A CONSERVATION LAW FOR LEARNING IN STRATEGIC SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard exploration diagnostics measure how much a policy moves, but not whether that movement remains informative after the data-generating process responds. We introduce *strategic observability* to measure how strongly policy motion separates latent environments after this response, and define *strategic capacity* as motion times squared observability. For binary environments, mutual information is bounded above by expected capacity and below by a truncated almost-sure capacity floor. Under local Fisher geometry, the quadratic exploration cost decomposes orthogonally into identifiable and null components. Their relative weights are the squared cosine and sine of the angle between policy motion and the identifiable subspace; the identifiable component controls learning up to the conditioning of the identification operator. Local characterisation results further select Fisher separation energy as the learning currency and the symmetric Fisher quadratic form as the cost. The finite-step quantities used in practice are measurable surrogates for, rather than exact copies of, these geometric objects. Empirically, six policies with exactly equal budget, average separation energy, switching, exposure, and action entropy separate only through minimum pairwise separation: the two with zero minimum perform at chance, whereas the remaining four achieve error below 14%. A language-model intervention likewise shows that greater motion along an uninformative direction can yield negligible separation.

1 INTRODUCTION

Exploration is usually evaluated by how much a learner varies its behavior. Action entropy, switching frequency, uncertainty bonuses, and distributional drift all quantify motion in one form or another. This is natural in classical sequential learning, where the observation law is fixed and exploration determines which part of that law is sampled (Wald, 1947; Lattimore and Szepesvári, 2020). Information-directed methods refine this view by asking how much information an action is expected to acquire relative to its cost (Russo and Roy, 2014; 2016). In both cases, however, the environment itself does not strategically reorganize in response to the learner's future policy.

That distinction matters in systems whose data are produced by responsive agents. Recommenders affect what creators produce and which content remains visible (Chaney et al., 2018); deployed policies can change the distribution on which they are subsequently trained (Perdomo et al., 2020); and preference-learning systems determine which outputs evaluators see (Christiano et al., 2017; Ouyang et al., 2022). Strategic classification similarly models agents who alter observable features in response to a decision rule (Hardt et al., 2016). These literatures show that policy and data distribution must be studied jointly. They leave a more basic measurement question: when the data-producing agents respond, which part of the learner's exploration still distinguishes the latent environment?

The issue is easiest to see when two policy changes have the same size but different directions. One change may induce responses that differ sharply across candidate environments; another may induce nearly identical responses under every candidate. Motion-based diagnostics assign the same exploration value to both, although only the first resolves uncertainty. Related failures of excitation are familiar in closed-loop identification (Borkar and Varaiya, 1979; Kumar and Becker, 1982; Mandl, 1974; Lai and Wei, 1982), while strategic exploration mechanisms emphasize that information production depends on how agents respond to incentives (Sellke and Slivkins, 2023; Mansour

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.

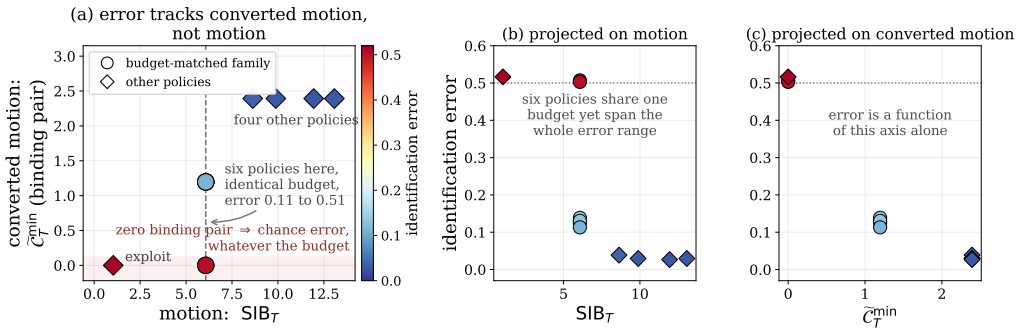


Figure 1: **Exploration under strategic feedback is two-dimensional, and only one dimension predicts learning.** Eleven policies ($|\Omega| = 4$, $T = 200$), placed by motion spent (SIB_T) against the part of it the environment leaves observable at the binding pair ($\widetilde{\mathcal{C}}_T^{\min}$); colour is identification error, red at chance. *(a)* The colour boundary is horizontal: spending more buys nothing in a direction the environment answers identically. *(b,c)* Six policies (circles) share a budget *exactly* by construction, not tuning (Prop. 16) along with switching, exposure, entropy and average energy, yet span the whole error range and separate on the vertical axis alone. The response map here is one we wrote, and $\widetilde{\mathcal{C}}_T$ is an unnormalised surrogate for (5) (§4.2).

et al., 2020). Our goal is to isolate this distinction as a measurable property of the policy-induced experiment.

We therefore separate exploration into two dimensions. The first is the policy's motion, measured by a step cost Δ_t. The second is *strategic observability*, $\mathfrak{O}_t \in [0, 1]$, which measures how much cross-environment separation the realized direction of motion produces after the strategic response. Their product defines the strategic capacity

$$\mathcal{C}_T(\Pi) = \sum_{t=1}^{T} \mathfrak{O}_t^2 \Delta_t.$$

The square is not chosen for convenience: locally, both information and Fisher cost are squared norms. Capacity is therefore large only when the policy both moves and moves in a direction that the environment answers differently.

Figure 1 illustrates why the second dimension cannot be replaced by a standard exploration diagnostic. Six policies have exactly the same budget, switching frequency, exposure, action entropy, and average pairwise separation energy. They differ only in the minimum separation over hypothesis pairs. The two policies with a zero minimum remain at chance, whereas the four with a positive minimum achieve error below 14%. Thus the relevant bottleneck is not total motion, or even average informativeness, but whether the least-separated pair remains distinguishable. We develop this observation in three steps. First, Theorem 1 compares mutual information with accumulated strategic capacity. For binary environments,

$$c \min\{\operatorname{ess\,inf} \mathcal{C}_T, 1\} \le I(\omega; Y_{1:T} \mid \Pi) \le C\, \mathbb{E}[\mathcal{C}_T].$$

The expectation and essential-infimum qualifiers reflect different directions of the comparison, and the lower bound is truncated because binary information saturates at $\log 2$. The information-accumulation argument is classical in spirit; the strategic contribution is to connect each increment of separation to motion selected and paid for by the policy.

Second, Theorem 2 gives the geometric content of the conservation law. Under local Fisher geometry, the quadratic exploration cost splits orthogonally as $q_t = \mathcal{C}_t^{\mathrm{geo}} + \mathcal{W}_t^{\mathrm{geo}}$: an identifiable component that the environments answer differently and a null component that they answer identically. The identifiable share is the squared cosine of the angle between policy motion and the identifiable subspace. It controls learning up to the conditioning of the identification operator (Proposition 3). Characterisation results then select Fisher separation energy as the local learning currency and the symmetric Fisher quadratic form as the corresponding cost (Appendix Q).

Third, Corollary 4 translates the local geometry into a finite-step quantity that can be measured without differentiating the response map:

$$\mathrm{SIB}_T = \mathcal{C}_T + \mathcal{W}_T, \qquad \mathcal{W}_T = \sum_t (1 - \mathfrak{O}_t^2)\Delta_t.$$

The identity alone is algebraic; its value comes from the information bounds and from its local comparison with the Fisher decomposition. This distinction also separates our object from the information ratio: the latter evaluates the information obtained under an induced experiment, whereas strategic capacity asks how effectively policy motion creates an informative experiment in the first place.

We study the framework in a decentralized commons, incentivized exploration, strategic RLHF, iterative preference learning, a hierarchical identification task, and an intervention with a language-model responder. Across these settings, the recurring comparison is between how much a policy moves and how much hypothesis separation that motion preserves. Additional results on capacity frontiers, near-optimal learning, admissible budgets, and stabilization appear in Appendix R. Together, the results turn the question of whether exploration is informative into a property of the induced strategic experiment rather than a conclusion drawn from motion alone. A full discussion, including model collapse, recommender feedback loops, information design and off-policy evaluation, is in Appendix T.

2 Strategic Feedback Systems and the Information Budget

Strategic feedback channels. Let $\omega \in \Omega$ be an unknown environment parameter. A policy Π selects actions A_t given history H_{t-1}. The *predictive kernel* at time t is $K_t^{\omega,\Pi}(\cdot \mid H_{t-1}) := \mathsf{Law}_\omega^\Pi(Y_t \in \cdot \mid H_{t-1})$. An environment is a *strategic feedback channel* if some pair of policies agreeing on A_t a.s. still induce different $K_t^{\omega,\cdot}$ with positive probability: the kernel depends on the policy's *continuation*, not just its current action.

Strategic information budget. The policy pays for the displacement it makes at each step. Writing $\bar{K}_t^\omega := \mathsf{K}(\cdot \mid \mathrm{BR}_t(\bar{u}_t, \omega), \bar{u}_t)$ for the kernel the baseline would have induced and $K_t^{\omega,\Pi}$ for the enacted one, the step cost is

$$\Delta_t^{\mathrm{KL}}(\Pi) := \sup_\omega \mathbb{E}\big[\mathrm{KL}(\bar{K}_t^\omega \,\|\, K_t^{\omega,\Pi}) \;\big|\; \mathcal{F}_{t-1}\big], \qquad \mathrm{SIB}_T^{\mathrm{KL}}(\Pi) := \sum_{t=1}^T \Delta_t^{\mathrm{KL}}(\Pi). \tag{1}$$

Two features matter. The cost is *conditional* on the history, so Δ_t^{KL} is $\mathcal{F}_{t-1}$-measurable and SIB_T, $\mathcal{C}_T$ are random variables; every bound below carries an explicit expectation, essential-infimum or almost-sure qualifier. And it is indexed by the *displacement*, not by a transition between consecutive kernels, giving one cost per step for $t = 1, \ldots, T$ and matching the count of observations; transitions would give $T - 1$ costs and leave the last observation uncharged. Analogously define Δ_t^{TV}, $\mathrm{SIB}_T^{\mathrm{TV}}$.

Timing: baseline, displacement, endpoint. Three objects must be kept distinct; conflating them makes the theory vacuous. The policy holds a *baseline continuation* $\bar{u}_t$, chooses a *displacement* δu_t, and enacts only the *endpoint* $u_t = \bar{u}_t + \delta u_t$. Agents respond to the enacted continuation through $x_t^\omega = \mathrm{BR}_t(u_t, \omega)$ and the observation is generated *after* the displacement, $Y_t \sim \mathsf{K}(\cdot \mid \mathrm{BR}_t(u_t, \omega), u_t)$ with $u_t = \bar{u}_t + \delta u_t$; write $K_t^{\omega,\Pi}$ for this endpoint kernel. A policy is *asymptotically strategically stable* if $d_{\mathrm{tail}}(u_{t:\infty}, u_{t+1:\infty}) \to 0$ a.s.

Realised separation, and the modulus derived from it. The primitive quantity is the *realised endpoint separation*

$$S_t := \mathrm{TV}\big(K_t^{\omega_0,\Pi}, K_t^{\omega_1,\Pi}\big) = \mathrm{TV}\big(\mathsf{K}(\cdot \mid \mathrm{BR}_t(\bar{u}_t + \delta u_t, \omega_0),\, u_t),\; \mathsf{K}(\cdot \mid \mathrm{BR}_t(\bar{u}_t + \delta u_t, \omega_1),\, u_t)\big). \tag{2}$$

It is $\mathcal{F}_{t-1}$-measurable, vanishes when $\delta u_t = 0$ by Assumption 2, and is exactly what an intervention measures. Our theorem is stated for the accumulated square $\sum_t S_t^2$; everything else is derived from it.

The *strategic observability modulus* is that separation per unit of motion,

$$\mathfrak{O}_t \;:=\; \frac{S_t}{\sqrt{C_{\mathrm{up}}}\,\|\delta u_t\|_\beta}, \qquad \mathfrak{O}_t := 0 \text{ when } \delta u_t = 0, \tag{3}$$

normalised by $\sqrt{C_{\mathrm{up}}}$ (Assumption 5) so that $\mathfrak{O}_t \in [0,1]$. This is a definition, not an approximation: the identity $S_t = \mathfrak{O}_t\sqrt{C_{\mathrm{up}}}\,\|\delta u_t\|_\beta$ holds by construction, with no remainder and no small-step requirement. Reading it as a limiting derivative would reintroduce an $o(\|\delta u_t\|_\beta)$ term that cannot be discarded when steps are not asymptotically small.

Remark 20 in Appendix O explains why the modulus must be directional: an operator norm is a supremum, reports what the *best* direction would achieve, and would credit the exploration mirage of §4.2 with capacity it does not have.

The one-step drift $\Delta_t^{\mathrm{KL}}(\Pi)$ therefore admits an exact partition:

$$\Delta_t^{\mathrm{KL}}(\Pi) \;=\; \underbrace{\mathfrak{O}_t^2\,\Delta_t^{\mathrm{KL}}(\Pi)}_{\text{signal: informative motion}} \;+\; \underbrace{(1-\mathfrak{O}_t^2)\,\Delta_t^{\mathrm{KL}}(\Pi)}_{\text{waste: kernel-projected motion}}\;. \tag{4}$$

Summing the signal term gives the central object of this paper, the *strategic capacity*

$$\mathcal{C}_T(\Pi;\,\omega_0,\omega_1) := \sum_{t=1}^{T} \mathfrak{O}_t^2\;\Delta_t^{\mathrm{KL}}(\Pi), \tag{5}$$

and the *waste component* $\mathcal{W}_T(\Pi;\,\omega_0,\omega_1) := \sum_{t=1}^{T}(1-\mathfrak{O}_t^2)\,\Delta_t^{\mathrm{KL}}(\Pi) \geq 0$. Three regimes are immediate: $\mathfrak{O}_t = 0$ (all motion is informationally inert), $\mathfrak{O}_t > 0, \Delta_t > 0$ (effective exploration), and $\Delta_t \to 0$ (collapse). SIB measures total motion; strategic capacity measures the motion that clears the observation channel; $\mathcal{W}_T$ measures the motion the channel cannot resolve.

The partition (4) is algebraically immediate, any weight in $[0,1]$ splits drift into two nonnegative parts. The content is not the split but that the first term tracks information: Theorem 1 shows this particular weight makes the signal share track mutual information across strategic feedback systems. Under local smoothness a *distinct* geometric decomposition has a squared-cosine weight (§3.2); the finite-step weight $\mathfrak{O}_t^2$ is comparable to that geometric quantity in the small-displacement regime, not identical with it. The *strategic capacity efficiency* $\eta_T(\Pi) := \mathcal{C}_T(\Pi)/\mathrm{SIB}_T(\Pi) \in [0,1]$ is a normalised surrogate for the fraction of exploration effort that converts to knowledge, making the question "*what fraction of this system's exploration is producing knowledge?*" a number. It lies in $[0,1]$ only because $\mathfrak{O}_t$ carries the $\sqrt{C_{\mathrm{up}}}$ normalisation; every estimator here reports the unnormalised $\widetilde{\eta}_T$, which can exceed one.

3 Main Results

3.1 The Conservation Law: Motion Partitions into Signal and Waste

Is there a fundamental limit on what a learner can know when the data-generating process responds to the learner? Performative prediction characterizes where such systems converge; no prior result characterizes what they can learn. The conservation law states that limit.

The theorem assumes dominated predictive kernels (Assumption 1), the no-static-signature condition at the baseline (Assumption 2: responses coincide across environments at $\bar{u}_t$, so separation comes from the displacement alone), and reverse-Pinsker regularity (Assumption 6), which supplies the upper constant. It does *not* assume drift separability $\mathbb{E}[\mathrm{KL}(K_t^{\omega,\Pi}\|K_t^{\omega',\Pi})] \leq C\,\Delta_t^{\mathrm{KL}}$ directly; Proposition 6 derives that from the primitive continuity conditions. Passing to capacity additionally uses Assumption 5.

Theorem 1 (Information tracks accumulated separation)**.** *Let $\Omega = \{\omega_0,\omega_1\}$ be binary with a uniform prior; see Remark 21 for $|\Omega| > 2$. Write $\mathcal{S}_T := \sum_{t=1}^{T} S_t^2$ for the accumulated squared endpoint separation (2). Under Assumptions 1–2 and 6, for every policy Π,*

$$c\,\min\!\big\{\operatorname{ess\,inf}\mathcal{S}_T,\;1\big\} \;\leq\; I(\omega;Y_{1:T}\mid\Pi) \;\leq\; C\,\mathbb{E}\big[\mathcal{S}_T\big], \tag{6}$$

with c absolute and $C = \kappa$ the reverse-Pinsker constant of Assumption 6.

Three features are deliberate. The sum runs over all T observations, matching the chain rule, so no boundary convention is needed. The two sides carry different qualifiers because $\mathcal{S}_T$ is random: information is an expectation and is bounded by the expected accumulation, while the lower bound runs through a pathwise argument (Lemma 8) and needs an almost-sure floor. And the theorem is stated for $\mathcal{S}_T$ rather than capacity, because $\mathcal{S}_T$ is what the proof controls and what an intervention measures; capacity enters as a corollary.

3.2 The Conservation Law as an Orthogonal Decomposition

The identity $\mathrm{SIB}_T = \mathcal{C}_T + \mathcal{W}_T$ is bookkeeping: $\mathcal{W}_T$ is a residual, and any weight in $[0,1]$ would produce one. Under smoothness there is a nearby triple for which it is Pythagoras, and this subsection is the substantive conservation law of the paper. Two qualifications belong in the statement rather than after it: the exact decomposition holds for an *environment-averaged* cost, not for the supremum defining Δ_t^{KL}, and the geometric capacity is *comparable to*, not equal to, the finite-step capacity (5).

Local geometry. Fix t, suppose the map from continuation to observation kernel is differentiable at $\bar{u}_t$, and write D_t^ω for its derivative under environment ω. By Assumption 2 both environments induce the same kernel at the baseline, so there is one base point and one tangent space, and only the derivatives differ. Set $\bar{D}_t := \frac{1}{2}(D_t^{\omega_0} + D_t^{\omega_1})$ and $J_t := D_t^{\omega_0} - D_t^{\omega_1}$, the *identification operator*. Equip the tangent space with the Fisher metric G_t, under which local KL is half the squared norm. With $v := \delta u_t$ the step cost under ω is $\frac{1}{2}\|D_t^\omega v\|_{G_t}^2$, and since $D_t^{\omega_{0,1}} v = \bar{D}_t v \pm \frac{1}{2} J_t v$ it depends on $J_t v$, not on the common motion alone. A supremum over ω is a maximum of two quadratic forms, hence not quadratic, and a non-quadratic cost has no orthogonal decomposition. We therefore average:

$$G_{\Delta,t} := \tfrac{1}{2}\big[(D_t^{\omega_0})^* G_t D_t^{\omega_0} + (D_t^{\omega_1})^* G_t D_t^{\omega_1}\big] = \bar{D}_t^* G_t \bar{D}_t + \tfrac{1}{4} J_t^* G_t J_t, \quad q_t(v) := \tfrac{1}{2} v^\top G_{\Delta,t}\, v, \tag{7}$$

and $\|v\|_{\Delta,t}^2 := q_t(v)$. Keep two objects apart: the exact averaged cost $\bar{\Delta}_t^{\mathrm{KL}} := \frac{1}{2}\sum_\omega \mathrm{KL}(\bar{K}_t^\omega \| K_t^{\omega,\Pi})$ and its quadratic part q_t, with $\bar{\Delta}_t^{\mathrm{KL}} = q_t(\delta u_t) + o(\|\delta u_t\|^2)$. The exact decomposition is about q_t; statements about $\bar{\Delta}_t^{\mathrm{KL}}$ carry the remainder. Little is lost: Ω is binary, so $\bar{\Delta}_t^{\mathrm{KL}} \le \Delta_t^{\mathrm{KL}} \le 2\bar{\Delta}_t^{\mathrm{KL}}$ to second order. We assume *nondegeneracy*, $\ker \bar{D}_t \cap \ker J_t = \{0\}$, a displacement outside it changes nothing under either environment which is exactly positive definiteness of $G_{\Delta,t}$, and holds in cases where the common-motion pullback $\langle \bar{D}_t \cdot, \bar{D}_t \cdot \rangle_{G_t}$ is degenerate.

Theorem 2 (Conservation as Pythagoras)**.** *Let $P_{\mathrm{id},t}, P_{\mathrm{null},t}$ be the $\langle\cdot,\cdot\rangle_{\Delta,t}$-orthogonal projections onto $(\ker J_t)^{\perp_\Delta}$ and $\ker J_t$, and set $\mathcal{C}_t^{\mathrm{geo}} := \|P_{\mathrm{id},t}\delta u_t\|_{\Delta,t}^2$, $\mathcal{W}_t^{\mathrm{geo}} := \|P_{\mathrm{null},t}\delta u_t\|_{\Delta,t}^2$. Then $q_t(\delta u_t) = \mathcal{C}_t^{\mathrm{geo}} + \mathcal{W}_t^{\mathrm{geo}}$ exactly, $\mathcal{C}_t^{\mathrm{geo}}/q_t(\delta u_t) = \cos^2\theta_t \in [0,1]$ for θ_t the angle between the realised motion and the identifiable subspace, and J_t annihilates the null component.*

Waste is thus a null-space component, not a leftover: motion both environments answer identically. What is not automatic is that the identifiable component is what the observer learns from.

Proposition 3 (Separation is comparable to geometric capacity)**.** *Let $\sigma_{\min,t}, \sigma_{\max,t}$ be the extreme singular values of J_t on $(\ker J_t)^{\perp_\Delta}$, from $\langle\cdot,\cdot\rangle_{\Delta,t}$ to G_t. Then $\sigma_{\min,t}^2\,\mathcal{C}_t^{\mathrm{geo}} \le \|J_t\delta u_t\|_{G_t}^2 \le \sigma_{\max,t}^2\,\mathcal{C}_t^{\mathrm{geo}}$, and always $\sigma_{\max,t}^2 \le 8$, a bound that is attained. If moreover $\sigma_{\min,t}^2 \ge \sigma^2 > 0$ for all t (*uniform conditioning*), then $I(\omega; Y_{1:T} \mid \Pi) \asymp \sum_t \mathcal{C}_t^{\mathrm{geo}}$ in the local regime, with constants σ^2, 8 and the reverse-Pinsker constant.*

The asymmetry, stated literally: $\|J_t\delta u_t\|_{G_t}^2 \le 8\,\mathcal{C}_t^{\mathrm{geo}}$ needs no hypothesis, because the averaged cost already contains $\frac{1}{4}J_t^* G_t J_t$, while $\|J_t\delta u_t\|_{G_t}^2 \ge \sigma^2\,\mathcal{C}_t^{\mathrm{geo}}$ is an assumption on the conditioning of J_t and fails as it becomes singular on its own co-kernel.

What this does not say. It does not say $\mathfrak{O}_t^2 = \cos^2\theta_t$: the finite-step modulus is a secant total-variation rate, $\cos\theta_t$ an exact angle, and they are equal only under a conformality condition on the *pair* $(\bar{D}_t, J_t)$ that we do not assume conformality of J_t alone does not suffice. Nor may Δ_t be read off the common-motion pullback, which ignores $J_t\delta u_t$ and is degenerate whenever $\bar{D}_t$ is. Appendix P quantifies both failures and verifies the statements above on 4000 random instances (`verify_geometry.py`).

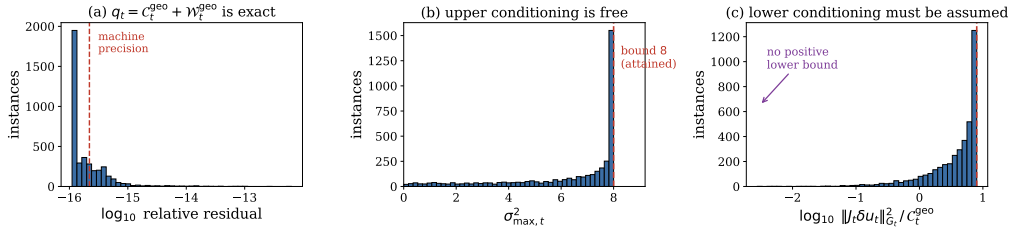


Figure 2: **The local geometry, checked numerically** on 4000 random instances $(G_t, \bar{D}_t, J_t, \delta u_t)$ satisfying the nondegeneracy condition (9.8% of raw draws violate it and are rejected). *(a)* The decomposition of Theorem 2 is exact: the relative residual never exceeds 6.3×10^{-13}. *(b)* The upper conditioning constant obeys $\sigma^2_{\max,t} \leq 8$ with no violations and mass piled at the boundary, so that the direction of Proposition 3 needs no hypothesis and the constant is tight. *(c)* The ratio $\|J_t \delta u_t\|^2_{G_t} / \mathcal{C}_t^{\mathrm{geo}}$ spans $[0.0028, 8]$ with a long tail towards zero: there is no positive uniform lower bound, which is why uniform conditioning is assumed rather than derived, and why the equivalence degrades as J_t becomes singular on its own co-kernel (`make_fig_geometry_emp.py`).

Relation to the finite-step theorem. The hierarchy is: an exact law for $(q_t, \mathcal{C}_t^{\mathrm{geo}}, \mathcal{W}_t^{\mathrm{geo}})$, which needs a derivative; an equivalence with information up to the conditioning of J_t; and a measurable surrogate $(\Delta_t^{\mathrm{KL}}, \mathcal{C}_T, \mathcal{W}_T)$ that needs no differentiability, which is what makes it measurable by intervention, and is comparable to the first within the factor 2 of the averaging bridge, the singular-value constants and the second-order remainders, in the local small-displacement regime. The experiments report the surrogate, the only one an intervention can compute.

Theorem 1 is, on its own, an information-accumulation result of a familiar kind: binary mutual information comparable to accumulated squared conditional separation and we do not claim the comparison as new. What is new is the reading the strategic setting forces on it: the separation at each step is produced by a displacement the policy chose and paid for, so the accumulation resolves into a cost and a conversion rate, and the rate can be zero while the spend is large (Remark **??**, Appendix O).

Corollary 4 (Conservation, and capacity as the tracked quantity)**.** *Add Assumption 5. Then the drift partitions exactly,*

$$\mathrm{SIB}_T^{\mathrm{KL}}(\Pi) \;=\; \mathcal{C}_T(\Pi; \omega_0, \omega_1) + \mathcal{W}_T(\Pi; \omega_0, \omega_1), \qquad \mathcal{W}_T := \sum_t (1 - \mathfrak{O}_t^2)\, \Delta_t^{\mathrm{KL}} \;\geq\; 0,$$

and capacity is equivalent to accumulated separation, $\mathcal{C}_T \leq \mathcal{S}_T \leq (C_{\mathrm{up}}/c_{\mathrm{low}})\, \mathcal{C}_T$ pathwise, so (6) holds with $\mathcal{S}_T$ replaced by $\mathcal{C}_T$ and the constants inflated by $C_{\mathrm{up}}/c_{\mathrm{low}}$.

Proof. The partition is the identity $\Delta_t = \mathfrak{O}_t^2 \Delta_t + (1 - \mathfrak{O}_t^2)\Delta_t$ summed over t, non-negative because $\mathfrak{O}_t \in [0, 1]$; it is algebraic and carries no content by itself. For the equivalence, $\mathfrak{O}_t$ is defined by (3) so that $S_t^2 = \mathfrak{O}_t^2\, C_{\mathrm{up}} \|\delta u_t\|_\beta^2$ identically, and Assumption 5 bounds $\|\delta u_t\|_\beta^2$ between $\Delta_t^{\mathrm{KL}}/C_{\mathrm{up}}$ and $\Delta_t^{\mathrm{KL}}/c_{\mathrm{low}}$. □

The truncation on the left is not a technicality. With $|\Omega| = 2$ we have $I(\omega; Y_{1:T} \mid \Pi) \leq \log 2$ for every T, while $\mathcal{S}_T$ grows with the horizon, so no lower bound linear in $\mathcal{S}_T$ can hold once $\mathcal{S}_T \gtrsim 1$. Accumulated separation and information agree in the identification-limited regime $\mathcal{S}_T = O(1)$ and separation continues to accumulate after information has saturated. Via Corollary 4 the same holds for capacity, and in particular waste generates no information. Full proof in Appendix D.

Remark 21 in Appendix O treats $|\Omega| > 2$. The one-sided bound $I \leq C\, \mathrm{SIB}$ follows immediately from $\mathrm{SIB} = \mathcal{C} + \mathcal{W} \geq \mathcal{C}$ and (6). Remark 22 in Appendix O shows no other exponent admits the identity, independently of the squared-norm argument of §3.2. Two immediate corollaries: the bound form $I \leq C\, \mathbb{E}[\mathrm{SIB}_T^{\mathrm{KL}}]$, with a matching lower bound when the waste vanishes and capacity has a floor, and a Fano-type error floor are in Appendix O.

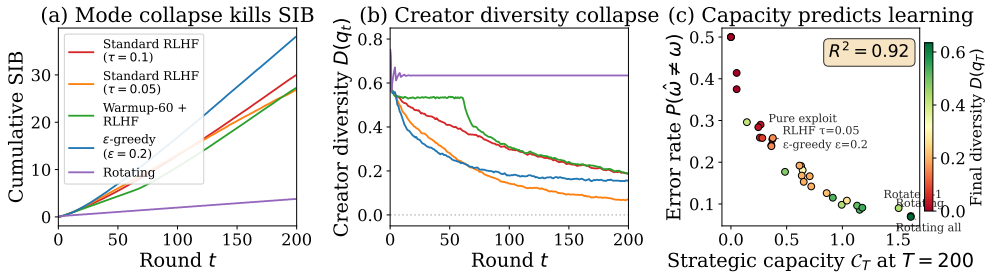


Figure 3: **A strategic RLHF model, where we wrote the response map.** *(a)* Every policy accumulates budget, and the aggressive ones accumulate it fastest. *(b)* Creator diversity nonetheless collapses under exploitation, so the responses pool across environments and the preference signal is silenced. *(c)* Identification error is predicted by separation energy ($R^2 = 0.92$ over the 29 policies clearing the $\widetilde{\mathcal{C}}_T > 0.01$ threshold used for the fit) and not by the budget ($R^2 = 0.07$), which is the dissociation the theory asks for. Colour encodes final diversity (`_rlhf_regen.py`). The response map is ours, so this tests the mechanism, not any claim about deployed systems.

4 Experiments

Three experiments appear here and the rest in Appendix S: whether standard diagnostics can be fooled while capacity cannot (§4.2), whether the modulus can be measured by intervention rather than inferred (§4.3), and whether the dissociation survives in model-internal studies (§4.1).

4.1 Synthetic Models: Strategic Capacity Predicts Error, SIB Does Not

Two model-internal studies, reported in full in Appendix S, establish the basic dissociation. Figure 3 shows the first: in a strategic RLHF model, aggressive exploitation collapses creator diversity and silences the preference signal, and separation energy predicts identification error across the policy sweep at $R^2 = 0.92$ while raw budget gives $R^2 = 0.07$. In an iterative LLM preference loop, blind and targeted perturbation spend budgets within a factor of 1.5 (8.86 against 13.09) with evaluator diversity equal to three decimals (0.476), yet differ by a factor of ten in information gain and eleven in separation energy.

4.2 Exploration Mirage: Budget Held Exactly Fixed

The previous experiments use $|\Omega| = 2$. We upgrade to $|\Omega| = 4$ with *hierarchical distinguishability*: environments split into a coarse partition (separable via styles s_A, s_B) and fine partitions (requiring s_A, s_C or s_B, s_D). A policy varying only along the coarse axis resolves half the uncertainty and stalls on the fine splits. Full specification in Appendix M.1.

The structure permits a counterexample in which every motion quantity we report is held exactly fixed (Figure 1). Consider the six policies that strictly alternate between two of the four styles: same switching frequency, exposure counts, freshness trajectory and action entropy, and, by Proposition 16, as an exact consequence of the design rather than approximately, identical budget and identical average pairwise separation energy. They differ in one quantity, the minimum over hypothesis pairs, and that is what predicts whether they learn.

Two scope notes. The plotted quantity is separation energy $\widetilde{\mathcal{C}}_T = \sum_t S_t^2$ with S_t the exact Gaussian total variation (14), not $\mathcal{C}_T$, which needs the unidentified C_{up}. And Assumption 2 fails globally here (θ_s^ω depends on ω at fixed s), but for each zero-minimum policy both played styles respond identically across its unresolved pair, so it holds for *that* pair, which is what Corollary 36 requires.

4.3 An Interventional Test in a Real Strategic System

Observational logs record no counterfactual, we never see how the responder would have behaved under a different continuation, which is exactly what $\mathfrak{O}_t$ compares, so the remedy is a different

Styles alternated	SIB_T	$\widetilde{\mathcal{C}}_T^{\mathrm{avg}}$	$\widetilde{\mathcal{C}}_T^{\min}$	Info (bits)	Error
$s_A \leftrightarrow s_B$	6.0714	1.5949	**0.000**	0.934	0.507 ± 0.011
$s_C \leftrightarrow s_D$	6.0714	1.5949	**0.000**	0.935	0.503 ± 0.011
$s_A \leftrightarrow s_C$	6.0714	1.5949	1.196	1.529	0.139 ± 0.008
$s_A \leftrightarrow s_D$	6.0714	1.5949	1.196	1.528	0.125 ± 0.007
$s_B \leftrightarrow s_C$	6.0714	1.5949	1.196	1.541	0.131 ± 0.008
$s_B \leftrightarrow s_D$	6.0714	1.5949	1.196	1.540	0.113 ± 0.007

Table 1: **A budget-matched mirage.** All six two-style alternations ($|\Omega| = 4$, $T = 200$, 2,000 runs; `matched_pair_exacttv.py`). Budget and average separation energy are *exactly* equal across all six observed spread 0 in both, and equal by Proposition 16 rather than by coincidence as are switching frequency, exposure counts and action entropy. The only quantity that differs is the binding pair $\widetilde{\mathcal{C}}_T^{\min}$, and it separates the two groups completely: zero gives chance error, positive gives error below 0.14. None of the motion diagnostics we report distinguishes these policies; the minimum over hypothesis pairs does. The $\widetilde{\mathcal{C}}_T$ columns are the unnormalised accumulated squared separation $\sum_t S_t^2$ with S_t the exact Gaussian total variation (14), not the defined capacity $\mathcal{C}_T$ (Appendix M.16).

policy	steps	SIB_T	$\widetilde{\mathcal{C}}_T$	$\widetilde{\eta}_T$	id. error
TARGETED	13	42.98	17.409	0.405	0.000
BLIND	13	61.94	0.076	0.001	0.343

Table 2: Interventional test with `Qwen2.5-7B-Instruct` as the responder; chance error is 0.500. The modulus is measured by perturbation, not estimated from a proxy. Columns are unnormalised as in Table 1.

experiment, not a better statistic. Here the responder is a real language model whose behaviour we do not design but whose response map we may *query*. A recommender announces a linear promotion rule weighting technical depth and length; a creator of latent type $\omega \in \{\omega_0, \omega_1\}$, the types differing only in the cost of depth, reports the depth and length of its next post; and the learner observes the axis its own policy promotes, as a recommender sees engagement only on what it surfaces. Two policies alternate their plan at the *same per-step drift*, TARGETED varies the depth weight, BLIND the length and at each step we perturb the realised continuation by δ along its own direction and divide the induced separation by δ. That is the unnormalised directional rate underlying $\mathfrak{O}_t$, not the normalised modulus of (3), which would need C_{up} and $\|\cdot\|_\beta$. Over 13 matched steps (Table 2) BLIND spends 44% *more* motion and accumulates 0.4% of the separation energy, erring at 0.343 against 0.000, chance being 0.500; every motion diagnostic ranks it as exploring more. Full protocol in Appendix S.

4.4 A Parameter-Free Conservation Prediction

The preceding experiments test whether capacity orders policies. A separate question is whether the theory predicts a quantitative level without fitting a conversion coefficient. We study a decentralized learning commons in which N agents privately choose information-producing effort while sharing the resulting posterior. Proposition 37 predicts that equilibrium and socially optimal precision grow with the same exponent, but that the planner creates a log-precision advantage of exactly $\frac{1}{3}\log N + o(1)$. This gap contains no fitted parameter.

Across $N \in \{2, 5, 10, 20, 50\}$ and 18 combinations of discounting and noise, the measured wedge agrees with $\frac{1}{3}\log N$ to within 2.6×10^{-4} (Table 3). The result is stronger than a correlation between budget and learning: it tests the level of a comparative prediction fixed before simulation. It also gives the conservation law an economic interpretation. Information-producing motion is a public good, and decentralization changes how much of it is financed without changing the underlying growth exponent. Appendix E gives the model and proof, and Appendix S gives the simulation protocol.

N	measured wedge	$\frac{1}{3}\log N$	error
2	0.23083	0.23105	-2.2×10^{-4}
5	0.53622	0.53648	-2.6×10^{-4}
10	0.76729	0.76753	-2.4×10^{-4}
20	0.99837	0.99858	-2.1×10^{-4}
50	1.30385	1.30401	-1.6×10^{-4}

Table 3: **A parameter-free prediction.** The measured equilibrium–planner wedge in log precision, averaged over 18 discounting/noise cells at each N, matches the $\frac{1}{3}\log N$ prediction of Proposition 37.

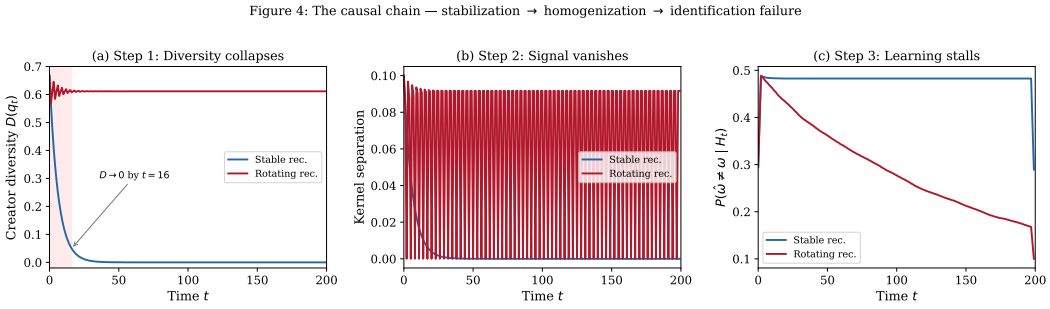


Figure 4: **The causal chain of strategic collapse.** Under stable recommendations, (a) creator diversity vanishes, (b) cross-environment kernel separation follows, and (c) identification stalls near chance. Rotation preserves diversity and the observation channel. The model uses $K = 3$ content types, $T = 200$, and 3,000 Monte Carlo runs.

4.5 Stabilization Destroys the Observation Channel

Strategic capacity also predicts a dynamic failure mode: a policy can continue to act while its induced observation channel becomes uninformative. In a recommender model with adaptive creators, stable recommendations first collapse creator diversity, then eliminate cross-environment kernel separation, and finally halt identification (Figure 4). Rotating recommendations preserve diversity and reduce error to 0.17, while the stable policy remains near chance. The ordering of the panels matters: the experiment exhibits the mechanism proposed by Theorem 11, rather than only an endpoint correlation between motion and error.

4.6 The Dissociation Persists with Weak Preference Signals

The matched construction uses an idealized preference matrix to isolate the mechanism exactly. To test whether the conclusion depends on that structure, we replace it with a matrix estimated from summarize-from-feedback comparisons: three latent preference criteria over nine length–extractiveness buckets, with signal spread only ± 0.12 around chance. Across four noise levels and five independent master seeds per level, separation energy remains a better predictor of information gain than SIB, with $R^2_{\widetilde{C}_T} - R^2_{\mathrm{SIB}} \in [0.07, 0.11]$ and non-overlapping 95% confidence intervals throughout. The absolute fit is lower than in the idealized model, as expected under weak signal, but the ordering of the two diagnostics is stable. This connects the exact counterexample to a substantially noisier preference structure derived from human-feedback data.

5 Discussion and Future Work

Our results suggest that exploration in strategic systems should be evaluated by the information preserved after the environment responds, not by policy motion alone. Strategic capacity provides a common language for diagnosing this distinction across learning settings. Future work should develop scalable estimators for capacity, extend the theory beyond binary and locally smooth environments, and design policies that actively direct exploration toward identifiable dimensions.

AI USE STATEMENT

We used generative AI tools (Claude) to edit manuscript, write and review the code, to draft and check several appendix derivations, and to assist with running and monitoring experiments. The research questions, the method, and the experimental design are the authors'. All derivations were verified by the authors, all numbers are generated from logged results by scripts the authors reviewed, and the authors take full responsibility for the content.

REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ShjMHfmPs0.

D. Bergemann and S. Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

D. Bergemann and J. Välimäki. Information acquisition and efficient mechanism design. *Econometrica*, 70(3):1007–1033, 2002.

D. Bergemann and J. Välimäki. Dynamic mechanism design: An introduction. *Journal of Economic Literature*, 57(2):235–274, 2019.

P. Bolton and C. Harris. Strategic experimentation. *Econometrica*, 67(2):349–374, 1999.

V. S. Borkar and P. P. Varaiya. Adaptive control of Markov chains, I: Finite parameter set. *IEEE Trans. Automatic Control*, 24(6):953–957, 1979.

G. Brown, S. Hod, and I. Kalemaj. Performative prediction in a stateful world. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6045–6061, 2022. URL https://proceedings.mlr.press/v151/brown22a.html.

S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023. arXiv:2307.15217, 2023.

A. J. B. Chaney, B. M. Stewart, and B. E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *RecSys*, pages 224–232, 2018.

Y.-K. Che and J. Hörner. Optimal design for social learning. Technical Report 2000, Cowles Foundation for Research in Economics, Yale University, 2015.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, pages 4299–4307, 2017.

D. Drusvyatskiy and L. Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2023.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *ICML*, pages 1097–1104, 2011.

P. Frazier, D. Kempe, J. Kleinberg, and R. Kleinberg. Incentivizing exploration. In *EC*, pages 5–22, 2014.

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In *ITCS*, pages 111–122, 2016.

N. Immorlica, J. Mao, A. Slivkins, and Z. S. Wu. Bayesian exploration with heterogeneous agents. In *The Web Conference*, pages 751–761, 2020.

Z. Izzo, L. Ying, and J. Zou. How to learn when data reacts to your model: Performative gradient descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4641–4650, 2021.

E. Kamenica and M. Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

G. Keller, S. Rady, and M. Cripps. Strategic experimentation with exponential bandits. *Econometrica*, 73(1):39–68, 2005.

J. Kirschner and A. Krause. Information directed sampling and bandits with heteroscedastic noise. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 358–384, 2018.

I. Kremer, Y. Mansour, and M. Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012, 2014.

P. R. Kumar and A. Becker. A new family of optimal adaptive controllers for Markov chains. *IEEE Trans. Automatic Control*, 27(1):137–146, 1982.

T. L. Lai and C. Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

P. Mandl. Estimation and control in Markov chains. *Advances in Applied Probability*, 6(1):40–60, 1974.

D. Manheim and S. Garrabrant. Categorizing variants of Goodhart's law, 2018. arXiv:1803.04585, 2018.

Y. Mansour, A. Slivkins, and V. Syrgkanis. Bayesian incentive-compatible bandit exploration. *Operations Research*, 68(4):1132–1161, 2020.

C. Mendler-Dünner, J. Perdomo, T. Zrnic, and M. Hardt. Stochastic optimization for performative prediction. In *NeurIPS*, 2020.

J. Miller, S. Milli, and M. Hardt. Strategic classification is causal modeling in disguise. In *ICML*, pages 6917–6926, 2020.

G. Neu, M. Papini, and L. Schwartz. Optimistic information-directed sampling. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 3970–4006, 2024. URL https://proceedings.mlr.press/v247/neu24a.html.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

A. Pan, E. Jones, M. Jagadeesan, and J. Steinhardt. Feedback loops with language models drive in-context reward hacking. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39154–39200, 2024. URL https://proceedings.mlr.press/v235/pan24d.html.

J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *ICML*, pages 7599–7609, 2020.

D. Precup, R. S. Sutton, and S. P. Singh. Eligibility traces for off-policy policy evaluation. In *ICML*, pages 759–766, 2000.

D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In *NeurIPS*, pages 1583–1591, 2014.

D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *JMLR*, 17 (68):1–30, 2016.

M. Sellke and A. Slivkins. Incentivizing exploration with selective data disclosure. In *EC*, 2023.

I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024. doi: 10.1038/s41586-024-07566-y.

J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger. Defining and characterizing reward hacking. In *NeurIPS*, 2022.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

N. N. Čencov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, 1982.

A. Wald. *Sequential Analysis*. John Wiley & Sons, 1947.

Y. Xu and A. Zeevi. Bayesian design principles for frequentist sequential learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38768–38800, 2023. URL https://proceedings.mlr.press/v202/xu23u.html.

B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

A NOTATION

Quantities are grouped by what they are: an exact object of the theory, its local quadratic counterpart, or an unnormalised empirical estimate. A tilde always marks the last of these.

Symbol	Kind	Meaning
Finite-step objects (no differentiability required)		
$\bar{u}_t,\ \delta u_t,\ u_t$	path	baseline continuation, displacement, enacted endpoint; $u_t = \bar{u}_t + \delta u_t$
$K_t^{\omega,\Pi}$	kernel	endpoint kernel $\mathsf{K}(\cdot \mid \mathrm{BR}_t(u_t,\omega), u_t)$
$\bar{K}_t^{\omega}$	kernel	kernel the baseline would have induced
S_t	random	realised endpoint separation $\mathrm{TV}(K_t^{\omega_0,\Pi}, K_t^{\omega_1,\Pi})$
$\mathcal{S}_T$	random	accumulated squared separation $\sum_t S_t^2$
Δ_t^{KL}	random	step cost $\sup_\omega \mathbb{E}[\mathrm{KL}(\bar{K}_t^{\omega} \| K_t^{\omega,\Pi}) \mid \mathcal{F}_{t-1}]$
$\mathrm{SIB}_T^{\mathrm{KL}}$	random	budget $\sum_{t=1}^T \Delta_t^{\mathrm{KL}}$
$\mathfrak{O}_t$	random	observability modulus $S_t/(\sqrt{C_{\mathrm{up}}}\|\delta u_t\|_\beta) \in [0,1]$
$\mathcal{C}_T,\ \mathcal{W}_T$	random	capacity $\sum_t \mathfrak{O}_t^2 \Delta_t^{\mathrm{KL}}$ and waste $\sum_t (1-\mathfrak{O}_t^2)\Delta_t^{\mathrm{KL}}$
η_T	random	efficiency $\mathcal{C}_T/\mathrm{SIB}_T \in [0,1]$
$\bar{\mathcal{C}}_T,\ \mathcal{C}_T^{\min}$	random	average and minimum pairwise capacity over $\binom{M}{2}$ pairs
Local geometric objects (require smoothness; §3.2)		
$\bar{D}_t,\ J_t$	operator	common motion $\frac{1}{2}(D_t^{\omega_0} + D_t^{\omega_1})$; identification operator $D_t^{\omega_0} - D_t^{\omega_1}$
$G_t,\ G_{\Delta,t}$	metric	Fisher metric on the kernel tangent; averaged cost metric (7)
$\bar{\Delta}_t^{\mathrm{KL}}$	exact	environment-averaged KL cost
$q_t(v)$	quadratic	its quadratic part $\frac{1}{2}v^\top G_{\Delta,t} v$; $\bar{\Delta}_t^{\mathrm{KL}} = q_t + o(\|\delta u_t\|^2)$
$\mathcal{C}_t^{\mathrm{geo}},\ \mathcal{W}_t^{\mathrm{geo}}$	quadratic	$\|P_{\mathrm{id},t}\delta u_t\|_{\Delta,t}^2$ and $\|P_{\mathrm{null},t}\delta u_t\|_{\Delta,t}^2$; sum to q_t exactly
θ_t	angle	between δu_t and $(\ker J_t)^{\perp_\Delta}$; $\mathcal{C}_t^{\mathrm{geo}}/q_t = \cos^2\theta_t$
$\sigma_{\min,t}, \sigma_{\max,t}$	constant	extreme singular values of J_t on $(\ker J_t)^{\perp_\Delta}$; $\sigma_{\max,t}^2 \le 8$ always
$\|J_t\delta u_t\|_{G_t}^2$	quadratic	*separation energy*, the object Chentsov canonicalises
Frontiers and empirical estimates		
$\mathcal{F}_T(B)$	number	$\sup\{I(\omega; Y_{1:T} \mid \Pi) : \mathbb{E}[\mathrm{SIB}_T] \le B\}$
$\mathcal{C}_{\mathrm{mean}}^*,\ \mathcal{C}_{\mathrm{floor}}^*$	number	$\sup \mathbb{E}[\mathcal{C}_T]$ and $\sup \operatorname{ess\,inf} \mathcal{C}_T$ at budget B; may differ arbitrarily
α	constant	floor regularity: $\operatorname{ess\,inf} \mathcal{C}_T \ge \alpha \mathbb{E}[\mathcal{C}_T]$ (Definition 29)
$\widetilde{\mathcal{C}}_T,\ \widetilde{\mathcal{C}}_T^{\min}$	estimate	*unnormalised* $\sum_t S_t^2$; equals $\mathcal{C}_T$ only after dividing by C_{up}
$\widetilde{\eta}_T$	estimate	$\widetilde{\mathcal{C}}_T/\mathrm{SIB}_T$; need not lie in $[0,1]$
$\mathcal{M}_T$	estimate	mirage ratio $1/\widetilde{\eta}_T$

Three objects are deliberately kept apart throughout and should not be conflated: *separation energy* $\|J_t\delta u_t\|_{G_t}^2$, which the learning-currency axioms canonicalise; *cost geometry* q_t, which the cost axioms canonicalise and which decomposes orthogonally; and the *finite-step surrogate* $(\Delta_t^{\mathrm{KL}}, \mathcal{C}_T, \mathcal{W}_T)$, which is what an intervention can measure and which is comparable to the first two only in the local regime and only up to the conditioning of J_t.

B ASSUMPTIONS

Assumption 1 (Dominated predictive kernels)**.** For each t and policy Π, the family $\{K_t^{\omega,\Pi}(\cdot \mid h) : \omega \in \Omega\}$ is dominated by a common σ-finite measure.

Assumption 2 (No static signature at the baseline)**.** Let $\bar{u}_t$ be the *baseline* continuation at time t: the continuation before the step, which is never enacted. At the baseline the strategic responses coincide across environments,

$$\mathrm{BR}_t(\bar{u}_t, \omega) = \mathrm{BR}_t(\bar{u}_t, \omega') \qquad \text{for all } \omega, \omega',$$

while the responses at the enacted endpoint $u_t = \bar{u}_t + \delta u_t$ are permitted to differ. Environments are indistinguishable from a static snapshot and separate only through the displacement.

The condition is imposed at the baseline and nowhere else. Imposing it at the enacted endpoint would equate $K_t^{\omega_0,\Pi}$ and $K_t^{\omega_1,\Pi}$, force $S_t = 0$ at every step, and empty every result in this paper:

the observation kernel takes the strategic state and control as its only arguments, so responses that agree at a path induce identical kernels there. Imposing it at *every* path is worse still, since functions agreeing everywhere have equal derivatives and the condition would then also contradict its own reading, that environments differ in how they respond to variation.

Assumption 3 (Strategic response decomposition)**.** There exist $L_{\mathrm{dir}}, L_{\mathrm{cont}} \geq 0$ and $\rho \in (0,1)$ such that $d_X(\mathrm{BR}_t(u,\omega), \mathrm{BR}_t(u',\omega')) \leq L_{\mathrm{dir}}\, d_\Omega(\omega,\omega') + L_{\mathrm{cont}} \sum_{s\geq t} \rho^{s-t} d_U(u_s, u'_s)$.

Assumption 4 (KL-Lipschitz observation kernels)**.** There exist $G_x, G_u < \infty$ such that $\mathrm{KL}(\mathsf{K}(\cdot|x,u)\|\mathsf{K}(\cdot|x',u')) \leq G_x\, d_X(x,x') + G_u\, d_U(u,u')$.

Assumption 5 (Two-sided local drift regularity)**.** Write $\delta u_t := u_t - \bar{u}_t$ for the realised displacement from the baseline continuation to the enacted endpoint, the same indexing as the step cost (1) and *not* the difference between consecutive continuations, which would give $T-1$ displacements against T observations and let $\|\cdot\|_\beta$ be defined as follows. For each ω, history h_{t-1} and time t, let $N_t^\omega(h_{t-1}) := \{v : \mathsf{K}(\cdot \mid \mathrm{BR}_t(\bar{u}_t + sv, \omega), \bar{u}_t + sv) = \bar{K}_t^\omega \text{ for all small } s\}$ be the directions along which environment ω sees no change at all, and set

$$N_t(h_{t-1}) \;:=\; \bigcap_{\omega\in\Omega} N_t^\omega(h_{t-1})$$

the *common* null space, the directions invisible to *every* environment. We assume $N_t(h_{t-1}) \equiv N$ does not vary with t or the history, fix any norm on the quotient $\mathcal{U}/N$, and let $\|v\|_\beta$ be the quotient norm of v. This is one norm, not a family, and it is $\mathcal{F}_{t-1}$-measurable because it does not depend on the history at all. Taking the intersection rather than a single N_t^ω matters: motion invisible to one environment but not the other produces separation and must be charged for. Where the common null space does vary with the history, the results below hold with $\|\cdot\|_\beta$ replaced by a measurable seminorm dominating every quotient norm, at the cost of the constants; we do not pursue that generality. There exist $0 < c_{\mathrm{low}} \leq C_{\mathrm{up}} < \infty$, depending only on the channel, such that for every t *and every history* h_{t-1} equivalently, almost surely, not merely in expectation

$$c_{\mathrm{low}}\, \|\delta u_t\|_\beta^2 \;\leq\; \Delta_t^{\mathrm{KL}}(\Pi) \;\leq\; C_{\mathrm{up}}\, \|\delta u_t\|_\beta^2.$$

Both halves are assumptions. The upper half does not follow from Assumptions 3–4: Assumption 4 is linear in d_X and d_U whereas the bound here is quadratic in $\|\delta u_t\|_\beta$, and no relation between those metrics making the composition quadratic has been supplied. The upper half is what normalises $\mathfrak{O}_t$ into $[0,1]$, so it is load-bearing and is assumed outright. The lower half is a genuine additional restriction: it says the continuation-to-kernel map is a quadratic bi-Lipschitz embedding on its image, so that a policy cannot move an observable amount while paying arbitrarily little drift. It is this half that licenses passing from a drift budget back to a displacement magnitude, a step the one-sided bound of Assumption 4 does *not* support. Quotienting is essential rather than cosmetic: motion in the null space changes nothing observable, contributes zero to Δ_t^{KL}, and would falsify the lower bound if it were included. The condition can fail at points where the channel degenerates, where distinct continuations become observationally indistinguishable to second order and we do not claim it holds universally.

Assumption 6 (Reverse Pinsker regularity)**.** There exists $\kappa < \infty$, uniform over histories, such that $\mathrm{KL}(\mathsf{K}(\cdot|x,u)\|\mathsf{K}(\cdot|x',u)) \leq \kappa\, \mathrm{TV}(\mathsf{K}(\cdot|x,u), \mathsf{K}(\cdot|x',u))^2$. We state this as a hypothesis and do not offer the usual gloss. Uniformly bounded likelihood ratios do *not* suffice: they give a bound linear in total variation, and differences carried on events of vanishing probability can make $\mathrm{KL}/\mathrm{TV}^2$ diverge. A quadratic reverse bound needs more, for instance a finite alphabet whose atoms are bounded below uniformly, and κ then depends on that lower bound. Readers should treat Assumption 6 as a genuine restriction on the observation channel rather than a regularity formality.

The converse comparison requires no assumption: by Pinsker's inequality (Tsybakov, 2009), for all x, x', u,

$$\mathrm{KL}(\mathsf{K}(\cdot|x,u)\|\mathsf{K}(\cdot|x',u)) \;\geq\; 2\,\mathrm{TV}(\mathsf{K}(\cdot|x,u), \mathsf{K}(\cdot|x',u))^2. \tag{8}$$

Inequality (8) is universal rather than an imposed regularity condition, so any step whose proof uses only the $\mathrm{KL} \gtrsim \mathrm{TV}^2$ direction needs no lower-regularity hypothesis. The split runs by *direction*, not by result. In the frontier sandwich (Theorem 30) the *lower* direction avoids Assumption 6; the *upper* direction does not, because it passes from information to expected capacity through the upper half of (6), whose constant is the reverse-Pinsker κ. The necessity results of §R.2 use the upper direction and therefore inherit it too.

Remark 5 (Why these must be pathwise). Assumptions 5 and 6 are imposed for every history rather than after taking expectations. The lower bound of Theorem 1 runs through Lemma 8, which needs the accumulated conditional separation $A = \sum_t H_t^2$ to be bounded below almost surely (or, in the weaker form, with high probability, paying the deviation term $\sqrt{P^0(A < a)}$). Control of A in expectation alone is not enough: the event on which separation is small can carry most of the mass of the likelihood-ratio integral. We flag this because the two-sided bound is stated with an essential infimum on the left, and that qualifier is only meaningful if the underlying regularity holds history-by-history.

Assumption 7 (Continuation continuity of strategic responses)**.** There exist nondecreasing $\Gamma_{\mathrm{dir}}, \Gamma_{\mathrm{cont}}$ with $\Gamma_{\mathrm{dir}}(0) = \Gamma_{\mathrm{cont}}(0) = 0$ such that $d_X(\mathrm{BR}_t(u,\omega), \mathrm{BR}_t(u',\omega')) \le \Gamma_{\mathrm{dir}}(d_\Omega(\omega,\omega')) + \Gamma_{\mathrm{cont}}(d_{\mathrm{tail}}(u,u'))$.

Assumption 8 (Continuity of observation kernels)**.** There exists nondecreasing Γ_K with $\Gamma_K(0) = 0$ such that $\mathrm{TV}(\mathsf{K}(\cdot|x,u), \mathsf{K}(\cdot|x',u')) \le \Gamma_K(d_X(x,x') + d_U(u,u'))$.

C Sufficient Conditions for Drift-Controlled Separability

Proposition 6 (Endogenous kernel separability)**.** *Under Assumptions 3–4, for any policy* Π *and* ω, ω'*,*

$$\mathbb{E}[\mathrm{KL}(K_t^{\omega,\Pi} \| K_t^{\omega',\Pi})] \le \underbrace{G_x L_{\mathrm{dir}}\, d_\Omega(\omega,\omega')}_{\textit{direct}} + \underbrace{G_x L_{\mathrm{cont}} \sum_{s \ge t} \rho^{s-t} \mathbb{E}[d_U(u_s, u_{s+1})]}_{\textit{policy-mediated}}.$$

Remark 7 (What this does and does not give). Under Assumption 2 the direct term vanishes at the baseline, so cross-environment separation there is controlled entirely by policy motion, which is the qualitative content the theory needs. We emphasise two limits. First, this proposition does *not* establish Assumption 2: that is a structural condition on the response map at the baseline, and no bound on $\mathbb{E}[\mathrm{KL}]$ implies it. Second, the policy-mediated term is a discounted sum of future control increments, whereas the analytic form used in the proofs is stated against the single temporal drift Δ_t^{KL}. Passing from one to the other requires a summability condition relating the tail of the control path to its one-step increment, which we assume rather than derive; for geometrically contracting paths ($\rho < 1$ with increments decaying at least as fast) the two agree up to a factor $(1-\rho)^{-1}$, but we do not claim this holds generally.

D Full Proof of the Conservation Law (Theorem 1)

Proof of Theorem 1. Partition identity. By construction, for every t, $\mathfrak{O}_t^2 \in [0,1]$, so $\Delta_t^{\mathrm{KL}} = \mathfrak{O}_t^2\, \Delta_t^{\mathrm{KL}} + (1 - \mathfrak{O}_t^2)\, \Delta_t^{\mathrm{KL}}$ with both terms non-negative. Summing gives the partition of Corollary 4 with $\mathcal{W}_T \ge 0$ and equality to zero iff $\mathfrak{O}_t = 1$ on the support of Δ_t.

Upper bound on MI. By the MI chain rule,

$$I(\omega; Y_{1:T} \mid \Pi) = \sum_{t=1}^{T} \mathbb{E}[\mathrm{KL}(K_t^{\omega,\Pi} \| \bar{K}_t^{\Pi})].$$

Both sides carry T terms: the chain rule has one per observation, and $\mathcal{S}_T = \sum_{t=1}^T S_t^2$ has one separation per observation. No boundary convention is required. Because the step cost (1) is indexed by the displacement rather than by a transition between consecutive kernels, SIB_T and $\mathcal{C}_T$ also carry T terms, and the translation in Corollary 4 is term-by-term. By convexity of KL and Jensen, each term is at most $\sup_{\omega'} \mathbb{E}[\mathrm{KL}(K_t^{\omega,\Pi} \| K_t^{\omega',\Pi})]$; with $|\Omega| = 2$ the supremum is attained at the single competing hypothesis. Passing from a drift budget back to a displacement magnitude requires the *lower* half of Assumption 5: from $c_{\mathrm{low}} \|\delta u_t\|_\beta^2 \le \Delta_t^{\mathrm{KL}}$ we obtain $\|\delta u_t\|_\beta \le \sqrt{\Delta_t^{\mathrm{KL}} / c_{\mathrm{low}}}$. We flag this explicitly because Assumption 4, which controls KL *given* a displacement, does not support the converse inference, and using it in that direction would be a mistake. By Assumption 2 the cross-environment separation vanishes at $\delta = 0$, so the separation map is linear to first order, and by the definition of the modulus (3) as a finite difference along the realised step, the separation

$S_t := \mathrm{TV}(K_t^{\omega_0,\Pi}, K_t^{\omega_1,\Pi})$ equals $\mathfrak{O}_t\sqrt{C_{\mathrm{up}}}\|\delta u_t\|_\beta$ identically, whence

$$\mathfrak{O}_t^2\,\Delta_t^{\mathrm{KL}} \;\le\; S_t^2 \;\le\; \frac{C_{\mathrm{up}}}{c_{\mathrm{low}}}\,\mathfrak{O}_t^2\,\Delta_t^{\mathrm{KL}}. \tag{9}$$

The left inequality uses the *upper* half of Assumption 5 ($\Delta_t^{\mathrm{KL}} \le C_{\mathrm{up}}\|\delta u_t\|_\beta^2$, hence $\|\delta u_t\|_\beta^2 \ge \Delta_t^{\mathrm{KL}}/C_{\mathrm{up}}$); the right uses the lower half. Squaring the right-hand side and applying reverse Pinsker (Assumption 6),

$$\sup_{\omega'}\mathbb{E}[\mathrm{KL}(K_t^{\omega,\Pi}\|K_t^{\omega',\Pi})] \;\le\; \kappa\,S_t^2 \;\le\; \kappa\,\tfrac{C_{\mathrm{up}}}{c_{\mathrm{low}}}\,\mathfrak{O}_t^2\,\Delta_t^{\mathrm{KL}}(\Pi),$$

and summing gives $I(\omega; Y_{1:T} \mid \Pi) \le C\,\mathcal{C}_T$ with $C = \kappa\,C_{\mathrm{up}}/c_{\mathrm{low}}$. The constant is not universal: it degrades as the channel approaches degeneracy ($c_{\mathrm{low}} \downarrow 0$). Motion orthogonal to $\hat{v}_t$ contributes nothing here, not because it is annihilated by an identification operator, an argument available only under an operator-norm reading of $\mathfrak{O}_t$ (Remark 20), but simply because S_t is by construction the separation generated along $\hat{v}_t$.

Lower bound on MI. Let P^0, P^1 denote the laws of $Y_{1:T}$ under ω_0, ω_1 with the policy held fixed. Under the uniform prior, $I(\omega; Y_{1:T} \mid \Pi) = \mathrm{JSD}(P^0, P^1)$. We route through Hellinger rather than KL, and we must do so in a way that survives adaptivity: because Π selects u_t from the realised history, the conditional laws at step t depend on $h_{t-1} := y_{1:t-1}$ and the coordinates are not independent. Lemma 8 supplies the required tensorisation.

Lemma 8 (Hellinger accumulation under an adaptive policy)**.** *Fix a policy Π and let P^0, P^1 be the laws of $Y_{1:T}$ under ω_0, ω_1, with $P^1 \ll P^0$ (Assumption 1). Write $H_t^2 := H^2\big(K_t^{\omega_0,\Pi}(\cdot \mid h_{t-1}), K_t^{\omega_1,\Pi}(\cdot \mid h_{t-1})\big)$ for the conditional squared Hellinger distance at step t, and $A := \sum_{t=1}^T H_t^2$. Then for every $a > 0$,*

$$H^2(P^0, P^1) \;\ge\; 1 - e^{-a} - \sqrt{P^0(A < a)}.$$

In particular, if $A \ge a$ holds P^0-a.s., then $H^2(P^0, P^1) \ge 1 - e^{-a} \ge (1 - e^{-1})\min\{a, 1\}$.

Proof. Let $Z_t := \prod_{s\le t}\sqrt{\mathrm{d}K_s^{\omega_1}/\mathrm{d}K_s^{\omega_0}}$ evaluated at the realised y_s, so that the Hellinger affinity is $\rho(P^0, P^1) = \mathbb{E}_{P^0}[Z_T]$ and $H^2 = 1 - \rho$. Each H_t^2 is $\mathcal{F}_{t-1}$-measurable, since the policy fixes u_t and hence both conditional kernels, before Y_t is drawn. This predictability is what the argument needs. By definition of the conditional affinity,

$$\mathbb{E}_{P^0}[Z_t \mid \mathcal{F}_{t-1}] \;=\; Z_{t-1}\,(1 - H_t^2) \;\le\; Z_{t-1}\,e^{-H_t^2}.$$

Therefore $V_t := Z_t \exp\big(\sum_{s\le t} H_s^2\big)$ satisfies $\mathbb{E}_{P^0}[V_t \mid \mathcal{F}_{t-1}] \le V_{t-1}$: it is a non-negative supermartingale with $V_0 = 1$, so $\mathbb{E}_{P^0}[Z_T e^A] \le 1$.

Split on $\{A \ge a\}$. On that event $Z_T e^a \le Z_T e^A$, so $\mathbb{E}_{P^0}[Z_T \mathbf{1}\{A \ge a\}] \le e^{-a}$. On the complement, Cauchy–Schwarz and $\mathbb{E}_{P^0}[Z_T^2] = \mathbb{E}_{P^0}\big[\prod_t \mathrm{d}K_t^{\omega_1}/\mathrm{d}K_t^{\omega_0}\big] = 1$ give $\mathbb{E}_{P^0}[Z_T \mathbf{1}\{A < a\}] \le \sqrt{P^0(A < a)}$. Adding the two and subtracting from 1 gives the claim. The final inequality uses $1 - e^{-a} \ge (1 - e^{-1})\min\{a, 1\}$. □

With the convention $H^2 = 1 - \rho$, the standard comparison $\mathrm{TV}(P, Q) \le \sqrt{2}\,H(P, Q)$ gives $H_t^2 \ge S_t^2/2$. Thus no pointwise likelihood-ratio bound is used in this direction. The left inequality of (9) gives $S_t^2 \ge \mathfrak{O}_t^2\Delta_t^{\mathrm{KL}}$ pathwise. Hence $A \gtrsim \mathcal{C}_T$ whenever the per-step sandwich holds P^0-a.s., and Lemma 8 yields $H^2(P^0, P^1) \gtrsim \min\{\mathcal{C}_T, 1\}$. Finally $\mathrm{JSD} \gtrsim H^2$, giving

$$I(\omega; Y_{1:T} \mid \Pi) \;\ge\; c\,\min\{\mathcal{C}_T,\, 1\}.$$

Two points deserve emphasis. The quantities $\mathfrak{O}_t$ and Δ_t^{KL} are themselves $\mathcal{F}_{t-1}$-measurable, so $\mathcal{C}_T$ is a random variable and the statement above is the a.s. version; the deviation term $\sqrt{P^0(A < a)}$ in Lemma 8 is what one pays if the per-step sandwich holds only with high probability rather than surely. And no independence is assumed anywhere: the supermartingale step uses only that the policy commits to u_t before observing Y_t, which is what non-anticipativity means.

Three remarks on this direction. First, it is the only place the *left* half of (9) is used, and that half is what the directional definition buys: under an operator-norm $\mathfrak{O}_t$ the inequality $S_t^2 \geq \mathfrak{O}_t^2 \Delta_t^{\mathrm{KL}}$ is false, since a supremum over directions cannot lower-bound the separation produced by the realised one. Second, the truncation is unavoidable and not an artefact of the proof: $\mathrm{JSD} \leq \log 2$ for binary Ω. Recovering a bound linear in $\mathcal{C}_T$ over a long horizon requires the hypothesis space to grow with T; that is exactly the amplified construction $|\Omega| = 2^T$ used in Appendix F, and we do not claim it here. Third, adaptivity is handled explicitly by Lemma 8 rather than assumed away: the policy's dependence on realised observations is precisely what makes H_t^2 predictable, and the supermartingale argument uses no independence between coordinates.

Combining the two bounds gives (6). □

E Proof of the Commons Growth Law (Proposition 37)

Proof of Proposition 37. Write $K' := m/(c\sigma^2)$, so that the effort rule $A_t = m\,\mathrm{MB}_t/c$ and the backward recursion $\mathrm{MB}_t = 1/(\sigma^2\tau_t^2) + \delta\,\mathrm{MB}_{t+1}$ give the exact relation

$$A_t \;=\; \frac{K'}{\tau_t^2} \;+\; \delta A_{t+1} \qquad\Longrightarrow\qquad A_t \;=\; K' \sum_{j\geq 0} \frac{\delta^{\,j}}{\tau_{t+j}^2}. \tag{10}$$

Two-sided control of A_t. Since τ is nondecreasing, $\tau_{t+j} \geq \tau_t$, so (10) gives immediately $A_t \leq K'/\big((1-\delta)\tau_t^2\big)$. For the reverse direction, suppose inductively $\tau_{t+j} \leq \tau_t(1 + j/t)^{1/3}(1 + o(1))$. Then

$$\sum_{j\geq 0} \delta^{\,j}\Big(1+\tfrac{j}{t}\Big)^{-2/3} \;=\; \frac{1}{1-\delta}\Big(1 - O\big(\tfrac{1}{t(1-\delta)}\big)\Big),$$

because the geometric weight concentrates on $j = O(1/(1-\delta))$. Hence

$$A_t \;=\; \frac{K'}{(1-\delta)\,\tau_t^2}\Big(1 + O\big(\tfrac{1}{t(1-\delta)}\big)\Big). \tag{11}$$

Integration. By definition $\tau_t - \tau_{t-1} = A_t/\sigma^2$, so (11) yields $\tau_t - \tau_{t-1} = (K/\tau_t^2)(1 + O(\cdot))$ with $K = K'/(\sigma^2(1-\delta)) = m/(c\sigma^4(1-\delta))$. Multiplying by τ_t^2 and summing (equivalently, integrating $\tau^2\,d\tau = K\,dt$, whose discretisation error is $O(\tau^{-3})$ per step and hence summable),

$$\tfrac{1}{3}\tau_t^3 \;=\; Kt\Big(1 + O\big(\tfrac{1}{t(1-\delta)}\big)\Big) \;+\; \tfrac{1}{3}\tau_0^3,$$

and since $\tau_0^3 = O(1)$ while $Kt \to \infty$, this gives $\tau_t = (3Kt)^{1/3}(1 + O(1/(t(1-\delta))))$, which is (17); taking logarithms gives the second form.

(i) Universality of the exponent. Positing $\tau_t = Ct^a$ in $\tau_t - \tau_{t-1} = K/\tau_t^2$ forces $a - 1 = -2a$, i.e. $a = 1/3$, by matching powers of t alone. The parameters $N, \delta, \sigma^2, c, m$ appear only in K, hence only in the prefactor $C = (3K)^{1/3}$.

(ii) The wedge. The multiplier m enters (17) only through $K \propto m$. With $m = N$ under equilibrium and $m = N^2$ under the social optimum,

$$\log \tau_t^{\mathrm{social}} - \log \tau_t^{\mathrm{eq}} = \tfrac{1}{3}\log\frac{N^2}{N} + O\big(\tfrac{1}{t(1-\delta)}\big) = \tfrac{1}{3}\log N + O\big(\tfrac{1}{t(1-\delta)}\big).$$

Because the correction term in (17) depends on the regime only through K, and enters $\log \tau_t$ additively at the same order in both, it cancels to leading order in the difference, which is why the wedge is numerically far more accurate at finite t than either level (§S.1). □

Remark 9 (Scope). The exponent $1/3$ is a consequence of the pairing (flow payoff $-1/\tau$, quadratic cost). A flow payoff $-\tau^{-q}$ would give $a = 1/(q+2)$ by the same power matching; the structure of the argument, and the fact that the governance multiplier enters only the prefactor, are unchanged. The wedge $\frac{1}{3}\log N$ is specific to the equilibrium/planner multipliers being N and N^2.

F Full Proof of Canonicality (Theorem 17)

Proof of Theorem 17. Let $F_T(\Pi) = \sum_{t=1}^{T} f_t(\{K_t^{\omega,\Pi}\}_\omega, \{K_{t+1}^{\omega,\Pi}\}_\omega)$ be an admissible budget. We must exhibit a family on which $F_T(\Pi) \geq c\,\mathrm{SIB}_T^{\mathrm{TV}}(\Pi)$. By temporal additivity (A2) it suffices to bound each one-step contribution on that family; by environment blindness (A5), f_t is a function of the temporal increment alone, so the construction below pins down the same f_t that a general system with the same increment would present. What (A5) does *not* give is invariance across increments with equal scalar drift, which is why the conclusion is stated on $\mathcal{E}$ rather than pointwise (Remark 23).

Step 1: Constructing the extremal system. Fix a target per-step drift $\delta \in (0,1)$ and a horizon T. Take $\Omega = \{0,1\}^T$ with the uniform prior, so the coordinates $\omega_1, \ldots, \omega_T$ are i.i.d. fair bits. Let step s be informative about coordinate ω_s and about nothing else: the kernel at step s is $K_s^{\omega,\Pi} = Q_{\omega_s}$ with $\mathrm{TV}(Q_0, Q_1) = \delta$, and the baseline kernel there is $\bar{K}_s = \frac{1}{2}(Q_0 + Q_1)$, so that $\Delta_s^{\mathrm{TV}} = \delta/2$ at every step. The construction is the same at every s; in particular the kernel increment presented to f_s is identical across s, which is what Step 3 uses.

Step 2: Le Cam lower bound, on all T coordinates at once. Because the coordinates are independent and Y_s depends on ω only through ω_s, the pairs (ω_s, Y_s) are independent across s and

$$I(\omega; Y_{1:T}) \;=\; \sum_{s=1}^{T} I(\omega_s; Y_s) \;\geq\; T\,\frac{\delta^2}{2\ln 2},$$

each summand by the two-point bound for a uniform binary pair at total variation δ. This is the step that defeats the additive constant in (A4), and it is why the construction is amplified: with a single informative step the left-hand side is capped at $\log 2$ and (A4) absorbs it, yielding no bound on f_t at all.

Step 3: Applying information domination. By temporal additivity (A2), $F_T = \sum_s f_s$, and by strategic nullity (A3) nothing outside the constructed steps contributes. By (A4),

$$T\,\frac{\delta^2}{2\ln 2} \;\leq\; C\sum_{s=1}^{T} f_s + C.$$

By construction every step presents the same increment, and by environment blindness (A5) f_s is a function of that increment alone, so $f_s = f$ for all s and the sum is Tf. Dividing by T,

$$f \;\geq\; \frac{\delta^2}{2C\ln 2} \;-\; \frac{1}{T}.$$

Choosing $T \geq T_0(\delta, C) := 4C\ln 2/\delta^2$ makes the subtracted term at most half the first, so $f \geq \delta^2/(4C\ln 2)$. Since $\Delta_s^{\mathrm{TV}} = \delta/2$, this is $f \geq c_0\,(\Delta_s^{\mathrm{TV}})^2$ with $c_0 = 1/(C\ln 2)$, a constant depending only on the domination constant of (A4) and not on δ or on the member of the family. The family $\mathcal{E}$ is therefore indexed by (δ, T) with $T \geq T_0(\delta, C)$, and the conclusion is uniform over it; it says nothing about horizons below T_0.

Step 4: Identifying the budget. Observe that the quadratic quantity already *is* the KL budget: Pinsker gives $\Delta_s^{\mathrm{KL}} \geq 2(\Delta_s^{\mathrm{TV}})^2$ and reverse Pinsker (Assumption 6) gives $\Delta_s^{\mathrm{KL}} \leq \kappa(\Delta_s^{\mathrm{TV}})^2$, so $(\Delta_s^{\mathrm{TV}})^2 \asymp \Delta_s^{\mathrm{KL}}$. Summing over s on the constructed family gives $F_T \geq c\,\mathrm{SIB}_T^{\mathrm{KL}}$ there, with $c = c_0/\kappa$, and hence the non-negligibility conclusion. Reverse Pinsker (Assumption 6) is used in this last step, and is therefore listed among the theorem's hypotheses. □

G Full Proof of the Frontier Sandwich (Theorem 30)

Proof of Theorem 30. Throughout, the budget constraint is on the expectation, $\mathbb{E}[\mathrm{SIB}_T] \leq B$. Mutual information is a number, not a random variable, so each side is matched to the qualifier the corresponding half of (6) supplies: an expectation on the upper, an essential infimum on the lower.

Upper bound. By Theorem 1 and the pathwise inequality $\mathrm{SIB}_T = \mathcal{C}_T + \mathcal{W}_T \geq \mathcal{C}_T$, every feasible policy has $I(\omega; Y_{1:T} \mid \Pi) \leq C\,\mathbb{E}[\mathcal{C}_T(\Pi)] \leq C\,\mathbb{E}[\mathrm{SIB}_T(\Pi)] \leq CB$. Taking the supremum over feasible Π gives $\mathcal{F}_T(B) \leq CB$, and the same chain stopped one step earlier gives $\mathcal{F}_T(B) \leq C\,\mathcal{C}^*_{\mathrm{mean}}(B)$.

Lower bound. Fix $\varepsilon > 0$ and choose Π_ε with $\mathbb{E}[\mathrm{SIB}_T(\Pi_\varepsilon)] \le B$ and $\operatorname{ess\,inf} \mathcal{C}_T(\Pi_\varepsilon) \ge \mathcal{C}^*_{\mathrm{floor}}(B) - \varepsilon$, which is possible by definition of $\mathcal{C}^*_{\mathrm{floor}}$ as a supremum over the same feasible set. The lower bound of (6) is stated for the essential infimum and holds for *every* policy, so applying it to Π_ε,

$$\mathcal{F}_T(B) \;\ge\; I(\omega; Y_{1:T} \mid \Pi_\varepsilon) \;\ge\; c\, \min\{\mathcal{C}^*_{\mathrm{floor}}(B) - \varepsilon,\; 1\}.$$

Letting $\varepsilon \downarrow 0$ gives the claim. The floor frontier is what appears here and not $\mathcal{C}^*_{\mathrm{mean}}$: a policy with large expected capacity carried by rare histories has no almost-sure floor and the lower bound says nothing about it. □

Remark 10 (Why there is no per-step construction). One might instead try to prove the lower bound by building a policy Π^* that, at each step, perturbs along a direction attaining the observability modulus, and then summing the per-step separations. That route has two defects. It constructs a perturbation whose drift is *at most* the allocated δ_t while accounting for it as though it were exactly δ_t; and, more seriously, it chooses the perturbation direction separately at each t, whereas $\mathfrak{O}_s$ for $s > t$ is defined relative to the continuation path and so changes when the policy is perturbed at time t. The per-step maximisers need not be simultaneously realisable by any single policy, so the sum over steps is not the strategic capacity of an actual policy.

The route we take avoids this issue rather than repairing it, and buys less than a construction would. Because the lower bound of Theorem 1 holds for an arbitrary given policy, we obtain a lower bound by applying it to a policy approaching the *floor*-capacity frontier $\mathcal{C}^*_{\mathrm{floor}}(B)$; the constant is then inherited from Theorem 1 rather than being a separate $1/C_{\mathrm{up}}$. What this does *not* do is exhibit a policy attaining the information frontier, or show that capacity-maximising policies attain it. The upper bound involves $\mathcal{C}^*_{\mathrm{mean}}(B)$, and a floor-capacity maximiser need not have large mean capacity while a mean-capacity maximiser need not have any floor at all, so neither maximiser is shown to be information-optimal (Remark 31).

Theorem 11 (Stabilization removes continuation-mediated separation)**.** *Under Assumptions 7–8, if* Π *is asymptotically strategically stable, then*

$$\limsup_{t\to\infty} \sup_{\omega,\omega'} \mathrm{TV}(K_t^{\omega,\Pi}, K_t^{\omega',\Pi}) \le \sup_{\omega,\omega'} \Gamma_K(\Gamma_{\mathrm{dir}}(d_\Omega(\omega,\omega'))) \quad \text{a.s.}$$

Thus stabilization removes the continuation-mediated component of separation. Total collapse requires the no-static-signature condition $\Gamma_{\mathrm{dir}} \equiv 0$*; otherwise an irreducible direct residual may remain.*

H Full Proof of the Stabilization Theorem

Proof of Theorem 11. Fix t, H_{t-1}, and ω, ω'. Since $K_t^{\omega,\Pi} = \mathsf{K}(\cdot \mid x_t^\omega, u_t)$, Assumption 8 gives $\mathrm{TV}(K_t^{\omega,\Pi}, K_t^{\omega',\Pi}) \le \Gamma_K(d_X(x_t^\omega, x_t^{\omega'}))$.

Apply the triangle inequality:

$$\begin{aligned} d_X(x_t^\omega, x_t^{\omega'}) \le\; & d_X(\mathrm{BR}_t(u_{t:\infty}, \omega), \mathrm{BR}_t(u_{t+1:\infty}, \omega')) \\ & + d_X(\mathrm{BR}_t(u_{t+1:\infty}, \omega'), \mathrm{BR}_t(u_{t:\infty}, \omega')). \end{aligned}$$

By Assumption 7, the first term is bounded by $\Gamma_{\mathrm{dir}}(d_\Omega(\omega,\omega')) + \Gamma_{\mathrm{cont}}(d_{\mathrm{tail}}(u_{t:\infty}, u_{t+1:\infty}))$ and the second by $\Gamma_{\mathrm{cont}}(d_{\mathrm{tail}}(u_{t:\infty}, u_{t+1:\infty}))$. Taking sup over ω, ω' and sending $t \to \infty$: $d_{\mathrm{tail}} \to 0$ by stabilization, so $\Gamma_{\mathrm{cont}}(\cdot) \to 0$. □

Proposition 12 (Stabilization kills the conversion rate)**.** *Under Assumptions 7–8* and *the reverse-Pinsker regularity of Assumption 6, if* Π *is asymptotically strategically stable, then: (i)* $\Delta_t^{\mathrm{KL}} \to 0$*, (ii) the predictive diameter decays to the irreducible direct residual, and (iii)* $\mathfrak{O}_t^2 \Delta_t \to 0$*.*

Remark 13 (Why reverse Pinsker is needed here). Assumptions 7–8 control total variation, and TV convergence does not by itself imply KL convergence: the likelihood ratio can blow up on a vanishing set. Claim (i) is a statement about Δ_t^{KL}, so it requires a route back from TV to KL, which is exactly Assumption 6 ($\mathrm{KL} \le \kappa\, \mathrm{TV}^2$, available whenever likelihood ratios are uniformly bounded). The continuity assumptions alone are not sufficient for claim (i).

Figure 4: Burn-in front-loads the information budget

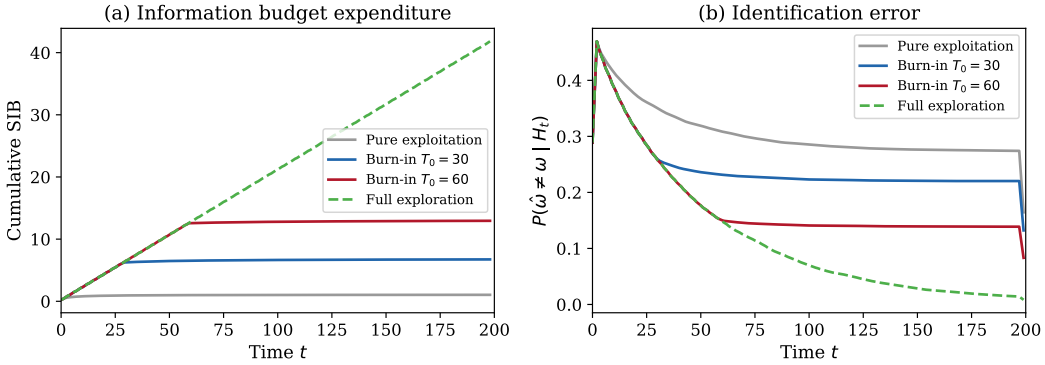


Figure 5: **Burn-in front-loads the information budget** (illustrates Corollary 4). (a) Cumulative SIB. (b) Identification error flattens when SIB stops growing. Error tracking SIB here rather than capacity is specific to this example, in which $\mathfrak{O}_t \approx 1$ throughout, so budget and capacity nearly coincide; the theory predicts tracking of capacity, not of SIB.

I Canonical Examples in Full

Example I: Tragedy of the learning commons. Consider $N \geq 2$ agents choosing exploration effort $a_{i,t}$ at private cost $c(a_{i,t})$. A public signal $Y_t \sim \mathcal{N}(\omega, \sigma^2/A_t)$ depends on aggregate effort $A_t = \sum_i a_{i,t}$. Each agent equates *private* marginal benefit to cost, while the social planner equates N times the benefit. By strict convexity, equilibrium effort is strictly below the optimum. Moreover, posterior variance decay implies that marginal incentives vanish over time, so equilibrium stabilizes (Lemma 15, Appendix K). By Theorem 11, stabilization collapses the observation channel: the information budget is bounded, no estimator can consistently identify ω, and the root cause is that SIB is a public good whose production cost is private.

Example II: Incentivized exploration as prepaid SIB. A centralized planner recommends actions subject to Bayesian incentive compatibility (BIC). The conservation law applies identically: burn-in works because it concentrates distributional motion into an initial phase where $\mathfrak{O}_t > 0$, spending the canonical budget up front. After burn-in, the policy stabilizes and no further information is generated. The planner does not escape the law; the planner *finances* it. In a three-arm bandit ($\omega \in \{0, 1\}$, one dominant-but-uninformative arm, two informative arms), burn-in concentrates budget where the observability modulus is positive (Figure 5). This is an illustration, not an attainment claim: Theorem 30 is a sandwich between two capacity frontiers and exhibits no policy reaching the upper one (Remark 31).

J SIB-Experimentation Game

We present an abstract game-theoretic formulation of the tragedy: N players choose noise levels η_i, aggregate noise determines signal quality through concave $G(\cdot)$, and individual utility is $U_i = V_i G(\sum_j \eta_j) - \alpha\eta_i^2$. The symmetric Nash equilibrium satisfies $\eta_{\text{Nash}} < \eta_{\text{Soc}}$, and $\eta_{\text{Nash}} \to 0$ as $N \to \infty$ even when η_{Soc} remains bounded away from zero.

K Tragedy of the Commons: Full Proofs

Lemma 14 (Under-exploration)**.** *Equilibrium aggregate effort satisfies $A_t^{\text{eq}} < A_t^{\text{opt}}$ at every history.*

Proof. Each agent equates private marginal benefit to $c'(a_{i,t})$; the planner equates N times that benefit to the same cost. Strict convexity forces equilibrium below the optimum. □

Lemma 15 (Equilibrium stabilization). *Along any equilibrium path,* $A_t^{\mathrm{eq}} \to \bar{A}$ *and* $|A_{t+1}^{\mathrm{eq}} - A_t^{\mathrm{eq}}| \to 0$.

Proof. Posterior variance satisfies $\mathrm{Var}(\omega|H_t) = (V_0^{-1} + \sum_{s=1}^{t} A_s/\sigma^2)^{-1}$, so marginal value of effort decays. Under discounting and strict convexity, the Bellman equation admits a stationary response; contraction implies convergence. □

L Recommender Model Details

The recommender has $K = 4$ types. Types 0,1 are informative about ω ($\mu_\omega[0] \in \{0.9, 0.1\}$); types 2,3 are uninformative ($\mu_\omega[k] = 0.5$). Creators adapt at rate $\lambda = 0.15$: $q_{t+1} = (1-\lambda)q_t + \lambda e_{r_t}$. Click probability is $\mathrm{cp}_t = \alpha\mu_\omega[r_t]D(q_t) + (1-\alpha)q_{t,r_t}$ where $D(q_t) = 1 - \max_k q_{t,k}$ and $\alpha = 0.35$.

Figure 6 demonstrates that action diversity $\neq$ information: "active but blind" (199 switches, 1.0 bit entropy) generates SIB $= 0.40$ and error $= 0.50$; "rare but targeted" (19 switches, 0.81 bits) achieves SIB $= 14.25$ and error $= 0.32$.

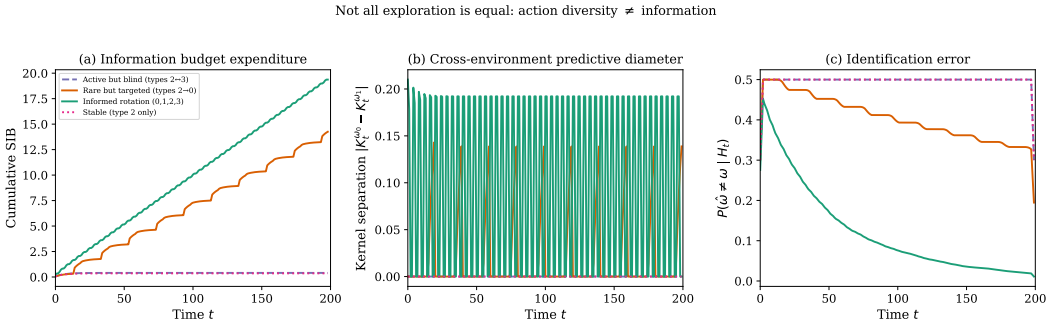


Figure 6: **Action diversity $\neq$ information** (is consistent with Theorem 1 and Corollary 4; Theorem 17 is an extremal statement about budgets and is not what this figure bears on). (a) Cumulative SIB. (b) Kernel separation. (c) Identification error.

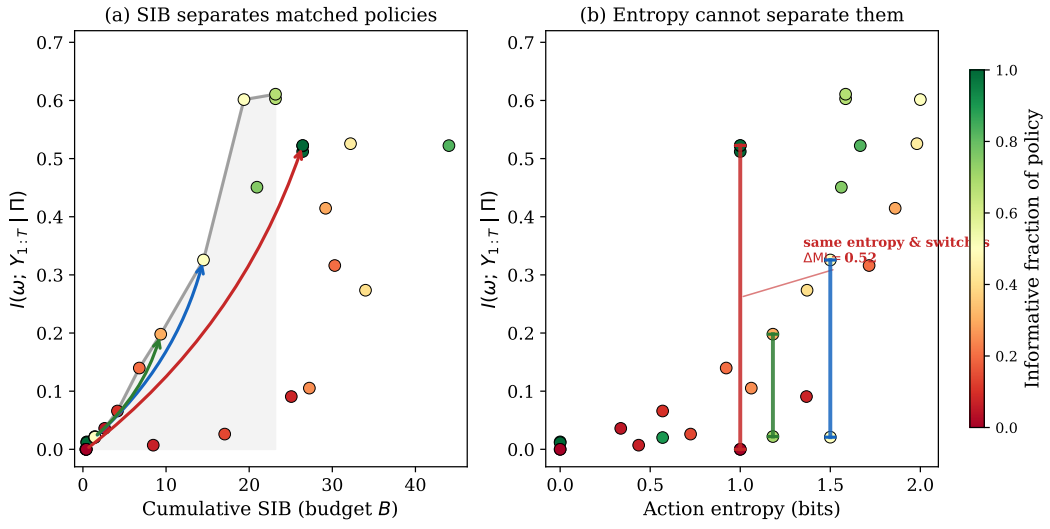


Figure 7: **SIB separates policies that entropy cannot.** (a) Under SIB, matched policies are separated: "Rot-01" (SIB $= 26.4$) vs. "Rot-23" (SIB $= 0.4$). (b) Under action entropy, these pairs collapse onto the same coordinate.

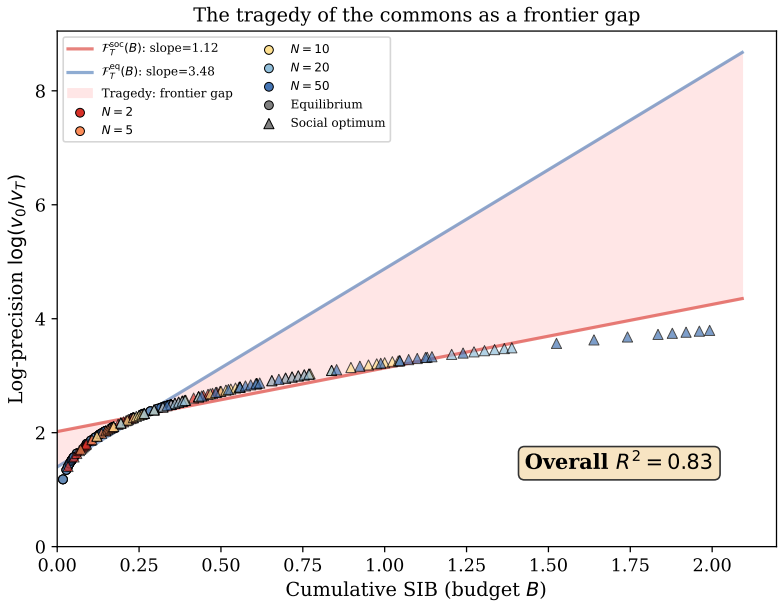


Figure 8: **The tragedy as a frontier gap.** Equilibrium (circles) and social-optimum (triangles) fall on linear frontiers with slopes 3.48 and 1.12 ($R^2 = 0.93$ and 0.89).

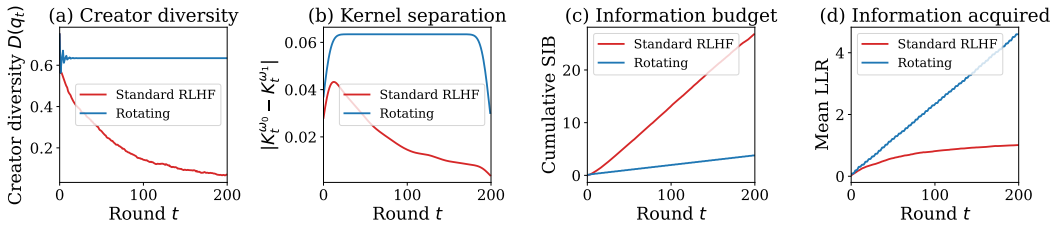


Figure 9: **The collapse mechanism in RLHF.** Standard RLHF (red) vs. Rotating (blue). (a) Diversity. (b) Kernel separation. (c) SIB. (d) Information acquired.

M The Exploration Mirage in Full

M.1 Environment Specification

This appendix provides the full specification for the exploration mirage experiment (§T.1), which upgrades the two-environment design to $|\Omega| = 4$ with hierarchical distinguishability.

M.2 World Parameter Space

$|\Omega| = 4$ environments with *hierarchical distinguishability structure*:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}, \qquad \Pr(\omega) = \tfrac{1}{4} \ \ \forall\, \omega.$$

The four environments encode a **two-level identification tree**:

- **Coarse split** (easy): $\{\omega_1, \omega_2\}$ vs. $\{\omega_3, \omega_4\}$: distinguishable via style-pair (s_A, s_B).
- **Fine split** (hard): ω_1 vs. ω_2; requires style-pair (s_A, s_C).
- **Fine split** (hard): ω_3 vs. ω_4; requires style-pair (s_B, s_D).

A policy varying only across (s_A, s_B) resolves the coarse partition but stalls on the fine partition—**partial identification**.

M.3 Content Styles (Actions)

$K = 4$ content styles: $\mathcal{A} = \{s_A, s_B, s_C, s_D\}$. At each round t the policy selects $\pi_t \in \Delta(\mathcal{A})$; a style $a_t \sim \pi_t$ is presented.

M.4 Evaluator Response Model

The evaluator has a latent preference vector $\theta^\omega \in \mathbb{R}^K$ (Table 4).

Environment	Preference vector θ^ω	Intuition
ω_1	$(+1, -1, +1, -1)$	Likes A, C; dislikes B, D
ω_2	$(+1, -1, -1, +1)$	Likes A, D; dislikes B, C
ω_3	$(-1, +1, +1, -1)$	Likes B, C; dislikes A, D
ω_4	$(-1, +1, -1, +1)$	Likes B, D; dislikes A, C

Table 4: Preference vectors for the four-environment mirage experiment.

Observation model. When style a_t is presented:

$$Y_t = \theta^\omega_{a_t} \cdot f_t(a_t) + \sigma\, \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, 1), \tag{12}$$

where $f_t(s) = (1 - \lambda)^{n_{t-1}(s)}$ is the evaluator's **freshness** for style s and $n_t(s) = \sum_{\tau=1}^{t} \mathbf{1}[a_\tau = s]$ is the exposure count.

Predictive kernel. $K_t^{\omega,\Pi}(\cdot \mid H_{t-1}) = \mathcal{N}(\theta^\omega_{a_t} \cdot f_t(a_t),\ \sigma^2)$.

Strategic response map. $x_t^\omega = \mathrm{BR}_t(u_{t:\infty}, \omega) = \theta^\omega \odot f_t$, where $f_t = (f_t(s_A), \ldots, f_t(s_D))$. Adaptation attenuates signal on overexposed styles.

M.5 Distinguishability Structure

Style pair	What it distinguishes	Why
(s_A, s_B)	$\{\omega_1, \omega_2\}$ vs. $\{\omega_3, \omega_4\}$	$\theta^\omega_A - \theta^\omega_B = +2$ for $\omega_{1,2}$; $= -2$ for $\omega_{3,4}$
(s_A, s_C)	ω_1 vs. ω_2 (given coarse)	$\theta^\omega_C = +1$ for ω_1; $= -1$ for ω_2
(s_B, s_D)	ω_3 vs. ω_4 (given coarse)	$\theta^\omega_D = -1$ for ω_3; $= +1$ for ω_4
(s_C, s_D)	Nothing new	Redundant for identification
(s_A, s_D)	Partial: conflates coarse and fine	$\theta^\omega_A - \theta^\omega_D$: $+2, 0, 0, -2$

Table 5: Pairwise distinguishability structure in the mirage environment.

A policy alternating only $s_A \leftrightarrow s_B$ generates maximal pairwise variation but can only resolve the coarse split, it stalls at 1 bit of identification (out of 2 needed). This is the *exploration mirage*.

M.6 Default Parameters

M.7 Policy Definitions

Six policies, designed a priori. Policies 1–4 are *mirage policies* (high standard diagnostics, low information); Policies 5–6 are *effective policies*.

M.8 Policy 1: Pure Exploit (Π_{exploit})

At each t, play the style with highest posterior expected reward: $a_t = \arg\max_{s \in \mathcal{A}} \mathbb{E}[\theta^\omega_s \mid H_{t-1}]$. Converges to a single style; collapse regime ($\mathfrak{D}_t \to 0$, $\Delta_t \to 0$).

Parameter	Symbol	Default	Rationale
Environments	$\lvert\Omega\rvert$	4	2 bits; hierarchical structure
Styles	K	4	Minimum for hierarchical distinguishability
Horizon	T	200	Matches §T.2–T.1
Adaptation rate	λ	0.25	Matches §T.2
Observation noise	σ	1.0	Unit noise
MC replications	—	2000	Matches §T.1
Prior	—	Uniform(Ω)	Standard

Table 6: Default parameters for the mirage experiment.

M.9 Policy 2: High-ε Greedy (Π_ε)

$\pi_t(s) = (1-\varepsilon)\cdot \mathbf{1}[s = s_t^*] + \varepsilon/K$ with $\varepsilon = 0.5$. Undirected exploration; high SIB, moderate $\mathcal{C}_T$.

M.10 Policy 3: Blind Alternation (Π_{blind})

Deterministically alternate $s_A, s_B, s_A, s_B, \ldots$ Maximal pairwise variation on the A–B axis, but the A–B comparison resolves only the coarse split. **Primary mirage policy.**

Monte Carlo estimates ($\lambda = 0.25$, $\sigma = 1$, $T = 200$, 2,000 runs): $\text{SIB}_T = 6.07$, $\widetilde{\mathcal{C}}_T^{\min} = \mathbf{0.000}$ (both fine splits), info $= 0.934$ bits.

M.11 Policy 4: Entropy-Matched Random (Π_{random})

$\pi_t(s) = 1/K$ for all s. Maximum entropy, undirected variation.

M.12 Policy 5: Targeted Rotation (Π_{targeted})

Deterministic 4-cycle: $a_t = \text{cycle}(s_A, s_B, s_C, s_D)$. Resolves both coarse and fine splits; balanced freshness decay. **Matched partner to Π_{blind}.**

Monte Carlo estimates ($\lambda = 0.25$, $\sigma = 1$, $T = 200$, 2,000 runs; `regen_mirage_pairwise.py`, `recompute_sib_baseline.py`): $\text{SIB}_T = 13.07$, $\widetilde{\mathcal{C}}_T^{\min} = \mathbf{2.392}$ (all six pairs positive), info $= 1.877$ bits.

M.13 Policy 6: Posterior-Guided ($\Pi_{\text{posterior}}$)

Adaptive: at each t, select $a_{t+1} = \arg\max_s \text{EIG}_t(s)$ where $\text{EIG}_t(s) = H[p_t] - \mathbb{E}_Y[H[p_t^{(s,Y)}]]$. Oracle upper bound on information efficiency.

M.14 The Budget-Matched Family

Write $\Theta \in \{\pm 1\}^{4\times 4}$ for the hypothesis-by-style matrix of §M.1, whose columns are $s_A = (+,+,-,-)$, $s_B = (-,-,+,+)$, $s_C = (+,-,+,-)$, $s_D = (-,+,-,+)$, and $\text{diff}_k(s) := |\theta_s^{\omega_i} - \theta_s^{\omega_j}| \in \{0,2\}$ for the k-th hypothesis pair (ω_i, ω_j). For distinct styles $s \neq s'$ let $\Pi_{ss'}$ be the policy that plays s at even t and s' at odd t.

Proposition 16 (Budget and average energy are invariant across the family)**.** *For every pair of distinct styles $s \neq s'$:*

- *(i) the freshness process, exposure counts, switching frequency and action entropy of $\Pi_{ss'}$ do not depend on which two styles are chosen;*
- *(ii) $\text{SIB}_T(\Pi_{ss'})$ takes the same value for all six pairs;*
- *(iii) the average pairwise separation energy $\widetilde{\mathcal{C}}_T(\Pi_{ss'}) = \frac{1}{6}\sum_k \widetilde{\mathcal{C}}_T(\Pi_{ss'}; k)$ takes the same value for all six pairs;*
- *(iv) $\widetilde{\mathcal{C}}_T^{\min}(\Pi_{ss'}) = 0$ if and only if $\{s, s'\}$ fails to separate some hypothesis pair, which happens exactly for $\{s_A, s_B\}$ and $\{s_C, s_D\}$.*

Proof. (i) Strict alternation gives each of the two styles $\lceil T/2 \rceil$ or $\lfloor T/2 \rfloor$ exposures and never touches the other two, so e_t and $f_t = (1-\lambda)^{e_t}$ are the same sequences up to relabelling the styles; switching frequency is 1 and action entropy $\log 2$ regardless.

(ii) By (13) the step cost is $\sup_\omega (\theta^\omega_{s'} f_t(s') - \theta^\omega_s f_t(s))^2/2\sigma^2$. Every entry of Θ has modulus 1, and inspection of the four columns shows that for each of the six style pairs there is a row ω on which the two entries have opposite sign. The supremum is therefore attained there and equals $(f_t(s') + f_t(s))^2/2\sigma^2$, which by (i) is the same sequence for every pair. Summing over t gives the claim.

(iii) By (14) the per-step separation depends on the pair only through $\mathrm{diff}_k(a_t) \in \{0, 2\}$, so with $h(z) := \big(2\Phi(z/2) - 1\big)^2$ and $w_t := h\big(2f_t(a_t)/\sigma\big)$ the energy of hypothesis pair k is $\sum_t \mathbf{1}[\mathrm{diff}_k(a_t) = 2]\, w_t$; the argument that follows uses only that w_t does not depend on k, so it is unchanged by replacing the quadratic proxy with the exact h. Each column of Θ separates exactly four of the six hypothesis pairs, s_A and s_B separate $(0,2), (0,3), (1,2), (1,3)$, and s_C and s_D separate $(0,1), (0,3), (1,2), (2,3)$, so for any style a, $\#\{k : \mathrm{diff}_k(a) = 2\} = 4$. Exchanging the order of summation,

$$\sum_{k=1}^{6} \sum_t \mathbf{1}[\mathrm{diff}_k(a_t) = 2]\, w_t \;=\; \sum_t w_t \,\#\{k : \mathrm{diff}_k(a_t) = 2\} \;=\; 4 \sum_t w_t,$$

independently of which styles appear, and of the form of h. Dividing by six and applying (i) gives the same average for every style pair.

(iv) $\widetilde{\mathcal{C}}^{\min}_T = 0$ iff some k has $\mathrm{diff}_k(s) = \mathrm{diff}_k(s') = 0$. From the rows of diff, the pairs $(0,1)$ and $(2,3)$ vanish exactly on $\{s_A, s_B\}$ and the pairs $(0,2)$ and $(1,3)$ exactly on $\{s_C, s_D\}$; every other style pair meets all six. □

The interest of the construction is that (ii) and (iii) hold *exactly*, so the comparison in Table 1 is not a matched pair produced by tuning. Budget, average energy, switching frequency, exposure and action entropy are all pinned by the design, and the only free quantity is the minimum over hypothesis pairs. Numerically, under the exact total variation (14), the observed spread of SIB_T across the six is 0 and that of $\widetilde{\mathcal{C}}_T$ is also 0 (`matched_pair_exacttv.py`), consistent with the proposition.

The heterogeneous comparison. For reference, Table 7 reports the six structurally different policies covering a wide range of budgets. It makes the weaker point that budget fails to *order* policies; the matched family above makes the stronger one that budget cannot distinguish them at all.

Policy	SIB_T	$\widetilde{\mathcal{C}}^{\mathrm{avg}}_T$	$\widetilde{\mathcal{C}}^{\min}_T$	Info (bits)	Error
Pure Exploit	1.0	1.015	**0.000**	0.86	0.516
ε-Greedy (0.5)	8.6	3.190	2.392	1.87	0.038
Blind Alternation	6.1	1.595	**0.000**	0.93	0.507
Uniform Random	9.9	3.190	2.392	1.88	0.029
Targeted Rotation	13.1	3.190	2.392	1.88	0.029
Posterior-Guided	11.9	3.190	2.392	1.88	0.026

Table 7: Structurally heterogeneous policies, same environment and settings. Budgets differ by more than a factor of ten, so this table cannot separate the effect of spending from the effect of where it is spent; that is what Table 1 is for. Separation energy tracks information ($R^2 = 0.98$) better than the budget does (0.77), but with six points neither figure carries much weight.

M.15 Matched-Pair Analysis

The centerpiece comparison: Π_{blind} vs. Π_{targeted}.

Blind alternation has *lower* SIB than targeted (6.07 vs. 13.07), so a naive practitioner might predict it is more efficient. The mirage is twofold: (i) blind's SIB is entirely channeled into the coarse split where $\mathfrak{O}_t > 0$; (ii) the fine-split strategic capacity is **identically zero** because s_A and s_B carry zero differential signal for ω_1 vs. ω_2 or ω_3 vs. ω_4.

Metric	Π_{blind}	Π_{targeted}	Gap
SIB_T (baseline to endpoint)	6.07	13.07	2.2× (targeted *higher*)
Fine-split $\widetilde{\mathcal{C}}_T(\omega_0, \omega_1)$	**0.000**	2.392	∞
Fine-split $\widetilde{\mathcal{C}}_T(\omega_2, \omega_3)$	**0.000**	2.392	∞
Coarse pairs (each of the other four)	2.392	2.392–4.784	—
Min pairwise $\widetilde{\mathcal{C}}_T$	**0.000**	2.392	∞
Avg pairwise $\widetilde{\mathcal{C}}_T$	1.595	3.190	2.0×
Expected info gain	0.934 bits	1.877 bits	2.0×
Identification error	0.507	0.030	—

Table 8: Matched-pair comparison. Monte Carlo estimates under the corrected estimator at the settings of Table 1 ($T = 200$, $\lambda = 0.25$, $\sigma = 1$, 2,000 runs), regenerated by `regen_mirage_pairwise.py` from the same code path, so the aggregate columns agree with Table 1 by construction.

Why the gap exists. Π_{blind} alternates $s_A \leftrightarrow s_B$, generating kernel separation only on the (A, B) dimension. Once the coarse split resolves, $\mathfrak{O}_t$ on the fine axis drops to 0; the policy burns SIB but $\mathfrak{O}_t^2 \Delta_t \approx 0$, so $\mathcal{C}_T$ stalls. Π_{targeted} cycles all four styles, maintaining nonzero $\mathfrak{O}_t$ on both axes.

M.16 Metric Definitions

All metrics instantiate §2; where they depart we say so. Throughout, a tilde marks an *unnormalised* empirical quantity, to keep it distinct from the defined object it estimates.

SIB (baseline to endpoint). $\text{SIB}_T^{\text{KL}}(\Pi) = \sum_{t=1}^{T} \Delta_t^{\text{KL}}(\Pi)$ with Δ_t^{KL} the cost of the *displacement* at step t, as in (1). The baseline continuation here is "keep playing the style played last", so baseline and endpoint are evaluated at the same freshness state f_t and the cost isolates the switch rather than the passage of time:

$$\Delta_t^{\text{KL}} = \sup_{\omega} \frac{\big(\theta^{\omega}_{a_{t-1}} f_t(a_{t-1}) - \theta^{\omega}_{a_t} f_t(a_t)\big)^2}{2\sigma^2}, \qquad t = 1, \dots, T, \quad a_0 := a_1, \tag{13}$$

giving T costs against T observations. An earlier version of this appendix instead summed over *consecutive kernels*, $(\theta^{\omega}_{a_t} f_t(a_t) - \theta^{\omega}_{a_{t+1}} f_{t+1}(a_{t+1}))^2/2\sigma^2$, which is the indexing §2 rejects: it yields $T - 1$ costs and charges the policy for freshness decay it did not choose.

Realised separation (exact Gaussian total variation). The observation at step t is $\mathcal{N}(\mu_t^{\omega}, \sigma^2)$ with $\mu_t^{\omega} = \theta^{\omega}_{a_t} f_t(a_t)$, so for the pair (ω_i, ω_j) the total variation between the two endpoint kernels is

$$S_t(\omega_i, \omega_j) \;=\; 2\,\Phi\Big(\frac{|\mu_t^{\omega_i} - \mu_t^{\omega_j}|}{2\sigma}\Big) - 1 \;=\; 2\,\Phi\Big(\frac{|\theta^{\omega_i}_{a_t} - \theta^{\omega_j}_{a_t}|\, f_t(a_t)}{2\sigma}\Big) - 1, \tag{14}$$

evaluated at the style *actually played* at t. Note that the argument of Φ is not a usable substitute for the total variation here: the correct first-order coefficient is $1/(\sigma\sqrt{2\pi})$, not $1/(2\sigma)$, and the gaps in this model are not small, at the largest gap ($|\Delta\mu| = 2$, $\sigma = 1$) the linearisation gives 1.000 against the exact 0.6827. All separation energies use (14). The step cost is unaffected, since $\Delta_t^{\text{KL}} = (\Delta\mu)^2/2\sigma^2$ is exact for Gaussians. There is no maximisation over styles: a style that is never played carries no information however fresh it is, and taking a maximum would be the operator norm that Remark 20 rules out.

Separation energy. The estimator reports $\widetilde{\mathcal{C}}_T(\Pi; \omega_i, \omega_j) = \sum_{t=1}^{T} S_t(\omega_i, \omega_j)^2$, which is $\mathcal{C}_T$ of (5) only up to the channel constant C_{up} of Assumption 5, not identified here. We therefore write $\widetilde{\mathcal{C}}_T$ and never call the column $\mathcal{C}_T$. For $|\Omega| = 4$:

$$\widetilde{\mathcal{C}}_T(\Pi) = \frac{1}{\binom{4}{2}} \sum_{i<j} \widetilde{\mathcal{C}}_T(\Pi; \omega_i, \omega_j), \qquad \widetilde{\mathcal{C}}_T^{\min}(\Pi) = \min_{i<j} \widetilde{\mathcal{C}}_T(\Pi; \omega_i, \omega_j).$$

Information gain. $I_T = \log|\Omega| - H[p_T]$, where $H[p_T] = -\sum_{\omega} p_T(\omega) \log p_T(\omega)$.

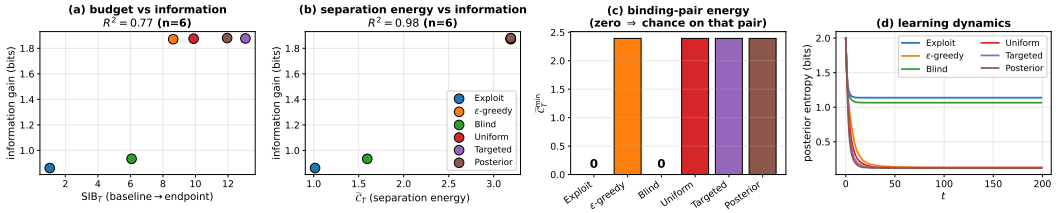


Figure 10: **The exploration mirage** (bears on Theorem 1, Corollary 4 and Corollary 36). (a) budget vs. information gain ($R^2 = 0.77$); (b) separation energy vs. information gain ($R^2 = 0.98$); (c) binding-pair energy $\widetilde{\mathcal{C}}_T^{\min}$, zero for Pure Exploit and Blind; (d) posterior entropy: Blind stalls near 1 bit, Targeted falls to $\sim$0. Six policies, so neither R^2 carries much weight; the dissociation is carried by panel (c), where two policies are exactly zero.

Identification error. $\mathrm{err}_T = \Pr(\hat{\omega}_T \neq \omega)$, $\hat{\omega}_T = \arg\max_\omega p_T(\omega)$.

Empirical efficiency and mirage ratio.

$$\widetilde{\eta}_T(\Pi) = \frac{\widetilde{\mathcal{C}}_T(\Pi)}{\mathrm{SIB}_T(\Pi)}, \qquad \mathcal{M}_T(\Pi) = \frac{1}{\widetilde{\eta}_T(\Pi)}. \tag{15}$$

These are the unnormalised analogues of η_T and its reciprocal. Because C_{up} is missing, $\widetilde{\eta}_T$ need not lie in $[0, 1]$—for Pure Exploit it is 1.91—so it is comparable across policies within a fixed channel but is not the efficiency of §2. Small $\widetilde{\eta}_T$ still indicates an exploration mirage; large $\widetilde{\eta}_T$ does not certify the converse.

Standard diagnostics (shown as misleading). Action entropy $H[\pi_t]$, action switches $\sum_t \mathbf{1}[a_t \neq a_{t-1}]$, evaluator diversity $D = 1 - \sum_s (\bar{\pi}(s))^2$.

M.17 Supplementary Mirage Figure

M.18 Monte Carlo Protocol

For each policy Π and each $\omega \in \Omega$: run $N_{\mathrm{MC}} = 2{,}000$ trajectories of length $T = 200$; record $(a_t, Y_t, f_t, p_t)_{t=1}^T$; report mean $\pm$ std and 95% CI.

M.19 Consistency Checks

1. Conservation: $I_T \leq C\,\mathbb{E}[\mathrm{SIB}_T]$ for all policies (Corollary 4).
2. Information-frontier sandwich (Theorem 30).
3. Information gap: $I_T^{\mathrm{targeted}} / I_T^{\mathrm{blind}} > 3$.

M.20 Predictions Examined

This is a simulation, so nothing here validates a theorem; the table records which prediction each observation bears on, and against which statement.

Statement	What the mirage experiment shows
Theorem 1 (information sandwich)	$I_T \leq C\,\mathbb{E}[\mathcal{C}_T]$ holds for all 6 policies
Corollary 4 (finite-step partition)	$\widetilde{\mathcal{C}}_T$ tracks I_T ($R^2 = 0.98$) better than SIB_T (0.77)
Corollary 36 (binding pair)	$\mathcal{C}_T^{\min} = 0$ forces chance error; blind attains it
Theorem 11 (stabilization)	Pure exploit: $\mathcal{C}_T \to 0$; blind collapses on the fine pairs

N Synthetic Plug-In Capacity Estimation

Observables. The estimator sees only what a real pipeline logs: the action sequence $a_{1:T}$, the observation sequence $Y_{1:T}$, the hypothesis class Θ (the learner's own model class), and the noise

scale σ. It does not see the true environment ω, the adaptation rate λ, the latent freshness state $f_t(\cdot)$, or any counterfactual response.

Step 1: strategic-state identification. In the mirage environment the latent state of the played arm satisfies $\mathbb{E}[Y_t^2] = f_t(a_t)^2\,\theta^2 + \sigma^2$ with $\theta^2 = 1$ across the hypothesis class, so the observation stream carries an unbiased fingerprint of the squared strategic state. We recover λ by empirical Bayes: for each candidate λ on a grid (120 points in $[0.01, 0.9]$), run the Bayesian filter over $\omega \in \Omega$ with candidate freshness $f_t = (1-\lambda)^{e_t}$ (exposure counts e_t are derived from the action log), accumulate the one-step-ahead predictive log-likelihood $\sum_t \log p(Y_t \mid Y_{1:t-1}, \lambda)$, and select the maximizer $\hat{\lambda}$.

Step 2: plug-in strategic capacity. Reconstruct $\hat{f}_t(s) = (1-\hat{\lambda})^{e_t(s)}$ from the action log; form plug-in means $\hat{\mu}_t^\omega = \theta^\omega_{a_t}\hat{f}_t(a_t)$ for each hypothesis; compute $\widehat{\Delta}_t = \max_\omega(\hat{\mu}_t^\omega - \hat{\mu}_{t+1}^\omega)^2/(2\sigma^2)$, $\widehat{\mathfrak{D}}_t^{(i,j)} = |\theta^i_{a_t} - \theta^j_{a_t}|\hat{f}_t(a_t)/(2\sigma)$, and $\widehat{\mathcal{C}}_T^{(i,j)} = \sum_t(\widehat{\mathfrak{D}}_t^{(i,j)})^2\widehat{\Delta}_t$; average (resp. minimize) over pairs for $\widehat{\mathcal{C}}_T$ (resp. $\widehat{\mathcal{C}}_T^{\min}$), and set $\widehat{\mathrm{SIB}}_T = \max_\omega \sum_t(\hat{\mu}_t^\omega - \hat{\mu}_{t+1}^\omega)^2/(2\sigma^2)$, $\hat{\eta}_T = \widehat{\mathcal{C}}_T/\widehat{\mathrm{SIB}}_T$.

Validation results. 6 policies $\times$ 4 environments $\times$ 2,000 runs, $T = 200$, $\lambda = 0.25$, $\sigma = 1$. Per-trajectory log-correlation between $\widehat{\mathcal{C}}_T$ and oracle $\mathcal{C}_T$: 0.94 (Spearman 0.80). Policy-level strategic capacity ordering: Kendall $\tau = 1.0$. Predictive power for information gain: $R^2 = 0.658$ (estimated) vs. 0.679 (oracle) vs. 0.591 (SIB). Median relative error of $\widehat{\mathcal{C}}_T$: 18–35% across policies. Estimated binding-pair efficiency is exactly 0 for Pure Exploit and Blind Alternation (the mirage policies), matching ground truth. Although $\hat{\lambda}$ is individually noisy (mean 0.34 vs. true 0.25; $n = 48{,}000$ trajectories), strategic capacity estimates remain accurate in this model: misestimated adaptation rescales $\widehat{\mathfrak{D}}_t^2$ and $\widehat{\Delta}_t$ together, partly offsetting the error in their product.

Scope. The protocol requires (i) an identifiable parametric form for the strategic response, and (ii) a known hypothesis class. These are strong, model-specific requirements and are not automatically available to a likelihood-based learner. The synthetic protocol does not require the response map to be queryable. Extending consistency guarantees beyond the exposure-decay family is a natural next step.

O Additional Consequences and Remarks

O.1 The Admissibility Axioms and the Extremal Canonicality Theorem

Admissible information budgets. A nonnegative functional $F_T(\Pi)$ is *admissible* if it satisfies: (A1) *Process invariance*:F_T depends on Π only through the induced kernels; (A2) *Temporal additivity*: F_T decomposes into one-step contributions; (A3) *Strategic nullity*: $F_T = 0$ when kernels do not depend on Π; (A4) *Uniform information domination*, there exists $C < \infty$ with $I(\omega; Y_{1:T} \mid \Pi) \le C\,F_T(\Pi) + C$ for all systems and policies; (A5) *Environment blindness* each f_t depends on the kernel path only through its temporal increment, not through any comparison across ω. Axiom (A5) formalises what it means to be a *budget*: a measure of what the policy spends, not of what it learns. It is indispensable rather than decorative; Remark 23 shows that without it the minimality claim below is false, with the paper's own strategic capacity functional as the counterexample.

Could a different notion of exploration budget yield a different theory? The next result says no: any admissible information currency dominates SIB, so the partition above is not an artifact of our accounting choices.

Theorem 17 (Extremal non-negligibility of temporal KL drift)**.** *Let $F_T(\Pi)$ be an admissible information budget satisfying (A1)–(A5), and assume reverse-Pinsker regularity (Assumption 6). Then there exists $c > 0$ and a family $\mathcal{E}$ of strategic feedback systems on which $F_T(\Pi) \ge c\,\mathrm{SIB}_T^{\mathrm{KL}}(\Pi)$. Consequently no admissible budget is uniformly negligible relative to the strategic information budget: there is no admissible F_T with $F_T = o(\mathrm{SIB}_T^{\mathrm{KL}})$ across all systems.*

Theorem 17 is a statement about the extremal family, not about every system; we do *not* claim $F_T \ge c\,\mathrm{SIB}_T^{\mathrm{KL}}$ pointwise, and axiom (A5) cannot be dropped, strategic capacity itself satisfies (A1)–(A4),

violates (A5), and is zero on systems with a positive budget. Details and the counterexample are in Remark 23, Appendix O. Remark 24 in Appendix O explains why the budget must be the KL one rather than the TV one. The proof constructs a two-environment system in which all observations outside step t are uninformative and then amplifies to $|\Omega| = 2^T$ independently informative steps, so that information grows linearly in T and the additive constant in (A4) becomes negligible; a single step cannot close the argument, because with $|\Omega| = 2$ information is capped at $\log 2$ and that constant absorbs it. Full proof in Appendix F. A sharper statement, which replaces "not negligible" by "equivalent" at the cost of smoothness, is the pair of characterisation theorems in Appendix Q.

O.2 Two Immediate Corollaries

Corollary 18 (Bound-form conservation)**.** $I(\omega; Y_{1:T} \mid \Pi) \le C\,\mathbb{E}[\mathrm{SIB}_T^{\mathrm{KL}}(\Pi)]$. *If in addition* $\mathcal{W}_T(\Pi) = 0$ *almost surely and capacity has an almost-sure floor, then in the identification-limited regime the matching lower bound of Theorem 1 applies with* SIB_T *in place of* $\mathcal{C}_T$*. The converse fails and we do not assert it:* $\mathcal{W}_T > 0$ *does not preclude two-sided comparability when* $\mathcal{W}_T \le a\,\mathcal{C}_T$*, and* $\mathcal{W}_T = 0$ *does not deliver it once information has saturated.*

Corollary 19 (Strategic uncertainty principle)**.** *For* $|\Omega| = M$ *and uniform prior:* $\inf_{\hat\omega} \Pr(\hat\omega \neq \omega) \ge 1 - (C\,\mathrm{SIB}_T^{\mathrm{KL}} + \log 2)/\log M$.

O.3 Remarks

Remark 20 (Why directional, and not an operator norm). Two alternatives are tempting and both are wrong for Theorem 1. Defining $\mathfrak{O}_t$ as the separation attained at unit perturbation confuses a value with a rate, and the two differ without bound once TV saturates. Defining it as the *operator norm*

$$\mathfrak{O}_t^{\mathrm{op}} := \frac{1}{\sqrt{C_{\mathrm{up}}}} \limsup_{\delta \downarrow 0} \sup_{\|v\|_\beta \le 1} \tfrac{1}{\delta}\mathrm{TV}(\cdot)$$

is worse, because an operator norm is a supremum: it reports what the *best* available direction would achieve, and can only upper-bound what the realised direction achieves. A policy whose motion lies in the kernel of the identification operator has $\mathfrak{O}_t^{\mathrm{op}} > 0$ and $\Delta_t > 0$ while generating no cross-environment separation at all, so no lower bound of the form $I \ge c\sum_t (\mathfrak{O}_t^{\mathrm{op}})^2 \Delta_t$ can hold. That is precisely the "active but blind" policy of Section 4.2: under the operator norm the theory would credit the mirage with strategic capacity it does not have. The directional modulus is the realised rate by construction, and is the quantity our estimators compute. The operator norm does have a natural upper role—it bounds what *any* policy could extract at the current baseline—but no result in this paper is stated in terms of it, and in particular Theorem 30 does not use it.

Remark 21 (More than two hypotheses). Strategic capacity as defined in (5) is a *pairwise* quantity, while $I(\omega; Y_{1:T} \mid \Pi)$ is not, so the two-sided bound is stated for binary Ω under a uniform prior and no single pair can characterize total mutual information when $|\Omega| > 2$. For finite Ω the statement is applied pairwise. The lower bound then holds with the best-separated pair *up to the prior mass of that pair*: conditioning on $\omega \in \{\omega_i, \omega_j\}$, an event of probability $2/M$ under the uniform prior, and applying Le Cam inside it gives

$$I(\omega; Y_{1:T} \mid \Pi) \;\gtrsim\; \frac{2}{M} \max_{(\omega_i,\omega_j)} \min\big\{\mathcal{C}_T(\omega_i, \omega_j),\, 1\big\},$$

not a bound with a universal constant. The factor $2/M$ is the same one that appears in Corollary 36. Identification of the full hypothesis set is governed instead by the *binding pair* $\min_{(\omega,\omega')} \mathcal{C}_T(\omega, \omega')$, which is the quantity our estimators report (Section 4) and the one that drives Corollary 36. We do not claim a two-sided characterization of $I(\omega; Y_{1:T} \mid \Pi)$ by any single pairwise strategic capacity for $|\Omega| > 2$.

Remark 22 (Why the exponent is 2). The weighting $\mathfrak{O}_t^2$ in (5) is forced, not chosen: the modulus (3) is a TV quantity while drift is KL, and the two directions of (6) use the two directions of the TV–KL comparison (Pinsker for the upper bound, reverse-Pinsker regularity, Assumption 6, for the lower). For $\mathcal{C}_T^{(p)} := \sum_t \mathfrak{O}_t^p\, \Delta_t$ with $p \neq 2$, one direction fails with any such constants. A witness suffices: fix $\Delta_t \equiv \Delta$ and $\mathfrak{O}_t \equiv \mathfrak{O}$ across T steps, so $\mathcal{C}_T^{(p)}/\mathcal{C}_T = \mathfrak{O}^{\,p-2}$. Letting $\mathfrak{O} \downarrow 0$ with Δ fixed sends this ratio to ∞ for $p < 2$ and to 0 for $p > 2$, while $I \asymp \mathcal{C}_T$ by (6)—the witness drives $\mathfrak{O} \downarrow 0$ with T and Δ fixed, so $\mathcal{C}_T \to 0$ and the truncation in the lower bound is inactive throughout. Hence $\mathcal{C}_T^{(p)}$ cannot

two-sided-bound I with constants independent of the instance. What follows from Corollary 34 is weaker than one might want: every functional that does two-sided-bound I is dominated by $\mathcal{C}_T$ in expectation unconditionally, and agrees with it up to channel-regularity constants only on policy classes with floor regularity (Definition 29). The exponent argument above is therefore a statement about the family $\mathcal{C}_T^{(p)}$ specifically, not a proof that $\mathcal{C}_T$ is the unique two-sided functional in general.

Remark 23 (What this does *not* say, and a counterexample). The theorem is a statement about the extremal family $\mathcal{E}$, not about every system. We do not claim $F_T \geq c\,\mathrm{SIB}_T^{\mathrm{KL}}$ pointwise: the proof lower-bounds f_t on a constructed pair of kernels, and carrying that bound to an arbitrary pair with the same scalar drift would require a further axiom— that f_t depend on the kernel increment only through a divergence, or be monotone under garbling.

Axiom (A5) cannot be dropped. Without it the pointwise claim is outright false, and the counterexample is internal to this paper: take $F_T := \mathcal{C}_T$, the strategic capacity itself. It satisfies (A1), (A4) but violates (A5), since $\mathfrak{O}_t$ is defined by comparing environments. Now consider a system whose kernels drift in time but do not depend on ω. Then $\mathfrak{O}_t \equiv 0$, so $\mathcal{C}_T = 0$, while $\mathrm{SIB}_T^{\mathrm{KL}} > 0$: the policy spends budget and converts none of it. No $c > 0$ satisfies $0 \geq c\,\mathrm{SIB}_T^{\mathrm{KL}}$. This is not an exotic construction but the waste-only corner of the paper's own decomposition $\mathrm{SIB} = \mathcal{C} + \mathcal{W}$, which already implies $\mathcal{C}_T \leq \mathrm{SIB}_T$ and so means SIB_T cannot be minimal among functionals that include strategic capacity. Axiom (A5) is what excludes them.

Remark 24 (The budget must be the KL one). The corresponding statement for $\mathrm{SIB}_T^{\mathrm{TV}}$ is *false*. Take $F_T := \mathrm{SIB}_T^{\mathrm{KL}}$, which is admissible: (A1)–(A3) are immediate and (A4) is Corollary 18. By Pinsker (8) and its reverse (Assumption 6) $\Delta_t^{\mathrm{KL}} \asymp (\Delta_t^{\mathrm{TV}})^2$, so $\mathrm{SIB}_T^{\mathrm{KL}}/\mathrm{SIB}_T^{\mathrm{TV}} \to 0$ as the per-step drifts shrink and no constant c can satisfy $\mathrm{SIB}_T^{\mathrm{KL}} \geq c\,\mathrm{SIB}_T^{\mathrm{TV}}$. This is not an artefact of the proof: Le Cam gives $\Theta(\delta^2)$ mutual information from a step of TV size δ, so any argument routed through (A4) yields a quadratic bound, and $\sum_t (\Delta_t^{\mathrm{TV}})^2 \asymp \mathrm{SIB}_T^{\mathrm{KL}}$ is exactly where it lands. The KL budget is the one used throughout this paper, in particular $\eta_T := \mathcal{C}_T/\mathrm{SIB}_T^{\mathrm{KL}}$.

P THE LOCAL GEOMETRY: PROOFS AND FAILURE MODES

This appendix proves Theorem 2 and Proposition 3, and records what goes wrong in the version

Proof of Theorem 2. By nondegeneracy $G_{\Delta,t} \succ 0$, so $\langle\cdot,\cdot\rangle_{\Delta,t}$ is an inner product and $\mathbb{R}^d = \ker J_t \oplus (\ker J_t)^{\perp_\Delta}$ is an orthogonal direct sum. Writing $\delta u_t = P_{\mathrm{id},t}\delta u_t + P_{\mathrm{null},t}\delta u_t$ and expanding the norm, the cross term vanishes by orthogonality, which is the displayed identity. The ratio $\mathcal{C}_t^{\mathrm{geo}}/q_t(\delta u_t)$ lies in $[0,1]$ because both summands are nonnegative and is the squared cosine of the angle between δu_t and the identifiable subspace by definition of an orthogonal projection. Finally $P_{\mathrm{null},t}\delta u_t \in \ker J_t$, so $J_t P_{\mathrm{null},t}\delta u_t = 0$ and $J_t\delta u_t = J_t P_{\mathrm{id},t}\delta u_t$. □

Proof of Proposition 3. Let B be a basis of $(\ker J_t)^{\perp_\Delta}$ orthonormal for $\langle\cdot,\cdot\rangle_{\Delta,t}$. Since $\|v\|_{\Delta,t}^2 = \frac{1}{2}v^\top G_{\Delta,t}v$, orthonormality means $\frac{1}{2}B^\top G_{\Delta,t}B = I$, that is

$$B^\top G_{\Delta,t}B = 2I, \tag{16}$$

and we keep this convention throughout; writing $B^\top G_{\Delta,t}B = I$ instead would be orthonormality for $v^\top G_{\Delta,t}v$, a different metric. Put $M := B^\top J_t^* G_t J_t B$. For $w = Bc$ in the subspace, $\|J_t w\|_{G_t}^2 = c^\top M c$ and $\|w\|_{\Delta,t}^2 = \|c\|^2$ by (16), so the Rayleigh quotient $\|J_t w\|_{G_t}^2/\|w\|_{\Delta,t}^2$ ranges over $[\lambda_{\min}(M), \lambda_{\max}(M)]$; set $\sigma_{\min,t}^2 := \lambda_{\min}(M)$ and $\sigma_{\max,t}^2 := \lambda_{\max}(M)$, and apply this to $w = P_{\mathrm{id},t}\delta u_t$, using $J_t\delta u_t = J_t w$ from the previous proof. The universal upper bound is immediate from (7): since $\bar{D}_t^* G_t \bar{D}_t \succeq 0$ we have $G_{\Delta,t} \succeq \frac{1}{4}J_t^* G_t J_t$, hence by (16) $2I = B^\top G_{\Delta,t}B \succeq \frac{1}{4}M$, so $\sigma_{\max,t}^2 = \lambda_{\max}(M) \leq 8$, with equality when $\bar{D}_t$ vanishes on the identifiable subspace. The final claim combines the sandwich with Theorem 1 and the total-variation-to-Fisher comparison of Assumption 6. □

Q TWO CHARACTERISATION THEOREMS

The extremal bound of Theorem 17 is weak: it says no admissible budget is negligible beside SIB_T on a constructed family. Under the local geometry of §3.2 one can say considerably more, and about the right object. Both results below are corollaries of Chentsov's theorem (Čencov, 1982): on a finite sample space the Fisher metric is, up to a positive scalar, the unique Riemannian metric invariant under sufficient statistics and monotone under Markov morphisms.

A caution about which axiom does the work. Monotonicity under garbling is not sufficient to pin down the form. Many quadratic forms contract under lossy garbling — $\sum_i v_i^2$, $\sum_i v_i^2/p_i^2$ and $\sum_i v_i^2/\sqrt{p_i}$ all do, in 20,000 random trials each. What separates the Fisher form is *invariance* under a lossless refinement: a sufficient statistic discards nothing, so an admissible currency must be unchanged by it, and the three alternatives above change by 39%, 98% and 25% respectively (`verify_chentsov.py`). The invariance axiom is therefore load-bearing and we state it explicitly rather than folding it into data processing.

The setting Chentsov's theorem needs. The axioms below are stated for a currency defined on a family of experiments, not for a single number, and we make that explicit because the strength of the conclusion depends on it. Fix a finite observation alphabet $\mathcal{Y}$. A *one-step experiment* is a pair of probability vectors on $\mathcal{Y}$; the currency assigns to each such pair, at each base point, a value, and it is required to do so coherently across all finite $\mathcal{Y}$ simultaneously. A *Markov morphism* is a stochastic map $\mathcal{Y} \to \mathcal{Y}'$ applied to both members of the pair; a *congruent embedding* is one that is invertible on the image, i.e. a sufficient statistic. Axiom (L4) means the value is unchanged by every congruent embedding, (L5) that it does not increase under any Markov morphism, and (L7) that the value on one fixed reference pair is the same for all $\mathcal{Y}$—so the normalising constant is shared across time, histories and alphabets rather than chosen per step. Chentsov's theorem applies to this data and returns a single positive multiple of the Fisher form, the same multiple everywhere. Without the coherence requirement the axioms would only pin the form down alphabet by alphabet, and the theorems below would not follow; we state it rather than leaving the appeal to Chentsov informal.

Definition 25 (Learning currency)**.** A sequential functional G_T is a *learning currency* if: (L1) each one-step contribution depends only on the pair of endpoint kernels $(K_t^{\omega_0,\Pi}, K_t^{\omega_1,\Pi})$; (L2) it vanishes exactly when those kernels coincide; (L3) it is twice differentiable and locally quadratic in a smooth perturbation; (L4) it is invariant under sufficient statistics; (L5) it is non-increasing under garbling; (L6) contributions add over steps; (L7) it takes a fixed value on a reference binary experiment.

Theorem 26 (Separation energy is canonical among learning currencies)**.** Local, second order. *Under the smoothness of §3.2, every learning currency satisfies, for each t,*

$$g_t \;=\; c\,\|J_t\,\delta u_t\|_{G_t}^2 \;+\; o(\|\delta u_t\|^2),$$

with c fixed by the normalisation (L7) and the remainder uniform over a neighbourhood of the base point. Consequently, along any sequence of policies with $\max_t \|\delta u_t\| \to 0$ and T fixed, or uniformly over policies whose displacements lie in a neighbourhood small enough that the remainders are dominated by the leading terms,

$$G_T(\Pi) \;\asymp\; \sum_{t=1}^{T} \|J_t\,\delta u_t\|_{G_t}^2 \;\asymp\; \sum_{t=1}^{T} S_t^2,$$

with constants depending only on the normalisation (L7) and, for the second comparison, on the reverse-Pinsker constant of Assumption 6. If in addition the uniform conditioning of Proposition 3 holds, $\sigma^2_{\min,t} \geq \sigma^2 > 0$ on $(\ker J_t)^{\perp_\Delta}$ for all t, then also $G_T(\Pi) \asymp \sum_t \mathcal{C}_t^{\mathrm{geo}}$, with constants σ^2 and 8.

Proof. By (L1) and (L3) the one-step contribution is a quadratic form on the tangent space at the common baseline kernel; by (L4) and (L5) with Chentsov's theorem it is a positive multiple of the Fisher form, the multiple fixed by (L7). The tangent vector separating the environments is $J_t\,\delta u_t$, so the contribution is $c\,\|J_t\delta u_t\|_{G_t}^2$, using (L2) to rule out a constant term; the comparison with S_t^2 is the two-sided relation between total variation and the Fisher norm under Assumption 6. Summing over t by (L6) gives the first two comparisons. The third is Proposition 3, not Theorem 2: what Chentsov

identifies is *separation energy*, and passing from it to the Pythagorean component requires two-sided singular-value control of J_t on the identifiable subspace. The upper direction is unconditional ($\sigma^2_{\max,t} \le 8$); the lower is the assumption. Without it the theorem still canonicalises $\sum_t \|J_t \delta u_t\|^2_{G_t}$, and that is the object we claim is canonical. □

Definition 27 (Cost currency)**.** A sequential functional Cost_T is a *cost currency* if it satisfies (L3)–(L7) and, in place of (L1)–(L2): (C1) *environment symmetry*—each contribution is the average over $\omega \in \Omega$ of a functional of the displacement of K_t^ω from baseline to endpoint, so that the two environments enter on the same footing and neither is singled out; and (C2) it vanishes exactly when both displacements are zero.

Theorem 28 (The symmetric local cost is canonical among cost currencies)**.** Local, second order, in the same sense as Theorem 26. *Every cost currency satisfies* $\mathrm{Cost}_T(\Pi) \asymp \sum_{t=1}^T \bar\Delta_t^{\mathrm{KL}}$ *under the same small-displacement qualification, where* $\bar\Delta_t^{\mathrm{KL}}$ *denotes the exact environment-averaged KL and* (7) *is its quadratic part. The operational budget is equivalent to that object to second order in the binary local regime,* $\sum_t \bar\Delta_t^{\mathrm{KL}} \le \mathrm{SIB}_T^{\mathrm{KL}}(\Pi) \le 2\sum_t \bar\Delta_t^{\mathrm{KL}}$ *exactly but is not itself the canonical object: it takes a supremum over* Ω*, is therefore not a quadratic form, and satisfies none of the axioms that Chentsov's theorem needs.*

Proof. By (L3) each summand of the average in (C1) is a quadratic form in $D_t^\omega \delta u_t$; by (L4), (L5) and Chentsov it is a positive multiple of the Fisher form $\|D_t^\omega \delta u_t\|^2_{G_t}$, the multiple fixed by (L7) and, by (C1), the same for both environments. Averaging over Ω gives a multiple of $\delta u_t^\top G_{\Delta,t}\, \delta u_t$ by (7), which is $q_t(\delta u_t)$, and hence $\bar\Delta_t^{\mathrm{KL}}$ up to that multiple and the second-order remainder; (C2) rules out a constant term. Summing by (L6) and applying $\bar\Delta_t^{\mathrm{KL}} \le \Delta_t^{\mathrm{KL}} \le 2\bar\Delta_t^{\mathrm{KL}}$ gives the claim.

The averaging in (C1) is not cosmetic. The operational budget takes a supremum over ω, and a maximum of two quadratic forms is not quadratic, so no Chentsov argument applies to it directly and it admits no orthogonal decomposition; what the binary prior buys is that the two differ by at most a factor of 2, which is why the characterisation survives as a comparison but not as an identity. □

Together, the two theorems give the structure the identity of Corollary 4 was reaching for:

$$\underbrace{\text{canonical symmetric cost}}_{\sum_t \bar\Delta_t^{\mathrm{KL}}} = \underbrace{\text{identifiable component}}_{\sum_t \mathcal{C}_t^{\mathrm{geo}}} + \underbrace{\text{null component}}_{\sum_t \mathcal{W}_t^{\mathrm{geo}}},$$

exactly, with the left-hand side characterised by the cost axioms and the first summand comparable to the learning currency the other axioms characterise. The two characterisations use the same classical theorem applied to two aspects of the same map: J_t, on which the environments disagree, and the symmetric average of the two derivatives, which treats them alike. Three quantities are kept apart throughout: *separation energy* $\|J_t \delta u_t\|^2_{G_t}$, which Chentsov canonicalises; *cost geometry* $\bar\Delta_t^{\mathrm{KL}}$ with quadratic part q_t, which the cost axioms canonicalise; and the *orthogonal projected capacity* $\mathcal{C}_t^{\mathrm{geo}}$, which is what the first is comparable to and the second decomposes.

R Additional Theoretical Results

R.1 The Strategic Capacity Frontiers

Two frontiers, because capacity is random. Capacity is a random variable, so a supremum over policies needs an ordering, and the two orderings the theorem supplies do not agree. We therefore define both

$$\mathcal{C}^*_{\mathrm{mean}}(B) := \sup\big\{\mathbb{E}[\mathcal{C}_T(\Pi)] : \mathbb{E}[\mathrm{SIB}_T(\Pi)] \le B\big\}, \qquad \mathcal{C}^*_{\mathrm{floor}}(B) := \sup\big\{\operatorname{ess\,inf} \mathcal{C}_T(\Pi) : \mathbb{E}[\mathrm{SIB}_T(\Pi)] \le B\big\},$$

and keep them apart. A policy can have large expected capacity that is carried by rare histories while its capacity is zero on most of them, so $\mathcal{C}^*_{\mathrm{floor}}(B) \le \mathcal{C}^*_{\mathrm{mean}}(B)$ with the inequality generally strict. Every frontier statement below is sandwiched between the two and we never substitute one for the other.

Definition 29 (Floor-to-mean regularity)**.** A class of policies satisfies α-*floor regularity* for $\alpha \in (0, 1]$ if $\operatorname{ess\,inf} \mathcal{C}_T(\Pi) \ge \alpha\, \mathbb{E}[\mathcal{C}_T(\Pi)]$ for every Π in the class.

This is the single hypothesis that converts every sandwich below into an equivalence, and we state it once rather than assuming it silently in each. It holds with $\alpha = 1$ when capacity is almost surely constant, and fails exactly when capacity is carried by rare histories. We do not verify it for any system in this paper, and every result that needs it says so.

Is the limit attainable, or merely an upper bound? Impossibility results are cheap if nothing approaches them. The next theorem does *not* settle this. It sandwiches the frontier between the two capacity frontiers, and because $\mathcal{C}^*_{\text{floor}}$ and $\mathcal{C}^*_{\text{mean}}$ can differ arbitrarily the sandwich is not a tight characterisation; we say so plainly in Remark 31 and label the result accordingly.

We additionally require: (iii) *strategic response decomposition* (Lipschitz BR; Assumption 3) and (iv) *KL-Lipschitz observation kernels* (Assumption 4). Both bounds are corollaries of Theorem 1 and therefore inherit its hypotheses, including the two-sided drift regularity of Assumption 5 and the reverse-Pinsker condition of Assumption 6. Full statements are in Appendix B.

Theorem 30 (Information-frontier sandwich)**.** *Under Assumptions 1–6, for any two environments ω_0, ω_1 with uniform prior, the strategic information frontier satisfies*

$$c\,\min\{\mathcal{C}^*_{\text{floor}}(B),\,1\} \;\le\; \mathcal{F}_T(B) \;\le\; C\,\mathcal{C}^*_{\text{mean}}(B) \;\le\; C\,B,$$

where $\mathcal{F}_T(B) := \sup_{\Pi:\,\mathbb{E}[\text{SIB}_T]\le B} I(\omega; Y_{1:T} \mid \Pi)$, an expected *budget, the same constraint that defines $\mathcal{C}^*_{\text{mean}}$ and $\mathcal{C}^*_{\text{floor}}$ above, since a pathwise constraint would shrink the feasible set and break the comparison and $C_{\text{up}} = (G_x L_{\text{cont}}/(1-\rho))^2$.*

Proof sketch. Both directions are corollaries of Theorem 1; no separate construction is required. The upper bound is $I \le C\,\mathbb{E}[\mathcal{C}_T] \le C\,\mathbb{E}[\text{SIB}_T] \le CB$. For the lower bound, take any feasible Π whose *essential infimum* capacity is within ε of $\mathcal{C}^*_{\text{floor}}(B)$ and apply the lower bound of (6) to that policy directly. Full proof in Appendix G. □

Remark 31 (The sandwich is not a tightness result). Theorem 30 bounds $\mathcal{F}_T(B)$ below by $\mathcal{C}^*_{\text{floor}}(B)$ and above by $\mathcal{C}^*_{\text{mean}}(B)$, and these are different functionals: a policy whose capacity is carried by rare histories has large mean and zero floor, so the inequality $\mathcal{C}^*_{\text{floor}} \le \mathcal{C}^*_{\text{mean}}$ can be strict and we know of no bound on the ratio. Three things follow that we do not want read into the statement. The lower-bound proof selects a policy near the *definition* of $\mathcal{C}^*_{\text{floor}}$; it constructs nothing and it does not exhibit a policy attaining the upper frontier. Nothing here says a capacity-aligned policy converts budget at a maximal rate. And the result is not an achievability theorem in the coding-theoretic sense, which is why we have renamed it.

The gap closes in one case we can state: if $\mathcal{C}_T(\Pi)$ is almost surely constant for every policy in the feasible set then $\mathcal{C}^*_{\text{floor}} = \mathcal{C}^*_{\text{mean}}$ and $\mathcal{F}_T(B) \asymp \mathcal{C}^*(B)$. Adaptive policies in systems with memory are exactly the case this excludes, and it is the case we care about. What would close it in general is a concentration hypothesis under which $\mathcal{C}^*_{\text{mean}}(B) \lesssim \mathcal{C}^*_{\text{floor}}(B)$, or a construction attaining the upper frontier. We have neither, and we flag the question as open rather than presenting the sandwich as a characterisation.

R.2 Necessity: Strategic Capacity Is Unavoidable

Could a clever policy learn optimally while ignoring strategic capacity? Could a future paper propose a different strategic-exploration metric and characterize optimal learning with it? The results above say strategic capacity is *sufficient* to account for learning. This section says it is *necessary* in one direction and, without an extra hypothesis, only in that direction: no diagnostic can certify learning that capacity does not already certify (Corollary 33, unconditional), and no near-optimal policy can spend negligible expected capacity (Corollary 32). Whether a *different* metric could two-sided-characterize learning without agreeing with capacity is settled only under floor regularity (Definition 29); we flag the general case as open. Call a policy Π with $\mathbb{E}[\text{SIB}_T(\Pi)] \le B$ γ-*information-optimal* at budget B if $I(\omega; Y_{1:T} \mid \Pi) \ge \gamma\,\mathcal{F}_T(B)$ for some $\gamma \in (0, 1]$. The constraint is on the expectation, the same feasible set that defines $\mathcal{F}_T(B)$, $\mathcal{C}^*_{\text{mean}}$ and $\mathcal{C}^*_{\text{floor}}$; using the pathwise constraint $\text{SIB}_T \le B$ here instead would compare a policy against the supremum over a strictly larger set and make γ-optimality harder to attain than the definition intends.

Corollary 32 (Near-optimal learners are capacity-seeking)**.** *Under Assumptions 1–6, every γ-information-optimal policy satisfies, in the identification-limited regime $\mathcal{C}^*_{\text{floor}}(B) \le 1$,*

$$\mathbb{E}[\mathcal{C}_T(\Pi)] \;\ge\; \frac{\gamma\,c}{C}\,\min\{\mathcal{C}^*_{\text{floor}}(B),\,1\},$$

where c and C are the lower and upper constants in (6). The constant is $\gamma c/C$, not $\gamma/(C_{\mathrm{up}}C)$: the lower bound now comes from Theorem 1 rather than from the discarded per-step construction whose constant was $1/C_{\mathrm{up}}$. This inherits the hypotheses of Theorem 1 through Theorem 30, reverse Pinsker included.

Proof. $C\,\mathbb{E}[\mathcal{C}_T(\Pi)] \geq I(\omega; Y_{1:T} \mid \Pi) \geq \gamma\,\mathcal{F}_T(B) \geq \gamma\, c\, \min\{\mathcal{C}^*_{\mathrm{floor}}(B), 1\}$, using the upper bound in (6) and the lower bound in Theorem 30. Every term in this chain is deterministic: I and $\mathcal{F}_T$ are numbers, $\mathcal{C}^*_{\mathrm{floor}}$ is a supremum of essential infima, and the only random quantity, $\mathcal{C}_T(\Pi)$, appears under an expectation. In the identification-limited regime $\mathcal{C}^*_{\mathrm{floor}}(B) \leq 1$ this reads $\mathbb{E}[\mathcal{C}_T(\Pi)] \geq (\gamma c/C)\,\mathcal{C}^*_{\mathrm{floor}}(B)$; once information has saturated the conclusion weakens to $\mathbb{E}[\mathcal{C}_T(\Pi)] \geq \gamma c/C$, since a policy cannot be forced to spend strategic capacity it can no longer convert. The conclusion binds the mean capacity and not the floor: a near-optimal policy must spend capacity on average, and the argument does not force it to do so on every history. □

Contrapositively: a policy whose expected strategic capacity is a vanishing fraction of $\mathcal{C}^*_{\mathrm{floor}}(B)$ forfeits information-optimality, *regardless of how its exploration is motivated or scored*. The same logic constrains diagnostics, not just policies.

Corollary 33 (No capacity-free certificate of learning)**.** *Let G_T be any (possibly random) functional of the induced kernel process such that*

$$I(\omega; Y_{1:T} \mid \Pi) \;\geq\; c_G\,\mathbb{E}[G_T(\Pi)]$$

for all strategic feedback systems satisfying Assumptions 1–6 and all policies (i.e., high G_T is consistent with learning). Then $\mathbb{E}[G_T(\Pi)] \leq (C/c_G)\,\mathbb{E}[\mathcal{C}_T(\Pi)]$: every valid certificate of learning is dominated in expectation by strategic capacity. The premise is stated on $\mathbb{E}[G_T]$ because mutual information is a number; a pathwise premise $I \geq c_G G_T$ would need G_T almost surely bounded to be taken in expectation, and we do not assume that.

Proof. $c_G\,\mathbb{E}[G_T(\Pi)] \leq I(\omega; Y_{1:T} \mid \Pi) \leq C\,\mathbb{E}[\mathcal{C}_T(\Pi)]$, the second inequality being the upper half of (6), which carries the reverse-Pinsker constant of Assumption 6 and the two-sided drift regularity of Assumption 5 used to pass from accumulated separation to capacity. Both are therefore hypotheses of this corollary, and the assumption range above states them. □

Corollary 34 (Uniqueness of two-sided characterizations)**.** *Let G_T satisfy $c_1\mathbb{E}[G_T(\Pi)] \leq I(\omega; Y_{1:T} \mid \Pi) \leq c_2\mathbb{E}[G_T(\Pi)]$ across all systems and policies satisfying Assumptions 1–6. Then*

$$\frac{c}{c_2}\;\min\big\{\operatorname{ess\,inf}\mathcal{C}_T(\Pi),\,1\big\} \;\leq\; \mathbb{E}[G_T(\Pi)] \;\leq\; \frac{C}{c_1}\,\mathbb{E}[\mathcal{C}_T(\Pi)]\,:$$

This is not yet an equivalence. *The left-hand side carries the essential infimum of $\mathcal{C}_T$ and the right-hand side its mean, and by Definition 29 those can differ arbitrarily, so the display does* not *give $\mathbb{E}[G_T] \asymp \mathbb{E}[\mathcal{C}_T]$. It becomes an equivalence under α-floor regularity: if $\operatorname{ess\,inf}\mathcal{C}_T \geq \alpha\,\mathbb{E}[\mathcal{C}_T]$ then, in the identification-limited regime where $\mathbb{E}[\mathcal{C}_T] \leq 1$,*

$$\frac{c\,\alpha}{c_2}\,\mathbb{E}[\mathcal{C}_T(\Pi)] \;\leq\; \mathbb{E}[G_T(\Pi)] \;\leq\; \frac{C}{c_1}\,\mathbb{E}[\mathcal{C}_T(\Pi)],$$

so any functional that two-sided-characterizes information agrees with strategic capacity up to constants depending on the channel's regularity and on α, none of them universal. Without floor regularity the conclusion is only the one-sided domination of Corollary 33. Once $\mathbb{E}[\mathcal{C}_T] \gtrsim 1$ even that weakens, because mutual information has saturated and cannot distinguish functionals that continue to grow.

Proof. The display combines both directions of (6) with both directions of the hypothesis; the truncation on the left is inherited from the lower bound of Theorem 1. For the equivalence, floor regularity gives $\operatorname{ess\,inf}\mathcal{C}_T \geq \alpha\mathbb{E}[\mathcal{C}_T]$, and in the regime $\mathbb{E}[\mathcal{C}_T] \leq 1$ the truncation $\min\{\cdot, 1\}$ is inactive on $\alpha\mathbb{E}[\mathcal{C}_T] \leq 1$, so the left-hand side is at least $(c\alpha/c_2)\mathbb{E}[\mathcal{C}_T]$. Both steps fail without the hypothesis: the first because the floor can be zero while the mean is not, the second because the truncation binds. □

Three remarks. First, these proofs are short because necessity is a structural consequence of the two-sided bound, not an additional axiom. Entropy, switching frequency, and SIB itself all fail the premise of Corollary 33—none of them lower-bounds information (§4.2 exhibits the failure)—so none can certify learning. Among the diagnostics we examined, capacity is the only one that can; this is a statement about that list, not a uniqueness theorem, which would need floor regularity. Second, Corollary 34 constrains what a replacement metric could look like, but only under floor regularity: within the no-static-signature class, below saturation, and for policy classes satisfying Definition 29, any functional that two-sided-characterizes learning is equivalent to strategic capacity up to constants depending on the channel and on α. Without that hypothesis what survives is one-sided: every valid certificate is dominated by capacity, and no functional can certify learning while capacity is zero. We are not able to rule out a replacement metric that agrees with information but not with capacity on policy classes whose capacity is carried by rare histories, and we do not claim otherwise. Outside the no-static-signature class the question is open entirely. Third, on assumption load: all three corollaries inherit the reverse-Pinsker regularity of Assumption 6. The reason is that reverse Pinsker sits in the *upper* bound of (6), which converts information into accumulated squared total variation, and every one of the three corollaries uses that direction. It is the *lower* bound whose constant c is universal, by Le Cam's two-point argument, so the asymmetry runs the opposite way from the one previously stated. Pinsker's inequality (8) supplies the KL→TV direction and is indeed universal, but that is not the direction these corollaries need.

Multi-hypothesis impossibility. The results above are pairwise. They extend to M-ary identification, and the extension singles out exactly the aggregate capacities reported in our experiments. Write $\bar{\mathcal{C}}_T := \binom{M}{2}^{-1} \sum_{i<j} \mathcal{C}_T(\Pi; \omega_i, \omega_j)$ for the average pairwise strategic capacity and $\mathcal{C}_T^{\min} := \min_{i<j} \mathcal{C}_T(\Pi; \omega_i, \omega_j)$ for the binding pair. Both are random variables, so the displays below carry $\mathbb{E}[\cdot]$ on the upper side and an essential infimum on the lower; mutual information and Bayes error are numbers and cannot be bounded by a random quantity unqualified.

Theorem 35 (Fano-type multi-hypothesis bound)**.** *Let $|\Omega| = M \geq 2$ with uniform prior. Under Assumptions 1–6—the full set, because the per-pair step $\mathbb{E}[\mathrm{KL}(P_i \| P_j)] \leq C\, \mathcal{C}_T(\omega_i, \omega_j)$ uses the reverse-Pinsker constant of Assumption 6 and the two-sided drift regularity of Assumption 5 that translates separation into capacity*

$$I(\omega; Y_{1:T} \mid \Pi) \;\leq\; C\, \frac{M-1}{M}\, \mathbb{E}[\bar{\mathcal{C}}_T] \;\leq\; C\, \mathbb{E}[\bar{\mathcal{C}}_T], \qquad \textit{and hence} \qquad \inf_{\hat{\omega}} \Pr(\hat{\omega} \neq \omega) \;\geq\; 1 - \frac{C\, \mathbb{E}[\bar{\mathcal{C}}_T] + \log 2}{\log M}.$$

Proof. With uniform prior, $I = \frac{1}{M} \sum_i \mathbb{E}[\mathrm{KL}(P_i \,\|\, \bar{P})]$ where P_i is the trajectory law under ω_i and $\bar{P}$ the mixture. By convexity of KL in its second argument, $\mathrm{KL}(P_i \| \bar{P}) \leq \frac{1}{M} \sum_j \mathrm{KL}(P_i \| P_j)$. By the chain rule and the per-pair upper-bound argument of Theorem 1, $\mathbb{E}[\mathrm{KL}(P_i \| P_j)] \leq C\, \mathbb{E}[\mathcal{C}_T(\Pi; \omega_i, \omega_j)]$, which is where Assumptions 5 and 6 enter; $\bar{\mathcal{C}}_T$ below denotes the expected average pairwise capacity. Averaging over the $M(M-1)$ ordered pairs gives the first display; Fano's inequality gives the second. □

Theorem 35 strengthens the uncertainty principle of Corollary 19, since $\bar{\mathcal{C}}_T \leq \mathrm{SIB}_T$: the budget in the error bound can be replaced by its strategic capacity share. A Le Cam two-point argument localizes the impossibility to the binding pair.

Corollary 36 (The mirage is forced by the binding pair)**.** *Under Assumptions 1–6, for any pair (ω_i, ω_j) and uniform prior over M environments,*

$$\inf_{\hat{\omega}} \Pr(\hat{\omega} \neq \omega) \;\geq\; \frac{1}{M}\Big(1 - \sqrt{C\, \mathbb{E}[\mathcal{C}_T(\Pi; \omega_i, \omega_j)]/2}\Big) \;\geq\; \frac{1}{M}\Big(1 - \sqrt{C\, \mathbb{E}[\mathcal{C}_T^{\min}]/2}\Big),$$

and conditionally on $\omega \in \{\omega_i, \omega_j\}$ the identification error is at least $\frac{1}{2}\big(1 - \sqrt{C\, \mathbb{E}[\mathcal{C}_T(\Pi; \omega_i, \omega_j)]/2}\big)$. In particular, zero binding-pair strategic capacity forces chance-level error on that pair.

Proof. Restrict to the event $\omega \in \{\omega_i, \omega_j\}$ (probability $2/M$); the induced binary Bayes error is at least $\frac{1}{2}(1 - \mathrm{TV}(P_i, P_j))$ by Le Cam. Pinsker gives $\mathrm{TV}(P_i, P_j) \leq \sqrt{\mathrm{KL}(P_i \| P_j)/2} \leq \sqrt{C\, \mathcal{C}_T(\Pi; \omega_i, \omega_j)/2}$, using the per-pair bound of Theorem 1. □

Corollary 36 converts the headline experiment from an observation into a theorem: Blind Alternation has $\mathcal{C}_T^{\min} = 0$ (Table 1), so its error on the fine pair is at least $1/2$. In this zero-capacity case the bound is exact so the observed 50.7% is the bound rather than an empirical accident; for positive capacity the numerical constant C is not sharp and we claim no attainment. It also explains why $\mathcal{C}_T^{\min}$, not average strategic capacity, is the identification bottleneck, and why $\bar{\mathcal{C}}_T$ is the correct predictor of total information (Theorem 35), precisely the two aggregates used in §4.

R.3 Canonical Examples

Two examples bracket the law's economic content; full treatments are in Appendix I. In the *tragedy of the learning commons*, N agents choose exploration effort at private cost while a public signal $Y_t \sim \mathcal{N}(\omega, \sigma^2/A_t)$ depends on aggregate effort: equilibrium effort falls strictly below the optimum, marginal incentives vanish as the posterior sharpens, equilibrium stabilizes (Lemma 15, Appendix K), and by Theorem 11 the continuation-mediated part of observational separation vanishes, SIB is a public good whose production cost is private. In *incentivized exploration*, a BIC planner escapes nothing: burn-in works precisely because it concentrates distributional motion where $\mathfrak{D}_t > 0$, spending the budget up front before the policy stabilizes.

The commons admits a closed form, which pins down exactly what governance costs in information terms.

Proposition 37 (Commons growth law and the free-riding wedge)**.** *Consider the Gaussian learning commons with flow payoff $-1/\tau_t$, quadratic effort cost c, discount $\delta \in (0,1)$, and effort multiplier m ($m = N$ under myopic equilibrium, $m = N^2$ under the social optimum). Put $K := m/(c\,\sigma^4(1-\delta))$. Then*

$$\tau_t \;=\; (3Kt)^{1/3}\Big(1 + O\big(\tfrac{1}{t(1-\delta)}\big)\Big), \qquad \log \tau_t \;=\; \tfrac{1}{3}\log(3Kt) + O\big(\tfrac{1}{t(1-\delta)}\big). \tag{17}$$

Consequently: (i) universality of the exponent—*the growth exponent $1/3$ is determined by power matching alone and is independent of N, δ, σ^2, c and m, which enter only through the prefactor K; and* (ii) the free-riding wedge—

$$\log \tau_t^{\text{social}} - \log \tau_t^{\text{eq}} \;=\; \tfrac{1}{3}\log N \;+\; O\big(\tfrac{1}{t(1-\delta)}\big). \tag{18}$$

Proof in Appendix E.

Equation (18) sharpens the usual reading of the tragedy. Decentralisation does not merely underproduce the budget; it costs a *fixed, closed-form quantity of knowledge*, $\frac{1}{3}\log N$, however the parameters are set. And because the $O(\cdot)$ correction has the same form in both regimes, it cancels in the difference: the comparative statement is accurate long before either absolute level has reached its asymptote (§S.1). The contrast is the principle: *someone must pay the information budget* decentralized systems underproduce it; centralized designers finance it through deliberate non-stationarity.

S Additional Experiments

S.1 Conservation: MI is Bounded by SIB

We sweep the Gaussian commons over $N \in \{2, 5, 10, 20, 50\}$, $\delta \in \{0.3, 0.5, \ldots, 0.99\}$, and $\sigma^2 \in \{0.5, 1, 2\}$, computing both the myopic equilibrium and the social optimum ($T = 300$): 90 parameter regimes, 180 points.

The law binds, and the closed form predicts the level with no free parameters. Log-precision rises with cumulative SIB in both regimes (slope $+2.90$ on $\log \text{SIB}$, $p = 1.2 \times 10^{-31}$), and the bound $I \le C\,\text{SIB}$ holds with $C = 27.9$. More sharply, Proposition 37 also predicts an absolute precision level from (N, δ, σ^2, m) alone.

The wedge is exact; the exponent is asymptotic. The predicted wedge $\frac{1}{3}\log N$ matches the measured one to 2×10^{-4} at every N (Table 9), whereas the absolute level is still 5×10^{-2} from its asymptote at $T = 300$ and the measured growth exponent is ≈ 0.37 rather than $1/3$. Both facts are predicted: the exponent approaches $1/3$ only as $t(1-\delta) \to \infty$ (measured local exponent $0.435 \to 0.394 \to 0.368 \to 0.353 \to 0.344 \to 0.339$ over $T = 50 \ldots 3200$), while the $O(\cdot)$ correction cancels in the between-regime difference.

N	measured wedge	$\frac{1}{3}\log N$	error
2	0.23083	0.23105	-2.2×10^{-4}
5	0.53622	0.53648	-2.6×10^{-4}
10	0.76729	0.76753	-2.4×10^{-4}
20	0.99837	0.99858	-2.1×10^{-4}
50	1.30385	1.30401	-1.6×10^{-4}

Table 9: Free-riding wedge in log-precision, averaged over the 18 (δ, σ^2) cells at each N. Predicted by Proposition 37 with no fitted quantity.

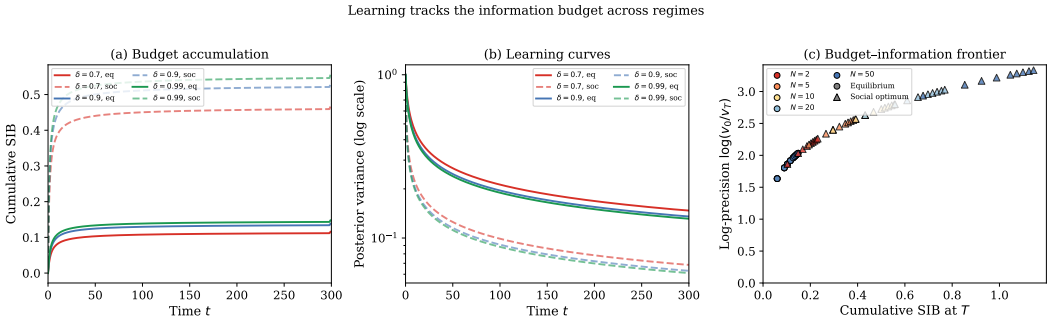


Figure 11: **Learning tracks the information budget across regimes** (illustrates Theorem 1 and Proposition 37; note this sweep has $\mathfrak{O}_t \approx 1$ throughout, so it tests conservation, not the strategic capacity weighting). (a) Cumulative SIB over time for equilibrium (solid) and social optimum (dashed). (b) Posterior variance stalls when SIB stops growing. (c) Budget–information frontier across 90 combinations. Equilibrium and social-optimum points occupy *distinct* loci ($F(2, 176) = 27.96$, $p = 2.9 \times 10^{-11}$); the zero-parameter prediction of Proposition 37.

Decentralised and centralised production do *not* convert at the same rate. The two regimes do not fall on a common curve: adding a regime indicator and its interaction to the log SIB regression is highly significant ($F(2, 176) = 27.96$, $p = 2.9 \times 10^{-11}$), with equilibrium slope $+2.23$ against social slope $+3.39$. This is the empirical face of Proposition 37: since SIB depends on the *ratio* path A_{t+1}/A_t while precision depends on the *level*, the two regimes occupy different loci, and pooling them is a category error rather than a demonstration of universality. What is universal is the growth exponent, not the conversion rate. Additionally, Figure 6 (Appendix L) is consistent with the extremal bound: "active but blind" policies with 199 switches and 1.0 bit of action entropy generate zero information (SIB $= 0.40$), while "rare but targeted" policies (19 switches, 0.81 bits) achieve SIB $= 14.25$—only kernel-changing variation counts.

S.2 Observability Determines Conversion

In a three-arm bandit ($\omega \in \{0, 1\}$, Gaussian noise $\sigma^2 = 1$), arm 0 is dominant but uninformative; arms 1,2 separate the environments. We compare pure exploitation, burn-in ($T_0 = 30, 60$), and full exploration (5,000 MC runs). Figure 5 (Appendix I) illustrates the frontier interpretation: burn-in concentrates budget where $\mathfrak{O}_t > 0$ and error drops rapidly; after burn-in, $\mathfrak{O}_t \approx 0$ and learning stalls. Figure 8 shows the tragedy as a frontier gap: equilibrium requires $3.1\times$ more SIB per unit of information than the social optimum ($R^2_{\text{eq}} = 0.93$).

S.3 Best-of-n Is Not a Strategic Feedback Channel

Gao et al. (2023) find that gold reward as a function of distance from the initial policy takes *different* functional forms for best-of-n and for RL, writing $d := \sqrt{\mathrm{KL}(\pi \| \pi_{\text{init}})}$ and fitting $R_{\text{bon}}(d) = d(\alpha_{\text{bon}} - \beta_{\text{bon}} d)$ against $R_{\text{RL}}(d) = d(\alpha_{\text{RL}} - \beta_{\text{RL}} \log d)$. They report the difference without accounting for it.

Our framework draws the relevant line. Best-of-n draws every sample from π_{init} and reweights; the observation kernel never depends on the policy's continuation, so by the definition in §2 it is *not* a strategic feedback channel, and Theorem 11 does not apply to it. Its degradation is the static selection effect of taking a maximum over noisy proxy scores. Reinforcement learning's continuation shapes the data it subsequently receives, so it is such a channel, and the collapse theorem does apply. Two distinct mechanisms should not share a functional form.

We state the limits of this plainly. It explains why the forms differ, not why they take the particular shapes they do; the classification is forced by our definitions, but so is very little else, and no numerical prediction follows. We also do not have an account of their two other regularities — that α_{RL} is nearly independent of proxy size while α_{bon} is not, and that a KL penalty raises proxy reward without moving the gold-reward frontier. The second is, if anything, uncomfortable for us: a central claim of this paper is that two policies spending the same budget can learn very different amounts, which makes the algorithm-independence of the gold–KL curve a fact we would have expected to come out otherwise.

T Extended Related Work

Sequential learning under endogenous data. Classical sequential analysis (Wald, 1947) and bandit theory (Lattimore and Szepesvári, 2020) assume exogenous data; we model the observation kernel as a function of the policy's continuation, and when that dependence is absent the conservation law degenerates to the ordinary mutual-information chain rule.

Strategic and performative learning. Performative prediction (Perdomo et al., 2020) and strategic classification (Hardt et al., 2016) characterize convergence to stable points; our collapse theorem provides the information-theoretic complement—convergence removes the continuation-mediated part of observational separation, leaving only whatever the environments reveal at a fixed continuation.

Information ratios, feedback loops, and information design. The information ratio (Russo and Roy, 2014; 2016) plays a structurally analogous role, splitting exploration into information and cost, but it is a regret-to-information ratio whereas capacity is a property of the induced experiment. RLHF (Christiano et al., 2017; Ouyang et al., 2022) creates the feedback loops the theory governs; training on model-generated data causes *model collapse* (Shumailov et al., 2024; Alemohammad et al., 2024), and recommender loops narrow diversity analogously (Chaney et al., 2018). Information design (Bergemann and Morris, 2019) and Bayesian persuasion (Kamenica and Gentzkow, 2011) ask how a principal controls the information environment; we invert this. Minimax bounds via Le Cam and Fano (Tsybakov, 2009; Yu, 1997) hold for fixed distributions; ours extends them to the endogenous setting. Estimating counterfactuals from logged data is off-policy evaluation (Precup et al., 2000; Dudík et al., 2011).

Closed-loop identification. That greedy control can leave a model accurate on-policy and ignorant off-policy is classical (Borkar and Varaiya, 1979; Kumar and Becker, 1982; Mandl, 1974; Lai and Wei, 1982) and we claim no priority for it. Those results establish that identifiability *can* fail and design excitation to prevent it; what $\mathfrak{O}_t$ adds is a scalar for how much identification a given unit of motion buys, so a policy can be scored rather than certified. We do not claim the law recovers them as corollaries. A fuller discussion, including persistent excitation, is in Appendix T.

Sequential learning under endogenous data. Classical sequential analysis (Wald, 1947) and bandit theory (Lattimore and Szepesvári, 2020) assume exogenous data; we model the observation kernel as a function of the policy's continuation, and the conservation law reduces to the standard MI chain rule in the exogenous case.

Strategic and performative learning. Performative prediction (Perdomo et al., 2020) and strategic classification (Hardt et al., 2016) characterize convergence to stable points; our collapse theorem provides the information-theoretic complement—convergence removes the continuation-mediated part of observational separation, leaving whatever the environments reveal at a fixed continuation. Follow-up work on performative and decision-dependent settings (Mendler-Dünner et al., 2020;

Miller et al., 2020; Izzo et al., 2021; Brown et al., 2022; Drusvyatskiy and Xiao, 2023) studies convergence and optimization, not the information-theoretic limits we establish. Strategic experimentation (Bolton and Harris, 1999; Keller et al., 2005) studies free-riding; our tragedy of the commons extends this to the SIB framework. Incentivized exploration (Sellke and Slivkins, 2023; Mansour et al., 2020; Immorlica et al., 2020; Kremer et al., 2014; Frazier et al., 2014) characterizes exploration costs within particular mechanism classes; we identify a universal constraint ($I \leq C\,\mathrm{SIB}_T$) applying to all policies simultaneously. The exploration mirage (§4.2) is a formal instance of Goodhart's law (Manheim and Garrabrant, 2018): standard exploration diagnostics diverge from the true objective.

Information-directed sampling and the information ratio. Russo and Roy (2014) introduced the information ratio as a complexity measure for bandits, with a comprehensive analysis for Thompson sampling (Russo and Roy, 2016); strategic capacity plays a structurally analogous role, decomposing exploration into an information component and a cost component. The key difference: the information ratio operates under exogenous data, while strategic capacity accounts for endogenous distortion of the observation channel. Extensions (Kirschner and Krause, 2018; Xu and Zeevi, 2023; Neu et al., 2024) all assume actions affect which data is collected but not the data-generating process itself—precisely the gap our framework fills.

RLHF and reward model collapse. RLHF (Christiano et al., 2017; Ouyang et al., 2022) creates exactly the strategic feedback loops our theory governs: the policy shapes which outputs evaluators see, which shapes the reward signal. Gao et al. (2023) document reward-model overoptimization. Section S.3 accounts for why their best-of-n and RL curves take different functional forms, and records two of their findings we cannot explain. Reward hacking (Skalse et al., 2022; Pan et al., 2024) and RLHF's fundamental limitations (Casper et al., 2023) receive a unified information-theoretic account: the reward signal is bounded by SIB, and policy convergence endogenously silences it.

Model collapse and feedback loops. Training on model-generated data causes *model collapse* (Shumailov et al., 2024; Alemohammad et al., 2024), and recommender feedback loops narrow content diversity analogously (Chaney et al., 2018). Our collapse theorem provides the mechanism with the distinction that collapse here arises from *strategic anticipation*, not merely data recycling.

Information design and Bayesian persuasion. Information design (Bergemann and Morris, 2019) and Bayesian persuasion (Kamenica and Gentzkow, 2011) study how a principal controls the information environment; our framework inverts the perspective, asking how much information the *learner* can extract when agents respond to the learning policy. The tragedy of the commons (§R.3) connects directly: SIB is a public good with private production cost, linking to dynamic mechanism design for information acquisition (Bergemann and Välimäki, 2002; 2019; Che and Hörner, 2015).

Information-theoretic limits. Minimax bounds via Le Cam and Fano (Tsybakov, 2009; Yu, 1997) hold for fixed distributions; our information-frontier sandwich extends them to the endogenous setting.

Closed-loop identification. That exploitation can leave a model accurate on-policy and ignorant off-policy is classical (Borkar and Varaiya, 1979; Kumar and Becker, 1982; Mandl, 1974; Lai and Wei, 1982), and the conservation law offers a related interpretation rather than a derivation: those results establish that identifiability can fail and design excitation to prevent it, while $\mathfrak{O}_t$ scores how much identification a unit of motion buys. The reward-biased MLE of Kumar and Becker (1982) induces channel motion of the kind the law rewards; we have not shown a formal mapping between the two, and we do not claim the law recovers these results as corollaries.

T.1 Preference Collapse in Iterative LLM Alignment

In an LLM preference loop ($K = 4$ styles, adaptive evaluator, $T = 200$, 2,000 MC runs), Blind and Targeted perturbation spend budgets in ratio 0.68 (8.86 vs. 13.09) with evaluator diversity equal to three decimals (0.476), yet differ $10\times$ in information gain (LLR 0.188 vs. 1.893) and $2.2\times$ in error

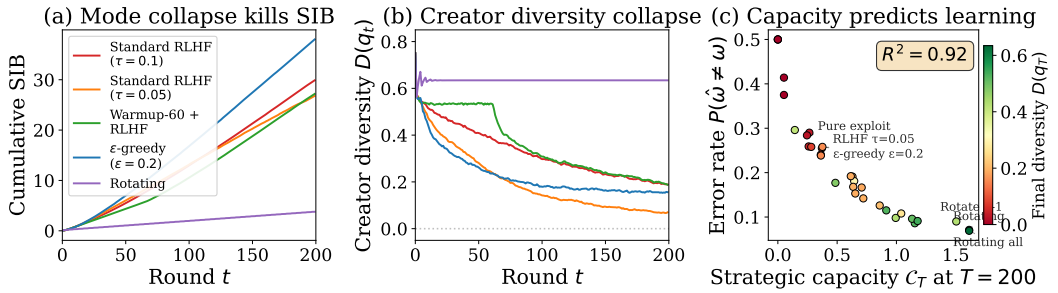


Figure 12: **Conservation law in RLHF.** (a) ε-greedy burns $10\times$ more SIB. (b) Creator diversity collapses under exploitation. (c) $\mathcal{C}_T$ predicts error ($R^2 = 0.92$); raw SIB does not.

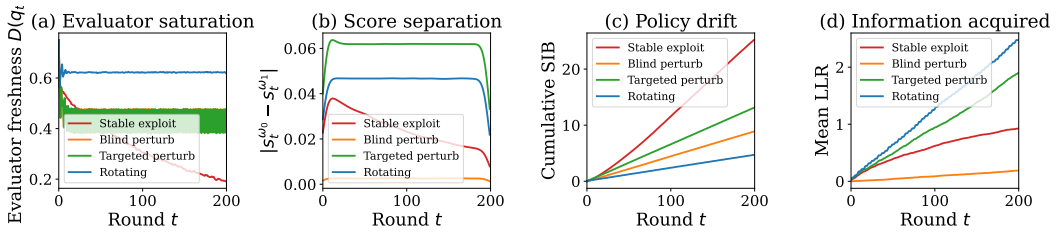


Figure 13: **Blind vs. targeted variation in LLM preference learning.** (a) Evaluator freshness. (b) Score separation: only Targeted and Rotating sustain nonzero separation. (c) Cumulative SIB. (d) Targeted acquires $10\times$ more information than Blind at comparable SIB.

(37.8% vs. 17.0%); separation energy accounts for it, $\widetilde{\mathcal{C}}_T = 0.356$ against 3.905, a factor of 11.0 (Figure 13).

T.2 Reward Model Collapse in RLHF

In a strategic RLHF model ($K = 4$ styles, creator adaptation $\lambda = 0.25$, $\alpha = 0.50$, $T = 200$), ε-greedy burns $10\times$ more SIB than a Rotating policy yet achieves $3.7\times$ worse identification error. Strategic capacity $\mathcal{C}_T$ predicts error across the 31 policies swept (29 of which clear the $\mathcal{C}_T > 0.01$ threshold used for the fit) with $R^2 = 0.92$; raw SIB is close to uninformative, $R^2 = 0.07$. Figure 12(a–c) shows the mechanism: aggressive exploitation collapses creator diversity, silencing the preference signal. Rotating achieves error 6.9% on SIB of only 3.8, the lowest among policies that explore at all, because it maintains $\mathfrak{D}_t > 0$ throughout. The single-style policies spend less still (SIB ≈ 0.02) but never move the channel and err at 37–50%.