FINITE ROLLOUT GROUPS AS SCORE-CONDITIONED POLICY-GRADIENT ORACLES

Anonymous authors
Paper under double-blind review

ABSTRACT

A finite rollout group determines which population policy-gradient fields are available, not only how accurately they are estimated. We study iid on-policy groups over finite outcomes with outcome-measurable, permutation-equivariant, stop-gradient coefficients and no direct policy input. Conditioning leaves $G-1$ peers: the conditional coefficient fields are exactly polynomials of degree at most $G-1$. Removing common baselines separates field representation from objective existence. Ordinary reward standardization has an exact boundary: for every finite $G \geq 2$ and every positive strictly increasing variance denominator, a scalar simplex potential exists if and only if rewards have at most two distinct levels. Count-based rules yield exact self-inclusion and centering transformations. Under affine reuse of two fixed coefficient rules across linear reward weights, with positive response, we further prove $\Theta(G^{-2})$ minimax loss at the best surrogate optimum over a fixed strongly concave target class. Controlled Qwen2.5 studies at 1.5B and 7B identify count-rule objectives: all 65 method–seed endpoints are closest to their assigned optimum among three analytic candidates. Separate tests recover a rollout threshold, self-inclusion fixed points, and the predicted zero-to-nonzero circulation transition from two to three reward levels under ordinary standardization.

1 INTRODUCTION

A rollout group can change the expected policy-gradient update, even when sampling error vanishes. We ask: *which population fields can a group of G rollouts represent, and when are those fields gradients of an objective?*

A binary example illustrates the distinction. Let p be the success probability. With two iid rollouts, averaging the scores of successful trajectories, with zero update when both fail, has expectation $(2-p)\nabla_\theta p$. Subtracting the group-average score only when the group contains a success changes the expectation to $\nabla_\theta p$. These updates optimize different functions of p. Section 5.1 extends this calculation to count-based rules and identifies the corresponding mismatch in MaxRL v1 (Tajwar et al., 2026).

Conditioning on the scored outcome leaves $G-1$ multinomial peers. The conditional fields are exactly the polynomials of degree at most $G-1$, realized by peer-count lookups. This defines the group's *score-conditioned policy-gradient oracle*: its observations determine the available population fields. The construction applies the fixed-sample polynomial principle (Kimpara et al., 2023); binary Bernstein representations and inverse design already appear in Davis & Recht (2025) and Zheng (2026).

With several outcomes, a represented field need not have an objective. Removing common baselines yields a simplex differential whose integrability decides objective existence. For ordinary within-group reward standardization with a positive, strictly increasing variance denominator, a scalar potential exists if and only if rewards take at most two distinct values, for every finite $G \geq 2$. A third reward level creates an obstruction absent in binary objective representations (Section 4).

Self-inclusion, centering, and occupancy weighting also select different objectives (Section 5). A separate consequence concerns optimum loss: when two fixed binary coefficient rules combine

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.

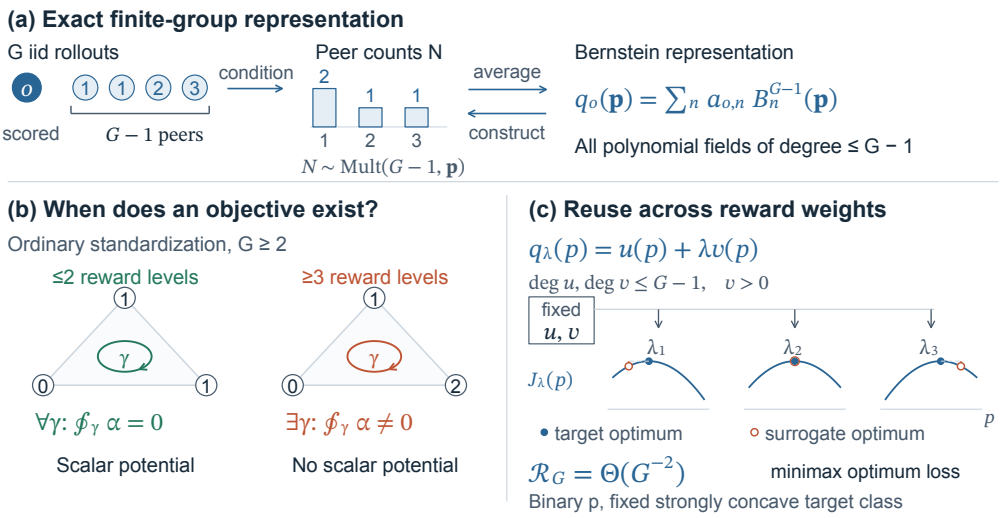


Figure 1: **An exact representation and two consequences.** (a) Under Section 2's assumptions, conditioning leaves $N \sim \mathrm{Mult}(G-1, p)$, with Bernstein weights $B_n^{G-1}(p) = \Pr(N = n)$, $|n| = G-1$. The coefficients $q_o(p) = \mathbb{E}[A_i \mid O_i = o]$ are exactly the polynomial fields of degree at most $G-1$ via peer-count construction (Section 3). (b) For ordinary standardization with a positive, strictly increasing variance denominator, a simplex potential exists exactly when rewards have at most two levels; $\alpha = \sum_{a<K}(q_a - q_K)\,dz_a$. The $K = 3$ examples contrast vanishing integrals on every closed loop with the existence of a nonzero loop. (c) Fixing the binary rules u, v before varying λ, with $v > 0$, yields minimax loss $\mathcal{R}_G = \Theta(G^{-2})$ at the best surrogate optimum over Section 6's fixed target class. The curves $J_\lambda(p)$ illustrate reuse and optimum gaps schematically.

affinely across linear reward weights, with positive response, the degree limit yields tight $\Theta(G^{-2})$ worst-case loss at the best surrogate optimum over a fixed strongly concave target class (Section 6).

Figure 1 summarizes the results; controlled Qwen and iid-group experiments test the fields (Section 7).

2 Finite-Group Policy-Gradient Estimators

Fix a prompt x and finite outcomes $\mathcal{O} = \{1, \ldots, K\}$. Write

$$\Delta_K = \{p : p_o \geq 0,\ \textstyle\sum_o p_o = 1\}, \qquad \Delta_K^\circ = \{p : p_o > 0,\ \textstyle\sum_o p_o = 1\}$$

for the closed simplex and its interior. Draw $Y_i \sim \pi_\theta(\cdot \mid x)$ iid and set $O_i = T(Y_i)$ for a deterministic canonicalizer. Distinct trajectories may share an outcome; in MultiRoute and MultiSolve, malformed responses map to an explicit invalid outcome. The sampled score is the gradient of the trajectory's log probability:

$$s_i = \nabla_\theta \log \pi_\theta(Y_i \mid x).$$

We study the estimator

$$\widehat{g} = \frac{1}{G}\sum_{i=1}^{G} A_i(O_{1:G}, U)\, s_i. \tag{1}$$

The fixed rule A_i is outcome-measurable, permutation-equivariant across group members, treated as stop-gradient, and satisfies $\mathbb{E}|A_i| < \infty$ and $\mathbb{E}[|A_i|\|s_i\|] < \infty$. It has no direct dependence on p or θ. The auxiliary U has a fixed law independent of θ, p, and all rollouts. We assume finite trajectory support or dominated score regularity permitting differentiation of outcome probabilities; the experiments use bounded generation.

For $p \in \Delta_K^\circ$, conditioning gives

$$p_o = \Pr(O = o), \qquad \mu_o = \mathbb{E}[s_i \mid O_i = o] = \nabla_\theta \log p_o, \qquad \sum_o p_o \mu_o = 0. \tag{2}$$

Define $q_o^{\mathrm{cond}}(p) = \mathbb{E}[A_i \mid O_i = o]$ on the interior. Because A_i observes Y_i only through O_i and the peers and U are independent of Y_i, the coefficient and score are conditionally independent given O_i. Thus

$$\mathbb{E}[\widehat{g}] = \sum_o p_o q_o^{\mathrm{cond}}(p)\mu_o = \sum_o \bar{q}_o(p)\nabla_\theta p_o, \tag{3}$$

where Section 3 constructs the polynomial extension $\bar{q}_o$ of q_o^{cond}. Boundary statements use this extension, not conditioning on a zero-probability outcome. The random estimator itself always uses the sampled score s_i.

The policy has *full simplex tangent rank* when the Jacobian $D_\theta p$ has rank $K-1$ at the points under consideration. This condition is needed to identify coefficients from parameter-space gradients, not to establish their polynomial degree. Off-policy reuse, token-dependent credit, learned critics, and coefficients differentiated through θ are outside Equation 1's information and differentiation assumptions.

3 Which fields can a finite group represent?

Let $C_o = \sum_i \mathbf{1}\{O_i = o\}$ and let e_o be the oth basis vector. For a focal outcome o and peer count n, let $\mathcal{C}_{o,n}$ contain all ordered configurations with $O_i = o$ and $C - e_o = n$. Define the fixed count-cell average

$$a_{o,n} = \mathbb{E}_U\left[\frac{1}{|\mathcal{C}_{o,n}|}\sum_{O_{1:G}\in\mathcal{C}_{o,n}} A_i(O_{1:G}, U)\right]. \tag{4}$$

It is independent of p and θ, including for cells that have zero probability at a boundary point.

Theorem 1 (Conditional polynomial representation and converse)**.** *Under Section 2's assumptions, each conditional coefficient $q_o^{\mathrm{cond}}(p)$ agrees on Δ_K° with a polynomial of total degree at most $G-1$. Conversely, every componentwise polynomial field of degree at most $G-1$ has a permutation-equivariant peer-count realization. The Bernstein representation defines a continuous extension to Δ_K.*

Construction. Conditioning leaves the peer count $N = C - e_o \sim \mathrm{Mult}(G-1, p)$. Hence

$$\bar{q}_o(p) = \sum_{|n|=G-1} a_{o,n}\binom{G-1}{n}\prod_{j=1}^K p_j^{n_j}. \tag{5}$$

Conversely, expand a target polynomial in this degree-$(G-1)$ Bernstein basis and use $A_i = a_{O_i, C-e_{O_i}}$; lower-degree terms are degree-elevated. Full proofs are in Appendices C.2–C.3. Below, q_o denotes the polynomial extension unless conditioning is explicit.

For binary $O_i \in \{0,1\}$, let $G = 3$ and $A_i = O_i\prod_{j\neq i} O_j$. Given a scored success, its coefficient is one exactly when both peers succeed: $q_1(p) = p^2$, $q_0(p) = 0$, and $\mathbb{E}[\widehat{g}] = p^2\nabla_\theta p$. No two-rollout rule in this class represents that field throughout $(0,1)$, because its coefficient difference is affine. Averaging arbitrarily many independent two-rollout groups does not remove this expressivity limit.

Common shifts leave the update unchanged because $\sum_o \nabla_\theta p_o = 0$. Write $q \sim q + b\mathbf{1}$, and let $[q]$ denote this equivalence class. Define $\mathfrak{F}_G$ as the classes admitting a polynomial representative of degree at most $G-1$. At full simplex tangent rank these are exactly the representable population fields modulo common baselines.

Corollary 1 (Rollout hierarchy)**.** *For $K \geq 2$,*

$$\mathfrak{F}_1 \subsetneq \mathfrak{F}_2 \subsetneq \cdots, \qquad G_{\min}([q]) = 1 + \deg_\Delta[q]. \tag{6}$$

Here $\deg_\Delta[q]$ is the minimum polynomial degree of a representative, or ∞ if none exists, with $\deg_\Delta[0] = 0$.

Strictness follows from the binary differential $p^G\,dp$, which belongs to $\mathfrak{F}_{G+1}$ but not $\mathfrak{F}_G$. For larger K, the same witness is obtained by varying one coordinate on an interior simplex slice.

4 WHEN DOES THE FIELD HAVE AN OBJECTIVE?

By an objective we mean a single function $J(p)$ whose gradient agrees with the expected update throughout the domain. The differential relevant to that update is $\alpha = \sum_o q_o \, dp_o$. In simplex coordinates $z_a = p_a$ for $a < K$, with $p_K = 1 - \sum_{a<K} z_a$, it becomes $\alpha = \sum_{a<K} \omega_a \, dz_a$, where $\omega_a = q_a - q_K$.

Theorem 2 (Common baselines and integrability)**.** *At full simplex tangent rank, two coefficient fields induce the same policy gradient exactly when they differ by $b(p)\mathbf{1}$. A represented field has a scalar simplex potential on Δ_K° if and only if*

$$\partial_{z_b} \omega_a = \partial_{z_a} \omega_b \qquad (a, b < K). \tag{7}$$

When q has degree at most $G - 1$, the potential can be chosen polynomial of degree at most G.

The quotient removes directions invisible to simplex-tangent changes. Integrating a closed one-form on the convex simplex interior proves the second statement (Appendix C.4): an objective exists exactly when line integrals are path independent. The test is automatic for $K = 2$; with more outcomes, nonzero closed-loop work rules out a potential. A differentiable batch loss does not guarantee this property. Its stop-gradient coefficients have a policy-dependent sampling distribution; Theorem 2 tests the resulting population field.

4.1 ORDINARY REWARD STANDARDIZATION: AN EXACT CLASSIFICATION

For fixed rewards r_o, normalize the original reward within each group:

$$\bar{r} = \frac{1}{G} \sum_i r_{O_i}, \qquad V_r = \frac{1}{G} \sum_i (r_{O_i} - \bar{r})^2, \qquad A_i = \frac{r_{O_i} - \bar{r}}{D_G(V_r)}. \tag{8}$$

The function $D_G : [0, \infty) \to (0, \infty)$ is finite-valued, strictly increasing and independent of p, θ. Both $\sqrt{v + \epsilon}$ and $\sqrt{v} + \epsilon$, with $\epsilon > 0$, qualify. The rule is stop-gradient.

Theorem 3 (Reward-level classification)**.** *For every finite $G \geq 2$, Equation 8 admits a scalar potential on the full simplex interior if and only if $|\{r_1, \ldots, r_K\}| \leq 2$. With two reward levels its potential is a polynomial of degree at most $G - 1$; with one level the field is zero.*

Figure 1(b) summarizes this boundary; E4 in Section 7 samples the same $K = 3$ two-versus-three-level transition as a finite-group recovery test.

Two reward levels. Let t be the high-reward mass. Equal-reward outcomes share coefficients $q_H(t)$ or $q_L(t)$, so $\alpha = [q_H(t) - q_L(t)] \, dt$ is integrable regardless of the number of answers. For reward gap $\delta > 0$ and $G = 2$, $q_H - q_L = \delta / [2D_2(\delta^2/4)]$, a constant. The objective is then a positive multiple of t.

Why a third level changes the answer. Choose $r_i < r_k < r_j$, with gaps $a = r_k - r_i > 0$ and $b = r_j - r_k > 0$. Near the middle-reward vertex e_k, a first derivative of the peer law replaces one middle-reward peer by a low- or high-reward peer. The relevant complete groups are therefore one low outcome, one high outcome, or one of each, with all other members at the middle reward. Their variances are

$$v_a = \frac{(G-1)a^2}{G^2}, \qquad v_b = \frac{(G-1)b^2}{G^2}, \qquad v_{ab} = \frac{(G-1)(a^2+b^2) + 2ab}{G^2}.$$

In the chart using k as reference, the polynomial extension satisfies

$$\begin{aligned} \mathcal{C}(e_k) &:= \left[\partial_{p_j}(q_i - q_k) - \partial_{p_i}(q_j - q_k)\right]_{e_k} \\ &= (G-1)\left[\frac{a}{D_G(v_a)} + \frac{b}{D_G(v_b)} - \frac{a+b}{D_G(v_{ab})}\right] > 0. \end{aligned} \tag{9}$$

Since $v_{ab} > v_a, v_b$, the mixed group discounts the combined reward difference more than either single excursion. Polynomial continuity transfers the nonzero residual to nearby full-support interior points, without conditioning on a null event. Appendix A gives the binary potential and zero-variance conventions. Binary GRPO objective representations (Davis & Recht, 2025) thus have an exact reward-level boundary, rather than extending to this normalization with arbitrary rewards.

5 SELF-INCLUSION AND THE OBJECTIVES OF COUNT RULES

The conditional law also explains why an unbiased reward statistic can produce a different score-weighted update when it is reused on its own samples. For binary outcomes, let C be the full-group success count and let $a_0, \dots, a_G$ be a fixed lookup. Its marginal expectation is $F_G[a](p) = B_G[a_0, \dots, a_G](p)$, where B_n denotes the binomial Bernstein polynomial. Attaching this lookup to a scored success changes the count law from $\mathrm{Bin}(G, p)$ to $1 + \mathrm{Bin}(G-1, p)$.

Theorem 4 (Self-inclusion transform)**.** *The success-conditioned coefficient is*

$$S_G[a](p) = B_{G-1}[a_1, \dots, a_G](p) = F_G[a](p) + \frac{1-p}{G} F'_G[a](p). \tag{10}$$

An exact degree-m marginal coefficient needs at least m samples; its peer-only policy-gradient realization needs at least $m+1$ total rollouts.

For $F_G(p) = p^m$, reuse adds $(m/G)(1-p)p^{m-1}$ to the coefficient. Leave-one-out (LOO) instead evaluates the statistic on independent peers. It restores the intended conditional mean, at the cost of an additional rollout when the degree bound is tight. The Bernstein derivative identity proves the transform (Appendix C.5); Section 7 tests its predicted change in a learned fixed point.

5.1 EXACT OBJECTIVES AND THE EFFECT OF CENTERING

For an occupancy budget $B \geq 1$, define the probability that an outcome appears at least once and its associated potential by

$$h_B(t) = 1 - (1-t)^B, \qquad \Psi_B(t) = \int_0^t \frac{h_B(u)}{u}\, du, \qquad J_B(p) = \sum_o r_o \Psi_B(p_o). \tag{11}$$

The integrand has continuous value B at zero. The set-level objective is $U_G(p) = \sum_o r_o h_G(p_o)$. Empirical inverse-probability scaling (EIPS) assigns $R_i = G r_{O_i}/C_{O_i}$ to each observed outcome. We consider the raw rule $A_i = R_i$, same-group centering $A_i = R_i - G^{-1}\sum_j R_j$, and the set-marginal rule $A_i = G r_{O_i} \mathbf{1}\{C_{O_i} = 1\}$.

Theorem 5 (Objectives of count-based rules)**.** *Under Section 2's assumptions, the rules in Table 1 have the listed conditional coefficients modulo common baselines, and their expected updates equal the gradients of the listed objectives. For GOC, the oracle additionally observes $L-1$ independent on-policy probe groups of size G; only the main group carries scores.*

For raw EIPS, the average coefficient mass assigned to outcome o is $r_o \mathbf{1}\{C_o > 0\}$. Conditioning on a scored occurrence gives $q_o = r_o h_G(p_o)/p_o$. Centering subtracts a common baseline and the outcome-specific term $r_o(1-p_o)^{G-1}$. The identity $h_G(t) - t(1-t)^{G-1} = h_{G-1}(t)$ then gives J_{G-1}. Thus centering changes the population objective, not only estimator noise. Full derivations are in Appendix C.6.

Geometric occupancy correction (GOC) raises the effective observation budget using independent probes. At $L = 2$, it multiplies a main-group EIPS coefficient by $1 + \mathbf{1}\{o \text{ absent from the probe}\}$. Independence changes the presence weight from h_G to $h_G(2 - h_G) = h_{2G}$. Additional probes give the h_{GL} weight in the table; Appendix C.7 specifies the general rule. The marginal $\Psi'_B(t) = \sum_{k=0}^{B-1}(1-t)^k$ makes the budget effect explicit: the relative weight of rare outcomes changes with B.

Table 1: **Different group computations select different objectives.** Here $t = p_o$ and $L \geq 1$. The count before/after the slash is the number of observed/scored rollouts. Centering requires $G \geq 2$; the other identities also hold at $G = 1$. Coefficients are shown modulo a common baseline, with continuous values at $t = 0$.

Rule	Rollouts	Conditional coefficient	Objective
Raw EIPS	G/G	$r_o h_G(t)/t$	J_G
Centered EIPS	G/G	$r_o h_{G-1}(t)/t$	J_{G-1}
Set-marginal	G/G	$G r_o (1-t)^{G-1}$	U_G
GOC(G, L)	GL/G	$r_o h_{GL}(t)/t$	J_{GL}

The centering distinction in MaxRL v1. For binary rewards, let C_+ count successes and $\bar{s} = G^{-1}\sum_i s_i$. MaxRL v1's raw estimator (Tajwar et al., 2026) is the binary EIPS specialization. Its on-policy Algorithm 1 also subtracts $\bar{s}$, but skips both terms when $C_+ = 0$. Although $\mathbb{E}[\bar{s}] = 0$, the success-gated subtraction has mean

$$\mathbb{E}[\mathbf{1}\{C_+ > 0\}\bar{s}] = \frac{1}{G}\nabla_\theta h_G(p) = (1-p)^{G-1}\nabla_\theta p. \tag{12}$$

The estimator in Algorithm 1 of the cited v1 follows J_{G-1} rather than J_G. The gate ties the subtracted baseline to the same samples that carry the scores. Appendix D.1 gives the calculation, including the all-failure event.

Corollary 2 (Inverse design)**.** *At full simplex tangent rank, an objective is exactly realizable at budget G if and only if its simplex differential has a polynomial coefficient representative of degree at most $G-1$. The Bernstein coefficients then give a constructive peer-count lookup.*

For example, $\sum_o r_o \log p_o$, with $r \neq 0$ and $K \geq 2$, has a nonpolynomial reciprocal differential, so no fixed finite budget realizes it throughout the interior (Appendix C.8). Approximation is possible away from zero. An unbounded GOC construction reaches the log field under the moment and expected-cost conditions in Appendix D.2.

6 An optimum-loss consequence of the degree limit

A field can be difficult to represent while its optimum is easy to attain: the two-rollout field $p^\star - p$ has any prescribed Bernoulli optimum $p^\star$. To obtain a loss bound, we therefore constrain how one rule serves several objectives. Consider $J_\lambda(p) = J(p) + \lambda p$ and require $A_i^{(\lambda)} = A_i^{(0)} + \lambda A_i^{(1)}$: the designer may know the entire J, but fixes both rules before the linear reward weight λ is selected. Theorem 1 gives

$$q_\lambda(p) = u(p) + \lambda v(p), \qquad u, v \in \mathcal{P}_{G-1}, \quad v(p) > 0 \quad (p \in I), \tag{13}$$

where q_λ is the binary coefficient difference and $\mathcal{P}_d$ denotes polynomials of degree at most d. Positive v is an additional assumption; $v \equiv 1$ adds the linear reward gradient exactly. Fix $I = [1/4, 3/4]$, $\Lambda = [-3/8, 3/8]$, and the budget-independent class

$$\mathcal{J} = \{J \in C^2(I) : J(1/2) = J'(1/2) = 0, \quad 1/2 \leq -J'' \leq 3/2\}. \tag{14}$$

Let $x_J(\lambda)$ maximize J_λ, and let $M_{u,v}(\lambda)$ be all global maximizers of the surrogate $\Phi_{u,v,\lambda}(p) = \int_{1/2}^{p}[u(t) + \lambda v(t)]\,dt$. Define

$$\mathcal{R}_G = \sup_{J\in\mathcal{J}} \inf_{u,v \text{ as in } (13)} \sup_{\lambda\in\Lambda} \min_{y\in M_{u,v}(\lambda)} [J_\lambda(x_J(\lambda)) - J_\lambda(y)]. \tag{15}$$

The minimum grants the surrogate its best optimum in a tie, so the loss persists even after favorable global optimization of the surrogate.

Theorem 6 (Optimum loss under reward-affine reuse)**.** *For every $G \geq 2$,*

$$\frac{1}{64\pi^2(G+3)^2} \leq \mathcal{R}_G \leq \frac{81\pi^2}{64G^2}. \tag{16}$$

The upper bound is realizable with $v \equiv 1$. The lower bound holds on $G+2$ weights for an explicit target in $\mathcal{J}$ and covers every surrogate global optimum at a hard weight.

Figure 1(c) isolates this family-level consequence from the single-field representation hierarchy.

Proof idea. A linear derivative plus an oscillation of amplitude $\Theta(G^{-1})$ keeps the target strongly concave. Tracking all $G+2$ selected optima within that amplitude would force a degree-G polynomial to change sign $G+1$ times. Strong concavity turns the resulting distance into $\Omega(G^{-2})$ loss. Conversely, a Jackson polynomial approximates J' within $O(G^{-1})$; Lemma 2 bounds optimum loss by the squared derivative error. Appendix B proves both bounds, including boundary optima, and gives a separate polynomial threshold example.

The restriction is on a reusable family of population fields, not on all learning algorithms. Rules redesigned at each weight can encode its optimum, and standardizing a newly mixed reward is generally not affine in that weight. The fixed target class bounds curvature, not higher smoothness; its worst-case target can vary with G.

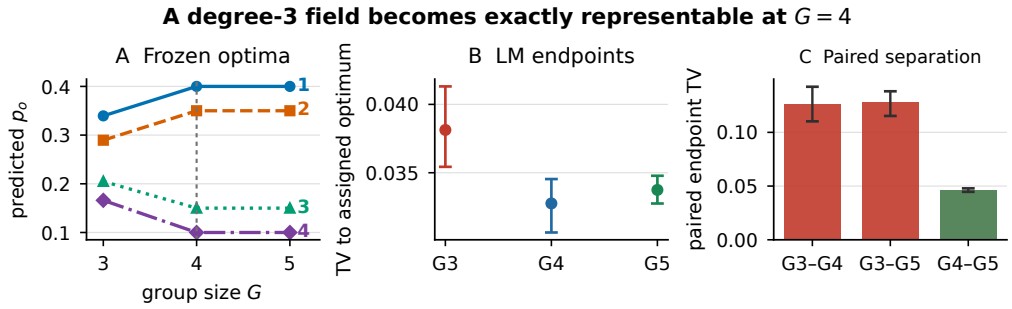


Figure 2: **The cubic field becomes exactly representable at $G = 4$.** G3 uses a quadratic projection; G4/G5 use the exact field (Appendix D.3). Error bars show descriptive 95% bootstrap intervals over five paired seeds.

7 Experiments: Identifying the Predicted Fields

We test four predictions about population fields and their consequences for finite computation. Exact finite-support checks establish the implemented fields; free-generation Qwen studies compare distributions with predicted optima; iid-group sampling tests circulation. These are identification tests, not estimator rankings. Targets were fixed before the reported runs. Appendix E records the protocols and development-informed calibration.

For each test prompt, let $p^\star$ be the predicted optimum and $\widehat{p}$ the estimated outcome distribution. We report

$$\mathrm{TV}(\widehat{p}, p^\star) = \frac{1}{2} \sum_o |\widehat{p}_o - p^\star_o|,$$

averaged over prompts. Intervals are descriptive 95% percentile bootstraps over training seeds ($n = 5$ at 1.5B, $n = 3$ at 7B), or independent sampling replicates ($n = 64$ in E4). Raw EIPS, set-marginal, GOC and the E1 lookup use scalar control variates computed from other independent groups. They leave the core mean fields unchanged; their policy-dependent sampling law is distinct from the fixed-law auxiliary U in Equation 1.

Exact field checks. Exhaustive enumeration of a length-four binary autoregressive policy checks the self-inclusion transform: a full-count coefficient can be marginally exact but biased after score conditioning, while LOO matches the intended field to numerical precision. A separate trie-restricted Qwen audit makes leaf probabilities and trajectory scores enumerable. Raw, centered, set-marginal, LOO and GOC/direct-budget gradients match their analytic fields, while full-count reuse has the predicted deviation. These checks identify the estimator before any interpretation of a training endpoint (Appendix E).

E1: a field's minimum rollout budget. A frozen cubic quotient field has minimum exact budget $G = 4$. G3 implements its unique degree-two projection under uniform simplex measure; G4 uses an exact lookup and G5 a degree elevation of the same field. The projection residual is 0.053, while G4/G5 fields agree to numerical precision. On MultiRoute with Qwen2.5-1.5B and five paired seeds, assigned-optimum TV is 0.038 [0.035, 0.041] for G3, 0.033 [0.031, 0.035] for G4, and 0.034 [0.033, 0.035] for G5. The cross-threshold paired-TV intervals start at 0.110, above the G4–G5 upper endpoint 0.048 (Figure 2). The target optima are separated by TV 0.121, and the endpoint separation tracks the distinction between the chosen projection and the exact field.

E2: objective selection across output surfaces and model sizes. Theorem 5 predicts centered G4 $\to J_3$; raw G4, centered G5, and GOC$(2,2) \to J_4$; and set G4 $\to U_4$. MultiRoute exposes four labeled routes; MultiSolve asks for the four signed solutions to two absolute-value equations. Both have four valid outcomes plus invalid output. All three panels use rewards $(1, .95, .70, .65, 0)$ and the same analytic targets. For example, the fourth valid outcome has optimal probability 0.0434 under J_3, 0.1052 under J_4, and 0.1942 under U_4. Thus objective selection predicts a change in

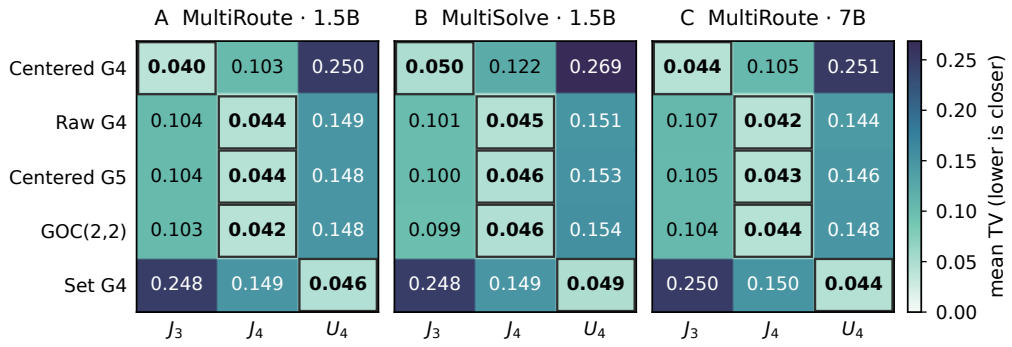


Figure 3: **Five rules select the three predicted target distributions.** Cells are mean per-prompt TV to the maximizers of J_3, J_4, and U_4. All 65 method–seed endpoints are closest to the assigned target; the figure shows means over five seeds in the 1.5B panels and three in the 7B panel.

the allocation among valid answers, not merely in the invalid-output rate. We use unconstrained generation, a shared warm start within each panel, 1,600 updates, and 120 score-bearing rollouts per update; GOC adds 120 forward-only probes. The two 1.5B panels use five paired seeds per method, and the MultiRoute 7B panel uses three (Qwen Team, 2024).

All 65 method–seed endpoints are nearest their assigned optimum among J_3, J_4 and U_4 (Figure 3): 25 endpoints in each 1.5B panel and 15 at 7B. The three J_4 implementations also cluster together: the largest upper endpoint of their pairwise-TV intervals is 0.068 on MultiSolve and 0.063 at 7B, below the distinct-target separation. The three different computations assigned to J_4 agree in target, whereas changing the group rule or the centered budget separates targets.

E3: self-inclusion moves the fixed point. For $F_m(p) = 2^{-m} - p^m$, the intended root is $1/2$. Full-count reuse at $G = m$ instead has root $2^{-m/(m-1)}$; LOO at $G = m + 1$ retains $1/2$. Free A/B generation with five paired seeds for each $m \in \{2, 3, 4\}$ gives full-count means $(.238, .337, .382)$, compared with predictions $(.250, .354, .397)$; LOO means are $(.514, .522, .500)$, compared with $.500$. All six aggregate errors are below the pre-specified 0.03 radius (Figure 4, left; Table 3).

E4: ordinary standardization crosses the reward-level boundary. We directly test Theorem 3 at $K = 3, G = 4$, comparing rewards $(0, 1, 2)$ and $(0, 1, 1)$ on the same outcome space. Both use ordinary realized-reward standardization. Across 64 independent replicates, the conditions share sampled iid counts within each pair. At the largest frozen budget, three-level loop work is 0.0305 [0.0304, 0.0305] versus the exact finite-quadrature target 0.03049; the two-level control is 0.0000 [-0.0001, 0.0000]. Their paired difference is 0.0305 [0.0304, 0.0305]. The three-level mixed-partial estimate is -3.1983 [-3.3293, -3.0690] versus the finite-difference target at $h = .01$, -3.0428. Over five budgets, the four descriptive finite-range RMSE slopes range from -0.52 to -0.49. The samples illustrate recovery; the classification remains analytic.

8 Related work

Finite-sample estimability and policy-gradient objectives. Kimpara et al. (2023) connect fixed-sample estimability with polynomial structure for black-box proper losses. For binary rewards, Davis & Recht (2025) derive induced objective derivatives in a Bernstein basis and construct weights for a prescribed derivative; Zheng (2026) studies finite-rollout surrogate expressivity. Our analysis retains that finite-sample principle and characterizes the finite-outcome field class under our assumptions. The additional issue is objective existence: in more than one simplex dimension, a represented field need not be a gradient. The reward-level classification resolves this issue for ordinary variance normalization.

Approximation and objective loss. Mohri et al. (2026) construct unbiased polynomial answer-level reward estimators and minimax approximations in distributional alignment games. Their

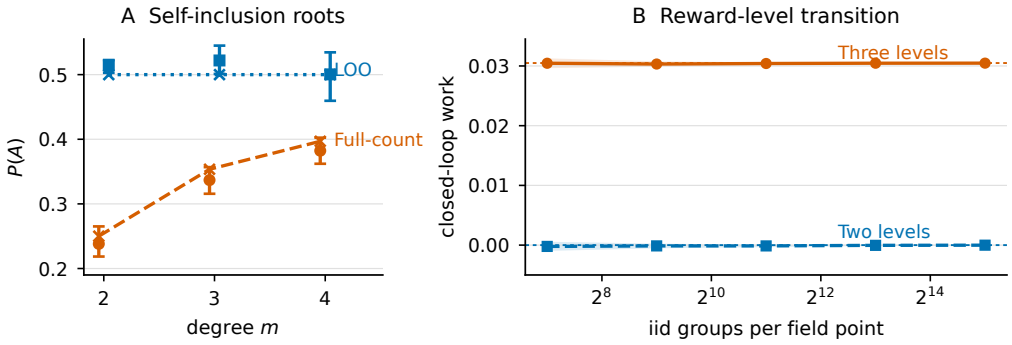


Figure 4: **Self-inclusion roots and the reward-level boundary.** Left: roots and Qwen endpoints (five seeds). Right: loop work (64 paired replicates). Error bars and shading show descriptive 95% bootstrap intervals. At $K = 3$, population work is nonzero with three reward levels and zero with two. Right-hand dashed lines are exact finite-quadrature targets.

$\Theta(G^{-2})$ weighted gradient-bias limit concerns a logarithmic target. Our bound concerns loss in a smooth true objective evaluated at the best surrogate optimum, when two coefficient rules must be reused affinely across reward weights. The restriction and loss differ from a uniform field-approximation error: a different field can share a target's optimum. The lower bound uses alternating optimal responses; the upper bound uses classical polynomial approximation.

Group-relative rewards and normalization. GRPO uses within-group relative rewards (Shao et al., 2024). Zhou et al. (2026) analyze a fixed gradient kernel as a U-statistic; Liu et al. (2026) identify self-inclusion bias and LOO corrections; Mroueh (2026) derives binary-reward losses and dynamics; and Qu et al. (2026) study target projections. The transform in Theorem 4 maps any binary full-count lookup from its marginal to its score-conditioned law. MaxRL v1 (Tajwar et al., 2026) supplies the binary raw-EIPS specialization; Equation 12 identifies the objective change in its success-gated centering implementation. MDP-GRPO studies variance-related pathologies (Salmani-Zarchi et al., 2026), and SoftmaxGRPO gives a three-level non-conservative example for exponential advantages (Hernandez et al., 2026). Our classification concerns ordinary reward standardization, all finite $G \geq 2$, and every positive strictly increasing variance denominator.

Distributional objectives and additional update mechanisms. IPS, ReCo, and UCPO address outcome concentration (Sinha et al., 2026; Park et al., 2026; Lochab et al., 2026); SetPO, Pass@K policy optimization, and RSPO use set- or risk-sensitive signals (Li et al., 2026; Walder & Karkhanis, 2025; Zhang et al., 2025). Table 1 identifies the objectives of the count rules defined here, rather than ranking those methods. Token weighting, optimizer scaling, and credit/rank effects introduce additional mechanisms (Fontana et al., 2026; Mishra et al., 2026) beyond the iid stop-gradient field analyzed in this paper.

9 Discussion and Limitations

Finite groups constrain field representation, objective existence and objective selection. Conditioning makes these constraints explicit, including the reward-level boundary and the loss under affine reuse.

The estimator class assumes finite outcomes, iid on-policy rollouts, outcome-only coefficients and stop-gradient differentiation. Full simplex tangent rank is needed for equivalence with parameter-space gradients; low-rank policies can hide the obstruction. The reuse bound additionally requires positive reward response and affine composition. It provides no estimator-variance or finite-time guarantee. Non-integrability likewise implies neither divergence nor a performance loss.

The Qwen studies cover one model family. E1 compares a selected projection with exact realizations; E4 uses paired $K = 3, G = 4$ samples to illustrate the ordinary-standardization boundary. Both headline classifications are analytic, not inferred from these experiments.

Reproducibility statement

The supplement contains exact-analysis and training code, finite supports, Qwen audits, compact result aggregates, dependency requirements and reproduction commands. Theorems 3 and 6 are proved in Appendices A and B; proofs of the finite-group characterization and objective identities are in Appendix C. Appendix D gives additional constructions, and Appendix E specifies experimental protocols and development-informed calibration. Model revisions and configuration hashes are recorded in the supplementary manifests. All empirical figures and tables use fixed observation sets; E4 uses pre-specified sampling budgets and replicate IDs with paired count streams for its two reward conditions. Deterministic checks of the theoretical results are separate from these experiments and require no model inference or training. Large checkpoints and verbose rollout streams remain external; reported aggregates and non-identifying protocol metadata are included in the anonymous bundle. The supplement's `README.md` maps each experiment and figure to its configuration, results, and reproduction command; E4's budgets and replicate IDs are in `configs/pr39_ordinary_standardization.yaml`.

AI use statement

Generative AI tools were used for writing assistance and literature retrieval. Writing assistance included editing language and organization, formatting references and LaTeX, and refining the presentation of figures and captions. Literature retrieval included formulating search queries, identifying related work, and locating primary sources to check bibliographic details and descriptions of prior work. The authors take full responsibility for the accuracy and originality of the submission, including its references.

Ethics statement

The experiments use synthetic controlled tasks and existing public model checkpoints, with no human subjects or personal data. The main risk is overgeneralizing controlled results. The claims therefore specify the estimator class, reward-normalization convention, reuse restriction and finite experimental setting. The anonymous supplement is supplied under review-only terms; a later public release must preserve the applicable model, data, code and artifact licenses.

References

Damek Davis and Benjamin Recht. What is the objective of reasoning with reinforcement learning? *arXiv preprint arXiv:2510.13651*, 2025. URL https://arxiv.org/abs/2510.13651v1.

Aleksandar Fontana, Marco Simoni, Giulio Rossolini, Paolo Mori, and Andrea Saracino. On the hidden objective biases of group-based reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 109–121. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-short.11. URL https://aclanthology.org/2026.acl-short.11/.

Jefferson Hernandez, Jaywon Koo, Zilin Xiao, Chen Wei, and Vicente Ordonez. SoftmaxGRPO: Learning to reason using softmax advantage group estimation. *arXiv preprint arXiv:2608.09271*, 2026. URL https://arxiv.org/abs/2608.09271v1. Accepted at the Conference on Language Modeling (COLM), 2026.

Dhamma Kimpara, Rafael Frongillo, and Bo Waggoner. Proper losses for discrete generative models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17015–17040. PMLR, 2023. URL https://proceedings.mlr.press/v202/kimpara23a.html.

Chenyi Li, Yuan Zhang, Bo Wang, Guoqing Ma, Wei Tang, Haoyang Huang, and Nan Duan. SetPO: Set-level policy optimization for diversity-preserving LLM reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61114.

Wei Liu, Jiawei Xu, Yingru Li, Longtao Zheng, Tianjian Li, Qian Liu, and Junxian He. Dr. Kernel: Reinforcement learning done right for Triton kernel generations. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65609.

Anamika Lochab, Bolian Li, and Ruqi Zhang. Uniform-correct policy optimization: Breaking RLVR's indifference to diversity. *arXiv preprint arXiv:2605.00365*, 2026. URL https://arxiv.org/abs/2605.00365v1.

Amritansh Mishra, Supriyo Chakraborty, and Berkcan Kapusuzoglu. On the policy gradient foundations of group relative policy optimization: Credit assignment, gradient sparsity, and rank collapse. *arXiv preprint arXiv:2606.29238*, 2026. URL https://arxiv.org/abs/2606.29238v1. Presented at the ICML 2026 Workshop on Foundations of Deep Generative Models (FoGen), a non-archival workshop.

Mehryar Mohri, Jon Schneider, and Yutao Zhong. Generalized distributional alignment games for unbiased answer-level fine-tuning. *arXiv preprint arXiv:2605.02435*, 2026. URL https://arxiv.org/abs/2605.02435v1.

Youssef Mroueh. Reinforcement learning with verifiable rewards: GRPO's loss, dynamics, and success amplification. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/60548.

Junoh Park, Junseo Hwang, Wonguk Cho, and Taesup Kim. ReCo: Reweighting GRPO against distributional concentration. *arXiv preprint arXiv:2607.26862*, 2026. URL https://arxiv.org/abs/2607.26862v1.

Yun Qu, Qi Wang, Yixiu Mao, Heming Zou, Yuhang Jiang, Yingyue Li, Wutong Xu, Lizhou Cai, Weijie Liu, Clive Bai, Kai Yang, Yangkun Chen, Saiyong Yang, and Xiangyang Ji. Listwise policy optimization: Group-based RLVR as target-projection on the LLM response simplex. *arXiv preprint arXiv:2605.06139*, 2026. URL https://arxiv.org/abs/2605.06139v2.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115v1.

Mohammad Mahdi Salmani-Zarchi, Zahra Rahimi, Heshaam Faili, and Mohammad Javad Dousti. MDP-GRPO: Stabilized group relative policy optimization for multi-constraint instruction following. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42784–42797. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1982. URL https://aclanthology.org/2026.acl-long.1982/.

Rikhav Shah, John Urschel, and Nicholas West. What is Jackson's constant? *arXiv preprint arXiv:2607.23375*, 2026. URL https://arxiv.org/abs/2607.23375v1.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300v3.

Abhijeet Sinha, Sundari Elango, and Dianbo Liu. Expected return causes outcome-level mode collapse in reinforcement learning and how to fix it with inverse probability scaling. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63579.

Fahim Tajwar, Guanning Zeng, Yueer Zhou, Yuda Song, Daman Arora, Yiding Jiang, Jeff Schneider, Russ Salakhutdinov, Haiwen Feng, and Andrea Zanette. Maximum likelihood reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65332. Our analysis refers to arXiv:2602.02710v1, available at https://arxiv.org/abs/2602.02710v1.

Christian Walder and Deep Tejas Karkhanis. Pass@K policy optimization: Solving harder reinforcement learning problems. In *Advances in Neural Information Processing Systems*, volume 38, pp. 152416–152445. Curran Associates, Inc., 2025. doi: 10.52202/085713-5096. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/df8a1a6306c3bcc89f6078759feadf62-Abstract-Conference.html.

Kaichen Zhang, Shenghao Gao, Yuzhong Hong, Haipeng Sun, Junwei Bao, Hongfei Jiang, Yang Song, Dingqian Hong, and Hui Xiong. RSPO: Risk-seeking policy optimization for Pass@k and Max@k metrics in large language models. *arXiv preprint arXiv:2508.01174*, 2025. URL https://arxiv.org/abs/2508.01174v1.

Yifu Zheng. RL2ML: Finite-rollout surrogate objectives from reinforcement learning to maximum likelihood. *arXiv preprint arXiv:2605.30154*, 2026. URL https://arxiv.org/abs/2605.30154v1.

Hongyi Zhou, Kai Ye, Erhan Xu, Jin Zhu, Ying Yang, Shijin Gong, and Chengchun Shi. Demystifying group relative policy optimization: Its policy gradient is a U-statistic. *arXiv preprint arXiv:2603.01162*, 2026. URL https://arxiv.org/abs/2603.01162v3.

A REWARD-LEVEL CLASSIFICATION FOR ORDINARY NORMALIZATION

For the rule in Equation 8, write its denominator as D_G to allow a different fixed function at each group size. Each D_G is finite-valued, positive and strictly increasing; no continuity or smoothness of D_G is needed. The coefficients for each realized count cell are constants, so their expectations remain polynomial in p. Write $\alpha = \sum_o q_o \, dp_o$ for the simplex one-form.

Proof of Theorem 3. We first record the finite polynomial representation, then prove each direction.

Finite polynomial extension. For a focal outcome o and peer count $n \in \mathbb{N}^K$ with $|n| = G - 1$, set

$$m_o(n) = \frac{r_o + \sum_\ell n_\ell r_\ell}{G}, \qquad a_o(n) = \frac{r_o - m_o(n)}{D_G(v_o(n))},$$
$$v_o(n) = \frac{(r_o - m_o(n))^2 + \sum_\ell n_\ell (r_\ell - m_o(n))^2}{G}.$$

The coefficients $a_o(n)$ are finite constants independent of p. Conditioning on $O_1 = o$ leaves $G - 1$ iid peers, so

$$q_o(p) = \sum_{|n|=G-1} \binom{G-1}{n} p^n a_o(n). \tag{17}$$

Thus each q_o has a polynomial extension $\bar{q}_o$ to the closed simplex. All evaluations and derivatives at boundary points below concern this extension, not conditional expectations on null events.

Sufficiency: one or two reward levels. If every reward is equal, all realized coefficients vanish. Otherwise suppose the reward values are $\ell < h$. Let $\delta = h - \ell > 0$, $H = \{o : r_o = h\}$, and

$$t = \sum_{o \in H} p_o, \qquad s_m = \frac{\delta^2 m(G-m)}{G^2} \quad (0 \le m \le G).$$

A group with m high-reward observations has variance s_m. Its high-reward coefficient is $\delta(G - m)/(GD_G(s_m))$, and its low-reward coefficient is $-\delta m/(GD_G(s_m))$. All high-reward outcomes consequently share a conditional coefficient $q_H(t)$; all low-reward outcomes share $q_L(t)$. For a high focal outcome, the number of high peers is $B \sim \mathrm{Bin}(G-1, t)$. The all-high case contributes zero, giving

$$q_H(t) = \frac{\delta}{G} \sum_{b=0}^{G-2} \binom{G-1}{b} t^b (1-t)^{G-1-b} \frac{G-1-b}{D_G(s_{b+1})}.$$

For a low focal outcome, the all-low case contributes zero, giving

$$q_L(t) = -\frac{\delta}{G} \sum_{b=1}^{G-1} \binom{G-1}{b} t^b (1-t)^{G-1-b} \frac{b}{D_G(s_b)}.$$

Apply $(G-1-b)\binom{G-1}{b} = (G-1)\binom{G-2}{b}$ to the first sum, and $b\binom{G-1}{b} = (G-1)\binom{G-2}{b-1}$ to the second. Then

$$q_H(t) = (1-t) W_G(t), \qquad q_L(t) = -t W_G(t), \tag{18}$$

where

$$W_G(t) = \frac{\delta(G-1)}{G} \sum_{b=0}^{G-2} \binom{G-2}{b} t^b (1-t)^{G-2-b} \frac{1}{D_G(\delta^2(b+1)(G-1-b)/G^2)}. \tag{19}$$

The sum has strictly positive coefficients and degree at most $G - 2$. Since $\sum_{o \in H} dp_o = dt$ and $\sum_{o \notin H} dp_o = -dt$,

$$\alpha = q_H(t)\, dt - q_L(t)\, dt = W_G(t)\, dt.$$

Therefore

$$J(p) = \int_0^{\sum_{o \in H} p_o} W_G(u)\, du \tag{20}$$

is a potential of degree at most $G - 1$. Notice that this argument permits arbitrarily many outcome labels with repeated rewards.

A vertex derivative identity. For the necessity argument, fix a reference outcome k and use the simplex coordinates $z_u = p_u$ for $u \neq k$, with $p_k = 1 - \sum_{u \neq k} z_u$. If $u \neq k$, expansion along $p = (1-\eta)e_k + \eta e_u$ gives

$$\partial_{z_u} \bar{q}_o|_{e_k} = (G-1)\left[A_o\big(e_o + e_u + (G-2)e_k\big) - A_o\big(e_o + (G-1)e_k\big)\right]. \tag{21}$$

Here $A_o(c)$ denotes the coefficient of a focal member with outcome o in a group with count vector c. To see the identity, the peer configurations with no u have probability $(1-\eta)^{G-1}$, those with exactly one u have probability $(G-1)\eta(1-\eta)^{G-2}$, and the others contribute $O(\eta^2)$. This differentiates a finite polynomial and requires no regularity of D_G.

Necessity: three distinct reward levels. Choose i, k, j with $r_i < r_k < r_j$, and write $a = r_k - r_i > 0$ and $b = r_j - r_k > 0$. Translation by $-r_k$ leaves centered rewards and variances unchanged, so the selected reward differences are $-a, 0, b$. Consider the following three group count vectors:

$$c_a = e_i + (G-1)e_k, \qquad c_b = e_j + (G-1)e_k, \qquad c_{ab} = e_i + e_j + (G-2)e_k.$$

Their respective within-group variances are

$$v_a = \frac{G-1}{G^2}a^2, \qquad v_b = \frac{G-1}{G^2}b^2, \qquad v_{ab} = \frac{(G-1)(a^2+b^2)+2ab}{G^2}. \tag{22}$$

In the reference-k chart, define the mixed-partial residual

$$\mathcal{C}_{ij}^{(k)}(p) = \partial_{z_j}(\bar{q}_i - \bar{q}_k) - \partial_{z_i}(\bar{q}_j - \bar{q}_k).$$

Applying Equation 21 to its four terms and canceling the all-k group yields

$$\frac{\mathcal{C}_{ij}^{(k)}(e_k)}{G-1} = \big(A_i(c_{ab}) - A_j(c_{ab})\big) - \big(A_i(c_a) - A_k(c_a)\big) + \big(A_j(c_b) - A_k(c_b)\big).$$

For two members of the same realized group, the mean cancels in their coefficient difference. Thus

$$\mathcal{C}_{ij}^{(k)}(e_k) = (G-1)\left[\frac{a}{D_G(v_a)} + \frac{b}{D_G(v_b)} - \frac{a+b}{D_G(v_{ab})}\right]. \tag{23}$$

The mixed group has strictly larger variance than either single-excursion group:

$$v_{ab} - v_a = \frac{(G-1)b^2 + 2ab}{G^2} > 0, \qquad v_{ab} - v_b = \frac{(G-1)a^2 + 2ab}{G^2} > 0.$$

Strict monotonicity and positivity of D_G therefore give

$$\mathcal{C}_{ij}^{(k)}(e_k) = (G-1)\left[a\left(\frac{1}{D_G(v_a)} - \frac{1}{D_G(v_{ab})}\right) + b\left(\frac{1}{D_G(v_b)} - \frac{1}{D_G(v_{ab})}\right)\right] > 0.$$

If $dJ = \alpha$ throughout the simplex interior, its chart mixed partials would agree there. The residual, being a polynomial, would be zero throughout the interior and, by continuity, at e_k, a contradiction. Equivalently, the positive boundary value gives nonzero residuals at sufficiently nearby full-support interior points. This also covers $K > 3$: the selected three outcomes determine the displayed vertex derivative, and the other outcomes can all be assigned arbitrarily small positive mass.

Finally, the same residual is a nonzero polynomial in $K-1$ chart variables. It cannot vanish on a nonempty open set. A potential on such a set would require it to vanish there, proving the stronger local statement. □

Normalization conventions. The theorem applies to $D_G(v) = \sqrt{v+\epsilon}$ and $D_G(v) = \sqrt{v} + \epsilon$ for every fixed $\epsilon > 0$. Using the unbiased sample variance instead of divisor G replaces v by $Gv/(G-1)$ inside D_G and preserves strict monotonicity. It also applies to the unstabilized rule $D_G(v) = \sqrt{v}$ provided every zero-variance group is assigned $A_i = 0$ for every member i. More generally, the same proof works for any positive, finite, strictly increasing D_G on $(0, \infty)$ with this zero-group convention: all three variances in Equation 23 and all variances used in Equation 19 are strictly positive.

Scope and necessity of the assumptions. The restriction $G \geq 2$ is necessary; with $G = 1$, stabilized centering always gives zero. Strictly increasing normalization is also substantive: for a constant denominator $d_0 > 0$, any reward vector satisfies

$$\alpha = \frac{G-1}{Gd_0}\, d\left(\sum_o p_o r_o\right),$$

so three or more reward levels can be conservative without variance-dependent normalization. The classification concerns the iid, stop-gradient, sequence-score population field for a fixed prompt. Under the score identity, $\mathbb{E}[\widehat{g}] = \sum_o q_o \nabla_\theta p_o$, a full-simplex-rank policy map transfers the obstruction to the parameter-space field. Rank-deficient restrictions may hide it. Non-existence of this potential does not imply non-convergence, cyclic behavior, or lack of reward improvement, and does not rule out different metrics or distribution-dependent rescalings.

B Optimum-loss bounds for reward-affine reuse

We use the interval $I = [a, b] = [1/4, 3/4]$, center $c = 1/2$, half-width $H = 1/4$, and weight set $\Lambda = [-3/8, 3/8]$ from Section 6. The fixed target class is Equation 14 and the risk is Equation 15. The kernels may depend on the whole base objective J, but the same two kernels serve every λ:

$$A_i^{(\lambda)} = A_i^{(0)} + \lambda A_i^{(1)}. \tag{24}$$

The base objective is distributional; only the added term λp is linear expected reward. No uniform bound on derivatives above order two is imposed on the target class.

Lemma 1 (Exact polynomial model for an affine-reused primitive)**.** *The expected update has the form*

$$\mathbb{E}[\widehat{g}_\lambda] = q_\lambda(p)\nabla_\theta p, \qquad q_\lambda(p) = u(p) + \lambda v(p), \qquad u, v \in \mathcal{P}_{G-1}, \tag{25}$$

where $\mathcal{P}_d$ is the space of real polynomials of degree at most d. Conversely every pair $u, v \in \mathcal{P}_{G-1}$ is realized by kernels satisfying (24)*.*

Proof. For a component kernel and a focal outcome $o \in \{0, 1\}$, conditioning leaves $G-1$ Bernoulli peers. Averaging over auxiliary randomness and ordered configurations with the same peer count gives constants $a_{o,k}$ independent of p. The conditional coefficient is

$$q_o(p) = \sum_{k=0}^{G-1} a_{o,k} \binom{G-1}{k} p^k (1-p)^{G-1-k}.$$

The Bernoulli score identity gives the quotient coefficient $q_1 - q_0$. Apply this argument separately to $A^{(0)}$ and $A^{(1)}$. For the converse write

$$u(p) = \sum_{k=0}^{G-1} b_k \binom{G-1}{k} p^k (1-p)^{G-1-k}, \quad v(p) = \sum_{k=0}^{G-1} d_k \binom{G-1}{k} p^k (1-p)^{G-1-k}.$$

With $N_i = \sum_{j \neq i} O_j$, set $A_i^{(\lambda)} = O_i(b_{N_i} + \lambda d_{N_i})$. Then the failure coefficient is zero and the success coefficient is exactly $u + \lambda v$. These are finite lookups and are permutation-equivariant. □

Write $\mathcal{A}_G = \{(u, v) : u, v \in \mathcal{P}_{G-1},\ v > 0 \text{ on } I\}$. The positivity condition $v > 0$ is substantive. Without it, the zero field would make every point a surrogate optimum and defeat the best-tie loss. Likewise, permitting separate redesign at each weight would allow the field $x_J(\lambda) - p$, which realizes each prescribed optimum with two rollouts. The lower bound instead quantifies over a single pair of kernels before the weight is selected.

B.1 Lower bound: alternating optimal responses

Fix G. Set

$$h = \frac{b-a}{G+3} = \frac{1}{2(G+3)}, \qquad \eta = \frac{h}{2\pi}, \qquad \omega = \frac{\pi}{h}, \qquad x_j = a + (j+1)h \quad (0 \leq j \leq G+1).$$

Let $\kappa = \cos(\omega(c - x_0))$ and define

$$f_G(p) = c - p + \eta[\cos(\omega(p - x_0)) - \kappa], \qquad J_G^{\mathrm{hard}}(p) = \int_c^p f_G(t)\, dt. \tag{26}$$

The superscript distinguishes this constructed objective from the manuscript's occupancy-objective notation J_G. Since $f_G(c) = 0$ and

$$f_G'(p) = -1 - \tfrac{1}{2}\sin(\omega(p - x_0)) \in [-3/2, -1/2],$$

we have $J_G^{\mathrm{hard}} \in \mathcal{J}$ for every G. At the selected points,

$$f_G(x_j) = c - x_j + \eta[(-1)^j - \kappa].$$

Choose $\lambda_j = -f_G(x_j)$. The derivative bound and $f_G(c) = 0$ give $|\lambda_j| \le (3/2)|x_j - c| < 3/8$, so all tilts are in Λ. The unique maximizer of $J_G^{\mathrm{hard}}(p) + \lambda_j p$ is x_j.

Now take an arbitrary $(u, v) \in \mathcal{A}_G$. Suppose for contradiction that for every j there exists a surrogate global maximizer y_j satisfying $|y_j - x_j| < \eta$. The x_j are separated by h and have boundary margin at least h. Because $2\eta = h/\pi < h$, the points y_j are interior and strictly increasing. Their first-order necessary conditions are

$$u(y_j) + \lambda_j v(y_j) = 0.$$

Consider the polynomial

$$P(p) = u(p) + (p - c + \eta\kappa)v(p), \qquad \deg P \le G. \tag{27}$$

Substitution of $\lambda_j = -f_G(x_j)$ yields

$$P(y_j) = [y_j - x_j + \eta(-1)^j]v(y_j). \tag{28}$$

The values have strictly alternating signs, because $v(y_j) > 0$ and $|y_j - x_j| < \eta$. In particular P is not the zero polynomial. There are $G + 2$ ordered points, so the intermediate value theorem forces at least $G + 1$ distinct roots, contrary to $\deg P \le G$.

Consequently some j obeys

$$\operatorname{dist}(x_j, M_{u,v}(\lambda_j)) \ge \eta.$$

The true tilted objective is $1/2$-strongly concave and has interior stationary point x_j. Therefore for *every* $y \in M_{u,v}(\lambda_j)$,

$$(J_G^{\mathrm{hard}} + \lambda_j \,\mathrm{id})(x_j) - (J_G^{\mathrm{hard}} + \lambda_j \,\mathrm{id})(y) \ge \frac{1}{4}|y - x_j|^2 \ge \frac{\eta^2}{4} = \frac{1}{64\pi^2(G+3)^2}.$$

This is uniform in (u, v), proving the lower bound after the infimum and supremum in (15).

Remark 1 (An exact endpoint benchmark for the finite hard family)**.** *For the $G + 2$ selected tilts, choose $u(p) = c - p - \eta\kappa$ and $v = 1$. Its unique surrogate optimum is $y_j = x_j - \eta(-1)^j$, an interior point at distance exactly η. Thus the minimax maximum* optimum distance *on this particular finite family is exactly η. Direct integration of* (26) *gives the same true loss at each selected tilt:*

$$(J_G^{\mathrm{hard}} + \lambda_j \,\mathrm{id})(x_j) - (J_G^{\mathrm{hard}} + \lambda_j \,\mathrm{id})(y_j) = \eta^2\big(2\sin(1/2) - 1/2\big).$$

This is an explicit feasible comparison, not a claim that this pair is the exact minimizer of the objective-loss risk.

B.2 Matching upper bound

Lemma 2 (Uniform field error controls every surrogate optimum)**.** *Let J be μ-strongly concave and continuously differentiable on a compact interval I. Let q be continuous with $\|q - J'\|_{\infty,I} \le \varepsilon$, and put $Q(p) = \int_c^p q(t)\, dt$. For any real λ, let $x = \arg\max_I[J(p) + \lambda p]$ and take any $y \in \arg\max_I[Q(p) + \lambda p]$. Then*

$$|y - x| \le \varepsilon/\mu, \qquad 0 \le J_\lambda(x) - J_\lambda(y) \le \frac{\varepsilon^2}{2\mu}. \tag{29}$$

No concavity of Q or interiority of either optimizer is needed.

Proof. First-order necessary optimality of y on the interval gives $(q(y)+\lambda)(x-y) \le 0$. Strong concavity at y implies

$$\begin{aligned} J_\lambda(x) - J_\lambda(y) &\le (J'(y)+\lambda)(x-y) - \frac{\mu}{2}|x-y|^2 \\ &\le \varepsilon|x-y| - \frac{\mu}{2}|x-y|^2 \le \frac{\varepsilon^2}{2\mu}. \end{aligned}$$

For the distance statement write $d = y - x$ and use strong monotonicity of $-J'$, together with target and surrogate first-order conditions:

$$\mu d^2 \le [J'(x) - J'(y)]d \le -[J'(y)+\lambda]d \le [q(y) - J'(y)]d \le \varepsilon|d|.$$

Here $(J'(x)+\lambda)d \le 0$ and $(q(y)+\lambda)d \ge 0$ remain valid at constrained boundaries. Divide by $|d|$ if $d \neq 0$. □

We give a self-contained bound using the classical Jackson-kernel construction; see Shah et al. (2026) for refined constants.

Lemma 3 (An explicit polynomial approximation bound)**.** *If f is L-Lipschitz on $[c-H, c+H]$ and $G \ge 2$, there is a polynomial $q \in \mathcal{P}_{G-1}$ with*

$$\|q - f\|_\infty \le \frac{3\pi LH}{G}. \tag{30}$$

Proof. Set $F(\theta) = f(c + H\cos\theta)$. This function is even, 2π-periodic, and LH-Lipschitz in circular distance. For an integer $m \ge 1$ let

$$D_m(t) = \frac{\sin(mt/2)}{\sin(t/2)}, \quad Z_m = \frac{2m^3+m}{3}, \quad K_m(t) = \frac{D_m(t)^4}{Z_m},$$

with removable values supplied by continuity. The finite geometric-sum identity gives

$$D_m(t)^2 = \sum_{k=-(m-1)}^{m-1} (m-|k|)e^{ikt}.$$

Thus D_m^4 is an even trigonometric polynomial of degree $2m-2$, whose constant Fourier coefficient is

$$m^2 + 2\sum_{k=1}^{m-1} k^2 = Z_m.$$

It follows that $K_m \ge 0$ and $(2\pi)^{-1}\int_{-\pi}^{\pi} K_m(t)\,dt = 1$. For $|t| \le \pi$, $|D_m(t)| \le \min(m, \pi/|t|)$. Splitting at π/m gives

$$\begin{aligned} \frac{1}{2\pi}\int_{-\pi}^{\pi} |t|K_m(t)\,dt &\le \frac{1}{\pi Z_m}\left[\int_0^{\pi/m} tm^4\,dt + \int_{\pi/m}^{\pi} \frac{\pi^4}{t^3}\,dt\right] \\ &= \frac{\pi(m^2 - 1/2)}{Z_m} \le \frac{3\pi}{2m}. \end{aligned}$$

The convolution

$$S_m F(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} K_m(t)F(\theta - t)\,dt$$

is even and has trigonometric degree at most $2m-2$. Consequently it is an algebraic polynomial $q(c + H\cos\theta)$ of the same degree, using $\cos(k\theta) = T_k(\cos\theta)$. Its uniform error is at most $3\pi LH/(2m)$ by the Lipschitz condition and the moment bound. Choose $m = \lfloor (G+1)/2 \rfloor$. Then $2m - 2 \le G - 1$ and $2m \ge G$, proving (30). □

Completion of Theorem 6. For $J \in \mathcal{J}$, take $f = J'$, $L = 3/2$, and $H = 1/4$ in Lemma 3. The resulting polynomial satisfies

$$\|q - J'\|_{\infty,I} \le \frac{9\pi}{8G}.$$

Choose $u = q$, $v = 1$, and realize these polynomials by Lemma 1. Lemma 2 with $\mu = 1/2$ then gives, simultaneously for every $\lambda \in \Lambda$ and every $y \in M_{q,1}(\lambda)$,

$$J_\lambda(x_J(\lambda)) - J_\lambda(y) \le \frac{81\pi^2}{64G^2}.$$

Taking the infimum and supremum proves the upper bound in (16). □

B.3 An exact polynomial threshold under exact reward addition

For this section impose *exact additive reward composition*: $q_\lambda(p) = q(p) + \lambda$. The surrogate is $Q(p) + \lambda p$ with $Q' = q \in \mathcal{P}_{G-1}$. This is the subclass $v \equiv 1$.

Theorem 7 (Polynomial target with positive optimum loss below exact budget)**.** *For every $G \ge 2$ there is a degree-$(G+1)$ polynomial $J_G^{\mathrm{poly}} \in \mathcal{J}$ and a fixed set $\Lambda_G \subset \Lambda$ of $G+1$ tilts such that*

$$\frac{7}{65536G^4} \le \inf_{q\in\mathcal{P}_{G-1}} \max_{\lambda\in\Lambda_G} \min_{y\in\arg\max_I [Q(p)+\lambda p]} \left[J_{G,\lambda}^{\mathrm{poly}}(x_G(\lambda)) - J_{G,\lambda}^{\mathrm{poly}}(y)\right] \le \frac{9}{16384G^4}, \quad (31)$$

where $Q' = q$ and $x_G(\lambda)$ is the unique true optimum. At budget $G+1$, a single peer-count base kernel represents $(J_G^{\mathrm{poly}})'$ exactly; adding λO_i realizes every tilted field and gives zero population optimum loss. The positive lower bound in (31) *is uniform over all base kernels with G rollouts, provided the same base kernel is reused and the tilt is added exactly.*

Proof. Let T_G be the degree-G Chebyshev polynomial and set

$$\eta_G = \frac{1}{32G^2}, \qquad f_G^{\mathrm{poly}}(p) = c - p - \eta_G[T_G(4p-2) - T_G(0)], \qquad J_G^{\mathrm{poly}}(p) = \int_c^p f_G^{\mathrm{poly}}(t)\,dt.$$

The identities $T_G(\cos\theta) = \cos(G\theta)$ and $|\sin(G\theta)| \le G|\sin\theta|$ imply $\|T_G'\|_{\infty,[-1,1]} \le G^2$, including endpoint limits. Therefore

$$7/8 \le -(f_G^{\mathrm{poly}})' \le 9/8, \qquad f_G^{\mathrm{poly}}(c) = 0, \qquad \deg f_G^{\mathrm{poly}} = G.$$

In particular $J_G^{\mathrm{poly}} \in \mathcal{J}$.

Put $\rho_G = 1 - 1/(4G^2)$ and choose contracted Chebyshev extrema

$$x_j = c + \frac{\rho_G}{4}\cos(j\pi/G), \quad \lambda_j = -f_G^{\mathrm{poly}}(x_j), \quad 0 \le j \le G.$$

These x_j are strictly decreasing interior points. By the derivative bound,

$$(-1)^j T_G(4x_j - 2) \ge \frac{3}{4}. \quad (32)$$

Indeed contraction changes the Chebyshev argument by at most $1/(4G^2)$, so it changes the polynomial value by at most $1/4$. The boundary margin of the extreme points is $(1-\rho_G)/4 = 2\eta_G$. The minimum adjacent spacing is at least

$$\frac{\rho_G}{4}(1 - \cos(\pi/G)) \ge \frac{\rho_G}{2G^2} \ge 15\eta_G.$$

Here $\rho_G \ge 15/16$ and $1 - \cos(\pi/G) = 2\sin^2(\pi/(2G)) \ge 2/G^2$. Also $|\lambda_j| \le (9/8)|x_j - c| < 9/32 < 3/8$, so $\Lambda_G = \{\lambda_0, \dots, \lambda_G\} \subset \Lambda$.

Suppose $q \in \mathcal{P}_{G-1}$ had a surrogate maximizer y_j within $\eta_G/2$ of each x_j. They would be interior and remain strictly ordered. Stationarity gives $q(y_j) = f_G^{\mathrm{poly}}(x_j)$. The polynomial

$$P(p) = q(p) + (p - c) - \eta_G T_G(0), \qquad \deg P \le G - 1,$$

therefore satisfies

$$P(y_j) = y_j - x_j - \eta_G T_G(4x_j - 2).$$

By (32) these $G+1$ values have strictly alternating signs, giving at least G distinct roots, an impossibility. Thus some tilt has optimum distance at least $\eta_G/2$ for every surrogate maximizer. The true objective is $7/8$-strongly concave, so the loss is at least

$$\frac{7}{16}(\eta_G/2)^2 = \frac{7}{65536G^4}.$$

For the feasible upper bound, choose the linear base field $q_0(p) = c - p + \eta_G T_G(0)$. At tilt λ_j its unique optimum is $y_j = x_j + \eta_G T_G(4x_j - 2)$, within η_G of x_j and still interior. The true curvature is at most $9/8$, giving loss at most $(9/16)\eta_G^2 = 9/(16384G^4)$. Finally f_G^{poly} has degree exactly G, so Lemma 1 realizes it with $G+1$ rollouts. Exact addition of λp preserves the realization for all tilts. □

C Proofs for the finite-group characterization and objective map

C.1 Notation and score identities

For fixed x, the outcome probability is $p_o = \sum_{y:T(y)=o} \pi_\theta(y \mid x)$, with the sum replaced by the corresponding integral when needed. Under Section 2's finite-support or dominated score-regularity assumption, differentiation on Δ_K° gives $\nabla p_o = p_o \mathbb{E}[\nabla \log \pi_\theta(Y \mid x) \mid O = o] = p_o \mu_o$. Consequently $\sum_o p_o \mu_o = \nabla \sum_o p_o = 0$. For Equation 1, exchangeability makes every summand have the same expectation. Conditional on O_i, the sampled score is independent of the peers and auxiliary randomness, and A_i uses Y_i only through O_i. Hence $\mathbb{E}[A_i s_i \mid O_i] = \mathbb{E}[A_i \mid O_i]\mathbb{E}[s_i \mid O_i]$, which gives Equation 3. The estimator itself uses sampled scores, not their conditional means.

C.2 Conditional representation

Fix member i and define $a_{o,n}$ by Equation 4. This structural definition is meaningful for every count cell without conditioning on a possibly null event. For $p \in \Delta_K^\circ$, conditional on $O_i = o$, the remaining $G-1$ outcomes are independent with law p. Their count $N = C - e_o$ has probability

$$\Pr(N = n) = \binom{G-1}{n} \prod_j p_j^{n_j}, \qquad |n| = G - 1.$$

Ordered configurations with the same peer count have equal conditional weight. Averaging their coefficients and the independent U therefore gives the constant $a_{o,n}$. Total expectation proves $q_o^{\mathrm{cond}} = \bar{q}_o$ on the interior, with $\bar{q}_o$ given by Equation 5. Finite K gives finitely many count cells; thus the expression is a polynomial of degree at most $G-1$ and defines a continuous extension to the closed simplex. Integrability of $A_i s_i$ justifies the score-conditioning step, yielding

$$\mathbb{E}[A_i s_i] = \sum_o p_o q_o^{\mathrm{cond}} \mu_o = \sum_o \bar{q}_o \nabla p_o.$$

No conditional expectation is assigned to a zero-probability outcome.

C.3 Bernstein/count-lookup converse

The monomials p^α with $|\alpha| \le G-1$, restricted to $\sum_j p_j = 1$, span the polynomials of degree at most $G-1$ on the simplex. For a degree-d monomial, multiplying by $(\sum_j p_j)^{G-1-d} = 1$ and expanding degree-elevates it into the multinomial Bernstein basis of degree $G-1$. Each target component consequently has coefficients $a_{o,n}$. Set

$$A_i(O_{1:G}) = a_{O_i, C - e_{O_i}}.$$

This finite lookup is permutation-equivariant, has conditional mean q_o on the interior, and has polynomial extension q_o on the boundary. Its coefficients are finite, so the score moment assumption in Section 2 gives the required product integrability. This proves the converse in Theorem 1.

C.4 QUOTIENT, INTEGRABILITY, AND THE HIERARCHY

In coordinates $z_a = p_a$ for $a < K$,

$$\sum_o q_o \, dp_o = \sum_{a<K} (q_a - q_K) \, dz_a.$$

The annihilator of the full simplex tangent is exactly the span of $\mathbf{1}$, giving coefficient identification modulo $b(p)\mathbf{1}$ at full tangent rank. A C^2 potential has equal mixed partials, which proves the necessary direction of Theorem 2. Conversely, the simplex interior is convex. When the one-form is closed, the integral

$$J(z) = J(z^0) + \int_0^1 \langle \omega(z^0 + t(z - z^0)), z - z^0 \rangle \, dt$$

is path-independent and differentiates to ω. For polynomial ω, integrating term by term raises degree by at most one. At $K = 2$ the domain is an interval and every such one-form has a primitive. Rank-deficient policy maps can have a larger annihilator.

For reference, define the potential class, modulo additive constants,

$$\mathfrak{P}_G = \{J : [dJ] \in \mathfrak{F}_G\}.$$

It is the conservative part of the represented-field class: its members have polynomial representatives of degree at most G on the simplex. The inclusions $\mathfrak{F}_G \subseteq \mathfrak{F}_{G+1}$ follow from polynomial-degree containment. For strictness on any $K \geq 2$, take the one-form $p_1^G \, dp_1$. Restricting to an interior line with fixed $p_3, \ldots, p_K$ shows that it cannot have a degree-$(G-1)$ representative; its binary quotient would otherwise be a lower-degree polynomial equal to p_1^G on an open interval. The converse realizes it at $G + 1$. This also proves $G_{\min}([q]) = 1 + \deg_\Delta[q]$. For inverse design, Theorem 1 supplies the necessary polynomial representative and Theorem 2 identifies its quotient with dJ. Conversely, expanding any such representative in the Bernstein basis gives the required lookup, proving Corollary 2.

C.5 SELF-INCLUSION TRANSFORM

Write $F_G[a] = \sum_{c=0}^{G} a_c \binom{G}{c} p^c (1-p)^{G-c}$. Conditional on a score-bearing success, $C = 1 + \mathrm{Bin}(G-1, p)$, so $S_G[a] = B_{G-1}[a_1, \ldots, a_G]$. The Bernstein derivative identity gives

$$F'_G[a] = G B_{G-1}[a_1 - a_0, \ldots, a_G - a_{G-1}].$$

Collecting adjacent coefficients gives $S_G = F_G + (1-p)F'_G/G$, as in Theorem 4. Endpoint values follow from the polynomial extensions. An exact degree-m marginal polynomial requires at least m samples, whereas a peer polynomial of that degree requires $G - 1 \geq m$. A leave-one-out lookup evaluates this peer statistic directly.

C.6 FINITE OCCUPANCY OBJECTIVES

All conditional identities below are on Δ_K°; boundary values mean polynomial extensions. With raw EIPS, the total coefficient on an observed outcome is r_o times its presence indicator. The expected update is $\sum_o r_o h_G(p_o)\mu_o = \nabla J_G$. For same-group centering, the mean is $\bar{R} = \sum_k r_k \mathbf{1}\{C_k > 0\}$. Set $H(p) = \sum_k r_k h_{G-1}(p_k)$. Then

$$\mathbb{E}[\bar{R} \mid O_i = o] = H(p) + r_o(1 - p_o)^{G-1}.$$

The first term is common across outcomes and contributes zero expected score. Centering therefore subtracts $\sum_o r_o p_o (1-p_o)^{G-1} \mu_o$ from the raw field. The identity

$$h_G(p) - p(1-p)^{G-1} = h_{G-1}(p)$$

proves the shift to J_{G-1}. For the set-marginal rule, a scored outcome o gets coefficient Gr_o only when none of the peers has that outcome. Thus $q_o = Gr_o(1-p_o)^{G-1} = \partial_{p_o} U_G$. These establish the first three identities in Theorem 5.

Table 2: Represented fields and rollout budgets. An implementation budget need not equal the minimum budget of its represented field.

Objective or field	Quotient degree	Realization
Genuine degree-d field	d	minimum $G = d + 1$
Raw EIPS: J_G	at most $G - 1$	group G
Centered EIPS: J_{G-1}	at most $G - 2$	group G
Set-marginal: U_G	at most $G - 1$	group G
GOC(G, L): J_{GL}	at most $GL - 1$	GL observations
Reward-proportional log	nonpolynomial	no fixed finite budget

C.7 Clipping and geometric occupancy correction

For occupancy clip $\epsilon > 0$, let $d_n = 1/\max((n+1)/G, \epsilon)$. Conditioning gives

$$q_{o,G,\epsilon} = r_o B_{G-1}[d_0, \ldots, d_{G-1}](p_o).$$

Let $\Psi_{G,\epsilon}$ be the zero-at-origin antiderivative of this Bernstein multiplier without r_o. The represented objective is $\sum_o r_o \Psi_{G,\epsilon}(p_o)$. Its boundary marginal is $\Psi'_{G,\epsilon}(0) = \min(G, 1/\epsilon)$, but its full polynomial generally differs from simply replacing G by that boundary value.

For finite GOC, observe L independent on-policy groups of size G: one score-bearing main group and $L - 1$ forward-only probe groups $P_1, \ldots, P_{L-1}$. With $a_o = (1 - p_o)^G$, define

$$T_{o,L} = \sum_{t=0}^{L-1} \prod_{\ell=1}^{t} \mathbf{1}\{o \notin P_\ell\}.$$

The empty product is one. Independence implies

$$\mathbb{E}T_{o,L} = \sum_{t=0}^{L-1} a_o^t = \frac{1 - a_o^L}{1 - a_o} = \frac{h_{GL}(p_o)}{h_G(p_o)}.$$

Multiplication of the main-group presence contribution by $T_{o,L}$ yields $r_o h_{GL}(p_o)\mu_o$, and hence ∇J_{GL}. Probes enter only the coefficient; their scores do not enter backpropagation. This expanded oracle has GL observations rather than the primitive's G observations. Shared and independent probes give the same mean but different second moments. For unbounded probing, $\mathbb{E}T_o = 1/h_G(p_o)$ and the mean becomes $\sum_o r_o \mu_o = \nabla_\theta \sum_o r_o \log p_o$. Appendix D.2 supplies explicit sufficient cost and moment conditions for this limit.

C.8 Fixed-budget log-matching impossibility

Proposition 1 (Fixed-budget log obstruction)**.** *Let $K \geq 2$ and $r \in \mathbb{R}^K \setminus \{0\}$. No fixed finite-$G$ estimator in Equation 1 represents $dJ_{\log}$ throughout Δ_K°, where $J_{\log}(p) = \sum_o r_o \log p_o$. Equivalently, $J_{\log} \notin \mathfrak{P}_G$ for every finite G. On a truncated domain $p_o \geq p_{\min} > 0$, polynomial approximation is possible. At full tangent rank the obstruction also applies to $\nabla_\theta J_{\log}$.*

Proof. Choose $i \neq j$ with $r_i \neq 0$ and fix the other probabilities so that $p_i + p_j = \rho > 0$. Parameterize the interior line by $p_i = t, p_j = \rho - t$. The required quotient is $r_i/t - r_j/(\rho - t)$. If a finite-group rule represented it, Theorem 1 would give a polynomial $P(t)$ satisfying

$$t(\rho - t)P(t) = r_i(\rho - t) - r_j t \qquad (0 < t < \rho).$$

This polynomial identity extends to $t = 0$, implying $0 = r_i \rho$, a contradiction. It allows signed rewards and does not treat the boundary as interior. Uniform approximation on the full interior is impossible because the target is unbounded while fixed polynomials extend continuously to the compact simplex. On a compact truncated domain, polynomial approximation applies to the continuous target field. □

C.9 THE EIPS-STANDARDIZED CIRCULATION WITNESS

The secondary circulation witness standardizes inverse-count coefficients, rather than ordinary rewards:

$$R_i = \frac{G r_{O_i}}{C_{O_i}}, \qquad \bar{R} = \frac{1}{G}\sum_j R_j, \qquad V_G = \frac{1}{G}\sum_j (R_j - \bar{R})^2, \\ A_i = \frac{R_i - \bar{R}}{\sqrt{V_G + 10^{-6}}}. \tag{33}$$

All quantities are stop-gradient and V_G is the realized within-group second central moment with divisor G.

Proposition 2 (EIPS-standardized witness)**.** *For Equation 33 with $K = 3$, $G = 4$, and $r = (.7, 1.2, 1.9)$, the mixed-partial residual $\partial_{p_2}(q_1 - q_3) - \partial_{p_1}(q_2 - q_3)$ at $p = (.2, .3, .5)$ lies in $(-1.473583, -1.473582)$. The field has no scalar simplex potential.*

Exact finite-count enumeration gives this algebraic residual; the symbolic certificate in the supplement provides the radical expression and 60-digit value. The sign convention is the negative of the usual exterior-derivative coefficient. Thus counterclockwise work integrates its negative. On the rectangle with vertices $(.15, .25), (.25, .25), (.25, .35), (.15, .35)$, a 4,096-panel-per-edge trapezoidal integration gives work 0.0149346922. Using the sampled path's 32 panels changes the exact-field quadrature by 9.71×10^{-8}. The nonzero algebraic residual proves non-exactness; the quadrature is a corroborating diagnostic. Raw and centered controls have absolute residual and work below 2.3×10^{-15} in the same float64 audit. These numbers apply to this EIPS rule and stabilizer, not to the ordinary-reward rule of Theorem 3.

D ADDITIONAL DERIVATIONS AND THE EXPLICIT E1 WITNESS

D.1 THE EFFECT OF SUCCESS-GATED SCORE CENTERING

Fix one prompt with success probability $p \in (0, 1)$, binary indicators b_i, count $C_+ = \sum_i b_i$, and trajectory scores s_i. Write

$$\widehat{g}_{\mathrm{raw}} = \begin{cases} C_+^{-1} \sum_i b_i s_i, & C_+ > 0, \\ 0, & C_+ = 0, \end{cases} \qquad \bar{s} = G^{-1} \sum_i s_i.$$

The raw estimator has mean $[h_G(p)/p]\nabla_\theta p$. The on-policy centering in MaxRL v1 (Tajwar et al., 2026, Section 4.3, Algorithm 1) instead implements

$$\widehat{g}_{\mathrm{gate}} = \widehat{g}_{\mathrm{raw}} - \mathbf{1}\{C_+ > 0\}\bar{s}.$$

The score identity for the joint iid group gives

$$\mathbb{E}\left[\mathbf{1}\{C_+ > 0\} \sum_i s_i\right] = \nabla_\theta \Pr(C_+ > 0) = \nabla_\theta h_G(p).$$

Consequently, for $G \geq 2$,

$$\mathbb{E}[\widehat{g}_{\mathrm{gate}}] = \left[\frac{h_G(p)}{p} - (1-p)^{G-1}\right] \nabla_\theta p = \frac{h_{G-1}(p)}{p} \nabla_\theta p.$$

This is the binary J_{G-1} field in Theorem 5. Unconditional subtraction of $\bar{s}$ is field-neutral; multiplying it by the same-group success indicator is not. At $G = 2$, the raw and gated coefficients are respectively $2 - p$ and 1, giving the introductory example. Appendix C.6 derives the corresponding transformation for multiple outcomes.

D.2 A SUFFICIENT MOMENT AND COST CONDITION FOR UNBOUNDED GOC

The limit in Appendix C.7 can be implemented with a random stopping time. Fix a prompt and an interior policy, finite K, and $\mathbb{E}\|s_i\| < \infty$. For an independent stream of probe groups, let T_o be the index of the first probe containing outcome o. Then

$$\Pr(T_o > t) = (1 - h_G(p_o))^t, \qquad \mathbb{E}T_o = \frac{1}{h_G(p_o)} < \infty.$$

The finite multiplier is $T_{o,L} = \min(T_o, L)$. Stopping after all required outcomes have appeared uses at most $\max_o T_o$ probe groups, with $\mathbb{E}\max_o T_o \le \sum_o 1/h_G(p_o) < \infty$. This establishes finite expected rollout cost at the fixed interior policy.

Let $\bar{s}_o = C_o^{-1}\sum_{i:O_i=o} s_i$ when $C_o > 0$ and zero otherwise, and set $M_o = \mathbb{E}[\mathbf{1}\{C_o > 0\}\|\bar{s}_o\|]$. The score moment assumption gives $M_o < \infty$. For every L, the estimator norm is bounded by $D = \sum_o |r_o| T_o \mathbf{1}\{C_o > 0\}\|\bar{s}_o\|$. Independence of the probes from the main group gives $\mathbb{E}D = \sum_o |r_o| M_o/h_G(p_o) < \infty$. Dominated convergence therefore permits $T_{o,L} \uparrow T_o$ inside the estimator expectation, yielding

$$\sum_o r_o h_G(p_o)\mathbb{E}[T_o]\mu_o = \sum_o r_o \mu_o = \nabla_\theta \sum_o r_o \log p_o.$$

These are sufficient pointwise conditions. Averaging over prompts or taking limits along changing policies additionally requires an integrable envelope for the cost and score-weighted terms. The fixed-policy cost bound is not uniform as an outcome probability approaches zero.

D.3 The frozen degree-three field used in E1

Let outcome 5 be invalid, set $a = (.40, .35, .15, .10)$ and $\kappa = (0, 0, 20, 20)$, and define

$$J_\star(p) = -\frac{1}{4}\left[\sum_{o=1}^{4}\left\{\frac{(p_o - a_o)^2}{2} + \frac{\kappa_o(p_o - a_o)^4}{4}\right\} + p_5\right]. \tag{34}$$

Its coefficients are $q_o = -\frac{1}{4}[(p_o - a_o) + \kappa_o(p_o - a_o)^3]$ for $o \le 4$ and $q_5 = -1/4$. These reproduce the frozen peer-count lookup registry. Every term inside the brackets is nonnegative, so the unique optimum is $(a, 0)$. The potential is strictly concave along simplex tangent directions, and the quotient $q_o - q_5$ has genuine degree three.

The G3 construction is the componentwise L^2 projection of $q_o - q_5$ onto total degree at most two under the uniform probability measure on the five-outcome simplex. A coordinate has marginal law $\mathrm{Beta}(1, 4)$; the cubic projection is

$$\mathrm{proj}_{\le 2}(t^3) = t^2 - \frac{1}{4}t + \frac{1}{84}.$$

Expanding q_o and replacing each p_o^3 by this polynomial, while keeping $q_5 = -1/4$, gives the frozen G3 field. This also equals the full multivariate projection: the conditional expectation of any total-degree-two monomial given p_o is a polynomial of degree at most two in p_o. The quotient L^2 residual is $\sqrt{5}/42 \approx 0.05323971375$; the projected optimum is $(.3395, .2895, .2049, .1660, 0)$. G4 uses the degree-three Bernstein coefficients of the exact field, and G5 uses their ordinary degree elevation. The potential, projection measure, coefficient scale, and lookups are fixed across the reported training runs.

E EXPERIMENTAL PROTOCOLS AND ADDITIONAL RESULTS

E.1 EXACT FINITE-SUPPORT CHECKS

The sequence audit uses float64 and exhaustively enumerates all length-four binary sequences and multinomial outcome counts. Its two deterministic canonicalizers are first-one position and Hamming weight. Policy seeds, logit scales, group sizes, and degrees were fixed before evaluation; all cases remain in the aggregates.

The constrained real-model audit uses pinned Qwen2.5 checkpoints (Qwen Team, 2024). Cert24 is a 24-point arithmetic task with a certified finite set of valid expressions. Its support and token trie are fixed before scoring; `configs/pr07a_constrained_cert24.yaml` specifies the audit. Dynamic programming computes exact leaf probabilities and edge scores, and direct autograd independently differentiates each objective. The adapter pullback is evaluated at zero delta without an optimizer step. Figures 5 and 6 summarize these checks.

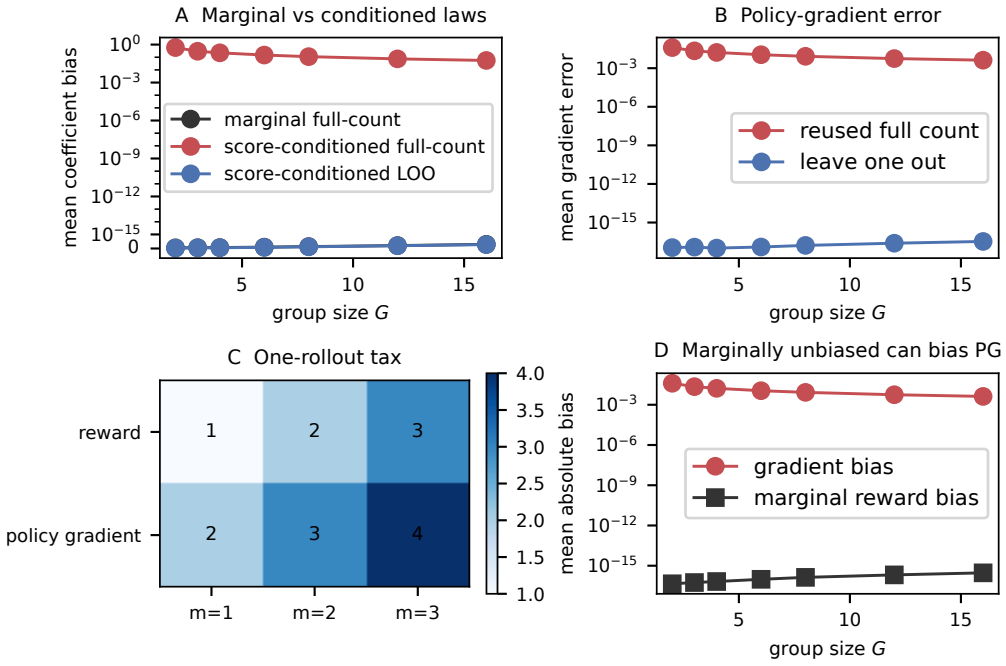


Figure 5: **Self-inclusion changes the score-conditioned law.** The full-count falling-factorial lookup is exact under the marginal group law. Reusing it on a scored sample produces the transform in Theorem 4, while leave-one-out reaches float64 precision across the fixed grid.

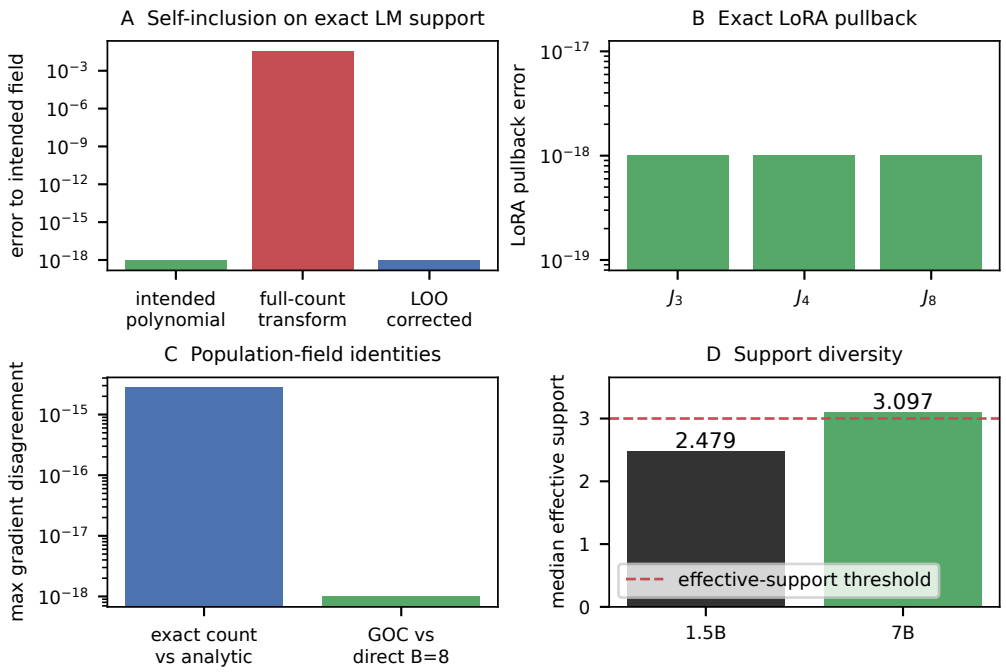


Figure 6: **Exact gradients on constrained Qwen support.** Certified trie supports allow enumeration of leaf probabilities and score/autograd gradients. These checks verify the full-count, leave-one-out, finite-budget and GOC population identities on that support. In D, effective support is $\exp(H(p))$, summarized by its median across tasks; the dashed line marks the pre-specified minimum of 3. This diversity criterion is separate from gradient agreement. The maximum probability normalization error $|\sum_o p_o - 1|$ is 3.0×10^{-15}.

E.2 Unconstrained training and control variates

The initial MultiRoute development configuration has 2,000/200/500 train/validation/test graphs, exactly four enumerated paths per graph, and 100 hashed test anchors. Prompts list the paths under common labels A–D; free generation returns a label and all other text is mapped to invalid. The initial rewards are $(1, .9, .8, .65)$, with zero for invalid output. The objective-map panels instead use the separately frozen vector $(1, .95, .70, .65)$, again with zero for invalid; every method in those panels uses this same reward configuration.

Raw G4, centered G4 and centered G5 consume 120 completions per update for 1,600 updates. AdamW uses learning rate 10^{-5}, zero weight decay, and a cosine tail from update 800 to 10^{-6}. Temperature and top-p are both one. There is no RL gradient-norm clipping, replay, KL penalty or constrained decoder. Gradient norms are logged without changing the update. The shared rank-16, alpha-32 LoRA warm start is selected using only the validation threshold on the invalid-output rate; graph topologies, split and anchor IDs, rewards and targets are fixed across methods.

Every uncentered coefficient method—raw EIPS, set-marginal, GOC and the E1 peer lookup—uses the same independent-group scalar control variate. Each focal group uses the mean coefficient of the other contemporaneous independent main groups as its baseline. Conditional on batch prompts, this baseline is independent of the focal group, so

$$\mathbb{E}\left[B_{-g}\sum_{i\in g} s_i\right] = 0.$$

Its law depends on the policy. It is therefore an external field-neutral variance reduction, not the fixed-law U in Equation 1. Theorems 1 and 5 identify each core field before this subtraction. Same-group-centered EIPS instead uses the objective-changing transformation of Theorem 5 and receives no external baseline.

Under this reward configuration, the compute comparison uses Raw G4, Centered G5 and GOC$(2, 2)$. For every two-rollout score-bearing main group, GOC generates an independent two-rollout probe from the same pre-update snapshot. The probe enters only the detached coefficient

$$2r_o C_o^{-1}[1 + \mathbf{1}\{o \text{ absent from probe}\}].$$

Probe tokens are not rescored and do not enter backpropagation. Centered G5 and GOC use all five seeds. A fresh Raw seed-0 run checks the J_4 conclusion against the five existing Raw runs; training trajectories are not assumed to be bitwise reproducible. Figure 7 separates score-bearing and probe costs.

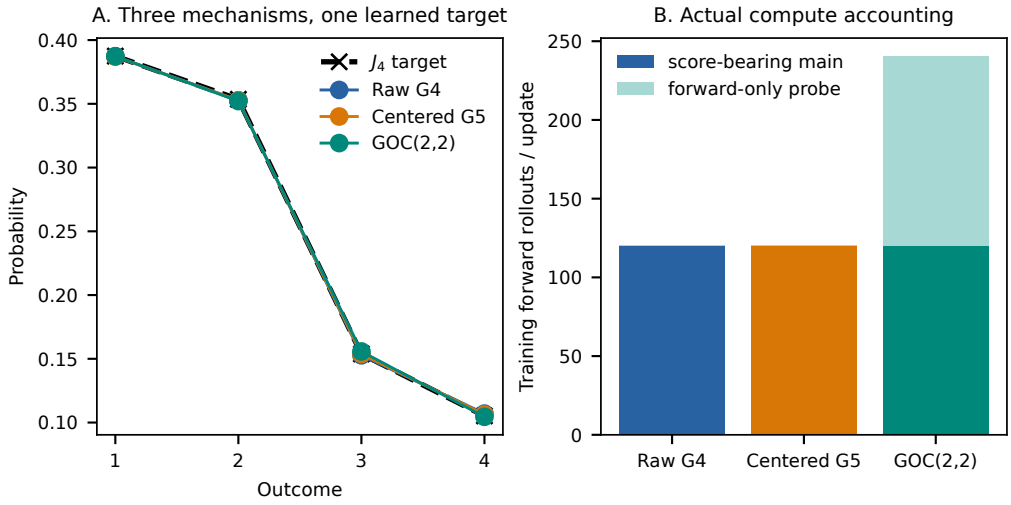


Figure 7: **Three computations, one predicted objective.** Left: learned distributions and the J_4 optimum. Right: score-bearing and forward-only rollouts. Only the main rollouts carry scores or enter backpropagation.

The action interface and finite-optimization schedule were fixed after examining superseded runs. The manifest records this use of model results; rewards, estimators, targets, paired seeds and anchors were not changed. The reported comparison is therefore a validation under the revised protocol, not a claim that every engineering choice preceded all model observations.

E.3 Output-surface replication

MultiSolve-4 replaces labeled graph routes with the equations $|x| = a$ and $|y| = b$ for positive two-digit magnitudes. The four signed integer pairs are the complete valid support; a strict one-line parser maps all other text to a fifth invalid outcome. The deterministic 2,000/200/500 split and 100 test anchors isolate variable-swap clusters. Complete-support SFT quartets give each prompt all four valid answers once. The final shared warm start passes the validation gates: invalid rate 0.00125, outcome range $[0.03875, 0.45750]$, and multi-outcome fraction 0.98.

The reward vector, exact $J_3/J_4/U_4$ targets, five methods, seeds, 1,600-update horizon and 120 score-bearing rollouts per update are retained. Surface calibration and the final sign-to-reward permutation are disclosed in the manifest; neither raw generations nor post-update pilot results were inspected. All 25 retained post-calibration runs use unconstrained generation from the same frozen code revision. This panel changes the output surface rather than testing reasoning generalization. GOC alone adds 120 forward-only probes per update.

E.4 7B scale replication

The scale replication keeps the MultiRoute panel, canonicalizer, reward vector, anchors, exact targets and five estimator constructions unchanged, replacing the base model by revision-pinned Qwen2.5-7B-Instruct. Three paired seeds produce 15 runs. A new rank-16, alpha-32

shared SFT adapter passes at update 50 with invalid rate 0.00125, aggregate valid outcome range $[0.22375, 0.27750]$, and multi-outcome prompt fraction 1.0. Each method uses 1,600 updates and 120 score-bearing rollouts per update; GOC adds 120 forward-only probes. The runs share one frozen code revision and warm-start hash. Final adapter checkpoints were not retained. Full trajectories, evaluations, provenance and compute records are retained.

E.5 Rollout expressivity threshold

E1 uses the unchanged MultiRoute panel and the separable strongly concave potential specified in Appendix D.3. Its quotient field has degree three and its exact optimum is $(.40, .35, .15, .10, 0)$. The degree-two projection of the full quotient under uniform simplex measure has optimum $(.3395, .2895, .2049, .1660, 0)$. Exact enumeration fixes the G3 projection residual, the lookup coefficients, the minimum exact size G4, and the G5 degree elevation before model runs. G3, G4 and G5 share a fresh rank-16 SFT adapter and five paired seeds. Each run uses 2,400 updates, 120 unconstrained score-bearing completions per update and the independent-group control variate, without auxiliary rollouts, clipping, replay or KL penalty. Final adapter checkpoints were not retained. Final metrics average 512 samples on each of the 100 inherited anchors. The analysis includes all 15 runs and all five pre-specified questions. This experiment compares the fixed exact field to its selected projection; the reward-affine minimax family in Section 6 is a separate analytic construction.

E.6 Self-inclusion fixed points

The binary task has 128 fixed prompt paraphrases and a total A/B mapping. The canonicalizer maps an output to A if its first alphabetic character is A (case-insensitive), and to B otherwise, including outputs with no letters. There is no separate invalid outcome in this binary task. For $m \in \{2, 3, 4\}$, full-count uses $G = m$ and coefficient

$$\mathbf{1}\{O_i = A\}(2^{-m} - \mathbf{1}\{C = m\}).$$

Leave-one-out uses $G = m + 1$ and removes the score-bearing sample from the peer count. Each of the six conditions uses five seeds, 300 updates and 60 completions per update. Evaluation occurs every 25 updates; the final estimate uses 256 samples per prompt. All conditions are retained and the aggregate uses 10,000 seed-bootstrap replicates.

Table 3: **Self-inclusion predicts the learned fixed point.** Mean [descriptive 95% seed-bootstrap interval] over five seeds; the last column tests the pre-specified absolute-error radius of 0.03.

m	method	G	theory p	empirical p	abs. err.	$< .03$
2	Full	2	0.2500	0.238 [0.219, 0.265]	0.012	yes
2	LOO	3	0.5000	0.514 [0.503, 0.523]	0.014	yes
3	Full	3	0.3536	0.337 [0.316, 0.357]	0.017	yes
3	LOO	4	0.5000	0.522 [0.501, 0.545]	0.022	yes
4	Full	4	0.3969	0.382 [0.362, 0.402]	0.015	yes
4	LOO	5	0.5000	0.500 [0.460, 0.534]	0.000	yes

E.7 Sampled circulation

E4 uses ordinary realized-reward standardization with $K = 3$, $G = 4$, and $D_G(v) = \sqrt{v + 10^{-6}}$, with within-group variance computed using divisor G. The residual point is $(.2, .3, .5)$, and the counterclockwise loop is the rectangle $(.15, .25) \to (.25, .25) \to (.25, .35) \to (.15, .35)$. It compares three levels $(0, 1, 2)$ with two levels $(0, 1, 1)$ while keeping the outcome space fixed. At each field point, N complete iid multinomial groups estimate

$$\widehat{q}_o = (NGp_o)^{-1} \sum_g C_{g,o} A_o(C_g).$$

Exchangeability makes this estimator unbiased; it uses known design probabilities but no exact conditional expectation. The conditions share the same sampled count arrays in each replicate and field point. Across 64 replicates, five budgets range from 128 to 32,768 groups. For three levels, direct

differentiation of the exact count polynomial gives residual -3.0425775, agreeing with an independently constructed symbolic polynomial to machine precision; the finite-difference target at $h = .01$ is -3.0427710. Its 32-interval-per-edge loop target is $.03049031$, while the separately computed 4096-interval reference is $.03049028$. For two levels, the exact residual is zero and both loop calculations have magnitude below 10^{-16}. Every RMSE uses the corresponding finite-difference or finite-quadrature target. The theorem's all-G certificate is separately checked at its stated middle-reward vertex, where its value 5.3710966 agrees with count-polynomial differentiation to 8.9×10^{-16}. The vertex certificate, interior analytic residual, finite-difference target, and two loop quadratures are reported separately. Fitted slopes describe only the sampled budget range; the samples illustrate recovery and do not prove the classification.

The EIPS-standardized witness of Proposition 2 provides secondary exact evidence: at rewards $(.7, 1.2, 1.9)$ its analytic residual is -1.4735827, and its closed-loop work with 32 panels per edge is $.01493479$. This witness uses inverse-count coefficients and is distinct from ordinary reward standardization.

Separate exploratory trajectory. After the circulation analysis, a separately frozen protocol follows 64 stochastic categorical-logit ascent paths from $(.2, .3, .5)$. Each of 400 steps uses 2,048 new full iid groups and step size $.05$; no exact field enters an update. The sampled mean stays within 5.3×10^{-4} of exact RK4, ending at $(.06706, .30394, .62900)$ versus $(.06713, .30398, .62889)$; the middle probability falls to $.29232$ and returns to $.30394$ in the retained supplementary trajectory artifact. Neither oscillation nor a limit-cycle endpoint was a pre-specified criterion, and this trajectory is not part of the E4 reward-level test.

E.8 What each type of evidence establishes

Table 4: Analytic results and empirical checks; finite checks do not replace proofs.

Result	Supporting check
Degree/converse (Theorem 1)	Enumeration; E1 selected-field threshold
Baseline quotient (Theorem 2)	Exact finite-support gradients
Ordinary normalization (Theorem 3)	E4 paired iid field recovery; proof is analytic
Self-inclusion (Theorem 4)	Exact audits; E3 binary roots
Count-rule objectives (Theorem 5)	Exact audits; E2 in both tasks
Reuse loss (Theorem 6)	Deterministic proof checks
Polynomial threshold (Theorem 7)	Symbolic construction, distinct from E1
EIPS standardization (Proposition 2)	Secondary exact certificate