WHEN IS A POLICY-GRADIENT BATCH RELIABLE? IDENTIFICATION LIMITS UNDER SEVERE UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

In reinforcement learning (RL), an agent learns a parameterized policy π_θ for sequential decision making by maximizing the expected cumulative return $J(\theta)$. Policy-gradient algorithms act on gradients estimated from finite trajectory batches, so a basic operational question is whether the resulting update points with high probability in the population-gradient direction. Existing work reduces gradient variance, adapts batch sizes, or controls policy improvement under explicit concentration assumptions, but these approaches do not answer a prior identification question: does the information conventionally retained about a one-trajectory gradient estimator determine its finite-batch wrong-direction probability? We show that, in general, it does not. Bounded atom-free REINFORCE problems can match any fixed jointly finitely generated menu of conventional one-rollout descriptors together with complete first- and second-order policy-gradient functions while having arbitrarily different strict directional risks. Under fixed Chernoff information and a bounded log-likelihood-ratio radius, we derive a sharp ambiguity law; a fixed pair of finite-horizon sequential REINFORCE Markov decision processes (MDPs) retains distinct directional large-deviation exponents despite exact matching of the stated policy-gradient and environment descriptors. We then give a non-oracle finite-sample certificate (without supplying the true population-gradient direction) using an independent confidence cone and held-out one-sided moment-generating-function information. Experiments show a regime-dependent picture: standard control settings can be close to Gaussian, whereas a controlled rare-critical family exhibits familywise-confirmed failures and a separately selected Grid2Op configuration survives an independent predeclared confirmation. The results position finite-batch policy-gradient reliability as an identification problem before it is a calibration problem.

1 INTRODUCTION

In reinforcement learning (RL), an agent learns a parameterized policy π_θ for sequential decision making by maximizing the expected cumulative return $J(\theta)$. Policy-gradient (PG) methods estimate $\nabla J(\theta)$ from sampled trajectories and update θ in the estimated ascent direction; from REINFORCE to modern actor-only and actor–critic variants, they optimize policies directly from trajectory data (Williams, 1992; Greensmith et al., 2004; Schulman et al., 2015; 2017). With a finite batch, however, the update direction is random rather than the population gradient itself. Reliability is therefore usually treated as an estimation-quality problem: reduce variance, improve the control variate, increase the batch, or appeal to an asymptotic approximation. This leaves a prior question unresolved: *what information is actually sufficient to know that a finite batch is unlikely to point the wrong way?*

The issue is easiest to see in a rare-event control example. Imagine that most trajectories produce moderate, mutually consistent gradient contributions, while a small fraction contain a large adverse outcome; power-flow collapse, safety violation, or other high-impact transition. Such constraints are central in safe/constrained reinforcement learning (RL), and power-network control is a concrete sequential domain in which avoiding severe failures matters (García & Fernández, 2015; Achiam et al., 2017; Marot et al., 2021). A batch that misses or over-represents those trajectories can point opposite to the population gradient even when its mean scale and covariance look ordinary. Large

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.

policy-gradient variance has long motivated baselines and control variates (Greensmith et al., 2004; Schulman et al., 2015), and pronounced heavy-tailed actor gradients have been measured in proximal policy optimization (PPO) (Garg et al., 2021). Conversely, a large tail discrepancy does not imply that every practical policy-gradient problem is unsafe; whether it matters depends on the regime and on the information used to certify the update.

Existing approaches address adjacent pieces through variance reduction, safe/adaptive batching, and moment-based distributional robustness (Greensmith et al., 2004; Kallus & Uehara, 2020; Huang & Jiang, 2020; Papini et al., 2017; 2022; Bertsimas & Popescu, 2005; Vandenberghe et al., 2007; Delage & Ye, 2010); Section 2 reviews them. None resolves our prior identification question: if two REINFORCE problems agree on the retained summaries, must they agree on finite-batch wrong-direction risk?

Our answer is no, and the consequence is operational: if the same retained summaries are compatible with two different wrong-direction risks, then no batch-sizing or reliability rule that sees only those summaries can know which guarantee it is making. The strength of the negative result is therefore what matters. We do not stop at the familiar statement that mean and variance fail to determine a tail. We construct bounded atom-free REINFORCE problems that match arbitrary fixed finite menus of conventional one-rollout descriptors and match the complete first- and second-order policy-gradient functions, yet have very different finite-batch directional risk. We then ask how much ambiguity remains after conditioning on genuine likelihood-level information and obtain a sharp law. A fixed matched pair is lifted to genuinely sequential finite-horizon MDPs, where the directional large-deviation exponents remain different. Finally, rather than concluding that reliability is unknowable, we identify additional one-sided information that is sufficient for a finite-sample certificate.

Empirically, we deliberately separate benign controls, a controlled significant-rare-event mechanism, and external discovery followed by fresh Grid2Op confirmation. This supports a regime-dependent (not universal-failure) claim: *conventional second-order information cannot certify finite-batch directional reliability in general.*

Contributions. We make five contributions. (i) We formulate finite-batch directional reliability as an identification problem for the gradient statistic used by the algorithm. (ii) We prove broad exact non-identification and a sharp likelihood-conditioned ambiguity law. (iii) We realize the separation in a fixed family of sequential REINFORCE MDPs and show robustness beyond the canonical scalar/atomic witness. (iv) We give a non-oracle certificate that combines an independent confidence cone with held-out one-sided moment-generating-function (MGF) information. (v) We test the resulting picture prospectively across benign, controlled rare-event, and external sequential settings, reporting negative and inconclusive panels alongside failures.

2 Related work and the unresolved gap

Policy-gradient estimation and variance reduction. REINFORCE provides unbiased score-function gradient estimators from sampled trajectories (Williams, 1992). Their variance motivated classical baselines and actor–critic control variates (Greensmith et al., 2004), generalized advantage estimation (Schulman et al., 2015), stochastic variance-reduced policy gradients (Papini et al., 2018), and statistically efficient on/off-policy estimators (Kallus & Uehara, 2020; Huang & Jiang, 2020). These methods target variance, mean-square error, or asymptotic efficiency. Our target is different: the probability that a fixed finite batch lies in the population downhill half-space. Two estimators can agree on first- and second-order information while differing on that directional tail.

Safe updates and adaptive batch sizing. Adaptive-batch stochastic optimization chooses sample size using estimated gradient quality (Byrd et al., 2012; Bollapragada et al., 2018). Safe RL more broadly asks for satisfactory performance or constraint satisfaction during learning or deployment (García & Fernández, 2015); constrained policy optimization gives near-constraint-satisfaction guarantees for policy updates (Achiam et al., 2017). Closer to our batch question, Papini et al. (2017) jointly select step and batch size to obtain high-probability policy improvement, and Papini et al. (2022) extend this program using smoothing-policy assumptions and explicit gradient-variance bounds. These guarantees are constructive under stated smoothness, boundedness, and concentration structure; the latter work also notes that worst-case problem constants can be large. Our negative

theory asks a logically prior question: whether commonly retained summaries identify the exact finite-batch directional risk at all. Our positive certificate then supplies one valid information interface when additional one-sided information is available.

Non-Gaussian gradients, directions, and large deviations. Empirical work shows that PPO actor gradients can be strongly heavy-tailed and that clipping partly mitigates this behavior (Garg et al., 2021). Directional stochastic-optimization methods such as sign-based gradients emphasize sign reliability rather than magnitude (Bernstein et al., 2018), but impose noise conditions tailored to their own convergence guarantees. Large-deviation theory has also been brought to policy-gradient iterates under structural assumptions such as Polyak–Lojasiewicz geometry (Jongeneel et al., 2024). We use classical Chernoff and large-deviation tools for a different purpose: to prove that two matched REINFORCE problems can retain different finite-batch directional tails and different directional exponents.

Moment information, ambiguity sets, and partial identification. Sharp probability bounds from moments are classical and can be computed by convex or semidefinite optimization in important cases (Bertsimas & Popescu, 2005; Vandenberghe et al., 2007); distributionally robust optimization similarly optimizes decisions over distributions consistent with moment information (Delage & Ye, 2010). More generally, partial-identification analysis asks what set of target values remains compatible with incomplete information or assumptions (Manski, 2003). Statistical comparison of experiments and Chernoff information characterize distinguishability at the level of full experiments (Blackwell, 1953; Chernoff, 1952). We make no new general moment-problem, partial-identification, or information-theory claim. The gap we address is policy-gradient specific: broad exact matching inside bounded REINFORCE problems, a sharp ambiguity law after fixing likelihood-level information, persistence in a fixed sequential MDP pair, and a finite-sample certificate tied to the statistic actually averaged by the algorithm.

3 Methodology and Theory

We first formalize the operational event, then ask what conventional information identifies it. The negative results progress from a broad exact-matching theorem to a sharp likelihood-conditioned law and a sequential realization; the final subsection gives a constructive finite-sample certificate. Full proofs and auxiliary constructions are in the appendix.

3.1 Problem formulation and the identification question

For a fixed environment/MDP instance M, let θ denote the policy parameter, let $J(\theta)$ be the expected return of the differentiable stochastic policy π_θ, and let

$$g(\theta) = \nabla J(\theta), \qquad \widehat{g}_1(\theta), \dots, \widehat{g}_n(\theta) \overset{\text{iid}}{\sim} P_\theta, \qquad \bar{g}_n = \frac{1}{n}\sum_{r=1}^{n} \widehat{g}_r(\theta),$$

where $\widehat{g}_r$ is a one-trajectory REINFORCE-type gradient estimator with $\mathbb{E}\widehat{g}_r = g(\theta)$. We study the *strict finite-batch wrong-direction risk*

$$\mathcal{R}_n^{\text{str}}(M, \theta) := \mathbb{P}(\langle \bar{g}_n, g(\theta) \rangle < 0)\,. \tag{1}$$

This is a local first-order reliability quantity: a strict reversal places the sampled update in the population downhill half-space. Under the usual differentiability/smoothness conditions, for all sufficiently small positive step sizes such an update decreases J to first order. Equation (1) does not assert global convergence or characterize the return after an arbitrary finite step.

For a retained descriptor map D, we say that D *identifies* $\mathcal{R}_n^{\text{str}}$ on a model class if $D(M) = D(M')$ implies $\mathcal{R}_n^{\text{str}}(M, \theta) = \mathcal{R}_n^{\text{str}}(M', \theta)$ for every compatible pair. Non-identification means that the same reported information can correspond to different operational risks.

Hard instances are indexed by an endpoint $i \in \{A, B\}$ and a hidden gradient sign $s \in \{-1, +1\}$. Let P_i^s be the law of one directional-gradient observation. An *environment-level* information quantity at endpoint i compares the trajectory laws under $s = +1$ and $s = -1$. A *statistic-level* quantity instead concerns the induced law of the gradient observation used by the algorithm. Matching across A and B

means equality of the stated quantity; it never means equality of the full experiments unless explicitly stated. We write $\mathcal{R}^{\mathrm{str}}_{i,n}$ (and, in the sequential construction, $\mathcal{R}^{\mathrm{str}}_{i,H,n}$) for the corresponding risk at endpoint i when the policy/reference parameter is understood. In the sign-symmetric LR-aligned constructions below,

$$L := \log \frac{dP_i^+}{dP_i^-}$$

denotes the one-sample log-likelihood ratio (LR), and $S_n := \sum_{r=1}^n L_r$.

Our target is the strict event in (1). We use the fair-tie functional $\mathcal{R}_n^{1/2} = \mathbb{P}(S_n < 0) + \frac{1}{2}\mathbb{P}(S_n = 0)$ only to invoke classical binary-testing identities. Ties vanish in the atom-free (non-atomic) class of Theorem 2. The finite-support sequential construction instead analyzes literal strict risk: zero-gradient atoms are not counted, and its finite-n lower bound uses a strictly negative window.

The first question is qualitative. Suppose a researcher reports any fixed finite collection of ordinary one-rollout descriptors. Can that collection determine strict finite-batch risk for every compatible PG problem? The definition below excludes only pathological scalar encodings of an entire probability law.

Definition 1 (Finitely generated one-rollout descriptor menu)**.** *In the sign-symmetric Bhattacharyya-tilted LR representation used below, the tilted measure has density proportional to $\sqrt{dP_+\, dP_-}$; let ν denote the law of the positive LR magnitude $|L|$ under that tilted measure. A descriptor D is* finitely generated *if, after fixing policy-side constants, there are finitely many continuous functions $h_1, \dots, h_d$ and a deterministic map Φ such that*

$$D(\nu) = \Phi\left(\int h_1\, d\nu, \dots, \int h_d\, d\nu\right).$$

A finite menu is jointly finitely generated if one finite list $h_1, \dots, h_d$ generates all of its coordinates. This class contains any fixed finite set of bounded reward/gradient moments and any fixed finite collection of total-variation (TV), Kullback–Leibler (KL), Hellinger, or Rényi-affinity coordinates in the PG realization below, while deliberately excluding a scalar that encodes the complete law.

The theorem fixes the batch size, the desired separation factor, and the descriptor menu. It allows the two compatible PG instances to depend on those choices. This is the correct quantifier order for failure of universal identification. In the Bernoulli-logit constructions below, $\eta \in \mathbb{R}$ denotes the scalar policy logit,

$$\sigma(\eta) := \frac{1}{1 + e^{-\eta}},$$

and H denotes the finite decision horizon.

Theorem 1 (No fixed finitely generated menu identifies strict risk)**.** *Fix any even $n \geq 2$, $K > 1$, and any jointly finitely generated menu from Definition 1. There exist bounded atom-free $H = 2$ REINFORCE endpoint pairs that match the entire menu exactly, match Chernoff/Hellinger and any prescribed finite Rényi menu, and have common policy score/Fisher geometry and identical complete functions $J_s(\eta)$, $\nabla J_s(\eta)$, and $\mathbb{E}\widehat{g}(\eta)^2$, yet*

$$\frac{\mathcal{R}^{\mathrm{str}}_{B,n}}{\mathcal{R}^{\mathrm{str}}_{A,n}} > K.$$

Appendix B constructs continuous likelihood-ratio (LR) microcells that preserve every requested integral coordinate while moving probability across the strict batch-sign boundary. Thus the theorem is stronger than the elementary statement that two moments do not determine a tail: the endpoints can match a prescribed richer descriptor menu and their complete PG mean and second-moment functions. It does *not* provide one pair that works for all n; Section 3.3 supplies that separate result.

3.2 Sharp ambiguity under fixed likelihood information

The preceding theorem permits an arbitrary finite menu. We next impose two strong controls on the complete hidden-sign testing problem and quantify exactly how much ambiguity remains. For laws P_+, P_- with densities p_+, p_-, their Chernoff information is

$$C(P_+, P_-) = -\log \inf_{0 \leq t \leq 1} \int p_+^t p_-^{1-t}.$$

This is an environment-level optimal-testing exponent, not an estimator variance. Let $L = \log(dP_+/dP_-)$. Let R denote the Bhattacharyya-tilted law of L, obtained from the measure proportional to $\sqrt{dP_+\, dP_-}$. An *LR-radius bound* B means $|L| \le B$ under R; it does not assert that the essential radius equals B.

Fix $C_0 > 0$ and write

$$\beta = e^{-C_0}, \quad m = e^{C_0}, \quad a(C_0) = 2\,\mathrm{arcosh}(m).$$

Let $\mathcal{A}_{\mathrm{ac}}(C_0, B)$ be the symmetric absolutely continuous laws R supported on $[-B, B]$ and satisfying

$$\mathbb{E}_R \cosh(L/2) = m. \tag{2}$$

For this sign-symmetric class, (2) fixes Chernoff information at C_0. For $R \in \mathcal{A}_{\mathrm{ac}}(C_0, B)$, let $\mathcal{R}_n^{\mathrm{str}}(R)$ denote the strict sign risk of the corresponding hidden-sign pair. The ratio

$$\Gamma_n^{\mathrm{str}}(C_0, B) := \frac{\sup_{R \in \mathcal{A}_{\mathrm{ac}}(C_0,B)} \mathcal{R}_n^{\mathrm{str}}(R)}{\inf_{R \in \mathcal{A}_{\mathrm{ac}}(C_0,B)} \mathcal{R}_n^{\mathrm{str}}(R)}.$$

Theorem 2 (Sharp strict-risk conditioning phase law)**.** *For every fixed $C_0 > 0$, every $n \ge 2$, and every $B > a(C_0)$,*

$$\boxed{\Gamma_n^{\mathrm{str}}(C_0, B) = \Theta_{C_0}\left(\min\{\sqrt{n}, e^{B/4}\}\right).} \tag{3}$$

Here $\Theta_{C_0}(\cdot)$ allows multiplicative constants depending on C_0; these constants can be chosen uniformly for C_0 in compact subsets of $(0, \infty)$.

The theorem answers a finite-n extremal question. With C_0 and an LR-radius upper bound held fixed, the largest compatible risk divided by the smallest compatible risk grows like $\sqrt{n}$ until the LR constraint forces saturation at order $e^{B/4}$. In particular, the $\sqrt{n}$ branch is not indefinite growth at fixed B. This result does not claim that practitioners know C_0 or B; they are conditioning variables used to locate the identification boundary.

The proof starts from the exact tilted identity for the optimal equal-prior Bayes error e_n^* based on n observations,

$$e_n^* = \frac{1}{2}\beta^n \mathbb{E}_R \exp\left(-\frac{1}{2}\left|\sum_{r=1}^{n} L_r\right|\right), \tag{4}$$

then combines Fourier control, Kolmogorov–Rogozin anti-concentration, matching rare-spike and anti-concentrated constructions, and atom-free exact-C_0 approximation. Appendix C contains the proof. The tools are classical; the contribution is the extremal envelope under joint (C_0, B) conditioning.

The phase law is also attainable inside PG geometry rather than only as an abstract testing class.

Corollary 1 (Matched-PG saturation)**.** *For every fixed $C_0 > 0$, there is a bounded $H = 2$ Bernoulli-logit REINFORCE class with common reference-policy Fisher $1/4$, common directional gradient mean and variance, and common complete functions $J_s(\eta)$, $\nabla J_s(\eta)$, and $\mathbb{E}\widehat{g}(\eta)^2$, such that whenever $B \ge B_\star(C_0, n) = O_{C_0}(\log n)$ its strict-risk envelope is $\Theta_{C_0}(\sqrt{n})$.*

This corollary realizes the unsaturated $\sqrt{n}$ branch inside an LR-aligned PG construction while fixing the listed first- and second-order PG quantities. It does not assert that every PG realization attains both branches of Theorem 2. The reference-policy gradient-sample bound is $1/2$ and the uniform-in-policy bound is one; Appendix D gives the constants and construction.

3.3 Sequential REINFORCE persistence

The finite-menu theorem may choose its witness after n is specified. We now hold the endpoint laws fixed while the batch size grows. The construction uses the remaining distributional freedom after four one-stage coordinates have been matched. Those coordinates determine the objective, its gradient, second-order PG geometry, reward moments, and one environmental testing quantity. Higher features of the transition law remain free, and they change the cumulant-generating function (CGF) of the directional gradient. Because a Cramér exponent depends on the full CGF rather than only its first two derivatives at zero, the exponent can vary without a variance change.

For $T = H - 1$, let $p = \sigma(\eta) \in (0, 1)$ and initialize $X_0 = 0$. Each stage samples $A_t \sim \text{Bernoulli}(p)$, sets $U_t = 2A_t - 1$, draws independent transition noise $W_t \sim P_{i,s}$, and evolves

$$X_t = X_{t-1} + \frac{U_t W_t}{T}, \qquad R_H = X_T, \qquad \widehat{g}_H = R_H \sum_{t=1}^{T} (A_t - p).$$

The reward is the same deterministic terminal-position map at both endpoints; only the transition-noise laws differ. Matching normalization, $\mathbb{E}W = \mu$, $\mathbb{E}W^2 = m_2$, and the Bhattacharyya coefficient matches the complete $J_s^{(H)}(\eta)$, $\nabla J_s^{(H)}(\eta)$, reward second moment, gradient second moment/covariance, and score/entropy/Fisher functions for every finite η. The trajectory reveals every W_t, so its LR is a T-term sum. If C_0 and B_0 denote the one-stage Chernoff information and LR-radius bound, respectively, then the corresponding trajectory quantities scale as $C_H = TC_0$ and $B_H = TB_0$. Appendix F gives the exact rational support. Table 1 summarizes exactly what the sequential endpoints match and what remains free.

The next theorem uses the same two one-stage endpoint laws for every horizon. For each chosen H, those laws induce an H-horizon MDP pair that is fixed independently of batch size n; the statement is not that one literal finite-horizon MDP has several horizons. For endpoint $i \in \{A, B\}$, write

$$I_i^{(H)} := -\lim_{n\to\infty} \frac{1}{n} \log \mathcal{R}_{i,H,n}^{\text{str}},$$

when the limit exists, and define the exponent gap $\Delta_H := I_B^{(H)} - I_A^{(H)}$.

Theorem 3 (One fixed genuinely sequential pair separates at every finite horizon)**.** *There is an explicit positive rational continuum of bounded one-stage transition laws with five paired support cells such that, for every finite $H \geq 2$, all quantities listed above are exactly matched while the strict wrong-direction Cramér exponent I_H varies strictly along the continuum. The same two one-stage endpoints A, B therefore induce, for every H, an H-horizon pair fixed independently of batch size n and satisfying*

$$I_A^{(H)} < I_B^{(H)}, \qquad \lim_{n\to\infty} \frac{1}{n} \log \frac{\mathcal{R}_{A,H,n}^{\text{str}}}{\mathcal{R}_{B,H,n}^{\text{str}}} = \Delta_H > 0.$$

The nonconstancy persists on a support/skew neighborhood for each fixed H. In the normalized lift,

$$\Delta_H = \frac{\kappa_*}{H-1} + o(H^{-1}), \qquad \kappa_* = 0.07128879336\ldots,$$

so the gap is positive for every finite horizon but is not horizon-uniform.

Corollary 2 (Finite-batch consequence for the same fixed pair)**.** *For fixed finite $H \geq 2$ and any $0 < \gamma_H < \Delta_H$, there are computable $c_H > 0$ and $n_{0,H} < \infty$ such that*

$$\frac{\mathcal{R}_{A,H,n}^{\text{str}}}{\mathcal{R}_{B,H,n}^{\text{str}}} \geq c_H e^{\gamma_H n}, \qquad n \geq n_{0,H}. \tag{5}$$

For any fixed H, the theorem identifies the asymptotic exponent and the corollary converts it into an eventual finite-n inequality. Only $c_H, n_{0,H}, \gamma_H$ depend on the requested bound; the endpoint

Quantity	A	B	Consequence
$J(\eta)$ and $\nabla J(\eta)$	matched	matched	same objective and population direction
Gradient second moment/covariance	matched	matched	same Gaussian/variance inputs
Score, entropy, and Fisher functions	matched	matched	same stated policy geometry
Reward first two moments; C_H, B_H; pointwise LR map	matched	matched	same listed environment descriptors
Higher transition/gradient law	differs	differs	different directional CGF and Cramér exponent

Table 1: **What the sequential endpoints match and what remains free.** The theorem matches the listed functions and coordinates exactly, not the full trajectory experiment or LR spectrum.

pair for that horizon does not depend on n. Appendix F gives the complete proof of Theorem 3 and Corollary 2, including the strictly negative tilted-window argument and the no-tie finite-n bound. Common-batch exponential risk separation does not imply exponential target-confidence batch complexity or end-to-end training complexity.

Why the construction is genuinely sequential. For $H \geq 3$, the terminal return interacts with every score term; deleting a stage changes the score, Fisher, return, LR, and gradient laws, and the estimator contains cross-stage products. For each fixed H, the exponent ordering persists on a horizon-dependent open policy neighborhood and under the variance-optimal deterministic time-indexed scalar baseline. The proof does not cover arbitrary state-dependent learned baselines. In this family a wrong-direction ordinary PG update decreases the objective for every positive step size, but this is a one-update statement rather than a training-dynamics guarantee.

3.3.1 Structural robustness beyond the canonical witness

The preceding witness is reflected, atomic, and scalar. The next theorem asks whether the separation disappears once those three conveniences are removed. It preserves exact endpoint matching while perturbing the transition law and policy geometry; persistence follows from strict derivative margins on compact parameter sets.

Theorem 4 (De-specialized all-horizon realization)**.** *A single sufficiently small perturbation removes exact reflection symmetry and atomic transition noise, and yields a two-parameter state/history-dependent policy whose Fisher matrix is full rank at the reference parameter $\theta = 0$. For every finite $H \geq 2$, the perturbed endpoint family preserves the stated exact matching coordinates for all finite policy parameters and retains a nonconstant strict-risk exponent at the reference parameter.*

Appendix G gives the complete construction and proof, including the exact rank and uniform stability certificates. The result establishes structural robustness, or non-isolation, of the separation. It does not establish prevalence in naturally occurring deep RL, match the complete LR spectrum, or cover arbitrary control variates and arbitrary MDP perturbations.

3.4 Constructive finite-sample certification

The negative results identify insufficient summaries, but they do not make reliability unknowable. The required information must be tied to the one-sided law of the statistic actually averaged. Let τ denote a trajectory under hidden-sign laws P_+, P_- and let $G = T(\tau)$ be a scalar statistic with induced laws Q_+, Q_-. Write $C(\tau)$ and $C(G)$ for the Chernoff informations of the trajectory experiment and the induced statistic, respectively.

Proposition 1 (Estimator-information hierarchy)**.** *For any measurable statistic $G = T(\tau)$, let $e_n^{\mathrm{mean}}(G)$ be the equal-prior error of the fixed sign-of-empirical-mean rule and $e_n^*(G)$ the optimal equal-prior error based on G^n. Then $e_n^{\mathrm{mean}}(G) \geq e_n^*(G)$ for every n and*

$$\limsup_{n\to\infty} -\frac{1}{n} \log e_n^{\mathrm{mean}}(G) \leq C(G) \leq C(\tau). \tag{6}$$

If the empirical-mean exponent exists, denote it by I_{mean}. In the reflected finite-support class of Appendix I, $I_{\mathrm{mean}} = C(G)$ iff $\log(dQ_+/dQ_-)(g) = \kappa g$ for some $\kappa > 0$ on the support. If G is sufficient for the trajectory binary experiment, then $C(G) = C(X)$.

Appendix H proves the data-processing and finite-n statements, while Appendix I proves the reflected-class equality characterization. The right inequality is classical data processing and the finite-n inequality is optimality of the LR test among decisions based on G^n. The proposition separates two losses: $\tau \mapsto G$ may discard information, and empirical averaging may fail to exploit all information retained in G. Consequently, environment-level Chernoff information alone cannot generally stand in for the directional tail of a particular PG estimator. The operational obstacle is that the event itself uses the unknown mean $\mu := \mathbb{E}\widehat{g} = \nabla J$. The certificate uses three independent roles for data:

1. a direction sample constructs a confidence ball, hence a cone containing $\mu/\|\mu\|$;
2. a validation sample upper-bounds one MGF value for the worst direction in that cone;
3. a future independent batch receives the resulting risk guarantee.

Independence prevents the direction or MGF parameter from being selected using the same observations that certify it.

Theorem 5 (Non-oracle future-batch certificate)**.** *Let $X \in \mathbb{R}^d$ be one i.i.d. trajectory-gradient vector with nonzero mean μ and known deterministic bound $\|X\| \le B$. Split pilot data into independent parts S_1, S_2. From S_1, construct a random Euclidean ball $C_1 = \{v : \|v - c\| \le r\}$ satisfying $\mathbb{P}(\mu \in C_1) \ge 1 - \delta_1$. If $\|c\| > r$, define $u_0 = c/\|c\|$, $\theta = \arcsin(r/\|c\|)$ and the direction cone*

$$\mathcal{U}_1 = \{u : \|u\| = 1,\ \angle(u, u_0) \le \theta\};$$

otherwise take $\mathcal{U}_1$ to be the whole unit sphere. For $x \in \mathbb{R}^d$ define the robust directional statistic

$$Z_1(x) = \inf_{u \in \mathcal{U}_1} \langle x, u \rangle, \qquad |Z_1(x)| \le B. \tag{7}$$

Let $\widehat{\lambda} \le 0$ be any S_1-measurable choice. Using only S_2, form a $(1-\delta_2)$ upper confidence bound U on $M_{\widehat{\lambda}} := \mathbb{E}[e^{\widehat{\lambda} Z_1(X)} \mid S_1]$ and set $I_{\mathrm{cert}} = [-\log U]_+$. Then, with probability at least $1-\delta_1-\delta_2$ over the pilot data, simultaneously for every independent future batch size $n \ge 1$, where $\bar{X}_n := n^{-1} \sum_{i=1}^n X_i$,

$$\mathbb{P}\big(\langle \bar{X}_n, \mu \rangle \le 0 \mid S_1, S_2\big) \le e^{-n I_{\mathrm{cert}}}. \tag{8}$$

Thus $n \ge \lceil \log(1/\alpha)/I_{\mathrm{cert}} \rceil$ is a valid sufficient future batch size whenever $I_{\mathrm{cert}} > 0$.

Because

$$\{\langle \bar{X}_n, \mu \rangle < 0\} \subseteq \{\langle \bar{X}_n, \mu \rangle \le 0\},$$

the same bound is also a valid upper bound on the strict wrong-direction risk. The proof propagates direction uncertainty rather than supplying the oracle direction. On $\{\mu \in C_1\}$, the true unit direction lies in $\mathcal{U}_1$, so $Z_1(x) \le \langle x, \mu/\|\mu\| \rangle$ pointwise. Wrong direction therefore implies a lower-tail event for Z_1, to which Chernoff's bound applies. We instantiate C_1 with the dimension-free vector empirical-Bernstein ball of Martinez-Taboada & Ramdas (2026). Because $\widehat{\lambda}$ is chosen before S_2 is observed, validation needs coverage at one MGF point rather than over a post-hoc search. Appendix H gives the closed forms and proof. The guarantee is simultaneous over all independent future batch sizes, conditional on a successful pilot. It becomes vacuous when the direction cone is too broad or the MGF upper bound is at least one. It requires a known valid deterministic bound B; replacing B by the observed maximum would invalidate the guarantee.

Valid alternative when hard support is unrealistic. Under the known all-direction sub-exponential condition $\log \mathbb{E} e^{\lambda \langle v, X - \mu \rangle} \le \nu^2 \lambda^2 / [2(1 - b|\lambda|)]$, an independent pilot lower bound $L \le \|\mu\|$ gives $\mathbb{P}(\langle \bar{X}_n, \mu \rangle \le 0 \mid \text{pilot}) \le \exp\{-nL^2/[2(\nu^2 + bL)]\}$ for every future n (Appendix H.5). This removes hard support, not tail assumptions: (ν, b) must be known or externally justified. Appendix H also gives an anytime-valid confidence-set rule for a realized update direction; we make no stopping-rule priority claim.

4 Experiments and results

We evaluate two questions separately: whether the identification gap is visible in ordinary control regimes, and whether it becomes operational under rare/high-impact uncertainty. All headline risks are estimated from batches independent of reference-gradient construction and calibration. Confirmatory threshold tests use exact one-sided Clopper–Pearson (CP) bounds with the predeclared multiplicity correction for the corresponding family; descriptive Gaussian comparisons are never promoted into confirmatory tests. Table 2 reports benign, failure, and inconclusive panels rather than only favorable cases, and discovery is kept separate from confirmation.

The prospective MountainCar/Acrobot follow-up broadens empirical scope without producing a new failure: only one estimator case passes the reference gate, the stable Acrobot baseline case has full-rank $(14/14)$ covariance and close Gaussian/direct risks, and the analytical-envelope certificate abstains throughout. The learned state-dependent baseline is empirical scope only.

In the controlled significant-rare-event (SRE) family, with rare-event probability p, ordinary-event scale $b_{\mathrm{ord}} > 0$, Bernoulli policy probability q, and severity $\rho > 0$, the rare adverse magnitude is $L_{\mathrm{rare}} = \rho(1-p)b_{\mathrm{ord}}/p$ while the population directional mean is proportional to $q(1-q)(1-$

Setting	Independent units	Result / inferential role
PPO CartPole	5 seeds, 20 frozen checkpoints	226-D gradients. At $n = 16$, median empirical/Gaussian discrepancy is $2.78\times$, but Gaussian screening makes no conditional empirical false reliability declaration at $\alpha \in \{.10, .05, .01\}$: benign control.
MountainCar/Acrobot follow-up	6 frozen environment–policy cases, 2 estimators	Prospective. One of 12 estimator cases passes the reference gate: full-rank 14-D Acrobot with a frozen state baseline. Direct risks at $n = 8, 16, 32$: $16.9\%, 9.62\%, 3.74\%$; Gaussian: $18.5\%, 10.24\%, 3.65\%$. Others inconclusive.
Exact CliffWalking	120 prespecified cases	No familywise-confirmed threshold failure at $\alpha \in \{.05, .01\}$: benign exact-reference control.
Rare-critical SRE	140 prespecified cases	Seven familywise-confirmed failures at $\alpha = .01$ and one at .05; maximum screen/Gaussian discrepancy $82.7\times$. Controlled mechanism, not prevalence evidence.
Grid2Op V1 $\rightarrow$ V2	V1 discovery; V2 100k fresh batches	V2: $1{,}258/100{,}000$ wrong at $n = 16$: 1.258%, one-sided 95% CP lower $1.201\% > 1\%$. Independent Gaussian calibration 0.998% is secondary descriptive; only direct risk is confirmatory.
MuJoCo / broad V1 controls	45 MuJoCo audits; 125/152 V1 cases complete	MuJoCo has no reference-stable checkpoint; Highway is partial (21/48 complete); Finance/Safety mostly fail reference gates. Inconclusive/missing, not null evidence.

Table 2: **Empirical landscape.** Independent-sample headline results; full protocols, intervals, multiplicity corrections, and inconclusive panels are in Appendix Q.

$p)b_{\mathrm{ord}}(1-\rho)$. Thus the confirmed $\rho \in \{.02, .05, .10\}$ regimes retain most ordinary-event signal rather than relying on near-cancellation. Across 140 prespecified cases, seven case–alpha outcomes are familywise-confirmed at $\alpha = .01$ and one at $\alpha = .05$; the strongest $\alpha = .01$ case has Gaussian risk 3.67×10^{-4}, confirmation risk 0.0306, and familywise lower bound 0.0291.

Certificate practicality. The constructive theorem is valid only when its direction cone and one-sided MGF bound are informative. Across 540 estimator cases from 180 exactly enumerable MDPs, a 10k/10k direction–MGF split produces a positive certificate in 69.6% of cases; among positives, the median certified/exact Cramér-exponent ratio is 0.186 and the median sufficient batch size for 1% future risk is 200. Increasing only the direction split to 20k raises positivity to 76.5% and the median ratio to 0.227. The dominant practical limitation is the deterministic envelope: on a prespecified 180-case sensitivity panel, inflating the tight bound from $B_\star$ to $2B_\star$ reduces positive certificates from 59.4% to 18.9%, and $5B_\star$ leaves only 1.1% positive. The prospective MountainCar/Acrobot follow-up is even harsher: all predeclared analytical-envelope certificates abstain because the direction cone is too broad. We therefore report validity and usefulness separately rather than treating abstention as failure of coverage. Grid2Op V2 is the only external threshold confirmation; V1 is discovery evidence only. The selected configuration was frozen and retested with no sample reuse; its 20k-episode reference is highly stable, with all eight shard means of one sign and $|\widehat{g}|/\mathrm{SE} = 82.94$. The direct risk and independent Gaussian calibration fall on opposite sides of the prespecified 1% threshold (Table 2); only the direct-risk statement is confirmatory. Because `l2rpn_case14_sandbox` uses `test=True`, we interpret V2 as an external sequential-mechanism replication rather than production-grid or prevalence evidence (Marot et al., 2021; Grid2Op developers, 2026).

5 Conclusion and scope

Finite-batch policy-gradient reliability is an identification problem before it is a calibration problem. We show that broad conventional summaries can leave directional risk unidentified, characterize the remaining ambiguity sharply under likelihood-level conditioning, realize persistent separation in matched sequential MDPs, and identify additional one-sided information sufficient for certification. The scope is deliberately limited: the finite-menu theorem applies to jointly finitely generated descriptor menus, the sharp law to its stated sign-symmetric atom-free class, and the sequential theorem to the listed PG and LR coordinates without a horizon-uniform exponent-gap claim. The results do not constitute a theorem for arbitrary deep policies or learned control variates; the latter are empirical scope only, and certification assumes fixed-policy i.i.d. sampling with a valid deterministic envelope or externally justified tail parameters. Empirically, benign and severe regimes coexist: the SRE benchmark yields familywise-confirmed failures and one Grid2Op configuration survives fresh V2 confirmation, alongside incomplete or reference-inconclusive controls. Future work should identify observable regime variables that predict when second-order information is sufficient and when additional tail information is required.

REPRODUCIBILITY STATEMENT

The supplementary material contains complete proofs, theorem-scope verification tables, executable algebraic/numerical theorem checks, empirical protocol tables, evidence summaries for the rare-event and Grid2Op studies, and commands for verifying the supplied artifacts. The sequential Grid2Op confirmation is reported separately from its V1 discovery sweep, and the primary/secondary inferential hierarchy is preserved.

AI USE STATEMENT

AI tools were used for research literature review and writing refinement. The mathematical statements and proofs, code, and experiment design and results analysis were done by the author who takes full responsibility for all of them in the submission.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 2017.

R. R. Bahadur and R. Ranga Rao. On deviations of the sample mean. *The Annals of Mathematical Statistics*, 31(4):1015–1027, 1960. doi: 10.1214/aoms/1177705674.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *ICML*, pp. 560–569, 2018.

Andrew C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941. doi: 10.1090/S0002-9947-1941-0003498-3.

Dimitris Bertsimas and Ioana Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005. doi: 10.1137/S1052623401399903.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Raghu Bollapragada, Richard Byrd, and Jorge Nocedal. Adaptive sampling strategies for stochastic optimization. *SIAM Journal on Optimization*, 28(4):3312–3343, 2018. doi: 10.1137/17M1154679.

Richard H. Byrd, Gillian M. Chin, Jorge Nocedal, and Yuchen Wu. Sample size selection in optimization methods for machine learning. *Mathematical Programming*, 134(1):127–155, 2012. doi: 10.1007/s10107-012-0572-5.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010. doi: 10.1287/opre.1090.0741.

Carl-Gustav Esseen. On the liapounoff limit of error in the theory of probability. *Arkiv för Matematik, Astronomi och Fysik*, 28A(9):1–19, 1942.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015. URL `https://jmlr.org/papers/v16/garcia15a.html`.

Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, J. Zico Kolter, Zachary C. Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3610–3619, 2021.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004.

Grid2Op developers. Available environments. Grid2Op 1.12.4 documentation, 2026. URL `https://grid2op.readthedocs.io/en/latest/available_envs.html`. Accessed 2026-09-16.

Jiawei Huang and Nan Jiang. From importance sampling to doubly robust policy gradient. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 4434–4443, 2020.

Wouter Jongeneel, Daniel Kuhn, and Mengmeng Li. A large deviations perspective on policy gradient algorithms. In *Proceedings of the 6th Annual Learning for Dynamics and Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pp. 916–928, 2024.

Tomas Juškevičius. The sharp form of the kolmogorov–rogozin inequality and a conjecture of leader–radcliffe. In *Bulletin of the London Mathematical Society*, 2024. doi: 10.1112/blms.13132.

Nathan Kallus and Masatoshi Uehara. Statistically efficient off-policy policy gradients. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 5089–5100, 2020.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer New York, 2003. doi: 10.1007/b97478.

Antoine Marot, Benjamin Donnot, Gabriel Dulac-Arnold, Adrian Kelly, Aidan O'Sullivan, Jan Viebahn, Mariette Awad, Isabelle Guyon, Patrick Panciatici, and Camilo Romero. Learning to run a power network challenge: a retrospective analysis. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pp. 112–132, 2021.

Diego Martinez-Taboada and Aaditya Ramdas. Empirical bernstein in smooth banach spaces. *The Annals of Applied Probability*, 36(3):2199–2236, 2026. doi: 10.1214/25-AAP2277.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Adaptive batch size for safe policy gradients. In *Advances in Neural Information Processing Systems 30*, pp. 3591–3600, 2017.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4026–4035, 2018.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients. *Machine Learning*, 111:4081–4137, 2022. doi: 10.1007/s10994-022-06232-6.

B. A. Rogozin. An estimate for concentration functions. *Theory of Probability and Its Applications*, 6(1):94–97, 1961. doi: 10.1137/1106009.

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

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

Mark Towers, Ariel Kwiatkowski, Jordan K. Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments, 2024.

Lieven Vandenberghe, Stephen Boyd, and Katherine Comanor. Generalized chebyshev bounds via semidefinite programming. *SIAM Review*, 49(1):52–64, 2007. doi: 10.1137/S0036144504440543.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

A SIGN-SYMMETRIC EXPERIMENTS AND THE TILTED-RISK IDENTITY

Fix $C_0 > 0$ and $B > 0$. Let L denote the coordinate variable and let R be a symmetric probability law for L on $[-B, B]$ satisfying

$$\mathbb{E}_R \cosh(L/2) = m = e^{C_0}. \tag{9}$$

Set $\beta = m^{-1}$ and define two probability measures, with respect to R,

$$\frac{dP_+}{dR}(\ell) = \beta e^{\ell/2}, \qquad \frac{dP_-}{dR}(\ell) = \beta e^{-\ell/2}. \tag{10}$$

Symmetry and (9) imply that both integrate to one, and $\log(dP_+/dP_-) = L$. The pair is invariant under reflection $\ell \mapsto -\ell$ and exchange of the hypotheses. Its Chernoff objective is symmetric around $1/2$ and log-convex, so the minimizer is $1/2$ and the Chernoff information is $C_0 = -\log \beta$.

For n iid observations write $S_n = \sum_{j=1}^n L_j$. Since

$$\min\{dP_+^n, dP_-^n\} = \beta^n e^{-|S_n|/2}\, dR^n,$$

the equal-prior Bayes risk is

$$e_n^*(R) = \frac{1}{2}\beta^n H_n(R), \qquad H_n(R) := \mathbb{E}_{R^n} e^{-|S_n|/2}. \tag{11}$$

By reflection symmetry the likelihood-ratio rule is an equalizer. If R is atom-free then S_n is atom-free, hence

$$P_+(S_n < 0) = P_-(S_n > 0) = e_n^*(R). \tag{12}$$

Thus the classical Bayes/minimax testing quantity becomes the literal strict wrong-direction probability whenever the PG realization has $G = cL$ with $c > 0$ and R is atom-free.

B PROOF OF THE STRICT FINITE-MENU PG NO-GO THEOREM

We prove Theorem 1 directly with continuous LR microcells; no limiting statement is needed.

B.1 FEATURE-MATCHED CONTINUOUS MICROCELLS

Fix a jointly finitely generated menu as in Definition 1, with common continuous generating list. After policy-side constants are fixed, enlarge that list to continuous functions $h_0, \dots, h_d$ of the positive LR magnitude, including $h_0(\ell) = \cosh(\ell/2)$, the first two reward/gradient moment coefficient functions needed for the complete $J_s(\eta)$, $\nabla J_s(\eta)$, and $\mathbb{E}\widehat{g}(\eta)^2$ matching claimed in the theorem, and all coefficient functions needed for any additional requested TV, KL, Rényi, reward, or stochastic-gradient coordinates. Let $r \le d + 2$ be a conservative Carathéodory support bound including normalization.

For completeness, the standard information coordinates named in Definition 1 really are finite linear features in this representation. If ν is the law of the positive LR magnitude under the symmetric tilted law and

$$\beta = \left(\int \cosh(\ell/2)\, d\nu(\ell) \right)^{-1},$$

then direct substitution into $dP_\pm = \beta e^{\pm L/2} dR$ gives

$$\mathrm{TV}(P_+, P_-) = \beta \int \sinh(\ell/2)\, d\nu(\ell), \tag{13}$$

$$\mathrm{KL}(P_+ \| P_-) = \beta \int \ell \sinh(\ell/2)\, d\nu(\ell), \tag{14}$$

$$\mathcal{A}_\alpha(P_+, P_-) := \int dP_+^\alpha dP_-^{1-\alpha} = \beta \int \cosh((\alpha - \tfrac{1}{2})\ell)\, d\nu(\ell). \tag{15}$$

Thus Hellinger/Bhattacharyya corresponds to $\alpha = 1/2$, any prescribed finite Rényi-affinity menu adds finitely many continuous functions, and TV/KL each add one. Under the bounded PG realization below, every prescribed finite reward or gradient moment similarly becomes an integral of a

continuous coefficient function of ℓ. The theorem therefore covers these conventional finite menus without treating a scalar encoding of the complete law as an admissible "descriptor."

Fix the prescribed even $n \geq 2$. Define

$$c_n = \binom{n}{n/2} 2^{-n} (3/4)^n r^{1-n}.$$

Choose an integer J large enough that

$$J \geq \frac{8e^{1/2}K}{c_n}, \tag{16}$$

and put $M = 4 \log J$. For $j = 1, \ldots, J$, let Q_j be the uniform law on the positive interval

$$I_j = [jM - \delta, jM + \delta],$$

where

$$0 < \delta \leq \min\{1/n, M/(2n)\}. \tag{17}$$

The intervals are disjoint. Define the cell feature vector

$$v_j = (\mathbb{E}_{Q_j} h_0, \ldots, \mathbb{E}_{Q_j} h_d).$$

Let $\nu_A = J^{-1} \sum_j Q_j$. Its feature vector $\bar{v} = J^{-1} \sum_j v_j$ lies in the convex hull of the v_j. Carathéodory therefore gives weights $w_j \geq 0$, supported on at most r cells and summing to one, with

$$\sum_j w_j v_j = \bar{v}.$$

Let $\nu_S = \sum_j w_j Q_j$ and

$$\nu_B = \frac{3}{4}\nu_S + \frac{1}{4}\nu_A.$$

Both ν_A and ν_B are absolutely continuous and have exactly the same retained feature vector. In particular they have the same $\mathbb{E}\cosh(L/2)$ after adding an independent equiprobable sign, hence the same Bhattacharyya coefficient, Hellinger distance, and Chernoff information.

B.2 Strict finite-batch separation

Let the signed tilted LR draw be $L = \varepsilon\ell$, with ε an independent Rademacher sign and $\ell \sim \nu_i$. Under endpoint A, condition on all signs and on the first $n-1$ base-cell indices. The equation for a zero integer base sum determines the final index uniquely if it has a solution, so the probability that the signed base centers sum to zero is at most $1/J$. Otherwise the base-center sum has magnitude at least M, while the continuous perturbations change the total by at most $n\delta \leq M/2$. Therefore

$$H_n(\nu_A) \leq \frac{1}{J} + e^{-(M - n\delta)/2} \leq \frac{2}{J}. \tag{18}$$

For ν_B, consider the event that all n magnitude draws come from the sparse component, all choose the same sparse cell, and exactly $n/2$ signs are positive. On this event the signed sum has magnitude at most $n\delta \leq 1$. If the nonzero sparse weights are $w_1, \ldots, w_q$, $q \leq r$, then

$$H_n(\nu_B) \geq e^{-1/2} (3/4)^n \binom{n}{n/2} 2^{-n} \sum_{j=1}^{q} w_j^n \tag{19}$$

$$\geq e^{-1/2} c_n, \tag{20}$$

where convexity gives $\sum_j w_j^n \geq q^{1-n} \geq r^{1-n}$. Since the Chernoff normalizer β is exactly matched, (11) and (16) imply

$$\frac{e^*_{B,n}}{e^*_{A,n}} \geq \frac{e^{-1/2} c_n J}{2} > K.$$

The LR laws are absolutely continuous, so by (12) these are literal strict risks.

B.3 Exact bounded PG realization

Let $D_L = JM + \delta$ and use a two-action $H = 2$ MDP. The terminal transition variable is the continuous LR L under the hidden-sign experiment and is independent of the initial Bernoulli-logit action A. Use the common deterministic reward

$$R(A, L) = (2A - 1)\frac{L}{D_L}.$$

At $\eta_0 = 0$,

$$G = R(A, L)(A - 1/2) = \frac{L}{2D},$$

so the empirical REINFORCE sign is exactly the LR sign rule and $|R| \le 1$, $|G| \le 1/2$. Any additional finite requested PG moment coordinate is matched by adding its continuous LR coefficient function to the feature list before Carathéodory. By construction the mandatory feature list already matches the first two moments of $X = L/D_L$, which gives for every $p = \sigma(\eta)$

$$J_s(\eta) = (2p - 1)\mathbb{E}_s X, \qquad \nabla J_s(\eta) = 2\mathbb{E}_s X\, p(1 - p),$$
$$\mathbb{E}\widehat{g}(\eta)^2 = \mathbb{E}X^2\, p(1 - p),$$

identically across the two endpoints. The score law and Fisher geometry are common because the policy is common. This proves Theorem 1.

C Proof of the sharp atom-free strict-risk phase diagram

We first derive the envelope for all symmetric tilted laws, then prove that the atom-free exact-C_0 subclass has the same infimum and supremum.

C.1 Universal lower prefactor

Fix $C_0 > 0$, $m = e^{C_0}$, $\beta = e^{-C_0}$, and let R be any symmetric tilted law satisfying (9). Since $\cosh(x/2) \ge 1 + x^2/8$,

$$v := \mathbb{E}L^2 \le V_0 := 8(m - 1).$$

For even n, the positive Fourier representation

$$e^{-|x|/2} = \frac{1}{\pi}\int_0^\infty \frac{\cos(tx)}{t^2 + 1/4}\, dt \tag{21}$$

gives

$$H_n(R) = \frac{1}{\pi}\int_0^\infty \frac{\varphi(t)^n}{t^2 + 1/4}\, dt, \qquad \varphi(t) := \mathbb{E}_R[\cos(tL)].$$

For this proof, write

$$\Gamma_n^H(C_0, B) := \frac{\sup_R H_n(R)}{\inf_R H_n(R)},$$

where the supremum and infimum range over all symmetric tilted laws satisfying the same (C_0, B) constraints.

On $0 \le t \le (nv)^{-1/2}$, $\varphi(t) \ge 1 - t^2 v/2$; for even n this yields a contribution $\Omega((nv)^{-1/2})$. If $nv < 4$, Jensen and Cauchy–Schwarz instead give a positive constant. Thus

$$H_n(R) \ge c_e(C_0) n^{-1/2} \tag{22}$$

for every even $n \ge 2$. For odd $n \ge 3$, the triangle inequality and independence imply

$$H_n(R) \ge H_{n-1}(R)\mathbb{E}e^{-|L|/2}.$$

Pointwise $e^{-|L|/2} \ge [2\cosh(L/2)]^{-1}$, and Jensen yields $\mathbb{E}e^{-|L|/2} \ge (2m)^{-1}$. Hence

$$H_n(R) \ge c_0(C_0) n^{-1/2} \qquad (n \ge 2). \tag{23}$$

C.2 THE LR RADIUS LIMITS CONCENTRATION

Let $a = 2\,\mathrm{arcosh}(m)$, $b_0 = a/2$, and $u_0 = \cosh(a/4) < m$. For $B > a$ define

$$\tau_{C_0}(B) = \frac{m - u_0}{\cosh(B/2) - u_0}.$$

From the moment constraint,

$$\mathbb{P}_R(|L| > b_0) \geq \tau_{C_0}(B). \tag{24}$$

By symmetry an interval of length a cannot contain both tails, so its one-step concentration function satisfies $Q_a(L) \leq 1 - \tau_{C_0}(B)/2$. Kolmogorov–Rogozin (Rogozin, 1961; Juškevičius, 2024) gives

$$Q_a(S_n) \leq \frac{C(C_0)}{\sqrt{n\tau_{C_0}(B)}}.$$

Partitioning the line into intervals of length a and summing the exponentially decaying kernel gives

$$H_n(R) \leq \min\left\{1, \frac{A(C_0)}{\sqrt{n\tau_{C_0}(B)}}\right\}. \tag{25}$$

Together with (23), this proves the envelope upper bound

$$\Gamma_n^H(C_0, B) \leq C(C_0)\min\{\sqrt{n}, \tau_{C_0}(B)^{-1/2}\}.$$

C.3 MATCHING LOWER CONSTRUCTIONS FOR EVERY PARITY

Let

$$q_B = \frac{m - 1}{\cosh(B/2) - 1}$$

and consider the symmetric tilted laws

$$R_{\mathrm{spike}} = (1 - q_B)\delta_0 + \frac{q_B}{2}(\delta_{+B} + \delta_{-B}), \qquad R_{\mathrm{rad}} = \tfrac{1}{2}(\delta_{+a} + \delta_{-a}).$$

Both satisfy the exact same moment constraint (9). For even k, the characteristic-function argument around zero gives

$$H_k(R_{\mathrm{spike}}) \geq c\min\{1, (kq_B)^{-1/2}\}.$$

For odd n, $H_n(R_{\mathrm{spike}}) \geq (2m)^{-1}H_{n-1}(R_{\mathrm{spike}})$. For the Rademacher law, the maximal binomial atom and a geometric summation give

$$H_n(R_{\mathrm{rad}}) \leq A(C_0)n^{-1/2}$$

for both parities. Since $q_B = \Theta_{C_0}(\tau_{C_0}(B))$ for $B > a$,

$$\Gamma_n^H(C_0, B) \geq c(C_0)\min\{\sqrt{n}, \tau_{C_0}(B)^{-1/2}\}. \tag{26}$$

Finally $\tau_{C_0}(B) = \Theta_{C_0}(e^{-B/2})$, proving the all-law version of the phase diagram.

C.4 ATOM-FREE DENSITY AT FIXED CHERNOFF INFORMATION

Lemma 1 (Moment-constrained atom-free density)**.** *Fix $C_0 > 0$ and $B > a(C_0)$. Symmetric absolutely continuous probability laws on $[-B, B]$ satisfying* (9) *are weakly dense in the set of all symmetric probability laws satisfying that same equation.*

Proof. Let R be any target law. Approximate it weakly by symmetric absolutely continuous laws R_k^0 supported on $[-B, B]$ (replace every atom by a sufficiently narrow symmetric microcell, moving boundary cells inward). Write $m_k = \mathbb{E}_{R_k^0}\cosh(L/2)$; bounded continuity gives $m_k \to m$.

Because $1 < m < \cosh(B/2)$, choose two fixed symmetric absolutely continuous anchor laws Q_- and Q_+ supported respectively in sufficiently small neighborhoods of 0 and of $\{-B, +B\}$ such that

$$\mathbb{E}_{Q_-}\cosh(L/2) < m < \mathbb{E}_{Q_+}\cosh(L/2).$$

If $m_k < m$, mix R_k^0 with Q_+ using the unique coefficient $\alpha_k \in [0,1]$ that restores moment m; if $m_k > m$, use Q_-. Explicitly, for the upper anchor,

$$\alpha_k = \frac{m - m_k}{\mathbb{E}_{Q_+} \cosh(L/2) - m_k},$$

and analogously for the lower anchor. Since $m_k \to m$ and the anchor moment gaps are fixed, $\alpha_k \to 0$. The corrected law $R_k = (1-\alpha_k)R_k^0 + \alpha_k Q_\pm$ is symmetric, absolutely continuous, has exact moment m, and converges weakly to R. □

The map

$$R \mapsto H_n(R) = \int e^{-|\ell_1 + \cdots + \ell_n|/2}\, dR^{\otimes n}$$

is continuous under weak convergence because its integrand is bounded and continuous on the compact product space. Lemma 1 therefore implies that the infimum and supremum of H_n over the atom-free exact-C_0 class equal those over the full conditioned class. In the atom-free class S_n has a density and no tie. Combining with (12) and (26) proves Theorem 2.

D Proof of strict saturation with the PG moments fixed

This proof starts from the positive finite-support construction used to match the PG constraints exactly, then makes both endpoint LR laws atom-free without relaxing a single matched equation.

D.1 PG moment coordinates

Fix $C_0 > 0$, $\beta = e^{-C_0}$, $m = e^{C_0}$, $h = 1 - \beta$, and $a = 2\,\mathrm{arcosh}(m)$. For a symmetric tilted LR law with magnitude $|L| = Bx$, $x \in [0,1]$, define a probability measure ν by

$$\nu(dx) = \beta \cosh(Bx/2) R_{|L|}(B\,dx).$$

Then

$$R_{|L|}(B\,dx) = m\,\mathrm{sech}(Bx/2)\nu(dx).$$

For the positive hidden sign,

$$\mathbb{E}_+ \frac{L}{B} = \mathbb{E}_\nu[x \tanh(Bx/2)], \qquad \mathbb{E}_+ \left(\frac{L}{B}\right)^2 = \mathbb{E}_\nu x^2.$$

We impose the four linear equations

$$\mathbb{E}_\nu 1 = 1, \quad \mathbb{E}_\nu\,\mathrm{sech}(Bx/2) = \beta, \quad \mathbb{E}_\nu[x\tanh(Bx/2)] = \frac{3h}{4}, \quad \mathbb{E}_\nu x^2 = \frac{29h}{48}. \tag{27}$$

They imply exactly

$$\mathbb{E}_s G = s\frac{3h}{8} = s\mu_0(C_0), \qquad \mathbb{E}_s G^2 = \frac{29h}{192} = m_2(C_0)$$

for $G = L/(2B)$.

D.2 Positive finite-support endpoints

For endpoint A, take normalized magnitudes $\{0, 1/2, 3/4, 1\}$. With

$$F_B(x) = (1, \mathrm{sech}(Bx/2), x\tanh(Bx/2), x^2)^\top,$$

solve the 4×4 system whose target is $(1, \beta, 3h/4, 29h/48)^\top$. As $B \to \infty$ the unique weights converge to

$$(\beta, h/3, h/3, h/3),$$

all positive.

For endpoint B, fix $\rho = 1/24$ at $x = a/B$ and solve the same four equations for weights on $\{0, 1/2, 3/4, 1\}$. The remaining weights converge to

$$(23\beta/24, h/12, 2h/3, 5h/24),$$

again strictly positive. Hence for sufficiently large B both exact finite-B solutions are positive, uniformly on compact nondegenerate C_0 intervals.

For the induced tilted LR law, endpoint B has total mass $1/24$ at magnitude a (mass $1/48$ at each sign). Endpoint A has nonzero tilted mass

$$q_A \le m\,\mathrm{sech}(B/4) \le 2me^{-B/4}.$$

Thus for $B \ge 4\log(4mn)$, endpoint A has $H_n = \Omega(1)$, while Kolmogorov–Rogozin applied to endpoint B gives $H_n = O_{C_0}(n^{-1/2})$.

D.3 Atom-free exactification with all constraints preserved

For endpoint A, replace each magnitude atom x_j by an absolutely continuous microcell law $Q_{j,\delta}$ supported in a disjoint interval of radius δ around x_j (use $[0,\delta]$ at zero). Let

$$\bar{F}_{j,\delta} = \mathbb{E}_{Q_{j,\delta}} F_B(X).$$

The matrix with columns $\bar{F}_{j,\delta}$ converges to the original invertible feature matrix as $\delta \downarrow 0$. Therefore it remains invertible for all sufficiently small δ, and the unique weights solving the *same exact target* in (27) converge to the original strictly positive weights. Endpoint A is thus atom-free and matches all four constraints exactly.

For endpoint B, also replace the distinguished a/B atom by a microcell, retain its coefficient $\rho = 1/24$, and solve the remaining 4×4 system using the averaged feature columns. Again invertibility and positivity persist for small δ. Converting ν to R multiplies its density by the positive continuous factor $m\,\mathrm{sech}(Bx/2)$, so the resulting signed LR laws are atom-free.

Choose δ small enough that $B\delta \le \min\{1/n, a/8\}$. For endpoint A, the low-magnitude tilted mass converges to the original near-one mass. On the event all n draws fall in that cell, $|S_n| \le nB\delta \le 1$; hence $H_n(A) = \Omega_{C_0}(1)$. For endpoint B, the two smeared clusters around $\pm a$ retain masses bounded below by a positive C_0-independent constant near $1/48$ and remain separated. Any interval of length $a/2$ excludes at least one cluster, so Kolmogorov–Rogozin gives $H_n(B) = O_{C_0}(n^{-1/2})$. Both laws are atom-free, therefore the ratio is a literal strict-risk ratio. The upper bound follows from Theorem 2. This proves Corollary 1.

D.4 PG realization

Use the common Bernoulli-logit policy and terminal LR transition variable, independent of the action. The deterministic reward

$$R(A, L) = (2A - 1)L/B$$

gives $G = L/(2B)$ at $p = 1/2$. Equations (27) yield the exact common values μ_0, m_2. For arbitrary $p = \sigma(\eta)$,

$$J_s(\eta) = 2s\mu_0(2p-1), \quad \nabla J_s(\eta) = 4s\mu_0 p(1-p), \quad \mathbb{E}\widehat{g}(\eta)^2 = 4m_2 p(1-p),$$

so all PG quantities in the theorem are matched pointwise in the policy parameter.

E A common-jitter strict-risk robustness lemma

The atom-free theorems above solve the tie issue intrinsically. The following independent lemma is useful for interpreting any lattice example.

Lemma 2 (Common continuous PG jitter)**.** *Let G be a bounded scalar PG sample and let U be an independent, symmetric, bounded continuous random variable. Put $G_\varepsilon = G + \varepsilon U$. For every fixed batch n,*

$$\mathbb{P}\left(\sum_{i=1}^n G_{\varepsilon,i} < 0\right) \longrightarrow \mathbb{P}\left(\sum_{i=1}^n G_i < 0\right) + \frac{1}{2}\mathbb{P}\left(\sum_{i=1}^n G_i = 0\right).$$

If G is lattice and ε is small enough that the jitter cannot change the sign of any nonzero batch sum, equality holds exactly.

Proof. Condition on the unperturbed batch sum. Negative and positive sums remain negative and positive in the limit; at a zero sum, the continuous symmetric jitter sum is negative with probability exactly $1/2$. The exact lattice statement follows when the perturbation radius is smaller than the minimum nonzero batch-sum magnitude. □

For a Bernoulli-logit PG sample at $p = 1/2$, the jitter has an exact bounded MDP realization by replacing the reward with

$$R_\varepsilon = R + 2\varepsilon(2A - 1)U.$$

Then $G_\varepsilon = G + \varepsilon U$. Because $\mathbb{E}U = 0$, the objective and population gradient are unchanged for every policy parameter; the second-moment function receives the same additive term in both compared environments. The supplied script `run_strict_tie_check.py` verifies exact equality in the paper's finite lattice examples when the nonzero-sum sign cannot flip.

F ARBITRARY-HORIZON NATURAL REINFORCE THEOREM

We prove Theorem 3 and its open-policy, deterministic-baseline, and finite-batch consequences, including Corollary 2. The reward is ordinary terminal position and never contains the likelihood ratio. Throughout, $T = H - 1$.

F.1 LOCAL PAIRED-SUPPORT COORDINATES AND THE EXACT MATCHED MANIFOLD

Fix positive magnitudes x_j and skews $t_j \in (-1, 1)$ and define, under positive hidden sign,

$$P_q(+x_j) = \frac{q_j(1 + t_j)}{2}, \qquad P_q(-x_j) = \frac{q_j(1 - t_j)}{2}. \tag{28}$$

The negative hidden-sign law is the reflection. The one-stage hidden-sign LR is

$$\ell(+x_j) = \log \frac{1 + t_j}{1 - t_j}, \qquad \ell(-x_j) = -\ell(+x_j),$$

so its pointwise formula is independent of q. Reflection symmetry puts the Chernoff optimizer at $1/2$, with Bhattacharyya coefficient

$$\beta(q) = \sum_j q_j \sqrt{1 - t_j^2}.$$

The four linear local features are therefore

$$1, \qquad x_j t_j, \qquad x_j^2, \qquad \sqrt{1 - t_j^2}.$$

Let $A(x, t)$ be the corresponding $4 \times k$ matrix. If $\operatorname{rank} A = 4$, then for a fixed target vector $c \in \mathbb{R}^4$ the positive set

$$\mathcal{M}_c = \{q > 0 : Aq = c\}$$

is locally an affine manifold with tangent $\ker A$ and dimension $k - 4$.

We use the same exact rational design as the $H = 2$ theorem:

$$x = \left(\frac{1}{9}, \frac{1}{3}, \frac{5}{9}, \frac{7}{9}, 1\right), \qquad t = \left(\frac{7}{25}, \frac{3}{5}, \frac{20}{29}, \frac{8}{17}, \frac{4}{5}\right).$$

Then

$$\sqrt{1 - t^2} = \left(\frac{24}{25}, \frac{4}{5}, \frac{21}{29}, \frac{15}{17}, \frac{3}{5}\right),$$

the local LR radius is $B_0 = \log 9$, and the first four feature columns have determinant

$$-\frac{43168}{14974875} \neq 0.$$

An exact null vector is

$$v = \left(1, -\frac{867}{5465}, -\frac{4089}{5465}, -\frac{4556}{5465}, \frac{4047}{5465}\right). \tag{29}$$

H	$I_A^{(H)}$	$I_B^{(H)}$	Δ_H	$(H-1)\Delta_H$
2	.18717	.21423	.02705	.02705
4	.17277	.18877	.01600	.04801
8	.17790	.18629	.00839	.05876
16	.18109	.18541	.00432	.06479
32	.18283	.18502	.00219	.06798

Table 3: Representative certified rates for the same explicit matched endpoints at increasing genuine horizons. The theorem proves positivity for every finite H; the final column approaches $\kappa_* \approx 0.07129$.

Let

$$D = 1094199182219124711,$$
$$a_0 = \frac{214905531270189760}{D}, \qquad b_0 = \frac{234577057138365671}{D},$$
$$q_0 = (a_0, a_0, a_0, b_0, a_0).$$

For

$$q_\delta = q_0 + \delta v, \qquad \delta \in [-19/100, 1/4], \tag{30}$$

all masses remain strictly positive and all four local constraints are exact. The common local Chernoff coefficient is

$$\beta_0 = \frac{126117818247319762697}{158658881421773083095}, \qquad C_0 = -\log \beta_0 \approx 0.22953996.$$

The endpoints are $q_A = q_{-19/100}$ and $q_B = q_{1/4}$. Table 3 reports representative certified rates for these same fixed endpoints across increasing horizons.

F.2 A GENUINELY SEQUENTIAL H-STEP LIFT AND ITS EXACT PG GEOMETRY

At every stage $t = 1, \dots, T$, draw $A_t \sim \text{Bernoulli}(p)$ with common $p = \sigma(\eta)$, set $U_t = 2A_t - 1$, draw $W_t \sim P_{q,s}$ independently, and evolve

$$X_t = X_{t-1} + \frac{U_t W_t}{T}, \qquad X_0 = 0.$$

There are no intermediate rewards and the deterministic terminal reward is $R = X_T$. The trajectory score and vanilla REINFORCE sample are

$$S_T = \sum_{t=1}^{T} (A_t - p), \qquad \widehat{g}_H = R S_T.$$

Write $r = 2p - 1$, $v = p(1-p)$, $\mathbb{E}_{P_{q,+}} W = \mu$, and $\mathbb{E}W^2 = m_2$. Reflection sends μ to $s\mu$. Since $R = T^{-1} \sum U_t W_t$,

$$J_s^{(H)}(\eta) = s\mu r, \qquad \nabla J_s^{(H)}(\eta) = 2s\mu v.$$

Also

$$\mathbb{E}R^2 = \frac{m_2}{T} + \frac{T-1}{T}\mu^2 r^2. \tag{31}$$

To compute the REINFORCE second moment, put $Z = \sum U_t$, so $S_T = (Z - Tr)/2$. Conditional on the actions,

$$\mathbb{E}[R^2 \mid U_1, \dots, U_T] = \frac{(m_2 - \mu^2)T + \mu^2 Z^2}{T^2}.$$

The centered Bernoulli moments are

$$\mathbb{E}S_T^2 = Tv, \qquad \mathbb{E}(Z - Tr)^3 = -8Trv,$$
$$\mathbb{E}(Z - Tr)^4 = 16\{Tv(1 - 6v) + 3T^2 v^2\}.$$

Substituting these identities gives

$$\mathbb{E}[(\widehat{g}_H)^2] = m_2 v + \frac{T-1}{T}\mu^2 \left[(T-4)r^2 v + 8v^2\right]. \tag{32}$$

Thus every q in (30) has the same complete J, gradient, reward-second-moment, gradient-second-moment/variance, score-law, entropy, and Fisher functions. The trajectory Fisher is

$$F_H(\eta) = \mathbb{E}S_T^2 = Tv.$$

The terminal reward satisfies $|R| \le 1$ for every H.

This lift is sequential rather than padded. When $T \ge 2$, the score has $T+1$ support values and

$$\widehat{g}_H = \frac{1}{T}\sum_{t,j} U_t W_t (A_j - p)$$

contains nonzero cross-stage terms $t \ne j$. Equation (32) also contains an explicit T-dependent cross-stage contribution. Deleting one stage changes this law and the score/Fisher geometry.

F.3 Trajectory information

The observed trajectory contains (A_t, X_{t-1}, X_t), so

$$W_t = TU_t(X_t - X_{t-1})$$

is recovered stage by stage. The policy probabilities are identical under the two hidden signs and cancel in their likelihood ratio. Hence

$$L_H(\tau) = \sum_{t=1}^{T} \ell(W_t). \tag{33}$$

The pointwise trajectory LR map and its support are therefore common across q. Independence gives

$$C_H = TC_0, \qquad B_H = T\log 9. \tag{34}$$

The probabilities assigned to LR values still differ with q; equality of the LR spectrum, full experiment, or all Rényi divergences is not claimed.

F.4 Exact rollout-gradient MGF at the reference policy

At $p = 1/2$,

$$S_T = \frac{1}{2}\sum_{t=1}^{T} U_t, \qquad G_{T,q} := \widehat{g}_H = \frac{1}{2T}\left(\sum_t U_t\right)\left(\sum_t U_t W_t\right).$$

Let

$$M_q(z) = \mathbb{E}_q e^{zW}.$$

Condition on k plus actions, write $d = 2k - T$ and $c_k = \lambda d/(2T)$. Independence of the W_t yields

$$\Psi_{T,q}(\lambda) = 2^{-T}\sum_{k=0}^{T}\binom{T}{k} M_q(c_k)^k M_q(-c_k)^{T-k}. \tag{35}$$

The rollout gradient is nonconstant, has positive mean $\mu/2$, and has negative support, so $\Psi''_{T,q}(\lambda) > 0$ and its unique minimizer $\lambda_{T,q}$ is finite and negative. Define

$$I_T(q) = -\log \Psi_{T,q}(\lambda_{T,q}).$$

For a matched tangent v, let $h(z) = DM_q(z)[v]$. Since $\partial_\lambda \Psi(\lambda_{T,q}) = 0$, the envelope theorem gives

$$DI_T(q)[v] = -\frac{D_q\Psi_{T,q}(\lambda_{T,q})[v]}{\Psi_{T,q}(\lambda_{T,q})}. \tag{36}$$

F.5 TWO EXACT LOCAL CERTIFICATES

For the rational null direction (29), put $y = e^{z/9}$. Exact algebra gives

$$h(z) = \frac{(y-1)^3(y+1)^3}{54650y^9} H(y), \tag{37}$$

where

$$H(y) = 36423y^{12} + 75769y^{10} + 83493y^8 + 52659y^6 + 18243y^4 - 81y^2 - 4047.$$

For $|z| \le 7/5$, $y \ge e^{-7/45} > 5/6$, and

$$H(y) \ge 18243y^4 - 81y^2 - 4047 > 18243(5/6)^4 - 81 - 4047 > 0.$$

Thus

$$\operatorname{sign} h(z) = \operatorname{sign} z, \qquad |z| \le 7/5. \tag{38}$$

We also need

$$\frac{h(-a)}{M_\delta(-a)} + \frac{h(a)}{M_\delta(a)} < 0, \qquad 0 < a \le 7/5, \tag{39}$$

for every $\delta \in [-19/100, 1/4]$. Clearing the positive denominator $M_\delta(a)M_\delta(-a)$ gives

$$N_\delta(a) = h(a)M_\delta(-a) + h(-a)M_\delta(a).$$

After extracting the positive factor $(y-1)^4(y+1)^4/y^{18}$, the remaining palindromic polynomial becomes a degree-seven polynomial in

$$w = y^2 + y^{-2} - 2 = 4\sinh^2(a/9).$$

For $a \le 7/5$, the elementary bound $\cosh x \le (1 - x^2/2)^{-1}$ at $x = 14/45$ gives $w < 0.11$. The polynomial is affine in δ. At the two endpoints $\delta = -19/100$ and $1/4$, summing every positive coefficient at $w = 0.11$ gives, after normalization by the absolute constant coefficient, the strict upper bounds

$$-0.2279950004\ldots, \qquad -0.9247125615\ldots,$$

respectively. Linearity in δ proves (39) on the whole continuum. These are exact rational inequalities before decimal display; the certificate script records the fractions.

F.6 A HORIZON-UNIFORM BRACKET FOR EVERY CRAMÉR MINIMIZER

Put $m_\delta(z) = M'_\delta(z)/M_\delta(z)$. At $\lambda = -L < 0$, pair a conditional action count with imbalance $r = |2k - T|/T$ and set $a = Lr/2$. The sign of its λ derivative is the sign of

$$Q_L(a, \delta) = \left(1 + \frac{2a}{L}\right) m_\delta(-a) - \left(1 - \frac{2a}{L}\right) m_\delta(a). \tag{40}$$

After multiplication by the positive product $M_\delta(a)M_\delta(-a)$, $Q_L(a,\delta)/a$ has an even Taylor series in a whose coefficients are rational quadratic functions of δ.

At $L = 14/5$, exact maximization of the first fifteen coefficients over $\delta \in [-19/100, 1/4]$ makes every coefficient strictly negative; the least-negative upper bound is below -8.2×10^{-26}. The absolute remainder for $a \le 7/5$ is below 2×10^{-19}. Hence $Q_{14/5} < 0$. At $L = 8/5$, the exact lower polynomial through degree 28 is greater than 0.23 for $a \le 4/5$, while the absolute remainder is below 2×10^{-26}; hence $Q_{8/5} > 0$. Therefore every noncentral term in (35) has negative derivative at $-14/5$ and positive derivative at $-8/5$; the central term, when present, has zero derivative. Strict convexity gives the uniform bracket

$$-\frac{14}{5} < \lambda_{T,\delta} < -\frac{8}{5} \qquad (T \ge 1). \tag{41}$$

All coefficient and tail inequalities are reproduced exactly in the anonymous checker.

F.7 Strict exponent variation for every finite horizon

Fix the minimizer in (41). For any noncentral conditional term, let $r \in (0, 1]$ be the absolute action imbalance and $a = -\lambda r/2 \in (0, 7/5)$. The majority-action factors evaluate $M_\delta(-a)$. Up to a positive multiplier, the δ log-derivative of that conditional term is

$$(1+r)\frac{h(-a)}{M_\delta(-a)} + (1-r)\frac{h(a)}{M_\delta(a)}.$$

Rewrite it as

$$\left[\frac{h(-a)}{M_\delta(-a)} + \frac{h(a)}{M_\delta(a)}\right] + r\left[\frac{h(-a)}{M_\delta(-a)} - \frac{h(a)}{M_\delta(a)}\right].$$

The first bracket is strictly negative by (39); the second is strictly negative by (38). Thus every noncentral conditional term decreases strictly with δ. The central term, if T is even, has $a = 0$ and derivative zero. Consequently

$$\partial_\delta \Psi_{T,q_\delta}(\lambda_{T,\delta}) < 0$$

for every integer $T \geq 1$. Equation (36) proves

$$I_T'(\delta) > 0$$

throughout the full continuum and for every finite horizon $H = T + 1$.

For every fixed H, strict positivity of this derivative at an interior compact subinterval, together with rank and positivity margins and analytic dependence on (x, t, q), gives an open neighborhood of support/skew designs with the same non-identifiability property. This open set is in the finite-dimensional support/skew parameter topology for that fixed H; no universal open set over arbitrary MDPs is asserted.

F.8 Horizon dependence and the $1/(H-1)$ law

The exponent gap is positive for each finite H but cannot stay uniformly separated in this normalized lift. Put

$$Z_T = T^{-1/2}\sum_{t=1}^{T} U_t, \qquad K_q(z) = \log M_q(z), \qquad c = \frac{\lambda Z_T}{2\sqrt{T}}.$$

Conditioning on Z_T, the log of the product factor in (35) is

$$\mathcal{L}_{q,T} = \frac{T + \sqrt{T}Z_T}{2}K_q(c) + \frac{T - \sqrt{T}Z_T}{2}K_q(-c).$$

All matched q share the first two cumulants μ and $\sigma^2 = m_2 - \mu^2$. If $\kappa_3(q), \kappa_4(q)$ denote the next cumulants, Taylor expansion gives

$$\mathcal{L}_{q,T} = A(\lambda)Z_T^2 + \frac{B_q(\lambda)}{T}Z_T^4 + O\left(\frac{|Z_T|^6}{T^2}\right), \tag{42}$$

where

$$A(\lambda) = \frac{\mu\lambda}{2} + \frac{\sigma^2\lambda^2}{8}, \qquad B_q(\lambda) = \frac{\kappa_3(q)\lambda^3}{48} + \frac{\kappa_4(q)\lambda^4}{384}.$$

The remainder is uniform for λ in a compact neighborhood of the limiting minimizer. Split at $|Z_T| \leq T^{1/8}$. On the central event, bounded support gives the displayed uniform Taylor remainder; outside it, Hoeffding gives super-polynomially small probability, while the relevant exponential weights are uniformly integrable because $A(\lambda)$ is strictly negative near the limiting minimizer. Hence the Rademacher central limit theorem (CLT) and dominated convergence give locally uniform convergence

$$\Psi_{T,q}(\lambda) \longrightarrow [1 - 2A(\lambda)]^{-1/2}.$$

The limiting minimizer and exponent are

$$\lambda_\infty = -\frac{2\mu}{\sigma^2}, \qquad I_\infty = \frac{1}{2}\log\frac{m_2}{\sigma^2}. \tag{43}$$

To compare the two matched endpoints, subtract their expansions before taking the limit. The distributional correction of Z_T is common and cancels. Uniform integrability gives

$$\log \Psi_{B,T}(\lambda) - \log \Psi_{A,T}(\lambda) = \frac{B_B(\lambda) - B_A(\lambda)}{T} \frac{\mathbb{E}[Z^4 e^{A(\lambda)Z^2}]}{\mathbb{E}[e^{A(\lambda)Z^2}]} + o(T^{-1}),$$

locally uniformly for standard normal Z. Since the leading limiting objective has a nondegenerate unique minimizer, optimizing changes this difference only by $o(T^{-1})$. At λ_∞,

$$\frac{\mathbb{E}[Z^4 e^{AZ^2}]}{\mathbb{E}[e^{AZ^2}]} = \frac{3}{(1-2A)^2}.$$

Therefore

$$I_{B,T} - I_{A,T} = \frac{\kappa_*}{T} + o(T^{-1}), \tag{44}$$

with

$$\kappa_* = -[B_B(\lambda_\infty) - B_A(\lambda_\infty)] \frac{3}{(1-2A(\lambda_\infty))^2}.$$

For the exact endpoints, direct rational substitution gives

$$\kappa_* = 0.07128879336108669\ldots > 0, \qquad I_\infty = 0.184658441443\ldots.$$

Thus $\Delta_H = \kappa_*/(H-1) + o(H^{-1})$. This proves the stated horizon dependence but is not a universal impossibility result for all possible sequential constructions.

F.9 Open-policy persistence

For arbitrary $p \in (0,1)$, condition on k plus actions. The trajectory score is $S = k - Tp$ and, with $c = \lambda S/T$,

$$\Psi_{T,q,p}(\lambda) = \sum_{k=0}^{T} \binom{T}{k} p^k (1-p)^{T-k} M_q(c)^k M_q(-c)^{T-k}. \tag{45}$$

This finite-support MGF is analytic in (p, λ, q) and strictly convex in λ. For every fixed H, the reference-policy gap is strictly positive, so the implicit-function theorem and continuity give an $\epsilon_H > 0$ on which the ordering persists. No horizon-uniform ϵ is claimed.

F.10 Variance-optimal deterministic baseline

Allow an arbitrary deterministic time-indexed baseline vector $b = (b_1, \ldots, b_T)$:

$$\widehat{g}_{H,b} = RS_T - \sum_{t=1}^{T} b_t (A_t - p).$$

The score coordinates are independent, centered, and have covariance vI_T. Exchangeability makes $\mathbb{E}[(A_t - p)RS_T]$ independent of t, so the variance-minimizing vector has equal coordinates. It is therefore enough to optimize one scalar b. Using $Z = Tr + 2S_T$,

$$\mathbb{E}[RS_T^2] = \frac{s\mu}{T}\mathbb{E}[ZS_T^2] = s\mu r v(T-2), \qquad \mathbb{E}S_T^2 = Tv.$$

Hence

$$b^*_{s,H}(p) = s\mu(2p-1)\frac{T-2}{T}. \tag{46}$$

This is common across endpoints for a fixed hidden sign, recovers $s\mu(1-2p)$ at $T=1$, and equals zero at $p = 1/2$ for all H. The baselined MGF is analytic and strictly convex near $p = 1/2$, so its strict endpoint gap persists locally in p. For $H > 2$, state-dependent baselines are a larger class and are not covered.

Finally, $J_s^{(H)}(\eta) = s\mu(2\sigma(\eta) - 1)$ is strictly increasing in $s\eta$. Thus for every finite η, $\alpha > 0$, and realized batch gradient $\bar{g}$,

$$J_s^{(H)}(\eta + \alpha\bar{g}) < J_s^{(H)}(\eta) \iff s\bar{g} < 0 \iff \bar{g}\,\nabla J_s^{(H)}(\eta) < 0.$$

F.11 Finite-batch exponential separation for every fixed horizon

Fix H and endpoint A. Let $\lambda_A < 0$ minimize the one-rollout gradient MGF, let $I_A = -\log M_A(\lambda_A)$, and define the Cramér tilt

$$\frac{dR_A^*}{dP_A}(g) = \frac{e^{\lambda_A g}}{M_A(\lambda_A)}.$$

Under R_A^* the gradient sample has mean zero. Let σ_A^2 be its variance and

$$\beta_{3,A} = \frac{\mathbb{E}_{R_A^*}|G|^3}{\sigma_A^3}.$$

For any $a > 0$, on the strictly negative window $-a\sigma_A\sqrt{n} < S_n < 0$, change of measure gives

$$\mathcal{R}^{\mathrm{str}}_{A,H,n} \geq e^{-nI_A - a|\lambda_A|\sigma_A\sqrt{n}} R_A^*\{-a\sigma_A\sqrt{n} < S_n < 0\}.$$

Berry–Esseen implies

$$\mathcal{R}^{\mathrm{str}}_{A,H,n} \geq e^{-nI_A - a|\lambda_A|\sigma_A\sqrt{n}} \left[\Phi(a) - \frac{1}{2} - \frac{2C_{\mathrm{BE}}\beta_{3,A}}{\sqrt{n}}\right]_+. \tag{47}$$

For endpoint B, exponential Markov gives

$$\mathcal{R}^{\mathrm{str}}_{B,H,n} \leq e^{-nI_B}.$$

Thus, for any $0 < \gamma_H < \Delta_H$, choose $n_{0,H}$ large enough that the bracket in (47) is at least a fixed positive c_H and

$$a|\lambda_A|\sigma_A\sqrt{n} \leq (\Delta_H - \gamma_H)n.$$

Then for every $n \geq n_{0,H}$,

$$\frac{\mathcal{R}^{\mathrm{str}}_{A,H,n}}{\mathcal{R}^{\mathrm{str}}_{B,H,n}} \geq c_H e^{\gamma_H n}.$$

The deterministic verifier computes representative certificates at $H = 2, 4, 8, 16$. No odd-batch assumption is needed because the lower event is strictly negative.

G Proof of the all-horizon de-specialization theorem

G.1 Asymmetric LR cells and the exact matched manifold

Represent the local hidden-sign experiment on ten cells. Let $q_\ell = P_-(\ell) > 0$, let $r_\ell > 0$ be fixed, and set $P_+(\ell) = r_\ell q_\ell$. A cell has transition/reward center y_ℓ. At the Chernoff interpolation parameter $a_0 = 1/2$, define

$$\begin{aligned} F_y(q) = \big(&\sum q_\ell, \sum r_\ell q_\ell, \sum q_\ell y_\ell, \sum r_\ell q_\ell y_\ell, \\ &\sum q_\ell y_\ell^2, \sum r_\ell q_\ell y_\ell^2, \sum q_\ell\sqrt{r_\ell}, \sum q_\ell\sqrt{r_\ell}\log r_\ell\big). \end{aligned}$$

The first six coordinates fix both hidden-sign normalizations and first two moments. The seventh fixes the Bhattacharyya coefficient; the eighth makes a_0 stationary for $\sum q_\ell r_\ell^a$. Strict convexity in a then makes $1/2$ the unique Chernoff minimizer.

At the signed-cell representation of the explicit reflected construction, take the 8×8 minor formed by columns $\{1, 3, 4, 5, 7, 8, 9, 10\}$. In exact arithmetic its determinant is

$$\begin{aligned} &-\frac{621236224}{645700815}\log 7 - \frac{9312555008}{9685512225}\log 3 \\ &\quad + \frac{130929664}{387420489}\log 5 + \frac{34056470528}{9685512225}\log 2 \\ &= 0.0526859264\ldots \neq 0. \end{aligned}$$

The nonzero claim is exact: after clearing the common denominator, equality to zero would imply $2^{34056470528}5^{3273241600} = 3^{9312555008}7^{9318543360}$, contradicting unique prime factorization. Thus $D_q F_y$ has rank eight. The same minor remains nonsingular after every sufficiently small support

perturbation. Fixing the original target $c = F_{y_0}(q_0)$, a fixed-minor solve gives an analytic local section $q(y)$ and an analytic two-dimensional affine matched set

$$\{q > 0 : F_y(q) = c\}.$$

A matched tangent $v(y)$ can be selected continuously with $v(y_0) = v_0$.

To rule out a hidden reflection coordinate, take

$$d = (1, -.2, -.6, .9, .4, -1, -.7, .1, -1.2, 1.1)$$

and $y(\epsilon) = y_0 + \epsilon d$. For small ϵ the order of the ten support points is unchanged. A finite subset of $\mathbb{R}$ is sent to a reflected set by a nondegenerate affine map iff it is centrally symmetric; for sorted points this requires all five extreme-pair sums to coincide. Their first derivatives in the direction d are not equal. Thus every sufficiently small nonzero ϵ is genuinely non-reflected, and translation, rescaling, sign reversal, or relabeling cannot restore reflection.

G.2 Absolutely continuous bounded transition laws

Choose $\eta > 0$ sufficiently small and replace cell ℓ by the uniform interval of half-width η centered at $(1 - \eta)y_\ell$. The intervals remain disjoint and lie in $[-1, 1]$. Use the same conditional uniform density inside a cell under both hidden signs. Consequently the pointwise LR is exactly $\log r_\ell$ throughout cell ℓ and all Chernoff coordinates above remain exact.

Equivalently, if $V \sim \mathrm{Unif}[-1, 1]$ is independent, then conditional on the cell,

$$W = (1 - \eta)y_\ell + \eta V.$$

Therefore

$$\mathbb{E}W = (1 - \eta)\mathbb{E}Y, \qquad \mathbb{E}W^2 = (1 - \eta)^2\mathbb{E}Y^2 + \eta^2/3,$$

so endpoint equality of the first two moments survives exactly. The local positive-sign MGF is

$$M_{q,y,\eta}(z) = \frac{\sinh(\eta z)}{\eta z} \sum_\ell r_\ell q_\ell e^{(1-\eta)y_\ell z},$$

with the continuous extension 1 at $z = 0$.

G.3 Uniform stability of the all-horizon rate certificate

The proof of Theorem 3 establishes its all-H conclusion through two strict inequalities on compact parameter domains: a common negative/positive derivative bracket for every Cramér minimizer and a strict sign for the matched-tangent derivative of the conditional rollout MGF. These inequalities have positive margins; see the exact polynomial bounds in the preceding appendix sections and the deterministic certificate.

The functions appearing in those inequalities depend continuously, in the uniform C^1 topology on the relevant compact z interval, on (y, η, q, v). The fixed-minor construction makes $q(y)$ and $v(y)$ continuous. Hence there is a neighborhood $\mathcal{U}$ of the explicit point $(y_0, 0, q_0, v_0)$ in which both strict inequalities remain valid simultaneously. This is a uniform statement before the horizon enters the action-count mixture, so it preserves the same minimizer bracket and the sign

$$DI_H(q)[v] > 0$$

for every integer $H \geq 2$. Choosing any sufficiently small nonzero asymmetric ϵ and positive η whose induced tuple lies in $\mathcal{U}$ gives a fixed non-reflected absolutely continuous local family with exponent non-identifiability at every finite horizon. Numerical values in the supplement stress this open-neighborhood theorem but are not used to infer the quantifier over H.

G.4 A genuinely two-dimensional state/history-dependent policy

Let C_1 be an independent Rademacher state coordinate and, for $\rho > 0$, define

$$\phi_t(C_t) = (1, \rho C_t), \qquad \pi_\theta(A_t = 1 \mid C_t) = \sigma(\theta^\top \phi_t(C_t)), \qquad C_{t+1} = C_t U_t.$$

In parallel, use the same sequential transition $X_t = X_{t-1} + U_t W_t/T$ and reward $R = X_T$. Reachable logits are $\theta_1 \pm \rho\theta_2$, so the reachable-policy Jacobian has rank two. At $\theta = 0$, the trajectory Fisher is

$$F_H(0) = \frac{T}{4}\,\mathrm{diag}(1, \rho^2),$$

which is positive definite. Scores are not confined to a fixed line, and for $T \geq 2$ the random action U_t changes C_{t+1} and therefore future policy features. Thus this is neither a rotation of a rank-one policy nor a direct product of independent scalar gadgets.

The action/context path law is endpoint independent. Conditional on such a path,

$$\mathbb{E}[R \mid A, C] = \frac{\mu}{T}\sum_t U_t,$$

and $\mathbb{E}[R^2 \mid A, C]$ is determined only by the common local (μ, m_2). Hence $J(\theta)$, $\nabla J(\theta)$, $\mathbb{E}[\widehat{g}\widehat{g}^\top]$, $\mathrm{Cov}(\widehat{g})$, score law, Fisher geometry, and the first two reward moments are identical across endpoints for every finite θ.

At $\theta = 0$, the C_1 symmetry gives

$$\nabla J(0) = (\mu/2, 0).$$

Consequently $\langle \widehat{g}, \nabla J(0)\rangle$ is a positive multiple of the first-coordinate statistic $R\sum_t U_t/2$, exactly the directional variable covered by the preceding all-H rate certificate. Combining this policy with the asymmetric continuous matched environment proves Theorem 4.

H Estimator-information hierarchy and non-oracle certificates

H.1 Data processing from trajectories to the algorithmic statistic

Let P_+, P_- be hidden-sign trajectory laws and let $G = T(\tau)$ be any measurable scalar statistic of trajectory τ, with induced laws Q_+, Q_-. For $a \in (0, 1)$ and any common dominating measure ξ, define the Chernoff affinity

$$A_a(P_+, P_-) = \int p_+^a p_-^{1-a}\, d\xi.$$

A deterministic statistic is a Markov kernel, so data processing for Rényi/f-divergence affinities gives $A_a(Q_+, Q_-) \geq A_a(P_+, P_-)$. Taking negative logs and optimizing over a yields

$$C(G) \leq C(\tau).$$

For n i.i.d. copies of G, let $e_n^*(G)$ be the equal-prior Bayes error of the optimal test and $e_n^{\mathrm{mean}}(G)$ the error of the fixed sign-of-empirical-mean rule with the paper's tie convention. For every finite n, $e_n^{\mathrm{mean}}(G) \geq e_n^*(G)$. Chernoff's theorem gives $-n^{-1}\log e_n^*(G) \to C(G)$, hence

$$\limsup_{n\to\infty} -\frac{1}{n}\log e_n^{\mathrm{mean}}(G) \leq C(G) \leq C(X).$$

If the empirical-mean limit exists we denote it by I_{mean}. In the reflected finite-support class, Section I proves that the first inequality is equality iff the induced LR is linear in G. If the trajectory LR is measurable with respect to G, then G is sufficient for this binary experiment and $C(G) = C(X)$.

H.2 From a confidence ball to a direction cone

Let $\mu \neq 0$ and suppose a random Euclidean ball

$$C = \{v : \|v - c\| \leq r\}$$

contains μ. If $\|c\| \leq r$, the ball contains the origin and gives no nontrivial uniform directional information, so define $\mathcal{U}$ to be the whole unit sphere. Suppose instead that $\|c\| > r$. Elementary tangent geometry shows that every direction $u = v/\|v\|$ with $v \in C$ lies in the spherical cap

$$\mathcal{U} = \{u : \|u\| = 1,\ \angle(u, u_0) \leq \theta\}, \qquad u_0 := \frac{c}{\|c\|}, \quad \theta := \arcsin\frac{r}{\|c\|}. \tag{48}$$

Indeed, the maximum angle is achieved by a ray tangent to the ball, giving a right triangle with hypotenuse $\|c\|$ and opposite side r.

For any x, define $Z_{\mathcal{U}}(x) = \inf_{u\in\mathcal{U}}\langle x, u\rangle$. Write

$$a = \langle x, u_0\rangle, \qquad b = \|x - au_0\|.$$

When $\|c\| > r$, minimizing over the cap reduces to minimizing $a\cos\phi - b\sin\phi$ over $\phi \in [0, \theta]$. Therefore

$$Z_{\mathcal{U}}(x) = \begin{cases} -\|x\|, & a < 0 \text{ and } b \le (-a)\tan\theta, \\ a\cos\theta - b\sin\theta, & \text{otherwise.} \end{cases} \tag{49}$$

For the whole sphere, $Z_{\mathcal{U}}(x) = -\|x\|$. In all cases, if $\|x\| \le B$ then $|Z_{\mathcal{U}}(x)| \le B$.

If the true direction $u_\star = \mu/\|\mu\|$ belongs to $\mathcal{U}$, then pointwise

$$Z_{\mathcal{U}}(x) \le \langle x, u_\star\rangle. \tag{50}$$

Averaging preserves the inequality. Hence

$$\left\{\langle \bar{X}_n, \mu\rangle \le 0\right\} = \left\{\langle \bar{X}_n, u_\star\rangle \le 0\right\} \subseteq \left\{\frac{1}{n}\sum_{i=1}^{n} Z_{\mathcal{U}}(X_i) \le 0\right\}. \tag{51}$$

This deterministic containment is the key step that removes the oracle direction.

H.3 A DIMENSION-FREE EMPIRICAL-BERNSTEIN CONFIDENCE BALL

Our theorem accepts any valid confidence set. For the experiments we instantiate the Euclidean ($D = 1$) empirical-Bernstein construction of Martinez-Taboada & Ramdas (2026). If $\|X_i\| \le B$, let predictable weights $w_i \in (0, 0.8]$, weighted centers

$$\bar{\mu}_0 = 0, \qquad \bar{\mu}_i = \frac{\sum_{j=1}^{i} w_j X_j}{\sum_{j=1}^{i} w_j},$$

and $\psi_E(w) = -\log(1-w) - w$. Their Corollary 1 implies that, with probability at least $1 - \delta_1$, simultaneously in t,

$$\|\bar{\mu}_t - \mu\| \le \frac{\frac{1}{4B}\sum_{i=1}^{t}\psi_E(w_i)\|X_i - \bar{\mu}_{i-1}\|^2 + 4B\log(2/\delta_1)}{\sum_{i=1}^{t} w_i}. \tag{52}$$

For a fixed pilot size m_1 we use their Section 4.1 tuning

$$w_i = \min\left\{\sqrt{\frac{2(4B)^2\log(2/\delta_1)}{\widehat{\sigma}_{i-1}^2 m_1}}, \frac{1}{2}\right\}, \qquad \widehat{\sigma}_i^2 = \frac{B^2/4 + \sum_{j=1}^{i}\|X_j - \bar{\mu}_{j-1}\|^2}{i+1}.$$

This is not our concentration result: Martinez-Taboada & Ramdas (2026) prove the confidence set and show that its leading width matches oracle Bernstein in Euclidean/Hilbert spaces. Our use of it is solely to obtain a data-dependent cone for the unknown PG direction.

H.4 PROOF OF THEOREM 5

Condition on the first split S_1. The ball C_1, cone $\mathcal{U}_1$, robust statistic Z_1, and tilt $\widehat{\lambda} \le 0$ are now fixed. On the event $E_1 = \{\mu \in C_1\}$, $u_\star \in \mathcal{U}_1$, so event containment (51) holds for every independent future batch size.

Use the independent second split $X'_1, \dots, X'_{m_2}$ and define

$$W_j = e^{\widehat{\lambda} Z_1(X'_j)}, \qquad a = e^{-|\widehat{\lambda}|B}, \qquad b = e^{|\widehat{\lambda}|B}.$$

Conditional on S_1, the W_j are i.i.d. in $[a, b]$. Let $\widehat{M}$ be their sample mean and $\widehat{V}$ their Bessel-corrected sample variance. A one-sided empirical-Bernstein inequality (Maurer & Pontil, 2009) gives, with conditional probability at least $1 - \delta_2$,

$$M_{\widehat{\lambda}} := \mathbb{E}[W \mid S_1] \le U := \min\left\{b, \widehat{M} + \sqrt{\frac{2\widehat{V}\log(2/\delta_2)}{m_2}} + \frac{7(b-a)\log(2/\delta_2)}{3(m_2-1)}\right\}. \tag{53}$$

The crucial point is that $\widehat{\lambda}$ is S_1-measurable, so no simultaneous post-selection bound over a CGF grid is required on S_2. In the implementation, split 1 chooses $\widehat{\lambda}$ by minimizing its empirical robust log-MGF over a fixed negative grid; split 2 alone validates the selected point.

On E_1 and the MGF event E_2, for any future independent batch,

$$\mathbb{P}(\langle \bar{X}_n, \mu\rangle \le 0 \mid S_1, S_2) \le \mathbb{P}\left(\sum_{i=1}^{n} Z_1(X_i) \le 0 \mid S_1\right) \le M_{\widehat{\lambda}}^n \le U^n.$$

The middle inequality is Markov/Chernoff because $\widehat{\lambda} \le 0$. Setting $I_{\mathrm{cert}} = [-\log U]_+$ proves (8). Finally, $\mathbb{P}(E_1 \cap E_2) \ge 1 - \delta_1 - \delta_2$ by conditioning and a union bound.

An oracle-direction scalar certificate is recovered as the special case $\mathcal{U}_1 = \{\mu/\|\mu\|\}$; we retain it only as a diagnostic benchmark, not as the operational guarantee.

H.5 Unbounded-support alternative under a known tail envelope

The deterministic support bound in Theorem 5 is one sufficient way to obtain finite-sample validity, but it is not logically necessary. The following conservative alternative replaces support by an externally justified joint tail model. It is not distribution-free and is not claimed as a new concentration inequality.

Proposition 2 (Non-oracle directional certificate under a sub-exponential vector envelope)**.** *Let $X \in \mathbb{R}^d$ be i.i.d. with mean $\mu \ne 0$. Suppose there are known $\nu > 0$ and $b \ge 0$ such that for every unit vector v and every $|\lambda| < 1/b$ (with the usual interpretation when $b = 0$),*

$$\log \mathbb{E} e^{\lambda\langle v, X-\mu\rangle} \le \frac{\nu^2\lambda^2}{2(1-b|\lambda|)}.$$

From an independent pilot $X_1, \dots, X_m$, let $c = m^{-1}\sum_i X_i$. If r is chosen so that

$$2\,5^d \exp\left[-\frac{mr^2}{8(\nu^2 + br/2)}\right] \le \delta, \qquad L = (\|c\| - r)_+,$$

then with pilot probability at least $1-\delta$, every independent future batch satisfies

$$\mathbb{P}\big(\langle \bar{X}_n, \mu\rangle \le 0 \mid X_{1:m}\big) \le \exp\left[-\frac{nL^2}{2(\nu^2 + bL)}\right], \qquad n \ge 1. \tag{54}$$

For the sub-Gaussian special case $b = 0$, one may take

$$r = \nu\sqrt{\frac{8\{d\log 5 + \log(2/\delta)\}}{m}}.$$

Proof. A Euclidean $1/2$-net of the unit sphere has cardinality at most 5^d. If $\|c - \mu\| \ge r$, one net direction has projection at least $r/2$. The assumed directional Bernstein envelope plus a union bound gives the displayed pilot failure probability. On the resulting event, $L \le \|\mu\|$. For $u = \mu/\|\mu\|$, wrong direction implies $\langle \bar{X}_n - \mu, u\rangle \le -\|\mu\| \le -L$. Applying the same one-dimensional envelope in direction u and optimizing the Chernoff parameter yields (54). The map $x \mapsto x^2/(\nu^2 + bx)$ is increasing on $x \ge 0$, so replacing $\|\mu\|$ by L is conservative. □

The proposition removes hard support at the cost of a stronger parametric tail assumption and a dimension-dependent pilot radius. We do not claim that unknown (ν, b) can be upper-certified from arbitrary finite data without additional assumptions; fitting them empirically would produce a diagnostic, not the guarantee above.

H.6 When a Gaussian approximation is justified

The impossibility results concern universal identification, not the usefulness of Gaussian approximations under additional regularity. Let the scalar directional variable Z have mean $\mu_Z > 0$, variance $\sigma_Z^2 > 0$, and $\rho = \mathbb{E}|Z - \mu_Z|^3 < \infty$. Let Φ denote the standard normal cumulative distribution function. The classical Berry–Esseen theorem (Berry, 1941; Esseen, 1942) gives a universal constant C_{BE} such that

$$\left|\mathbb{P}(\bar{Z}_n \le 0) - \Phi\left(-\frac{\sqrt{n}\,\mu_Z}{\sigma_Z}\right)\right| \le \frac{C_{\mathrm{BE}}\rho}{\sigma_Z^3\sqrt{n}}. \tag{55}$$

Table 4: Non-oracle future-risk certificate over 180 MDPs (540 estimator cases). Medians involving I_{cert} are over positive certificates only.

pilot split	positive	median I_{cert}/I	median I_{cert}	median $n_{1\%}$
10k/10k	69.63%	.186	.0231	200
20k/10k	76.48%	.227	.0255	181

For atom-free Z, strict and weak zero-threshold events coincide. Thus covariance alone does not universally identify reliability, while covariance together with a controlled standardized third absolute moment gives an explicit regime in which the Gaussian approximation is valid. The five-seed PPO-trained CartPole audit in Appendix H.14 is designed around this distinction rather than around a claim of ubiquitous Gaussian failure.

Proposition 3 (Zero-extra-pilot adaptive direction certificate)**.** *Let $(C_t)_{t\geq 1}$ be any confidence sequence for μ satisfying $\mathbb{P}(\forall t : \mu \in C_t) \geq 1-\delta$. At any data-dependent time τ, certify only if $\inf_{v\in C_\tau}\langle \bar{X}_\tau, v\rangle > 0$. Then*

$$\mathbb{P}\big(\exists \tau : \text{ certificate issued and } \langle \bar{X}_\tau, \mu\rangle \leq 0\big) \leq \delta.$$

The same holds over finitely many predeclared checkpoints using simultaneous confidence sets with total error budget δ.

H.7 Proof of Proposition 3

Let $E = \{\mu \in C_t \ \forall t\}$, which has probability at least $1-\delta$. If the rule certifies at any (possibly data-dependent) time τ, then on E

$$\langle \bar{X}_\tau, \mu\rangle \geq \inf_{v\in C_\tau} \langle \bar{X}_\tau, v\rangle > 0.$$

Therefore the event "certificate issued and true direction nonpositive" is contained in E^c and has probability at most δ. For finitely many predeclared checkpoints one may instead construct separate $(1-\delta_k)$ confidence balls with $\sum_k \delta_k \leq \delta$ and apply the same argument on their simultaneous coverage event. The experiment uses this finite-checkpoint form so that every tested stopping batch is explicit.

H.8 Non-oracle future-risk experiments and pilot economics

All certification inputs are sampled gradient vectors. Exact trajectory enumeration is used only after certificate construction to compute ground-truth μ and the exact Cramér exponent for evaluation. Because the tabular models have finite trajectory support, we use the exact deterministic support maximum $B = \max \|X\|$ as the required bound; in a model-free application a known analytical reward/score bound can replace it, potentially more conservatively. A sample maximum would not be valid.

Table 4 summarizes the complete 180-MDP panels. With a 10k/10k split, 376 of 540 cases (69.63%) have positive certificates; all 540 sampled confidence balls contain the enumerated gradient and all 540 certificates lie below the oracle exponent. The latter frequencies are falsification checks, not substitutes for theorem coverage.

The pre-specified 3×20-MDP pilot-economics subset gives Table 5. The conclusion is deliberately not "20k is cheap": the bound is conservative, and the intended use is fixed-policy auditing/batch planning when such a pilot can be amortized at a checkpoint.

H.9 Adaptive realized-update experiments

We test predeclared checkpoints $64, 128, 256, \ldots, 32768$ with a simultaneous 95% confidence budget. In panel seed 20260830 all 180 estimator cases certify, no false certificate is observed, and the stopping-count distribution is

$$(512:5,\ 1024:40,\ 2048:87,\ 4096:30,\ 8192:11,\ 16384:6,\ 32768:1).$$

Table 5: Pilot economics on the three-panel sensitivity subset.

total pilot	positive	median I_{cert}/I	median $n_{1\%}$
5k	16.1%	.033	549
10k	47.8%	.074	454
15k	60.0%	.118	294
20k	71.7%	.154	257
30k	80.6%	.215	209

Table 6: Sensitivity of the non-oracle future-risk certificate to deterministic-bound inflation. "Cone usable" means $\|c\| > r$ after the vector confidence step. Certificate medians are over positive certificates only. All 180 confidence balls cover their enumerated reference gradients and every produced certificate is below the exact oracle exponent at every multiplier.

bound	positive cert.	cone usable	median $r/\|\mu\|$	median I_{cert}/I	median $n_{1\%}$
$1B_\star$	59.4%	98.9%	.167	.153	266
$2B_\star$	18.9%	97.2%	.221	.127	172
$5B_\star$	1.1%	90.0%	.475	.078	175
$10B_\star$	0.0%	55.0%	.934	—	—
$20B_\star$	0.0%	9.4%	1.860	—	—

In independent panel 20270830 all 180 cases again certify with zero observed false certificates, with

$$(512 : 6,\ 1024 : 44,\ 2048 : 88,\ 4096 : 36,\ 8192 : 6).$$

Both panels have median stopping batch 2048. A separate three-panel 20-MDP sensitivity subset also certifies all 180 cases with zero observed false certificates and median 2048. These observations assess conservatism/sample economics; Proposition 3 supplies the statistical guarantee.

H.10 Deterministic-envelope sensitivity and where conservatism enters

Theorem 5 requires a deterministic norm envelope B valid for future trajectory-gradient vectors. To measure the price of looseness rather than treating B as an innocuous constant, we rerun the same 10k/10k non-oracle certificate on a pre-specified three-panel subset containing 60 MDPs and all three estimator variants (180 cases total). We replace the tight support bound $B_\star$ by $mB_\star$ for $m \in \{1, 2, 5, 10, 20\}$ everywhere the validity proof requires the envelope. Table 6 reports the complete result.

The early collapse is not explained only by direction-cone geometry. At $2B_\star$, 97.2% of confidence balls still exclude the origin and therefore define nontrivial direction cones, yet only 18.9% of complete certificates remain positive. Inflation of B also widens the deterministic range $[e^{-|\lambda|B}, e^{|\lambda|B}]$ used by the one-point MGF confidence bound in (53); this scalar validation step is therefore a major source of conservatism. At larger multipliers both mechanisms matter. We do not replace B by the observed sample maximum, because that would not preserve the theorem's future-sample guarantee. Figure 1 visualizes this bound-sensitivity effect together with the separate high-dimensional matched-second-order stress test.

H.11 A 65-dimensional continuous-context matched-second-order stress test

The tabular experiments are deliberately small enough for exact enumeration. We therefore add a separate controlled stress family in which exact trajectory enumeration is unavailable. Contexts are sampled continuously as $x \sim \mathrm{Unif}([-1, 1]^8)$. A fixed nonlinear random-feature map $\phi(x) \in \mathbb{R}^{65}$ consists of 64 tanh features plus a bias coordinate and is normalized to unit Euclidean norm. Only the 65-dimensional Bernoulli-logit output parameter is trainable:

$$p_\theta(x) = \sigma(\theta^\top \phi(x)), \qquad X = (A - p_\theta(x))R\,\phi(x).$$

For a bounded advantage signal $a(x) > 0$, reward is

$$R = (2A - 1)\frac{a(x)}{2} + \xi,$$

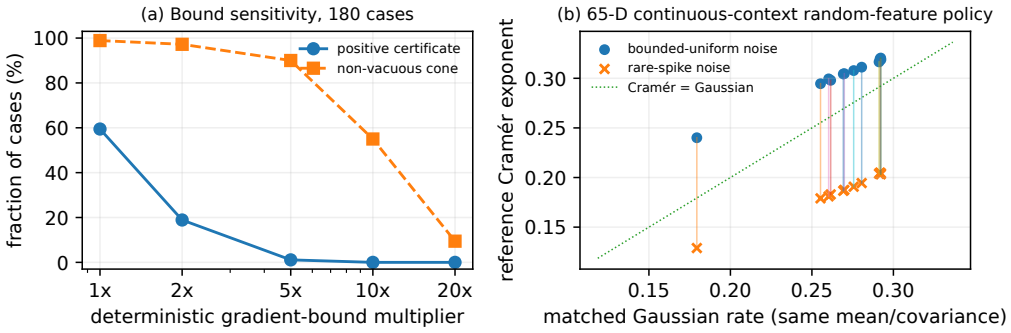

Figure 1: **Controlled stress diagnostics retained in the appendix.** Left: non-oracle certificate non-vacuity versus deterministic-bound inflation. Right: 65-dimensional continuous-context paired laws with exactly matched PG mean/covariance but separated reference Cramér exponents.

Table 7: Continuous-context 65-dimensional random-feature-policy stress test. Within each of the 12 pairs, PG mean/covariance and the Gaussian second-order rate are equal analytically by (57). Cramér rates are high-precision Monte Carlo references, not exact enumeration.

quantity	result
pre-specified matched pairs / cases	12 / 24
median benign/rare Cramér-exponent ratio	1.619
range of paired Cramér-exponent ratios	[1.553, 1.862]
median rare/benign strict-risk ratio at $n = 16$	4.966
paired Cramér 95% intervals disjoint	12 / 12
non-oracle certificate positive	24 / 24
median positive $I_{\mathrm{cert}}/I_{\mathrm{ref}}$	.514
certificate below reference exponent	24 / 24

where ξ is independent of (x, A) and has mean zero. Conditional on x,

$$\mathbb{E}[X \mid x] = p(1-p)a(x)\phi(x), \tag{56}$$

$$\mathbb{E}[XX^\top \mid x] = p(1-p)\left(\frac{a(x)^2}{4} + \mathrm{Var}(\xi)\right)\phi(x)\phi(x)^\top. \tag{57}$$

Thus two noise laws with the same first two moments produce *exactly* the same population PG mean and full one-sample second moment/covariance, as well as the same policy/Fisher geometry, for every fixed feature/policy seed.

For each of 12 pre-specified seeds we use two bounded noise laws with exactly common variance 0.0256: (i) centered uniform noise and (ii) a rare symmetric spike, taking ± 0.8 with probability 0.02 each and zero otherwise. Each case uses 500,000 projected-gradient reference samples to estimate the Cramér exponent, 100,000 independent future batches at $n = 16$, and a 10k/10k non-oracle certificate pilot. Write I_{ref} for the high-precision reference Cramér exponent. Table 7 reports only pre-specified aggregate statistics; no seed was selected after observing outcomes.

Cramér-reference uncertainty is audited with an independent 100k-sample MGF computation: the median estimated standard error of I is 0.00263, the maximum is 0.00322, and all 12 paired 95% intervals remain disjoint. The largest absolute difference between the 500k reference estimate and the independent recomputation is 0.00682. Because the second-order equality is analytic, small raw Monte Carlo differences in empirical Gaussian-rate estimates are sampling noise rather than an approximate-matching claim.

The experiment is intentionally a *stress test*: the rare-spike law is designed to alter higher-order tails while preserving second-order PG information. It therefore addresses survival of the mechanism outside tiny exact enumeration and at higher dimension, but it is not evidence that ordinary deep actor–critic training is generically non-Gaussian.

Table 8: Guarantee comparison with close wrong-direction/adaptive-sampling work. "Approx." denotes a CLT/Student-t or Gaussian modeling step rather than a distribution-free finite-sample guarantee for the exact future-batch risk in this paper.

Method	target	validity	assumptions / distinction
Boyles et al.; Welling; Korattikara et al.	update-direction reliability; adaptive subset	Approx.	CLT/Student-t testing; closest prior work in wrong-direction intent.
Byrd et al.; Bollapragada et al.	gradient accuracy / descent direction	finite-sample under stated tests	variance, norm, or inner-product sampling conditions; optimization rather than PG-risk identification.
Papini et al.	policy improvement	finite-sample	PG smoothness/concentration plus adaptive batch and step size; stronger objective event.
Aolaritei et al.	anytime stochastic-gradient-descent (SGD) stopping	time-uniform	observable suboptimality/optimizer-distance confidence sequences in convex/strongly-convex settings.
Gaussian / split t	direction evidence or future-risk prediction	Approx.	first/second-order modeling; empirically useful, not a universal identifier.
Ours: future risk	every independent future-batch wrong-direction risk	**finite-sample**	no true-gradient direction supplied; bounded fixed-policy i.i.d. gradients and independent validation split.
Ours: adaptive direction	realized update direction at stopping time	**time-uniform**	any valid vector confidence sequence; elementary consequence, not a new general stopping framework.

Table 9: Three-panel rate-diagnostic replication. Spearman correlations compare each predicted rate with observed $-\log \widehat{\mathcal{R}}_n/n$; MAE measures calibration error. C=Cramér, G=Gaussian, B=Bernstein.

n	ρ_C	ρ_G	ρ_B	MAE_C	MAE_G	MAE_B
4	.9891	.9947	.9738	.3124	.3251	.3830
8	.9972	.9977	.9781	.1775	.1902	.2481
16	.9992	.9973	.9785	.1015	.1142	.1721

H.12 Closest operational baselines and guarantee boundaries

Table 8 compares the target of each method rather than treating unlike guarantees as a numerical leaderboard. Earlier stochastic-optimization work already makes stochastic-direction reliability operational; safe-PG work certifies a stronger policy-improvement event; recent confidence-sequence work gives anytime-valid optimization stopping rules. Our distinct positive object is a finite-sample upper bound on the *future independent-batch wrong-direction probability* without supplying the true gradient direction.

Numerical baseline diagnostic on the 65-dimensional panel. A split one-sided Student-t/CLT direction diagnostic accepts all 24 cases and its selected direction has median cosine 0.999974 with the high-precision reference gradient. At $n = 16$, a Gaussian plug-in has mean absolute error (MAE) 0.0527 against the observed finite-n log-risk rate, compared with 0.1126 for an oracle Gaussian rate and 0.1359 for the reference Cramér exponent. This is *not* evidence against the Cramér theory: I is the asymptotic exponent and need not minimize finite-n prediction error at $n = 16$. It is instead a useful negative result showing that Gaussian approximations can be excellent in controlled settings. The contribution of the Cramér/CGF layer is identification of distribution-sensitive tail geometry and a finite-sample non-oracle upper certificate, not universal finite-n MAE dominance.

H.13 Oracle rate diagnostics: useful for mechanism, not certification

The following experiment retains the true direction only to test whether large-deviation geometry predicts observed risk. It is not the operational certificate. Three independently seeded panels contain 180 random tabular MDPs and 540 estimator/environment cases per batch size across vanilla REINFORCE, an exact-value baseline, and generalized advantage estimation (GAE) with $\lambda = 0.95$. Figure 2 visualizes the rate diagnostic, and Table 9 reports its aggregate calibration and ranking statistics.

Monte-Carlo censoring at $n = 16$. Each case uses $100{,}000$ independent future batches. Exactly three of 540 cases have zero failures. Removing them gives Spearman correlations 0.99920 (Cramér), 0.99729 (Gaussian), and 0.97818 (Bernstein), with MAEs 0.10128, 0.11273, and 0.16935. Replacing the zero counts by conservative one-sided 95% Clopper–Pearson upper risks gives corre-

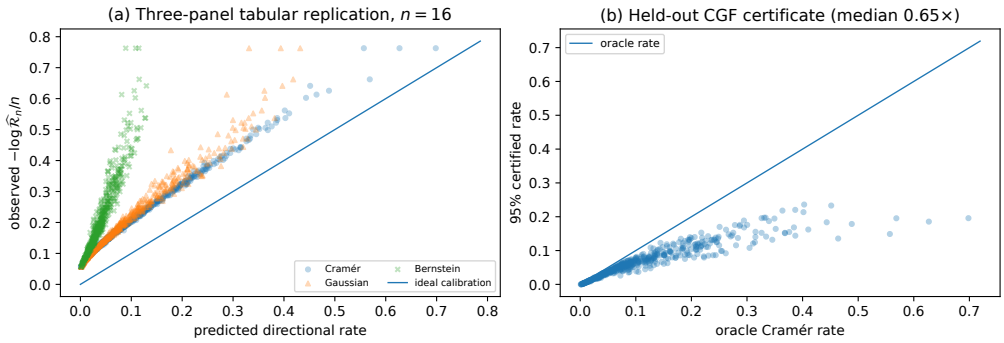

Figure 2: **Oracle-direction rate diagnostic.** At $n = 16$ (left), the exact Cramér rate, Gaussian second-order rate, and rigorous Bernstein mean/variance/range rate are compared with observed finite-batch directional rate. Cramér is best calibrated; Gaussian ranking is nevertheless already strong. Right: the old oracle-projected held-out CGF certificate is shown only as a diagnostic baseline. The operational non-oracle certificate is Theorem 5.

lations 0.99921, 0.99733, and 0.97848, with MAEs 0.10102, 0.11358, and 0.17144. The headline comparison is therefore not a Monte-Carlo-floor artifact.

Gaussian-neighborhood stress test. Among same-estimator/same-regime $n = 16$ cases with relative Gaussian-rate gap at most 1%, 150 pairs qualify; their median exact-rate ratio is 1.073, 95th percentile 1.215, and maximum 1.670. The maximizing pair has Gaussian rates 0.3314 and 0.3284 but exact Cramér rates 0.5570 and 0.3335. This descriptive post-replication analysis shows that strong global Gaussian ranking does not imply identification by second-order information.

H.14 Five-seed PPO-trained CartPole checkpoint audit

This experiment tests external validity without selecting environments or checkpoints for tail behavior. We use a dependency-free vectorized implementation of the standard CartPole-v1 dynamics and train five independent categorical policies with PPO. The policy network is $4 \to \tanh(32) \to 2$, giving 226 trainable policy parameters. Training uses 64 vectorized environments, GAE(0.95), the standard clipped PPO surrogate, and checkpoints fixed in advance at updates $0, 5, 15, 30$ for seeds $100, \ldots, 104$.

What is audited. At each frozen checkpoint we do *not* reuse PPO minibatches. We draw fresh complete episodes and form the unbiased score-function gradient for the undiscounted episodic-return objective, using the learned state-value function only as an action-independent control variate. Thus the audit targets a standard frozen-policy PG object covered by our i.i.d. theory, not PPO's multi-epoch clipped-surrogate optimization. Twenty thousand independent episodes (four shards of 5000) define the Monte-Carlo reference mean $g_\star$ and another disjoint 20k episodes define the audit gradient law. The cosine between the first reference shard mean and the remaining reference-shard mean has minimum 0.99758 and median 0.99950 over the 20 checkpoints, supporting directional stability of $g_\star$ without treating it as exact.

For audit gradients G_i we set $Z_i = \langle G_i, g_\star/\|g_\star\|\rangle$. For $n \in \{1, 2, 4, 8, 16, 32, 64\}$, 10^5 Monte-Carlo batches are generated by sampling with replacement from the independent 20k-trajectory audit law. We restrict confirmatory interpretation to $n \le 32$. The reported Clopper–Pearson intervals are therefore exact for the conditional Monte-Carlo resampling experiment; they do *not* include uncertainty from replacing the population law by the finite 20k-trajectory audit law. A zero failure count is never reported as zero probability: for logarithms/plots we use the fixed continuity value $(k + 1/2)/(N + 1)$, while the count interval itself remains Clopper–Pearson. Table 10 summarizes the resulting checkpoint-level Gaussian-calibration discrepancies.

The enormous maxima are not headline evidence: one $n = 16$ checkpoint has zero failures among the 10^5 conditional resampling batches. Excluding that point leaves the $n = 16$ median discrepancy

Table 10: Five-seed CartPole audit over 20 pre-specified PPO-trained checkpoints. "Factor" is the median of $\max\{\widehat{R}_n/R_n^{\mathrm{G}}, R_n^{\mathrm{G}}/\widehat{R}_n\}$ across checkpoints; the last column is the fraction for which the Gaussian plug-in lies outside the conditional resampling 95% interval.

n	median factor	maximum factor	outside cond. 95% interval
1	1.07	2.08	1.00
2	1.08	3.49	.90
4	1.27	9.66	.85
8	1.72	76.46	.95
16	2.78	3189.85	1.00
32	7.20	240.97	1.00

at $2.55\times$, so the aggregate conclusion does not depend on the continuity convention. From $n = 4$ onward the Gaussian plug-in is predominantly conservative (18/20 checkpoints at $n = 4$, 19/20 at $n = 8$, and 20/20 at $n = 16, 32$). Median projected skewness across seeds decreases monotonically across the pre-specified training stages: $4.48, 4.07, 2.72, 1.99$ at updates $0, 5, 15, 30$.

Predeclared matched-checkpoint search. Before inspecting risk ratios, we declared a pair eligible when the relative differences in Gaussian rate, projected mean, and projected variance are at most 5%, 15%, and 15%, respectively. Among all $\binom{20}{2} = 190$ pairs, only two qualify. Their maximum risk ratio is 1.30 at $n = 16$ and 1.76 at $n = 32$; no qualifying pair reaches $2\times$. We therefore do not use this experiment to claim a natural approximately matched analogue of the exact construction.

Reliability-decision diagnostic. Across the 120 checkpoint/batch cases with $n \le 32$, a Gaussian plug-in rule declares 46, 33, and 13 cases reliable at thresholds $\alpha = .10, .05, .01$, respectively; none has conditional empirical-law risk above the corresponding threshold. The analogous empirical-Cramér asymptotic diagnostic declares 33, 23, and 14 cases, again with no false declaration. These are descriptive diagnostics, not formal future-risk certificates. In particular, the CartPole experiment shows substantial probability miscalibration without showing that Gaussian screening is unsafe on this panel.

Batch-size interpretation. If two projected-gradient laws have Cramér exponents $I_1 > I_2 > 0$, the leading large-deviation approximation $R_n \asymp e^{-nI}$ implies an equal-target batch-size ratio $n_2/n_1 \approx I_1/I_2$. Thus the median $1.62\times$ exponent ratio in the separate 65-dimensional matched-second-order stress test corresponds asymptotically to about a $1.62\times$ batch-size multiplier for equal directional reliability. This translation is asymptotic and is not asserted for small n without finite-n correction.

H.15 Behavior-point PPO identity

Let $r_\theta = \pi_\theta(a \mid s)/\pi_{\theta_{\mathrm{old}}}(a \mid s)$ and treat the advantage as stop-gradient. At $\theta = \theta_{\mathrm{old}}$, $r_\theta = 1$ lies in the interior of the clipping interval, so

$$\nabla_\theta L_{\mathrm{PPO}}\big|_{\theta_{\mathrm{old}}} = \widehat{A}\,\nabla \log \pi_{\theta_{\mathrm{old}}}(a \mid s).$$

Thus the GAE experiment also describes the first PPO clipped-surrogate minibatch gradient at the behavior checkpoint. No multi-epoch PPO claim follows.

I LR alignment and estimator efficiency

Let P be full-support on a finite symmetric set, $Q(w) = P(-w)$, and define

$$\beta = \sum_w \sqrt{P(w)Q(w)}, \qquad C = -\log \beta,$$

$$M_P(\lambda) := \mathbb{E}_P e^{\lambda W}, \qquad I_W(0) := -\log \inf_\lambda M_P(\lambda).$$

The normalized Hellinger midpoint

$$H(w) = \frac{\sqrt{P(w)Q(w)}}{\beta}$$

is symmetric, hence mean zero. If $\lambda_\star$ minimizes the MGF and

$$R_\star(w) = \frac{P(w)e^{\lambda_\star w}}{M_P(\lambda_\star)},$$

then $\mathbb{E}_{R_\star} W = 0$, and for any mean-zero R,

$$D(R\|P) = D(R\|R_\star) - \log M_P(\lambda_\star) \geq I_W(0).$$

Thus $I_W(0)$ is the KL projection value onto the mean-zero affine set. Since H is feasible and symmetry cancels the antisymmetric LR term,

$$D(H\|P) = C,$$

so $I_W(0) \leq C$.

Equality forces $H = R_\star$ by strict convexity of KL. Comparing their formulas yields

$$\frac{1}{2}\log\frac{Q(w)}{P(w)} = c + \lambda w.$$

Antisymmetry forces $c = 0$, hence

$$\log\frac{P(w)}{Q(w)} = -2\lambda w = \kappa w$$

with $\kappa > 0$. Conversely, if the LR is κw, the Hellinger midpoint equals the exponential tilt at $-\kappa/2$, so $I_W(0) = C$. This proves the reflected-class equality statement in Proposition 1.

J Common-jitter strict-risk robustness

The common-jitter lemma is included because it independently shows that fair-tie lattice examples can be interpreted as literal strict-risk limits. If a scalar one-step PG sample G is replaced by $G + \varepsilon Z$ with a common symmetric atom-free nuisance Z independent of the endpoint and hidden sign, then ties vanish. When the original n-sample sum is separated from zero except on exact ties and ε is sufficiently small relative to the nonzero lattice gap, strict risk equals the original fair-tie risk exactly after averaging over the symmetric jitter. This lemma is not used by the atom-free headline phase theorem or by the odd-batch natural theorem.

K Full-rank vector lift

Let the scalar endpoint law be G with common mean $\mu > 0$ and common second moment m_2. In a d-context MDP, draw a context C uniformly from $\{1, \ldots, d\}$ and let only the corresponding logit parameter control the binary action. On context 1, use the scalar hard instance; on the remaining contexts use a common bounded symmetric construction chosen so that the coordinate second moments match. The resulting one-trajectory estimator has

$$\mathbb{E}\widehat{g} = \frac{\mu}{d}e_1,$$

and the policy Fisher is

$$F_\pi = \frac{1}{4d}I_d.$$

With the common orthogonal noise choice used in the supplement, the covariance is a positive scalar multiple of I_d, hence full rank with condition number one. Conditional on the number $K \sim \mathrm{Binomial}(n, 1/d)$ of informative-context trajectories, the directional error is exactly the scalar K-sample error, so

$$\mathcal{R}_{i,n}^{(d)} = \sum_{k=0}^{n}\binom{n}{k}d^{-k}(1-d^{-1})^{n-k}\mathcal{R}_{i,k}.$$

An orthogonal reparameterization maps e_1 to any prescribed unit vector u without changing Fisher and covariance eigenvalues. Choosing dense u makes every mean-gradient coordinate nonzero. The target remains the one-dimensional directional event $\langle \bar{g}, g\rangle < 0$; we do not claim a d-hypothesis testing problem.

L CONTROLLED WINDY LINEWORLD FAMILY

The controlled family uses

$$x_{t+1} = x_t + U_t W_t, \qquad R = x_2,$$

with two ordinary Bernoulli-logit actions. All 41 transition-noise laws share

$$\mathbb{E}W = 3/20, \qquad \mathbb{E}W^2 = 1/10,$$

while higher-order probabilities vary along an exact moment-preserving family. Hence the objective, gradient, trajectory Fisher, one-step PG mean/variance, and complete REINFORCE second-moment function are identical throughout the family. The scripts `run_windy_family_study.py` and `run_windy_family_study_fast.py` enumerate the exact one-step gradient distributions and exact strict batch risks by convolution; optimization trajectories use fixed seeds and the same step size across all instances.

For the final 41-instance run, exact strict risks span

$$\begin{aligned} [0.0235062, 0.0487101] \quad & (n = 10), \\ [0.00484190, 0.0136342] \quad & (n = 20), \\ [0.000939536, 0.00399910] \quad & (n = 30). \end{aligned}$$

The Gaussian/signal-to-noise-ratio (SNR) proxy is constant across the family because the first two moments are fixed. The Cramér rate, which uses the complete one-step MGF, tracks $-\log \mathcal{R}_n^{\mathrm{str}}/n$ with correlations at least 0.9993 in these three batch sizes. A fixed-learning-rate 30-update simulation gives a target-success spread of about 4.3 percentage points. These experiments test a theorem-derived prediction; they are not used to establish the theorems.

M PLUG-IN RELIABILITY DIAGNOSTIC

Let P be a one-step directional-gradient law and let P^{*n} denote its n-fold convolution. Strict risk is the probability assigned by P^{*n} to $(-\infty, 0)$. For any P, Q,

$$|\mathcal{R}_n^{\mathrm{str}}(P) - \mathcal{R}_n^{\mathrm{str}}(Q)| \le \mathrm{TV}(P^{\otimes n}, Q^{\otimes n}) \le 1 - (1 - \mathrm{TV}(P, Q))^n \le n\mathrm{TV}(P, Q).$$

For a known k-point support, the empirical histogram $\widehat{P}_m$ satisfies the Weissman et al. bound

$$\mathbb{P}\Big(\|\widehat{P}_m - P\|_1 > \epsilon_1\Big) \le (2^k - 2)e^{-m\epsilon_1^2/2}.$$

Since TV is half the L_1 distance, a sufficient calibration-size bound follows after a conservative constant simplification. The implementation performs exact empirical convolution, not a Gaussian approximation. The calibration study reports absolute error distributions over fixed-seed repeated calibration samples and compares them with the common Gaussian/SNR proxy.

N NUMERICAL PHASE-DIAGRAM SEARCH PROTOCOL

The phase-diagram computation is an adversarial numerical search, not a proof. For each (C_0, B, n), it discretizes symmetric tilted LR magnitudes on a 28-point grid. Every two-support candidate and 500 fixed-seed random three-support candidates are solved so that $\mathbb{E}\cosh(L/2) = e^{C_0}$ exactly on the grid. FFT convolution computes H_n and the largest/smallest discovered values. The full configuration contains 90 cells: three Chernoff levels, six LR radii per level, and five batch sizes. Intermediate results are cached for resume. The scripts save CSV, PDF, and PNG outputs and record the ratio to the theoretical phase-shape proxy. Because the search is restricted to low-support grid laws, failure to attain the analytic constants is not evidence against the theorem; its purpose is to test the predicted two-branch dependence without selecting only the analytic extremizers.

O DETAILED RELATED-WORK POSITIONING

Table 11 expands the concise comparison in Section 2. The distinction is between improving or bounding a gradient estimator under an assumed stochastic model and asking whether a retained information class identifies the finite-batch directional tail in the first place.

Lineage and representative work	Main object / guarantee	What remains unresolved for our target
Score-function and efficient PG: Williams (1992); Greensmith et al. (2004); Schulman et al. (2015); Kallus & Uehara (2020); Huang & Jiang (2020)	Unbiasedness, variance reduction, or mean-square/asymptotic efficiency of gradient estimators	These criteria do not identify the exact finite-batch probability that the sampled gradient enters the population downhill half-space; our no-go persists after matching complete first- and second-order PG functions.
Safe/constrained and adaptive PG: García & Fernández (2015); Achiam et al. (2017); Papini et al. (2017; 2022)	Constraint/performance guarantees and adaptive step/batch selection under explicit regularity and concentration assumptions	Complementary constructive guarantees. We ask the prior identification question, then derive a separate future-batch certificate from validated one-sided information.
Non-Gaussian/directional gradients: Garg et al. (2021); Bernstein et al. (2018)	Empirical heavy-tail characterization or algorithms that emphasize gradient signs	Motivates directional reliability, but does not give exact matched REINFORCE models, a sharp ambiguity law, or sequential non-identification.
Large deviations: Bahadur & Ranga Rao (1960); Jongeneel et al. (2024)	Exponential rates for rare sample-mean or optimization events under a fixed data-generating model	We use large deviations comparatively: two matched sequential REINFORCE models retain different directional exponents.
Moment bounds, distributionally robust optimization (DRO), and partial identification: Bertsimas & Popescu (2005); Vandenberghe et al. (2007); Delage & Ye (2010); Manski (2003)	Feasible target sets, worst-case event probabilities, or decisions over ambiguity classes	Supplies the broad incomplete-information perspective. Our contribution is PG-specific exact matching, a likelihood-conditioned sharp ambiguity law, sequential persistence, and a constructive certificate.
Experiment comparison / information: Blackwell (1953); Chernoff (1952)	Distinguishability and ordering of full statistical experiments	We use these tools to distinguish environment-level information from the directional tail of the statistic actually averaged by the algorithm.

Table 11: Closest methodological lineages and the precise gap addressed here. The paper does not claim priority for the classical statistical tools themselves.

P CLAIM BOUNDARIES AND SCOPE NOTES

Theorem 2 is a statistical statement about sign-symmetric atom-free experiments with fixed (C_0, B). Corollary 1 proves the $\Theta(\sqrt{n})$ large-B branch after additionally fixing concrete PG geometry; we do not claim that those extra moment constraints leave the small-B branch unchanged.

Theorem 3 is different: its ordinary reward is not LR aligned, so environmental Chernoff information does not pin the empirical-gradient exponent. The same matched continuum works for every finite horizon. Theorem 4 shows that the all-H mechanism survives simultaneous removal of reflection symmetry, finite atomic noise, and rank-one/shared-scalar policy geometry, but the resulting class is still deliberately structured (binary actions, compact cell mixtures with a finite LR alphabet, designed finite-state history features, additive terminal reward); it is not a theorem about arbitrary MDPs. Its matched reward quantities are the first two reward moments and the reward bound, not the full reward law. Same pointwise trajectory LR map is weaker than same LR spectrum. The open-policy persistence result gives a horizon-dependent neighborhood, not a global policy theorem. General-H baseline robustness is for deterministic time-indexed scalar baselines; state-dependent baselines are outside scope. Corollary 2 concerns common-batch risk ratios and does not imply an exponential ratio of target-confidence batch sizes or full adaptive training complexities. The exponent gap decays as $\Theta(1/H)$ in this normalized lift, so no horizon-uniform constant gap is claimed.

Q OPERATIONAL EVIDENCE UNDER SEVERE UNCERTAINTY

This appendix records the empirical hierarchy used in Section 4. The experiments have distinct roles: benign controls test whether failure is ubiquitous; the SRE family isolates a rare/high-impact mechanism; external V1 is hypothesis-generating; and Grid2Op V2 is the only external confirmatory test. Discovery data are never pooled into V2 confidence statements.

Q.1 BENIGN, NULL, AND INCONCLUSIVE PANELS

Classic control. A prospective audit considered CartPole, MountainCar, and Pendulum with five seeds and nine frozen checkpoints per seed family (45 audits), using four independent reference-gradient shards, a minimum pairwise reference-direction cosine gate of 0.995, disjoint calibration and screening samples, and fresh direct-risk confirmation. No robust threshold failure survived the protocol. A higher-precision follow-up on CartPole and Pendulum used 40,000 reference trajectories,

a 60,000-trajectory pool, and 5,000 direct batches per tested point. Nine of 30 checkpoints were reference-stable, yielding 57 eligible nontrivial tail rows; the maximum empirical/Gaussian ratio was 1.1849, and no factor-two miscalibration occurred.

Q.2 Prospective recognized-control follow-up: MountainCar and Acrobot

To broaden the real-control audit without searching for a favorable failure, we froze a second classic-control protocol on 17 September 2026 before its full outcomes were inspected To broaden the real-control audit without searching for a favorable failure, we prespecified and fixed a second classic-control protocol on 17 September 2026 before inspecting its full outcomes. It uses the documented Gymnasium `MountainCar-v0` and `Acrobot-v1` dynamics (Towers et al., 2024); all six environment–policy configurations are retained. Each environment is evaluated under three fixed categorical policies (uniform, weak momentum, strong momentum), with both vanilla return-to-go REINFORCE and an independently fitted, frozen state-only linear baseline. The policy-gradient dimensions are 6 for MountainCar and 14 for Acrobot.

For every environment–policy pair, independent streams supply 5,000 baseline-fit episodes, 40,000 reference episodes, 40,000 Gaussian-calibration episodes, a 10k/10k split for the non-oracle certificate, and 5,000 disjoint fresh batches at each $n \in \{8, 16, 32\}$. A reference is called stable only if four shard means have minimum pairwise cosine at least 0.99 and $\|\widehat{g}\|/\sqrt{\mathrm{tr}(\widehat{\Sigma})/N} \geq 6$. Threshold tests at $\alpha \in \{.05, .01\}$ use one-sided Clopper–Pearson bounds with Bonferroni budget $0.05/72$ over all predeclared environment–policy–estimator–batch–threshold hypotheses. Maximal discrepancies are descriptive only.

Eleven of the 12 estimator cases fail the frozen reference gate and are therefore inconclusive. The sole stable case is Acrobot with the strong-momentum policy and frozen state baseline: minimum reference-shard cosine 0.9921, reference signal-to-trace-SE 22.20, and empirical gradient-covariance rank $14/14$. Its direct and independent Gaussian risks are:

n	direct risk (95% CP)	Gaussian	familywise threshold conclusion
8	.1690 [.1587, .1797]	.1850	$>.05$ and $>.01$
16	.0962 [.0882, .1047]	.1024	$>.05$ and $>.01$
32	.0374 [.0323, .0430]	.0365	$<.05$ and $>.01$

The largest descriptive direct/Gaussian ratio across these three batch sizes is 1.025 (largest absolute gap 0.0160), so this stable case is a non-failure for Gaussian calibration despite substantial finite-batch directional risk. The non-oracle certificate is vacuous for all twelve cases under the predeclared analytical pathwise envelope: the direction cone never excludes the full sphere. This negative result is consistent with the bound-inflation sensitivity in Appendix H; it demonstrates that the theorem's validity does not by itself make worst-case deterministic envelopes practically useful. The follow-up therefore contributes three pieces of scope evidence at once: full-rank vector gradients, an independently learned state-only baseline, and an explicit example where second-order calibration is adequate while the rigorous certificate abstains.

Exact-reference CliffWalking. The tabular audit spans 120 cases: four grid sizes, five stochasticity levels, and six policy-ϵ settings. Population finite-horizon gradients are computed exactly by dynamic programming. With 80,000 calibration trajectories, 120,000 independent screening trajectories, 20,000 bootstrap batch-risk replicates, and 10,000 fresh confirmation batches, no familywise-confirmed threshold failure occurs at $\alpha = .05$ or $.01$. The maximum screening/Gaussian risk ratio is 1.0803 at $\alpha = .05$ and 0.9618 at $\alpha = .01$.

MuJoCo and broad external controls. A 45-audit Hopper/Walker2d/HalfCheetah panel does not yield a reference-stable checkpoint under its predeclared four-shard gate and is therefore inconclusive rather than negative. In the external V1 suite, Finance has one stable reference among 48 cases and no selected risk test; Safety-Gym has none among 32. Highway completes 21 of 48 cases before post-launch computational censoring after approximately 44 hours; 13 completed cases are reference-stable and none is a screening candidate. The remaining 27 Highway cases are treated as missing, not as negative outcomes.

Q.3 CONTROLLED SIGNIFICANT-RARE-EVENT BENCHMARK

The significant-rare-event (SRE) benchmark is designed to separate population signal from finite-batch tail behavior. Let p denote the rare-event probability, $b_{\mathrm{ord}} > 0$ the ordinary-event scale, q the Bernoulli policy probability, and $\rho > 0$ the severity parameter. Define the rare-event magnitude

$$L_{\mathrm{rare}} = \rho \frac{(1-p)b_{\mathrm{ord}}}{p}.$$

For the directional REINFORCE statistic used in the benchmark, the population gradient reduces to

$$g = q(1-q)(1-p)b_{\mathrm{ord}}(1-\rho).$$

Hence $\rho = .02, .05, .10$ retains, respectively, 98%, 95%, and 90% of the ordinary-event population signal. The confirmed failures are therefore not produced by tuning the mean close to zero.

The prospective grid contains 140 cases:

$$H \in \{5, 10, 20, 50\}, \quad p \in \{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\},$$
$$\rho \in \{.02, .05, .10, .25, .50, .97, 1.03\}.$$

Population means and variances are computed analytically. Gaussian batch selection uses only these second-order quantities. Screening and confirmation use independent random streams. Confirmation uses a one-sided Hoeffding lower bound with Bonferroni correction over the full 140×2 case–alpha family.

At $\alpha = .01$, 92 cases receive a Gaussian-selected batch size, 11 independent screens exceed the operational threshold, and seven are familywise-confirmed. At $\alpha = .05$, 96 cases are selected, two screen candidates occur, and one is familywise-confirmed. The confirmed $\alpha = .01$ keys are

`H5_p0.0001_rho0.020`, `H5_p0.0001_rho0.100`, `H5_p0.001_rho0.050`,
`H10_p0.0001_rho0.020`, `H10_p0.0001_rho0.050`, `H10_p0.001_rho0.050`,
`H20_p0.0001_rho0.050`.

The confirmed $\alpha = .05$ key is `H5_p0.0001_rho0.050`. Table 12 gives representative confirmed cases from the two operational thresholds. Write n_G for the batch size selected by the Gaussian second-order rule. The maximum descriptive selected batch-size ratio is 2. These numbers characterize a controlled mechanism and should not be read as a prevalence estimate for modern deep RL.

(H, p, ρ)	L_{rare}	skewness	n_G	Gaussian	screen	confirm / familywise LB
$(5, 10^{-4}, .02)$	199.98	-62.63	1024	.000367	.030315	.030561 / .029092
$(10, 10^{-4}, .02)$	199.98	–	2048	.000611	.020605	.020391 / .018922
$(5, 10^{-4}, .05)$	499.95	-83.49	2048	.020700	.060940	.061384 / .059915

Table 12: Representative confirmed SRE cases. The first two rows use $\alpha = .01$; the last uses $\alpha = .05$. The maximum independent-screen/Gaussian ratio over the $\alpha = .01$ selected panel is 82.686, attained by the first row.

Q.4 EXTERNAL V1 DISCOVERY SUITE

The V1 suite contains 152 prospectively enumerated cases: Highway (48), Grid2Op (24), Finance (48), and Safety-Gym (32). A case is reference-stable only if four independent shard means have the same nonzero sign and $|\widehat{\mu}| > 4\,\mathrm{SE}$. Calibration, screening, and confirmation use independent seeds. A single familywise budget $0.05/(152 \times 2)$ is used for the V1 case–alpha family.

A total of 125 cases complete. Grid2Op completes 24/24 and all are reference-stable. At $\alpha = .01$, one Grid2Op screening candidate occurs for penalty 1000, overload threshold 1.05, policy parameter $\theta = -3$, and V1 seed 9514: $n_G = 16$, Gaussian risk 0.004958, screen risk 0.012, fresh-confirmation risk 0.011. Its sweep-wide one-sided lower bound is $0.002812 < 0.01$, so the case is *not* a V1 confirmatory success. A different V1 case has the largest descriptive screen/Gaussian ratio, $9.344\times$, but is not a threshold candidate. These two facts must not be conflated.

Highway completes 21/48 cases; the remaining 27 workers were terminated after approximately 44 hours of continuous computation. This was post-launch computational censoring after partial data were visible, not a predeclared time cap; no prospective full-panel inference is made for Highway. Finance and Safety-Gym are inconclusive because the reference gate rarely or never passes.

Q.5 GRID2OP SEMANTICS AND V2 INDEPENDENT CONFIRMATION

Grid2Op is an open sequential power-network simulation framework (Marot et al., 2021). The selected environment is `l2rpn_case14_sandbox` with `test=True`. The framework documentation marks these environments for testing/education and cautions against extensive studies because they contain little data (Grid2Op developers, 2026); we therefore use it as an external sequential stress environment rather than a production-power-system benchmark.

A fresh semantic audit examines 55 episodes and the first 20 early terminations. All 20 have nonempty backend exceptions, with `is_illegal=false` and `is_ambiguous=false`; the exceptions report power-flow non-convergence, and post-termination observations have zero connected lines out of 20. We therefore describe the event conservatively as power-flow divergence/grid game-over, not as a literal real-world blackout.

The V2 configuration is fixed from V1 before V2 outcomes are generated: $\theta = -3$, overload threshold 1.05, logit slope 8, action equal to disconnecting the connected line with maximum line-loading ratio ρ, failure penalty 1000, and maximum horizon 96. V2 reuses no V1 samples. Its primary hypothesis is

$$H_0 : \mathcal{R}_{16}^{\mathrm{str}} \leq .01 \qquad \text{versus} \qquad H_1 : \mathcal{R}_{16}^{\mathrm{str}} > .01$$

with a single one-sided level-0.05 test. Reference, calibration, and confirmation use disjoint random streams.

The reference stage uses 20,000 episodes in eight shards and requires all nonzero shard means to have the same sign and $|\widehat{\mu}|/\mathrm{SE} > 6$. It yields mean -1475.6076, standard deviation 2516.0544, standard error (SE) 17.7912, and z-score 82.9403, with all eight shard means negative. The independent 50,000-episode calibration sample yields projected mean 1484.1743, SD 2551.0383, and the *secondary descriptive* Gaussian risk 0.00997814 at $n = 16$.

The primary confirmation comprises exactly 100,000 independent batches of 16 episodes, or 1.6 million episodes. There are 1,258 strict wrong-direction batches:

$$\widehat{\mathcal{R}}_{16}^{\mathrm{str}} = 0.01258, \qquad \mathrm{CP}_{95\%}^{\mathrm{lower}} = 0.0120057569 > 0.01.$$

The predeclared primary criterion is therefore met. The Gaussian number is not retroactively part of that primary test; it is reported only to show that a separate second-order calibration falls on the opposite side of the operational threshold.