INFORMATION ARCHITECTURE AND DECISION LEARNABILITY IN STRATEGIC POPULATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study a multi-agent information design problem in which a coordinator repeatedly designs signals to influence the behavior of strategic agents. The coordinator may value different agent groups differently, but does not know their type compositions and observes only aggregate responses. Communication thus affects both the induced behavior and what the coordinator can learn about the underlying population. For a finite experiment library, sublinear regret at every population in an observational equivalence class requires a common optimal experiment. In a canonical game, public signals cannot distinguish populations with disjoint optimal experiment sets, leading to linear worst-case regret over a fixed population pair. A single population-independent private policy achieves at most logarithmic regret at each fixed population, with each architecture evaluated against its own oracle. We characterize the informational limits of public feedback and show how targeted signals can enable local identification. We derive a Kullback–Leibler (KL) regret lower bound that accounts for information gained under optimal play, compute its canonical value, and construct a matching policy. The resulting logarithmic coefficient may be positive or vanish, while alternatives approaching a decision boundary yield a local minimax regret of order $\sqrt{T}$. In a nonlinear benchmark, information-theoretic calculations predict decision errors and explain why equally costly private probes differ in informativeness. Adaptive experiments on a population pair show how exploration incurs regret at one population while preventing persistent decision errors at the other.

1 INTRODUCTION

Consider a service platform whose users belong to subscription groups and differ in their service preferences. A coordinator acting as the sender sends availability messages by group, while its telemetry records only the aggregate usage of each service. Users respond to these messages and to congestion caused by others' choices. Because groups carry different value weights, aggregate usage does not directly reveal the coordinator's realized reward. The coordinator must therefore choose whom to inform while learning the unknown population composition from aggregate responses. We ask how message targeting affects the learnability of an optimal policy and the regret incurred in learning it.

We formalize this problem as repeated Bayesian persuasion with unknown class compositions (Kamenica & Gentzkow, 2011; Bergemann & Morris, 2019). Each signal changes receivers' posterior beliefs and their equilibrium behavior, jointly determining the sender's value and observation kernel (Figure 1). Under class-neutral utilities, public signals induce the same response from receivers of the same type across classes, while the sender may value their classes differently. Reallocating types across classes can therefore preserve aggregate feedback yet change the optimal experiment. Repeated public observations cannot resolve this ambiguity.

Learning under partial feedback requires a common optimal decision among observationally indistinguishable environments (Hanawal et al., 2016; Parisi et al., 2024). Here signaling and strategic equilibrium determine those equivalence classes, the architecture's *observational quotient*. When a class has no common optimum, its feedback cannot support sublinear regret at every population. Our two-class game gives an explicit separation: every public learner incurs linear worst-case regret over a fixed pair, whereas one population-independent private policy achieves at most logarithmic regret at

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.

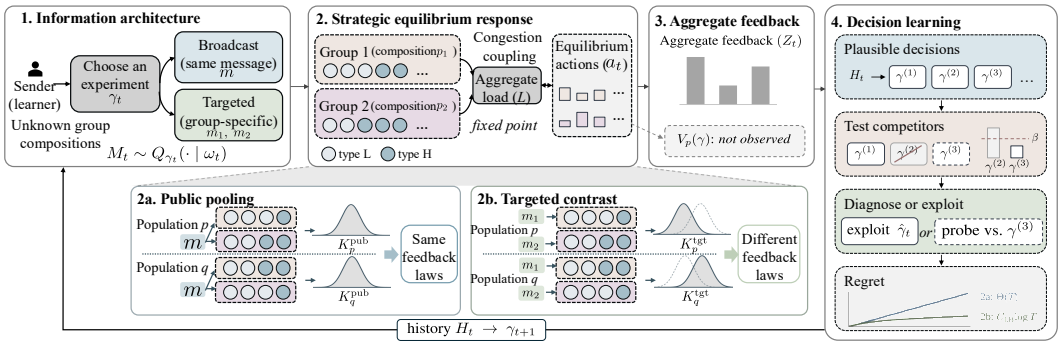


Figure 1: Public and targeted signals shape equilibrium feedback and the information available for learning optimal decisions.

each fixed population. Each architecture uses its own oracle, separating learning costs from static value differences.

Once signals distinguish populations requiring different decisions, learning costs depend on which experiments provide that information. An optimal experiment may itself distinguish the relevant alternatives; otherwise, the sender must incur regret to explore. This gives the two layers of our analysis. *Decision resolution* asks whether every observational class admits a common optimum, without requiring full population identification. *Learning complexity* quantifies the cost of distinguishing competing decisions through Kullback–Leibler (KL) allocations, accounting for information acquired under optimal play. Our contributions are as follows.

1. **Equilibrium-induced observational collapse.** Public signaling pools class-specific compositions, whereas targeted contrasts can provide local identification (Theorem 2). We connect this equilibrium-induced loss of information to decision non-resolution and linear regret (Theorem 1), with a finite witness of at most $|\Gamma|$ populations (Corollary 1).
2. **Exact information geometry and exploration cost.** Using structured-bandit KL allocations (Graves & Lai, 1997; Combes et al., 2017; Komiyama et al., 2015), we identify the canonical alternatives left indistinguishable under optimal play. We derive exact regret coefficients and a matching policy (Theorems 3–4), including zero-cost information under exploitation and positive exploration costs whose factor of four follows from the decision-boundary geometry.
3. **Architecture-dependent canonical regret regimes.** Every public learner incurs linear worst-case regret over a fixed pair, while one population-independent private policy achieves at most logarithmic regret at each fixed population, against each architecture's own oracle. Shrinking alternatives give a two-point $\Theta(\sqrt{T})$ boundary rate (Theorem 5); the fixed-instance separation persists under small perturbations (Proposition 2). Nonlinear experiments test the information supplied by equally costly probes and the regret cost of resolving decision ambiguity.

2 Related Work

Unsupervised sensor selection characterizes learnability through an optimal-action map of the observable sensor distribution (Hanawal et al., 2016, Proposition 1). Monitored MDPs require a common optimal policy within each indistinguishability class (Parisi et al., 2024, extended version, Appendix B). We derive these classes from signaling and equilibrium aggregation, then quantify their canonical learning costs.

Partial monitoring relates observability to regret when losses and feedback differ (Bartók et al., 2014; Kirschner et al., 2023). Instance-dependent KL allocations extend bandit lower bounds to controlled sensing, structured bandits, and stochastic partial monitoring (Lai & Robbins, 1985; Graves & Lai, 1997; Combes et al., 2017; Komiyama et al., 2015). Sequential design and best-arm identification supply allocation and stopping principles (Chernoff, 1959; Nitinawarat et al., 2013; Naghshvar & Javidi, 2013; Garivier & Kaufmann, 2016; Elfving, 1952). Related decision-learning settings include

decision-focused learning, adaptive contracts, and strategic recommendation (Capitaine et al., 2026; Ho et al., 2014; Chen et al., 2023; Kleine Buening et al., 2024).

Bayesian persuasion optimizes signals (Kamenica & Gentzkow, 2011; Bergemann & Morris, 2019); online variants learn utilities, types, priors, or private information (Castiglioni et al., 2020; 2021; 2023; Bernasconi et al., 2023; Bacchiocchi et al., 2024; Lin & Li, 2025; Vurankaya & Topcu, 2026). Zu et al. (2025) study a state distribution unknown to sender and receivers; our sender knows the state prior but learns group compositions. Learning receivers (Lin & Chen, 2025) differ from our fixed equilibrium-response setting.

Principals can induce myopic agents to explore unknown arm rewards through information disclosure (Kremer et al., 2014; Mansour et al., 2015) or payments (Frazier et al., 2014). Here the sender learns population composition from equilibrium aggregate feedback, which can conceal differences relevant to its unobserved reward.

Bayes-correlated equilibrium and multi-channel persuasion study information in games (Bergemann & Morris, 2016; Arieli & Babichenko, 2019; Babichenko et al., 2022). With known types, Shaki et al. (2025) use type symmetry to compute optimal public, private, and semi-private policies under externalities. Our unknown class compositions can conceal decision-relevant differences in aggregate feedback. In routing games, Massicot & Langbort (2022) compare signaling efficiency, while Verbree & Cherukuri (2021) infer priors and population fractions from Wardrop flows. Hossain et al. (2025) learn classifiers under externalities. Performative prediction studies deployment-induced distribution changes (Hardt & Mendler-Dünner, 2025); here signaling changes the observation kernel through equilibrium.

3 Information Design with Strategic Populations

We consider a nonatomic unit-mass population of receivers partitioned into sender-observable classes $g \in [G]$. Class g has mass $w_g > 0$, with $\sum_g w_g = 1$, and a fixed type composition $p_g \in \Delta^{K-1}$ unknown to the sender. Write the population composition as $p = (p_1, \dots, p_G)$. Receivers take aggregate loads as given. Each round follows

$$H_{t-1} \longrightarrow \gamma_t \longrightarrow \omega_t \longrightarrow M_t \longrightarrow \mathbf{a}_t \longrightarrow Z_t.$$

The sender uses past history H_{t-1} to publicly commit to an experiment γ_t from a finite library Γ. It observes the state $\omega_t \sim \nu$, where ν is a known common prior independent of p, and draws the message profile $M_t \sim Q_{\gamma_t}(\cdot \mid \omega_t)$. Each message kernel Q_γ is population independent. Bayesian updating gives posteriors that average to ν (Bayes plausibility). Equilibrium actions $\mathbf{a}_t$ induce aggregate loads L_t, and the sender observes feedback Z_t.

A type-k receiver chooses a to maximize $\mathbb{E}[u_k(a, L, \omega) \mid m]$ given its own message m. The load L creates an anonymous externality. Class affects targeting and sender value, but public pooling assumes class-neutral receiver utility. Receivers know p when predicting conditional loads. This benchmark fixes their equilibrium response while the sender learns, separating its estimation problem from receiver learning. A unique equilibrium or fixed population-independent selection rule induces value $V_p(\gamma)$ and kernel K_p^γ for the complete transcript $O_t = (\omega_t, M_t, Z_t)$. Class loads, types, and realized reward are unobserved.

Each round involves a fresh receiver cohort with the same fixed composition p, so the sender learns about the population without receiver adaptation across rounds. Conditional on γ_t, the transcript O_t follows the stationary kernel $\mathsf{K}_p^{\gamma_t}$ independently of past history. An adaptive policy uses population-independent kernels $\gamma_t \sim \pi_t(\cdot \mid H_{t-1})$, where $H_{t-1} = (\gamma_s, O_s)_{s<t}$ may include internal randomization.

The sender's optimal set, one-step gap, and pseudo-regret are

$$V_p^\star = \max_{\gamma \in \Gamma} V_p(\gamma), \qquad \Gamma^\star(p) = \arg\max_{\gamma \in \Gamma} V_p(\gamma),$$

$$\Delta_p(\gamma) := V_p^\star - V_p(\gamma), \qquad R_p^\pi(T) = \mathbb{E}_p^\pi\left[\sum_{t=1}^{T} \Delta_p(\gamma_t)\right].$$

Because reward is unobserved, two populations may produce identical feedback laws while having different experiment values and optimal sets. If an integrable reward with mean $V_p(\gamma)$ were observed,

equal transcript laws across all experiments would imply identical values and optimal sets. The architecture specifies the feasible library Γ and feedback channel, with oracle $V_p^\star$ defined within that library.

4 Observational Quotients and Decision Resolution

On population space $\mathcal{P} = \prod_g \Delta^{K-1}$, strategic transcript laws induce the equivalence relation

$$p \equiv_\Gamma q \iff \mathsf{K}_p^\gamma = \mathsf{K}_q^\gamma \quad \forall \gamma \in \Gamma, \qquad \rho_\Gamma(p) = [p]_\Gamma.$$

The map ρ_Γ sends populations to the quotient $\mathcal{Q}_\Gamma = \mathcal{P}/\equiv_\Gamma$. We call the architecture *decision resolving* if every class admits a common optimum:

$$\bigcap_{q \in [p]_\Gamma} \Gamma^\star(q) \neq \varnothing \quad \forall p.$$

Equivalently, an optimal selector s_Γ chooses a common optimum for each quotient class without recovering the population. Pairwise intersections of optimal sets alone are insufficient.

Adaptivity cannot distinguish populations whose kernels agree for every available experiment: induction gives the same history law at each population. If their optimal sets have no common element, every chosen experiment incurs a positive gap somewhere in the family. Summing these gaps yields the following bound.

Theorem 1 (Linear regret under observational equivalence)**.** *Let $p_1, \dots, p_m$ be mutually observationally equivalent populations and suppose Γ is finite. Every adaptive randomized policy induces the same full-history law at all p_i. If $\bigcap_{i=1}^m \Gamma^\star(p_i) = \varnothing$, then, with*

$$\Delta_{\mathrm{fam}} = \min_{\gamma \in \Gamma} \sum_{i=1}^m \Delta_{p_i}(\gamma) > 0,$$

every adaptive policy satisfies $\sum_i R_{p_i}^\pi(T) \geq \Delta_{\mathrm{fam}} T$, and hence $\max_i R_{p_i}^\pi(T) \geq \Delta_{\mathrm{fam}} T/m$.

Corollary 1 (Finite witness of decision non-resolution)**.** *If an observational class has no common optimum, it contains a fixed family of at most $|\Gamma|$ populations satisfying the linear regret bound of Theorem 1 with $\Delta_{\mathrm{fam}} > 0$.*

Appendix A proves both results. Decision resolution is necessary for sublinear regret throughout $\mathcal{P}$, but does not specify a learning rule.

Under public signals, class-neutral utility and a common posterior give the same type-specific best response across classes. The aggregate fixed-point problem then depends only on $\bar{p} = \sum_g w_g p_g$. Theorem 2 states when equilibrium selection preserves this dependence. In simplex coordinates, the pooling map is $A = w^\top \otimes I_{K-1}$, where $\otimes$ denotes the Kronecker product.

For private identification, let $h(\theta)$ be the mean of fixed contrasts from independent targeted observations, with Jacobian $J_h = \nabla h$. Write r_Γ for the Fisher rank in simplex coordinates. The density and moment assumptions for these diagnostics are given in Appendix A.2.

Theorem 2 (Identifiability under public and private signaling)**.** *Suppose receivers have class-neutral utilities and face an anonymous strategic externality. Assume that public best responses are unique or selected by a fixed class-invariant rule based only on the best-response problem. Assume also that aggregate fixed points under public messages are unique or selected by a fixed rule depending only on the fixed-point problem. If feedback is generated from aggregate behavior through a population-independent observation kernel, the complete public transcript law depends on p only through $\bar{p}$. Thus $r_{\mathrm{pub}} \leq K - 1$ and $\dim \ker A = (G-1)(K-1)$. For a targeted private library, a full column rank contrast Jacobian J_h achieves the maximal local rank $r_{\mathrm{priv}} = G(K-1)$. At zero externality, this holds when targeted contrasts isolate each class and the K type-specific contrast vectors are affinely independent within every class. If the contrast Jacobian depends continuously on the externality parameters, this rank condition persists near zero externality.*

Public pooling equates observations along entire fibers, whereas the private rank condition only separates nearby populations. Full rank need not exclude distant equivalents, and deficient rank

need not prevent a common optimum. We next check whether the pooled differences affect optimal decisions.

For an attainable pooled composition $\bar{p}$, the fiber $\mathcal{F}(\bar{p}) = \{q : \sum_g w_g q_g = \bar{p}\}$ has the same public observations throughout. By Theorem 1, sublinear regret on this fiber requires a common public optimum. A decision boundary crossing an exactly hidden direction in $\mathcal{N}_{\mathrm{pub}} = \ker A$ can violate this condition; Appendix A.3 gives the local derivative criterion. A zero Fisher direction alone need not preserve exact observation laws.

5 Learning Complexity: Information and Regret

Learning which experiment is optimal may require exploratory plays of suboptimal experiments. We quantify this exploration cost using the adaptive change-of-measure argument of Graves & Lai (1997). The relevant alternatives are populations with disjoint optimal sets:

$$\mathrm{Alt}(p) = \{q : \Gamma^\star(q) \cap \Gamma^\star(p) = \varnothing\}.$$

A uniformly efficient policy must choose experiments from different optimal sets under p and each alternative. Binary testing therefore requires at least $(1 - o(1)) \log T$ history KL. If $N_\gamma(T)$ counts selections of γ, the adaptive KL chain rule writes this divergence as expected counts times one-round KL. The allocation variable x_γ represents $\mathbb{E}_p[N_\gamma(T)]/\log T$ along a common convergent subsequence. This leads to

$$C_{\mathrm{LB}}(p) = \inf_{x \in [0,\infty]^\Gamma} \left\{ \sum_{\gamma : \Delta_p(\gamma) > 0} x_\gamma \Delta_p(\gamma) : \sum_\gamma x_\gamma D_{\mathrm{KL}}(\mathsf{K}_p^\gamma \| \mathsf{K}_q^\gamma) \geq 1 \ \forall q \in \mathrm{Alt}(p) \right\}. \tag{1}$$

We use $\infty \cdot 0 = 0$ and $\infty \cdot d = \infty$ for $d > 0$. Optimal experiments played $\Theta(T)$ times have $x_\gamma = \infty$, contributing information at zero regret.

Theorem 3 (Instance-dependent regret lower bound)**.** *Suppose Γ is finite and all experiment values are finite. Assume that complete transcript kernels have densities with respect to a common measure and that the relevant KL divergences are finite. If a horizon-indexed adaptive policy family is uniformly efficient (for every fixed q and $a > 0$, $R_q^{\pi^{(T)}}(T) = o(T^a)$), then*

$$\liminf_{T \to \infty} \frac{R_p^{\pi^{(T)}}(T)}{\log T} \geq C_{\mathrm{LB}}(p).$$

Appendix B.1 proves the testing bound and passes from normalized counts to a feasible allocation.

Following structured-bandit confusing alternatives (Combes et al., 2017), define the populations invisible under every optimal experiment:

$$\mathrm{Conf}_0(p) = \{q \in \mathrm{Alt}(p) : D_{\mathrm{KL}}(\mathsf{K}_p^\gamma \| \mathsf{K}_q^\gamma) = 0 \ \forall \gamma \in \Gamma^\star(p)\}.$$

Optimal play separates alternatives outside this set at zero regret cost; those inside require suboptimal experiments. Under finite KL, $C_{\mathrm{LB}}(p) = 0$ exactly when $\mathrm{Conf}_0(p) = \varnothing$ (Proposition 1). With multiple optima, this lower bound need not be attainable. Appendix B groups these constraints by competing decisions, motivating the nonlinear tests.

6 Canonical Game: Decision Resolution and Regret

We now derive the equilibrium kernels and values in a two-class game, evaluate the lower bound, and construct a policy that attains it.

Two equal-mass classes have type-H fractions $p = (x, y) \in [0, 1]^2$, with $s_L = 1, s_H = 2$. The state is uniform on $\{-1, +1\}$, and a type-k receiver chooses $a \geq 0$ with payoff

$$u_k(a) = a(20 + s_k \omega) - \frac{a^2}{2} - 8aL.$$

The sender observes only (ω, Z) with $Z = L + \epsilon$, $\epsilon \sim \mathcal{N}(0, \tau^2)$. Writing n for an uninformative message, the Bayes-plausible library contains no disclosure $N = (n, n)$, public disclosure $F = (\omega, \omega)$, and targeted disclosures $P_1 = (\omega, n)$ and $P_2 = (n, \omega)$.

Strict concavity ensures a unique best response, and the equilibrium lies in the interior (Appendix C.2). Therefore, for message m, the first-order condition is $a = 20 + s_k\mathbb{E}[\omega \mid m] - 8\mathbb{E}[L \mid m]$. Revealing messages fix ω; message n leaves its mean at zero. Aggregating these conditions gives the unique loads

$$L_\omega^N = L_0, \qquad L_\omega^F = L_0 + \omega\frac{2+x+y}{18},$$
$$L_\omega^{P_1} = L_0 + \omega\frac{1+x}{10}, \qquad L_\omega^{P_2} = L_0 + \omega\frac{1+y}{10}, \qquad L_0 = \frac{20}{9}.$$

The same aggregate measurement reveals $x + y$ under public disclosure and x or y under targeting.

The sender value is class sensitive but unobserved online:

$$V_p(\gamma) = \mathbb{E}\big[\omega\left(A_{1,\omega}^\gamma + \tfrac{1}{2}A_{2,\omega}^\gamma\right)\big] - c_\gamma,$$

where $A_{g,\omega}^\gamma := \sum_k p_{gk}a_{gk,\omega}^\gamma$ is class g's per-capita action before multiplying by its population mass. The experiment-dependent payoff offsets $(c_N, c_F, c_{P_1}, c_{P_2}) = (1/4, 1/2, 3/20, 0)$ align the public and private decision boundaries and make public experiments strictly dominated in the full library. The common $1/4$ shift cancels from public-oracle regret, so this calibration does not create the public learning lower bound (Appendix C.3). With $d = 2x - y - 1/2$, substituting the equilibrium gives

$$V(N) = -\tfrac{1}{4}, \quad V(F) = -\tfrac{1}{4} + \tfrac{d}{6}, \quad V(P_1) = \tfrac{1+4x}{20}, \quad V(P_2) = \tfrac{1+y}{10}.$$

Hence P_1 is optimal for $d > 0$, P_2 for $d < 0$, and they tie at $d = 0$. Public policies share this boundary but cannot distinguish populations along $x + y =$ constant: the hidden direction $(1, -1)$ changes d at rate three. In contrast, a pooled-load objective makes this rank-one library decision resolving, with distinct N- and F-optimal regions across fibers (Appendix C.5).

Optimal play already distinguishes some alternatives. For those that remain, equal-variance Gaussian KL is the squared mean difference divided by $2\tau^2$, so the closest decision-boundary limits determine the constraints. Minimizing regret per unit KL gives the constants below (Appendix C).

Corollary 2 (Canonical regret lower bound)**.** *For every off-boundary canonical population,*

$$C_{\mathrm{LB}}(p) = \begin{cases} 20\tau^2/d, & d > 0,\ x < 3/4, \\ 0, & d > 0,\ x \geq 3/4, \\ 80\tau^2/|d|, & d < 0. \end{cases}$$

P_1 reveals x but leaves y unknown. When $x \geq 3/4$, the constraint $y \leq 1$ gives $d \geq 0$, so P_1 remains optimal for every compatible y. Thus, for $d > 0$, optimal play distinguishes every fixed alternative with disjoint optima, including at $x = 3/4$. This satisfies the zero-coefficient condition without identifying the full population.

In the two positive-coefficient regions, the factor of four has a geometric explanation. When $d > 0$ and $x < 3/4$, optimal P_1 play fixes x, so the nearest decision-boundary point changes y by d. When $d < 0$, optimal P_2 play fixes y, and changing x by $|d|/2$ reaches the boundary. The latter change gives one quarter of the KL for the same gap $|d|/10$, hence four times the regret coefficient.

Under P_1 and P_2, the transformed observation $10\omega(Z - L_0) - 1$ has distribution $\mathcal{N}(x, 100\tau^2)$ and $\mathcal{N}(y, 100\tau^2)$, respectively. Thus deciding between the experiments reduces to testing the sign of d. A pilot chooses the likely optimum for additional sampling, which has no regret when the choice is correct. The rival is then sampled at geometric checkpoints until a Gaussian sign test accepts a decision. Its leading sample count matches the KL constraint; Gaussian tail bounds control the expected loss from incorrect commitments. When the estimated x is sufficiently close to or above $3/4$, a separate stopping rule attains the zero coefficient.

Theorem 4 (Asymptotically optimal canonical policy)**.** *There exists a horizon-dependent policy family independent of p that uses only P_1, P_2 and observes only (ω, Z). For every fixed canonical population, it satisfies*

$$R_p^{\pi^{(T)}}(T) = C_{\mathrm{LB}}(p)\log T + o(\log T).$$

The policy is uniformly efficient on $[0, 1]^2$.

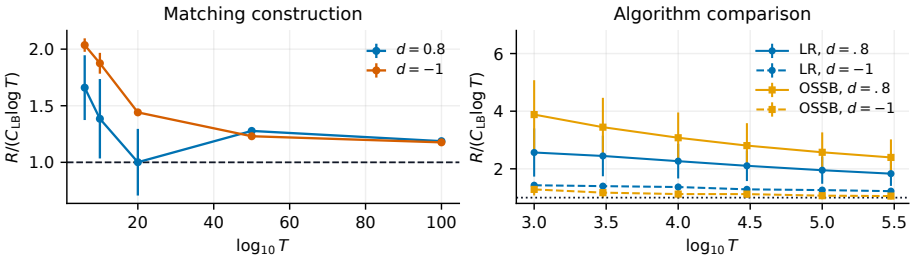


Figure 2: Canonical regret normalized by $C_{\mathrm{LB}}(p) \log T$, at $(x, y) = (.70, .10)$ ($d = .8$) and $(.10, .70)$ ($d = -1$); $\tau = .05$, P_1/P_2 feedback. Left: batched sign testing, 200 seeds. Right: Decision + LR and OSSB, 20 paired seeds; $\epsilon_{\mathrm{OSSB}} = \gamma_{\mathrm{OSSB}} = .05$. Lines: one; bars: standard errors (SE).

Appendix D specifies the policy and proves its expected regret bound in both positive- and zero-coefficient regions. Uniform efficiency here is defined at every fixed population; it does not assert a uniform logarithmic rate over populations approaching a decision boundary.

The architecture comparison follows by fixing populations $q^+ = (.8, .4)$ and $q^- = (.5, .7)$. They share $x + y = 1.2$ and every public transcript law, but their public optima are F and N. Theorem 1 gives summed regret at least $T/30$, hence maximum regret at least $T/60$ for every public learner on this pair. The private policy of Theorem 4 has regret $o(\log T)$ at q^+ and $400\tau^2 \log T + o(\log T)$ at q^-. Public and private regret use oracles $\{N, F\}$ and $\{P_1, P_2\}$, respectively. Thus the comparison isolates learning within each library from differences in static architecture value.

At $d = 0$, $\{P_1, P_2\}$ intersects every canonical optimal set, so $\mathrm{Alt}(p) = \varnothing$ and $C_{\mathrm{LB}}(p) = 0$. Yet $T^{-1/2}$-local alternatives remain statistically indistinguishable, with $O(1)$ history KL, and their $T^{-1/2}$ action gap yields $\sqrt{T}$ regret.

Theorem 5 (Canonical two-point local boundary rate)**.** *At $p_0 = (1/2, 1/2)$, let $p_{+,T} = (1/2, 1/2 - c/\sqrt{T})$ and $p_{-,T} = (1/2, 1/2 + c/\sqrt{T})$. For fixed $0 < c < 10\tau$ and every integer $T \geq \lceil 4c^2 \rceil$, every adaptive rule satisfies*

$$R^{\pi}_{p_{+,T}}(T) + R^{\pi}_{p_{-,T}}(T) \geq \frac{c}{10}\left(1 - \frac{c}{10\tau}\right)\sqrt{T}.$$

Uniform randomization between P_1, P_2 gives worst-case regret $c\sqrt{T}/20$ on this pair, so the two-point local minimax rate is $\Theta(\sqrt{T})$.

Appendix E gives the proof; Figure 5 illustrates the scale.

Small perturbations within the class-neutral quadratic family preserve pooling, private dominance, and the positive- and zero-coefficient regions (Proposition 2). Gaussian upper confidence bounds preserve the regret-rate separation (Corollary 3). Class-dependent utilities can break pooling; class neutrality is structural.

7 Numerical Experiments

In this section, we first examine the regret coefficients and boundary rate in the canonical game, where the theoretical characterization is explicit. We then use a nonlinear benchmark to examine the observation mechanism beyond this tractable setting. In all experiments, learners observe only O_t, while the true populations are used solely for offline evaluation. Since rewards are unobserved, baselines must infer decision values from aggregate feedback.

7.1 Canonical regret constants

Figure 2 compares batched sign testing with its predicted coefficient. At $T = 10^{100}$, the normalized regret ratios are 1.187 and 1.177, still above the asymptotic limit. These horizons are evaluated by simulating until commitment and adding the remaining rounds times the committed decision's gap.

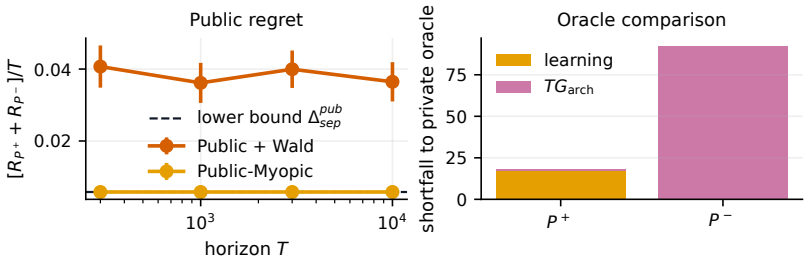


Figure 3: Public regret, 20 seeds. Left: $(R_{P^+} + R_{P^-})/T$ against the public oracle, mean with SE; line: normalized lower bound. Right: Public-Myopic regret and static value gap TG_{arch}, $T = 3000$.

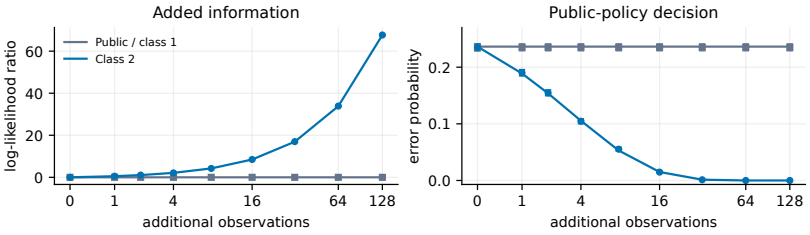


Figure 4: Known-pair test, 10000 replicates per endpoint. Left: added log-likelihood ratio supporting the true population (sign reversed at P^-). Right: error after four shared probes and n added observations. Predictions (lines) and means with 95% intervals (points); public and class-1 curves coincide.

We also compare Decision + LR with Optimal Sampling for Structured Bandits (OSSB) (Combes et al., 2017). The canonical LR rule fits the Gaussian likelihood and tests the two decision half-spaces (Appendix C.7). The matching guarantee in Theorem 4 applies to batched sign testing.

7.2 Nonlinear benchmark and protocol

The benchmark has four equal-mass classes, three types, and three services, with a uniform favorable-service state. Gumbel utility shocks give logit choice probabilities, and a strictly convex entropy-regularized congestion potential gives a unique equilibrium. The sender maximizes class-weighted returns minus congestion and signaling costs. It observes aggregate loads through $Z_t = L(M_t) + \xi_t$, with independent $\xi_t \sim \mathcal{N}(\mathbf{0}, (0.035)^2 I_3)$; class loads and reward are unobserved. The public oracle uses four public experiments; the full-library oracle adds eleven private experiments (Appendix F).

Numerical Fisher ranks are two for public and eight for private experiments (relative tolerance 10^{-6}; Figure 6, also comparing finer feedback). At congestion $\alpha = (.65, .75, .70)$, private full rank is a numerical finding beyond Theorem 2's continuation near zero externality.

Populations P^+, P^- share every public feedback law but have distinct public optima: disclosure and reveal probability .70. Theorem 1 gives $R_{P^+}^{\pi_{\text{pub}}}(T) + R_{P^-}^{\pi_{\text{pub}}}(T) \geq 5.817 \times 10^{-3} T$. Figure 3 separates public-learning regret from the static architecture gap TG_{arch}, where $G_{\text{arch}} = V_{\text{full}}^\star - V_{\text{pub}}^\star$.

We use known candidates P^+, P^- with equal priors and a terminal public-policy decision. Their public optima differ, while their full-library optimum coincides. After four shared initialization observations, one per class-targeted full-reveal probe, each branch adds n observations from public disclosure, class-1 probing, or class-2 probing. The two private probes have equal signaling cost .022. Because the candidates exchange classes 2 and 4, class-1 probing preserves their observational equivalence, whereas class-2 probing separates them.

Gaussian likelihood ratios give expected added information nD_γ, where D_γ is the one-round KL divergence between the candidates. It is zero for public and class-1 probes and .52869 for class-2 probing. Gaussian tails then predict decision error (Appendix F.3). At $n = 16$, class-2 probing gives

predicted error 1.497% and observed error 1.475% over 10000 replicates per endpoint. The other branches retain 23.50% error. Figure 4 compares all budgets: equally costly private probes can differ in whether they resolve the relevant population difference.

7.3 Adaptive decisions and exploration cost

We next remove the known-candidate assumption and let each learner estimate all eight population coordinates. The random suites mostly favor experiments that are informative as well as valuable: all 60 true optima across the two original suites have numerical local Fisher rank eight. Decision + LR averages only .12 additional diagnostics over $T = 3000$ in the first suite and .507 in an independent suite. This is consistent with sparse diagnostics; establishing $C_{\mathrm{LB}} = 0$ would require global separation of decision alternatives.

Decision + LR selects a diagnostic probe by variance reduction per estimated regret for the closest plausible decision competitor. It plays the estimated optimum when that experiment supplies sufficient information or the decision test passes. Greedy shares initialization, estimation, and refits. Across 30 populations and ten paired seeds, their mean regrets are .89 and 1.08, with paired difference interval $[-.567, -.0002]$. Decision + LR matches greedy in 295 of 300 runs, with 97% of the regret reduction coming from one corrected public-disclosure error. Appendix F.4 reports the independent suite and additional comparisons.

D-optimal policies select probes by Fisher log-determinant gain and exploit after testing. D-Optimal + LR incurs regret 54.20 with 995 diagnostics (LR/Wald variants: Figure 7, Appendix F.4).

To examine the case where exploration is costly, we halve congestion while keeping the same populations P^+, P^- and experiment library. Public disclosure becomes uniquely optimal at P^+, while a private experiment is optimal at P^-. Their public transcript laws remain equal, so optimal play at P^+ cannot distinguish this alternative. Thus $P^- \in \mathrm{Conf}_0(P^+)$, and the KL allocation gives $C_{\mathrm{LB}}(P^+) \geq .07276$ (Appendix F.5).

Table 1 reports 100 new paired seeds per population. At P^-, Decision + LR corrects all 16 terminal errors made by greedy and reduces mean regret. At P^+, both finish correctly, but Decision + LR incurs higher regret. Exploration therefore costs value at one population while preventing persistent public-policy errors at the other; a positive exploration-cost lower bound does not imply a finite-horizon advantage.

Table 1: Positive-cost pair, $T = 3000$, 100 new paired seeds per population. Regret includes initialization; play counts exclude it. Accuracy counts correct final decisions; $\pm$ denotes SE.

Population	Policy	Regret	Diagnostics	Suboptimal public plays	Accuracy
P^+	Decision + LR	$3.401 \pm .318$	119.16	0.00	100/100
P^+	Greedy	$.763 \pm .082$	0.00	0.00	100/100
P^-	Decision + LR	$.365 \pm .017$	3.58	0.00	100/100
P^-	Greedy	13.353 ± 2.902	0.00	486.18	84/100

8 Discussion

Signaling changes learnability through equilibrium observations. Public pooling can conceal differences that change the optimum; targeting separates them in the canonical game. Pooling relies on class-neutral responses under the stated equilibrium selection; class-dependent utilities can expose group composition. Receivers know p to isolate sender learning. Learning receivers would require history-dependent observation kernels.

Exact coefficients describe information costs, but normalized regret ratios of 1.187 and 1.177 at $T = 10^{100}$ limit finite-sample predictions. The random suites mostly supply information during exploitation; constructed conflicts expose exploration costs without guaranteeing finite-horizon gains. A general matching algorithm requires global decision-region searches and sequential likelihood control, beyond the local tests used in our nonlinear experiments.

REFERENCES

Itai Arieli and Yakov Babichenko. Private bayesian persuasion. *Journal of Economic Theory*, 182: 185–217, 2019. doi: 10.1016/j.jet.2019.04.008.

Yakov Babichenko, Inbal Talgam-Cohen, Haifeng Xu, and Konstantin Zabarnyi. Multi-channel bayesian persuasion. In *13th Innovations in Theoretical Computer Science Conference*, volume 215 of *Leibniz International Proceedings in Informatics*, pp. 11:1–11:2. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022. doi: 10.4230/LIPIcs.ITCS.2022.11.

Francesco Bacchiocchi, Matteo Bollini, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion without a clue. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2434. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8c1075ffea0abbf0bb94c9de43638e58-Abstract-Conference.html.

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014. doi: 10.1287/moor.2014.0663.

Dirk Bergemann and Stephen Morris. Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, 11(2):487–522, 2016. doi: 10.3982/TE1808.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019. doi: 10.1257/jel.20181489.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Francesco Trovò, and Nicola Gatti. Optimal rates and efficient algorithms for online Bayesian persuasion. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2164–2183. PMLR, 2023. URL https://proceedings.mlr.press/v202/bernasconi23a.html.

Aymeric Capitaine, Maxime Haddouche, Eric Moulines, Michael I. Jordan, Etienne Boursier, and Alain Oliviero Durmus. Online decision-focused learning. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10010611.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ba5451d3c91a0f982f103cdbe249bc78-Abstract.html.

Matteo Castiglioni, Alberto Marchesi, Andrea Celli, and Nicola Gatti. Multi-receiver online bayesian persuasion. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1314–1323. PMLR, 2021. URL https://proceedings.mlr.press/v139/castiglioni21a.html.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Regret minimization in online bayesian persuasion: Handling adversarial receiver's types under full and partial feedback models. *Artificial Intelligence*, 314:103821, 2023. doi: 10.1016/j.artint.2022.103821.

Siyu Chen, Jibang Wu, Yifan Wu, and Zhuoran Yang. Learning to incentivize information acquisition: Proper scoring rules meet principal-agent model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5194–5218. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23ah.html.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3): 755–770, 1959. doi: 10.1214/aoms/1177706205.

Richard Combes, Stefan Magureanu, and Alexandre Proutière. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1763–1771, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/e19347e1c3ca0c0b97de5fb3b690855a-Abstract.html.

G. Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262, 1952. doi: 10.1214/aoms/1177729442.

Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, pp. 5–22, 2014. doi: 10.1145/2600057.2602897.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027. PMLR, 2016. URL https://proceedings.mlr.press/v49/garivier16a.html.

Todd L. Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743, 1997. doi: 10.1137/S0363012994275440.

Manjesh Hanawal, Csaba Szepesvári, and Venkatesh Saligrama. Sequential learning without feedback. *arXiv preprint arXiv:1610.05394*, 2016. URL https://arxiv.org/abs/1610.05394.

Moritz Hardt and Celestine Mendler-Dünner. Performative prediction: Past and future. *Statistical Science*, 2025. doi: 10.1214/25-STS986. URL https://projecteuclid.org/journals/statistical-science/volume-40/issue-3/Performative-Prediction-Past-and-Future/10.1214/25-STS986.short.

Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, pp. 359–376. ACM, 2014. doi: 10.1145/2600057.2602880.

Safwan Hossain, Evi Micha, Yiling Chen, and Ariel D. Procaccia. Strategic classification with externalities. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/46079bebac3c5379e7a70b1f36ad6750-Abstract-Conference.html.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011. doi: 10.1257/aer.101.6.2590.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Linear partial monitoring for sequential decision making: Algorithms, regret bounds and applications. *Journal of Machine Learning Research*, 24(346):1–45, 2023. URL https://jmlr.org/papers/v24/22-1248.html.

Thomas Kleine Buening, Aadirupa Saha, Christos Dimitrakakis, and Haifeng Xu. Bandits meet mechanism design to combat clickbait in online recommendation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2ac879d1865475a7abc8dfc7a9c15c27-Abstract-Conference.html.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1792–1800, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/cd89fef7ffdd490db800357f47722b20-Abstract.html.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012, 2014. doi: 10.1086/676597.

T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. doi: 10.1016/0196-8858(85)90002-8.

Tao Lin and Yiling Chen. Generalized principal-agent problem with a learning agent. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/df3e24cc1755e332fab642871bc70a77-Abstract-Conference.html.

Tao Lin and Ce Li. Information design with unknown prior. In *16th Innovations in Theoretical Computer Science Conference*, volume 325 of *Leibniz International Proceedings in Informatics*, pp. 72:1–72:1. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. doi: 10.4230/LIPIcs.ITCS.2025.72.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pp. 565–582, 2015. doi: 10.1145/2764468.2764508.

Olivier Massicot and Cédric Langbort. Competitive comparisons of strategic information provision policies in network routing games. *IEEE Transactions on Control of Network Systems*, 9(4): 1589–1599, 2022. doi: 10.1109/TCNS.2021.3115718.

Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013. doi: 10.1214/13-AOS1144.

Sirin Nitinawarat, George K. Atia, and Venugopal V. Veeravalli. Controlled sensing for multi-hypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013. doi: 10.1109/TAC.2013.2261188.

Simone Parisi, Montaser Mohammedalamen, Alireza Kazemipour, Matthew E. Taylor, and Michael Bowling. Monitored Markov decision processes. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 1549–1557, 2024. URL https://arxiv.org/abs/2402.06819.

Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018. doi: 10.1561/2200000070.

Jonathan Shaki, Jiarui Gan, and Sarit Kraus. Bayesian persuasion with externalities: Exploiting agent types. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):14095–14102, 2025. doi: 10.1609/aaai.v39i13.33543.

Jasper Verbree and Ashish Cherukuri. Inferring the prior in routing games using public signalling, 2021. URL https://arxiv.org/abs/2109.05895.

I. Arda Vurankaya and Ufuk Topcu. Learning to persuade privately informed receivers, 2026. URL https://arxiv.org/abs/2607.28342.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938. doi: 10.1214/aoms/1177732360.

You Zu, Krishnamurthy Iyer, and Haifeng Xu. Learning to persuade on the fly: Robustness against ignorance. *Operations Research*, 73(1):194–208, 2025. doi: 10.1287/opre.2021.0529.

APPENDIX CONTENTS

Notation **14**

A Proofs of Theorems 1 and 2 **15**

A.1 Proof of Theorem 1 15

A.2 Proof of Theorem 2 15

A.3 Decision boundaries on pooling fibers 17

B Instance-Dependent Regret Lower Bound and Region-Wise Form **17**

B.1 Proof of Theorem 3 17

B.2 Exact region-wise form and zero-cost information 19

B.3 Canonical specialization 21

C Canonical $G = 2, K = 2$ Derivations **21**

C.1 Model and Bayesian best responses 21

C.2 Bayesian equilibrium and induced observations 22

C.3 Sender values, confusing parameters, and asymptotic regret 23

C.4 Public observational equivalence and linear regret 25

C.5 A rank-deficient architecture that resolves decisions 26

C.6 Stability under parameter perturbations 26

C.7 Canonical simulations 29

D Asymptotically Optimal Canonical Policy **31**

D.1 Asymptotic optimality of batched sign testing 31

E Canonical Boundary Proof **35**

E.1 Proof of Theorem 5 35

F The $G = 4, K = 3$ Benchmark **36**

F.1 Model, equilibrium, and experiment library 36

F.2 Population settings and observation kernels 37

F.3 Controlled test with known candidate populations 38

F.4 Decision-focused sampling with likelihood-ratio tests 39

F.5 A positive exploration-cost lower bound 41

F.6 Sensitivity to congestion strength 42

G AI Use Statement **44**

NOTATION

The table collects the main symbols. Symbols specific to the canonical game or nonlinear benchmark are identified in the last group.

Symbol	Meaning
Population, signals, and observations	
$G, K; g, k$	Numbers of observable classes and latent types; their respective indices.
w_g, p_g, p_{gk}	Class mass, class-specific type distribution, and fraction of type k within class g.
$p, \bar{p}$	Population $(p_1, \ldots, p_G)$ and pooled composition $\bar{p} = \sum_g w_g p_g$.
θ	Free population coordinates: $\theta_{gk} = p_{gk}$ for $k < K$; $p_{gK} = 1 - \sum_{k<K} \theta_{gk}$.
Γ, γ, Q_γ	Finite experiment library, an experiment, and its state-dependent message kernel.
ω_t, ν, M_t	State, known state prior, and message profile at round t.
$\mathbf{a}_t, L_t, Z_t$	Equilibrium actions, induced aggregate loads, and observed aggregate feedback.
$u_k(a, L, \omega)$	Receiver utility for type k, action a, aggregate load L, and state ω.
O_t, H_{t-1}	Transcript $O_t = (\omega_t, M_t, Z_t)$ and history $H_{t-1} = (\gamma_s, O_s)_{s<t}$.
$\mathsf{K}_p^\gamma, k_p^\gamma$	Complete transcript law under population p and experiment γ, and its density.
π, T	Adaptive experiment-selection policy and time horizon.
Decisions, identifiability, and learning costs	
$V_p(\gamma), V_p^\star$	Expected sender value and its maximum over the specified library.
$\Gamma^\star(p), S_p$	Optimal experiment set; S_p abbreviates $\Gamma^\star(p)$ in Appendix B.1.
$\Delta_p(\gamma), R_p^\pi(T)$	One-step value gap $V_p^\star - V_p(\gamma)$ and expected cumulative pseudo-regret.
Δ_{fam}	Minimum total gap $\min_{\gamma\in\Gamma} \sum_i \Delta_{p_i}(\gamma)$ for an indistinguishable population family.
$N_\gamma(T), x_\gamma$	Experiment count and its asymptotic allocation at the $\log T$ scale. An optimal experiment may have $x_\gamma = \infty$.
$p \equiv_\Gamma q, [p]_\Gamma$	Equality of all transcript laws in Γ, and the resulting observational equivalence class.
$\mathcal{P}, \mathcal{Q}_\Gamma, \rho_\Gamma, s_\Gamma$	Population space, observational quotient, quotient map, and an optimal selector on the quotient when one exists.
$\mathcal{F}(\bar{p}), A, \mathcal{N}_{\mathrm{pub}}$	Pooling fiber $\{q : \sum_g w_g q_g = \bar{p}\}$ and linear pooling map $A = w^\top \otimes I_{K-1}$, with null space $\mathcal{N}_{\mathrm{pub}} = \ker A$.
$\mathrm{Alt}(p)$	Populations whose optimal sets are disjoint from $\Gamma^\star(p)$.
$\mathrm{Conf}_0(p)$	Alternatives in $\mathrm{Alt}(p)$ indistinguishable under every experiment optimal at p.
$D_{\mathrm{KL}}, D_\gamma(p, q)$	Kullback–Leibler divergence; $D_\gamma(p, q) = D_{\mathrm{KL}}(\mathsf{K}_p^\gamma \| \mathsf{K}_q^\gamma)$.
$C_{\mathrm{LB}}(p)$	Value of the KL allocation problem, giving the asymptotic regret lower-bound coefficient.
h, J_h	Mean observation contrasts and their Jacobian with respect to θ.
$\mathcal{I}_{\Gamma_0}, r_{\Gamma_0}$	Fisher information matrix for sublibrary Γ_0 and its rank.
Canonical game and nonlinear benchmark	
x, y, d	Canonical type-H fractions and decision margin $d = 2x - y - 1/2$; x here is distinct from the allocation x_γ.
N, F, P_1, P_2	No disclosure, public disclosure, and disclosure targeted to class 1 or class 2 in the canonical game.
L_0, c_γ	Canonical baseline load $20/9$ and experiment-dependent payoff offset; in the nonlinear benchmark, c_γ is a signaling cost.
$A_{g,\omega}^\gamma, X_{g,e}$	Canonical per-capita class action and nonlinear class load on service e; $X_{g,e}$ includes class mass.
τ, σ_Z	Gaussian feedback noise standard deviations in the canonical game and nonlinear benchmark, respectively.
ϑ, κ	Canonical perturbation parameters (Appendix C.6) and nonlinear congestion multiplier (Appendix F.6).
$\widehat{\theta}_t, \mathcal{I}_t$	Fitted nonlinear population coordinates and accumulated information matrix.
$\mathcal{I}_\gamma(\widehat{p}_t)$	Expected information from one additional observation under experiment γ at the fitted population.
$b_{j,t}, g_{j,t}(\gamma)$	Value-gap gradient for competitor j and estimated variance reduction from experiment γ (Appendix F.4).

A Proofs of Theorems 1 and 2

A.1 Proof of Theorem 1

We first show by induction that the history law is the same at every population in the family. Let H_t contain the sender's internal randomization, selected experiments, and observed feedback through round t. At $t = 0$, all history laws agree. Assume they agree through $t - 1$. Conditional on a common history, the adaptive policy draws the same distribution over γ_t under every p_i. For every realized $\gamma_t \in \Gamma$, mutual observational equivalence gives the same $\mathsf{K}_{p_i}^{\gamma_t}$. Integrating over the common experiment distribution gives the same law of H_t across the family. Induction proves the claim for all T.

Let $N_\gamma(T)$ count selections of γ; its law is the same at every population in the family. Then

$$\begin{aligned}\sum_{i=1}^{m} R_{p_i}^{\pi}(T) &= \sum_{\gamma\in\Gamma} \mathbb{E}[N_\gamma(T)] \sum_{i=1}^{m} \Delta_{p_i}(\gamma) \\ &\geq \Delta_{\mathrm{fam}} \sum_{\gamma\in\Gamma} \mathbb{E}[N_\gamma(T)] = \Delta_{\mathrm{fam}} T.\end{aligned}$$

Every γ is suboptimal for at least one p_i because the common intersection is empty. Finiteness of Γ makes the minimum positive. Averaging gives the maximum-regret bound.

Proof of Corollary 1. Suppose $\bigcap_{q\in[p]_\Gamma} \Gamma^\star(q) = \varnothing$. For each $\gamma \in \Gamma$, choose one $q_\gamma \in [p]_\Gamma$ for which $\gamma \notin \Gamma^\star(q_\gamma)$. Removing duplicates gives W with $|W| \leq |\Gamma|$. Every experiment has a positive gap at some member, so $\delta_W > 0$. Theorem 1 gives the stated sum-regret bound and $\max_{q\in W} R_q^\pi(T) \geq \delta_W T/|W|$. The family and constant are independent of the policy and horizon.

The quotient factorization is equivalent to the common-intersection condition: a selector on a class belongs to every member's optimal set, and a nonempty intersection supplies such a selector. For the converse construction, fix an ordering of the finite library and choose the first experiment in each intersection. No equality of the complete optimal sets is needed.

This finite family also proves the necessity statement for an entire equivalence class. If a policy had $o(T)$ regret at every fixed population in the class, the sum of its regrets over the witness family would be $o(T)$. The displayed linear bound rules this out. The argument allows the regret rates to differ between populations; it does not require a uniform convergence rate over the class.

A.2 Proof of Theorem 2

Public pooling. Fix a public message m, its common posterior μ_m, and aggregate load L. Because an individual receiver has zero mass, it takes L as given. A type-k receiver in every class solves the same class-neutral problem

$$a_k(\mu_m, L) \in \arg\max_a \mathbb{E}_{\omega\sim\mu_m}[u_k(a, L, \omega)].$$

Uniqueness or the assumed class-invariant selection rule makes this response independent of g. Aggregate behavior can therefore be written

$$L(m) = \sum_{g,k} w_g p_{gk} a_k(\mu_m, L(m)) = \sum_k \bar{p}_k a_k(\mu_m, L(m)).$$

Thus populations with the same $\bar{p}$ define the same aggregate fixed-point problem. If its solution is unique, their equilibrium loads coincide. Otherwise, the stipulated rule selects the same solution from this identical problem. The state/message law is population independent, and the same observation channel is applied to identical aggregate behavior. The complete public transcript law therefore factors through $\bar{p}$.

Regularity for Fisher diagnostics. For Fisher ranks, assume differentiable transcript densities k_p^γ with a common dominating measure, square-integrable mean-zero scores, and valid differentiation under expectation. Use coordinates $\theta_{gk} = p_{gk}$, $k < K$. For private identification, stack fixed

contrasts of independent observations from Γ_0 into a finite-second-moment statistic S, with mean $h(\theta)$ and Jacobian $J_h = \nabla h(\theta)$. Define

$$\mathcal{I}_{\Gamma_0}(p) = \sum_{\gamma \in \Gamma_0} \mathbb{E}_p\big[\nabla \log k_p^\gamma(O) \nabla \log k_p^\gamma(O)^\top\big], \qquad r_{\Gamma_0}(p) = \operatorname{rank} \mathcal{I}_{\Gamma_0}(p).$$

The following argument proves the rank criterion using the score covariance identity.

Information rank. In tangent coordinates, $\bar{\theta} = A\theta$ with $A = w^\top \otimes I_{K-1}$. Since all class masses are positive, A has row rank $K-1$ and $\dim \ker A = (G-1)(K-1)$. Applying the chain rule to the log-density gives a public Fisher matrix of the form $A^\top \widetilde{\mathcal{I}} A$, where $\widetilde{\mathcal{I}}$ is the information in the pooled coordinates. Its rank is at most $K-1$ and its null space has dimension at least $(G-1)(K-1)$.

For the private criterion, collect the independent observations from Γ_0 into O and denote their joint density by f_θ. The joint score is $s_\theta(O) = \nabla_\theta \log f_\theta(O)$. Independence makes its Fisher information equal to the sum $\mathcal{I}_{\Gamma_0}$. Differentiating $\int f_\theta(o)\, do = 1$ gives $\mathbb{E}_\theta[s_\theta] = 0$. Since the contrast statistic $S(O)$ does not depend on the unknown parameter, differentiation under the integral gives

$$J_h(\theta) = \int S(o) \nabla_\theta f_\theta(o)^\top \, do = \mathbb{E}_\theta[S s_\theta^\top].$$

Centering the score yields the score covariance identity

$$J_h(\theta) = \operatorname{Cov}_\theta(S, s_\theta).$$

If the Fisher information had a nonzero null vector v, then

$$0 = v^\top \mathcal{I}_{\Gamma_0} v = \mathbb{E}_\theta[(v^\top s_\theta)^2].$$

The nonnegative squared score must therefore vanish almost surely. The covariance identity would then give $J_h v = 0$, contradicting full column rank. Thus any direction that changes the contrast mean must carry Fisher information, and a full column rank J_h gives full Fisher rank on the population tangent space.

Local identification. Differentiability of h converts full Jacobian rank into separation of nearby populations. At a point θ_0 where $J_h(\theta_0)$ has full column rank, let $s_{\min} > 0$ be its smallest singular value. The first-order expansion gives

$$h(\theta_0 + u) - h(\theta_0) = J_h(\theta_0)u + o(\|u\|), \qquad \|h(\theta_0 + u) - h(\theta_0)\| \geq \tfrac{s_{\min}}{2}\|u\|$$

for sufficiently small feasible u. Equal transcript laws imply equal contrast means, so no distinct nearby population is observationally equivalent to θ_0.

The lower bound follows because $\|J_h(\theta_0)u\| \geq s_{\min}\|u\|$, while differentiability makes the remainder smaller than $s_{\min}\|u\|/2$ in a sufficiently small neighborhood. This step establishes separation from θ_0 using observable contrast means; populations farther away require a separate comparison of their transcript laws.

Targeted contrasts. For the primitive construction, set the externality scale to $\alpha = 0$. Let $\mathbf{v}_{gk} \in \mathbb{R}^{m_g}$ be the stacked feedback contrast per unit mass of type k in class g. A targeted message to class g leaves every other class's response unchanged. In simplex coordinates $p_{gK} = 1 - \sum_{k<K} \theta_{gk}$, its contrast mean has derivative

$$H_g = w_g\big[\mathbf{v}_{g1} - \mathbf{v}_{gK}, \ldots, \mathbf{v}_{g,K-1} - \mathbf{v}_{gK}\big].$$

At zero externality, the mean contrast for class g is $w_g \sum_k p_{gk}\mathbf{v}_{gk}$, so its derivative with respect to another class's composition is zero. After grouping contrast rows by target class, $J_h(0) = \operatorname{diag}(H_1, \ldots, H_G)$. Affine independence of the K vectors $\mathbf{v}_{gk}$ means precisely that their $K-1$ differences from $\mathbf{v}_{gK}$ are linearly independent. Thus $\operatorname{rank} H_g = K-1$ for every g gives total rank $G(K-1)$.

For a smooth entropy-regularized equilibrium with a nonsingular Hessian, the implicit-function theorem gives a smooth equilibrium branch and hence continuous $J_h(\alpha)$. Choose a square minor of size $G(K-1)$ whose determinant is nonzero at $\alpha = 0$. Continuity keeps that determinant nonzero for sufficiently small externalities, even though the Jacobian need no longer be block diagonal.

A.3 Decision boundaries on pooling fibers

Write populations in free simplex coordinates θ, with $p(\theta)$ reconstructing the last type fraction in each class. The pooling null space $\ker A$ consists of exact unobservable directions: $A(\theta + tv) = A\theta$ whenever $Av = 0$ and $\theta + tv$ is feasible. Consequently, the pooling fibers lie inside observational equivalence classes. A common public optimum on every fiber is necessary for sublinear regret throughout each public equivalence class. If the public kernel map is injective in $\bar{p}$, the fibers are exactly these classes, so the fiber condition also suffices for the existence of a common optimum in each class. With unique optima, the optimum must be constant along each fiber.

Suppose i, j are the only optimal experiments in the public library at an interior population $p(\theta_0)$. Assume their value difference is differentiable and all experiment values are continuous. Let $f_{ij}(\theta) = V_{p(\theta)}(i) - V_{p(\theta)}(j)$ and take a tangent direction $v \in \ker A$ with $\nabla_\theta f_{ij}(\theta_0)^\top v \neq 0$. Interiority makes $\theta_\pm = \theta_0 \pm \varepsilon v$ feasible for sufficiently small $\varepsilon > 0$. The populations $p(\theta_+), p(\theta_-)$ have the same pooled composition since $Av = 0$.

At the tie, $f_{ij}(\theta_0) = 0$. A first-order Taylor expansion gives

$$f_{ij}(\theta_\pm) = \pm\varepsilon \nabla_\theta f_{ij}(\theta_0)^\top v + o(\varepsilon),$$

so i and j exchange their ordering across the two populations. Every other experiment has a strictly smaller value at $p(\theta_0)$. Finiteness of the library and continuity preserve these gaps for sufficiently small ε. Consequently, $p(\theta_+)$ and $p(\theta_-)$ have different unique public optima despite identical public transcript laws. Applying Theorem 1 to this fixed pair gives linear regret.

Taking the contrapositive, a common optimum on the fiber requires $\ker A \subseteq \ker(\nabla_\theta f_{ij}(\theta_0)^\top)$ at such a tie. This is a local first-order requirement, not a sufficient condition for a common optimum on the entire fiber. Vanishing derivatives at one point can leave higher-order sign changes. More generally, a zero Fisher direction need not preserve the exact observation law, whereas the linear pooling map gives exact equivalence along feasible fiber segments.

B Instance-Dependent Regret Lower Bound and Region-Wise Form

We prove Theorem 3 by combining the adaptive KL chain rule with a binary testing bound. We then express its information constraints over decision regions and remove the information supplied by optimal experiments. This reduction identifies when the logarithmic regret coefficient is zero and supplies the allocation problem solved in Appendix C.

B.1 Proof of Theorem 3

Let Θ denote the population space. Under the protocol of Section 3, the complete one-round kernel is

$$\mathsf{K}_p^\gamma = \mathcal{L}_p(O_t \mid \gamma_t = \gamma), \qquad O_t = (\omega_t, M_t, Z_t). \tag{B.1}$$

The policy selects γ_t before observing ω_t. For a fixed horizon, write $\pi = \pi^{(T)}$; its randomized action kernel is population independent at any given history. Any persistent random seed is included in that history. With $N_\gamma(T)$ the experiment count and $\Delta_p(\gamma)$ its regret gap,

$$R_p^\pi(T) = \sum_{\gamma \in \Gamma} \Delta_p(\gamma) \mathbb{E}_p^\pi N_\gamma(T). \tag{B.2}$$

Uniform efficiency means

$$R_q^{\pi^{(T)}}(T) = o(T^a) \quad \text{for every fixed } q \in \Theta \text{ and } a > 0. \tag{B.3}$$

Write the finite one-round divergences as

$$D_\gamma(p, q) := D_{\mathrm{KL}}(\mathsf{K}_p^\gamma \| \mathsf{K}_q^\gamma) < \infty. \tag{B.4}$$

The KL chain rule decomposes history divergence into conditional one-round divergences. The policy kernel is the same under both populations at a given history, so only the observation kernels contribute.

Lemma 1 (Adaptive transcript KL identity). *For every adaptive randomized policy π and $p, q \in \Theta$,*

$$D_{\mathrm{KL}}\big(\mathbb{P}_p^\pi(H_T)\|\mathbb{P}_q^\pi(H_T)\big) = \sum_{\gamma\in\Gamma} \mathbb{E}_p^\pi[N_\gamma(T)]D_\gamma(p,q). \tag{B.5}$$

Proof. Conditionally on a realized history h_{t-1}, the law of (γ_t, O_t) under population p is

$$\pi_t(d\gamma \mid h_{t-1})\,\mathsf{K}_p^\gamma(do),$$

and under q it is the same policy kernel followed by K_q^γ. The conditional KL of the two policy kernels is zero: the algorithm does not know which population generated the common realized history. The conditional KL increment is therefore

$$\sum_{\gamma\in\Gamma} \pi_t(\gamma \mid h_{t-1})D_\gamma(p,q).$$

The KL chain rule and expectation under $\mathbb{P}_p^\pi$ give

$$D_{\mathrm{KL}}(\mathbb{P}_p^\pi(H_T)\|\mathbb{P}_q^\pi(H_T)) = \sum_{t=1}^{T}\sum_{\gamma\in\Gamma} \mathbb{P}_p^\pi(\gamma_t=\gamma)D_\gamma(p,q),$$

which is (B.5). A persistent initial random seed has the same distribution under p and q and contributes zero KL. Conditional on this seed, the argument is unchanged. □

The testing event. To distinguish decisions when optima need not be unique, let

$$S_p := \Gamma^\star(p), \qquad \mathrm{Alt}(p) = \{q \in \Theta : \Gamma^\star(q) \cap S_p = \varnothing\}. \tag{B.6}$$

For fixed $q \in \mathrm{Alt}(p)$, finiteness of Γ gives the strict gaps

$$\delta_p = \min_{\gamma\notin S_p} \Delta_p(\gamma) > 0, \qquad \delta_{q,p} = \min_{\gamma\in S_p} \Delta_q(\gamma) > 0. \tag{B.7}$$

To lower-bound this history KL, we reduce decision learning to a binary test. Uniform efficiency and Markov's inequality control the probabilities of playing S_p for most rounds. KL data processing then converts their separation into the logarithmic bound.

Lemma 2 (Decision-separating event). *If $\{\pi^{(T)}\}$ is uniformly efficient and $q \in \mathrm{Alt}(p)$, then*

$$\liminf_{T\to\infty} \frac{D_{\mathrm{KL}}(\mathbb{P}_p^\pi(H_T)\|\mathbb{P}_q^\pi(H_T))}{\log T} \geq 1. \tag{B.8}$$

Proof. Use the event

$$E_T = \left\{ \sum_{\gamma\in S_p} N_\gamma(T) \geq T/2 \right\}.$$

Under p, the complement of this event requires at least $T/2$ plays with gap at least δ_p. Under q, the event requires at least $T/2$ plays with gap at least $\delta_{q,p}$. Markov's inequality and (B.2) give

$$\mathbb{P}_p^\pi(E_T^c) \leq \frac{2R_p^\pi(T)}{\delta_p T} \longrightarrow 0, \qquad \mathbb{P}_q^\pi(E_T) \leq \frac{2R_q^\pi(T)}{\delta_{q,p}T} = o(T^{a-1}) \tag{B.9}$$

for every $a > 0$. By data processing, the history KL is at least the binary KL between these event probabilities. Write kl for Bernoulli KL divergence. If $u_T = \mathbb{P}_p(E_T)$ and $v_T = \mathbb{P}_q(E_T)$, then

$$\mathrm{kl}(u_T, v_T) \geq u_T \log(1/v_T) - \log 2.$$

Indeed, expanding Bernoulli KL gives the term $u_T \log(1/v_T)$, a nonnegative term $(1-u_T)\log(1/(1-v_T))$, and minus the binary entropy, which is at most $\log 2$. Thus a test that accepts p with probability tending to one under p, but only polynomially small probability under q, requires logarithmic history KL. For fixed $a \in (0,1)$, (B.9) gives $\liminf \mathrm{kl}(u_T, v_T)/\log T \geq 1 - a$. Letting $a \downarrow 0$ proves (B.8). □

Combining Lemmas 1 and 2 gives, for every fixed $q \in \mathrm{Alt}(p)$,

$$\liminf_{T\to\infty} \frac{\sum_\gamma \mathbb{E}_p N_\gamma(T) D_\gamma(p,q)}{\log T} \geq 1. \tag{B.10}$$

From information to regret. Optimal experiments may be played linearly often, so their normalized counts $\mathbb{E}_p N_\gamma(T)/\log T$ need not have finite limits. For $x \in [0,\infty]^\Gamma$, use

$$\infty \cdot d = \begin{cases} \infty, & d > 0, \\ 0, & d = 0, \end{cases} \qquad \langle x, \Delta_p \rangle := \sum_{\gamma:\Delta_p(\gamma)>0} x_\gamma \Delta_p(\gamma). \tag{B.11}$$

Thus an infinite coordinate is costless only for an optimal experiment.

The constraints in (B.10) hold along every horizon sequence. We extract a common convergent subsequence of normalized counts and show that its limit is feasible for equation 1. This gives the claimed bound,

$$\boxed{\liminf_{T\to\infty} \frac{R_p^\pi(T)}{\log T} \geq C_{\mathrm{LB}}(p).} \tag{B.12}$$

Completion of the proof of Theorem 3. If the left side is infinite there is nothing to prove. Otherwise choose a subsequence T_n attaining its finite liminf and set

$$y_\gamma^{(n)} = \frac{\mathbb{E}_p N_\gamma(T_n)}{\log T_n}.$$

For every suboptimal γ, (B.2) and the positive fixed gap bound $y_\gamma^{(n)}$ along this subsequence. Because Γ is finite, a further subsequence makes every coordinate converge in the compact extended interval $[0,\infty]$. Call the limit x.

This subsequence is selected from the count vector alone, before fixing any alternative q. It is therefore the same subsequence for every information constraint. In particular, the argument does not combine separate coordinatewise lower limits, which could correspond to different horizons.

For each fixed $q \in \mathrm{Alt}(p)$, (B.10) and coordinatewise convergence imply

$$\sum_\gamma x_\gamma D_\gamma(p,q) \geq 1.$$

Indeed, every $D_\gamma(p,q)$ is finite and nonnegative: a coordinate with $x_\gamma = \infty$ either contributes ∞ when its KL is positive or zero when its KL is zero. Hence x is feasible for equation 1. The suboptimal coordinates are finite, so the finite-sum limit gives

$$\lim_{n\to\infty} \frac{R_p^\pi(T_n)}{\log T_n} = \sum_{\gamma:\Delta_p(\gamma)>0} x_\gamma \Delta_p(\gamma) \geq C_{\mathrm{LB}}(p),$$

which proves (B.12). □

If $\mathrm{Alt}(p) = \varnothing$, the zero allocation gives $C_{\mathrm{LB}}(p) = 0$. If an alternative is indistinguishable under every experiment, the allocation program is infeasible and Theorem 1 rules out uniform efficiency. The proof uses finite Γ and finite KL; compactness of Θ is not needed for the change-of-measure argument.

B.2 Exact region-wise form and zero-cost information

For each experiment $\eta \in \Gamma \setminus S_p$, define

$$\mathcal{B}_\eta(p) = \{q \in \mathrm{Alt}(p) : \eta \in \Gamma^\star(q)\}. \tag{B.13}$$

The regions may overlap and need not be closed, but they satisfy the exact finite cover

$$\mathrm{Alt}(p) = \bigcup_{\eta\in\Gamma\setminus S_p} \mathcal{B}_\eta(p). \tag{B.14}$$

An alternative with several optimal experiments belongs to each corresponding region.

For extended x, write $I_x(q) = \sum_\gamma x_\gamma D_\gamma(p,q)$ with the convention (B.11). The finite cover (B.14) makes the pointwise information constraints equivalent to

$$\inf_{q\in\mathcal{B}_\eta(p)} \sum_\gamma x_\gamma D_{\mathrm{KL}}(\mathsf{K}_p^\gamma \| \mathsf{K}_q^\gamma) \geq 1 \quad \forall \eta \notin \Gamma^\star(p). \tag{B.14a}$$

The objective remains that of equation 1. We use $\inf \varnothing = +\infty$; the inner infimum need not be attained.

Information from optimal experiments. Following the structured-bandit lower bound (Combes et al., 2017), consider alternatives indistinguishable under every optimal experiment:

$$\mathrm{Conf}_0(p) = \{q \in \mathrm{Alt}(p) : D_\gamma(p,q) = 0 \text{ for every } \gamma \in S_p\}. \tag{B.15}$$

Set all optimal coordinates to infinity at zero regret cost. Every alternative outside $\mathrm{Conf}_0(p)$ is then separated; only the residual constraints remain. This gives a semi-infinite linear program: finitely many allocation variables satisfy constraints indexed by the confusing parameters,

$$C_{\mathrm{LB}}(p) = \inf_{z \in \mathbb{R}_+^{\Gamma \setminus S_p}} \left\{ \sum_{\gamma \notin S_p} z_\gamma \Delta_p(\gamma) : \sum_{\gamma \notin S_p} z_\gamma D_\gamma(p,q) \geq 1,\ \forall q \in \mathrm{Conf}_0(p) \right\}. \tag{B.16}$$

When $\Gamma^\star(p)$ is not a singleton, this lower bound need not be tight: $C_{\mathrm{LB}}(p) = 0$ does not imply the existence of a uniformly efficient policy with $o(\log T)$ regret at p. The infinite coordinates represent counts that grow faster than $\log T$, such as the order-T observations collected during optimal play. They do not assert that a finite sample separates all alternatives at once. An alternative in $\mathrm{Conf}_0(p)$ has zero KL under every such observation, so only suboptimal experiments can satisfy its constraint.

Given a feasible z in (B.16), set $x_\gamma = \infty$ for every $\gamma \in S_p$ and $x_\gamma = z_\gamma$ otherwise. Every nonresidual alternative is then separated at infinite $\log T$-scale rate, while the residual constraints are supplied by z. Conversely, a finite-cost feasible x for equation 1 has finite suboptimal coordinates. At a residual alternative, its optimal coordinates contribute zero. Its suboptimal part is therefore feasible for (B.16). This proves equality.

Region-wise, one may equivalently use

$$\mathcal{B}_\eta^0(p) = \mathcal{B}_\eta(p) \cap \mathrm{Conf}_0(p)$$

in (B.16). These regions preserve the allocation problem when the optimum is set-valued.

Proposition 1 (Zero lower-bound coefficient)**.** *Assume all relevant KL divergences are finite. Then*

$$\boxed{C_{\mathrm{LB}}(p) = 0 \quad \Longleftrightarrow \quad \mathrm{Conf}_0(p) = \varnothing.} \tag{B.17}$$

Equivalently, there exists a full-support $\lambda \in \Delta(S_p)$ such that

$$\sum_{\gamma \in S_p} \lambda_\gamma D_\gamma(p,q) > 0 \quad \textit{for every fixed } q \in \mathrm{Alt}(p). \tag{B.18}$$

The positivity in (B.18) is pointwise and need not be uniform over an open decision region.

Proof. If $\mathrm{Conf}_0(p)$ is empty, (B.16) has no constraints and value zero; equivalently, let the normalized allocation diverge for every optimal experiment in equation 1. Conversely, fix $q \in \mathrm{Conf}_0(p)$ and define

$$M(p,q) = \max_{\gamma \notin S_p} \frac{D_\gamma(p,q)}{\Delta_p(\gamma)}.$$

Finiteness of the library, finite KL, and strictly positive suboptimal gaps make $M(p,q)$ finite. If $M(p,q) = 0$, no experiment distinguishes this alternative and the allocation is infeasible. Otherwise, every feasible z in (B.16) obeys

$$1 \leq \sum_{\gamma \notin S_p} z_\gamma D_\gamma(p,q) \leq M(p,q) \sum_{\gamma \notin S_p} z_\gamma \Delta_p(\gamma).$$

Hence $C_{\mathrm{LB}}(p) \geq 1/M(p,q) > 0$, proving (B.17). For finite S_p, a full-support λ has positive weighted KL at a fixed q exactly when some optimal experiment has positive KL there. This proves (B.18). □

Finite allocations and uniform separation. The infinite optimal coordinates in the lower-bound program separate each fixed alternative. A finite zero-cost allocation exists exactly when some $\lambda \in \Delta(S_p)$ obeys the stronger condition

$$\inf_{q \in \mathcal{B}_\eta(p)} \sum_{\gamma \in S_p} \lambda_\gamma D_\gamma(p,q) > 0 \quad \text{for every nonempty } \mathcal{B}_\eta(p). \tag{B.19}$$

If $\mathrm{Alt}(p) = \varnothing$, both sides hold vacuously: the zero allocation is feasible and any $\lambda \in \Delta(S_p)$ satisfies the empty family of conditions. Otherwise, let x be a finite feasible allocation with zero regret cost. Every suboptimal coordinate is zero, and $M = \sum_{\gamma \in S_p} x_\gamma$ is strictly positive. Define $\lambda_\gamma = x_\gamma / M$. Feasibility gives

$$\sum_{\gamma \in S_p} \lambda_\gamma D_\gamma(p, q) \geq \frac{1}{M} \qquad \text{for every } q \in \mathrm{Alt}(p),$$

which implies (B.19) on each region.

Conversely, suppose λ satisfies (B.19). The library has finitely many decision regions, so the smallest rate over the nonempty regions,

$$\kappa = \min_{\eta : \mathcal{B}_\eta(p) \neq \varnothing} \inf_{q \in \mathcal{B}_\eta(p)} \sum_{\gamma \in S_p} \lambda_\gamma D_\gamma(p, q),$$

is strictly positive. Set $x_\gamma = \lambda_\gamma / \kappa$ for $\gamma \in S_p$ and zero otherwise. The finite cover (B.14) makes this allocation feasible, and its cost is zero because it uses only optimal experiments. Condition (B.19) therefore requires a uniform positive separation. In contrast, (B.18) allows the separation to approach zero along a sequence of alternatives, as at the canonical endpoint below.

For a selected optimum $\gamma^0 \in S_p$, the closed comparison regions $\{q : V_q(\eta) \geq V_q(\gamma^0)\}$ may also include ties whose optimal sets intersect S_p. Imposing information constraints on these larger regions can therefore increase the allocation cost. The endpoint below shows why these ties matter.

B.3 Canonical specialization

Section C.3 evaluates the residual program (B.16) using the canonical equilibrium and Gaussian KL formulas, proving Corollary 2.

The endpoint $x = 3/4$. If $d > 0$ and $x = 3/4$, every population with a disjoint optimal set satisfies $x' < (y' + 1/2)/2 \leq 3/4$, hence $x' < x$ and $d_{P_1}(p, q) > 0$. Thus $\mathrm{Conf}_0(p) = \varnothing$ and the exact lower-bound constant is zero. Alternatives may still satisfy $x' \uparrow 3/4$ and $y' \uparrow 1$, so the uniform infimum of the P_1 KL is zero. The endpoint therefore separates the pointwise lower-bound program from finite uniform closed-region certification.

Appendix D proves that batched sign testing attains the canonical coefficient in expectation.

C Canonical $G = 2, K = 2$ Derivations

The derivation solves the message-contingent equilibrium directly, obtains the Gaussian transcript kernels, and evaluates the KL allocation by projecting residual alternatives onto the linear decision boundary. Continuity of the equilibrium and strict value inequalities then establish perturbation stability.

C.1 Model and Bayesian best responses

The canonical game has a nonatomic unit mass of receivers, split equally between observable classes $g \in \{1, 2\}$. The type-H fractions are $p = (x, y) \in [0, 1]^2$. With $s_L = 1$ and $s_H = 2$, the mean type sensitivities are

$$S_x = 1 + x, \qquad S_y = 1 + y, \qquad \bar{S} = 1 + \frac{x + y}{2}.$$

The state is uniform on $\{-1, +1\}$. A type-k receiver chooses $a \geq 0$ and obtains

$$u_k(a, L_\omega, \omega) = a(20 + s_k \omega) - \frac{a^2}{2} - 8 a L_\omega.$$

The four deterministic experiments have kernels

γ	$\omega = -1$	$\omega = +1$
N	(n, n)	(n, n)
F	$(-, -)$	$(+, +)$
P_1	$(-, n)$	$(+, n)$
P_2	$(n, -)$	$(n, +)$.

The two entries in each pair are the messages sent to classes 1 and 2. Experiment N reveals nothing, F reveals the state to both classes, and P_i reveals it only to class i. Each displayed pair has conditional probability one and every other pair has probability zero. The sender observes (ω, Z), where $Z = L_\omega + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \tau^2)$ independently over rounds. The message profile is already determined by the selected experiment and observed state. In particular, class loads and realized sender reward are not feedback.

Bayes' rule gives posterior means $\zeta(m) := \mathbb{E}[\omega \mid m]$, with $\zeta(+) = 1$, $\zeta(-) = -1$, and $\zeta(n) = 0$: a state label reveals the state, while n leaves the prior unchanged. For each class and experiment, the law of iterated expectations gives $\sum_m \Pr(m)\zeta(m) = \mathbb{E}[\omega] = 0$, verifying Bayes plausibility.

Strict concavity in a gives the unique pointwise response

$$a_{gk}(m) = \left[20 + s_k\zeta(m) - 8\mathbb{E}(L_\omega \mid m, g, \gamma, p)\right]_+ .$$

Here $[z]_+ = \max\{z, 0\}$ denotes the positive part. The conditional expectation includes the receiver's belief about the other class's private message and induced action. Solving this strategic fixed point gives the observation kernels used below.

C.2 Bayesian equilibrium and induced observations

We first solve the equilibrium for each experiment in Section C.1, then derive the induced observation kernels. Write

$$L_0 = \frac{20}{1+8} = \frac{20}{9}, \qquad A_{g,\omega} = \sum_k p_{gk} a_{gk,\omega}.$$

Here $A_{g,\omega}$ is the mean action within class g, before multiplication by its population mass. Since both classes have mass $1/2$, the aggregate load is $L_\omega = (A_{1,\omega} + A_{2,\omega})/2$. Keeping these two quantities separate is necessary because the sender weights classes differently even though feedback contains only their aggregate.

Set $S_1 = S_x = 1 + x$ and $S_2 = S_y = 1 + y$ for the two class-specific type means.

Substituting the posterior means into the best-response equation yields a linear candidate equilibrium. Positivity verifies the candidate, and monotonicity of aggregate best responses rules out additional equilibria with zero actions.

Lemma 3 (Unique message-contingent equilibrium)**.** *For every $(x, y) \in [0, 1]^2$, all equilibrium actions are strictly positive and the unique Bayesian equilibrium is as follows.*

For N, every receiver chooses L_0, so $A^N_{g,\omega} = L^N_\omega = L_0$. For F,

$$L^F_\omega = L_0 + \frac{\omega\bar{S}}{9}, \qquad a^F_{gk}(\omega) = L_0 + \omega\left(s_k - \frac{8\bar{S}}{9}\right), \qquad A^F_{g,\omega} = L_0 + \omega\left(S_g - \frac{8\bar{S}}{9}\right).$$

For P_i, $i \in \{1, 2\}$, class i is informed and the other class chooses L_0. With $S_1 = S_x$ and $S_2 = S_y$,

$$L^{P_i}_\omega = L_0 + \frac{\omega S_i}{10}, \qquad a^{P_i}_{ik}(\omega) = L_0 + \omega\left(s_k - \frac{4S_i}{5}\right), \qquad A^{P_i}_{i,\omega} = L_0 + \frac{\omega S_i}{5}.$$

Proof. Insert the posterior means from Section C.1 into the best response. Under N, $L = 20 - 8L$; under F, $L_\omega = 20 + \omega\bar{S} - 8L_\omega$. For P_i, the informed class conditions on the realized state. The other class always receives n, so its posterior remains the symmetric prior. Its type-dependent state term has mean zero, and all of its types use the same expected load $\bar{L} = \mathbb{E}[L_\omega]$. The aggregate fixed-point equation is therefore

$$L^{P_i}_\omega = \tfrac{1}{2}(20 + \omega S_i - 8L^{P_i}_\omega) + \tfrac{1}{2}(20 - 8\bar{L}).$$

Averaging over the prior yields $\bar{L} = L_0$. Solving these scalar equations and substituting back gives all displayed actions and loads. Under F, the largest absolute state coefficient is $\max_{s\in\{1,2\},\bar{S}\in[1,2]} |s - 8\bar{S}/9| = 10/9$. Under an informed private message it is $\max_{s\in\{1,2\},S\in[1,2]} |s - 4S/5| = 6/5$; all uninformed actions equal L_0. Hence the smallest candidate action over all experiments, types, states, and populations is

$$L_0 - \max\{0, 10/9, 6/5\} = \frac{20}{9} - \frac{6}{5} = \frac{46}{45} > 0,$$

so the positive part never binds.

It remains to exclude another, boundary equilibrium. Public experiments reduce, at each public information set, to $L = f(L)$ with f nonincreasing. Two roots would satisfy $(L - L')^2 = (L - L')(f(L) - f(L')) \leq 0$, hence coincide. For P_1, even allowing truncation, write

$$f_\omega(L) = \mathbb{E}_k[20 + s_k\omega - 8L]_+, \qquad u(\bar{L}) = [20 - 8\bar{L}]_+.$$

For two solutions, write $\delta_\omega = L_\omega - L'_\omega$, so $\bar{L} - \bar{L}' = \frac{1}{2}\sum_\omega \delta_\omega$. Subtracting their fixed-point equations gives

$$\delta_\omega = \tfrac{1}{2}\{f_\omega(L_\omega) - f_\omega(L'_\omega) + u(\bar{L}) - u(\bar{L}')\}.$$

Multiply by δ_ω and sum over ω. Each f_ω term is nonpositive by monotonicity. The remaining term is $(\bar{L} - \bar{L}')[u(\bar{L}) - u(\bar{L}')] \leq 0$. The left-hand side is $\sum_\omega \delta_\omega^2$, so both differences are zero. The proof for P_2 is identical. □

Observation kernels. Let $\phi_\tau(\cdot - \mu)$ be the Gaussian density with mean μ and variance τ^2, and let $m^\gamma(\omega_0)$ be the deterministic message profile specified by experiment γ in state ω_0. For $\omega_0 \in \{-1, +1\}$, the complete sender-transcript kernel is

$$\mathsf{K}^\gamma_{x,y}(\omega = \omega_0, M = m^\gamma(\omega_0), Z \in dz) = \frac{1}{2}\phi_\tau(z - \mu^\gamma_{\omega_0}(x, y))\, dz,$$

where

γ	$\mu^\gamma_{\omega_0}(x, y)$
N	$20/9$
F	$20/9 + \omega_0(2 + x + y)/18$
P_1	$20/9 + \omega_0(1 + x)/10$
P_2	$20/9 + \omega_0(1 + y)/10.$

These laws follow from equilibrium loads and include every sender-observed variable. The message $M = m^\gamma(\omega)$ is population independent and deterministic given (γ, ω). The map between (ω, Z) and (ω, M, Z) is bijective on the support, so their KL divergences are equal.

C.3 Sender values, confusing parameters, and asymptotic regret

To evaluate the regret allocation, we need both the sender values and the observation divergences. The sender value is

$$V_p(\gamma) = \mathbb{E}\big[\omega\{A^\gamma_{1,\omega} + \tfrac{1}{2}A^\gamma_{2,\omega}\}\big] - c_\gamma,$$

where c_γ is an experiment-dependent sender payoff offset and $(c_N, c_F, c_{P_1}, c_{P_2}) = (1/4, 1/2, 3/20, 0)$. The available policy set is $\Gamma = \{N, F, P_1, P_2\}$ and $d = 2x - y - 1/2$.

The offsets determine the decision regions. Before choosing their values, the equilibrium gives

$$V(F) - V(N) = \frac{1 + 2x - y}{6} - (c_F - c_N), \qquad V(P_1) - V(P_2) = \frac{1 + 2x - y}{10} - (c_{P_1} - c_{P_2}).$$

Thus $c_F - c_N = 1/4$ and $c_{P_1} - c_{P_2} = 3/20$ place both boundaries at $d = 0$. Starting from $(0, 1/4, 3/20, 0)$, the common $1/4$ increment to the public offsets makes N, F strictly dominated throughout the full library. This increment cancels in public-oracle regret but changes comparisons between public and private experiments, including their regret per unit information. Adding one common constant to all four offsets leaves every value gap unchanged. Section C.6 shows that small parameter perturbations preserve the three learning-cost regions; the exact constants and threshold $3/4$ depend on the calibration.

Substitution of the equilibrium class actions into the sender objective gives the values below. Because the state is symmetric, only the state-dependent slope of each action contributes.

The values are

$$V(N) = -\frac{1}{4}, \qquad V(F) = -\frac{1}{4} + \frac{d}{6}, \qquad V(P_1) = \frac{1 + 4x}{20}, \qquad V(P_2) = \frac{1 + y}{10}.$$

Thus $V(F) - V(N) = d/6$ and $V(P_1) - V(P_2) = d/10$. Over the public family the optimum is F for $d > 0$, N for $d < 0$, and both on $d = 0$. Over Γ, P_1 is uniquely optimal for $d > 0$, P_2 is uniquely optimal for $d < 0$, and both are optimal on $d = 0$; N and F are strictly dominated everywhere.

Proof. Only the coefficient of ω in each class action survives $\mathbb{E}_\omega[\omega A_{g,\omega}]$. Substitution of the equilibrium class actions gives

$$V(F) = S_x - \frac{8\bar{S}}{9} + \frac{1}{2}\left(S_y - \frac{8\bar{S}}{9}\right) - \frac{1}{2} = \frac{2S_x - S_y}{6} - \frac{1}{2},$$
$$V(P_1) = \frac{S_x}{5} - \frac{3}{20}, \qquad V(P_2) = \frac{S_y}{10},$$

and $V(N) = -c_N$. Using $S_x = 1 + x$ and $S_y = 1 + y$ yields the formulas. Their differences yield the decision regions. Finally, $V(N) = -1/4$, $V(F) \le 0$, $V(P_1) \ge 1/20$, and $V(P_2) \ge 1/10$, proving strict domination of the public policies. □

We next quantify information using the Gaussian KL formula: with common variance τ^2 and mean difference h, KL is $h^2/(2\tau^2)$.

KL divergences. For $p = (x, y)$ and $p' = (x', y')$, the complete transcript has one-round divergences

γ	$D_{\mathrm{KL}}(\mathsf{K}^\gamma_p \| \mathsf{K}^\gamma_{p'})$
N	0
F	$\dfrac{[(x'-x)+(y'-y)]^2}{648\tau^2}$
P_1	$\dfrac{(x'-x)^2}{200\tau^2}$
P_2	$\dfrac{(y'-y)^2}{200\tau^2}$.

Proof. Condition on the observed state and apply the Gaussian formula. The equilibrium means in Section C.2 differ respectively by 0, $\omega(\Delta x + \Delta y)/18$, $\omega\Delta x/10$, and $\omega\Delta y/10$. Averaging over the state gives the table. □

By the residual-program reduction (B.16), only alternatives invisible under optimal play can force positive regret cost. The KL table identifies these alternatives directly.

Confusing parameters. The alternatives invisible under every optimal experiment, defined in (B.15), are

$$\mathrm{Conf}_0(x, y) = \begin{cases} \{(x, y') : 2x - \frac{1}{2} < y' \le 1\}, & d > 0,\ x < 3/4, \\ \varnothing, & d > 0,\ x \ge 3/4, \\ \{(x', y) : (y + \frac{1}{2})/2 < x' \le 1\}, & d < 0, \\ \varnothing, & d = 0. \end{cases}$$

The closest alternatives in the nonempty cases are boundary limits, not members of the sets of confusing parameters.

Proof. If $d > 0$, equality of the P_1 observation kernels requires $x' = x$, so every confusing parameter keeps this coordinate fixed. Disjoint optimal sets require $d' = 2x - y' - 1/2 < 0$, giving the first set. It is nonempty exactly when $2x - 1/2 < 1$, i.e. $x < 3/4$. If $x \ge 3/4$, the P_1 observation and the constraint $y' \le 1$ rule out every population with a disjoint optimal set, proving $C_{\mathrm{LB}} = 0$. If $d < 0$, equality of the P_2 observation kernels requires $y' = y$, and an opposite decision requires $x' > (y + 1/2)/2$. This interval is always nonempty because its lower endpoint is at most $3/4$. At $d = 0$, the true optimal set is $\{P_1, P_2\}$ and every population's optimal set intersects it, so the disjoint-optimum condition itself makes the set of confusing parameters empty. □

Within each residual set, all informative KL terms grow with squared distance from the true population. The closest boundary limit therefore supplies the binding constraint. The resulting one-constraint linear program selects the experiment with the smallest regret–KL ratio.

To see why mixtures cannot improve this ratio, let k_γ denote the limiting KL at that boundary and let $r = \min_{\gamma: k_\gamma > 0} \Delta_p(\gamma)/k_\gamma$, with the minimum taken over suboptimal experiments. The constraint is $\sum_\gamma z_\gamma k_\gamma \geq 1$. Every feasible allocation has

$$\sum_\gamma z_\gamma \Delta_p(\gamma) \geq r \sum_\gamma z_\gamma k_\gamma \geq r.$$

Assigning $z_\gamma = 1/k_\gamma$ to a minimizing experiment and zero to the others attains this value. The boundary itself need not be an alternative: approaching it from the strict decision region gives the infimum that every feasible allocation must satisfy.

Evaluation of the lower-bound program. Substituting these sets into (B.16) gives, for every off-boundary population,

$$C_{\mathrm{LB}}(p) = \begin{cases} 20\tau^2/d, & d > 0,\ x < 3/4, \\ 0, & d > 0,\ x \geq 3/4, \\ 80\tau^2/|d|, & d < 0. \end{cases}$$

At $d = 0$, the confusing set is empty and $C_{\mathrm{LB}}(p) = 0$.

Proof. For $d > 0, x < 3/4$, confusing parameters satisfy $x' = x$. Write a confusing parameter as $y' = y + h$, $h > d$. Both informative suboptimal KL divergences increase with h^2, so the limiting constraint $h \downarrow d$ is the hardest. At that limit,

$$\mathrm{KL}_{P_2} = \frac{d^2}{200\tau^2}, \qquad \Delta_p(P_2) = \frac{d}{10},$$

and their ratio is $20\tau^2/d$. The public diagnostic has $\mathrm{KL}_F = d^2/(648\tau^2)$. Moreover $x < 3/4$ and feasibility imply $0 < d < 1$, while $V(P_1) - V(F) > 11/60$; hence

$$\frac{648\tau^2\{V(P_1) - V(F)\}}{d^2} > \frac{20\tau^2}{d}.$$

N has zero KL. The KL allocation problem is therefore a one-constraint linear program whose cheapest information source is P_2; mixing with F cannot improve the ratio. The case with no confusing parameters has no constraints and cost zero.

For $d < 0$, confusing parameters satisfy $y' = y$ and the boundary limit is $x_\partial = (y + 1/2)/2$, a displacement $|d|/2$. Thus

$$\mathrm{KL}_{P_1} = \frac{d^2}{800\tau^2}, \qquad \Delta_p(P_1) = \frac{|d|}{10},$$

giving $80\tau^2/|d|$. At the same limit $\mathrm{KL}_F = d^2/(2592\tau^2)$. Since $V(P_2) - V(F) > 7/20$ and $|d| \leq 3/2$, its regret–information ratio is again strictly larger. These comparisons solve the global KL program. □

The negative-side coefficient is four times larger at equal $|d|$. A change in x moves the decision margin twice as fast as the same change in y, so the closest alternative requires only half the coordinate displacement. Its Gaussian KL is four times smaller, while the regret gap remains $|d|/10$.

C.4 Public observational equivalence and linear regret

The kernels in Section C.2 show that N is population independent and F depends only on $x + y$. Its conditional means $20/9 \pm (2 + x + y)/18$ identify that sum. Thus the public observational fibers are exactly

$$\mathcal{F}_t = \{(u, t - u) : \max(0, t - 1) \leq u \leq \min(1, t)\}, \qquad t \in [0, 2],$$

with tangent direction $(1, -1)$ in the interior. Theorem 1 transfers this one-round equivalence to adaptive histories.

For the pair used in Section 6, $q^+ = (.8, .4)$ and $q^- = (.5, .7)$ both have $x + y = 1.2$, with $d^+ = .7$ and $d^- = -.2$. Since $V(F) - V(N) = d/6$, their unique public optima are F and N, respectively.

Let u_t be the probability of choosing F at round t, which is the same under both populations. Relative to their public oracles,

$$R^{\pi}_{q^+}(T) + R^{\pi}_{q^-}(T) = \sum_{t=1}^{T} \left[(1-u_t)\frac{7}{60} + u_t \frac{1}{30} \right] \geq \frac{T}{30}.$$

This is Theorem 1 with $\Delta_{\mathrm{fam}} = 1/30$.

C.5 A RANK-DEFICIENT ARCHITECTURE THAT RESOLVES DECISIONS

This analytic control keeps the canonical receiver game, population domain $[0,1]^2$, public library $\Gamma_{\mathrm{pub}} = \{N, F\}$, and aggregate Gaussian feedback. It changes the sender objective to depend on pooled load:

$$V^{\mathrm{ctrl}}_{x,y}(\gamma) = \mathbb{E}[\omega L^{\gamma}_{\omega}] - c^{\mathrm{ctrl}}_{\gamma}, \qquad c^{\mathrm{ctrl}}_{N} = 0, \qquad c^{\mathrm{ctrl}}_{F} = \frac{1}{6}.$$

Changing sender value does not change receiver utility or the equilibrium induced by a fixed experiment. Thus the loads derived in Section C.2 remain valid. With $s = x + y$, they give

$$V^{\mathrm{ctrl}}_{x,y}(N) = 0, \qquad V^{\mathrm{ctrl}}_{x,y}(F) = \frac{2+s}{18} - \frac{1}{6} = \frac{s-1}{18}.$$

Both experiments are uniquely optimal on nonempty open subsets of the population square: N for $s < 1$ and F for $s > 1$. At $s = 1$ they tie. The decision therefore changes across the domain even though one population direction is never observed.

The no-disclosure law is independent of (x, y). Under F, the conditional Gaussian mean is $L_0 + \omega(2+s)/18$, which identifies s from the transcript law. Hence two populations are observationally equivalent for the entire public library exactly when they have the same sum. Every such fiber has the same optimal set,

$$\Gamma^{\star,\mathrm{ctrl}}(x,y) = \begin{cases} \{N\}, & x + y < 1, \\ \{N, F\}, & x + y = 1, \\ \{F\}, & x + y > 1. \end{cases}$$

This proves decision resolution over the full square, including fibers containing distinct populations and fibers on the decision boundary.

For the conditional mean under F, both coordinate derivatives are $\omega/18$. Averaging its Gaussian Fisher information over the observed state and adding the zero information from N gives

$$\mathcal{I}^{\mathrm{ctrl}}_{\Gamma_{\mathrm{pub}}}(x,y) = \frac{1}{324\tau^2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad r_{\Gamma_{\mathrm{pub}}} = 1 < 2 = G(K-1).$$

The null direction $(1, -1)$ is tangent to the decision boundary $x + y = 1$: the value difference has gradient $(1, 1)/18$. In the original class-sensitive objective, the same public feedback hides a direction transverse to $2x - y - 1/2 = 0$. The control therefore holds the observation channel fixed and changes whether its hidden direction matters for the decision.

For concrete fibers, $(.1, .5)$ and $(.5, .1)$ both have N as their unique optimum; $(.5, .9)$ and $(.9, .5)$ both have F as their unique optimum. The respective F-minus-N gaps are $-1/45$ and $1/45$. Populations $(.2, .8)$ and $(.8, .2)$ lie on a tie fiber. All three fibers retain the same rank-one information matrix. These are exact consequences of the displayed equilibrium and value formulas, rather than estimates from the nonlinear benchmark.

C.6 STABILITY UNDER PARAMETER PERTURBATIONS

Keep the state prior, the four kernels N, F, P_1, P_2, class-neutral quadratic utility, and aggregate Gaussian feedback fixed in form, but let

$$\vartheta = (b, \lambda, s_L, s_H, w, r_1, r_2, c_N, c_F, c_1, c_2, \tau)$$

vary, where $w \in (0,1)$ is the mass of class 1, $s_H > s_L$, $r_1, r_2, \lambda, \tau > 0$, and

$$u_k(a, L, \omega) = a(b + s_k\omega) - \frac{a^2}{2} - \lambda a L, \qquad V_p(\gamma) = \mathbb{E}[\omega(r_1 A_1 + r_2 A_2)] - c_{\gamma}.$$

The canonical point is

$$\vartheta_0 = (20, 8, 1, 2, 1/2, 1, 1/2, 1/4, 1/2, 3/20, 0, \tau_0)$$

for any fixed $\tau_0 > 0$. Perturbations stay inside this structural family; in particular, they do not introduce class-dependent receiver utility or change the signal architecture.

The proof uses continuity on the compact population square to preserve strict action and value margins. Explicit best-response formulas track the affine decision boundary and its intersection with $y = 1$, which determines the threshold for $C_{\mathrm{LB}} = 0$.

Proposition 2 (Stability of the three regimes)**.** *There is a relative open neighborhood $\mathcal{U}$ of ϑ_0 such that, for every $\vartheta \in \mathcal{U}$:*

1. *the message-contingent equilibrium is unique and all actions are strictly positive for every $(x, y) \in [0, 1]^2$;*
2. *N, F remain public-pooling experiments and are strictly dominated by P_1, P_2;*
3. *the private optimum is P_1 or P_2, separated by a regular affine curve $D_\vartheta(x, y) = 0$ with nonzero normal; and*
4. *there is a continuous threshold $x_{\mathrm{free}}(\vartheta) \in (0, 1)$, with $x_{\mathrm{free}}(\vartheta_0) = 3/4$, such that*

region	*lower-bound coefficient*
$D_\vartheta > 0,\ x < x_{\mathrm{free}}(\vartheta)$	$C_{\mathrm{LB}} > 0,$
$D_\vartheta > 0,\ x \geq x_{\mathrm{free}}(\vartheta)$	$C_{\mathrm{LB}} = 0,$
$D_\vartheta < 0$	$C_{\mathrm{LB}} > 0.$

At $D_\vartheta = 0$, both private policies are optimal and $C_{\mathrm{LB}} = 0$ pointwise. The proposition preserves the qualitative regimes, not the canonical closed-form constants or the numerical location $3/4$.

Proof. Let $\delta = s_H - s_L$, $S_1 = s_L + \delta x$, $S_2 = s_L + \delta y$, and $L_0 = b/(1 + \lambda)$. Direct solution of the message-contingent best responses gives

γ	$L_\omega^\gamma - L_0$
F	$\dfrac{\omega\{wS_1 + (1-w)S_2\}}{1+\lambda},$
P_1	$\dfrac{\omega w S_1}{1 + w\lambda},$
P_2	$\dfrac{\omega(1-w)S_2}{1 + (1-w)\lambda}.$

The informed class-action slopes under P_1, P_2 are respectively $S_1/(1 + w\lambda)$ and $S_2/(1 + (1 - w)\lambda)$; the uninformed class has zero state slope. Hence

$$V_1 = A_\vartheta S_1 - c_1, \qquad A_\vartheta = \frac{r_1}{1 + w\lambda},$$
$$V_2 = B_\vartheta S_2 - c_2, \qquad B_\vartheta = \frac{r_2}{1 + (1-w)\lambda},$$

and therefore

$$D_\vartheta(x, y) = V_1 - V_2 = A_\vartheta(s_L + \delta x) - B_\vartheta(s_L + \delta y) - (c_1 - c_2). \tag{R.1}$$

Because $A_\vartheta, B_\vartheta, \delta > 0$, this boundary has a nonzero normal throughout the perturbation family. Written as $y = h_\vartheta(x)$, it is

$$h_\vartheta(x) = \frac{A_\vartheta(s_L + \delta x) - B_\vartheta s_L - (c_1 - c_2)}{B_\vartheta \delta}.$$

Its intersection with $y = 1$ occurs at

$$x_{\mathrm{free}}(\vartheta) = \frac{B_\vartheta(s_L + \delta) + (c_1 - c_2) - A_\vartheta s_L}{A_\vartheta \delta}, \tag{R.2}$$

which is continuous and equals $3/4$ at ϑ_0. At the canonical point,

$$D_{\vartheta_0}(0,0) = -\frac{1}{20}, \quad D_{\vartheta_0}(1,0) = \frac{3}{20}, \quad D_{\vartheta_0}(0,1) = -\frac{3}{20}, \quad D_{\vartheta_0}(1,1) = \frac{1}{20}.$$

Continuity preserves these four strict signs and $x_{\mathrm{free}}(\vartheta) \in (0,1)$ on a common neighborhood. Since D_ϑ increases with x and decreases with y, its inverse boundary intersects the interior of $[0,1]$ at every fixed $y \in [0,1]$. In particular, both private decision regions remain nonempty.

Under P_1, the transcript mean depends only on x, with nonzero coefficient $w\delta/(1+w\lambda)$. Under P_2, it depends only on y, with nonzero coefficient $(1-w)\delta/(1+(1-w)\lambda)$. If $D_\vartheta > 0$, optimal play of P_1 therefore fixes x. A population with the same x and a disjoint optimum exists exactly when $h_\vartheta(x) < 1$, giving the first two rows of the table. If $D_\vartheta < 0$, optimal P_2 fixes y; the inverse boundary is at most $x_{\mathrm{free}}(\vartheta) < 1$, so a population with the same y and a disjoint optimum always exists. Such a population is invisible under the true optimum, forcing exploration of suboptimal experiments in the finite Gaussian library and hence $C_{\mathrm{LB}} > 0$. When no residual population exists, every fixed alternative with disjoint optimal sets has positive KL under the optimal experiment. Its count normalized by $\log T$ may diverge at zero regret, giving $C_{\mathrm{LB}} = 0$.

It remains to justify that no other policy or equilibrium changes this geometry. At ϑ_0, the smallest candidate action is $46/45 > 0$ uniformly on the population square, and every private value exceeds every public value by at least $1/20$. The explicit linear candidate actions and their induced values are jointly continuous in (p, ϑ), with nonzero denominators near ϑ_0. Uniform continuity on the compact population square therefore permits one parameter neighborhood on which every action changes by less than $23/45$ and every private–public value gap changes by less than $1/40$. All actions remain positive and all these gaps remain positive throughout that same neighborhood. Intersecting it with the neighborhood preserving the corner signs gives a single $\mathcal{U}$ for all four conclusions.

For completeness, uniqueness also holds when a competing solution is allowed to have zero actions. Under P_1, define $f_\omega(L) = \mathbb{E}_k[b + s_k\omega - \lambda L]_+$ and $u(\bar{L}) = [b - \lambda\bar{L}]_+$. For two candidate load solutions, put $\delta_\omega = L_\omega - L'_\omega$ and $\bar{\delta} = \frac{1}{2}\sum_\omega \delta_\omega$. Subtract the fixed-point equations, multiply by δ_ω, and sum to obtain

$$\sum_\omega \delta_\omega^2 = w \sum_\omega \delta_\omega\{f_\omega(L_\omega) - f_\omega(L'_\omega)\} + 2(1-w)\bar{\delta}\{u(\bar{L}) - u(\bar{L}')\} \le 0.$$

Both terms are nonpositive because f_ω and u are nonincreasing. Thus the loads coincide, and strict concavity fixes each receiver's action. The P_2 argument interchanges the class masses; public experiments use the scalar monotonicity argument from Section C.2. Finally, on $D_\vartheta = 0$, the optimal set is $\{P_1, P_2\}$. Public policies remain dominated, so every population's optimal set intersects this set. The disjoint-optimum alternative set is empty. □

Stability of the regret separation. The same family also permits a logarithmic private-regret bound away from ties. The argument uses a linear transformation of private telemetry and the standard upper confidence bound (UCB) principle with Gaussian concentration. It establishes the regret order throughout a parameter neighborhood; the matching-coefficient construction remains the one analyzed at the canonical parameters.

Corollary 3 (Regret separation under parameter perturbations)**.** *After shrinking $\mathcal{U}$ if necessary, every $\vartheta \in \mathcal{U}$ admits a pair $q_\vartheta^+, q_\vartheta^-$, independent of T, with identical public transcript laws and distinct public optima. For some $a_\vartheta > 0$, every public learner satisfies*

$$R_{q_\vartheta^+}^{\pi_{\mathrm{pub}}}(T) + R_{q_\vartheta^-}^{\pi_{\mathrm{pub}}}(T) \ge a_\vartheta T.$$

For each known $\vartheta \in \mathcal{U}$, a population-independent policy family using only P_1, P_2 has $O_{\vartheta,p}(\log T)$ regret at every fixed population with a unique private optimum, and zero regret at private ties. The public and private regrets use their respective oracles.

Proof. Keep $q_\vartheta^+ = (.8, .4)$ and set

$$q_\vartheta^- = \left(.5,\ .4 + \frac{.3w}{1-w}\right).$$

For w near $1/2$, this pair lies in the population square and has the same weighted composition $wx + (1-w)y$. Hence its public transcript laws coincide. At ϑ_0, the public value gaps are $V_{q^+}(F) - V_{q^+}(N) = 7/60$ and $V_{q^-}(N) - V_{q^-}(F) = 1/30$. The populations and values vary continuously with ϑ, so both gaps remain positive on a smaller neighborhood. Taking a_ϑ to be their minimum, the common history law and the argument in Section C.4 give the displayed lower bound. Allowing the second population to vary with w keeps the pair on one pooling fiber.

For the private upper bound, write $w_1 = w$, $w_2 = 1 - w$ and use the load formulas above. Under experiment P_i, the observable transformation

$$Y_i = \frac{r_i}{w_i}\,\omega(Z - L_0) - c_i \sim \mathcal{N}\big(V_p(P_i), \varsigma_i^2\big)\,, \qquad \varsigma_i = \frac{r_i \tau}{w_i},$$

has known variance. Indeed, its mean is $r_i S_i/(1 + w_i\lambda) - c_i = V_p(P_i)$, and multiplying the independent Gaussian feedback noise by $\omega \in \{-1, 1\}$ preserves its distribution. This statistic is computed from telemetry; it does not require observing the sender's realized reward.

For horizon $T \geq 2$, sample each private experiment once. Thereafter select the larger index

$$\overline{Y}_i\big(N_i(t-1)\big) + \varsigma_i \sqrt{\frac{6 \log T}{N_i(t-1)}},$$

where $\overline{Y}_i(n)$ is the mean of the first n observations from experiment P_i. View these observations as two independent Gaussian streams. The Gaussian tail bound and a union bound over both streams and $1 \leq n \leq T$ imply, with probability at least $1 - 4/T^2$, that every sample mean differs from its value by at most the corresponding index radius. On this event, a suboptimal experiment P_i with gap $\Delta = |V_p(P_1) - V_p(P_2)| > 0$ can be selected only while its radius is at least $\Delta/2$. Its total count is therefore at most $1 + 24\varsigma_i^2 \log T/\Delta^2$. On the complement of this event, regret is at most $T\Delta$, giving

$$R_p^{\pi_\vartheta^{(T)}}(T) \leq \Delta + \frac{24\varsigma_i^2 \log T}{\Delta} + \frac{4\Delta}{T} = O_{\vartheta,p}(\log T).$$

At a tie, either private experiment incurs zero regret. Since public experiments remain strictly dominated, the private oracle also attains the full-library optimum. □

C.7 Canonical simulations

Canonical Decision + LR. After one observation from each private experiment, this policy computes the Gaussian maximum likelihood estimate (MLE) exactly over $[0, 1]^2$. It evaluates the generalized likelihood ratio (GLR) against the opposite decision half-space, intersected with that square, using threshold $\beta(t) = \log(t + 1) + 1.1 \log\log(t + e)$ at decision round t.

A second check uses normal-mixture confidence sequences obtained by mixing Gaussian exponential martingales (Howard et al., 2021). For either transformed stream in (D.5), with variance $\sigma^2 = 100\tau^2$ and local count n, the interval is the sample mean plus or minus

$$r(n, \alpha_t) = \frac{\sigma}{n}\sqrt{2(n + \rho)\left[\log(1/\alpha_t) + \frac{1}{2}\log\frac{n + \rho}{\rho}\right]}, \qquad \rho = 1,$$

intersected with $[0, 1]$. We use total error budget $\delta_t = 6(.05)/(\pi^2 t^2)$ and $\alpha_t = \delta_t/2$ for each coordinate. If the coordinate intervals are $[x_\ell, x_u]$ and $[y_\ell, y_u]$, the corresponding interval for d is $[2x_\ell - y_u - 1/2,\ 2x_u - y_\ell - 1/2]$. The learner accepts a decision when this interval excludes zero and agrees with the GLR. It updates these checks every round in the algorithm comparisons, including the zero-coefficient endpoint, and every 64 rounds in the boundary experiment after checking each of the first ten rounds. Between checks it retains the last test decision.

When no decision is accepted, the policy first checks whether $\widehat{d} > 0$ and

$$\widehat{x} \geq \frac{3}{4} - b_t, \qquad b_t = \min\left\{.05,\ 30\tau\sqrt{\frac{\log(\max\{e, t\})}{N_{P_1}(t-1)}}\right\}.$$

If so, it plays P_1. The shrinking buffer permits information collection under P_1 near the endpoint $x = 3/4$. Otherwise, the estimated KL allocation gives a target count $\beta(t)a_i(\widehat{p})$ for each private

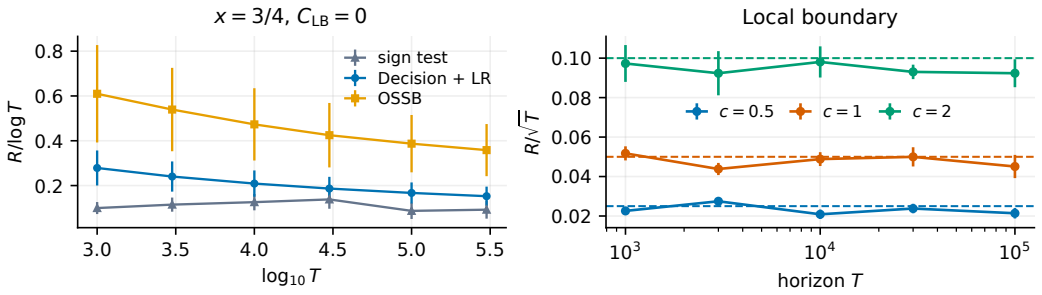


Figure 5: Canonical zero-coefficient and boundary results. Left: batched sign testing, Decision + LR, and OSSB at the $x = 3/4$ endpoint, where $C_{\mathrm{LB}} = 0$ ($\tau = .05$, 20 paired seeds). Right: $R/\sqrt{T}$ for Decision + LR, where R is mean regret over $p_{+,T}$ and $p_{-,T}$ ($\tau = .25$, 16 paired seeds). Dashed lines give $c/20$ for always choosing either private experiment. Error bars show standard errors.

experiment, where

$$(a_1(\widehat{p}), a_2(\widehat{p})) = \begin{cases} (0, 200\tau^2/\widehat{d}^2), & \widehat{d} > 0,\ \widehat{x} < 3/4, \\ (800\tau^2/\widehat{d}^2, 0), & \widehat{d} < 0, \\ (0, 0), & \text{otherwise.} \end{cases}$$

The learner selects the diagnostic experiment if its count is below both its target and $.1t$. On all remaining rounds it plays the fitted optimum. The estimate and allocation are updated every round, even when the decision tests are retained from an earlier check.

OSSB and myopic comparison. We specialize Optimal Sampling for Structured Bandits (OSSB) (Combes et al., 2017) to the two private Gaussian streams and the allocation a_i above. Sample means clipped to $[0, 1]$ give the fitted population. At round t, OSSB exploits if every count satisfies $N_{P_i}(t-1) \geq 1.05 a_i(\widehat{p}) \log(\max\{2, t\})$. Otherwise it increments the number E of non-exploitation rounds. If the smaller count is at most $.05E$, it samples that experiment to improve estimation. If not, it samples the experiment minimizing $N_{P_i}(t-1)/a_i(\widehat{p})$ among those with positive allocation. Myopic uses the same initial private observations as Decision + LR and subsequently plays its fitted optimum. Batched sign testing follows the horizon-dependent construction in Appendix D; the matching theorem applies to that policy.

Finite-horizon behavior. All methods in Figures 2 and 5 use the same P_1/P_2 Gaussian feedback and receive no population information. At $T = 3 \times 10^5$, the Decision + LR normalized regret ratios are 1.83 for $d = .8$ and 1.23 for $d = -1$; canonical OSSB reaches 2.40 and 1.05, respectively. Both methods choose the correct final decision in all 20 seeds. Myopic is omitted from the plot scale because it remains wrong in 25% of the $d = .8$ runs and incurs mean regret 6000.05. Batched sign testing is the policy analyzed in Appendix D.

An independent message-contingent equilibrium solver checks Bayes plausibility, actions, loads, values, and KL formulas on population grids. At $\tau = .25$, the batched policy's incorrect pilot decisions and cap hits remain nonnegligible at the evaluated horizons, so its unconditional regret ratios do not yet show clean convergence. In the boundary experiment, Decision + LR has $R/\sqrt{T}$ near .021, .045, and .093 for $c = .5, 1, 2$, while $R/\log T$ increases with the horizon. Here $|d| = c/\sqrt{T}$. Always choosing either private experiment incurs zero regret at one endpoint and $T|d|/10$ at the other, so the endpoint mean satisfies $R/\sqrt{T} = c/20$. The panel illustrates this boundary scale.

D Asymptotically Optimal Canonical Policy

D.1 Asymptotic optimality of batched sign testing

We construct the policy in Theorem 4 under the model of Sections C.1–C.3. The observation-noise variance $\tau^2 > 0$ is known. The policy uses the horizon, P_1, P_2, and the observed pair (ω, Z), but does not depend on the population.

Affine transformations of the two targeted observations give independent Gaussian streams for x and y. This permits sign tests for $d = 2x - y - 1/2$. A pilot identifies the likely optimum, heavy sampling estimates its coordinate, and the rival is sampled until a sign test passes. The allocation and stopping structure follows sequential experimental design and best-arm identification (Chernoff, 1959; Garivier & Kaufmann, 2016). Here the objective is expected regret over a fixed horizon, which also requires controlling the loss after an incorrect commitment. The bounds below use Gaussian tails over a finite checkpoint grid.

Put $\sigma^2 = 100\tau^2$. For all sufficiently large horizons, define

$$m_T = \left\lceil (\log T)^{2/3} \right\rceil, \qquad L_T = \left\lceil (\log T)^2 \right\rceil, \qquad \eta_T = (\log T)^{-1/4}, \qquad s_T = (\log T)^{-3/4}. \tag{D.1}$$

The slack η_T sets checkpoint spacing, while s_T allows the early stopping condition to apply at $x = 3/4$. Let $\mathcal{G}_T$ be the increasing list of distinct integers in

$$\left\{ \left\lceil m_T(1+\eta_T)^k \right\rceil : k \geq 0 \right\} \cap [m_T, L_T] \quad \text{together with } L_T, \tag{D.2}$$

and let $K_T = |\mathcal{G}_T|$. Then

$$K_T = O\Big((\log T)^{1/4} \log\log T \Big). \tag{D.3}$$

Use the deterministic threshold

$$b_T = \log(4K_T T \log T) = \log T + O(\log\log T). \tag{D.4}$$

The choices ensure $\eta_T^2 b_T \to \infty$ and $L_T s_T^2 \to \infty$, which control the grid and early stopping errors. For all sufficiently large T, $2L_T \leq T$. For the finitely many smaller horizons, always play P_1; this leaves the asymptotic statements unchanged.

These sample sizes separate three roles. The pilot uses $m_T = o(\log T)$ observations from each experiment, so its regret does not affect the leading coefficient. Sampling the likely optimum to L_T makes its estimation variance negligible compared with the variance from order-$\log T$ diagnostic observations. The geometric grid then locates the required diagnostic count with a relative error tending to zero. Step 2 below bounds the cost of choosing the wrong experiment for the L_T samples.

Batched sign testing. First sample each of P_1, P_2 to local count m_T. From these observations form

$$X_i = 10\omega_i(Z_i - 20/9) - 1 \quad (P_1), \qquad Y_i = 10\omega_i(Z_i - 20/9) - 1 \quad (P_2), \tag{D.5}$$

and $\widehat{d}_m = 2\bar{X}_{m_T} - \bar{Y}_{m_T} - 1/2$. If $\widehat{d}_m \geq 0$, sample P_1 to total count L_T; otherwise sample P_2 to total count L_T. If this selected experiment is P_1 and

$$\bar{X}_{L_T} \geq \frac{3}{4} - s_T, \tag{D.6}$$

play P_1 for the rest of the horizon.

Otherwise sample the other experiment. Whenever its total local count reaches $n \in \mathcal{G}_T$, let (n_1, n_2) be the current P_1, P_2 counts and compute

$$\widehat{d}(n_1, n_2) = 2\bar{X}_{n_1} - \bar{Y}_{n_2} - \frac{1}{2}, \qquad v(n_1, n_2) = \sigma^2 \left(\frac{4}{n_1} + \frac{1}{n_2} \right). \tag{D.7}$$

The Gaussian likelihood-ratio test for the sign of d uses the unrestricted parameter space $\mathbb{R}^2$. If

$$|\widehat{d}(n_1, n_2)| \geq \sqrt{2b_T v(n_1, n_2)}, \tag{D.8}$$

commit to P_1 for a positive sign and P_2 for a negative sign. If both streams reach L_T without crossing (D.8), commit according to the sign of $\widehat{d}(L_T, L_T)$.

The threshold in (D.8) follows from the Gaussian likelihood, rather than a separate approximation. At fixed counts (n_1, n_2), the log-likelihood loss from replacing the unrestricted estimate $(\bar{X}_{n_1}, \bar{Y}_{n_2})$ by (a, b) is

$$\frac{n_1(a - \bar{X}_{n_1})^2 + n_2(b - \bar{Y}_{n_2})^2}{2\sigma^2}.$$

The closest point with the opposite sign lies on $2a - b - 1/2 = 0$. Minimizing this weighted quadratic by a Lagrange multiplier gives loss $\widehat{d}(n_1, n_2)^2/(2v(n_1, n_2))$. Requiring it to exceed b_T is exactly (D.8). The test uses $\mathbb{R}^2$ for this minimization; the population constraint $y \le 1$ enters separately through the early stopping rule (D.6). Because the checkpoint reached is data-dependent, Step 3 controls errors by a union bound over deterministic count pairs.

We show that this policy is uniformly efficient on $[0, 1]^2$ and, for every fixed population $p = (x, y)$, satisfies

$$\boxed{R_p^{\pi^{(T)}}(T) = C_{\mathrm{LB}}(p) \log T + o(\log T).} \tag{D.9}$$

More explicitly, with $d = 2x - y - 1/2$,

$$R_p^{\pi^{(T)}}(T) = \begin{cases} \dfrac{20\tau^2}{d} \log T + o(\log T), & d > 0,\ x < 3/4, \\ o(\log T), & d > 0,\ x \ge 3/4, \\ \dfrac{80\tau^2}{|d|} \log T + o(\log T), & d < 0, \\ 0, & d = 0. \end{cases} \tag{D.10}$$

The rate is pointwise in p. At $x = 3/4$, the early stopping rule does not certify a whole neighborhood of the endpoint.

Proof of Theorem 4. We first control pilot and commitment errors. We then concentrate the diagnostic count around the KL allocation and bound the regret of exceptional paths to obtain convergence in expectation.

Step 1: Gaussian observation sequences. The equilibrium laws give, on a P_1 round,

$$Z = \frac{20}{9} + \omega \frac{1 + x}{10} + \epsilon,$$

so (D.5) satisfies $X = x + 10\omega\epsilon$. Similarly $Y = y + 10\omega\epsilon$ on a P_2 round. Each experiment is chosen before the fresh state and noise, so we can sample two independent, pre-generated sequences. Within each sequence, observations are independent and identically distributed (i.i.d.):

$$X_i \overset{\text{iid}}{\sim} \mathcal{N}(x, \sigma^2), \qquad Y_i \overset{\text{iid}}{\sim} \mathcal{N}(y, \sigma^2), \qquad \sigma^2 = 100\tau^2, \tag{D.11}$$

with the streams independent. Hence, at every deterministic count pair,

$$\widehat{d}(n_1, n_2) \sim \mathcal{N}\big(d, \sigma^2 \left(4/n_1 + 1/n_2\right)\big). \tag{D.12}$$

Step 2: pilot and additional sampling. At the pilot counts, the variance in (D.12) is $5\sigma^2/m_T$. For fixed $d \neq 0$, the Gaussian tail bound gives

$$\Pr_p\{\mathrm{sign}(\widehat{d}_m) \neq \mathrm{sign}(d)\} \le \exp\left\{-\frac{d^2 m_T}{10\sigma^2}\right\}. \tag{D.13}$$

If the pilot selects the optimum, the additional samples incur no regret. An incorrect selection uses at most L_T suboptimal rounds, whose expected contribution is bounded by

$$L_T \exp\{-c_p m_T\} = o(1) \tag{D.14}$$

for a population-dependent $c_p > 0$. The pilot itself uses exactly m_T suboptimal rounds and therefore contributes $o(\log T)$ regret.

Step 3: wrong commitment. Fix a potential checkpoint count pair and write $h = \sqrt{2b_T v(n_1, n_2)}$. If $d > 0$, a wrong crossing requires $\widehat{d} \le -h$. The Gaussian tail bound gives

$$\Pr_p\{\widehat{d} \le -h\} \le \frac{1}{2}\exp\left\{-\frac{(d+h)^2}{2v}\right\} \le \frac{1}{2}e^{-b_T}. \tag{D.15}$$

The negative case is symmetric. Reaching a count pair depends on the pilot data. A wrong crossing at that pair is contained in the unconditional Gaussian event in (D.15). A union bound over both pilot choices and all grid points gives

$$\Pr_p\{\text{wrong checkpoint commitment}\} \le K_T e^{-b_T} \le \frac{1}{4T\log T}. \tag{D.16}$$

At the cap (L_T, L_T), (D.12) has variance $5\sigma^2/L_T$, so a wrong plug-in sign has probability at most $e^{-c_p L_T}$. The early stopping rule can be wrong only when $d < 0$. Such a fixed population has $x < 3/4$. Once $s_T < (3/4 - x)/2$, its probability of an incorrect early stop is also at most $e^{-c_p L_T}$. Indeed, $d < 0$ implies $x < (y + 1/2)/2 \le 3/4$, so no fixed negative-margin population lies at $x = 3/4$. Thus the expected post-commitment regret is bounded by a constant times

$$T\left\{(4T\log T)^{-1} + 2e^{-c_p L_T}\right\} = o(1). \tag{D.17}$$

Step 4: diagnostic stopping time when $C_{\mathrm{LB}} > 0$. Fix p with $d < 0$, or $d > 0$ and $x < 3/4$. On the correct positive branch, $n_1 = L_T$ and P_2 is sampled diagnostically. On the correct negative branch, $n_2 = L_T$ and P_1 is diagnostic. Treat both cases by writing

$$v_n = \sigma^2\left(\frac{a_{\mathrm{diag}}}{n} + \frac{a_{\mathrm{opt}}}{L_T}\right), \qquad a_p = \frac{2\sigma^2 a_{\mathrm{diag}}}{d^2}, \tag{D.18}$$

where $(a_{\mathrm{diag}}, a_{\mathrm{opt}}) = (1, 4)$ on the positive branch and $(a_{\mathrm{diag}}, a_{\mathrm{opt}}) = (4, 1)$ on the negative branch. The nominal crossing count is $n_p^\star = a_p b_T$.

For each fixed p in these regions, $n_p^\star = \Theta(\log T)$, whereas $m_T = o(\log T)$ and $L_T/\log T \to \infty$. Hence, for all sufficiently large T,

$$m_T < (1 - 3\eta_T)n_p^\star < (1 + 3\eta_T)n_p^\star < L_T. \tag{D.19}$$

The nominal count is obtained by equating the leading threshold variance to d^2. In fact, for $h_n = \sqrt{2b_T v_n}$,

$$\frac{h_n^2}{d^2} = \frac{n_p^\star}{n} + \frac{2\sigma^2 a_{\mathrm{opt}} b_T}{d^2 L_T}.$$

The second term is $O(b_T/L_T) = o(\eta_T)$ at fixed p. At counts $(1 \mp 3\eta_T)n_p^\star$, the first term is $1 \pm 3\eta_T + O(\eta_T^2)$. Taking square roots, and using monotonicity in n, shows that for all sufficiently large T,

$$h_n \ge |d|(1 + \eta_T) \quad \text{if } n \le (1 - 3\eta_T)n_p^\star, \tag{D.20}$$

and

$$h_n \le |d|(1 - \eta_T) \quad \text{if } n \ge (1 + 3\eta_T)n_p^\star. \tag{D.21}$$

At an early checkpoint, crossing either boundary requires $|\widehat{d} - d| \ge h_n - |d|$. Its standardized distance satisfies

$$\frac{h_n - |d|}{\sqrt{v_n}} = \sqrt{2b_T}\left(1 - \frac{|d|}{h_n}\right) \ge \sqrt{2b_T}\frac{\eta_T}{1 + \eta_T}.$$

The two-sided Gaussian tail bound is therefore at most $2\exp\{-b_T\eta_T^2/(1 + \eta_T)^2\}$. At the first grid point after $(1 + 3\eta_T)n_p^\star$, failure to cross requires $|\widehat{d} - d| \ge |d| - h_n$, whose standardized distance is at least $\sqrt{2b_T}\eta_T/(1 - \eta_T)$. Both probabilities are bounded by $2e^{-c_p\eta_T^2 b_T}$ for a positive constant c_p.

These are unconditional Gaussian bounds at deterministic count pairs. As in Step 3, the event that the policy reaches a pair and crosses too early, or fails to cross at the late checkpoint, is contained in the corresponding Gaussian deviation event. No Gaussian law conditional on the pilot outcome is needed. A union bound over the checkpoint grid therefore controls all early crossings and the late failure together.

The geometric grid overshoots the late comparison point by a factor $1 + \eta_T + o(\eta_T)$. The ceiling error is $O(1/n_p^\star) = o(\eta_T)$, so the first such checkpoint is below $(1 + 5\eta_T)n_p^\star$ for large T. Let ν_T be the total count at commitment of the experiment that is diagnostic on the true-sign branch. On an early stop, set $\nu_T = m_T$. Adding the incorrect-pilot and early-stop events gives

$$\Pr_p\{\nu_T < (1 - 3\eta_T)n_p^\star\} + \Pr_p\{\nu_T > (1 + 5\eta_T)n_p^\star\} \le C_p K_T e^{-c_p \eta_T^2 b_T} + C_p e^{-c_p m_T} + C_p e^{-c_p L_T}. \tag{D.22}$$

The first term in (D.22) controls checkpoint deviations, the second controls an incorrect pilot, and the third controls the separate early stopping rule. For $d > 0, x < 3/4$, that rule requires an upward deviation of fixed positive size for large T, because eventually $3/4 - s_T > x + (3/4 - x)/2$. For $d < 0$, its error is bounded in Step 3. Since $\eta_T^2 b_T \asymp \sqrt{\log T}$ and $\nu_T \le L_T$, multiplying the right side of (D.22) by L_T gives $o(\log T)$. Also $\eta_T n_p^\star = o(\log T)$. On the complement of the event bounded in (D.22), $|\nu_T - n_p^\star| \le 5\eta_T n_p^\star$; hence

$$|\mathbb{E}_p \nu_T - n_p^\star| \le 5\eta_T n_p^\star + L_T \Pr_p\{\text{event in (D.22)}\} = o(\log T).$$

Therefore

$$\mathbb{E}_p \nu_T = a_p b_T + o(\log T) = a_p \log T + o(\log T). \tag{D.23}$$

To count every suboptimal play, define three events. Let B_T denote a correct pilot sign and S_T an erroneous early stop in the region $d > 0, x < 3/4$. Let W_T denote a wrong final decision caused by a checkpoint, the final test, or early stopping. On $B_T \cap S_T^c \cap W_T^c$, the total number of suboptimal selections $N_{\mathrm{sub}}(T)$ equals ν_T; on every path,

$$|N_{\mathrm{sub}}(T) - \nu_T| \le L_T \mathbf{1}\{B_T^c\} + L_T \mathbf{1}\{S_T\} + T\mathbf{1}\{W_T\}. \tag{D.24}$$

The preceding Gaussian bounds give

$$\mathbb{E}_p|N_{\mathrm{sub}}(T) - \nu_T| \le L_T e^{-c_p m_T} + L_T e^{-c_p L_T} + \frac{1}{4\log T} + 2Te^{-c_p L_T} = o(\log T). \tag{D.25}$$

The decomposition bounds the cost of an incorrect pilot by L_T rounds and that of an incorrect final decision by T rounds. Substituting the two values of a_{diag} in (D.18) yields

$$\mathbb{E}_p N_{P_2}(T) = \frac{200\tau^2}{d^2}\log T + o(\log T) \quad (d > 0, x < 3/4), \tag{D.26}$$

and

$$\mathbb{E}_p N_{P_1}(T) = \frac{800\tau^2}{d^2}\log T + o(\log T) \quad (d < 0), \tag{D.27}$$

where the displayed counts include every suboptimal play.

Step 5: region with $C_{\mathrm{LB}} = 0$ and decision boundary. If $d > 0, x > 3/4$, the probability that (D.6) fails after a correct pilot is at most $e^{-c_p L_T}$. At $x = 3/4$ it is at most

$$\Pr\{\bar{X}_{L_T} < x - s_T\} \le \exp\left\{-\frac{L_T s_T^2}{2\sigma^2}\right\} \le \exp\{-c_0\sqrt{\log T}\}. \tag{D.28}$$

Here $c_0 > 0$ depends only on τ. If the early stopping condition fails, the diagnostic phase has length at most L_T. Equation (D.14) bounds the cost of additional sampling after an incorrect pilot. Hence throughout $d > 0, x \ge 3/4$, the leading suboptimal term is the m_T-round pilot, giving $R_p(T) = O(m_T) = o(\log T)$. This is a pointwise result; at $x = 3/4$ the uniform regional KL infimum remains zero. If $d = 0$, both P_1, P_2 are optimal and the policy never uses N, F, so its regret is identically zero.

Step 6: constants and uniform efficiency. For $d \neq 0$, the gap between P_1 and P_2 is $|d|/10$. Multiplying (D.26)–(D.27) by this gap gives the positive-coefficient cases of (D.10); Step 5 gives the zero-coefficient cases. Corollary 2 then gives (D.9). For each fixed population, regret is $O(\log T)$, $o(\log T)$, or zero. Each rate is $o(T^a)$ for every $a > 0$, which establishes uniform efficiency on $[0, 1]^2$. □

E Canonical Boundary Proof

E.1 Proof of Theorem 5

Recall the boundary point $p_0 = (1/2, 1/2)$ and, for fixed $c > 0$, the alternatives

$$p_{+,T} = (1/2, 1/2 - c/\sqrt{T}), \qquad p_{-,T} = (1/2, 1/2 + c/\sqrt{T}).$$

Then $d(p_{+,T}) = c/\sqrt{T}$ and $d(p_{-,T}) = -c/\sqrt{T}$. The populations lie in $[0,1]^2$ whenever $T \geq \lceil 4c^2 \rceil$. Define the two-point local minimax regret by

$$\mathfrak{R}_T^{(2)}(c) = \inf_{\pi} \max_{p \in \{p_{+,T}, p_{-,T}\}} R_p^{\pi}(T),$$

where the infimum is over population-independent adaptive policies on the full canonical library. A policy may depend on T and the candidate pair, but not on which candidate is true.

We first bound the history KL using Lemma 1. Pinsker's inequality then bounds the total variation distance between the two history laws. The resulting testing error, multiplied by the action gap, gives the regret bound.

Lemma 4 (History divergence near the decision boundary)**.** *For every adaptive randomized policy,*

$$D_{\mathrm{KL}}(\mathbb{P}^{\pi}_{+,T} \| \mathbb{P}^{\pi}_{-,T}) \leq \frac{c^2}{50\tau^2}.$$

Proof. The alternatives differ only in y, by $2c/\sqrt{T}$. The exact one-round KL values are 0 for N and P_1, $c^2/(162\tau^2 T)$ for F, and $c^2/(50\tau^2 T)$ for P_2. Every conditional KL increment is therefore at most $c^2/(50\tau^2 T)$. Summing over T rounds proves the claim. □

Proof of Theorem 5. At every round the wrong private policy costs $c/(10\sqrt{T})$. For a randomized rule, augment the pre-action history by its population-independent round-t random seed. This adds zero conditional KL and makes the selected action measurable. Let A_t be the event that the rule selects P_1 at time t. Public actions have an even larger loss, so

$$R_+(T) + R_-(T) \geq \frac{c}{10\sqrt{T}} \sum_{t=1}^{T} \{\mathbb{P}_+(A_t^c) + \mathbb{P}_-(A_t)\}.$$

For any event A and probability laws P, Q, the two-point testing inequality follows directly from the definition of total variation:

$$P(A^c) + Q(A) = 1 - \{P(A) - Q(A)\} \geq 1 - \mathrm{TV}(P,Q), \qquad \mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|.$$

Apply this to the augmented pre-action history at each round. Its KL is no larger than the full-history KL, by data processing and independence of the fresh random seed. Pinsker's inequality, $\mathrm{TV}(P,Q) \leq \sqrt{D_{\mathrm{KL}}(P\|Q)/2}$, and the preceding lemma therefore give $\mathrm{TV} \leq c/(10\tau)$. Hence

$$R^{\pi}_{p_{+,T}}(T) + R^{\pi}_{p_{-,T}}(T) \geq \frac{c}{10}\left(1 - \frac{c}{10\tau}\right)\sqrt{T}.$$

The coefficient is positive for $0 < c < 10\tau$. For the upper bound, choose P_1 or P_2 with equal probability at every round, independently of feedback. At either population, the wrong experiment is chosen with probability $1/2$ and incurs gap $c/(10\sqrt{T})$. The expected regret is therefore $c\sqrt{T}/20$ at each endpoint. Since the larger regret is at least half the sum, taking the infimum over policies yields

$$\frac{c}{20}\left(1 - \frac{c}{10\tau}\right)\sqrt{T} \leq \mathfrak{R}_T^{(2)}(c) \leq \frac{c}{20}\sqrt{T}.$$

Thus the two-point local minimax rate is $\Theta(\sqrt{T})$. □

At the fixed point p_0, both private experiments are optimal and $C_{\mathrm{LB}}(p_0) = 0$. The local result instead concerns populations that approach p_0 with T: their action gap is of order $T^{-1/2}$, while the history KL remains bounded. It therefore describes a different limit from the pointwise logarithmic result.

F THE $G = 4, K = 3$ BENCHMARK

This appendix specifies the equilibrium and experiment library, the populations used in each comparison, and the likelihood calculations and adaptive rules.

F.1 MODEL, EQUILIBRIUM, AND EXPERIMENT LIBRARY

There are four equal-mass observable classes, three latent types, and three services. The state prior is uniform. Let $\mu_g(m_g)$ be the posterior state vector after message m_g, D the service-cost matrix, and α the congestion coefficients. A type-k receiver evaluates service e, whose aggregate load is L_e, using

$$\mathbb{E}[U_{gk,e} \mid m_g, \varepsilon_{gk,e}] = 1.8\mu_{g,e}(m_g) - D_{k,e} - \alpha_e \mathbb{E}[L_e(M) \mid m_g] + \varepsilon_{gk,e},$$

where ε is independent type-I extreme value with scale .60. Write $\sigma_{gk,e}(m_g)$ for the resulting probability of choosing service e. These probabilities depend only on type, class, and own message. Conditional loads average over the joint state/message law, retaining correlations between class messages. The equilibrium minimizes the ex-ante entropy-regularized congestion potential and is unique on the evaluated interior populations. The fixed-population and congestion audits give residuals below 10^{-11} and load changes below 5×10^{-12} across initializations.

The benchmark parameters are

$$D = \begin{pmatrix} 0 & .45 & .65 \\ .55 & 0 & .40 \\ .35 & .55 & 0 \end{pmatrix}, \quad \alpha = (.65, .75, .70), \quad B = \begin{pmatrix} 1.55 & .45 & .25 \\ .25 & 1.55 & .45 \\ .45 & .25 & 1.55 \\ 1.10 & .65 & 1.15 \end{pmatrix}.$$

The favorable state adds .75 to the matching service entry of B. Class masses are $1/4$, and the quality coefficient is 1.8.

Each class receives U, A, B, or C. A reveal-rate experiment sends the true state label with probability ρ_g and U otherwise. The complete 15-experiment library is

family	$\rho = (\rho_1, \rho_2, \rho_3, \rho_4)$	c_γ
public pool/intermediate/disclosure	$r(1,1,1,1),\ r \in \{0, .35, .70, 1\}$	$0, .040, .080, .170$
group probe g–55	$.55e_g$	.012
group probe g–100	e_g	.022
coordination 13 / 24	$(.75, .15, .75, .15)$ / $(.15, .75, .15, .75)$	.025
coordination gradient	$(.20, .40, .60, .80)$	.065

Public experiments use a common draw; private reveals are class-targeted and conditionally independent given the state. Their induced posteriors average to the uniform prior. For a fixed experiment, the equilibrium potential over own-message strategies is

$$\Phi(\sigma) = \mathbb{E}_{\omega,M}\Bigg[\frac{1}{2}\sum_e \alpha_e L_e(M)^2 + \sum_{g,k,e} \frac{p_{gk}}{4}\sigma_{gk,e}(M_g)\Big(D_{k,e} - 1.8\mathbf{1}\{\omega = e\} + .60\log\sigma_{gk,e}(M_g)\Big)\Bigg].$$

The load term is convex, and entropy makes Φ strictly convex on positive-probability information sets when all type masses are positive. To obtain the equilibrium conditions, fix a type and an own message m_g with positive probability. Differentiating Φ with respect to its service probabilities and dividing by the positive mass of this information set gives

$$D_{k,e} - 1.8\mu_{g,e}(m_g) + \alpha_e \mathbb{E}[L_e(M) \mid m_g] + .60\big(1 + \log\sigma_{gk,e}(m_g)\big) + \lambda_{gk,m_g} = 0,$$

where λ_{gk,m_g} enforces that the service probabilities sum to one. The conditional load averages over the other classes' messages given m_g; replacing it by an unconditional load would change the receiver's problem. Exponentiating and normalizing the first-order conditions yields the own-message equilibrium

$$\sigma_{gk}(m_g) = \mathrm{softmax}\left(\frac{1.8\mu_g(m_g) - D_k - \alpha \odot \mathbb{E}[L(M) \mid m_g]}{.60}\right),\ L_e(M) = \sum_{g,k} \frac{p_{gk}}{4}\sigma_{gk,e}(M_g).$$

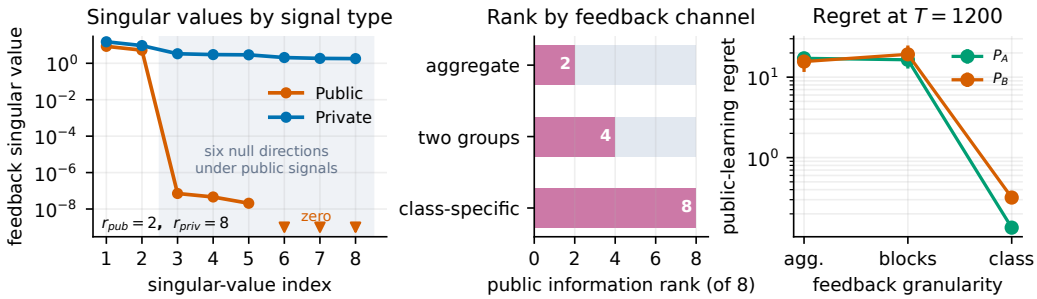


Figure 6: Observation Jacobian spectrum $s_i = \sqrt{\lambda_i(\mathcal{I}_\Gamma)}$, weighted by probability and scaled by noise. Left: public/private ranks are 2/8; triangles mark numerical zeros at the display floor 10^{-9}. Center: finer feedback raises public rank $2 \to 4 \to 8$ (relative tolerance 10^{-6}). Right: Public + Wald regret against public oracles for P_A, P_B (20 seeds). Grouping and channel dimension both change.

Here D_k is row k of D, and $\odot$ denotes componentwise multiplication. The class load on service e is

$$X_{g,e}(M) = \frac{1}{4}\sum_k p_{gk}\sigma_{gk,e}(M_g), \qquad L_e(M) = \sum_g X_{g,e}(M).$$

Thus $X_{g,e}$ includes the class mass $1/4$. The sender value is

$$V_p(\gamma) = \mathbb{E}\left[\sum_{g,e} R_{g,e,\omega} X_{g,e} - .22\sum_e L_e^2\right] - c_\gamma,$$

where $R_{g,e,\omega} = B_{g,e} + .75\mathbf{1}\{e = \omega\}$. Online feedback is $Z_t = L(M_t) + \xi_t$, with $\xi_t \sim \mathcal{N}(\mathbf{0}, (0.035)^2 I_3)$ independently across rounds.

F.2 Population settings and observation kernels

The 30-population comparison uses generator seed 20260828 and draws rows independently from $\mathrm{Dirichlet}(1.2, 1.2, 1.2)$. It retains the first populations whose type masses are all at least .035. Selection does not use decisions or performance. The four acquisition/test combinations and matched greedy share feedback, estimator, initialization, experiment library, populations, and seeds. Local Laplace Thompson sampling is an additional comparator. All full-library regret calculations use the oracle over all fifteen experiments. For Figure 6, let $a_1 = (.72, .18, .10)$, $a_2 = (.12, .73, .15)$, $a_3 = (.14, .16, .70)$, and $a_4 = (.34, .33, .33)$; the richer-feedback sweep fixes $P_A = (a_1, a_2, a_3, a_4)$ and $P_B = (a_2, a_3, a_1, a_4)$.

The left spectrum in Figure 6 is evaluated at $P_0 = \frac{1}{2}(a_1 + a_3, a_2 + a_4, a_1 + a_3, a_2 + a_4)$. To relate it to Fisher information, let $J_\gamma(\omega, M)$ be the Jacobian of the conditional feedback mean in the eight population coordinates. With feedback noise standard deviation $\sigma_Z = .035$, the Gaussian information formula gives

$$\mathcal{I}_\Gamma(p) = \frac{1}{\sigma_Z^2}\sum_{\gamma\in\Gamma} \mathbb{E}_{\omega,M|\gamma}\left[J_\gamma(\omega, M)^\top J_\gamma(\omega, M)\right].$$

This is the sum of the information from one observation per experiment in Γ. The state/message law is population independent, so it contributes no score. Conditional on a state/message profile, the Gaussian score is $J_\gamma^\top(Z - L)/\sigma_Z^2$; taking its covariance gives $J_\gamma^\top J_\gamma/\sigma_Z^2$. Averaging over profiles and summing over experiments yields the displayed formula.

Stacking the matrices $\sqrt{\Pr(\omega, M \mid \gamma)}J_\gamma(\omega, M)/\sigma_Z$ yields a Jacobian whose Gram matrix is $\mathcal{I}_\Gamma$. Its singular values are $s_i = \sqrt{\lambda_i(\mathcal{I}_\Gamma)}$, where λ_i are the eigenvalues in decreasing order. Effective rank counts those exceeding $10^{-6}s_1$. The same calculation applies to each feedback channel with its corresponding conditional mean.

Populations with identical public feedback. The fixed population pair is

$$P^+ = \begin{pmatrix} .0578235 & .5795482 & .3626283 \\ .3864226 & .0471035 & .5664739 \\ .8317217 & .0553204 & .1129579 \\ .1500459 & .7872957 & .0626584 \end{pmatrix},$$

with P^- obtained by exchanging rows 2 and 4 of P^+. Both pool to $(.3565034, .3673169, .2761796)$. In the order pool, intermediate-35, intermediate-70, and disclosure, their public-policy values are $(.89038, .96217, 1.03395, 1.03977)$ at P^+ and $(.98266, 1.03792, 1.09319, 1.08485)$ at P^-. Thus disclosure and intermediate-70 are the respective unique public optima, and intermediate-70 minimizes the sum of endpoint public regrets at .00581765.

The search enforces equal pooled composition, identical public feedback laws, and disjoint public optima before maximizing the separation gap; algorithm performance is not a selection criterion. Public factorization then makes every adaptive public history law identical.

F.3 Controlled test with known candidate populations

The candidates are the populations P^+, P^- defined above. Their unique public optima are disclosure and intermediate-70. The full fifteen-experiment library has the same optimum, coordination-13, at both endpoints. The terminal decision in this test is restricted to the public library; diagnostic collection can use the three specified public or private experiments.

Write $a_i = (\omega_i, m_i)$ for the state and message profile in observation i, and $h_{\pm}^{\gamma_i}(a_i)$ for its mean aggregate feedback. The distribution of a_i is independent of the candidate population. Conditional on these profiles, feedback noise is independent Gaussian with covariance $\sigma_Z^2 I_3$, where $\sigma_Z = .035$. The common state/message probabilities cancel from the likelihood ratio. For a profile a, the conditional KL and its average are

$$\delta_\gamma(a) = \frac{\|h_+^\gamma(a) - h_-^\gamma(a)\|_2^2}{2\sigma_Z^2}, \qquad D_\gamma = \mathbb{E}_a[\delta_\gamma(a)].$$

Expanding the Gaussian squared residuals gives

$$\Lambda_n = \sum_{i=1}^{4+n} \frac{(h_+^{\gamma_i}(a_i) - h_-^{\gamma_i}(a_i))^\top [Z_i - (h_+^{\gamma_i}(a_i) + h_-^{\gamma_i}(a_i))/2]}{\sigma_Z^2}.$$

For a single observation under P^+, write $Z_i = h_+^{\gamma_i}(a_i) + \xi_i$. Its contribution to Λ_n has conditional mean $\delta_{\gamma_i}(a_i)$ and variance $2\delta_{\gamma_i}(a_i)$. Under P^-, the mean changes sign and the variance is unchanged. Independence of the Gaussian noise therefore gives, with $S_n = \sum_{i=1}^{4+n} \delta_{\gamma_i}(a_i)$,

$$\Lambda_n \mid (a_i)_{i=1}^{4+n} \sim \mathcal{N}(S_n, 2S_n) \quad \text{and} \quad \mathcal{N}(-S_n, 2S_n).$$

The two laws correspond to P^+ and P^-, respectively, and include the four shared initialization observations. An equal-prior likelihood-ratio test selects P^+ for $\Lambda_n > 0$ and P^- for $\Lambda_n < 0$, randomizing at a tie. Because the public optima differ, its classification error equals its public-policy decision error. Gaussian tails give

$$e_n = \mathbb{E}_{(a_i)}\left[\Phi\left(-\sqrt{S_n/2}\right)\right],$$

where Φ is the standard normal cumulative distribution function. All three added experiments reveal deterministically to their specified recipients, so each has one message profile per state. We calculate e_n by summing over the three-state multinomial counts and the four initial state realizations. Averaging the conditional errors is necessary: substituting $\mathbb{E}[S_n]$ into the nonlinear Gaussian tail generally gives a different value.

For a fixed added experiment γ, the KL chain rule gives

$$\mathbb{E}_{P^+}[\Lambda_n - \Lambda_0] = nD_\gamma, \qquad \mathbb{E}_{P^-}[\Lambda_0 - \Lambda_n] = nD_\gamma.$$

Figure 4 averages the increment in favor of the true population: $\Lambda_n - \Lambda_0$ at P^+ and $\Lambda_0 - \Lambda_n$ at P^-. The four common initial probes have expected total KL 1.05738 and predicted error .236283.

Public disclosure has identical feedback laws by pooled composition. Class-1 probing also preserves equivalence: the informed class is unchanged, and the other classes enter aggregate behavior through their combined composition. Class-2 probing distinguishes the exchanged classes.

For public disclosure, class-1 probing, and class-2 probing, respectively, (D_γ, c_γ) equals $(0, .170)$, $(0, .022)$, and $(.528690, .022)$. Their mean full-library oracle gaps are .019738, .057549, and .075925. Thus equally priced private probes can carry different decision information, while signaling cost and regret cost remain distinct.

The simulation uses 10000 replicates per endpoint and budgets $n = 0, 1, 2, 4, 8, 16, 32, 64, 128$, with shared initialization, state, and Gaussian noise draws across branches. Exact conditional likelihoods are computed from the three-dimensional telemetry. The largest standardized discrepancies are 1.34 for mean added log-likelihood ratio and 1.42 for decision error; at $n = 16$, the predicted and observed increments are 8.4590 and 8.4837. This known-candidate calculation requires no population optimization.

F.4 Decision-focused sampling with likelihood-ratio tests

Estimation. Decision + LR and the matched full-library comparators begin with one full-reveal probe per class. At geometric checkpoints they fit the constrained Gaussian likelihood by sequential least squares programming (SLSQP), initialized at the previous estimate and the uniform population. Each type fraction is constrained to be at least .002. The first refit follows observation 1; after a refit at observation count t, the next occurs at $\lfloor 1.75t \rfloor + 1$. A feasible converged candidate replaces the previous estimate only if it improves the likelihood. The fitted population determines the current best experiment $\widehat{\gamma}_t$.

Let $a_s = (\omega_s, M_s)$ denote the observed state/message profile, and let $h_p^\gamma(a_s)$ be its equilibrium feedback mean. Up to a parameter-independent term, the log-likelihood is

$$\ell_t(p) = -\frac{1}{2\sigma_Z^2} \sum_{s=1}^{t} \|Z_s - h_p^{\gamma_s}(a_s)\|_2^2.$$

The population-independent state/message probabilities cancel from likelihood comparisons. Central differences at the fitted coordinates $\widehat{\theta}_t$ give the value gradients and conditional mean Jacobians $J_{\gamma_s}(a_s; \widehat{\theta}_t)$. The accumulated information used by the learner is

$$\mathcal{I}_t = 10^{-5} I_8 + \frac{1}{\sigma_Z^2} \sum_{s=1}^{t} J_{\gamma_s}(a_s; \widehat{\theta}_t)^\top J_{\gamma_s}(a_s; \widehat{\theta}_t).$$

The ridge term makes the matrix invertible before all directions have been sampled. This sum uses the profiles actually observed. In contrast, the one-round information $\mathcal{I}_\gamma(\widehat{p}_t)$ of a candidate experiment averages over its possible future profiles, as in Appendix F. Between refits, the learner collects new observations but retains the fitted population, information matrix, value gradients, and decision tests from the last refit.

Decision tests. At a refit checkpoint t, let ℓ_t be the log-likelihood and $\widehat{p}_t$ the fitted population. Motivated by the decision regions in equation B.14a, the log-GLR compares this fit with the best fit where a competitor matches or exceeds the plug-in value:

$$\mathcal{A}_{j,t} = \{p : V_p(j) \geq V_p(\widehat{\gamma}_t)\}, \qquad G_{j,t} = \ell_t(\widehat{p}_t) - \sup_{p \in \mathcal{A}_{j,t}} \ell_t(p).$$

These pairwise comparison regions include ties and populations where j beats $\widehat{\gamma}_t$ but is not globally optimal. They are larger than the corresponding optimality regions in equation B.14a.

To approximate the log-GLR, let $\mathcal{I}_t$ be the regularized accumulated Fisher information at the fitted coordinates $\widehat{\theta}_t$. Write $\dagger$ for the Moore–Penrose pseudoinverse. First-order error propagation (the delta method) approximates the variance of a value difference using its gradient:

$$\widehat{\Delta}_{j,t} = V_{\widehat{p}_t}(\widehat{\gamma}_t) - V_{\widehat{p}_t}(j), \qquad b_{j,t} = \nabla_\theta \left[V_p(\widehat{\gamma}_t) - V_p(j) \right] \big|_{\widehat{\theta}_t},$$

$$v_{j,t} = b_{j,t}^\top \mathcal{I}_t^\dagger b_{j,t}, \qquad G_{j,t}^{\text{local}} = \frac{\widehat{\Delta}_{j,t}^2}{2 v_{j,t}}.$$

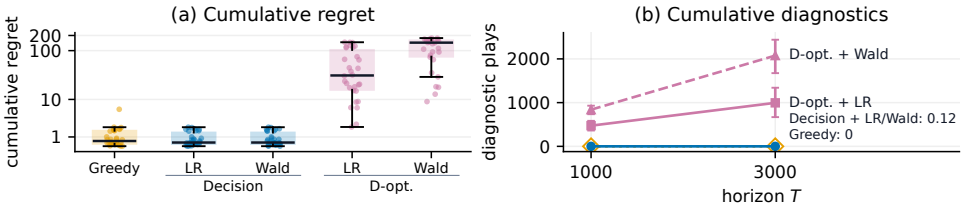


Figure 7: Adaptive policies (30 populations, 10 paired seeds). (a) Dots: per-population mean full-library regret; boxes: interquartile ranges with median lines. Values use $\log_{10}(1 + R)$ with original-scale ticks. (b) Mean diagnostics after four initialization plays, with 95% population-cluster bootstrap intervals.

For $v_{j,t} > 0$, the last expression is the minimum quadratic likelihood loss $\frac{1}{2}u^\top \mathcal{I}_t u$ subject to the linearized comparison $\widehat{\Delta}_{j,t} + b_{j,t}^\top u \le 0$, where u is a coordinate perturbation. The implementation floors the denominator at 10^{-30}. The quadratic approximation ranks competitors. When the smallest statistic reaches half the test threshold, a constrained likelihood search refines that comparison; the others retain their quadratic approximations. Failed or conflicting searches leave the competitor plausible, and negative likelihood differences are truncated to zero.

For a scalar comparison, a one-degree-of-freedom chi-square calibration motivated by Wilks' approximation (Wilks, 1938) gives the fixed log-LR threshold $\frac{1}{2}\chi^2_{1,.90} = 1.3528$. This is a heuristic calibration for constrained local searches, without a global or anytime guarantee. The Wald variant accepts the fitted decision when every value gap exceeds $1.7\sqrt{\log(t+3)}$ local standard errors.

Sampling. Among competitors still plausible after testing, let j have the smallest retained log-GLR statistic. For a candidate experiment γ, the predicted one-step decrease in the variance of this value difference is

$$g_{j,t}(\gamma) = \left[b_{j,t}^\top \mathcal{I}_t^\dagger b_{j,t} - b_{j,t}^\top \big(\mathcal{I}_t + \mathcal{I}_\gamma(\widehat{p}_t)\big)^\dagger b_{j,t} \right]_+ .$$

This calculation holds the fitted population and value gradient fixed while adding the expected information from one observation. Decision + LR chooses the diagnostic experiment maximizing

$$\frac{g_{j,t}(\gamma)}{\max\{.003,\ V_{\widehat{p}_t}(\widehat{\gamma}_t) - V_{\widehat{p}_t}(\gamma)\}}.$$

It plays the estimated optimum if the decision test passes or if this experiment provides at least .35 of the selected diagnostic probe's variance reduction. Otherwise it plays that diagnostic probe. The ratio prioritizes information about the competing decisions per unit estimated regret, while the .35 rule permits information gathering during exploitation.

Comparators. Greedy, called Initialized-Myopic in the implementation, always plays the fitted optimum after the same four probes and with the same estimator and refit schedule. Private-Myopic omits this shared initialization. Local Laplace Thompson sampling (TS) approximates the population posterior by $\mathcal{N}(\widehat{\theta}_t, \mathcal{I}_t^\dagger)$, projects a sampled coordinate vector to the simplex, and maximizes its linearized value (Russo et al., 2018). D-Optimal selects probes by one-step log-determinant Fisher gain and exploits after its LR or Wald decision test passes. Public + Wald restricts Decision + Wald to public experiments and initializes with one play of each public experiment. Thus the D-optimal comparisons change both probe selection and the condition for exploitation.

Metrics and additional results. Regret is computed from true equilibrium values for offline evaluation. Parameter error is $\|\widehat{\theta} - \theta\|_2$ in the eight free coordinates, and accuracy records whether the final fitted decision is optimal. Diagnostic counts exclude the four initialization probes. The first two random suites use 5000 population-cluster bootstrap resamples.

At $T = 3000$, D-Optimal + Wald has mean regret 115.16 and uses 2075 diagnostics; local Laplace TS has mean regret 1.39. Decision + LR and D-Optimal + LR use .12 and 995 diagnostics, respectively. Final accuracy ranges from .957 to .967 across the five policies in Figure 7. The 95% intervals for

Decision + LR minus D-Optimal + LR and local Laplace TS are $[-71.94, -36.72]$ and $[-.65, -.36]$. Against greedy, the means are .892 and 1.083, with paired interval $[-.567, -.0002]$. These small differences agree with the nearly identical behavior described in the main text.

A second 30-population suite gives mean regret .805 for Decision + LR and 1.986 for greedy, with paired difference interval $[-2.310, -.297]$. Both suites informed one shared estimator correction before the fixed rerun, so neither is an untouched holdout. All sixty true optima have numerical Fisher rank eight at relative singular-value tolerance 10^{-6}.

Independent population evaluation. We draw 30 additional populations from the same conditional Dirichlet distribution and run ten paired seeds per population through $T = 3000$. The estimator and acquisition settings are fixed before drawing these populations and are retained throughout evaluation. Uncertainty in the mean regret difference is measured by 20000 bootstrap resamples of the 30 population clusters, keeping each population's ten seed pairs together.

At $T = 3000$, mean regret is 1.158 for Decision + LR and 2.298 for greedy. The paired difference is -1.140, with population-cluster 95% interval $[-2.284, -.232]$. Their terminal decisions are correct in $274/300$ and $270/300$ runs, respectively. Decision + LR has lower regret in ten runs across seven populations and equal regret in the other 290 runs. The largest single run accounts for 27.8% of the total reduction. It uses .507 diagnostic observations per run on average after initialization. Thus the mean improvement comes from a small subset of runs, while most runs have the same regret under both policies.

F.5 A POSITIVE EXPLORATION-COST LOWER BOUND

To examine an instance where optimal play leaves a decision-relevant ambiguity, replace the congestion vector by $\alpha = (.325, .375, .350)$ and retain the full 15-experiment library. Use the seven-decimal population coordinates above, with P^- obtained by exchanging rows 2 and 4 of P^+. The pooled composition is exactly unchanged. Public disclosure uniquely maximizes value at P^+, with value 1.07119703 and gap .01536572 to the next-best experiment. At P^-, the unique optimum is the coordination experiment with reveal probabilities $(.75, .15, .75, .15)$, with value 1.13265082 and gap .02442277 to the next best. These value orderings are evaluated using the unique congestion equilibrium.

Theorem 2 gives equal complete transcript laws under every public experiment, including the unique optimum at P^+. Thus $P^- \in \mathrm{Conf}_0(P^+)$. Keeping only this alternative in the KL allocation gives a lower bound on the full program. Write $D_\gamma = D_{\mathrm{KL}}(\mathsf{K}^\gamma_{P^+} \| \mathsf{K}^\gamma_{P^-})$. Every feasible allocation must satisfy $\sum_\gamma x_\gamma D_\gamma \geq 1$, so

$$C_{\mathrm{LB}}(P^+) \geq \min_{\gamma: D_\gamma > 0} \frac{\Delta_{P^+}(\gamma)}{D_\gamma} \approx .07276686.$$

Numerical equilibrium and KL evaluation give the displayed approximation, with the minimum attained by the coordination experiment with reveal probabilities $(.15, .75, .15, .75)$: its gap is .02250828 and its KL divergence is .30932041. All experiments with positive KL are suboptimal at P^+, making the bound strictly positive. Additional alternatives may increase the allocation cost, so this calculation gives a lower bound on the full coefficient.

The ratio arises from a linear allocation problem with one information constraint. If $r = \min_{\gamma: D_\gamma > 0} \Delta_{P^+}(\gamma)/D_\gamma$, then $\Delta_{P^+}(\gamma) \geq r D_\gamma$ for every experiment. Multiplying by the allocation and summing gives cost at least $r \sum_\gamma x_\gamma D_\gamma \geq r$. Allocating $1/D_\gamma$ to a minimizing experiment attains r for this single-alternative problem. The online policy below does not receive the candidate pair or this allocation; it must estimate the population and find competing decisions from feedback.

Online behavior on the same pair. Following a ten-seed pilot, we evaluate Decision + LR and greedy on 100 new paired seeds at each population, with horizons $T \in \{100, 300, 1000, 3000\}$. The populations, algorithm settings, and sample size are fixed before these runs. Both learners estimate eight population coordinates and use all 15 experiments; the candidate pair is not supplied to them. They share four initialization observations and the estimator. Table 1 reports $T = 3000$.

Decision + LR minus greedy mean regret is 2.637 at P^+ and -12.988 at P^-. The respective 95% percentile bootstrap intervals are $[2.057, 3.284]$ and $[-18.964, -7.542]$, using 20000 paired seed

resamples separately at each fixed population. At P^-, Decision + LR has lower regret in 23 runs and ties greedy in 77. The largest run contributes 6.1% of the total reduction; omitting any single seed leaves a mean reduction of at least 12.322.

All 16 erroneous greedy runs at P^- select public disclosure throughout rounds 1001–3000. Its per-round gap is .0263573, giving mean regret growth $(16/100) \times 2000 \times .0263573 = 8.4343$, equal to the observed increase from 4.918 to 13.353 up to rounding. Decision + LR's mean regret remains .365 from $T = 100$ onward. The exact binomial 95% intervals for final error are $[9.43\%, 24.68\%]$ for greedy and $[0\%, 3.62\%]$ for Decision + LR. The observed growth is therefore accounted for by these persistent public-disclosure errors. Over the full horizon, the 16 terminal-error runs contribute 47,909 suboptimal public plays and the 84 terminal-correct runs contribute 709, giving the reported mean of 486.18. A correct final decision can therefore coexist with earlier public-disclosure errors.

F.6 Sensitivity to congestion strength

The half-congestion pair above was chosen to exhibit costly exploration. To examine its sensitivity to that choice, fix the same populations and experiment library and set $\alpha(\kappa) = \kappa(.65, .75, .70)$ for $\kappa \geq 0$. We evaluate the predetermined grid $\kappa \in \{0, .25, .5, .75, 1, 1.25, 1.5\}$. The sender payoff, signaling costs, telemetry noise, and learner settings are unchanged. At each grid point, we compare the two policies on 20 paired seeds per population through $T = 3000$.

At $\kappa = 0, .25, .5, .75$, public disclosure is uniquely optimal at P^+ and coordination 13 is uniquely optimal at P^-. Public pooling makes P^- invisible under the optimum at P^+, giving the respective single-alternative lower bounds .11006, .09661, .07277, and .04052. At $\kappa = 1, 1.25, 1.5$, coordination 13 is optimal at both populations, so this pair no longer gives a positive-cost lower bound. These latter points do not determine whether another population belongs to $\mathrm{Conf}_0(P^+)$.

Local parameter stability follows from continuity of the unique equilibrium. For a fixed experiment and population, the entropy-regularized potential is jointly continuous in κ and in the own-message strategies, over a compact feasible set. As $\kappa_n \to \kappa$, compactness gives convergent subsequences of equilibrium strategies, and continuity makes each limit a minimizer of the limiting potential. Strict convexity makes that minimizer unique. The equilibrium, values, and Gaussian observation means therefore vary continuously with κ. Whenever public disclosure and a different experiment are strict optima at the two populations, these inequalities persist under sufficiently small changes of κ. Public pooling holds throughout the family, preserving the decision-relevant ambiguity. The grid places a change of optimum between $\kappa = .75$ and 1.

Table 2: Congestion sweep with the fixed population pair and library ($T = 3000$, 20 paired seeds per endpoint). ΔR is mean regret of Decision + LR minus greedy, with pointwise 95% paired bootstrap intervals. Errors are terminal decision errors for Decision + LR / greedy, each out of 20. A dash means that this pair gives no positive lower bound on the full coefficient.

		P^+		P^-	
κ	Lower bound on $C_{\mathrm{LB}}(P^+)$	ΔR [95% interval]	Errors	ΔR [95% interval]	Errors
0	0.1101	0.992 [0.601, 1.599]	0/0	−23.053 [−32.296, −13.804]	0/10
0.25	0.0966	1.342 [0.929, 1.773]	0/0	−13.094 [−24.557, −3.182]	0/4
0.5	0.0728	1.533 [1.109, 1.986]	0/0	−8.442 [−20.184, −0.016]	0/2
0.75	0.0405	4.204 [2.861, 5.662]	0/0	0.000 [0.000, 0.000]	0/0
1	–	−0.343 [−0.457, −0.229]	14/18	0.000 [0.000, 0.000]	0/0
1.25	–	0.000 [0.000, 0.000]	7/7	0.000 [0.000, 0.000]	0/0
1.5	–	0.000 [0.000, 0.000]	9/9	0.000 [0.000, 0.000]	0/0

Table 2 reports every grid point. For $\kappa = 0, .25, .5$, Decision + LR incurs extra regret at P^+ and corrects greedy's terminal errors at P^-. At $\kappa = .75$, both methods finish correctly at both populations, while Decision + LR still incurs additional regret at P^+. A positive exploration-cost lower bound therefore does not imply a finite-horizon advantage over greedy on each tested population.

For $\kappa = 1, 1.25, 1.5$, the gap between coordination 13 and its closest competitor at P^+ decreases from 2.38×10^{-4} to 7.42×10^{-5}. At $\kappa = 1$, public disclosure accounts for seven of Decision +

LR's fourteen terminal errors and sixteen of greedy's eighteen. At $\kappa = 1.25, 1.5$, every terminal error selects coordination 24. Both methods have mean regret below 1.1 at these grid points, so low regret coexists with frequent errors in selecting the exact optimum.

G AI USE STATEMENT

We used generative artificial intelligence (AI) tools, including OpenAI Codex, as described below. The authors take responsibility for all content, including AI-assisted derivations, text, code, figures, and interpretation of results.

The authors originated the research question, core ideas, and mathematical models. They formulated the main theoretical claims and proposed the key lines of argument, which were developed with AI assistance. For some subsequent, more technically involved arguments, AI tools generated draft derivations that the authors reviewed and revised as needed.

The authors proposed the experimental questions and approach. AI tools wrote, implemented, debugged, and ran the simulation code from these specifications, and assisted with output analysis. The reported numerical results were obtained by executing the code.

The authors conducted literature searches and prepared figures, with AI assistance in finding additional references, checking bibliographic details, and revising figures, labels, captions, and explanations. AI tools also suggested improvements to manuscript organization and technical explanations, and assisted with a final language edit for readability, sentence length, and terminology consistency.