,b} - \lambda_b H_{j,b}(Q)] = \mathbb{E}[X_{j,b}]. \tag{40}$$

*The same holds for exact teammate marginalization.*

*Proof.* Conditional on $(\mathbf{s}_t, a_{i,t})$ and all objects frozen before the cluster, the realized and proposed teammate actions have the same law. Consequently $\mathbb{E}[d_{i,t}^{(K)}[Q] \mid \mathbf{s}_t, a_{i,t}] = 0$. The score $S_{i,t}^{(b)}$ is measurable under this conditioning, so its product with the correction has expectation zero. Summing over agents and time proves $\mathbb{E}[H_{j,b}(Q)] = 0$; multiplying by the frozen coefficient and subtracting from $X_{j,b}$ proves Equation (40). Exact marginalization uses the same conditional-centering identity. □

With conditionally independent proposals, the decision-level moments are

$$\mathbb{E}[d_i^{(K)}[Q] \mid \mathbf{s}, a_i] = 0, \qquad \mathrm{Var}(d_i^{(K)}[Q] \mid \mathbf{s}, a_i) = \left(1 + \frac{1}{K}\right) \mathrm{Var}_{\mathbf{a}_{-i} \sim \boldsymbol{\pi}_{-i}^{\mathrm{old}}} Q(\mathbf{s}, a_i, \mathbf{a}_{-i}). \tag{41}$$

These local moments do not determine the covariance of the aggregated block.

### F.3 Complete-Block Coefficient Geometry

For $V_b = \mathrm{tr}\,\mathrm{Cov}(X_{j,b}) > 0$ and $\omega_b > 0$, the complete-block reduction is a concave quadratic with optimum

$$\lambda_b^\star = \frac{\chi_b}{\omega_b}, \qquad \mathcal{D}_b(\lambda_b^\star) = \frac{\chi_b^2}{\omega_b}, \qquad 0 \le \frac{\chi_b^2}{\omega_b V_b} \le 1. \tag{42}$$

The final inequality is Cauchy–Schwarz. A coefficient helps precisely when $\lambda_b(2\chi_b - \lambda_b\omega_b) > 0$; any coefficient grid and score normalization must be fixed before screening.

### F.4 Proof of Proposition 4

For distinct actor blocks, write $\mathbf{X}_j = (X_{j,b})_b$ and $\mathbf{H}_j = (H_{j,b})_b$, with coordinatewise block coefficients $\boldsymbol{\lambda}$. Set $\mathcal{V}_X = \sum_b \mathrm{tr}\,\mathrm{Cov}(X_{j,b}) > 0$, $\chi_{A,b} = \mathbb{E}[X_{j,b}^\mathsf{T} H_{j,b}(Q_A) \mid A]$, $\omega_{A,b} = \mathbb{E}[\|H_{j,b}(Q_A)\|^2 \mid A]$, and $\mathcal{D}_A(\boldsymbol{\lambda}) = \sum_b (2\lambda_b\chi_{A,b} - \lambda_b^2\omega_{A,b})$. This is the blockwise form of the main-text scalar-coefficient notation.

*Proof.* Assume that the $N$ clusters are conditionally iid under the frozen behavior policy and have finite second moments. Let $M \le N$ be a bounded stopping time for the filtration containing every cluster, fit, gate decision, and resampling draw revealed by each look. The selected $Q_M$ and $\boldsymbol{\lambda}_M$ are measurable with respect to the stopped prefix and frozen on the suffix. We also assume $\mathbb{E}\|\boldsymbol{\lambda}_M \odot \mathbf{H}_j(Q_M)\|_2^2 < \infty$ and $\mathbb{E}[(N - M)|\mathcal{D}_M(\boldsymbol{\lambda}_M)|] < \infty$.

*Step 1 (conditional law of the suffix).* Write $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}_j]$, and let $\mathcal{F}_M$ denote the stopped-prefix sigma-field. Conditional on $\mathcal{F}_M$, the suffix clusters are iid under the frozen behavior policy. Fresh proposal streams are independent across those clusters and use the rule selected from the prefix. Theorem 9 therefore gives

$$\mathbb{E}[\mathbf{H}_j(Q_M) \mid \mathcal{F}_M] = \mathbf{0}, \qquad \mathbb{E}[\mathbf{X}_j - \boldsymbol{\lambda}_M \odot \mathbf{H}_j(Q_M) \mid \mathcal{F}_M] = \boldsymbol{\mu}.$$

*Step 2 (mean and covariance of the stopped prefix).* Conditioning Equation (12) on $\mathcal{F}_M$ gives

$$\mathbb{E}[\widehat{\mathbf{G}}^{\mathrm{R}} \mid \mathcal{F}_M] = \boldsymbol{\mu} + \frac{1}{N}\sum_{j\le M}(\mathbf{X}_j - \boldsymbol{\mu}).$$

The centered cluster contributions form a square-integrable martingale-difference sequence. Bounded optional stopping and martingale orthogonality yield

$$\mathbb{E}[\widehat{\mathbf{G}}^{\mathrm{R}}] = \boldsymbol{\mu}, \qquad \operatorname{tr}\operatorname{Cov}\Big(\mathbb{E}[\widehat{\mathbf{G}}^{\mathrm{R}} \mid \mathcal{F}_M]\Big) = \frac{\mathbb{E}[M]}{N^2}\mathcal{V}_X. \tag{43}$$

*Step 3 (covariance of the independent suffix).* By the definition of $\mathcal{D}_M$, one fresh corrected cluster has conditional covariance trace $\mathcal{V}_X - \mathcal{D}_M(\boldsymbol{\lambda}_M)$. Conditional independence across the suffix gives

$$\operatorname{tr}\operatorname{Cov}(\widehat{\mathbf{G}}^{\mathrm{R}} \mid \mathcal{F}_M) = \frac{N-M}{N^2}\{\mathcal{V}_X - \mathcal{D}_M(\boldsymbol{\lambda}_M)\}. \tag{44}$$

*Step 4 (total covariance).* Combining Equations (43) and (44) gives

$$\operatorname{tr}\operatorname{Cov}(\widehat{\mathbf{G}}^{\mathrm{R}}) = \frac{\mathcal{V}_X}{N} - \frac{1}{N^2}\mathbb{E}[(N-M)\mathcal{D}_M(\boldsymbol{\lambda}_M)].$$

The all-data sparse estimator has covariance trace $\mathcal{V}_X/N$; division proves Equation (13).

*Step 5 (abstention and discarded splitting).* If $\boldsymbol{\lambda}_M = \mathbf{0}$, every summand in Equation (12) equals $\mathbf{X}_j$, so the fallback is exact for the realized batch. For fixed $M$, averaging only the $N - M = qN$ corrected suffix clusters gives covariance ratio $(1-\bar{\rho})/q$, whereas recycling gives $1 - q\bar{\rho}$. □

For $0 < q < 1$, the exact fixed-split phase map is

$$\begin{aligned} R_{\mathrm{sparse}} &= 1, \qquad R_{\mathrm{recycle}} = 1 - q\bar{\rho}, \qquad R_{\mathrm{discard}} = \frac{1-\bar{\rho}}{q}, \\ R_{\mathrm{recycle}} \le R_{\mathrm{discard}} \quad &\Longleftrightarrow \quad \bar{\rho} \le \frac{1}{1+q}. \end{aligned} \tag{45}$$

Recycling improves on all-data sparse exactly when $\bar{\rho} > 0$, while suffix-only averaging requires $\bar{\rho} > 1-q$. Suffix-only averaging becomes preferable to recycling beyond the high-reduction crossover $\bar{\rho} > (1+q)^{-1}$.

The exact adaptive counterpart is

$$\operatorname{tr}\operatorname{Cov}(\widehat{\mathbf{G}}^{\mathrm{R}}) < \operatorname{tr}\operatorname{Cov}(\widehat{\mathbf{G}}^{0}) \quad \Longleftrightarrow \quad \mathbb{E}\left[\left(1 - \frac{M}{N}\right)\rho_M\right] > 0. \tag{46}$$

### F.5 Gate Construction and Proof of Corollary 10

After a model prefix of size $m_0$ freezes $Q$, each later screening cluster provides, for every candidate $\boldsymbol{\lambda}$ in a finite grid $\Lambda$,

$$Z_j(\boldsymbol{\lambda}) = \sum_b \{2\lambda_b X_{j,b}^{\mathsf{T}} H_{j,b} - \lambda_b^2 \|H_{j,b}\|_2^2\}, \qquad \mathbb{E}[Z_j(\boldsymbol{\lambda}) \mid \mathcal{F}_{m_0}] = \mathcal{D}(\boldsymbol{\lambda}). \tag{47}$$

Let $\mathcal{T}$ be the frozen look schedule and $n_m = m - m_0$. Suppose the gate observations are conditionally iid and finite endpoints $\ell_{\boldsymbol{\lambda}} \le Z_j(\boldsymbol{\lambda}) \le u_{\boldsymbol{\lambda}}$ are established before the gate stream. Write $R_{\boldsymbol{\lambda}} = u_{\boldsymbol{\lambda}} - \ell_{\boldsymbol{\lambda}}$ and let $\widehat{V}_m(\boldsymbol{\lambda})$ be the Bessel sample variance. With $x_r^{\mathrm{EB}} = \log(2|\Lambda||\mathcal{T}|/\delta_r)$ and $n_m \ge 2$, the empirical-Bernstein lower bound is

$$L_m^{\mathrm{EB}}(\boldsymbol{\lambda}) = \bar{Z}_m(\boldsymbol{\lambda}) - \sqrt{\frac{2\widehat{V}_m(\boldsymbol{\lambda})x_r^{\mathrm{EB}}}{n_m}} - \frac{7R_{\boldsymbol{\lambda}}x_r^{\mathrm{EB}}}{3(n_m-1)}. \tag{48}$$

It applies to each candidate $\boldsymbol{\lambda}$ (Maurer and Pontil, 2009). Support is fixed before screening; variance is estimated on the gate stream within the bound. A union bound covers the frozen $\mathcal{T} \times \Lambda$ pairs simultaneously. At the first positive bound, the maximizing candidate is frozen on the unseen suffix; if none crosses, $M = N$ and the method abstains. An alternative bound using prespecified conditional sub-gamma constants is given in Appendix F.5.

**Corollary 10** (Scheduled-look gate)**.** *Under the conditional-iid support condition above, with a frozen grid and look schedule, $L_m^{\mathrm{EB}}$ covers all grid–look pairs jointly with probability at least $1-\delta_r$. The same statement holds for the sub-gamma construction in Appendix F.5. On this event, any $(\mathcal{F}_m)$-adapted rule that stops at a positive bound and freezes $(M, \boldsymbol{\lambda}_M)$ before observing clusters $j > M$ weakly reduces the conditional covariance trace of that suffix. Every abstention is exactly the all-data sparse task-gradient estimator.*

Let $\mathcal{T}$ be the prespecified look schedule, $n_m = m - m_0$, and $x_r = \log(|\Lambda||\mathcal{T}|/\delta_r)$. For each grid point, assume predictable constants $(v_{\boldsymbol{\lambda}}, c_{\boldsymbol{\lambda}})$, fixed before the gate stream, such that

$$\log \mathbb{E}\Big[e^{\eta\{Z_j(\boldsymbol{\lambda})-\mathcal{D}(\boldsymbol{\lambda})\}} \mid \mathcal{F}_{j-1}\Big] \le \frac{\eta^2 v_{\boldsymbol{\lambda}}}{2(1-c_{\boldsymbol{\lambda}}\eta)}, \qquad 0 < \eta < c_{\boldsymbol{\lambda}}^{-1}. \tag{49}$$

The scale bounds are part of the method specification and must follow from a valid analytic envelope for the deployed statistic. An independent prefix may choose among envelopes already proved valid, but it cannot turn a plug-in scale estimate into a conditional moment-generating-function (MGF) bound. Runs without such an envelope are reported as empirical screens and do not invoke Corollary 10. If $\bar{Z}_m$ averages the $n_m$ observations available at look $m$, define

$$L_m(\boldsymbol{\lambda}) = \bar{Z}_m(\boldsymbol{\lambda}) - \sqrt{\frac{2v_{\boldsymbol{\lambda}}x_r}{n_m}} - \frac{c_{\boldsymbol{\lambda}}x_r}{n_m}. \tag{50}$$

For the support branch, conditionally iid gate observations and endpoints $\ell_{\boldsymbol{\lambda}}, u_{\boldsymbol{\lambda}}$ fixed in $\mathcal{F}_{m_0}$ give the Bessel variance

$$\widehat{V}_m(\boldsymbol{\lambda}) = \frac{1}{n_m - 1}\sum_{j=m_0+1}^{m}\{Z_j(\boldsymbol{\lambda}) - \bar{Z}_m(\boldsymbol{\lambda})\}^2.$$

Writing $R_{\boldsymbol{\lambda}} = u_{\boldsymbol{\lambda}} - \ell_{\boldsymbol{\lambda}}$, the lower bound is exactly Equation (48). This branch estimates variance on the gate stream but fixes its support before that stream; observed pilot extrema are not substituted for the endpoints. PLACID stops at the first scheduled look with a positive lower bound, selects the deterministic maximizer, and freezes it on the unseen suffix. If no lower bound crosses zero, it sets $M = N$ and $\boldsymbol{\lambda}_M = \mathbf{0}$.

A simple auditable sufficient condition is a deterministic pre-gate bound $|Z_j(\boldsymbol{\lambda})| \le b_{\boldsymbol{\lambda}}$ almost surely. Then the centered contrast has variance at most $b_{\boldsymbol{\lambda}}^2$ and absolute value at most $2b_{\boldsymbol{\lambda}}$, so Bernstein's MGF inequality permits $v_{\boldsymbol{\lambda}} = b_{\boldsymbol{\lambda}}^2$ and $c_{\boldsymbol{\lambda}} = 2b_{\boldsymbol{\lambda}}/3$. This statement requires a support bound implied by the frozen environment and actor interface; the maximum observed in a finite pilot is not such a bound. Charging Bernstein the worst-case range in place of the variance the frozen law has is what makes this envelope conservative rather than invalid. For the scalar coefficient in the support-EB study, write $A_j = \sum_b X_{j,b}^{\mathsf{T}} H_{j,b}$ and $C_j = \sum_b \|H_{j,b}\|_2^2$. If $|A_j| \le G$ and $0 \le C_j \le H$, then for every real $\lambda$,

$$-(2|\lambda|G + \lambda^2 H) \le Z_j(\lambda) \le 2|\lambda|G, \qquad R_\lambda = 4|\lambda|G + \lambda^2 H.$$

Thus an analytic actor/environment envelope supplies the asymmetric support used by Equation (48). Multiplying the statistic by a deterministic $c_B > 0$ multiplies the endpoints and range by $c_B$ and the sample variance by $c_B^2$, so the stopping sign is invariant.

*Proof. Step 1 (simultaneous coverage).* Conditional on $\mathcal{F}_{m_0}$, the sub-gamma Chernoff bound gives failure probability at most $\delta_r/(|\mathcal{T}||\Lambda|)$ for each scheduled look and grid point. For the support branch, apply Theorem 4 of Maurer and Pontil (2009) to the reflected variable $(u_{\boldsymbol{\lambda}} - Z_j)/R_{\boldsymbol{\lambda}} \in [0,1]$ (the $R = 0$ case is exact). Its upper deviation bound becomes the lower bound Equation (48) after rescaling. A union bound therefore covers every $(m, \boldsymbol{\lambda}) \in \mathcal{T} \times \Lambda$ with probability at least $1 - \delta_r$, without requiring independence across candidates or looks.

*Step 2 (deployment).* On this simultaneous event, the coefficient selected at the stopping look satisfies $\mathcal{D}_M(\boldsymbol{\lambda}_M) \ge L_M(\boldsymbol{\lambda}_M) > 0$. Comparing Equation (44) with the same realized prefix and an uncorrected suffix removes $(N-M)\mathcal{D}_M/N^2$ from the conditional covariance trace.

*Step 3 (abstention).* If no boundary is crossed, $M = N$ and $\boldsymbol{\lambda}_M = \mathbf{0}$. Equation (12) then equals the all-data sparse update sample by sample. Simultaneous coverage validates both the repeated looks and the finite-grid selection. □

## G Estimator-Level Positioning

**Learning from demonstrations, preferences, and offline data.** Bui et al. (2024) learn cooperative policies from demonstrations using factorized local and joint soft value functions. O-MAPL learns rewards and policies from offline preferences (Bui et al., 2025c); Preference-Guided Learning incorporates preferences into online sparse-reward MARL (Bui et al., 2025b); and ComaDICE addresses distribution shift in offline cooperative MARL (Bui et al., 2025a). These methods expand how a joint score or policy can be learned. Our question is downstream: for a frozen score and an on-policy shared actor, does the induced centered correction reduce the covariance of the complete gradient?

**Role diversity and parameter sharing.** Hu et al. (2022) use role diversity to diagnose parameter sharing, communication, and credit assignment. Their shared-value analysis bounds fitted-Q error under a shared function class. Here the actor is fixed, and the question is whether a centered credit correction improves the covariance of its complete policy-gradient estimate.

Zhao and Li study reward and likelihood-ratio aggregation with disjoint agent parameter blocks (Zhao and Li, 2026). Their comparison varies these aggregation supports. Here the behavior policy and shared coordinates are fixed, and we ask whether teammate marginalization preserves a local variance guarantee after summation.

**Control-variate interface.** For $\bar{Q}_i(\mathbf{s}, a_i) = \mathbb{E}_{\mathbf{a}_{-i}} Q(\mathbf{s}, a_i, \mathbf{a}_{-i})$, the exact-marginal, unit-coefficient actor term has the residual-plus-marginal form

$$S_i\{Y - B - Q + \bar{Q}_i\} = S_i(Y - B - Q) + S_i\bar{Q}_i. \tag{51}$$

The second term exposes the teammate-marginal object used by local critic methods, while the first retains the observed-outcome residual. Their sum preserves the target under score misspecification.

The teammate correction is a conditional control variate. Since $S_i$ is measurable given $(\mathbf{s}, a_i)$,

$$S_i d_i[Q] = S_i Q - \mathbb{E}_{\mathbf{a}_{-i}}[S_i Q \mid \mathbf{s}, a_i]. \tag{52}$$

Equation (52) makes explicit the teammate-action centering used by PLACID. Q-Prop, trajectory-wise control variates, and doubly robust policy gradients retain sampled outcomes while using approximate action values (Gu et al., 2017; Cheng et al., 2020; Huang and Jiang, 2020); multi-fidelity policy gradients do the same with a cross-simulator auxiliary (Liu et al., 2026). Factorized baselines condition on teammates and integrate the focal action (Wu et al., 2018). Optimal-baseline analysis instead minimizes per-agent trajectory-gradient variance and quantifies other-agent exploration noise (Kuba et al., 2021). PLACID evaluates a different object: the candidate correction after every agent–time term that moves one shared parameter block has been summed, retaining the cross terms in Equation (39). The local-to-total optimal-baseline derivation of Wu et al. (2018) assumes pointwise score orthogonality $S_i^{\mathsf{T}} S_j = 0$ for $i \neq j$ (their Section 4.2), including separate function approximators without weight sharing. Theorem 1 supplies a bounded exact-critic reversal with locally optimal coefficients when those shared-coordinate cross terms remain; Proposition 3 establishes the insufficiency of the specified local summaries. PLACID centers teammate actions and couples that outcome-retaining correction to one-pass prospective selection after complete shared-actor aggregation. Geffner and Domke select control-variate weights from a full gradient vector (Geffner and Domke, 2020); PLACID combines this selection geometry with shared-MARL aggregation and suffix-only deployment.

Lapeyre–Lelong's adaptive estimator $N^{-1}\sum_j H(\theta_{j-1}, X_j)$ explicitly uses one sample stream both to adapt $\theta$ and to estimate the mean; adaptive control-variate simulation develops the corresponding sequential tuning principle (Kim and Henderson, 2007; Lapeyre and Lelong, 2011). With the raw estimator encoded at $\theta = 0$, Equation (12) is the one-switch choice $\theta_{j-1} = 0$ for $j \leq M$ and $\theta_{j-1} = (Q_M, \boldsymbol{\lambda}_M)$ for $j > M$. PLACID supplies the target-neutral teammate correction and complete shared-actor selection unit for this predictable information pattern. Proposition 4 then derives the finite-$N$ covariance trace for this bounded one-switch information pattern, retaining the dependence between $M$ and $\rho_M$ and giving samplewise all-$N$ fallback.

Other reuse interfaces make different choices. StackMC and K-fold policy-gradient baselines correct samples out of fold through reciprocal fits (Tracey and Wolpert, 2016; Kota et al., 2017); split

control functionals omit their fitting set (Oates et al., 2017); symmetric cross-fitting applies reciprocal corrections (Chernozhukov et al., 2018; Kallus and Uehara, 2020). Recent methods adapt or stop CV pilots without retaining this binary raw-prefix/frozen-suffix interface (Wu and Kawai, 2026; Coons et al., 2025; Pan et al., 2026).

Table 1: Component-level precedents for the admission and reuse interface. This sub-interface combines outcome-retaining teammate centering, complete shared-actor selection, a retained raw prefix, and one-switch prospective deployment with exact fallback. Conditional sampling refinement is treated separately in Section 5.4 and Appendix D.

| Work | actor/control-variate interface | reuse or selection interface | interface supplied by PLACID |
|---|---|---|---|
| Q-Prop / DR-PG / TrajCV / MFPG | outcome plus auxiliary CV | action, trajectory, or simulator samples | teammate-centered shared-MARL deployment |
| Wu et al. | focal-action baseline | orthogonal scores for local-to-total optimality | nonorthogonal shared-block admission |
| Kuba et al. | per-agent trajectory-gradient variance | optimal baseline / fixed-estimator analysis | stopped complete-block correction admission |
| ASAE / PERLA | teammate-marginal critic signal | Monte Carlo teammate actions | retained outcome residual for any frozen score |
| GPAE | focal-marginal per-agent value operator | replay / off-policy correction | outcome-retaining on-policy correction |
| Geffner–Domke | full-vector CV-weight selection | generic estimator pool | raw-prefix / frozen-suffix contract |
| Kim–Henderson / Lapeyre–Lelong | predictable adaptive CV | same-stream sequential adaptation | one-switch MARL covariance law and exact fallback |
| StackMC / K-fold / split CF | learned CV fitted separately | reciprocal fits or omitted fit set | single retained selecting prefix |
| Wu–Kawai / Coons / EVPO | adaptive or stopped CV selection | fresh stopped batch or gated baseline | retained raw prefix with MARL block gate |
| PLACID | outcome-retaining teammate correction | one raw prefix; one frozen corrected suffix | complete prospective interface |

COMA integrates the focal action in a baseline (Foerster et al., 2018); ROLA, ASAE, and PERLA form critic-based local or teammate-marginalized actor signals (Xiao et al., 2021; Wan et al., 2021; Jafferjee et al., 2025), while GPAE builds a focal-marginal per-agent value operator (Kim et al., 2026); DAE trades explicit shaping against policy bias (Li et al., 2022). Centralized-critic analyses show why critic information and policy-gradient validity must be kept separate (Lyu et al., 2023). State-action and trajectory control variates retain a Monte Carlo residual but center different random variables (Tucker et al., 2018; Cheng et al., 2020). The actor-level distinction central to PLACID is explicit in Equations (51) and (52): the observed outcome remains in the estimator and an arbitrary fixed auxiliary enters only through a zero-mean correction.

## H Exact-Enumeration Checks

**What is being checked.** These experiments are algebraic validation rather than learning benchmarks. Small cooperative games are exhaustively enumerated so that policy-gradient means, covariance matrices, proposal moments, recycling identities, and abstention can be compared with exact population values. The goal is to verify that the implementation matches the estimator analyzed in the theory before moving to fitted scores and neural actors.

**All-team and sharp-window checks.** An 80-digit evaluation brackets the first reciprocal-binomial root for $n \in \{2, 3, 4, 5, 8, 16, 32, 64, 100\}$ and checks its monotonic derivative, $np_n \in (1/2, 2/3)$, bounded reward, nonzero local targets, and the exact shared correction moments. The largest residual in $\mathbb{E}[1/Z] = 0$ is below $3 \times 10^{-64}$. Balanced odd-team checks cover $n \in \{3, 5, 7, 11, 31\}$. The

sharp-window polynomial identity is evaluated on 48 four-action cells, including both sides of the boundary, with maximum discrepancy below $6 \times 10^{-82}$. Exact rational Gaussian moments verify the quadratic coordination example and finite budgets $K \in \{1, 2, 4, 8, 16\}$.

**Structural separations.** The four action atoms in Proposition 3 have identical local joint laws and prediction mean-squared error but opposite signs of complete-block reduction. Exact enumeration also checks that disjoint parameter supports remove cross-agent inner products, while shared supports retain them. These finite checks validate the constructions; the proofs establish the full parameterized results.

We verify the identities used by the current estimator through exhaustive enumeration in float64. For small cooperative games, the harness enumerates the joint action space, all teammate-resample tuples, and the outcome noise, producing exact means and covariance matrices. The target-preservation, finite-$K$ conditional-moment, and fixed-split checks cover 144 settings that vary team size, action space, interaction order, score error, policy skew, and $K$; the recycling rows exercise fixed and adaptively stopped screens.

Table 2: Exact-enumeration checks used by PLACID. Errors are maximum absolute float64 discrepancies; the populated recycling rows summarize 960 fixed-screen cells and 36 stopped settings.

| Identity | Settings | Max abs. error | Violations |
|---|---|---|---|
| $\mathbb{E}[g_i^{\lambda}(Q)] = \mathbb{E}[S_{i,t}Y]$ for frozen $Q$, $\lambda$, $K$ | 144 | $7.8 \times 10^{-16}$ | 0 |
| finite-$K$ conditional decision moments, Eq. (41) | 144 | $2.5 \times 10^{-14}$ | 0 |
| fixed discarded-split ratio $(1 - \bar{\rho})/q$ | 144 | $4.6 \times 10^{-14}$ | 0 |
| fixed-$M$ screen-and-recycle identity, Eq. (13) | 36 | $2.2 \times 10^{-16}$ | 0 |
| bounded-stopping screen-and-recycle identity, Eq. (13) | 36 | $2.2 \times 10^{-16}$ | 0 |
| samplewise all-data fallback at abstention | 36 | 0 exactly | 0 |

Table 3: One-at-a-time assumption perturbations over 16 games and 384 checks. Each row checks the corresponding alternative error law induced by removing one nominal condition. For the fitted-coefficient row, 2.2% is the maximum relative gap $|b_N/b_N^{(1)} - 1|$ at the largest batch in each sweep, where $b_N$ is the enumerated infinity norm of the bias and $b_N^{(1)}$ is its leading-order prediction. The other displayed errors are absolute numerical discrepancies.

| Perturbation | condition removed | primary identity affected | law check |
|---|---|---|---|
| resample follows realized action | resample independence | finite-$K$ moments | exact, $2.7 \times 10^{-15}$ |
| coefficient fitted on actor fold | exogenous coefficient | gradient mean | asymptotic, 2.2% |
| coupled behavior actions | factorized policy | gradient mean | exact, $7.4 \times 10^{-15}$ |
| baseline reads focal action | action-free baseline | gradient mean | exact, $2.5 \times 10^{-15}$ |
| correction repeated | one behavior-point use | optimal reduction | exact, $1.2 \times 10^{-13}$ |
| resample branches share a draw | branch independence | finite-$K$ variance | exact, $4.8 \times 10^{-15}$ |

These checks validate the algebra and its executable implementation. The repeated-batch study then measures the complete estimator error with a newly fitted score and selecting prefix in every replication; the neural actor-loop study evaluates prospective admission across policy refreshes.

### H.1 Local screens against exact shared truth

This study asks how often a screen formed on local terms admits a correction that harms the shared gradient. It enumerates 1,566 cooperative games with 2–4 agents, 2–3 actions, horizon 2–3, additive, pairwise and higher-order payoffs, six frozen credit-function families and three parameter layouts: one block per agent, two shared blocks, and one block shared by every agent. The exact teammate marginal is used and every trajectory is enumerated, so each screen's decision is scored against the exact complete-block reduction $\mathcal{D}(\lambda)$. Every screen chooses from the same grid $\{0\} \cup \Lambda$. The agent-diagonal screen maximizes the sum of the agents' own reductions and therefore drops the

cross-agent terms; the decision-level screen also drops cross-time terms; the complete-block screen is Equation (11) with population moments. A false deployment selects $\lambda \neq 0$ with $\mathcal{D}(\lambda) \leq 0$.

Table 4: Local screens against exact complete-block truth on 1,566 enumerated games, using population moments. "Mismatched $\lambda$" counts games in which both the agent-diagonal and the complete-block screen deploy but choose different coefficients. "Max. increase" is the largest shared covariance-trace increase caused by a deployed coefficient, relative to the raw estimator.

| | | Agent-diagonal | | | Decision-level | Complete block |
|---|---|---|---|---|---|---|
| Parameter layout | Games | False | Mismatched $\lambda$ | Max. increase (%) | False | False |
| One block per agent | 594 | 0 | 0 | 0.00 | 3 | 0 |
| Two shared blocks | 378 | 7 | 27 | 3.67 | 7 | 0 |
| One shared block | 594 | 35 | 97 | 17.56 | 34 | 0 |

With one parameter block per agent the omitted cross-agent terms vanish, and the agent-diagonal and complete-block screens agree in every game, as they must. Once agents share parameters, the agent-diagonal screen deploys harmful corrections and chooses different coefficients; with a single shared block it deploys 35 harmful corrections, one raising the shared covariance trace by $17.56\%$, and in $5\%$ of its deployments it retains at most $-11.2\%$ of the complete-block optimum. The decision-level screen fails even without sharing because it also drops cross-time terms. The complete-block screen makes no false deployment in any layout. Re-running the study on the current code reproduces every selector record exactly.

# I Detailed Algorithm and Implementation

**Implementation contract.** The guarantees depend on information flow within a policy refresh. The score and coefficient must be selected before the corrected suffix is observed; every prefix trajectory keeps its raw sparse contribution; proposal refinements may use the recorded suffix trajectory only after admission choices are frozen; and the certified correction is used for one actor step at the behavior policy. This appendix states that contract explicitly and records the recurrent, masking, normalization, and block-specific details used in the implementation.

**Fixed linear preconditioning.** The certificate concerns the one-step task gradient in actor coordinates. For a linear preconditioner $P$ fixed before screening, admission may instead screen $|PX_j|^2 - |P(X_j - \lambda H_j)|^2$ using valid transformed support bounds fixed before screening; its expectation is the covariance-trace reduction of the preconditioned gradient.

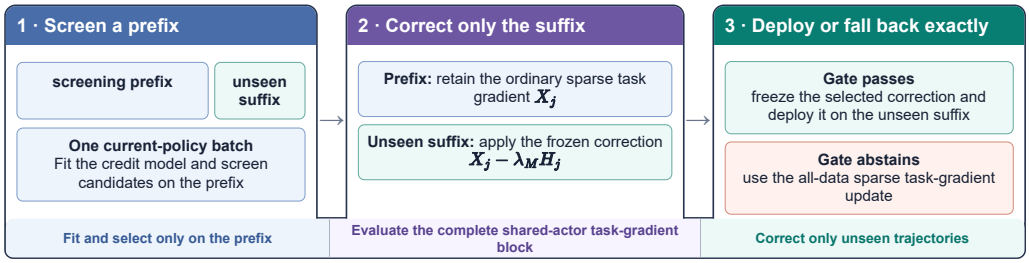


Figure 6: PLACID retains the screening prefix in the sparse task-gradient update and applies a frozen learned correction only to the independent suffix. If the screen abstains, the method recovers the all-data sparse update.

Algorithm 1 gives the certified refresh; PLACID-E uses the prespecified empirical screen in Section 5.3. The implementation details below apply to the admitted teammate-centered correction.

**Dependency clusters and information flow.** A complete episode is a dependency cluster when its reset is regenerative: conditional on the frozen refresh history, successive complete episode records

**Algorithm 1** PLACID-C at one behavior-policy refresh

**Require:** A fixed behavior policy, baseline, parameter blocks, and normalization; cluster budget $N$; model-prefix size $m_0$; coefficient grid $\Lambda$; scheduled evaluation times $\mathcal{T}$; teammate-sampling budget $K$; and confidence level $\delta_r$.

Fix the cluster ordering independently of returns, episode lengths, and completion times.
Fit and, if needed, select the credit function $Q$ using only clusters $1{:}m_0$, and then hold $Q$ fixed.
Establish valid bounds for the admission statistics, and initialize $M = N$ and $\lambda_M = 0$.
**for** each scheduled time $m \in \mathcal{T}$ with $m - m_0 \geq 2$ **do**
  Compute $L_m^{\mathrm{EB}}(\lambda)$ for all $\lambda \in \Lambda$ using the available screening clusters.
  **if** $\max_{\lambda \in \Lambda} L_m^{\mathrm{EB}}(\lambda) > 0$ **then**
    Set $M = m$, select the deterministic maximizing coefficient $\lambda_M$, and hold it fixed for the suffix.
    Stop screening.
  **end if**
**end for**
Retain the ordinary contribution $X_j$ for every $j \leq M$.
For each unseen suffix cluster $j > M$, generate teammate-action samples using the fixed sampling rule and form $X_j - \lambda_M H_j(Q_M)$.
Average all $N$ cluster contributions according to Equation (12). If no correction was admitted, the result is exactly $\widehat{G}^0$.
Take one full-batch actor step at the behavior policy using Equation (53).

Table 5: Allowed information flow within one frozen behavior-policy refresh. Every prefix trajectory retains its ordinary sparse contribution. Suffix data do not select the score, strength or budget; a prespecified refinement rule may condition its proposal strata on the recorded trajectory.

| Role | objects it may determine | actor contribution | fixed before next role |
|---|---|---|---|
| Model prefix $1{:}m_0$ | score family, $Q$, hyperparameters | $\mathbf{X}_j$ | $Q$ |
| Gate prefix $m_0{+}1{:}M$ | stopping look, $\boldsymbol{\lambda}_M$ | $\mathbf{X}_j$ | $M, \boldsymbol{\lambda}_M$ |
| Unseen suffix $M{+}1{:}N$ | strata under frozen rule | $\mathbf{X}_j - \boldsymbol{\lambda}_M \odot \mathbf{H}_j(Q)$ | choices precede proposals |
| Actor step | nothing | uniform $1/N$ aggregate | parameters move once |

have independent copies of the same law. Episodes coupled by persistent random state, a common random opponent state, or an initial latent not fixed in that conditioning remain together in one fixed-size cluster. Cluster construction and normalization are identical throughout a refresh. Before collection, fix a permutation $\Pi$ independently of the sigma-field generated by every cluster's states, actions, outcomes, lengths, and completion times; ordering by any realized cluster field is disallowed even when it ignores outcomes. Thus the ordered $N$ clusters have the same frozen behavior law. The score may use previously collected data, whose simulator steps and compute are accounted separately; only current-policy clusters enter Proposition 4. Baseline parameters, the scalar normalizer, and recurrent/mask handling rules are frozen before the stream; realized histories and masks remain pre-action fields stored inside their clusters. The suffix never selects $Q$, $K$, a coefficient, a stopping time, or a normalizer.

**Gate statistic and selection.** For every model candidate, hyperparameter selection is completed inside clusters $1{:}m_0$. The gate then observes complete block vectors from subsequent prefix clusters. One scalar $Z_j(\boldsymbol{\lambda})$ sums all actor blocks within the dependency cluster. The gate uses either the conditional sub-gamma bound in Equation (50), whose constants are fixed before the stream, or the support-EB bound in Equation (48), whose endpoints are fixed before the stream and whose sample variance is covered by the inequality. Each study fixes its model prefix, look schedule and level before sampling (Appendix J). The base certified protocol's five coefficients $\{0.1, 0.25, 0.5, 0.75, 1\}$ are scalars broadcast across actor blocks; ties favor the smallest strength. The signed-grid variant of Appendix J.2 mirrors this grid to $\{\pm 0.1, \pm 0.25, \pm 0.5, \pm 0.75, \pm 1\}$. Appendix M describes a secondary empirical coefficient calibration. The theory also permits a prespecified finite grid of blockwise coefficient vectors.

**Transporting an admission through sampling refinement.** The optional sampling step conditions on the recorded suffix but changes neither the selected exact correction nor its conditional mean. Its rule and budget are fixed before production proposals; any direction calculation is detached from the

actor loss. By Theorem 5, refining an iid realization after the same frozen prefix weakly reduces its covariance, so a valid iid admission remains valid. When reporting the exact identity in Equation (13), $\rho_M$ refers to the realization actually deployed; the admitted iid reduction is a lower bound after this refinement.

**Scalar actor surrogate.** Each agent's policy belongs to one connected parameter block; policies sharing any trainable parameter belong to the same block. Let $r_{i,t}(\theta) = \pi_{i,\theta}(a_{i,t} \mid s_{i,t})/\pi_i^{\mathrm{old}}(a_{i,t} \mid s_{i,t})$, and let $\mathrm{sg}$ detach its entire argument. For block $b = b(i)$, set

$$A^{\mathrm{R}}_{j,i,t} = Y_j - B_{j,t} - \mathbf{1}\{j > M\}\lambda_{M,b} d^{(K)}_{j,i,t}[Q_M].$$

Define the scalar surrogate

$$\mathcal{J}_{\mathrm{R}}(\theta) = \frac{1}{N}\sum\nolimits_{j,t,i} r_{i,t}(\theta)\,\mathrm{sg}(A^{\mathrm{R}}_{j,i,t}), \qquad \nabla_\theta \mathcal{J}_{\mathrm{R}}(\theta)|_{\theta=\theta^{\mathrm{old}}} = \widehat{G}^{\mathrm{R}}. \tag{53}$$

Equation (53), with these block-indexed advantages, has gradient $\widehat{\mathbf{G}}^{\mathrm{R}}$ at the collecting policy. Thus score evaluations, the baseline, sampled proposals, and coefficients supply fixed numerical advantages; differentiation follows only the actor likelihood. The clipped PPO surrogate has this same derivative at $r_{i,t} = 1$.

The neural implementation uses the fixed padded decision average $\mathcal{J}_{\mathrm{impl}} = c_{\mathrm{pad}}\mathcal{J}_{\mathrm{R}}$, where $c_{\mathrm{pad}} = 1/(T_{\max}|\mathcal{N}|)$. Its task-gradient is therefore $c_{\mathrm{pad}}\widehat{\mathbf{G}}^{\mathrm{R}}$; the sparse comparator has the identical scale. Covariances and reduction statistics scale by $c^2_{\mathrm{pad}}$, preserving all trace ratios, coefficient choices and samplewise fallback. In Shared-Relay, $T_{\max} = 1$, $|\mathcal{N}| = 2$ and $c_{\mathrm{pad}} = 1/2$. Reported block-vector diagnostics use the unscaled sums defined in Section 3.

**Actor implementation.** In the base iid branch, each focal agent's $K$ counterfactual teammate actions are iid draws from the exact factorized $\boldsymbol{\pi}^{\mathrm{old}}_{-i}$. In the optional stratified branch, draws are independent across equal-mass cells but not identically distributed; their equally weighted mixture preserves the same teammate expectation. Both branches respect the stored action masks and detach proposals from the actor gradient. Coupled focal streams retain their declared joint law within each dependency component, as specified in Appendix D. The prefix and suffix sparse terms use the same terminal outcome, baseline, and uniform cluster weight, followed by the common fixed normalization above. Only suffix terms include the frozen correction. The full-batch gradient is evaluated at the collecting policy before one optimizer step moves its parameters. A new rollout refreshes the gate. In the neural actor-loop study, the entropy term is common to both arms and subsequent PPO epochs use the sparse surrogate, as specified in Appendix J.

For a recurrent actor, reconstruct the hidden states from recorded histories and differentiate through the parameter-dependent history computation defining $\pi_{i,\theta}(a_{i,t} \mid h_{i,t})$. Recorded environment observations and action masks remain fixed inputs. The resulting score includes the recurrent encoder's parameter dependence in both $X$ and $H$; the numerical actor studies here use feedforward networks. Appendix L summarizes the native-horizon SMAC and MA-MuJoCo complete-gradient checks.

**Score interface.** The estimator requires frozen evaluations of a joint score. The reported experiments use outcome regression. The same interface accepts the published implicit $Q, V$ construction (Bui et al., 2025b) when fitted and selected inside the model prefix.

## J Experimental Details and Reporting

**Organization.** The experimental appendix mirrors the three levels separated in the main paper. Controlled studies measure complete-estimator error against known population gradients; native-simulator studies retain trajectory and fitting variability; and proposal-realization studies condition on fixed trajectories and scores to isolate auxiliary sampling error. Actor-loop experiments are reported separately because a covariance certificate for one refresh is not a return guarantee across training.

### J.1 EXACT RECYCLING AND GATE CALIBRATION

The exact harness enumerates fixed splits and bounded stopping rules, including $M \in \{m_0, \dots, N\}$, $\lambda = 0$, negative reduction, and data-dependent $Q_M, \lambda_M$. It checks the mean, conditional suffix identity, total-covariance identity, discarded-split ratio, and all-data fallback over 960 fixed-screen cells and 36 stopped settings. The grid includes null, beneficial, anti-aligned, and harmful corrections; fallback is bitwise over $16{,}385{,}958$ enumerated samples.

The one-step family draws a factorized joint action and reveals $Y = f(\mathbf{a}) + \epsilon$, with additive, pairwise, and higher-order payoff tables. The short-horizon family uses $s_{t+1} = (s_t + \sum_i a_{i,t}) \bmod |\mathcal{S}|$ and terminal feedback. The exact checks also cover disjoint-support and shared-support actor blocks.

### J.2 REPEATED COMPLETE-ESTIMATOR COMPARISON

**Frozen task family.** The direct study uses four game shapes: $(n, T) = (2, 2)$ with additive payoff, $(2, 3)$ and $(3, 2)$ with pairwise payoff, and $(3, 3)$ with higher-order payoff. Each has three states, two actions per agent, a fully shared tabular softmax actor, and four independent payoff/policy instances. State transitions are as above, and only $Y = \sum_t u(s_t, \mathbf{a}_t) + \epsilon$ is observed, with independent symmetric $\epsilon \in \{-0.5, +0.5\}$. Backward induction precomputes a state-value baseline; it and the behavior policy are fixed across replications. The suite uses $N \in \{8192, 32768\}$ and 256 independent complete batches in every cell. The initial schedule-development panel and final confirmation each use all four shapes and four instances per shape. Confirmation uses new payoff/policy instances and independently sampled batches.

The initial state is $s_0 = 0$. For each frozen game, draw independent standard-normal arrays in this order: policy logits $L \in \mathbb{R}^{1\times3\times2}$, a state offset $b_s$, singleton tables $w_i(s, a_i)$ in agent order, pair tables $w_{ij}(s, a_i, a_j)$ in lexicographic $i < j$ order when present, and finally a dense joint-action table $w_{\text{all}}$ for higher-order games. The shared behavior table is $\pi(a \mid s) = \text{softmax}(1.1L_{0,s,:})_a$, and

$$u(s, \mathbf{a}) = b_s + \sum_i w_i(s, a_i) + 0.7 \sum_{i<j} w_{ij}(s, a_i, a_j) + 0.7 w_{\text{all}}(s, \mathbf{a}).$$

Additive games omit both interaction terms; pairwise games omit the dense term. These payoff coefficients are drawn once per game, without rescaling, and reused at every time step. Arrays and exact moments use float64 on CUDA with PyTorch 2.11.0.

**Paired sampling.** The development and confirmation panels use independent random streams for game construction and batch sampling. Batches are generated in chunks of 16 replications: draw $16N$ categorical trajectory atoms with replacement, reshape to $(16, N)$, and draw independent outcome-noise signs of the same shape. Within each game and budget, all score families and estimators receive the same sampled batches.

**Fresh score fitting.** Every batch is divided into two halves, each with a model quarter followed by a gate quarter. A joint-action table is fit solely from its own model quarter:

$$\widehat{Q}(t, s, \mathbf{a}) = \frac{\sum_{j \in \mathcal{I}_{\text{fit}}} \mathbf{1}\{s_{j,t} = s, \mathbf{a}_{j,t} = \mathbf{a}\} Y_j}{1 + \sum_{j \in \mathcal{I}_{\text{fit}}} \mathbf{1}\{s_{j,t} = s, \mathbf{a}_{j,t} = \mathbf{a}\}}.$$

The denominator supplies a fixed unit ridge penalty. Additive regression sums the $n$ empirical $(t, s, a_i)$-conditional outcome means and subtracts $(n - 1)$ times the $(t, s)$-conditional mean, each with the same unit smoothing. Sign reversal applies to the fitted joint table. The exact teammate marginal is evaluated from each frozen fitted table and behavior policy. The score fit uses sampled terminal outcomes. The selector receives the independently established support endpoints and observed gate statistics; exact conditional-value tables and policy-gradient moments are reserved for setup and evaluation.

**Matched estimators.** Sparse uses every raw gradient. Default PLACID fits on the first quarter, screens once at the end of the second quarter, and uses the selected correction on the final half. Its matched discarded split uses that same score and coefficient but averages only the final half, including when the screen abstains. The early-crossing PLACID variant instead checks every 32 gate

observations and starts correcting after its first positive simultaneous bound. Both PLACID variants recover the all-$N$ sparse estimator at abstention. Ungated recycling uses the fixed $N/2$ boundary and $\lambda = 1$. Both reciprocal controls fit and screen in each half before correcting the opposite half, one with single looks and the other with early crossing. Their outputs average all $N$ clusters and also equal the all-data sparse estimator when both fold screens abstain.

The base positive-grid certified arms use $\Lambda = \{0.1, 0.25, 0.5, 0.75, 1\}$ and deterministic support endpoints from the fitted score's images of every supported trajectory and both outcome-noise signs. Support depends on the fitted score and environment support independently of the associated gate observations. The refresh level is $0.05$; cross-fitting allocates $0.025$ to each fold. A one-look screen pays only its five-candidate multiplicity; scheduled screens also pay for their frozen looks. Exact gradient moments enter the evaluation after these decisions.

**Full-vector endpoint and uncertainty.** For every complete batch and method, we record $\|\widehat{\mathbf{G}} - \boldsymbol{\mu}\|^2/(\mathcal{V}_{\mathrm{raw}}/N)$, where $\mathcal{V}_{\mathrm{raw}}$ is the trace of the per-cluster raw-gradient covariance, retaining all cross-fold terms in the cross-fitted vector. The PLACID prediction is $1 - (1 - M/N)\rho_M$, where the selected correction's $\rho_M$ is evaluated against the exact law after routing. Averages compare actual complete-estimator error with this prediction. Per-cell output also records coverage over every scheduled grid–look pair, false deployments, stopping fractions, coefficient choices, and sample-mean vectors. The 16 games are fixed strata with equal weight. We bootstrap 19,999 sets of complete-batch repetitions within each game; the same resampled indices are used across all score families and methods, preserving their pairing. The two default-PLACID versus sparse budget contrasts use $0.0125, 0.9875$ quantiles for a joint 95% Bonferroni interval family. Other displayed intervals are pointwise 95% intervals for this fixed suite. The confirmation analysis rule and all replication counts were locked before its outcomes.

Table 6: Independent confirmation: actual normalized mean-squared error (NMSE) of the complete estimator and pointwise 95% paired-stratified bootstrap intervals. The population sparse value is one; comparisons use paired differences on the same batches.

| Estimator | $N = 8192$ NMSE | $N = 32768$ NMSE |
|---|---|---|
| Sparse | 0.9934 [0.9624,1.0247] | 0.9920 [0.9617,1.0226] |
| Discarded split (one look) | 1.9075 [1.8500,1.9658] | 1.8536 [1.8005,1.9076] |
| PLACID-C (one look) | 0.9722 [0.9417,1.0028] | 0.9509 [0.9225,0.9795] |
| Cross-fit (one look) | 0.9541 [0.9242,0.9843] | 0.9183 [0.8908,0.9457] |
| PLACID-C (early crossing) | 0.9884 [0.9574,1.0196] | 0.9843 [0.9543,1.0145] |
| Cross-fit (early crossing) | 0.9863 [0.9553,1.0175] | 0.9811 [0.9512,1.0113] |
| Ungated recycling | 1.0629 [1.0313,1.0947] | 1.0627 [1.0319,1.0938] |
| PLACID-E (one look) | 0.9454 [0.9164,0.9749] | 0.9400 [0.9122,0.9680] |

Table 7: Confirmation NMSE by fitted-score family, averaging all 16 fixed games. Reversed applies the negative fitted joint table as an orientation control for the frozen positive coefficient grid; every arm receives the same batch within each game and budget.

| | $N = 8192$ | | | $N = 32768$ | | |
|---|---|---|---|---|---|---|
| Estimator | Joint | Additive | Reversed | Joint | Additive | Reversed |
| Sparse | 0.9934 | 0.9934 | 0.9934 | 0.9920 | 0.9920 | 0.9920 |
| Discarded split (one look) | 1.8066 | 1.9172 | 1.9987 | 1.6533 | 1.8897 | 2.0179 |
| PLACID-C (one look) | 0.9487 | 0.9744 | 0.9934 | 0.9002 | 0.9605 | 0.9920 |
| Cross-fit (one look) | 0.9116 | 0.9572 | 0.9934 | 0.8254 | 0.9375 | 0.9920 |
| PLACID-C (early crossing) | 0.9833 | 0.9886 | 0.9934 | 0.9748 | 0.9861 | 0.9920 |
| Cross-fit (early crossing) | 0.9788 | 0.9865 | 0.9934 | 0.9669 | 0.9845 | 0.9920 |
| Ungated recycling | 0.8951 | 0.9793 | 1.3142 | 0.8792 | 0.9759 | 1.3329 |
| PLACID-E (one look) | 0.8912 | 0.9516 | 0.9934 | 0.8757 | 0.9521 | 0.9920 |

**PLACID-E: matched empirical calibration.** PLACID-E evaluates empirical coefficient selection on the same confirmation batches, retaining all 16 games, three score families, both budgets and 256 repetitions. It uses the identical first-quarter fitted score and maximizes the mean of Equation (11)

on the second quarter over $\{0, 0.1, 0.25, 0.5, 0.75, 1\}$, with zero selected whenever the maximum is nonpositive. Both quarters remain raw, and only the second half receives the frozen correction. No population moment enters fitting or selection. This is a retrospective matched-control addition; the original seven methods and prespecified primary interval family are unchanged. Its 19,999 whole-batch bootstrap draws retain the same fixed-game strata and pairing across families and methods, with pointwise 95% intervals.

Its complete NMSE is $0.9454$ $[0.9164, 0.9749]$ and $0.9400$ $[0.9122, 0.9680]$, respectively. Relative to the certified one-look estimator on the same batches, it removes an additional $2.67$ $[2.26, 3.09]$ and $1.09$ $[0.84, 1.35]$ percentage points of sparse population error. It deploys on $7695/12288$ and $7680/12288$ refreshes. At the smaller budget, the selected direction has strictly negative population reduction in one joint-fit, 27 additive-fit and two reversed-fit refreshes; the larger budget has none. These counts audit the selected direction, separately from its realized batch error. Among the 30 selections, population reduction $\rho$ ranges from $-1.496\%$ to $-0.018\%$ and averages $-0.470\%$. Their half-batch suffix exposure gives mean and maximum penalties of $0.235\%$ and $0.748\%$ of sparse population error. Their harmful-only contribution averaged over all 12,288 refreshes is $0.000574$ percentage points. These are population covariance-accounting terms, not observed errors on individual batches. Tables 6 and 7 retain every score family. Thus empirical and certified admission are compared on both complete-estimator error and the population quality of the directions actually admitted.

Deterministic replay first reconstructs 73,728 original sparse, ungated and one-look certified records: coefficients and Boolean diagnostics agree exactly, gradient discrepancies are at most $4.45 \times 10^{-16}$, and NMSE discrepancies are at most $5.07 \times 10^{-14}$. The added empirical row then contributes 24,576 records on those same sampled batches. Its separately timed supplied-gradient core includes its own fit, correction, empirical gate and final vector: three warmups and ten repetitions on each game/family/budget timing cell give 960 measured calls, with median $0.0350/0.0454$ ms per refresh at the two budgets. This core endpoint has the same exclusions as the original computation study below.

**Budget and grid sweep.** The same 16 games, score families, level, split, and 256 repetitions were rerun at $N \in \{128, 256, \ldots, 4096\}$ alongside the two confirmation budgets, whose batches replay exactly (per-cell NMSE means agree to $9 \times 10^{-16}$), with the positive grid $\Lambda$ and the mirrored grid $\Lambda \cup (-\Lambda)$. Tables 8 and 9 report sparse population error removed with pointwise paired stratified 95% intervals; the sweep uses new bootstrap draws, while the locked confirmation analysis retains the original simultaneous intervals for positive-grid PLACID-C. The certified rule admits nothing below $N = 1024$ on either grid, admits 12,159 corrections with the positive grid and 17,942 with the signed grid, and none of them has negative population reduction. The empirical rule admits in 56–63% of refreshes with the positive grid and in 97–100% with the signed grid; at $N = 128$, 1,741 of its 6,918 positive-grid admissions and 3,648 of its 11,960 signed-grid admissions have negative population reduction, with harmful-only contributions of $0.44$ and $0.97$ percentage points against net removals of $2.16\%$ and $3.68\%$; both contributions fall below $0.02$ points from $N = 2048$ on. On the sign-reversed family the signed-grid rules remove the same error as on the correctly oriented score ($11.63\%$ for PLACID-E and $8.93\%$ for PLACID-C at $N = 32{,}768$). Pairing $Q$ with $-Q$ and $\lambda$ with $-\lambda$ leaves $X - \lambda H(Q)$ unchanged, explaining the matched sign-flip results. The moment-ratio comparators use the same gate quarter. Write $\hat{A} = \hat{\chi}$ and $\hat{C} = \hat{\omega}$ for the means of $X_j^{\mathsf{T}} H_j$ and $\|H_j\|^2$, and let $s_A, s_C$ be their sample standard deviations divided by the square root of the gate-quarter size. The plug-in rule uses $\hat{A}/\hat{C}$ when $\hat{C} > 0$ and zero otherwise. The positive and signed one-standard-error shrinkage rules use $(\hat{A} - s_A)_+/(\hat{C} + s_C)$ and $\operatorname{sign}(\hat{A})(|\hat{A}| - s_A)_+/(\hat{C} + s_C)$, respectively. Ungated correction admits every refresh and increases error at every budget when pooled with the reversed control. With the two fitted critics alone it removes error, but PLACID-E removes more at every budget (Table 10).

**Schedule-development panel.** The initial panel used early-crossing PLACID, its matched discarded split, early-crossing reciprocal cross-fitting, one-look recycling, ungated recycling, and sparse. One-look recycling removed $0.01797$ and $0.04061$ normalized error units at the two budgets; early-crossing recycling removed $0.00361$ and $0.00912$. This motivated the one-look default and a confirmation in which the reciprocal baseline receives the same schedule. The full initial comparison is retained in Table 11.

Table 8: Sparse population error removed (%) by the PLACID admission rules across batch budgets and grids on the confirmation suite (one-look rules; ± marks the signed grid); pointwise paired stratified 95% intervals. Admission columns give deployments out of 12,288 refreshes, with deployments of negative population reduction in parentheses.

| $N$ | PLACID-C | PLACID-C ± | PLACID-E | PLACID-E ± | PLACID-C ± admits | PLACID-E ± admits |
|---|---|---|---|---|---|---|
| 128 | +0.00 [0.00,0.00] | +0.00 [0.00,0.00] | +2.16 [1.65,2.68] | +3.68 [2.82,4.55] | 0 (0) | 11,960 (3,648) |
| 256 | +0.00 [0.00,0.00] | +0.00 [0.00,0.00] | +2.76 [2.24,3.29] | +4.51 [3.64,5.41] | 0 (0) | 11,919 (2,677) |
| 512 | +0.00 [0.00,0.00] | +0.00 [0.00,0.00] | +3.15 [2.59,3.72] | +5.21 [4.24,6.22] | 0 (0) | 11,977 (1,712) |
| 1,024 | +0.00 [-0.00,0.01] | +0.00 [0.00,0.00] | +3.65 [3.09,4.21] | +6.33 [5.36,7.29] | 0 (0) | 12,056 (930) |
| 2,048 | +0.65 [0.53,0.77] | +0.70 [0.55,0.86] | +3.98 [3.42,4.54] | +6.82 [5.86,7.78] | 1,510 (0) | 12,151 (493) |
| 4,096 | +1.74 [1.45,2.03] | +2.79 [2.31,3.28] | +4.40 [3.81,5.00] | +7.54 [6.50,8.58] | 2,809 (0) | 12,203 (221) |
| 8,192 | +2.12 [1.83,2.43] | +3.39 [2.90,3.89] | +4.80 [4.23,5.37] | +8.36 [7.38,9.35] | 3,916 (0) | 12,201 (83) |
| 32,768 | +4.11 [3.62,4.62] | +7.01 [6.20,7.85] | +5.21 [4.60,5.81] | +9.21 [8.16,10.28] | 9,707 (0) | 12,245 (20) |

Table 9: Comparators on the same batches: reciprocal cross-fitting (one look, positive grid), the continuous plug-in ratio $\hat{\chi}/\hat{\omega}$, the one-standard-error shrinkage rule and its signed version, and ungated correction; pointwise paired stratified 95% intervals.

| $N$ | Cross-fit | Plug-in $\hat{\chi}/\hat{\omega}$ | Shrinkage | Shrinkage signed | Ungated |
|---|---|---|---|---|---|
| 128 | +0.00 [0.00,0.00] | -99.25 [-168.49,-47.24] | -0.68 [-2.63,0.81] | -1.22 [-4.73,1.44] | -6.18 [-6.61,-5.74] |
| 256 | +0.00 [0.00,0.00] | -50.91 [-120.31,-9.35] | +0.82 [-0.74,1.90] | +0.85 [-2.26,2.92] | -6.49 [-6.95,-6.05] |
| 512 | +0.00 [0.00,0.00] | -0.78 [-5.96,2.75] | +2.41 [1.87,2.91] | +3.71 [2.61,4.70] | -6.75 [-7.22,-6.30] |
| 1,024 | +0.00 [0.00,0.00] | +5.70 [4.57,6.82] | +3.32 [2.88,3.76] | +5.64 [4.89,6.39] | -7.17 [-7.66,-6.69] |
| 2,048 | +0.83 [0.71,0.96] | +6.99 [5.90,8.09] | +3.87 [3.38,4.37] | +6.65 [5.82,7.49] | -6.50 [-6.93,-6.07] |
| 4,096 | +3.14 [2.79,3.51] | +7.68 [6.57,8.81] | +4.36 [3.82,4.91] | +7.49 [6.57,8.42] | -6.89 [-7.37,-6.42] |
| 8,192 | +3.93 [3.53,4.35] | +8.62 [7.56,9.68] | +4.81 [4.28,5.35] | +8.35 [7.43,9.26] | -6.95 [-7.39,-6.50] |
| 32,768 | +7.37 [6.74,8.03] | +9.30 [8.18,10.45] | +5.21 [4.60,5.83] | +9.17 [8.13,10.24] | -7.06 [-7.54,-6.61] |

**Computational accounting.** Every arm uses the same current-policy sample budget. Both PLACID schedules, discarded splitting and ungated recycling each fit one score; reciprocal cross-fitting fits two. Certified arms compute support and gradient statistics for each corresponding fit. On an RTX 4080 SUPER, isolated method calls receive precomputed trajectory gradients and a frozen teammate-marginal linear map, and include their own score fits, support, screen and output-vector construction. Each fixed game/budget cell supplies a batch of 16 refreshes shared by every method and score. After three warmups per method, ten timed repetitions rotate the method order and synchronize the GPU before and after each call. Times divided by 16 measure estimator-core throughput per refresh. Sampling, map construction, exact-truth evaluation, bootstrap and output writing are separate operations. All planned gate contrasts are vectorized; stopping controls which data may select a correction, while the recorded compute includes the complete gate allocation. The 6,720 timed calls cover every game, score, budget and method. Table 12 reports pooled medians and interquartile ranges; split by budget, one-look PLACID costs $0.105914/0.226925$ ms per refresh and its reciprocal counterpart costs $0.207340/0.443958$ ms. The separately measured cold map construction has median $1.043$ ms per game, and batch sampling has median $0.868$ ms per batch of 16. These measurements describe the estimator core with supplied trajectory gradients and the frozen marginal map.

### J.3 Changing-Actor Scope Check

This secondary check asks whether complete-block certification remains operational while the actor changes; it does not establish a general training-return improvement.

**Changing-actor Shared-Relay check.** Shared-Relay is a one-step cooperative game in which a worker predicts a uniform bit and a marker's action does not affect reward. Both agents update one shared actor. Ten paired runs compare PLACID with the same actor loop using zero correction, across 30 refreshes of $N = 8192$ episodes per run. The score is fitted on a 2,048-episode model prefix. The finite coefficient grid and scheduled looks are fixed before each gate stream, and exact enumeration of the one-step atoms supplies valid support endpoints. All 300 per-refresh certificates admit a correction; the held-out exposure-weighted covariance reduction is $7.065\%\,[6.320, 7.797]$. The paired difference in normalized return-curve area is $0.000005\,[-0.000004, 0.000015]$. Thus certification can operate while the actor changes, but the observed return difference is negligible.

Table 10: Fitted critic families only (joint table and additive fit; the sign-reversed orientation control is excluded): sparse population error removed (%) and paired differences with ungated correction (percentage points), with paired stratified 95% bootstrap intervals over the same batches.

| $N$ | PLACID-E $^{\pm}$ | PLACID-E | Ungated | PLACID-E $^{\pm}$ −ungated | PLACID-E −ungated |
|---|---|---|---|---|---|
| 128 | +3.07 [2.28,3.89] | +3.43 [2.68,4.21] | +1.58 [0.57,2.62] | +1.49 [0.82,2.18] | +1.85 [1.33,2.38] |
| 256 | +4.09 [3.28,4.90] | +4.26 [3.49,5.05] | +2.14 [1.15,3.14] | +1.95 [1.33,2.57] | +2.12 [1.63,2.62] |
| 512 | +4.82 [3.96,5.71] | +4.95 [4.12,5.79] | +3.47 [2.46,4.49] | +1.35 [0.82,1.90] | +1.48 [1.05,1.91] |
| 1,024 | +5.57 [4.70,6.45] | +5.54 [4.70,6.39] | +3.95 [2.94,4.96] | +1.62 [1.13,2.13] | +1.59 [1.20,1.98] |
| 2,048 | +6.03 [5.18,6.91] | +5.96 [5.14,6.82] | +4.56 [3.62,5.54] | +1.47 [1.06,1.88] | +1.40 [1.07,1.73] |
| 4,096 | +6.63 [5.71,7.54] | +6.59 [5.70,7.48] | +5.33 [4.31,6.33] | +1.30 [0.93,1.69] | +1.27 [0.97,1.57] |
| 8,192 | +7.43 [6.54,8.32] | +7.20 [6.35,8.05] | +5.62 [4.67,6.56] | +1.81 [1.41,2.23] | +1.58 [1.29,1.88] |
| 32,768 | +8.00 [7.07,8.94] | +7.81 [6.90,8.72] | +6.45 [5.45,7.45] | +1.55 [1.16,1.98] | +1.36 [1.07,1.67] |

Table 11: Initial schedule-development panel: actual NMSE with pointwise 95% intervals. It has the same fixed shapes and score families as confirmation but independent payoff/policy instances.

| Estimator | $N = 8192$ NMSE | $N = 32768$ NMSE |
|---|---|---|
| Sparse | 1.0216 [0.9890,1.0552] | 1.0119 [0.9808,1.0437] |
| Discarded split | 1.9599 [1.8987,2.0216] | 1.8699 [1.8144,1.9262] |
| PLACID-C (one look) | 1.0037 [0.9719,1.0365] | 0.9713 [0.9420,1.0012] |
| PLACID-C (early crossing) | 1.0180 [0.9856,1.0515] | 1.0028 [0.9720,1.0342] |
| Cross-fit (early crossing) | 1.0162 [0.9838,1.0495] | 0.9980 [0.9673,1.0292] |
| Ungated recycling | 1.1131 [1.0803,1.1468] | 1.0994 [1.0682,1.1309] |

## K Shared-Actor Training

This section gives the protocol behind Section 6.3 and the results for every suite–critic pair.

**Games.** The tabular suite is the 16-game suite of Section 6.1 ($n \in \{2, 3\}$, two actions, horizon two or three, three states, the same payoff tables and outcome noise). The team suite has 12 one-step games, four for each $n \in \{4, 8, 12\}$, with additive and pairwise payoff terms of the same family (interaction strength 0.7). The dispersion suite has 12 one-step games, three for each $(n, A) \in \{(3, 3), (4, 4), (5, 4), (6, 4)\}$: each agent chooses one of $A$ resources, resource $r$ has value $v_r \sim \mathrm{Unif}[0.5, 1.5]$, and the team outcome is $\sum_r v_r \mathbf{1}\{r \text{ is chosen by some agent}\}$. All games add symmetric two-point outcome noise of variance 0.25. Returns are normalized by the best and worst deterministic team policies.

**Actor and critics.** All agents share one actor with two tanh layers of width 32; its input is the one-hot time, state and agent identity. The last layer starts at 0.01 of the usual scale, so every run starts near the uniform policy. The actor ascends the estimated gradient with Adam (step 0.003) for 300 updates of $N = 64$ episodes. The neural critic has two tanh layers of width 64, reads time, state and the joint action, and takes 20 Adam steps (step 0.01) per update on the last eight batches. The coordination-graph critic is the exact ridge least-squares fit of main and pairwise joint-action effects on the same eight batches. Both critics are refitted after the update that used the current batch, so the correction at update $k$ depends only on earlier batches and has exact mean zero. The baseline is the critic's on-policy value, and the correction uses the exact teammate marginal of the fitted critic. Methods share architectures, initial weights and fitting budgets; each method's data and fitted critic evolve with its own policy.

**Admission in training.** The empirical, local and plug-in rules screen on the previous batch, re-evaluated under the current actor, critic and baseline, and correct the whole current batch. Let $\mathcal{F}_{k-1}$ contain the history before batch $k$. Its actor $\theta_k$, fitted critic, baseline rule and selected coefficients are fixed given this history. Fresh current-policy draws give zero-mean correction terms by Theorem 9, so the batch gradient estimate obeys $\mathbb{E}[\widehat{G}_k \mid \mathcal{F}_{k-1}] = \nabla_\theta J(\theta_k)$. Past-batch screening provides a predictable empirical coefficient. This lagged variant corrects all current data, with suffix fraction $q = 1$. PLACID-E maximizes the complete-vector empirical reduction over $\{0\} \cup \Lambda \cup (-\Lambda)$; local admission maximizes each agent's own reduction over the same grid and applies one coefficient per

Table 12: Isolated estimator-core throughput on RTX 4080 SUPER, with PyTorch 2.11.0/CUDA 12.8. Each value summarizes 960 timed batch calls, divided by 16 refreshes per batch. Method time includes its own score fits, support, screen and output gradient. Trajectory-gradient construction, frozen-map setup, sampling and diagnostic evaluation are outside this timing. Brackets give interquartile ranges.

| Estimator | score fits | ms/refresh: median [Q25,Q75] |
|---|---|---|
| Sparse | 0 | 0.0045 [0.0032,0.0076] |
| Discarded split (one look) | 1 | 0.1828 [0.1049,0.2220] |
| PLACID-C (one look) | 1 | 0.1915 [0.1059,0.2269] |
| Cross-fit (one look) | 2 | 0.3451 [0.2074,0.4440] |
| PLACID-C (early crossing) | 1 | 0.2084 [0.1106,0.2299] |
| Cross-fit (early crossing) | 2 | 0.4062 [0.2093,0.4444] |
| Ungated recycling | 1 | 0.0251 [0.0215,0.0325] |

agent; plug-in applies $\hat{\chi}/\hat{\omega}$; ungated uses $\lambda = 1$. Within each paired run, methods start from identical actor and critic weights and use shared uniforms for inverse-CDF sampling. Their sampled episodes follow their respective evolving policies. Each game has 32 paired runs.

**Exact measurements.** The score of every agent–time decision is a row of the actor Jacobian at the current weights, so every per-episode gradient is a coefficient vector times this Jacobian. Population moments therefore follow exactly from the Gram matrix of the Jacobian and the enumerated trajectory law. At every update we record the normalized return, the exact conditional MSE of each method given its selected coefficient (normalized by the sparse population MSE), and, along the sparse run, the local and shared moments of the learned critic and of the exact on-policy critic. A reversal is an update in which the per-agent optimal coefficients $\lambda_i = \chi_{ii}/\omega_{ii}$, each minimizing its own agent's variance, increase the trace covariance of the complete shared gradient. The code reproduces the witness of Proposition 3 exactly ($\mathrm{Var}\, X = 25/32$, local reductions $1/8$, $\mathcal{D}_{\pm}(1) = -1/4,\ 3/4$), and its exact moments agree with 200,000-episode Monte Carlo estimates. Return AUC is the mean normalized return over the 300 recorded policy iterates. Policy sharpness averages the maximum action probability over all represented time–state–agent inputs. The learned and exact diagnostics each use their own critic's policy-value baseline.

**Results.** Table 13 reports paired return-AUC differences with pointwise 95% intervals from 19,999 bootstrap draws of complete paired runs within fixed games. Table 14 gives exact complete-gradient MSE by learning phase, Table 15 reversal frequency by policy sharpness, and Table 16 the sizes of those reversals. On dispersion games with the coordination-graph critic, reversal occurs in $0.2\%$ of updates with the learned critic and $0.0\%$ with the exact critic.

| | | PLACID-E return-AUC difference ($\times 10^{-3}$) versus | | | |
|---|---|---|---|---|---|
| Suite | Critic | sparse | ungated | local | plug-in |
| Dispersion | graph | $+22.54\,[17.58, 27.55]$ | $-0.18\,[-3.66, 3.26]$ | $+2.10\,[-1.52, 5.67]$ | $+1.71\,[-2.26, 5.64]$ |
| Dispersion | neural | $+26.50\,[21.78, 31.22]$ | $-1.48\,[-5.13, 2.14]$ | $+2.46\,[-1.00, 5.85]$ | $+3.35\,[-0.57, 7.29]$ |
| Team | graph | $+5.20\,[3.59, 6.96]$ | $+0.27\,[-0.35, 1.15]$ | $+0.74\,[0.02, 1.66]$ | $+2.94\,[1.80, 4.17]$ |
| Team | neural | $+3.16\,[1.58, 4.70]$ | $+0.04\,[-1.02, 1.15]$ | $-0.67\,[-1.61, 0.24]$ | $+3.11\,[1.55, 4.78]$ |
| Tabular | neural | $+0.44\,[-0.82, 1.70]$ | $+0.71\,[-0.38, 2.18]$ | $-0.97\,[-2.25, 0.25]$ | $+1.96\,[0.66, 3.38]$ |
| Tabular | graph | $+0.67\,[-0.56, 1.84]$ | $-0.74\,[-1.81, 0.23]$ | $+0.25\,[-0.45, 1.00]$ | $+1.18\,[-0.05, 2.39]$ |

Table 13: Paired differences in normalized return-curve area with pointwise 95% bootstrap intervals over complete runs within fixed games; positive values favor lagged PLACID-E.

| Suite | Critic | Learning phase PLACID-E | ungated | local | plug-in | Converged phase PLACID-E | plug-in |
|---|---|---|---|---|---|---|---|
| Dispersion | graph | 0.58 | 0.58 | 0.58 | 0.59 | 0.99 | 20.66 |
| Dispersion | neural | 0.57 | 0.57 | 0.57 | 0.58 | 0.98 | 19.52 |
| Team | graph | 0.34 | 0.32 | 0.35 | 0.38 | 1.02 | 30.05 |
| Team | neural | 0.37 | 0.35 | 0.37 | 0.42 | 1.01 | 1518.62 |
| Tabular | neural | 0.81 | 0.78 | 0.79 | 78.48 | 1.01 | 3516.68 |
| Tabular | graph | 0.83 | 0.82 | 0.82 | 20.31 | 1.01 | 406.77 |

Table 14: Exact complete-gradient MSE of the deployed update divided by the sparse population MSE, averaged over updates. Learning and converged phases are updates with mean maximum action probability below and at least $0.9$.

| Mean max. probability | Team, graph | Team, neural | Tabular, neural | Tabular, graph |
|---|---|---|---|---|
| $[0.5, 0.6)$ | 0.0 / 0.0 | 1.0 / 0.0 | 1.7 / 0.0 | 0.0 / 0.0 |
| $[0.6, 0.7)$ | 0.0 / 0.0 | 0.0 / 0.0 | 0.7 / 2.0 | 0.9 / 2.0 |
| $[0.7, 0.8)$ | 0.0 / 0.0 | 0.0 / 0.0 | 6.8 / 8.2 | 7.4 / 7.9 |
| $[0.8, 0.9)$ | 0.3 / 0.0 | 0.2 / 0.0 | 17.8 / 18.5 | 15.9 / 18.4 |
| $[0.9, 0.95)$ | 2.2 / 0.0 | 0.7 / 0.0 | 28.8 / 32.7 | 35.6 / 33.9 |
| $[0.95, 0.99)$ | 3.0 / 0.1 | 2.8 / 0.0 | 22.8 / 29.6 | 33.2 / 30.0 |
| $[0.99, 1]$ | 19.8 / 0.2 | 6.6 / 0.0 | 35.9 / 63.4 | 69.0 / 63.0 |
| All updates | 12.3 / 0.1 | 4.5 / 0.0 | 26.0 / 38.1 | 41.8 / 38.5 |

Table 15: Percentage of sparse-training updates in which per-agent optimal credit increases shared-gradient covariance trace, learned critic / exact on-policy critic.

| Suite, trained critic | Diagnostic | Rate (%) | Median (%) | 99th (%) | Above 1% (%) |
|---|---|---|---|---|---|
| Dispersion, graph | Learned | 0.2005 | 0.0016 | 0.1743 | 0.0000 |
| Dispersion, graph | Exact | 0.0000 | — | — | 0.0000 |
| Dispersion, neural | Learned | 0.4644 | 0.0054 | 0.9986 | 0.0052 |
| Dispersion, neural | Exact | 0.0000 | — | — | 0.0000 |
| Team, graph | Learned | 12.2535 | 0.0528 | 1.1137 | 0.1293 |
| Team, graph | Exact | 0.1389 | 0.0045 | 0.1498 | 0.0000 |
| Team, neural | Learned | 4.5182 | 0.0351 | 1.9734 | 0.1042 |
| Team, neural | Exact | 0.0017 | 0.0082 | 0.0090 | 0.0000 |
| Tabular, neural | Learned | 26.0404 | 0.0366 | 9.8040 | 1.4342 |
| Tabular, neural | Exact | 38.0625 | 0.0315 | 7.3840 | 1.1270 |
| Tabular, graph | Learned | 41.7643 | 0.0592 | 5.3961 | 1.3392 |
| Tabular, graph | Exact | 38.5254 | 0.0298 | 6.9331 | 1.1992 |

Table 16: Reversal magnitudes under per-agent population-optimal coefficients on sparse-training trajectories. Increases are normalized by the raw covariance trace for each diagnostic. Rate and the last column count all updates; the last column requires an increase above 1%. Median and 99th percentile use only updates with a reversal. Dashes indicate no reversals.

### K.1 Finite teammate proposals

Practical counterfactual estimators replace the exact teammate marginal by $K$ sampled teammate actions. Fresh proposals add to $\mathbb{E}\|H_K\|^2$ while preserving $\mathbb{E}[X^\mathsf{T} H_K]$, so the shared optimum $\chi/\omega_K$ depends on the proposal budget. We rerun the team and dispersion suites with the graph critic, changing only the correction of decision $(t, i)$ to $Q(\mathbf{a}) - K^{-1}\sum_k Q(a_i, \tilde{\mathbf{a}}^k_{-i})$ with fresh draws from the current teammate policies; every other setting, game, paired run and architecture is that of the exact-marginal runs, and nothing was tuned. The protocol and decision rule were fixed before the runs. For independent proposal streams across decisions $d = (i, t)$, let $V_d = \mathrm{Var}_{\tilde{\mathbf{a}}_{-i}} Q(\mathbf{s}_t, a_{i,t}, \tilde{\mathbf{a}}_{-i})$. The proposal contribution to $\mathbb{E}\|H_K\|^2$ is $K^{-1}\sum_d \mathbb{E}[\|S_d\|^2 V_d]$. We weight these terms by the deployed coefficients when computing complete-gradient MSE. A 200,000-episode Monte Carlo

check reproduces it within 0.3% for $K \in \{1, 3\}$. Table 17 reports paired return-AUC differences as in Table 13.

| | | PLACID-E return-AUC difference ($\times 10^{-3}$) vs | | Ungated vs sparse | Complete-gradient NMSE | |
|---|---|---|---|---|---|---|
| $K$ | Suite | sparse | ungated | ($\times 10^{-3}$) | PLACID-E | ungated |
| 1 | Team | $+2.29\,[0.73, 3.95]$ | $+1.66\,[0.35, 3.00]$ | $+0.62\,[-1.09, 2.42]$ | 0.995 | 1.028 |
| 1 | Dispersion | $+8.94\,[3.80, 14.08]$ | $+9.08\,[4.73, 13.38]$ | $-0.14\,[-4.89, 4.60]$ | 0.802 | 1.047 |
| 2 | Team | $+2.28\,[0.57, 4.04]$ | $+0.23\,[-1.14, 1.56]$ | $+2.05\,[0.74, 3.55]$ | 0.967 | 0.959 |
| 2 | Dispersion | $+14.84\,[9.79, 19.73]$ | $+5.06\,[0.94, 9.19]$ | $+9.77\,[4.82, 14.59]$ | 0.753 | 0.842 |

Table 17: Shared-actor training with $K$ sampled teammate proposals (graph critic, 32 paired runs per game, 12 games per suite). Intervals are pointwise 95% bootstrap intervals over complete runs within fixed games; NMSE is the exact conditional complete-gradient MSE of the deployed update divided by the sparse population MSE, averaged over updates.

With one proposal, unit correction has complete-gradient NMSE above one on both suites. Lagged PLACID-E improves return AUC over sparse in the four graph-critic team/dispersion configurations and over unit correction on both suites at $K = 1$ and on dispersion at $K = 2$.

The recorded protocol also includes graph-critic tabular games at $K = 1$ and additive critics with exact teammate marginals on the team and dispersion suites. Table 18 gives the matched comparisons for all seven configurations, using the same paired-run analysis.

| | | | PLACID-E return-AUC difference ($\times 10^{-3}$) versus | | | |
|---|---|---|---|---|---|---|
| Suite | Critic | $K$ | Sparse | Ungated | Local | Plug-in |
| Team | Graph | 1 | $+2.29\,[0.73, 3.95]$ | $+1.66\,[0.35, 3.00]$ | $-0.36\,[-1.60, 0.86]$ | $+4.95\,[3.00, 7.11]$ |
| Dispersion | Graph | 1 | $+8.94\,[3.80, 14.08]$ | $+9.08\,[4.73, 13.38]$ | $-3.09\,[-7.00, 0.79]$ | $-4.45\,[-8.26, -0.66]$ |
| Team | Graph | 2 | $+2.28\,[0.57, 4.04]$ | $+0.23\,[-1.14, 1.56]$ | $-0.91\,[-2.30, 0.40]$ | $+3.00\,[1.17, 4.85]$ |
| Dispersion | Graph | 2 | $+14.84\,[9.79, 19.73]$ | $+5.06\,[0.94, 9.19]$ | $-2.69\,[-6.66, 1.27]$ | $-0.72\,[-4.61, 3.08]$ |
| Tabular | Graph | 1 | $+0.67\,[-0.43, 1.80]$ | $+0.23\,[-0.89, 1.32]$ | $+0.27\,[-0.60, 1.15]$ | $+5.35\,[3.49, 7.38]$ |
| Team | Additive | $\infty$ | $+2.20\,[0.40, 4.05]$ | $-0.55\,[-1.53, 0.56]$ | $+0.01\,[-1.22, 1.29]$ | $+2.07\,[0.80, 3.44]$ |
| Dispersion | Additive | $\infty$ | $+0.40\,[-3.93, 4.74]$ | $-2.76\,[-6.64, 1.22]$ | $+0.71\,[-3.35, 4.85]$ | $+1.50\,[-2.53, 5.65]$ |

Table 18: Matched paired comparisons for the finite-proposal and additive-critic protocol. Graph critics at $K = 1$ on team and dispersion games are the primary configurations; the others are secondary. Intervals are pointwise 95% bootstrap intervals over complete runs within fixed games.

## L Native-Simulator Complete-Gradient Scope Check

This fixed-policy study asks whether the conditional proposal gains in Appendix N are visible in the *complete* gradient after environment variation, score fitting, and admission. It uses native SMAC 8m (eight agents, masked discrete actions, horizon 120) and MA-MuJoCo CoupledHalfCheetah (two six-dimensional continuous-action agents, horizon 1000) (Samvelyan et al., 2019; de Witt et al., 2020). Both shared actors have two width-32 tanh layers. Three actor checkpoints per task were fixed before estimator evaluation.

At each checkpoint, 32 independent 64-episode batches are paired into 16 covariance replications. Within a batch, the first 16 episodes fit a fresh joint-observation/action outcome score, the next 16 screen a coefficient from $\{0, 0.1, 0.25, 0.5, 0.75, 1\}$, and only the final 32 receive the frozen correction. The model-prefix and gate-prefix episodes retain their raw gradient contributions. All methods use the same recorded episodes, baseline, score family, and $K = 4$ whole-joint teammate proposals. The complete shared-parameter vector is evaluated on every episode; no counterfactual simulator transitions are required. Covariance traces are estimated from half the squared distance between independent complete-batch gradients. The 19,999 paired bootstrap draws resample complete batch pairs within each checkpoint.

Empirical recycling reduces covariance relative to the all-data sparse estimator by $0.079\%\,[-0.007, 0.167]$ on SMAC and $0.047\%\,[-0.278, 0.398]$ on CoupledHalfCheetah. These intervals are inconclusive. On all 96 SMAC batches, the support-based empirical-Bernstein gate abstains, so its realized estimator is exactly the sparse estimator. The categorical support endpoints

are derived from fixed network-weight, mask, horizon, and score bounds before screening; their validity alone does not establish the episode-iid condition required for certificate interpretation. No analogous bounded envelope is assigned to the continuous Gaussian actor. This study therefore supplies a scope check, not evidence of native-task return improvement.

**Native cost.** On the first recorded batch at each of the three checkpoints, every method was timed from raw trajectories, including its own score fit, gate gradients, proposals, support construction and final complete gradient (one warmup, three rotated repetitions, CUDA). Charged with the recorded simulator collection of the same 64 episodes, including session startup and close, empirical recycling adds $0.57$–$0.86\%$ to the sparse time per batch on SMAC 8m and $2.30$–$3.29\%$ on CoupledHalfCheetah, and certified recycling adds $0.55$–$0.60\%$ on SMAC 8m. Reciprocal cross-fitting needs $1.33$–$2.08$ times the standalone computation of recycling, because it fits two scores and evaluates twice as many gate gradients. Timing uses float32 actor derivatives. On the last CoupledHalfCheetah checkpoint, the two accumulation paths differ by $4.35 \times 10^{-7}$ in relative gradient norm; a separate float64 replay agrees to $9.53 \times 10^{-16}$.

## M SMAC Calibration and Corrected-Observation Panel

A separate SMAC 8m reference actor probes whether better outcome prediction implies a better shared-gradient correction. The actor is prepared by native-reward PPO without any PLACID correction; the first prespecified readiness checkpoint wins 34 of 128 evaluation episodes. An observation adapter corrects an unseen-enemy feature-row error for this panel only. Earlier SMAC records retain their original observation implementation, and the two panels are not pooled.

The fresh panel has 32 independent groups of four 64-episode simulator sessions ($N = 256$ per group). The first session fits a joint outcome score, the second calibrates a coefficient from complete trajectory-gradient vectors, and the final two receive that frozen correction. Both a constant baseline $B_0 = 34/128$ and a separately fitted, frozen state-value baseline $V$ are evaluated on matched data. Complete groups, rather than episodes within a persistent simulator session, are the independent units for 16 paired covariance comparisons. The empirical regularized rule uses $(A - s_A)_+/(C + s_C)$, where $A$ and $C$ are the gate means of $X^\mathsf{T}H$ and $\|H\|^2$, and $s_A, s_C$ are their nominal one-standard-error penalties. This rule is a calibration proxy, not a finite-sample admission certificate.

Relative to sparse/$V$, its complete-gradient covariance reduction is $-0.406\%\,[-0.646, -0.041]$; the negative sign means increased covariance. Held-out trajectory Brier loss nevertheless falls from $0.193409$ under $B_0$ to $0.174792$ under $V$. Better prediction thus does not certify useful shared-gradient alignment. The corrected-observation actor and the first-session frozen scores also supply the separately locked proposal-realization confirmation in Appendix N.

## N Native-Trajectory Proposal Realization

We evaluate a frozen candidate teammate correction at unit coefficient on native SMAC 8m and CoupledHalfCheetah trajectories. The measured quantity is the covariance introduced by auxiliary action proposals, conditional on the complete recorded trajectory, actor and outcome score. This experiment holds the environment samples and exact-marginal correction fixed while changing its Monte Carlo realization. This diagnostic resamples the unweighted proposal vector $F_{\tau,m}$ defined below and does not require a positive gate decision. It is separate from the complete-gradient study, whose original SMAC certified gate abstains on all 96 batches. The SMAC records use the original observation implementation of Appendix L; the corrected-observation panel in Appendix M uses separate data.

**Frozen-policy protocol.** SMAC uses eight agents with masked categorical actions and a native episode limit of 120 steps. CoupledHalfCheetah uses two agents with six continuous action coordinates each and 1,000-step episodes. The shared actor has two 32-unit tanh hidden layers; the complete parameter blocks contain 4,110 and 2,182 trainable coordinates, respectively. The continuous policy retains the Gaussian latent action for likelihood-ratio scores and clips only the action executed by the simulator and evaluated by the outcome model. Its log standard deviation is fixed at $-0.7$. Both tasks use joint-observation/action outcome models with two 32-unit tanh hidden layers. Their original fits

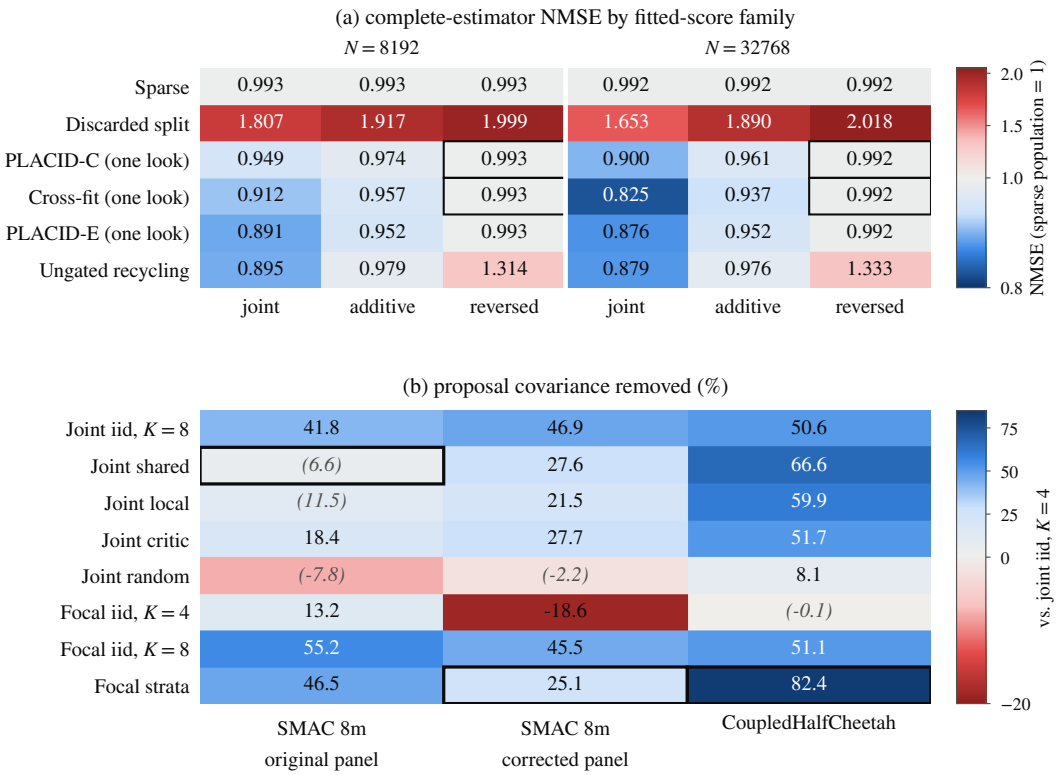


Figure 7: Two endpoints. (a) Positive-grid complete-estimator MSE by credit-function family and budget, normalized by the sparse population MSE (one). Outlined cells show sparse fallback under the reversed score. (b) Conditional proposal covariance removed on fixed native trajectories. The reference is joint iid with $K = 4$; positive values mean less teammate-sampling noise for the same frozen correction. Rows with $K = 8$ double the proposal budget. Gray parenthesized values have pointwise 95% intervals including zero; outlines mark prespecified primary comparisons.

use only the first 16 trajectories of each 64-trajectory batch, with 48 Adam steps and minibatches of at most 1,024 decisions. The fitted models and all actor parameters remain unchanged during proposal evaluation.

Each task contributes all three prescheduled actor checkpoints from native-reward preparation. At each checkpoint we take the third and fourth complete batches and evaluate the first and seventeenth trajectories of their untouched suffixes, after the 16-trajectory model prefix and the following 16 screening trajectories. This specifies four trajectories per checkpoint and 12 per task, disjoint from the earlier exploratory sampling pilot. All recorded time steps, available-action masks and alive-agent indicators are retained. The resulting SMAC trajectories contain 30–120 steps; all continuous trajectories contain 1,000 steps. No environment transition or model fit is added by this evaluation.

**Methods and dependency structure.** Joint-iid uses a common joint policy draw for all focal counterfactuals. Joint-shared, joint-local and joint-critic stratify that same joint noise vector using, respectively, the complete shared-score Gram, its diagonal, and critic sensitivity without actor-score weighting. Joint-random uses an independently drawn direction fixed before production proposals. Focal-iid and focal-stratified use independent proposal streams for each focal term; the latter normalizes that term's own sensitivity. We retain all nine alternatives in Table 20, including both joint and independent-focal eight-proposal controls. Every four-proposal method makes $4nT$ production counterfactual score evaluations on a trajectory of length $T$ with $n$ agents; the eight-proposal controls make twice that number.

The prespecified dependency-aware output is an alias for joint-shared in SMAC and focal-stratified in CoupledHalfCheetah, not an additional sampled alternative. For two agents, the first focal integrand

depends only on the second agent's auxiliary noise and vice versa. These blocks are independent under the joint-iid law, so splitting them preserves that law while permitting a separate stratum direction for each block. For the generic eight-agent joint score, focal integrands share auxiliary coordinates and the complete joint block is retained. Independent-focal alternatives in the latter setting provide a separately reported coupling comparison.

**Policy transports and direction construction.** At each recorded decision, fix the focal executed action and evaluate the score's action derivative with the teammates at their policy means. For Gaussian actions, the latent coordinates have scale $\exp(-0.7)$, the actor mean is a tanh output, and the sensitivity multiplies the executed-action derivative by that scale and the derivative of clipping at the mean. For categorical actions, use the soft probability vector as the expansion point, but discrete one-hot actions in every production score query. If action $a$ occupies CDF interval $(c_{a-1}, c_a]$, its exact first Gaussian moment is

$$\mathbb{E}[\mathbf{1}\{T(U) = a\}U] = \phi(\Phi^{-1}(c_{a-1})) - \phi(\Phi^{-1}(c_a)).$$

Contract this vector with the score's soft-action derivative to obtain each categorical sensitivity. This is a frozen direction heuristic, not a derivative of the discontinuous inverse-CDF map. The focal latent coordinates have zero sensitivity because that action is held fixed.

For a unit direction $v$ in a component's standard-normal input space, independently draw $W_k \sim \mathrm{Uniform}(0,1)$ and $R_k \sim \mathcal{N}(0, I)$ and set

$$U_k = v\Phi^{-1}\left(\frac{k - 1 + W_k}{K}\right) + (I - vv^{\mathsf{T}})R_k, \qquad k = 1, \ldots, K.$$

Residuals are independent across strata, components and recorded decisions. Categorical proposals apply the masked policy inverse CDF to $\Phi(U_{k,j})$; continuous proposals execute $\mathrm{clip}(m_j + \exp(-0.7)U_{k,j}, -1, 1)$. The complete score Gram uses all actor layers: if $h_{i,\ell-1}$ is a layer's input and $\delta_{i,\ell}$ its log-likelihood backward factor, then

$$\mathsf{G}_{ii'} = \sum_{\ell} (\delta_{i,\ell}^{\mathsf{T}} \delta_{i',\ell})(h_{i,\ell-1}^{\mathsf{T}} h_{i',\ell-1} + 1).$$

The added one includes bias coordinates. The leading direction of $A^{\mathsf{T}}\mathsf{G}A$ is computed exactly, with a fixed coordinate fallback at a zero matrix; the two-agent implementation uses its rank-two factorization. No proposal draw selects its own direction. Covariance preservation uses the cell-mixture identity and does not rely on the accuracy of either linearization.

**Full-trajectory endpoint and uncertainty.** For each fixed trajectory $\tau$ and method $m$, we generate 256 independent realizations of the full proposal-gradient vector

$$F_{\tau,m}^{(r)} = \sum_{t,i} S_{i,t} \frac{1}{K_m} \sum_{k=1}^{K_m} Q(s_t, a_{i,t}, \widetilde{a}_{-i,t}^{(r,k)}).$$

Every agent and time step is summed before taking a norm. The raw trajectory gradient and observed-action correction term are fixed under this conditioning, so the covariance of $F_{\tau,m}$ is precisely the proposal component of the correction covariance. We estimate its trace using 128 disjoint replicate pairs:

$$\widehat{V}_{\tau,m} = \frac{1}{128} \sum_{r=1}^{128} \frac{1}{2} \left\| F_{\tau,m}^{(2r-1)} - F_{\tau,m}^{(2r)} \right\|_2^2.$$

Task-level estimates average the 12 fixed-trajectory estimates before forming the reduction ratio. We use 19,999 bootstrap replicates, resampling pair indices within each fixed trajectory and sharing those indices across methods. The two primary task comparisons use 97.5% percentile intervals for the two-environment family; other reported intervals are pointwise 95% intervals.

**Descriptive matched-coupling contrast.** A secondary contrast calculated after confirmation compares focal-stratified directly with focal-iid, both at $K = 4$. Holding the independent-focal coupling fixed gives reductions of $38.43\%$ with interval $[26.80, 47.88]\%$ in SMAC and $82.45\%$ with interval $[81.89, 82.98]\%$ in CoupledHalfCheetah. These pointwise intervals describe the matched stratification operation and do not replace the prespecified SMAC joint-block comparison.

| Prespecified dependency-aware output | Reduction (%) | Family interval (%) |
|---|---|---|
| SMAC 8m: joint-shared | 6.62 | $[-14.81,\ 24.27]$ |
| CoupledHalfCheetah: focal-stratified | 82.43 | $[81.76,\ 83.09]$ |

Table 19: Primary conditional proposal-covariance comparisons against joint-iid with four proposals. The SMAC interval includes zero; the continuous-task interval excludes zero. Both are the original prespecified comparisons.

| Proposal realization | $K$ | SMAC 8m (%) | CoupledHalfCheetah (%) |
|---|---|---|---|
| Joint-iid | 4 | reference | reference |
| Joint-iid | 8 | $41.78\ [32.24, 49.78]$ | $50.62\ [48.98, 52.19]$ |
| Joint-shared | 4 | $6.62\ [-11.82, 22.05]$ | $66.61\ [65.53, 67.64]$ |
| Joint-local | 4 | $11.52\ [-4.08, 24.93]$ | $59.86\ [58.53, 61.11]$ |
| Joint-critic | 4 | $18.43\ [5.11, 30.11]$ | $51.67\ [50.04, 53.26]$ |
| Joint-random | 4 | $-7.76\ [-24.26, 6.71]$ | $8.05\ [4.97, 11.06]$ |
| Focal-iid | 4 | $13.15\ [0.15, 24.37]$ | $-0.12\ [-3.48, 3.12]$ |
| Focal-iid | 8 | $55.18\ [48.73, 60.70]$ | $51.12\ [49.39, 52.75]$ |
| Focal-stratified | 4 | $46.52\ [37.28, 54.25]$ | $82.43\ [81.83, 83.00]$ |

Table 20: All nine frozen proposal realizations, with reduction relative to joint-iid at $K = 4$ and pointwise 95% intervals. Positive values indicate lower conditional proposal covariance. Primary family-adjusted intervals appear separately in Table 19; this table does not select a replacement primary.

**Confirmation on new corrected-observation trajectories.** A separately locked panel repeats the realization comparison on twelve new trajectories: the first episode of the fourth session of twelve groups from the mixed-outcome panel of Appendix M, each with its group's archived global score, fitted on that group's first session only, and the corrected-observation actor. Focal-stratified realization against joint iid at $K = 4$ was fixed as the primary before execution, after the exploratory ranking above; the nine realizations, 256 replicates, 128 pairs and the bootstrap are unchanged. Table 21 reports every arm. The primary removes $25.09\%$ with 95% interval $[14.58, 34.58]$, and per-trajectory reductions range from $10.6\%$ to $54.6\%$. The earlier joint-shared primary removes $27.62\%$ $[17.05, 36.85]$ on this panel, and the matched independent-focal contrast is $36.85\%$ $[28.13, 44.51]$; independent-focal iid itself increases conditional covariance by $18.63\%$ relative to joint iid for eight agents. Doubled-budget iid remains the strongest comparison at $46.86\%$. The panel takes 38.6 seconds without environment episodes or score fits.

| Proposal realization | $K$ | New SMAC 8m panel (%) |
|---|---|---|
| Joint-iid | 4 | reference |
| Joint-iid | 8 | $46.86\ [39.13, 53.42]$ |
| Joint-shared | 4 | $27.62\ [17.05, 36.85]$ |
| Joint-local | 4 | $21.47\ [8.57, 32.65]$ |
| Joint-critic | 4 | $27.74\ [17.90, 36.40]$ |
| Joint-random | 4 | $-2.17\ [-17.73, 11.45]$ |
| Focal-iid | 4 | $-18.63\ [-34.41, -4.98]$ |
| Focal-iid | 8 | $45.55\ [37.91, 52.14]$ |
| Focal-stratified (primary) | 4 | $25.09\ [14.58, 34.58]$ |

Table 21: Confirmation on twelve new corrected-observation SMAC 8m trajectories: conditional proposal covariance reduction relative to joint-iid at $K = 4$. The prespecified primary uses a 95% percentile interval; other rows are pointwise 95% intervals.