AGGREGATE IN THE ADVANTAGE, NOT THE RATIO: A CANONICAL-FORM ANALYSIS OF COOPERATIVE MULTI-AGENT POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent policy optimization, exemplified by PPO-based methods, is a key branch of cooperative Multi-Agent Reinforcement Learning (MARL). A central design question is how many neighboring agents[1] to aggregate so as to effectively utilize global information for cooperation. This decision must be made along two dimensions: in the advantage (which agents' rewards contribute to the credit signal) and in the ratio (which agents' likelihood ratios form the clipped importance weight). We formalize these two design choices as a pair of support patterns, one for the advantage and one for the ratio, and prove a canonical structure: the expected multi-agent policy optimization objective depends on the two patterns only through a single combined quantity, the product of the ratio pattern and the advantage pattern. This yields two consequences: (i) Redundancy: the two patterns are interchangeable with respect to the resulting learning signal, so neither way of aggregating neighbors is inherently superior to the other. (ii) Variance Ordering: the advantage combines neighbors' rewards additively, so its variance grows only mildly and it admits an interior sweet spot where the aggregation range matches the coupling neighborhood, whereas the ratio combines neighbors' likelihood ratios multiplicatively, so its variance grows exponentially with the number of aggregated agents and brings no compensating reduction in bias. Together these results point to an unambiguous design principle: aggregate neighbors in the advantage, sized to the coupling neighborhood, and keep the ratio per-agent. This also explains why neighbor-based advantages are far more common than neighbor-based ratios across prior heuristic and empirical designs. We prove these results under specified assumptions and validate them across four carefully designed synthetic cooperative games and a real-world large-scale traffic-signal control task. The code for our experiments is available in an anonymous repository at `https://anonymous.4open.science/r/MAPO-803E`.

1 INTRODUCTION

Cooperative Multi-Agent Reinforcement Learning (MARL) built on policy optimization, represented by Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Trust Region Policy Optimization (TRPO) (Schulman et al., 2015a), faces a recurring question when updating each agent: How many other agents should be aggregated into that agent's update? This question is easy to overlook because it is answered twice, in two different places:

- **Advantage support** S^A: which agents' rewards are summed into agent i's advantage, the credit signal that indicates how good the joint action was for i; and
- **Ratio support** S^R: which agents' likelihood ratios are multiplied into agent i's clipped importance weight, the trust-region object that keeps the update near the behavior policy.

Both range from per-agent (aggregate no one else) to joint (aggregate everyone), and existing methods sit at different points on each: IPPO (De Witt et al., 2020) keeps both per-agent, while MAPPO

[1] In this paper, "neighbors" refer to agents whose actions influence one another, not merely those in physical proximity.

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.

(Yu et al., 2022) pairs a team advantage (all rewards) with a per-agent ratio. Traffic and networked methods use neighborhood advantages (Chu et al., 2019) with per-agent ratios. Sequential learning methods employ joint sequential ratio products while keeping the advantage per-agent (Kuba et al., 2022; Zhong et al., 2024). Casting the whole team as a single multi-action agent and running vanilla PPO on the factorized joint policy, with a joint ratio and joint advantage, yields a route taken by centralized single-agent reductions of MARL (Zhao et al., 2026; Zhao & Li, 2026). Yet the two supports have never been analyzed together, and there is no principle for where to sit on either, which this paper aims to address through canonical-form analysis.

Our starting point is that the two supports are not independent degrees of freedom. We prove (Theorem 1) that the expected multi-agent policy gradient depends on the pair (S^A, S^R) only through their matrix product $\widetilde{S} = S^R S^A$. This is a gauge freedom: S^A and S^R are interchangeable up to their product, so any cross-agent aggregation placed in the ratio can equally be placed in the advantage, and vice versa (Corollary 1). In particular, the ratio side carries no aggregation that the advantage side could not realize, which is the sense in which cross-agent ratios are redundant. What breaks this symmetry is variance, not signal. Aggregating rewards through the advantage is a sum of bounded terms (additive variance), whereas aggregating ratios is a product of importance weights, whose variance compounds multiplicatively in the number of factors (Lemma 2). Since the two routes realize the same expected gradient but the ratio route strictly inflates variance, the variance-optimal choice is to place all cross-agent aggregation in the advantage and keep the ratio per-agent (Corollary 2). It is in this precise, variance-ordered sense that cross-agent importance ratios are redundant: they add no signal that the advantage cannot, and cost variance that the advantage does not. This also explains a standing asymmetry in previous empirical and heuristic designs: neighborhood advantages are widely used, while local ratios are comparatively rare.

In conclusion, we contribute: (i) a canonical form showing that the two supports enter the expected gradient only through $\widetilde{S} = S^R S^A$, establishing their mutual redundancy; (ii) a variance-ordering that breaks the tie, since advantage sums are additive while ratio products are exponential, and that yields the design rule: aggregate in the advantage and keep the ratio per-agent; and (iii) an empirical validation across synthetic games and a real-world traffic signal control task that pin down when each effect appears. We discuss connections to prior work throughout, and provide an extended treatment in Appendix D.

2 PRELIMINARIES

Single-agent policy optimization. PPO (Schulman et al., 2017) optimizes a policy $\pi_{\boldsymbol{\theta}}$ by taking several gradient steps on a batch collected under a behavior policy $\pi_{\boldsymbol{\theta}_{\text{old}}}$, correcting the off-policy mismatch with a clipped importance ratio $\varrho(\boldsymbol{\theta}) = \pi_{\boldsymbol{\theta}}(a \mid s)/\pi_{\boldsymbol{\theta}_{\text{old}}}(a \mid s)$ and the advantage A, estimated by Generalized Advantage Estimation (GAE) (Schulman et al., 2015b):

$$L^{\text{PPO}}(\boldsymbol{\theta}) = \mathbb{E}\big[\min\big(\varrho A,\ \text{clip}(\varrho, 1-\epsilon, 1+\epsilon)A\big)\big]. \tag{1}$$

Two objects drive the update: the advantage A (the credit signal) and the ratio ϱ (the trust-region weight).

Cooperative multi-agent policy optimization. In a cooperative Markov game (Littman, 1994) (formalized in Appendix A) with n agents, each agent i has a policy $\pi^i_{\boldsymbol{\theta}}$; under Centralized-Training with Decentralized-Execution (CTDE) and Centralized-Training with Centralized-Execution (CTCE) schemes alike (Jin et al., 2025), the joint policy factorizes as $\pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid s) = \prod_i \pi^i_{\boldsymbol{\theta}}(a^i \mid s)$. Lifting (1) to n agents requires two design decisions that are usually made implicitly:

1. **Which reward drives agent i's advantage?** MAPPO (Yu et al., 2022) uses the team advantage (all agents share one A built from the global reward); independent learners (IPPO) (De Witt et al., 2020) use each agent's own reward; value-decomposition, difference-reward, and neighbor-advantage methods sit in between.
2. **Whose ratios enter agent i's clipped weight?** Independent and decentralized actors (IPPO, MAPPO) clip a per-agent ratio ϱ^i. A fully centralized controller that treats $\pi = \prod_i \pi^i$ as one policy clips the joint ratio $\prod_j \varrho^j$, in which every factor is differentiated jointly (Zhao et al., 2026). Sequential methods such as HAPPO (Kuba et al., 2022) form a compound ratio $\prod_{j \le i} \varrho^j$. We

study the jointly-differentiated case, whose supports range from per-agent ($S^R = I$) to fully joint ($S^R = \mathbf{1}\mathbf{1}^\top$).

These are the two knobs this paper isolates. We call the first the advantage support and the second the ratio support, and formalize both as $0/1$ matrices in Section 3.1.

Why the choice matters: a bias–variance tension. Both supports can reduce the same bias, but they pay for it very differently, a tension we preview here with a direct measurement of the multi-agent policy gradient and defer the full study to Sec. 4. On a single-step dense pairwise game ($n = 20$ agents, $K = 4$ actions each, ring coupling of radius $\rho^\star = 4$; full definition in Sec. 4.1), we fix a reference policy, estimate the true team-return policy gradient by Monte Carlo, and then form the clipped-surrogate gradient estimator induced by a given pair of supports. For that estimator we report three quantities (all in log scale): its squared bias, variance, and their sum, the Mean Square Error (MSE). A support of radius ρ means each agent aggregates its ρ-hop neighborhood on the coupling graph: $\rho = 0$ is per-agent (aggregate no one else), $\rho = \rho^\star$ exactly covers the true coupling neighborhood, and larger ρ over-aggregates.

Panels (a,b) vary one support at a time. Panel (a) sweeps the advantage radius ρ_A with the ratio held per-agent, and panel (b) sweeps the ratio radius ρ_R with the advantage held per-agent. The two bias curves are identical: enlarging either support removes the same missing-coupling bias, so on the expected gradient the two supports are interchangeable (a redundancy we prove in Sec. 3.1). The two variance curves differ sharply instead. The advantage aggregates rewards additively, so its variance grows gently and its MSE bottoms out near the coupling radius $\rho^\star$; the ratio aggregates likelihood ratios multiplicatively, so its variance grows much faster and its MSE turns up at a smaller radius. Panels (c,d) isolate this by fixing a matched effective support (so both realize the same expected gradient) and comparing the two ways of reaching it, aggregating through the advantage (path $\mathbf{P}$) versus through the ratio (path $\mathbf{Q}$). They plot the bias (c) and variance (d) of each path as the support grows; the bias tracks together while the variance separates by up to $9\times$. This is the question we take up: understanding why the benefit is the same yet the cost differs, and what that implies for where to sit on each axis.

3 Analysis: Two Supports, One Product

3.1 Problem Setup

We first analyze a single decision step. Remark 6 and the episodic experiment (Sec. 4) provide the finite-horizon extension.

Agents, policies, actions. Let $\mathcal{N} = \{1, \dots, n\}$ be the set of agents and $\mathcal{A} = \{1, \dots, K\}$ a finite action set. Agent i has policy $\pi^i_{\boldsymbol{\theta}}(\cdot) \in \Delta(\mathcal{A})$ with its own parameter block $\boldsymbol{\theta}^i$; the full parameter is $\boldsymbol{\theta} = (\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^n)$ with disjoint blocks. The joint policy factorizes as

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{a}) = \prod_{i=1}^{n} \pi^i_{\boldsymbol{\theta}}(a^i), \qquad \boldsymbol{a} = (a^1, \dots, a^n) \in \mathcal{A}^n. \tag{2}$$

Rewards and coupling. Agent i receives reward $r_i(\boldsymbol{a})$; the team return is $\mathrm{R}(\boldsymbol{a}) = \sum_i r_i(\boldsymbol{a})$. The coupling graph has symmetric adjacency $C \in \{0,1\}^{n\times n}$ with $C_{ii} = 1$, where $C_{ij} = 1$ iff r_i depends on a^j. We write $\partial i = \{j : C_{ij} = 1\}$. (For simplicity, we omit the dependence on the state in the reward notation here.)

Two supports. Fix a behavior parameter $\boldsymbol{\theta}_{\mathrm{old}}$ and, for a candidate $\boldsymbol{\theta}$, define the per-agent likelihood ratio $\varrho_j(\boldsymbol{a}) = \pi^j_{\boldsymbol{\theta}}(a^j)/\pi^j_{\boldsymbol{\theta}_{\mathrm{old}}}(a^j)$. Let $S^A, S^R \in \{0,1\}^{n\times n}$ be symmetric support matrices with $S_{ii} = 1$. The advantage and importance weight of agent i are given by

$$A_i = \sum_{j:\, S^A_{ij}=1} r_j(\boldsymbol{a}) - b_i, \qquad w_i(\boldsymbol{\theta}) = \prod_{j:\, S^R_{ij}=1} \varrho_j(\boldsymbol{a}), \tag{3}$$

with b_i any control variate independent of $\boldsymbol{a}$. Special cases include $S^A = I$ (independent advantage), $S^A = \mathbf{1}\mathbf{1}^\top$ (team advantage), $S^R = I$ (per-agent ratio), and $S^R = \mathbf{1}\mathbf{1}^\top$ (joint ratio $\prod_j \varrho_j$).

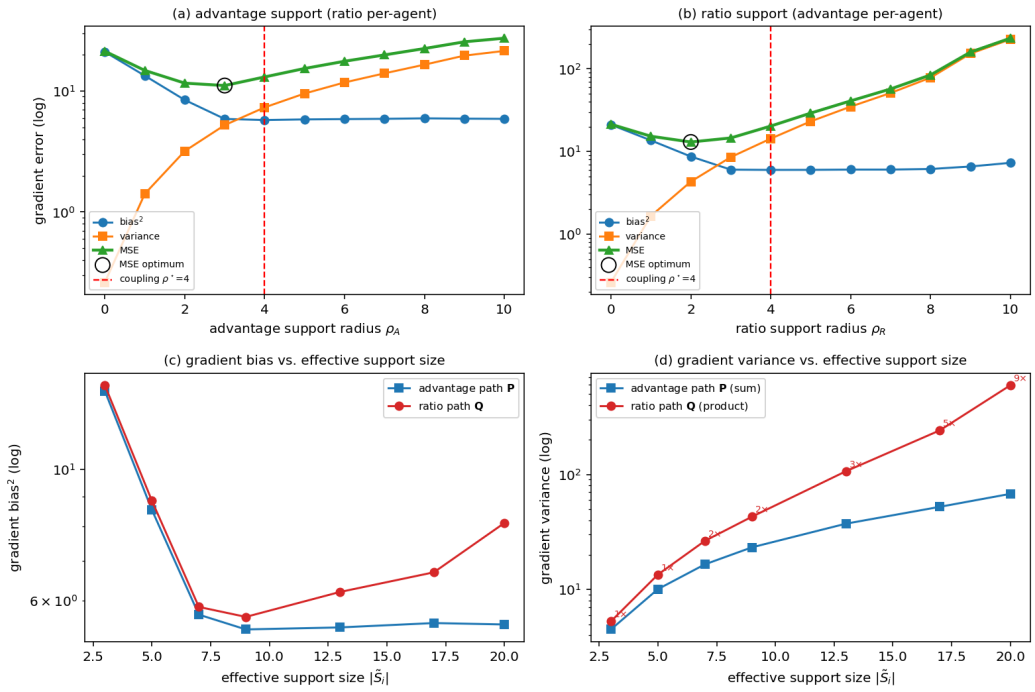


Figure 1: Bias, variance, and MSE of the multi-agent policy-gradient estimator on a single-step dense-pairwise game, as a function of the aggregation support (log axes throughout). (a) Sweep of the advantage-support radius ρ_A with a per-agent ratio; each panel plots the estimator's squared bias, variance, and their sum (MSE), with the MSE minimum marked (○) and the coupling radius $\rho^\star = 4$ shown dashed. (b) Sweep of the ratio-support radius ρ_R with a per-agent advantage. (c,d) At a matched effective support $\widetilde{S}$ = ring of increasing size, the advantage path $\mathbf{P}$ (S^A = ring, $S^R = I$) and the ratio path $\mathbf{Q}$ ($S^A = I$, S^R = ring): (c) squared bias and (d) variance of each path, the annotations in (d) giving the ratio $\mathrm{Var}(\mathbf{Q})/\mathrm{Var}(\mathbf{P})$.

Surrogate and gradient. The (unclipped) surrogate is $L(\boldsymbol{\theta}) = \sum_i \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[w_i(\boldsymbol{\theta})A_i]$, with A_i detached (constant in $\boldsymbol{\theta}$). Since $\nabla_{\boldsymbol{\theta}} w_i = w_i \sum_{j:S^R_{ij}=1} \nabla_{\boldsymbol{\theta}} \log \pi^j(a^j)$ and $\nabla_{\boldsymbol{\theta}^m} \log \pi^j = 0$ for $j \neq m$, we have

$$\boldsymbol{g}_m(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}^m} L = \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\Big[\Big(\sum_{i:\, S^R_{im}=1} w_i(\boldsymbol{\theta})A_i\Big)\nabla_{\boldsymbol{\theta}^m} \log \pi^m(a^m)\Big]. \tag{4}$$

3.2 Assumptions and Their Justification

Assumption 1 (Factorized policy / conditional action independence)**.** *Under both $\boldsymbol{\theta}_{\text{old}}$ and any candidate $\boldsymbol{\theta}$, the joint policy factorizes as in* (2)*; equivalently, given the parameters, actions are sampled independently across agents.*

Justification. This is not an extra modeling restriction but rather the defining structure of the methods under study. In most MARL algorithms (e.g. IPPO, MAPPO, HAPPO) each agent samples $a^i \sim \pi^i(\cdot \mid o^i)$ independently by construction, so (2) holds exactly. The assumption would fail only for policies with an explicitly autoregressive action head (e.g., a decoder that conditions a^i on realized $a^{<i}$). We exclude that case and note that it corresponds to a different (chain-rule) factorization (Wen et al., 2022). Moreover, Assumption 1 concerns conditional independence given parameters (and context); far from implying that the agents' actions are marginally uncorrelated, it is entirely consistent with them being correlated, through shared observations and, during learning, through the coupled reward.

Assumption 2 (Full support / finite second moment)**.** *$\pi^i_{\boldsymbol{\theta}}(a) > 0$ for all i, a and all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_{\text{old}}$, so each ϱ_j is finite with $\mathbb{E}[\varrho_j^2] < \infty$.*

Justification. Softmax policies, the standard parameterization, satisfy full support automatically. Finite second moments are required for any importance-weighted estimator to have finite variance and are standard in PPO and TRPO analysis.

Assumption 3 (Symmetric coupling)**.** *C is symmetric with $C_{ii} = 1$, and r_i depends on a^j iff $C_{ij} = 1$.*

Justification. Symmetry holds whenever coupling arises from shared or pairwise terms (shared resources, pairwise congestion or collision, common team reward), which covers the cooperative tasks of interest. The directional case is a routine extension (Remark 5) obtained by replacing C with the directed influence matrix. The main theorem does not require symmetry, which we adopt only to state the coupling radius cleanly.

Assumption 4 (Bounded rewards)**.** *$|r_i(\boldsymbol{a})| \le r_{\max} < \infty$ for all $i, \boldsymbol{a}$.*

Justification. This is standard and holds for any bounded reward. By truncation, it also applies to sub-Gaussian rewards up to negligible tails. It ensures that the advantage-side aggregation has variance $O(|S^A|)$ rather than exponential.

3.3 Ratio Moments

Lemma 1 (Unbiased weight)**.** *Under Assumptions 1–2, for any $S \subseteq \mathcal{N}$, $\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\big[\prod_{j\in S} \varrho_j\big] = 1$. In particular $\mathbb{E}[w_i] = 1$ for every ratio support.*

This follows because the behavior-policy expectation factorizes over agents and each per-agent ratio has unit mean; a proof is provided in Appendix B.1.

Lemma 2 (Multiplicative variance)**.** *Under Assumptions 1–2, with $\chi^2_j := \chi^2(\pi^j_{\boldsymbol{\theta}} \,\|\, \pi^j_{\text{old}}) = \mathbb{E}[\varrho_j^2] - 1 \ge 0$,*

$$\operatorname{Var}\Big(\prod_{j\in S} \varrho_j\Big) = \prod_{j\in S}(1+\chi^2_j) - 1. \tag{5}$$

Hence the weight variance is nondecreasing in S, and if $\chi^2_j \ge c > 0$ then it is $\ge (1+c)^{|S|} - 1$, i.e., exponential in $|S|$.

The multiplicative form follows from the independence of the per-agent ratios, which causes the second moments to factorize; the full derivation is given in Appendix B.2.

Remark 1. *At $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{old}}$, every $\chi^2_j = 0$ and $w_i \equiv 1$, so the ratio only activates once $\boldsymbol{\theta}$ departs from $\boldsymbol{\theta}_{\text{old}}$, i.e., across PPO's inner epochs. This is precisely why the two supports coincide on-policy (Thm. 1) yet diverge off-policy (Prop. 1).*

3.4 Main results

3.4.1 The expected gradient factorizes through $\widetilde{S} = S^R S^A$

Theorem 1 (Support factorization)**.** *Under Assumptions 1–3, at $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{old}}$ the expected gradient* (4) *is*

$$\boldsymbol{g}_m = \sum_{j:\, C_{jm}=1} \widetilde{S}_{mj}\, \mathbb{E}\big[r_j\, \nabla_{\boldsymbol{\theta}^m} \log \pi^m(a^m)\big], \qquad \widetilde{S} = S^R S^A, \tag{6}$$

where $\widetilde{S}$ is the matrix product of S^R and S^A. Thus $\boldsymbol{g}_m$ depends on (S^A, S^R) only through $\widetilde{S}$: any two support pairs with the same product $S^R S^A$ induce the same expected gradient.

The key step evaluates the surrogate gradient at $\boldsymbol{\theta}_{\text{old}}$, where $w_i{=}1$, and collects terms by the score-function identity; the complete proof is in Appendix B.3.

3.4.2 Canonical form and redundancy of cross-agent ratios

Corollary 1 (Per-agent ratio canonical form)**.** *Under Theorem 1, every estimator with supports (S^A, S^R) has, at the on-policy point, the same expected gradient as the* per-agent-ratio *estimator $(S^R{=}I,\ S^A{=}\widetilde{S})$ whose advantage is the reweighted quantity $A'_m = \sum_j \widetilde{S}_{mj} r_j$ with $\widetilde{S} = S^R S^A$. In particular the joint ratio $S^R = \mathbf{1}\mathbf{1}^\top$ paired with advantage S^A is gradient-equivalent to a per-agent ratio with advantage reweighted by $\mathbf{1}\mathbf{1}^\top S^A$. Hence cross-agent importance ratios add no expected-gradient signal beyond a linear reweighting of the advantage.*

This is immediate from Theorem 1 by reading off the coefficient of each score term (Appendix B.4).

Remark 2 (The team-advantage scalar)**.** *With team advantage $S^A = \mathbf{1}\mathbf{1}^\top$, per-agent ratio gives $\widetilde{S} = \mathbf{1}\mathbf{1}^\top$ while joint ratio gives $\widetilde{S} = \mathbf{1}\mathbf{1}^\top\mathbf{1}\mathbf{1}^\top = n\,\mathbf{1}\mathbf{1}^\top$. The two expected gradients are therefore* identical up to the global scalar n*: the joint ratio merely rescales the effective step by the support size. Any fair comparison must control for this scalar (we do so in Sec. 4 by matching effective step size).*

3.4.3 Variance Domination

We now compare the two canonical realizations of a target product $\widetilde{S} = C$:

$$\text{(P) advantage path: } S^A = C,\ S^R = I; \qquad \text{(Q) ratio path: } S^A = I,\ S^R = C. \tag{7}$$

Both realize the same expected gradient $\sum_{j\in\partial m}\mathbb{E}[r_j\nabla\log\pi^m]$ by Theorem 1.

Proposition 1 (Off-policy variance domination)**.** *Under Assumptions 1–4: (i) at $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{old}}$, P and Q have identical mean and identical variance; (ii) for $\boldsymbol{\theta} \neq \boldsymbol{\theta}_{\text{old}}$, the P-weight $\sum_{j\in\partial m} r_j$ is bounded by $|\partial m|\, r_{\max}$ with variance independent of $\boldsymbol{\theta}$, whereas the Q-weight $\sum_{i\in\partial m}\big(\prod_{j\in\partial i}\varrho_j\big)r_i$ has second moment $\mathbb{E}[(\prod_{j\in\partial i}\varrho_j)^2] = \prod_{j\in\partial i}(1+\chi_j^2)$, so once any $\chi_j^2 > 0$ its variance is bounded below by a positive multiple of $\prod_{j\in\partial m}(1+\chi_j^2) - 1$. Hence $\mathrm{Var}(\boldsymbol{g}_m^Q)/\mathrm{Var}(\boldsymbol{g}_m^P) \to \infty$ as the policy step or $|\partial m|$ grows.*

The bound follows by comparing the bounded P-weight to the Q-weight's exploding second moment from Lemma 2; see Appendix B.5.

Remark 3 (Clipping widens the gap)**.** *Reinstating the PPO clip applies a per-weight, 1-Lipschitz truncation. For P, the clipped object is a single low-variance ratio ϱ_m, so clipping rarely activates; for Q, it is the heavy-tailed product (Lemma 2), which triggers clipping more often, injecting truncation bias and leaving the retained variance elevated. Clipping thus cannot reverse Proposition 1 and typically amplifies it.*

Corollary 2 (Design rule)**.** *Among all (S^A, S^R) that realize a target unbiased product $\widetilde{S}$, the choice $S^R = I$ (per-agent ratio) with $S^A = \widetilde{S}$ minimizes the gradient-estimator variance. In particular, the joint (compound) ratio is weakly dominated on-policy and strictly dominated off-policy: it should be replaced by a per-agent ratio with the corresponding advantage reweighting.*

3.4.4 The Advantage Support: A Bias–Variance Tradeoff

At a matched effective support, the ratio support does not provide any bias reduction advantage over the advantage support: any bias it can remove is already removed by the advantage support at strictly lower variance (Corollary 1, Prop. 1). Consequently, its variance-optimal setting is the per-agent boundary $S^R = I$. The advantage support is qualitatively different: with the ratio fixed at per-agent, it trades bias against variance solely through S^A, and this tradeoff has an interior optimum. This asymmetry between the two supports is fundamental, and we can characterize the advantage side as sharply as the ratio side.

Proposition 2 (Advantage-support bias–variance tradeoff)**.** *Fix the per-agent ratio $S^R = I$ and consider the on-policy estimator $\boldsymbol{g}_m$ with advantage support S^A. Write the true (team-return) gradient block as $\boldsymbol{g}_m^\star = \sum_{j\in\partial m}\mathbb{E}[r_j\nabla_{\boldsymbol{\theta}^m}\log\pi^m]$. Then, under Assumptions 1–4:*

- *(i) (Bias) $\mathbb{E}[\boldsymbol{g}_m] = \sum_{j\in\partial m} S^A_{mj}\,\mathbb{E}[r_j\nabla_{\boldsymbol{\theta}^m}\log\pi^m]$, so $\boldsymbol{g}_m$ is unbiased iff $S^A_{mj} = 1$ for every coupled agent $j \in \partial m$; a support that misses a coupled agent omits its term and is biased.*
- *(ii) (Variance) Including an uncoupled agent $j \notin \partial m$ leaves the mean unchanged (its term has zero expectation by the score-function identity) but adds a nonnegative variance contribution, strictly positive whenever r_j is conditionally non-degenerate given a^m.*

Consequently, the mean-squared-error-optimal advantage support is exactly the coupling neighborhood ∂m: smaller supports are biased, larger supports inflate variance.

The bias and variance terms are computed separately and their sum minimized; the derivation is in Appendix B.6.

Remark 4 (The asymmetry, precisely)**.** *Propositions 1 and 2 together establish the central asymmetry of the paper as a theoretical result, not merely an empirical observation. With the ratio fixed to per-agent, the advantage support has a bias term that larger supports remove, yielding an interior MSE optimum at the coupling neighborhood. In contrast, the ratio support offers no bias-reduction benefit: by the canonical form (Corollary 1), any bias reduction it could achieve is equally attainable through the advantage, and by Proposition 1, at strictly lower variance. Hence, on the variance-optimal frontier, the ratio is never used to reduce bias, and its optimum is the boundary $S^R = I$. Although the two knobs appear symmetric in the surrogate* (3)*, they play categorically different roles.*

Remark 5 (Directional-coupling extension)**.** *If coupling is directional, replace C by the directed influence matrix D with $D_{ij} = 1$ iff r_i depends on a^j. Theorem 1 holds with the survival condition $D_{jm} = 1$ and $\widetilde{S} = S^R S^A$; the variance-optimal design uses $S^R = I$, $S^A = D^\top$.*

Remark 6 (Finite-horizon extension)**.** *In a finite-horizon Markov game, the argument applies per time step to the GAE advantage and the per-step ratio. The multiplicative variance of Lemma 2 then compounds across agents and time, so the joint ratio's variance grows in $n \cdot H$; the on-policy factorization of Theorem 1 is unchanged. The episodic experiment (Sec. 4) confirms the redundancy in this setting.*

4 Experiments

4.1 Experiment Setup

The theory yields three testable predictions. **(R) Redundancy:** estimators with the same product $\widetilde{S} = S^R S^A$ have identical expected gradients. **(D) Domination:** realizing a fixed $\widetilde{S}$ through the ratio product incurs strictly higher variance off-policy than through the advantage sum, with the gap growing multiplicatively in the support size, as previewed in Fig. 1. **(A) Asymmetry:** at the gradient-estimator level, the advantage support presents a bias–variance tradeoff with an interior MSE optimum at the coupling neighborhood (Prop. 2), whereas the ratio support has no bias effect and a boundary optimum at $\rho_R = 0$ (Prop. 1). During training, the advantage radius influences the learned policy only insofar as bias alters the equilibrium, decisively when the externality is not self-internalized and only mildly otherwise. The ratio radius, being bias-free, never changes the learned policy and affects only stability. We test these predictions through a multi-family training study that traces how each support affects the learned policy across four coupling structures and a scaling study that identifies the regime in which the ratio's variance penalty becomes manifest.

All environments are cooperative games with a prescribed coupling graph C, so the ground-truth coupling neighborhood is known exactly. Each agent i observes a phase or context and selects $a^i \in \{1, \ldots, K\}$. Per-agent rewards r_i are defined in Appendix C.1, and the team return is $\mathrm{R} = \sum_i r_i$. Ring neighborhoods of radius $\rho^\star$ define ∂i. We use four game families that span the regimes distinguished by the theory: a dense pairwise game (symmetric payoff over a ring neighborhood, with internalized externality); a directed dilemma, where a high-value action imposes a nuisance on downstream successors that the actor does not bear (a non-internalized externality); local congestion, where neighbors choosing the same resource split its value (crowding partly self-borne); and a block community game with all-to-all coupling inside disjoint blocks (a non-geometric graph). A more detailed description is provided in Appendix C.1. In addition, we provide a real-world large-scale traffic signal control validation in Appendix C.3.2.

4.2 Experiment Results

We first examine PPO training across the four coupling families, which differ in how agents interact. Each is a finite-horizon Markov game with phase-conditioned tabular actors trained by PPO (clip 0.2). We sweep the advantage radius ρ_A (with per-agent ratio fixed) and, separately, the ratio radius ρ_R (with team advantage fixed). Curves are converged over the final 40 iterations and averaged over 5 seeds.

The advantage radius selects the outcome only for non-internalized externalities (Fig. 2). Figure 2 plots, for each of the four coupling families, the final team return as a function of the advantage radius ρ_A (per-agent ratio fixed), with the true coupling radius $\rho^\star$ marked. The four families separate sharply. In the directed dilemma, the advantage radius is decisive: independent advantage ($\rho_A = 0$)

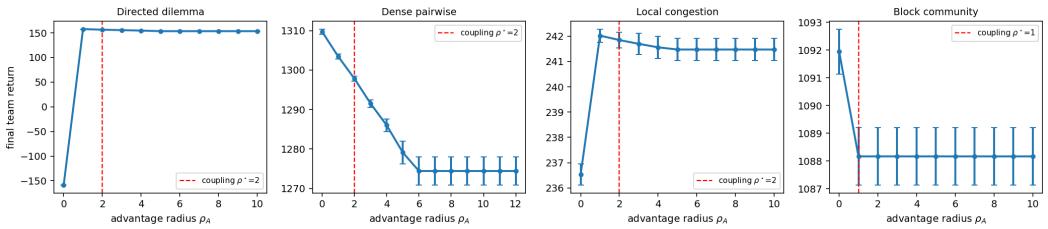


Figure 2: Advantage-support sweep across the four coupling families: final team return versus the advantage radius ρ_A, with the ratio held per-agent. The dashed vertical line marks the coupling radius $\rho^\star$. Bands indicate ± 1 standard deviation across seeds. In the directed dilemma the standard deviation is below 0.4 and smaller than the marker.

converges to the tragedy equilibrium (team return -159), while any $\rho_A \geq 1$ internalizes enough of the successor externality to reach the social optimum ($+158$). In the other three families, the effect is mild and the outcome-optimal radius is small ($\rho_A \in \{0, 1\}$): when an agent already bears (part of) the coupling it induces, whether through symmetric pairwise payoffs, shared congestion, or within-block rewards, the independent advantage is nearly unbiased, so enlarging ρ_A buys little signal and adds variance, and the return is flat or slightly decreasing in ρ_A. This is the training-level counterpart of the estimator tradeoff in Fig. 1a: the advantage radius should match the coupling neighborhood, but the outcome moves only when the un-internalized part of the externality is large enough to change the equilibrium. Full training curves for all four families are provided in Appendix C.2.

Moreover, redundancy implies that the ratio radius carries no benefit. Proposition 1 states that it carries a variance cost that grows multiplicatively with support size and with the per-update policy shift χ^2. Whether that cost is visible in training therefore depends on how far off-policy each update travels, which a trust-region method controls through its step size. We make this explicit by scaling the number of agents under two configurations that differ only in aggressiveness: a standard PPO setting ($\mathrm{lr} = 3 \times 10^{-3}$, batch 64) and an off-policy stress setting ($\mathrm{lr} = 0.15$, batch 16). Both use the team advantage and match the effective step across ratio supports (Remark 2), so any difference is due to variance, not step size.

Masked under conservative updates, revealed under stress (Fig. 3). Figure 3 varies the number of agents n (log axis) for each family and reports two quantities: the joint/per-agent final return ratio (solid, left axis), indicating how much return the joint ratio sacrifices relative to a per-agent ratio, and the joint ratio's clip fraction (dashed, right axis), under both the standard (blue) and off-policy stress (red) configurations. Under the standard configuration, the joint and per-agent ratios are statistically indistinguishable up to $n = 160$ (return ratio 1.00, joint clip fraction ≈ 0), with a 4–8% gap emerging only at $n = 320$: the trust region keeps χ^2 so small ($\chi^2_j \approx 6 \times 10^{-4}$ per update) that the variance factor $(1 + \chi^2)^n - 1$ is negligible even for n in the hundreds. Under the stress configuration, the same experiment reproduces the folklore instability across all four families: the joint ratio's clip fraction rises monotonically with n (reaching 0.2–0.45 at the largest n tested per family), while the per-agent ratio's clip fraction remains at 0 throughout, the signature of Lemma 2. A product of n ratios leaves the trust region almost surely once n is large, so it is clipped increasingly often and its gradient is throttled, whereas a single ratio never is. Where the coupling is sharp enough that this throttling starves the update of signal, as in the directed dilemma and dense pairwise families, the return degrades with it. Where the reward is more forgiving (local congestion, block community), the same clipping leaves the final return closer to parity, making the clip fraction the more universal diagnostic.

Remark 7 (Step size, not agent count, is the trigger)**.** *The variance factor $(1 + \chi^2)^n - 1$ has two levers: the support size n and the per-update shift χ^2. Fig. 3 shows they act together, since the gap grows with n but only once χ^2 is non-negligible. A conservative learning rate makes χ^2 vanishingly small, so even n in the hundreds is harmless. An aggressive rate makes it bite. This reconciles the theorem with the common observation that MAPPO's per-agent ratio and various joint-ratio schemes often perform comparably: they do, precisely in the near-on-policy regime where the penalty is dormant. The per-agent ratio weakly dominates always and strictly dominates whenever updates are pushed off-policy, at no cost, since the benefit of a larger ratio support is exactly zero.*

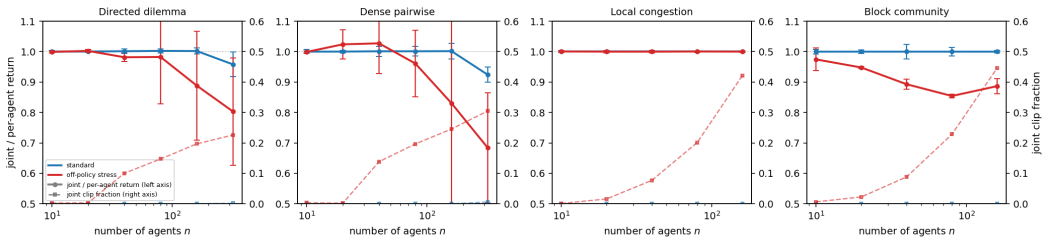


Figure 3: Agent-count scaling across the four coupling families. For each family and agent count n (log axis), solid curves report the joint/per-agent final return ratio (left axis), and dashed curves show the joint ratio's clip fraction (right axis). Results are shown under a standard PPO configuration (blue; $\mathrm{lr} = 3 \times 10^{-3}$, batch 64) and an off-policy stress configuration (red; $\mathrm{lr} = 0.15$, batch 16); both use the team advantage and match the effective step across ratio supports. Error bars indicate ± 1 standard deviation over seeds (on the standard curves they are smaller than the markers).

Remark 8 (Why we recommend against the joint ratio, not merely note its risk)**.** *It might seem that, since the joint ratio is harmless in the near-on-policy regime, one could simply use it with a small step. We caution against this for a practical reason that our controlled environments understate. Here we know the coupling, the reward scale, and the effective off-policyness, so we can see that the penalty is dormant. In a real large-scale system, one does not know these in advance: the effective χ^2 varies across states, agents, and training phases, and a minibatch that happens to be more off-policy, after a reward spike, an exploration burst, or a learning-rate warmup, can push the n-fold product out of the trust region and stall learning, with no diagnostic that distinguishes this from ordinary noise. Because the joint ratio offers zero upside over a per-agent ratio (Cor. 1) and an unbounded, hard-to-anticipate downside that grows with n, the asymmetry of the bet is decisive: keep the ratio per-agent. This is especially pertinent at the scales where cooperative MARL is deployed, such as traffic grids, sensor and robot fleets, and power networks, where n is in the hundreds or thousands and the $(1+\chi^2)^n$ factor is one adverse batch away from biting. Our 196-intersection traffic network bears this out directly: with $n = 196$, the joint ratio collapses even at standard settings, while per-agent and neighborhood ratios learn normally (Appendix C.3.2).*

5 Conclusion

In this paper, we decomposed the credit-assignment structure of multi-agent policy optimization into two independent design axes: an advantage support and a ratio support. Our analysis reveals that the expected policy gradient depends on these supports only through their matrix product $\widetilde{S} = S^R S^A$, establishing a gauge freedom that renders the two supports mutually redundant in terms of the gradient signal. This redundancy is broken by variance: aggregating through the advantage corresponds to a sum of rewards, yielding additive variance with an interior optimum at the coupling neighborhood, whereas aggregating through the ratio corresponds to a product of importance weights, incurring multiplicative, exponentially growing variance in the support size. These results lead to a clear and actionable design principle: aggregate neighbors in the advantage, sized according to the true reward coupling, and keep the ratio strictly per-agent. This rule applies universally to any factorized-policy PPO method and is supported by rigorous theoretical analysis. We validate our conclusions across four carefully designed cooperative MAMDP games and a real-world large-scale traffic signal control task, demonstrating both the correctness of the canonical form and the practical relevance of the variance asymmetry. More discussion is provided in Appendix E.

AI Use Statement

In this work, we used generative AI tools for paper polishing, including grammar and wording checks, verification and support of mathematical derivations, and coding. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

Ethics Statement

This work adheres to the principles outlined in the ICLR Code of Ethics.

Reproducibility Statement

The code for this paper is available at `https://anonymous.4open.science/r/MAPO-803E`.

References

Lucas N. Alegre. SUMO-RL. `https://github.com/LucasAlegre/sumo-rl`, 2019.

Michael Behrisch, Laura Bieker, Jakob Erdmann, and Daniel Krajzewicz. Sumo–simulation of urban mobility: an overview. In *Proceedings of SIMUL 2011, the third international conference on advances in system simulation*. ThinkMind, 2011.

Tianshu Chu, Jie Wang, Lara Codecà, and Zhaojian Li. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE transactions on intelligent transportation systems*, 21(3): 1086–1095, 2019.

Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Weiqiang Jin, Hongyang Du, Biao Zhao, Xingwu Tian, Bohang Shi, and Guang Yang. A comprehensive survey on multi-agent cooperative decision-making: Scenarios, approaches, challenges and perspectives. *arXiv preprint arXiv:2503.13415*, 2025.

Seongmin Kim, Giseung Park, Woojun Kim, Jiwon Jeon, Seungyul Han, and Youngchul Sung. Generalized per-agent advantage estimation for multi-agent policy optimization. *arXiv preprint arXiv:2603.02654*, 2026.

JG Kuba, R Chen, M Wen, Y Wen, F Sun, J Wang, and Y Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *ICLR 2022-10th International Conference on Learning Representations*, pp. 1046. The International Conference on Learning Representations (ICLR), 2022.

Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 4501–4510, 2020.

Yueheng Li, Guangming Xie, and Zongqing Lu. Difference advantage estimation for multi-agent policy gradients. In *International Conference on Machine Learning*, pp. 13066–13085. PMLR, 2022.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. Scalable multi-agent reinforcement learning for networked systems with average reward. *Advances in Neural Information Processing Systems*, 33:2074–2086, 2020a.

Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning of localized policies for multi-agent networked systems. In *Learning for Dynamics and Control*, pp. 256–266. PMLR, 2020b.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015a.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Muning Wen, Jakub Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. Multi-agent reinforcement learning is a sequence modeling problem. *Advances in Neural Information Processing Systems*, 35:16509–16521, 2022.

S Whiteson, M Samvelyan, T Rashid, CS De Witt, G Farquhar, N Nardelli, TGJ Rudner, CM Hung, PHS Torr, and J Foerster. The starcraft multi-agent challenge. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, pp. 2186–2188, 2019.

David H Wolpert and Kagan Tumer. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 4(02n03):265–279, 2001.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Zijian Zhao and Sen Li. Triple-BERT: Do we really need MARL for order dispatch on ride-sharing platforms? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=symgW6FhA6.

Zijian Zhao, Jing Gao, and Sen Li. Bridging marl to sarl: An order-independent multi-agent transformer via latent consensus. *arXiv preprint arXiv:2604.13472*, 2026.

Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32): 1–67, 2024.

APPENDIX CONTENTS

A MAMDP Formulation **13**

B Proofs **13**

B.1 Proof of Lemma 1 (unbiased weight) 14

B.2 Proof of Lemma 2 (multiplicative variance) 14

B.3 Proof of Theorem 1 (support factorization) 14

B.4 Proof of Corollary 1 (per-agent-ratio canonical form) 15

B.5 Proof of Proposition 1 (off-policy variance domination) 15

B.6 Proof of Proposition 2 (advantage bias–variance tradeoff) 15

C Experimental Details and Additional Results **16**

C.1 Experimental Setup 16

C.2 Extended Training-Curve Analysis 17

C.3 Real-World Traffic-Signal Control 18

D Related Work **21**

E Discussion **21**

A MAMDP FORMULATION

We formalize the cooperative setting as a *multi-agent Markov decision process* (MAMDP) (Littman, 1994), equivalently a cooperative Markov game. This appendix provides the full formulation summarized in the main text.

Definition. A MAMDP is a tuple $\big(\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, \{r_i\}_{i\in\mathcal{N}}, \gamma, \mu_0\big)$, where $\mathcal{N} = \{1, \dots, n\}$ is the set of agents; $\mathcal{S}$ is the state space; $\mathcal{A}_i$ is the finite action set of agent i, with joint action space $\mathcal{A} = \prod_i \mathcal{A}_i$ and joint action $\boldsymbol{a} = (a^1, \dots, a^n)$; $P(s' \mid s, \boldsymbol{a})$ is the transition kernel; $r_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is agent i's reward, with team reward $\mathrm{R}(s, \boldsymbol{a}) = \sum_i r_i(s, \boldsymbol{a})$; $\gamma \in [0, 1)$ is the discount factor; and μ_0 is the initial-state distribution. The setting is fully cooperative: all agents share the single team objective defined below.

Factorized policy. Each agent acts through its own policy $\pi^i_{\boldsymbol{\theta}^i}(a^i \mid s) \in \Delta(\mathcal{A}_i)$ with a disjoint parameter block $\boldsymbol{\theta}^i$, so $\boldsymbol{\theta} = (\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^n)$. Actions are conditionally independent across agents given the state, so the joint policy factorizes as

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid s) = \prod_{i=1}^{n} \pi^i_{\boldsymbol{\theta}^i}(a^i \mid s). \tag{8}$$

This factorization, equivalently independent action heads sharing a state encoder, is the single structural assumption used in our analysis. It also covers the centralized multi-action policy that treats the whole team as one agent whose action is $\boldsymbol{a}$.

Objective. The agents jointly maximize the expected discounted team return

$$J(\boldsymbol{\theta}) = \mathbb{E}_{s_0\sim\mu_0,\ \boldsymbol{a}_t\sim\pi_{\boldsymbol{\theta}}(\cdot|s_t),\ s_{t+1}\sim P(\cdot|s_t,\boldsymbol{a}_t)} \left[\sum_{t\geq 0} \gamma^t\, \mathrm{R}(s_t, \boldsymbol{a}_t) \right]. \tag{9}$$

Values and advantages. The state value is $V^\pi(s) = \mathbb{E}_\pi\big[\sum_{t\geq 0} \gamma^t \mathrm{R}(s_t, \boldsymbol{a}_t) \mid s_0 = s\big]$, and Q^π, the generalized advantage estimate (GAE), and the per-agent advantage are defined as usual from the per-agent rewards. The *advantage support* $S^A \in \{0, 1\}^{n\times n}$ selects which agents' rewards form agent i's advantage: with $\hat{A}_j$ the single-agent advantage built from r_j, agent i's aggregated advantage is $A_i = \sum_j S^A_{ij} \hat{A}_j$.

Coupling graph. The *coupling graph* $C \in \{0, 1\}^{n\times n}$ has $C_{ij} = 1$ iff r_i depends on a^j (with $C_{ii} = 1$). Its smallest neighborhood radius $\rho^\star$ is the *coupling radius*; $\partial i = \{j : C_{ij} = 1\}$ is the coupling neighborhood of agent i.

Clipped surrogate with two supports. Fixing a behavior policy $\boldsymbol{\theta}_{\mathrm{old}}$, the per-agent likelihood ratio is $\varrho^j = \pi^j_{\boldsymbol{\theta}}(a^j \mid s)/\pi^j_{\boldsymbol{\theta}_{\mathrm{old}}}(a^j \mid s)$. The *ratio support* $S^R \in \{0, 1\}^{n\times n}$ selects which agents' ratios enter agent i's importance weight, $w_i = \prod_{j:S^R_{ij}=1} \varrho^j$, and the (clipped) multi-agent PPO surrogate is $\sum_i \mathbb{E}\big[\min(w_i A_i,\ \mathrm{clip}(w_i, 1 \pm \epsilon) A_i)\big]$. The pair (S^A, S^R) is the object of study: $S^A = S^R = I$ corresponds to IPPO, $S^A = \mathbf{1}\mathbf{1}^\top, S^R = I$ to MAPPO, and $S^A = S^R = \mathbf{1}\mathbf{1}^\top$ to the fully centralized multi-action reduction.

B PROOFS

Throughout, all expectations are over $\boldsymbol{a} \sim \pi_{\boldsymbol{\theta}_{\mathrm{old}}}$ unless noted, and we write $\varrho_j = \pi^j_{\boldsymbol{\theta}}(a^j)/\pi^j_{\boldsymbol{\theta}_{\mathrm{old}}}(a^j)$ for the per-agent likelihood ratio and $s^m := \nabla_{\boldsymbol{\theta}^m} \log \pi^m(a^m)$ for the per-agent score.

B.1 Proof of Lemma 1 (unbiased weight)

Proof. Fix $S \subseteq \mathcal{N}$. By Assumption 1, the actions $\{a^j\}_{j\in S}$ are independent under $\pi_{\boldsymbol{\theta}_{\text{old}}}$, so the expectation of the product factorizes:

$$\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\Big[\prod_{j\in S}\varrho_j\Big] = \prod_{j\in S}\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\big[\varrho_j\big] = \prod_{j\in S}\sum_{a\in\mathcal{A}}\pi^j_{\boldsymbol{\theta}_{\text{old}}}(a)\,\frac{\pi^j_{\boldsymbol{\theta}}(a)}{\pi^j_{\boldsymbol{\theta}_{\text{old}}}(a)} \tag{10a}$$

$$= \prod_{j\in S}\sum_{a\in\mathcal{A}}\pi^j_{\boldsymbol{\theta}}(a) = \prod_{j\in S}1 = 1, \tag{10b}$$

where (10a) uses independence (Assumption 1) and the definition of ϱ_j, and the cancellation in (10b) is valid because $\pi^j_{\boldsymbol{\theta}_{\text{old}}}(a) > 0$ (Assumption 2); the last equality follows from the normalization of $\pi^j_{\boldsymbol{\theta}}$. Taking $S = \{j : S^R_{ij} = 1\}$ gives $\mathbb{E}[w_i] = 1$ for every ratio support. □

B.2 Proof of Lemma 2 (multiplicative variance)

Proof. By Lemma 1, $\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[\prod_{j\in S}\varrho_j] = 1$, so $\mathrm{Var}(\prod_{j\in S}\varrho_j) = \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[(\prod_{j\in S}\varrho_j)^2] - 1$. The squared product factorizes over agents by independence (Assumption 1):

$$\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\Big[\Big(\prod_{j\in S}\varrho_j\Big)^2\Big] = \prod_{j\in S}\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\big[\varrho_j^2\big], \tag{10}$$

$$\text{and}\qquad \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\big[\varrho_j^2\big] = \sum_{a\in\mathcal{A}}\pi^j_{\boldsymbol{\theta}_{\text{old}}}(a)\frac{\pi^j_{\boldsymbol{\theta}}(a)^2}{\pi^j_{\boldsymbol{\theta}_{\text{old}}}(a)^2} = \sum_a \frac{\pi^j_{\boldsymbol{\theta}}(a)^2}{\pi^j_{\boldsymbol{\theta}_{\text{old}}}(a)}. \tag{11}$$

The rightmost sum is $1+\chi^2\big(\pi^j_{\boldsymbol{\theta}} \,\|\, \pi^j_{\boldsymbol{\theta}_{\text{old}}}\big) = 1+\chi^2_j$ by the definition of the χ^2-divergence. Substituting this into (11) and subtracting 1 gives (5): $\mathrm{Var}(\prod_{j\in S}\varrho_j) = \prod_{j\in S}(1+\chi^2_j) - 1$. Since each factor $1+\chi^2_j \geq 1$, the product is nondecreasing in S; if $\chi^2_j \geq c > 0$ for all j, then $\prod_{j\in S}(1+\chi^2_j) - 1 \geq (1+c)^{|S|} - 1$, which is exponential in $|S|$. □

B.3 Proof of Theorem 1 (support factorization)

Proof. Evaluate the gradient (4) at $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{old}}$, where every $\varrho_j = 1$ and hence $w_i = 1$. Writing $A_i = \sum_j S^A_{ij} r_j - b_i$,

$$\boldsymbol{g}_m = \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\Big[\Big(\sum_{i:S^R_{im}=1} A_i\Big)\, s^m\Big] = \sum_{i:S^R_{im}=1}\sum_j S^A_{ij}\,\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\big[r_j\, s^m\big] \;-\; \sum_{i:S^R_{im}=1} b_i\,\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\big[s^m\big]. \tag{12}$$

The baseline term vanishes: b_i is constant in $\boldsymbol{a}$ and $\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[s^m] = \sum_{a^m}\pi^m_{\boldsymbol{\theta}_{\text{old}}}(a^m)\nabla_{\boldsymbol{\theta}^m}\log\pi^m(a^m) = \nabla_{\boldsymbol{\theta}^m}\sum_{a^m}\pi^m(a^m) = \nabla_{\boldsymbol{\theta}^m}1 = 0$. Next, by Assumption 3 (symmetric coupling) and conditional independence, $\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[r_j\, s^m] = 0$ whenever $C_{jm} = 0$: if r_j does not depend on a^m, conditioning on a^{-m} gives

$$\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\big[r_j\, s^m\big] = \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\Big[r_j\,\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\big[s^m \mid a^{-m}\big]\Big] = \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\big[r_j\big]\cdot\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\big[s^m\big] = 0, \tag{13}$$

using $\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[s^m \mid a^{-m}] = \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[s^m] = 0$ (action independence). Substituting (13) into (12) and exchanging the order of the i and j sums,

$$\boldsymbol{g}_m = \sum_{j:\,C_{jm}=1}\Big(\sum_i S^R_{im}S^A_{ij}\Big)\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\big[r_j\, s^m\big] = \sum_{j:\,C_{jm}=1}\big(S^R S^A\big)_{mj}\,\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\big[r_j\, s^m\big], \tag{14}$$

where the last step uses $S^R_{im} = S^R_{mi}$ (symmetry), so $\sum_i S^R_{mi}S^A_{ij} = (S^R S^A)_{mj} = \widetilde{S}_{mj}$. This is (6). Since $\widetilde{S} = S^R S^A$ is the only way (S^A, S^R) enter, any two pairs with the same product induce the same $\boldsymbol{g}_m$. □

B.4 Proof of Corollary 1 (per-agent-ratio canonical form)

Proof. The estimator $(S^R = I, S^A = \widetilde{S})$ has product $\widetilde{S}' = I \cdot \widetilde{S} = \widetilde{S}$, identical to that of (S^A, S^R). By Theorem 1, the expected gradient depends on the supports only through this product, so the two estimators share the same $\boldsymbol{g}_m$; concretely, reading from (6), both equal $\sum_{j:C_{jm}=1} \widetilde{S}_{mj} \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[r_j s^m]$, which is exactly the gradient of a per-agent-ratio estimator with advantage $A'_m = \sum_j \widetilde{S}_{mj} r_j$ (a nonnegative integer-weighted reward sum, hence admissible). For the joint ratio $S^R = \mathbf{1}\mathbf{1}^\top$, the reweighting is $\widetilde{S} = \mathbf{1}\mathbf{1}^\top S^A$. Thus, cross-agent ratios contribute nothing to $\boldsymbol{g}_m$ beyond the linear advantage reweighting $r \mapsto \widetilde{S} r$. □

B.5 Proof of Proposition 1 (off-policy variance domination)

Proof. Both paths in (7) realize the same product $\widetilde{S} = C$, so by Theorem 1 they share the mean $\boldsymbol{g}_m = \sum_{j\in\partial m} \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[r_j s^m]$. Write the per-step weights whose product with s^m is averaged: for the advantage path (P), $W_m^{\text{P}} = \sum_{j\in\partial m} r_j$; for the ratio path (Q), $W_m^{\text{Q}} = \sum_{i\in\partial m} \big(\prod_{j\in\partial i} \varrho_j\big) r_i$.

(i) On-policy. At $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{old}}$, every $\varrho_j = 1$, so $W_m^{\text{Q}} = \sum_{i\in\partial m} r_i = W_m^{\text{P}}$ pathwise; the two estimators are identical random variables and share mean and variance.

(ii) Off-policy. W_m^{P} has no $\boldsymbol{\theta}$ dependence and is bounded: $|W_m^{\text{P}}| \le |\partial m|\, r_{\max}$ (Assumption 4), so $\mathrm{Var}(W_m^{\text{P}})$ is a fixed constant. For Q, take a single summand $i \in \partial m$ and condition on s^m; by Assumption 4, there is $\delta > 0$ with $\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[r_i^2 \mid a^m] \ge \delta$ on a set of positive probability, and by independence of the ratios from r_i and Lemma 2,

$$\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\Big[\Big(\prod_{j\in\partial i} \varrho_j\Big)^2 r_i^2\Big] \;\ge\; \delta\, \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\Big[\Big(\prod_{j\in\partial i} \varrho_j\Big)^2\Big] \;=\; \delta \prod_{j\in\partial i} (1+\chi_j^2). \tag{15}$$

Hence the second moment of the Q-weight, and therefore $\mathrm{Var}(\boldsymbol{g}_m^{\text{Q}})$, is bounded below by a positive multiple of $\prod_{j\in\partial i}(1+\chi_j^2)$, which grows without bound as the policy step increases (each $\chi_j^2 \to \infty$) or as $|\partial m|$ grows (more factors). Since $\mathrm{Var}(\boldsymbol{g}_m^{\text{P}})$ stays bounded, $\mathrm{Var}(\boldsymbol{g}_m^{\text{Q}})/\mathrm{Var}(\boldsymbol{g}_m^{\text{P}}) \to \infty$. □

B.6 Proof of Proposition 2 (advantage bias–variance tradeoff)

Proof. Take $S^R = I$, so $\widetilde{S} = S^A$ and by Theorem 1 $\boldsymbol{g}_m = \sum_{j:C_{jm}=1} S^A_{mj}\, \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[r_j s^m]$. The unbiased (full-coupling) gradient is $\boldsymbol{g}_m^\star = \sum_{j\in\partial m} \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[r_j s^m]$, i.e., the case $S^A_{mj} = 1$ for all $j \in \partial m$.

Bias. Subtracting,

$$\boldsymbol{g}_m - \boldsymbol{g}_m^\star = \sum_{j\in\partial m} (S^A_{mj} - 1)\, \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[r_j s^m] = -\sum_{j\in\partial m:\; S^A_{mj}=0} \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[r_j s^m], \tag{16}$$

so the bias is exactly the sum of the score–reward couplings of the coupled agents omitted by S^A. It is zero iff $S^A \supseteq \partial m$ and strictly nonzero when a genuinely coupled agent with $\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[r_j s^m] \neq 0$ is dropped. Adding agents outside ∂m changes nothing, since $\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[r_j s^m] = 0$ there by (13).

Variance. The aggregated advantage $A_m = \sum_j S^A_{mj} r_j$ is additive. Including an extra agent j adds the term $r_j s^m$ to the estimator. This term has mean $\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[r_j s^m]$ and contributes

$$\Delta\mathrm{Var} = \mathrm{Var}\big(r_j s^m\big) + 2\,\mathrm{Cov}\Big(r_j s^m,\; \textstyle\sum_{k\in S^A_m \setminus j} r_k s^m\Big), \tag{17}$$

whose dominant, always-nonnegative part is $\mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}[r_j^2 (s^m)^2] \ge 0$ (strictly positive when r_j is conditionally nondegenerate). Thus each added agent raises the variance, while only agents in ∂m reduce the bias (16). The MSE $\|\boldsymbol{g}_m - \boldsymbol{g}_m^\star\|^2 + \mathrm{Var}(\boldsymbol{g}_m)$ is therefore minimized by including exactly the coupled agents, $S^A = \{\, j : C_{jm} = 1 \,\} = \partial m$: smaller supports are biased, larger supports inflate variance. □

C Experimental Details and Additional Results

C.1 Experimental Setup

C.1.1 Choice of Environments

Our study requires environments with two properties that standard cooperative benchmarks lack: a per-agent reward (so the advantage support is meaningful) and a known coupling graph (so the coupling neighborhood is defined). Popular suites such as StarCraft (Whiteson et al., 2019) and Google Research Football (Kurach et al., 2020) expose a single shared team reward and no explicit coupling structure; verifying the advantage-support axis there would require synthesizing per-agent signals via difference rewards, introducing approximation and forfeiting the ground-truth coupling graph. We therefore study synthetic cooperative games in which the coupling graph is prescribed exactly, allowing us to vary the coupling family, its range, and the agent count while keeping the ground truth fixed. Applying the same analysis to real systems that natively provide per-agent rewards and a physical coupling graph, of which traffic-signal control is a canonical example, is a natural direction we leave to future work. Our goal here is a clean verification of the mechanism, not benchmark leaderboard performance, so the environments are simple and fully specified.

The families, shown as Fig. 4, are chosen to separate the two axes the theory distinguishes, and to stress each in turn. They differ first in whether an agent internalizes its own externality: in the directed dilemma, an agent's action harms only its successors, not itself, so an independent advantage is badly biased and the advantage support is decisive for the learned policy; in dense pairwise and local congestion, the coupling term enters the actor's own reward (fully or partly), so an independent advantage is already near-unbiased and the advantage support mainly trades variance. This spread lets us show that the advantage radius matters for the outcome exactly when the externality is not self-internalized (Fig. 2). The families differ second in graph structure: the ring families (dilemma, pairwise, congestion) provide a geometric coupling with a well-defined radius, while block community gives a non-geometric, community-structured graph, verifying that matching the advantage support to the coupling neighborhood is about the coupling graph, not about spatial distance. The ratio-support conclusions (redundancy and variance ordering), by contrast, depend only on the number of factors and hold across all four, which is why we use every family for the ratio studies but highlight the dilemma for the advantage-outcome effect.

C.1.2 Coupling Families: Reward Definitions

All environments used in this paper are cooperative games with a prescribed coupling graph C, so the ground-truth coupling neighborhood is known exactly. For completeness, we briefly restate the notation. There are n agents; each agent i observes a phase or context and selects a discrete action $a^i \in \{1, \dots, K\}$, where K is the number of actions per agent. We write the joint action as $\boldsymbol{a} = (a^1, \dots, a^n)$. Each agent receives a per-agent reward $r_i(\boldsymbol{a})$, defined per family below, and the team return is their sum $\mathrm{R} = \sum_i r_i$. The coupling graph $C \in \{0,1\}^{n \times n}$ records reward dependencies, with $C_{ij} = 1$ iff r_i depends on a^j; its neighborhood $\partial i = \{j : C_{ij} = 1\}$ is the set of agents whose actions enter r_i. For the ring families, ∂i consists of the agents within graph distance $\rho^\star$ on a ring, and we refer to $\rho^\star$ as the coupling radius. Zero-mean Gaussian observation noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with standard deviation σ is added where noted.

Dense pairwise (symmetric, internalized). This family uses a symmetric pairwise payoff over the ring neighborhood,

$$r_i(\boldsymbol{a}) = \sum_{j \in \partial i,\, j \neq i} M[a^i, a^j] + \varepsilon_i, \qquad M = M^\top \in \mathbb{R}^{K \times K},\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{18}$$

where M is a fixed symmetric $K \times K$ payoff matrix whose entry $M[a^i, a^j]$ is the payoff agent i obtains from the action pair (a^i, a^j) with neighbor j, and ε_i is the observation noise defined above. Because M is symmetric, the same edge $M[a^i, a^j]$ contributes to both r_i and r_j; hence an agent's action affects its neighbors' reward as strongly as its own, so the externality is internalized. This game supplies the single-step results in Fig. 1 ($n = 20$, $K = 4$, $\rho^\star = 4$, $\sigma = 0.6$); the training study uses $n = 24$.

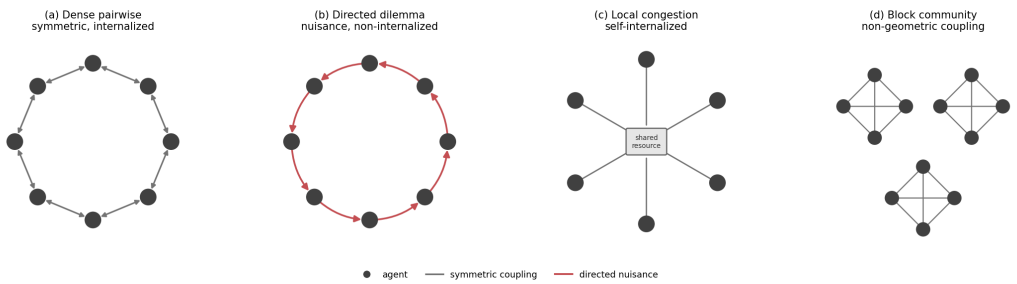


Figure 4: Schematic of the four coupling families (illustrated at a reduced scale). Nodes represent agents, and edges represent reward dependencies. **(a)** Dense pairwise: undirected edges couple ring neighbors symmetrically. **(b)** Directed dilemma: each agent transmits a one-way nuisance (red) to its downstream successors. **(c)** Local congestion: neighbors selecting the same resource share the reward. **(d)** Block community: all-to-all coupling within disjoint blocks, with no coupling across blocks.

Directed dilemma (non-internalized externality). On a directed ring, choosing a high-value action gives the actor a private benefit but imposes a nuisance on its $\rho^\star$ successors:

$$r_i(\boldsymbol{a}) = v(a^i) - c \sum_{j:\, i \in \partial^+ j} g(a^j), \tag{19}$$

where $v(a^i)$ is the private value agent i obtains from its own action, $g(a^j)$ is the nuisance that agent j's action imposes on its downstream neighbors, $c > 0$ is the weight of that nuisance, and $\partial^+ j$ denotes the $\rho^\star$ successors of j on the directed ring; the sum thus runs over those agents j whose successor set contains i, i.e., the predecessors that emit onto i. A rotating phase makes the tempting action time-varying. Since the actor does not bear the harm it causes, an independent advantage converges to the tragedy-of-the-commons equilibrium, whereas a neighborhood advantage recovers the social optimum. We use $n = 20$, $\rho^\star = 2$, $c = 3$.

Local congestion (self-internalized). Agents within ∂i that choose the same resource split its value:

$$r_i(\boldsymbol{a}) = \frac{v(a^i)}{\left|\{j \in \partial i : a^j = a^i\}\right|} + \varepsilon_i, \tag{20}$$

where $v(a^i)$ is the base value of the resource selected by action a^i, and the denominator counts the number of agents in the neighborhood ∂i (including i itself) that chose the same resource, so the resource value is shared equally among them. An agent that over-subscribes a popular resource thus depresses its own reward as well as its neighbors', so the crowding cost is partly self-borne. We use $n = 20$, $\rho^\star = 2$.

Block community (non-geometric coupling). Agents are partitioned into disjoint blocks; coupling is all-to-all within a block and absent across blocks, so C is block-diagonal, with an intra-block symmetric payoff matrix M as in the dense-pairwise case (18) (here ∂i is i's entire block rather than a ring neighborhood). This setup tests a coupling graph that is not a geometric ring. We use $n = 20$ and block size 5.

In every case, the reward parameters (M, v, g, block assignment) are fixed per environment. We report mean $\pm$ standard deviation over random seeds. Full hyperparameters are deferred to the appendix.

C.2 Extended Training-Curve Analysis

This subsection reports the full per-family training dynamics summarized in the main text: the advantage-support curves (complementing the sweep in Fig. 2) and the ratio-support redundancy curves. Together, they demonstrate, family by family, that enlarging the advantage support can change the learned policy, whereas enlarging the ratio support never does.

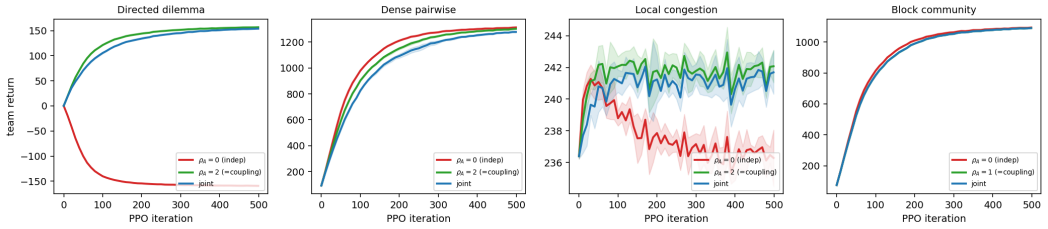


Figure 5: Advantage-support training curves across the four coupling families. Team return versus PPO iteration under three advantage supports: independent ($\rho_A = 0$), coupling-radius, and joint, with the ratio held per-agent. In the directed-dilemma and dense-pairwise panels, the seed variance is below 1.5% of the return scale, so the bands are narrower than the line width.

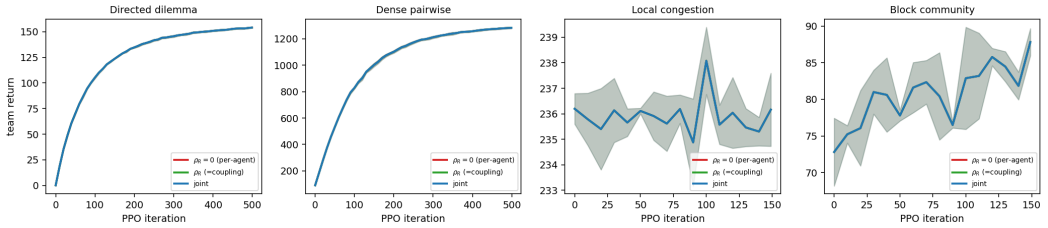


Figure 6: Ratio-support redundancy under standard PPO across the four coupling families. Team return versus PPO iteration with team advantage fixed. Each panel shows three ratio supports: per-agent ($\rho_R = 0$), coupling-radius, and joint.

Figure 5 plots the team return over PPO iterations for each family under three advantage supports, namely independent ($\rho_A = 0$), coupling-radius, and joint, with the ratio held per-agent. The directed dilemma shows a dramatic separation: independent advantage descends to the tragedy equilibrium, while coupling-radius and joint advantage climb to the social optimum. In the other three families, where the externality is self-internalized, independent advantage already learns effectively, and larger supports add variance rather than altering the outcome. This illustrates the advantage side of the asymmetry: the advantage support can shift the learned policy.

To demonstrate that the ratio radius is redundant, we fix the team advantage and sweep the ratio support $\rho_R \in \{\text{per-agent}, \text{neighborhood}, \text{joint}\}$ (Fig. 6). Within every family, the three learning curves are indistinguishable, overlapping to within seed noise with clip fraction 0 throughout, so enlarging the ratio support changes neither the trajectory nor the final outcome, exactly as predicted by Corollary 1. This is the ratio-side counterpart to Fig. 5: enlarging the advantage support can change the outcome, whereas enlarging the ratio support never does.

C.3 Real-World Traffic-Signal Control

Finally, we test the advantage-support prediction (Proposition 2) on a real cooperative traffic-signal control problem that natively provides both ingredients the theory requires: a per-agent reward (each intersection's local queue, delay, and throughput) and a physical coupling graph (the road network). We use the SUMO (Behrisch et al., 2011) Manhattan 28×7 grid ($n = 196$ traffic signals) via `sumo-rl` (Alegre, 2019), with the standard MLP-actor MAPPO from the toy suite; the only tunable parameter is the advantage support, built by k-hop expansion on the road-adjacency graph, with a per-agent ratio throughout.

Each intersection i is an agent with a local observation o_i (the standard `sumo-rl` encoding): a one-hot of the active green phase, a binary flag indicating whether the minimum green time has elapsed, and, for every incoming lane, the normalized vehicle density and queue length. The action a_i is discrete, the choice of the next green phase from that intersection's signal program, applied on a fixed 10 s control cycle with a 2 s yellow transition and a 5 s/50 s minimum/maximum green. The shared MLP actor maps o_i to a categorical distribution over a_i; agents share parameters but act

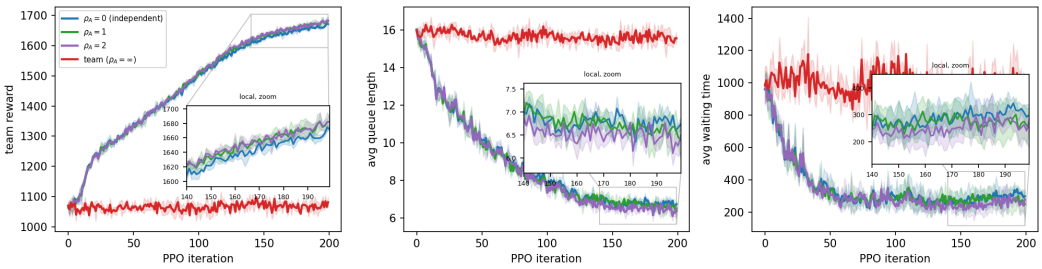


Figure 7: Traffic-signal control on a 196-intersection network. Advantage-support training curves (per-agent ratio) on all three metrics, namely team reward, average queue, and average waiting time, versus PPO iteration. Each panel compares three local advantage supports (independent ($\rho_A = 0$) and one/two-hop neighborhoods) against the global/team advantage ($\rho_A = \infty$). Insets zoom the last 60 iterations for the three local supports alone ($\rho_A = 0, 1, 2$).

Table 1: Traffic-signal control performance on 196 intersections.

advantage support	team return ↑	avg queue ↓	avg wait (s) ↓	clip frac
independent ($\rho_A = 0$)	1652.8±4.5	6.77±0.04	298.0±47.1	0.014
1-hop neighborhood	1663.7±7.5	6.68±0.21	275.9±57.7	0.016
2-hop neighborhood	1663.8±6.0	6.50±0.06	258.6±45.4	0.016
global / team ($\rho_A = \infty$)	1068.1±7.5	15.62±0.23	968.9±48.7	0.000

independently, so the joint policy factorizes as $\prod_i \pi^i(a_i)$, exactly as our analysis assumes. Each agent's reward is a purely local combination of its own traffic state:

$$r_i \;=\; w_v\,\bar{v}_i \;-\; w_q\,q_i \;-\; w_\omega\,\omega_i \;-\; w_p\,p_i, \qquad (w_v,\, w_q,\, w_\omega,\, w_p) = (3.0,\, 1.0,\, 0.3,\, 1.5), \quad (21)$$

where q_i is the total queue length, ω_i the total waiting time, p_i the inflow–outflow pressure $|\text{in} - \text{out}|$, and $\bar{v}_i$ the mean speed, each normalized per lane (by 50, 1000, 20 vehicles, and 15 m/s, respectively). Here r_i depends on other intersections only through physical spillover onto adjacent roads, so the reward coupling graph is exactly the road-network adjacency (each signal coupled to its up-to-four grid neighbors), and the range $\rho^\star$ is what our advantage support aims to match.

C.3.1 Advantage-Support Study: Global Advantage Collapses; Local Advantages Work (Fig. 7)

Figure 7 plots the three traffic metrics, namely team reward (the trained objective), average queue length, and average waiting time, over PPO iterations, comparing the global (team) advantage against three local advantage supports. Table 1 reports the converged metrics. A team advantage, the radius-∞ support that aggregates all 196 intersections' rewards as in vanilla MAPPO, fails to learn: its team return remains flat near its initial value while local advantages climb steadily, and at convergence it is 36% worse in return, with 2.4× the average queue and 3.7× the average waiting time (Table 1). This is Proposition 2(ii) at scale: summing 196 per-intersection rewards drowns each signal's own learning contribution in the noise of 195 others. In contrast, the three local supports, namely independent ($\rho_A = 0$) and one/two-hop neighborhoods, all learn well and closely track one another, with a mild monotone gain from including immediate neighbors. The insets confirm that these three curves differ only marginally and within the seed bands. The small spread among local radii indicates that traffic coupling is largely self-internalized (an intersection's own queue strongly reflects its own action), placing this system in the regime where the advantage-support bias is small and the dominant effect is the variance cost of an over-large support, exactly the failure mode exhibited by the team advantage. The team–local gap far exceeds the seed spread.

C.3.2 Ratio-Support Study: Redundancy and Variance Ordering

The advantage study above fixes a per-agent ratio and varies the advantage support. We now perform the converse, varying the ratio support among per-agent ($S^R = I$), 1-hop and 2-hop neighborhoods,

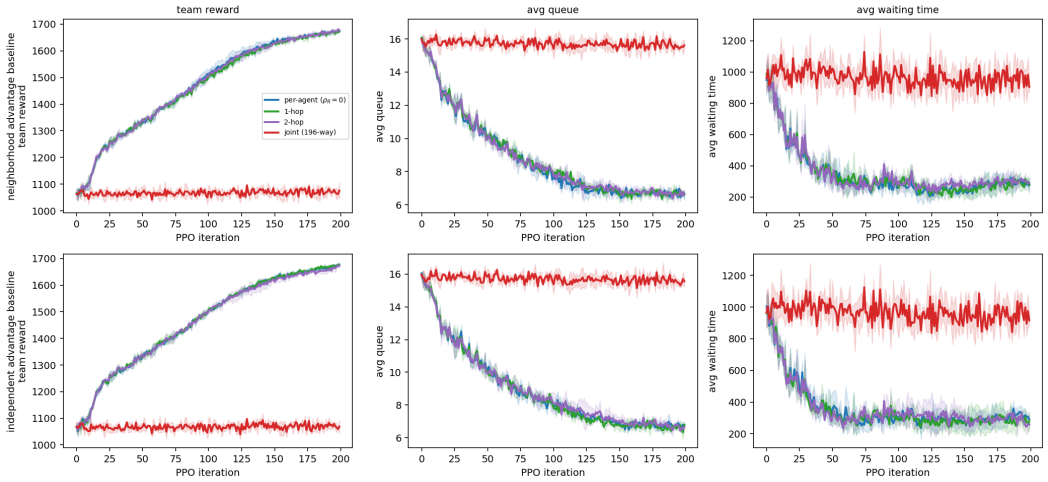


Figure 8: Ratio-support study on the 196-intersection network under a standard PPO configuration. Top row: the 1-hop neighborhood advantage held fixed. Bottom row: the independent ($S^A = I$) advantage held fixed. Columns show team reward, average queue, and average waiting time versus PPO iteration.

Table 2: Ratio-support study on the traffic network, under two fixed advantage baselines. Ratio supports: per-agent ($\rho_R = 0$), 1-hop, 2-hop, and joint (196-way).

advantage	ratio support	team return ↑	avg queue ↓	avg wait (s) ↓	clip frac
1-hop nbr	per-agent ($\rho_R = 0$)	1659.5±4.6	6.69±0.04	286.0±39.0	0.014
	1-hop	1657.1±2.7	6.71±0.06	280.3±36.6	0.016
	2-hop	1659.4±6.2	6.72±0.15	295.8±29.2	0.016
	joint (196-way)	1071.1±6.7	15.63±0.18	950.6±37.1	0.041
independent	per-agent ($\rho_R = 0$)	1656.5±6.5	6.77±0.09	301.4±42.9	0.015
	1-hop	1658.6±1.5	6.68±0.13	286.4±60.3	0.016
	2-hop	1649.7±6.1	6.79±0.12	291.0±8.1	0.015
	joint (196-way)	1070.6±7.0	15.63±0.21	945.8±46.1	0.041

and joint ($S^R = \mathbf{1}\mathbf{1}^\top$, the 196-way product). To verify that the ratio-side conclusion does not depend on how the advantage is chosen, we repeat the entire sweep under two fixed advantage baselines, the 1-hop neighborhood advantage and the independent ($S^A = I$) advantage. All runs use a standard PPO configuration. This isolates the ratio support on a real system and tests the two ratio-side predictions, redundancy (Corollary 1) and the variance ordering that worsens with n (Proposition 1, Fig. 3). Table 2 reports the final metrics of the two baselines.

Figure 8 plots, for each advantage baseline (rows) and each metric (columns), the training curve of every ratio support. The three small ratio supports, per-agent, 1-hop, and 2-hop, learn equally well and reach essentially the same reward, queue, and waiting time: enlarging the ratio support from 0 to 2 hops changes nothing about the learned policy, exactly as the redundancy corollary (Corollary 1) predicts. The joint 196-way ratio, in contrast, collapses outright, and it does so whether the advantage is the 1-hop neighborhood or fully independent. This is the sharpest real-system signature of the n-dependence in Fig. 3: with $n = 196$ agents, the joint ratio is a 196-way product whose variance factor $\prod_j (1 + \chi^2_j)$ is enormous, drowning the surrogate gradient in variance well before any benefit could accrue. Crucially, no off-policy stress is needed to expose this. Unlike the toy games, where $n \leq 24$ keeps the standard-configuration penalty dormant, the traffic network is large enough that the joint ratio fails under an ordinary, conservative PPO configuration. That the collapse and the redundancy are identical across the two advantage baselines confirms that the ratio-side conclusion is a property of the ratio support alone, independent of the advantage support with which it is paired.

D Related Work

Trust-region and proximal MARL. Extending PPO and TRPO to cooperative MARL requires deciding how the importance ratio is formed across agents. MAPPO (Yu et al., 2022) pairs a shared team advantage with a per-agent ratio and serves as a strong benchmark baseline. IPPO trains each agent independently, also with a per-agent ratio. At the other extreme, HATRPO and HAPPO (Kuba et al., 2022) update agents sequentially and clip a compound ratio, a product of the ratios of previously updated agents, justified by the multi-agent advantage-decomposition lemma. These methods span the ratio-support axis we study, from the per-agent extreme ($S^R = I$) to the fully joint extreme ($S^R = \mathbf{1}\mathbf{1}^\top$), yet the choice is typically made by architectural convention rather than analysis. Our results speak directly to this decision: the compound or joint ratio is gradient-redundant relative to a per-agent ratio with a suitably reweighted advantage (Theorem 1), and incurs strictly higher variance (Proposition 1).

Locality and scalability in networked MARL. A parallel line of work exploits the observation that, in networked systems, an agent's influence decays with graph distance. Qu et al. (2020b;a) formalize this as an exponential decay property and use it to prove that truncating each agent's value estimate to a κ-hop neighborhood incurs only $O(\rho^\kappa)$ error, yielding provably scalable actor–critic algorithms. This provides the theoretical grounding for our advantage-support axis: if reward coupling has finite range $\rho^\star$, then aggregating rewards over the $\rho^\star$-neighborhood is nearly unbiased, while a smaller support omits real externalities and a larger one only adds variance. We complement this literature by (i) separating the advantage support from the ratio support and (ii) showing that the two are redundant on the expected gradient, questions that are orthogonal to value-function truncation.

Counterfactual and difference credit. COMA (Foerster et al., 2018) and difference-reward methods (Wolpert & Tumer, 2001; Li et al., 2022) reduce variance by subtracting a counterfactual baseline whose expected gradient contribution is zero. Recent per-agent advantage estimators (Kim et al., 2026) prove policy-gradient invariance of such baselines under the factorized policy. These works concern the advantage or baseline and its zero-offset property; none addresses the ratio support or the factorization of the expected gradient through $\widetilde{S} = S^R S^A$, which is the object of our study.

Local advantages and local ratios. Several applied methods already restrict the advantage to a neighborhood without a general analysis. In traffic-signal control, MA2C (Chu et al., 2019) introduces a spatial discount factor that down-weights distant agents' rewards inside each local return, an advantage support tapered by graph distance. We observe an asymmetry in this literature: a rich set of methods aggregate the advantage locally, but we are not aware of any cooperative method that aggregates the ratio locally (e.g., a local importance-sampling weight $\prod_{j\in\partial i} \varrho^j$). Where a joint ratio does appear, it is typically not chosen for its own sake but inherited from casting the whole team as a single multi-action agent: running PPO on the factorized joint policy $\prod_i \pi^i$ treats $\boldsymbol{a}$ as one action and thus clips the product of all agents' ratios. This reduction is attractive because it turns MARL into single-agent RL, with its simpler objective and convergence story, and is used, for example, to make combinatorial cooperative tasks tractable as a single-controller problem. Our results explain the resulting high variance: the two supports are interchangeable on the expected gradient (Theorem 1), so the joint ratio buys nothing over a per-agent ratio, yet it is strictly the higher-variance route (Proposition 1), a product rather than a sum. This variance is therefore not a flaw of any particular single-agent-reduction algorithm but an inherent cost of aggregating on the ratio side, and it can be removed by moving the aggregation into the advantage. The field's revealed preference for local advantages over local ratios is exactly what this variance-ordering predicts. We subsume the local-advantage heuristics as particular advantage-support choices and provide the design principle behind them.

E Discussion

Existing methods on the bias–variance plane. The two supports provide a common coordinate system for methods that otherwise appear unrelated. IPPO sits at the low-variance, high-bias corner ($S^A = S^R = I$): both supports per-agent, so its estimator has minimal variance but ignores every externality. MAPPO moves the advantage to the team ($S^A = \mathbf{1}\mathbf{1}^\top, S^R = I$): it removes the

missing-coupling bias but, on a large team, pays the additive variance of summing all rewards, and as our 196-intersection experiment shows, this already fails when the team is large and only a local neighborhood is truly coupled. Neighborhood-advantage methods (spatial-discount and local-critic actor–critics) occupy the principled middle: S^A is a k-hop neighborhood, $S^R = I$, which our analysis identifies as the MSE-optimal choice when the coupling has range k. At the far corner lie single-agent reductions that treat $\prod_i \pi^i$ as one multi-action policy, for example CMAT (Zhao et al., 2026), which factorizes the joint policy through a latent consensus and then optimizes it with single-agent PPO on the joint action. Such methods effectively push both supports to fully joint ($S^A = S^R = \mathbf{1}\mathbf{1}^\top$), so $\widetilde{S} = n\,\mathbf{1}\mathbf{1}^\top$, which is the maximal effective support. In our terms, this is the least biased and most variant point in the plane: it aggregates every neighbor's reward and multiplies every neighbor's ratio. Our results indicate that the first is unnecessary beyond the coupling neighborhood and the second is strictly harmful. The elegant equivalence to single-agent RL is real at the level of the expected objective, but it is bought with a variance that, in a large system, need not pay off. This is not a defect of any particular such algorithm but an inherent property of the reduction, in the sense that the maximal-support corner is high-variance by construction, and our canonical form shows the variance is removable without changing what is learned, by moving all aggregation onto the advantage and sizing it to the coupling graph. This reframes per-agent versus centralized not as a binary but as a point in a two-dimensional support plane whose optimal location the theory pins down.

Does the result apply to multi-action (centralized) PPO? A natural concern is that our analysis concerns multi-agent PPO, whereas a centralized controller may treat the whole team as a single agent whose action is the joint $\boldsymbol{a} = (a^1, \ldots, a^n)$ and run vanilla single-agent PPO on it, the multi-action view underlying CMAT (Zhao et al., 2026). Our results apply directly to this setting, because the only structural assumption we use is that the policy factorizes across action components, $\pi(\boldsymbol{a}) = \prod_i \pi^i(a^i)$ (Assumption 1), i.e., independent categorical heads, which is exactly how such multi-action policies are built. Under that factorization, multi-action PPO is the corner $(S^A, S^R) = (\mathbf{1}\mathbf{1}^\top, \mathbf{1}\mathbf{1}^\top)$: a single scalar advantage (the team return) is broadcast to every action component, and the importance weight is the ratio of the joint action, $\prod_i \varrho^i$. The canonical form (Theorem 1) therefore governs it unchanged: the joint ratio is gradient-redundant relative to a per-agent ratio with a reweighted advantage, and strictly higher-variance (Prop. 1), so the design rule transfers verbatim: clip each action component's ratio separately rather than as one n-way product, which turns multi-action PPO back into a per-agent-ratio scheme (a MAPPO-style update) at no cost to the expected gradient. Two boundaries are worth stating. First, the factorization is essential: an autoregressive joint head (component a^j conditioned on $a^{<j}$) uses a different chain rule and is out of scope. Second, when the environment provides only a single non-decomposable team reward, the advantage cannot be localized (there is no per-agent reward to aggregate over a neighborhood), so the bias-reduction half of our rule is inapplicable. However, the variance half still holds, in that the joint ratio should be replaced by a per-agent ratio regardless, which is precisely the sense in which a MAPPO-style per-agent ratio dominates a fully centralized multi-action update even in the single-team-reward regime.

Two caveats the experiments make precise. The design rule is exact at the level of the expected gradient, but two qualifications govern how visibly it matters in training. First, the advantage support reshapes the learned policy only when the coupling externality is not already internalized by an agent's own reward: in the directed dilemma, independent advantage collapses to the tragedy equilibrium and a neighborhood advantage recovers the social optimum, whereas in symmetric-payoff, congestion, and block games, independent advantage is already near-unbiased and larger supports only add variance (Fig. 2). Second, the ratio's variance penalty is gated by the per-update policy shift χ^2: a conservative trust region keeps χ^2 so small that per-agent and joint ratios are indistinguishable even at hundreds of agents, and the penalty surfaces only as updates are driven off-policy (Fig. 3). Neither qualification weakens the design rule, since the ratio's benefit is exactly zero and its cost is non-negative throughout, but both explain why the folklore that per-agent ratios are more stable is only intermittently observed in practice.

Implications for large-scale deployment. The design principle is most consequential precisely where cooperative MARL is scaling: systems with hundreds to thousands of agents and local physical coupling, such as traffic-signal grids, sensor and robot fleets, power and communication networks, and warehouse logistics. Two concrete recommendations follow. *(1) Never aggregate the ratio.* At

these scales, the joint ratio's variance factor $(1 + \chi^2)^n$ is enormous the moment any batch drifts off-policy, and our 196-intersection result shows the analogous failure for an over-large advantage (a global team advantage never learns, while local advantages cut queues by $2.4\times$ and waiting time by $3.7\times$). Keep the ratio per-agent. Put every bit of cross-agent aggregation in the advantage, where the cost is additive. *(2) Size the advantage to the physical coupling.* Because these systems come with a known interaction graph (the road network, the communication topology, the spatial adjacency), the MSE-optimal advantage support is directly available, namely the k-hop neighborhood for a coupling of range k, rather than something to be tuned blindly. This turns a hyperparameter search into a modeling choice and connects to the scalable-MARL line (Qu et al., 2020b;a), whose exponential-decay property is exactly the condition under which a small advantage neighborhood is near-unbiased. Neither recommendation requires changing the network architecture or the training loop, only which rewards and which ratios enter each agent's update, so they are immediately actionable for existing MAPPO and IPPO codebases.

Future work: learning the support. Our design rule assumes the coupling neighborhood is known, which holds for physically networked systems but not in general. When the coupling graph is unknown or state-dependent, the open problem is to learn the right advantage support: which neighbors to aggregate, and at what radius, so as to minimize gradient MSE. This is a bias–variance model-selection problem, namely estimating each agent's influence graph (e.g., from reward sensitivities or attention over other agents' states) and including an agent in the support when its estimated coupling outweighs the variance its reward adds. A dynamic, per-state support that grows in strongly-coupled regimes and shrinks otherwise would sit at the MSE optimum adaptively, generalizing the fixed k-hop rule. The canonical form guarantees that whatever support is learned, it should be realized on the advantage and never on the ratio.