

A Sharp Verification Boundary for Pairwise Social Choice under Strategic Evidence Distortion

Kevin Fathi
fathikevin@protonmail.com

April 2026

Abstract

We study pairwise social choice when the publicly observed evidence law can be strategically distorted. For a contested pair (x, y) , each frame f is associated with a compact feasible set \mathcal{Q}_f of post-distortion evidence laws on a finite alphabet. The operational problem is robust binary classification between the two compact uncertainty sets

$$A_x := \bigcup_{f \in F_x} \mathcal{Q}_f, \quad A_y := \bigcup_{g \in F_y} \mathcal{Q}_g.$$

We prove a sharp dichotomy. If $A_x \cap A_y \neq \emptyset$, then for every sample size and every decision rule the worst-case typewise misranking error is at least $1/2$. If $A_x \cap A_y = \emptyset$, then the nearest-set empirical-distribution classifier has uniformly exponentially small error, with an explicit rate controlled by the total-variation gap between A_x and A_y . We then specialize to reverse-Kullback–Leibler budgets

$$\mathcal{Q}_f(C_S) := \{q \in \Delta(\mathcal{E}) : D_{\text{KL}}(q \| p_f) \leq C_S\}.$$

For this model the existence boundary is exact: writing

$$C_W(x, y) := \min_{f \in F_x, g \in F_y} C(p_f, p_g)$$

for the pairwise witness capacity, where C is Chernoff information, we prove

$$A_x(C_S) \cap A_y(C_S) = \emptyset \iff C_S < C_W(x, y).$$

Below this threshold the hidden-adversary problem has an exact asymptotic minimax exponent,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \inf_{\delta_n} \mathcal{E}_n(\delta_n) = -D^*(C_S),$$

where $D^*(C_S)$ is the minimum feasible Chernoff information over cross pairs. We also give a finite- n polynomial sandwich around this exponent. For dynamic attacks with ergodic feasibility, time variation and actor-side switching costs do not change the existence boundary because any successful aliasing strategy can be chosen constant. Under common-preference misalignment, the same condition is necessary and sufficient for evidence-based mechanism design. The paper makes no claim to solve Arrow-type impossibility theorems or the exact minimax-exponent problem for arbitrary compact uncertainty classes.

1 Introduction

The paper isolates a narrow problem and solves it exactly. Fix a contested pair (x, y) . Different latent frames support different rankings of that pair. Publicly observed evidence can help identify which side of the pair is justified, but only if the evidence law itself cannot be strategically distorted into cross-frame indistinguishability. The relevant question is therefore not the full social-choice problem and not the full robust-testing problem. It is the following verification problem:

When does there exist a data-based rule that robustly distinguishes the frames favoring x from the frames favoring y , uniformly over all strategically feasible evidence distortions?

The paper answers that question in three layers. First, for general compact uncertainty classes, the operational issue is whether the feasible evidence laws for the two sides overlap. If they overlap, robust verification is impossible: an adversary can make an x -frame and a y -frame generate the same evidence law, and no statistical rule can do better than chance on the worst cross pair. If they do not overlap, robust verification is possible, and a universal classifier based on the empirical distribution has exponentially small worst-case error. Second, in the reverse-KL model the overlap condition has an exact information-theoretic threshold. Let p_f be the baseline evidence law under frame f , and let C_S be the reverse-KL distortion budget. Then overlap occurs on a cross pair (f, g) if and only if $C_S \geq C(p_f, p_g)$, where C is Chernoff information. Taking the minimum over cross pairs produces a sharp pairwise threshold $C_S = C_W(x, y)$. Third, below that threshold the hidden-adversary problem has an exact asymptotic minimax exponent. The exponent is

$$D^*(C_S) := \min_{f \in F_x, g \in F_y} \inf_{u \in \mathcal{Q}_f(C_S), v \in \mathcal{Q}_g(C_S)} C(u, v),$$

the minimum feasible Chernoff information over cross pairs. Thus the KL model is solved not only at the level of existence but also at the level of asymptotic rate. This is the paper’s sharpest claim. It does not solve Arrow’s theorem, Gibbard–Satterthwaite, Myerson–Satterthwaite, or the exact minimax exponent problem for arbitrary compact uncertainty classes. It does something more limited and more defensible: it identifies when pairwise evidence-based verification survives strategic distortion, and in the reverse-KL model it identifies the exact exponential rate of that survival.

What is new. The contribution is the combination of four pieces into one operational criterion:

- (i) a robust-verification formulation of pairwise social choice under manipulable evidence;
- (ii) an exact Chernoff/KL threshold in the reverse-KL model;
- (iii) an exact asymptotic minimax exponent for the hidden-adversary reverse-KL problem; and
- (iv) a coordinatewise extension to mechanism design under preference misalignment.

What is not claimed. The paper does not claim a new social-welfare axiom, a new impossibility theorem of Arrow/Gibbard type, a universal theory of Goodhart effects, or a general AI-alignment theorem. It also does not claim the exact optimal exponent for arbitrary compact uncertainty classes. In the general compact model the paper proves existence of exponential recovery with a universal classifier, not sharp asymptotic optimality.

Lineage. The failure mode of expressed agreement under latent disagreement is the spurious-unanimity problem emphasized by Mongin [7]. The testing layer is classical robust hypothesis testing [5, 6]. The manipulation layer is structurally related to strategic classification [4]. The KL threshold itself is built on the Chernoff radius identity [1].

2 Model and operational criterion

Let \mathcal{X} be a finite set of alternatives and let \mathcal{F} be a finite set of frames. Fix a contested pair $(x, y) \in \mathcal{X}^2$ with $x \neq y$ and define

$$F_x := \{f \in \mathcal{F} : x \succ_f y\}, \quad F_y := \{g \in \mathcal{F} : y \succ_g x\}.$$

Only this pair matters in what follows. Let \mathcal{E} be a finite evidence alphabet. Under frame f , the baseline one-sample evidence law is a probability mass function $p_f \in \Delta(\mathcal{E})$.

Definition 2.1 (Feasible post-distortion laws). For each frame $f \in \mathcal{F}$, let $\mathcal{Q}_f \subseteq \Delta(\mathcal{E})$ be a nonempty compact set. The interpretation is that, when frame f is the true frame, the strategic actor can replace the baseline law p_f by any post-distortion law $q \in \mathcal{Q}_f$.

The robust pairwise decision problem depends only on the two unions

$$A_x := \bigcup_{f \in F_x} \mathcal{Q}_f, \quad A_y := \bigcup_{g \in F_y} \mathcal{Q}_g.$$

These are compact subsets of the simplex because they are finite unions of compact sets.

Remark 2.1 (Joint feasible families and the coordinatewise hull). Sometimes the primitive object is a compact set

$$\Sigma \subseteq (\Delta(\mathcal{E}))^{\mathcal{F}}$$

of jointly feasible evidence-law families $q = (q_f)_{f \in \mathcal{F}}$. The operational decision problem depends only on the coordinatewise projections

$$\mathcal{Q}_f := \text{proj}_f(\Sigma), \quad \Sigma^\square := \prod_{f \in \mathcal{F}} \mathcal{Q}_f.$$

The pairwise uncertainty sets are then $A_x = \cup_{f \in F_x} \text{proj}_f(\Sigma)$ and $A_y = \cup_{g \in F_y} \text{proj}_g(\Sigma)$. The diagonal criterion can be written equivalently as

$$A_x \cap A_y = \emptyset \iff \Sigma^\square \cap \mathcal{D}_{xy} = \emptyset,$$

where

$$\mathcal{D}_{xy} := \{q \in (\Delta(\mathcal{E}))^{\mathcal{F}} : \exists (f, g) \in F_x \times F_y \text{ with } q_f = q_g\}.$$

For a sample of size n , a decision rule is any measurable map

$$\delta_n : \mathcal{E}^n \rightarrow \{x, y\}.$$

Its worst-case typewise error is

$$\mathcal{E}_n(\delta_n) := \max \left\{ \sup_{f \in F_x} \sup_{q \in \mathcal{Q}_f} \mathbb{P}_{q^{\otimes n}}(\delta_n(E^{(n)}) = y), \sup_{g \in F_y} \sup_{q \in \mathcal{Q}_g} \mathbb{P}_{q^{\otimes n}}(\delta_n(E^{(n)}) = x) \right\}. \quad (2.1)$$

This is the operational notion used throughout the paper.

2.1 A universal classifier

Let

$$\hat{P}_n(e) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{E_i = e\} \quad (e \in \mathcal{E})$$

be the empirical distribution. For a compact set $A \subseteq \Delta(\mathcal{E})$, write

$$\text{dist}_1(r, A) := \inf_{a \in A} \|r - a\|_1.$$

Define the nearest-set classifier

$$\psi_n(E^{(n)}) = \begin{cases} x, & \text{dist}_1(\hat{P}_n, A_x) \leq \text{dist}_1(\hat{P}_n, A_y), \\ y, & \text{otherwise.} \end{cases} \quad (2.2)$$

The proof of exponential recovery uses a standard type-counting estimate.

Lemma 2.1 (Type bound). *Let $p \in \Delta(\mathcal{E})$ and let $\Gamma \subseteq \Delta(\mathcal{E})$ be any set. Then*

$$\mathbb{P}_{p^{\otimes n}}(\hat{P}_n \in \Gamma) \leq (n+1)^{|\mathcal{E}|} \exp\left(-n \inf_{r \in \Gamma} D_{\text{KL}}(r \| p)\right).$$

Proof. There are at most $(n+1)^{|\mathcal{E}|}$ empirical types on \mathcal{E} . For a type r achievable at sample size n , the probability of its type class under $p^{\otimes n}$ is at most $\exp(-n D_{\text{KL}}(r \| p))$; see, for example, the method-of-types bound in Csiszár and Körner [3, Chapter 2]. Summing over the types contained in Γ yields the claim. \square

Lemma 2.2 (Pinsker in ℓ_1 form). *For all $p, r \in \Delta(\mathcal{E})$,*

$$D_{\text{KL}}(r \| p) \geq \frac{1}{2} \|r - p\|_1^2.$$

Proof. This is Pinsker's inequality in the convention where total variation is $\|\cdot\|_1/2$. \square

3 Robust verification under compact uncertainty

The decisive fact is simple: robust verification is possible if and only if the cross-side feasible evidence laws are disjoint.

Theorem 3.1 (Operational dichotomy). *Let $A_x, A_y \subseteq \Delta(\mathcal{E})$ be the compact uncertainty sets defined above.*

(A) **Aliasing.** *If $A_x \cap A_y \neq \emptyset$, then for every $n \geq 1$ and every decision rule δ_n ,*

$$\mathcal{E}_n(\delta_n) \geq \frac{1}{2}.$$

(B) **Recovery.** *If $A_x \cap A_y = \emptyset$, define the gap*

$$d_{xy} := \inf_{p \in A_x, q \in A_y} \|p - q\|_1 > 0.$$

Then the nearest-set classifier (2.2) satisfies, for every $n \geq 1$,

$$\mathcal{E}_n(\psi_n) \leq (n+1)^{|\mathcal{E}|} \exp\left(-\frac{nd_{xy}^2}{8}\right).$$

In particular, the worst-case typewise error decays exponentially.

Proof. (A) Choose $r \in A_x \cap A_y$. Then there exist $f^* \in F_x$ and $g^* \in F_y$ such that $r \in \mathcal{Q}_{f^*} \cap \mathcal{Q}_{g^*}$. Fix any rule δ_n and write

$$\alpha := \mathbb{P}_{r^{\otimes n}}(\delta_n(E^{(n)}) = x).$$

Under the x -side type f^* , the typewise error is $1 - \alpha$. Under the y -side type g^* , the typewise error is α . Hence

$$\mathcal{E}_n(\delta_n) \geq \max\{\alpha, 1 - \alpha\} \geq \frac{1}{2}.$$

(B) Since A_x and A_y are compact and disjoint, $d_{xy} > 0$. Fix $p \in A_x$. If ψ_n misclassifies a sample generated from p , then by definition of ψ_n there exists some $q \in A_y$ such that

$$\|\hat{P}_n - q\|_1 \leq \text{dist}_1(\hat{P}_n, A_y) \leq \text{dist}_1(\hat{P}_n, A_x) \leq \|\hat{P}_n - p\|_1.$$

Therefore

$$d_{xy} \leq \|p - q\|_1 \leq \|p - \hat{P}_n\|_1 + \|\hat{P}_n - q\|_1 \leq 2\|\hat{P}_n - p\|_1,$$

so misclassification implies

$$\|\hat{P}_n - p\|_1 \geq \frac{d_{xy}}{2}.$$

Applying Lemma 2.1 with

$$\Gamma_p := \left\{ r \in \Delta(\mathcal{E}) : \|r - p\|_1 \geq \frac{d_{xy}}{2} \right\}$$

and then Lemma 2.2 yields

$$\mathbb{P}_{p^{\otimes n}}(\psi_n(E^{(n)}) = y) \leq (n+1)^{|\mathcal{E}|} \exp\left(-n \inf_{r \in \Gamma_p} D_{\text{KL}}(r \| p)\right) \leq (n+1)^{|\mathcal{E}|} \exp\left(-\frac{nd_{xy}^2}{8}\right).$$

This bound is uniform in $p \in A_x$. The argument for $q \in A_y$ is identical. Taking suprema gives the claim. \square

Corollary 3.2 (Bayes error under a frame prior). *Fix any prior μ_h on \mathcal{F} . If $A_x \cap A_y = \emptyset$, then the Bayes misranking probability of ψ_n conditional on h is at most*

$$(n+1)^{|\mathcal{E}|} \exp\left(-\frac{nd_{xy}^2}{8}\right).$$

If $A_x \cap A_y \neq \emptyset$, then for every n and every rule δ_n there exist $f \in F_x$ and $g \in F_y$ such that the conditional Bayes misranking probability is at least $\min\{\mu_h(f), \mu_h(g)\}$.

Proof. The first claim follows from Theorem 3.1(B) by averaging the typewise error bound over the prior. For the second, use the same pair (f^*, g^*) and common law r as in the proof of Theorem 3.1(A). \square

Remark 3.1 (What this theorem does and does not say). Theorem 3.1 solves the existence problem for robust verification under compact uncertainty. It does not claim the exact optimal exponent for arbitrary uncertainty sets. The explicit exponent in part (B) is a universal, constructive lower bound obtained from the ℓ_1 gap and Pinsker's inequality.

4 Reverse-KL budgets and the Chernoff boundary

We now specialize to the model that motivated the original manuscript.

Assumption 4.1 (Full support). For every $f \in \mathcal{F}$ and every $e \in \mathcal{E}$, $p_f(e) > 0$.

For $p, q \in \Delta(\mathcal{E})$ define the Chernoff information

$$C(p, q) := - \min_{s \in [0, 1]} \log \left(\sum_{e \in \mathcal{E}} p(e)^s q(e)^{1-s} \right). \quad (4.1)$$

Definition 4.1 (Reverse-KL feasible sets and pairwise witness capacity). For $C_S \geq 0$, define

$$\mathcal{Q}_f(C_S) := \{q \in \Delta(\mathcal{E}) : D_{\text{KL}}(q \| p_f) \leq C_S\}.$$

For the pair (x, y) define the witness capacity

$$C_W(x, y) := \min_{f \in F_x, g \in F_y} C(p_f, p_g).$$

We call the threshold $C_S = C_W(x, y)$ the *pairwise verification boundary*.

The KL specialization rests on the classical fact that Chernoff information is exactly the minimum common reverse-KL radius needed to hit both laws.

Lemma 4.1 (Chernoff information vanishes only on equality). *Under Assumption 4.1,*

$$C(p, q) = 0 \iff p = q.$$

Proof. If $p = q$, then the expression in (4.1) equals 1 for every s , so $C(p, q) = 0$. Conversely, if $C(p, q) = 0$, then in particular

$$\sum_{e \in \mathcal{E}} \sqrt{p(e)q(e)} = 1.$$

By Cauchy–Schwarz, equality can hold only when $p = q$. □

Lemma 4.2 (Tilted equalizer). *Under Assumption 4.1, if $p \neq q$ then there exists $s^* \in (0, 1)$ such that the tilted law*

$$r_{s^*}(e) := \frac{p(e)^{s^*} q(e)^{1-s^*}}{\sum_{u \in \mathcal{E}} p(u)^{s^*} q(u)^{1-s^*}}$$

satisfies

$$D_{\text{KL}}(r_{s^*} \| p) = D_{\text{KL}}(r_{s^*} \| q) = C(p, q).$$

Proof. Define

$$\Lambda(s) := \log \sum_{e \in \mathcal{E}} p(e)^s q(e)^{1-s}, \quad s \in [0, 1].$$

Exactly as in the standard proof of the Chernoff theorem, Λ is convex, $\Lambda'(0) = -D_{\text{KL}}(q \| p) < 0$, and $\Lambda'(1) = D_{\text{KL}}(p \| q) > 0$. Hence Λ has a minimizer $s^* \in (0, 1)$ with $\Lambda'(s^*) = 0$. Writing out the KL divergences of the tilted law gives

$$D_{\text{KL}}(r_{s^*} \| p) = -\Lambda(s^*) = D_{\text{KL}}(r_{s^*} \| q) = C(p, q).$$

□

Proposition 4.3 (Chernoff radius identity). *Under Assumption 4.1,*

$$\inf_{r \in \Delta(\mathcal{E})} \max\{D_{\text{KL}}(r\|p), D_{\text{KL}}(r\|q)\} = C(p, q) \quad \text{for all } p, q \in \Delta(\mathcal{E}).$$

Proof. If $p = q$, both sides are zero. Assume $p \neq q$. For any $r \in \Delta(\mathcal{E})$ and any $s \in [0, 1]$,

$$\max\{D_{\text{KL}}(r\|p), D_{\text{KL}}(r\|q)\} \geq (1-s)D_{\text{KL}}(r\|p) + sD_{\text{KL}}(r\|q).$$

A direct calculation gives

$$(1-s)D_{\text{KL}}(r\|p) + sD_{\text{KL}}(r\|q) = D_{\text{KL}}(r\|\bar{r}_s) - \log \sum_{e \in \mathcal{E}} p(e)^{1-s} q(e)^s,$$

where

$$\bar{r}_s(e) := \frac{p(e)^{1-s} q(e)^s}{\sum_{u \in \mathcal{E}} p(u)^{1-s} q(u)^s}.$$

Hence

$$\max\{D_{\text{KL}}(r\|p), D_{\text{KL}}(r\|q)\} \geq -\log \sum_{e \in \mathcal{E}} p(e)^{1-s} q(e)^s.$$

Taking the infimum over r and then the maximum over s yields

$$\inf_r \max\{D_{\text{KL}}(r\|p), D_{\text{KL}}(r\|q)\} \geq C(p, q).$$

The reverse inequality follows from Lemma 4.2: the tilted law r_{s^*} satisfies

$$\max\{D_{\text{KL}}(r_{s^*}\|p), D_{\text{KL}}(r_{s^*}\|q)\} = C(p, q).$$

□

We can now state the exact KL threshold.

Theorem 4.4 (Sharp reverse-KL boundary). *Assume Assumption 4.1. For each $C_S \geq 0$, define*

$$A_x(C_S) := \bigcup_{f \in F_x} \mathcal{Q}_f(C_S), \quad A_y(C_S) := \bigcup_{g \in F_y} \mathcal{Q}_g(C_S).$$

Then

$$A_x(C_S) \cap A_y(C_S) = \emptyset \iff C_S < C_W(x, y).$$

Equivalently, aliasing is feasible if and only if $C_S \geq C_W(x, y)$.

Proof. If $A_x(C_S) \cap A_y(C_S) \neq \emptyset$, then for some cross pair (f, g) there exists r such that

$$D_{\text{KL}}(r\|p_f) \leq C_S, \quad D_{\text{KL}}(r\|p_g) \leq C_S.$$

Therefore

$$C(p_f, p_g) = \inf_u \max\{D_{\text{KL}}(u\|p_f), D_{\text{KL}}(u\|p_g)\} \leq C_S$$

by Proposition 4.3. Taking the minimum over cross pairs gives $C_W(x, y) \leq C_S$. Conversely, if $C_S \geq C_W(x, y)$, choose a minimizing cross pair (f^*, g^*) with

$$C(p_{f^*}, p_{g^*}) = C_W(x, y).$$

By Proposition 4.3 and Lemma 4.2, there exists r^* such that

$$D_{\text{KL}}(r^*\|p_{f^*}) = D_{\text{KL}}(r^*\|p_{g^*}) = C_W(x, y) \leq C_S.$$

Thus $r^* \in \mathcal{Q}_{f^*}(C_S) \cap \mathcal{Q}_{g^*}(C_S)$, so $A_x(C_S) \cap A_y(C_S) \neq \emptyset$. □

Combining Theorems 3.1 and 4.4 gives the operational KL theorem.

Corollary 4.5 (Pairwise verification theorem under reverse KL). *Assume Assumption 4.1.*

- (a) *If $C_S \geq C_W(x, y)$, then every decision rule has worst-case typewise error at least $1/2$ for every sample size.*
- (b) *If $C_S < C_W(x, y)$, then the nearest-set classifier built from $A_x(C_S)$ and $A_y(C_S)$ has uniformly exponentially decaying worst-case typewise error.*

4.1 Worked example: binary evidence and two frames

Take $|\mathcal{E}| = 2$ and $|F_x| = |F_y| = 1$. Write

$$p := p_f = \text{Bern}(0.7), \quad q := p_g = \text{Bern}(0.3).$$

Then $q = (0.3, 0.7) = (1 - 0.7, 0.7)$ is the reflected law of $p = (0.7, 0.3)$, so the Chernoff minimizer is $s^* = \frac{1}{2}$ and

$$C_W(x, y) = C(p, q) = -\log\left(2\sqrt{0.7 \cdot 0.3}\right) \approx 0.08718.$$

Hence verification is possible exactly for $C_S < 0.08718$ and fails for $C_S \geq 0.08718$. At the boundary the tilted equalizer is

$$r_{1/2}(e) \propto \sqrt{p(e)q(e)},$$

which here gives

$$r_{1/2} = (0.5, 0.5) = \text{Bern}(0.5).$$

A direct calculation confirms

$$D_{\text{KL}}(\text{Bern}(0.5) \parallel \text{Bern}(0.7)) = D_{\text{KL}}(\text{Bern}(0.5) \parallel \text{Bern}(0.3)) \approx 0.08718 = C_W(x, y).$$

For $0 \leq C_S < C_W(x, y)$, symmetry implies that the nearest feasible cross pair is

$$q_f^*(C_S) = \text{Bern}(\theta_+(C_S)), \quad q_g^*(C_S) = \text{Bern}(1 - \theta_+(C_S)),$$

where $\theta_+(C_S) \in [0.5, 0.7]$ is the unique inner solution of

$$D_{\text{KL}}(\text{Bern}(\theta) \parallel \text{Bern}(0.7)) = C_S.$$

At $C_S = 0.05$ this gives

$$\theta_+(0.05) \approx 0.54972, \quad q_f^*(0.05) = \text{Bern}(0.54972), \quad q_g^*(0.05) = \text{Bern}(0.45028).$$

The induced ℓ_1 gap is

$$d_{xy}(0.05) = \|q_f^*(0.05) - q_g^*(0.05)\|_1 \approx 0.19889,$$

so the universal Pinsker exponent from Theorem 3.1(B) is

$$\frac{d_{xy}(0.05)^2}{8} \approx 0.00494.$$

The sharp KL exponent from Section 5 is

$$D^*(0.05) = C(q_f^*(0.05), q_g^*(0.05)) \approx 0.00497.$$

Table 1 reports the values of $D^*(C_S)$ and the Pinsker exponent $d_{xy}(C_S)^2/8$ across the feasible range. In this symmetric Bernoulli example the Pinsker exponent is numerically close to the sharp exponent away from the origin, but it is uniformly smaller and therefore not exact.

C_S	$\theta_+(C_S)$	$D^*(C_S)$	$d_{xy}(C_S)^2/8$	ratio
0	0.70000	0.08718	0.08000	1.08971
0.02	0.60602	0.02300	0.02248	1.02318
0.04	0.56599	0.00879	0.00871	1.00881
0.06	0.53496	0.00245	0.00244	1.00245
0.08	0.50865	0.00015	0.00015	1.00015
0.08718	0.50000	0	0	—

Table 1: Binary reverse-KL example with $p = \text{Bern}(0.7)$ and $q = \text{Bern}(0.3)$. All values are computed from the closed-form Bernoulli KL and Chernoff formulas.

Remark 4.1 (Observable versus hidden post-distortion laws). If the realized post-distortion family $q = (q_f)_{f \in \mathcal{F}}$ is known to the decision maker, the pairwise MAP rule has error exponent

$$\min_{f \in F_x, g \in F_y} C(q_f, q_g).$$

Section 5 shows that when the adversary's post-distortion choice is hidden but restricted by reverse-KL budgets, the exact asymptotic minimax exponent is $D^*(C_S)$.

5 Sharp minimax exponents for reverse-KL uncertainty

Fix $0 \leq C_S \leq C_W(x, y)$ and define the robust Chernoff separation

$$D^*(C_S) := \min_{f \in F_x, g \in F_y} \inf_{u \in \mathcal{Q}_f(C_S), v \in \mathcal{Q}_g(C_S)} C(u, v).$$

This quantity is the sharp KL analogue of the universal Pinsker exponent from Theorem 3.1(B).

Proposition 5.1 (Basic properties of $D^*(C_S)$). *Assume Assumption 4.1.*

- (a) The map $C_S \mapsto D^*(C_S)$ is nonincreasing on $[0, C_W(x, y)]$.
- (b) $D^*(0) = C_W(x, y)$.
- (c) $D^*(C_S) > 0$ for every $0 \leq C_S < C_W(x, y)$.
- (d) $D^*(C_W(x, y)) = 0$.
- (e) The map $C_S \mapsto D^*(C_S)$ is continuous on $[0, C_W(x, y)]$.

Proof. For a fixed cross pair (f, g) , write

$$\Phi_{fg}(t) := \inf_{u \in \mathcal{Q}_f(t), v \in \mathcal{Q}_g(t)} C(u, v), \quad D^*(t) = \min_{f \in F_x, g \in F_y} \Phi_{fg}(t).$$

Parts (a) and (b) are immediate. The feasible sets expand with t , so each Φ_{fg} is nonincreasing. At $t = 0$ we have $\mathcal{Q}_f(0) = \{p_f\}$, hence

$$D^*(0) = \min_{f \in F_x, g \in F_y} C(p_f, p_g) = C_W(x, y).$$

For part (c), fix $0 \leq t < C_W(x, y)$ and suppose that $D^*(t) = 0$. Then some cross pair (f, g) admits sequences $u_n \in \mathcal{Q}_f(t)$ and $v_n \in \mathcal{Q}_g(t)$ with $C(u_n, v_n) \rightarrow 0$. The product $\mathcal{Q}_f(t) \times \mathcal{Q}_g(t)$ is compact, so along a subsequence $(u_n, v_n) \rightarrow (u, v)$. For each $s \in (0, 1)$ the map

$$(u, v) \mapsto -\log \sum_{e \in \mathcal{E}} u(e)^s v(e)^{1-s}$$

is continuous, and

$$C(u, v) = \sup_{s \in (0, 1)} \left(-\log \sum_{e \in \mathcal{E}} u(e)^s v(e)^{1-s} \right),$$

so C is lower semicontinuous on $\Delta(\mathcal{E}) \times \Delta(\mathcal{E})$. Therefore

$$0 \leq C(u, v) \leq \liminf_{n \rightarrow \infty} C(u_n, v_n) = 0.$$

By Lemma 4.1, $u = v$, so $A_x(t) \cap A_y(t) \neq \emptyset$, contradicting Theorem 4.4. Hence $D^*(t) > 0$. For part (d), Theorem 4.4 gives a cross pair (f^*, g^*) and a common law $r^* \in \mathcal{Q}_{f^*}(C_W) \cap \mathcal{Q}_{g^*}(C_W)$. Then

$$D^*(C_W(x, y)) \leq C(r^*, r^*) = 0.$$

Nonnegativity of C forces equality. For part (e), it suffices to prove that each Φ_{fg} is continuous. Let $t_n \downarrow t$. Choose $(u_n, v_n) \in \mathcal{Q}_f(t_n) \times \mathcal{Q}_g(t_n)$ with

$$C(u_n, v_n) \leq \Phi_{fg}(t_n) + \frac{1}{n}.$$

Compactness gives a convergent subsequence with limit $(u, v) \in \mathcal{Q}_f(t) \times \mathcal{Q}_g(t)$. Lower semicontinuity of C yields

$$\liminf_{n \rightarrow \infty} \Phi_{fg}(t_n) \geq C(u, v) \geq \Phi_{fg}(t).$$

Monotonicity gives the reverse inequality, so $\Phi_{fg}(t_n) \rightarrow \Phi_{fg}(t)$. Now let $t_n \uparrow t$ and fix $\varepsilon > 0$. Choose $(u, v) \in \mathcal{Q}_f(t) \times \mathcal{Q}_g(t)$ with

$$C(u, v) \leq \Phi_{fg}(t) + \varepsilon.$$

For $\lambda \in (0, 1)$ set

$$u_\lambda := (1 - \lambda)u + \lambda p_f, \quad v_\lambda := (1 - \lambda)v + \lambda p_g.$$

Convexity of $r \mapsto D_{\text{KL}}(r \| p_f)$ and $r \mapsto D_{\text{KL}}(r \| p_g)$ gives

$$D_{\text{KL}}(u_\lambda \| p_f) \leq (1 - \lambda)t, \quad D_{\text{KL}}(v_\lambda \| p_g) \leq (1 - \lambda)t.$$

For each fixed $s \in [0, 1]$, the map $(u, v) \mapsto -\log \sum_e u(e)^s v(e)^{1-s}$ is convex, hence so is C as a pointwise supremum of convex functions. Therefore

$$C(u_\lambda, v_\lambda) \leq (1 - \lambda)C(u, v) + \lambda C(p_f, p_g).$$

Choose λ so small that the right-hand side is at most $\Phi_{fg}(t) + 2\varepsilon$. For all sufficiently large n , $t_n \geq (1 - \lambda)t$, so (u_λ, v_λ) is feasible for $\Phi_{fg}(t_n)$ and

$$\Phi_{fg}(t_n) \leq \Phi_{fg}(t) + 2\varepsilon.$$

Monotonicity gives $\Phi_{fg}(t_n) \geq \Phi_{fg}(t)$. Hence $\Phi_{fg}(t_n) \rightarrow \Phi_{fg}(t)$. Since D^* is the minimum of finitely many continuous maps, it is continuous. \square

Remark 5.1 (Computation of $D^*(C_S)$). For a fixed cross pair (f, g) and a fixed $s \in [0, 1]$, the map

$$(u, v) \mapsto -\log \sum_{e \in \mathcal{E}} u(e)^s v(e)^{1-s}$$

is convex on $\Delta(\mathcal{E}) \times \Delta(\mathcal{E})$ because $(u, v) \mapsto \sum_e u(e)^s v(e)^{1-s}$ is concave and $-\log$ is convex and decreasing. Hence the sublevel sets of

$$C(u, v) = \sup_{s \in [0, 1]} \left(-\log \sum_{e \in \mathcal{E}} u(e)^s v(e)^{1-s} \right)$$

are convex, so each pairwise problem defining $D^*(C_S)$ is quasiconvex over two reverse-KL balls. In practice one solves the $|F_x||F_y|$ pairwise problems and takes the minimum. A convenient numerical strategy is bisection on the Chernoff level together with convex feasibility tests. In the one-frame-per-side Bernoulli example of Subsection 4.3, the problem collapses to one-dimensional root finding.

Lemma 5.2 (Likelihood half-space bound). *Let P and Q be probability mass functions on a finite alphabet Ω . Then, for every $n \geq 1$,*

$$P^{\otimes n} \left(Q^{\otimes n}(Z^{(n)}) \geq P^{\otimes n}(Z^{(n)}) \right) \leq (n+1)^{|\Omega|} \exp(-nC(P, Q)).$$

Proof. Let \hat{R}_n be the empirical distribution of $Z^{(n)}$. If $Q^{\otimes n}(z^n) \geq P^{\otimes n}(z^n)$ and $P^{\otimes n}(z^n) > 0$, then every symbol used by \hat{R}_n has positive Q -mass and

$$\sum_{\omega \in \Omega} \hat{R}_n(\omega) \log Q(\omega) \geq \sum_{\omega \in \Omega} \hat{R}_n(\omega) \log P(\omega).$$

Equivalently,

$$D_{\text{KL}}(\hat{R}_n \| Q) \leq D_{\text{KL}}(\hat{R}_n \| P).$$

Fix $s \in [0, 1]$. Then

$$D_{\text{KL}}(\hat{R}_n \| P) \geq (1-s)D_{\text{KL}}(\hat{R}_n \| P) + sD_{\text{KL}}(\hat{R}_n \| Q)$$

and the same algebra as in the proof of Proposition 4.3 gives

$$(1-s)D_{\text{KL}}(\hat{R}_n \| P) + sD_{\text{KL}}(\hat{R}_n \| Q) \geq -\log \sum_{\omega \in \Omega} P(\omega)^{1-s} Q(\omega)^s.$$

Taking the supremum over s yields

$$D_{\text{KL}}(\hat{R}_n \| P) \geq C(P, Q).$$

Therefore

$$\left\{ Q^{\otimes n}(Z^{(n)}) \geq P^{\otimes n}(Z^{(n)}) \right\} \subseteq \left\{ D_{\text{KL}}(\hat{R}_n \| P) \geq C(P, Q) \right\}$$

up to a $P^{\otimes n}$ -null set. Lemma 2.1 gives the claim. \square

Theorem 5.3 (Observable-family upper bound). *Assume Assumption 4.1 and let $0 \leq C_S < C_W(x, y)$. Fix a feasible post-distortion family $\tilde{q} = (q_f)_{f \in \mathcal{F}}$ with $q_f \in \mathcal{Q}_f(C_S)$ for every f . Define*

$$\phi_n^{\text{obs}}(E^{(n)}) = \begin{cases} x, & \max_{f \in F_x} q_f^{\otimes n}(E^{(n)}) \geq \max_{g \in F_y} q_g^{\otimes n}(E^{(n)}), \\ y, & \text{otherwise.} \end{cases}$$

Then

$$\mathcal{E}_n(\phi_n^{\text{obs}}; \tilde{q}) \leq |F_x||F_y|(n+1)^{|\mathcal{E}|} \exp\left(-n \min_{f \in F_x, g \in F_y} C(q_f, q_g)\right) \leq |F_x||F_y|(n+1)^{|\mathcal{E}|} e^{-nD^*(C_S)},$$

where

$$\mathcal{E}_n(\delta_n; \tilde{q}) := \max \left\{ \sup_{f \in F_x} \mathbb{P}_{q_f^{\otimes n}}(\delta_n(E^{(n)}) = y), \sup_{g \in F_y} \mathbb{P}_{q_g^{\otimes n}}(\delta_n(E^{(n)}) = x) \right\}.$$

Proof. Fix $f \in F_x$. If $\phi_n^{\text{obs}}(E^{(n)}) = y$, then there exists $g \in F_y$ such that

$$q_g^{\otimes n}(E^{(n)}) \geq q_f^{\otimes n}(E^{(n)}).$$

Hence

$$\mathbb{P}_{q_f^{\otimes n}}(\phi_n^{\text{obs}}(E^{(n)}) = y) \leq \sum_{g \in F_y} \mathbb{P}_{q_f^{\otimes n}}(q_g^{\otimes n}(E^{(n)}) \geq q_f^{\otimes n}(E^{(n)})).$$

Lemma 5.2 gives

$$\mathbb{P}_{q_f^{\otimes n}}(\phi_n^{\text{obs}}(E^{(n)}) = y) \leq |F_y|(n+1)^{|\mathcal{E}|} \exp\left(-n \min_{g \in F_y} C(q_f, q_g)\right).$$

Taking the supremum over $f \in F_x$ yields the x -side bound. The y -side bound is symmetric, with $|F_x|$ in place of $|F_y|$. The displayed inequality follows because $\max\{|F_x|, |F_y|\} \leq |F_x||F_y|$ and

$$\min_{f \in F_x, g \in F_y} C(q_f, q_g) \geq D^*(C_S).$$

□

Theorem 5.4 (GLR upper bound under a hidden static adversary). *Assume Assumption 4.1 and let $0 \leq C_S < C_W(x, y)$. Define the generalized likelihood-ratio classifier*

$$\phi_n^{\text{GLR}}(E^{(n)}) = \begin{cases} x, & \sup_{p \in A_x(C_S)} p^{\otimes n}(E^{(n)}) \geq \sup_{q \in A_y(C_S)} q^{\otimes n}(E^{(n)}), \\ y, & \text{otherwise.} \end{cases} \quad (5.1)$$

Then, for every $n \geq 1$,

$$\mathcal{E}_n(\phi_n^{\text{GLR}}) \leq (n+1)^{|\mathcal{E}|} e^{-nD^*(C_S)}.$$

Proof. Fix $p \in A_x(C_S)$ and let \hat{P}_n be the empirical distribution. If $\phi_n^{\text{GLR}}(z^n) = y$ and $p^{\otimes n}(z^n) > 0$, then the defining maximum over $A_y(C_S)$ is attained at some $q_{z^n} \in A_y(C_S)$ and

$$q_{z^n}^{\otimes n}(z^n) \geq p^{\otimes n}(z^n).$$

As in the proof of Lemma 5.2, this implies

$$D_{\text{KL}}(\hat{P}_n \| p) \geq C(p, q_{z^n}) \geq \inf_{q \in A_y(C_S)} C(p, q).$$

Therefore

$$\left\{ \phi_n^{\text{GLR}}(E^{(n)}) = y \right\} \subseteq \left\{ D_{\text{KL}}(\hat{P}_n \| p) \geq \inf_{q \in A_y(C_S)} C(p, q) \right\}$$

up to a $p^{\otimes n}$ -null set. Lemma 2.1 yields

$$\mathbb{P}_{p^{\otimes n}}(\phi_n^{\text{GLR}}(E^{(n)}) = y) \leq (n+1)^{|\mathcal{E}|} \exp\left(-n \inf_{q \in A_y(C_S)} C(p, q)\right).$$

Taking the supremum over $p \in A_x(C_S)$ gives

$$\sup_{p \in A_x(C_S)} \mathbb{P}_{p^{\otimes n}}(\phi_n^{\text{GLR}}(E^{(n)}) = y) \leq (n+1)^{|\mathcal{E}|} e^{-nD^*(C_S)}.$$

The argument with $A_x(C_S)$ and $A_y(C_S)$ interchanged gives the same bound on the y -side error. Taking the maximum proves the theorem. \square

Lemma 5.5 (Finite- n type-class lower bound). *Let P and Q be distinct full-support probability mass functions on a finite alphabet Ω , and write $m := |\Omega|$. Then there exists a constant $c_* = c_*(P, Q, m) > 0$ such that, for every $n \geq 1$ and every decision rule $\delta : \Omega^n \rightarrow \{1, 2\}$,*

$$\max \left\{ P^{\otimes n}(\delta(Z^n) = 2), Q^{\otimes n}(\delta(Z^n) = 1) \right\} \geq c_* n^{-(m-1)/2} e^{-nC(P, Q)}.$$

Proof. By Lemma 4.2, there exists a full-support law R_* such that

$$D_{\text{KL}}(R_* \| P) = D_{\text{KL}}(R_* \| Q) = C(P, Q).$$

Set

$$r_{\min} := \min_{\omega \in \Omega} R_*(\omega) > 0.$$

Choose an n -type R_n with $\|R_n - R_*\|_{\infty} \leq 1/n$; such a type exists by rounding the coordinates of nR_* and adjusting one coordinate to restore the total mass. For

$$n_0 := \left\lceil \frac{2}{r_{\min}} \right\rceil$$

we have $R_n(\omega) \geq r_{\min}/2$ for every ω and every $n \geq n_0$. Let

$$K := \left\{ r \in \Delta(\Omega) : r(\omega) \geq \frac{r_{\min}}{2} \text{ for all } \omega \in \Omega \right\}.$$

Because $r \mapsto D_{\text{KL}}(r \| P)$ and $r \mapsto D_{\text{KL}}(r \| Q)$ are continuously differentiable on the compact set K , there exists $L_0 < \infty$ such that, for every $n \geq n_0$,

$$D_{\text{KL}}(R_n \| P) \leq C(P, Q) + \frac{L_0}{n}, \quad D_{\text{KL}}(R_n \| Q) \leq C(P, Q) + \frac{L_0}{n}.$$

Now fix $n \geq n_0$ and write $n_{\omega} := nR_n(\omega)$. The type class T_{R_n} has size

$$|T_{R_n}| = \frac{n!}{\prod_{\omega \in \Omega} n_{\omega}!}.$$

Stirling's bounds

$$n! \geq \sqrt{2\pi} n^{n+1/2} e^{-n}, \quad k! \leq e k^{k+1/2} e^{-k} \quad (k \geq 1)$$

imply

$$|T_{R_n}| \geq \frac{\sqrt{2\pi}}{e^m} n^{-(m-1)/2} \prod_{\omega \in \Omega} R_n(\omega)^{-1/2} e^{nH(R_n)} \geq \frac{\sqrt{2\pi}}{e^m} n^{-(m-1)/2} e^{nH(R_n)}.$$

Therefore

$$P^{\otimes n}(T_{R_n}) = |T_{R_n}| e^{-n(H(R_n) + D_{\text{KL}}(R_n \| P))} \geq \frac{\sqrt{2\pi}}{e^m} n^{-(m-1)/2} e^{-nC(P, Q) - L_0},$$

and the same bound holds with Q in place of P . Let

$$R_1 := \{z^n : \delta(z^n) = 1\}, \quad R_2 := \{z^n : \delta(z^n) = 2\}.$$

All sequences in T_{R_n} have the same $P^{\otimes n}$ -probability and the same $Q^{\otimes n}$ -probability. Hence either $|T_{R_n} \cap R_2| \geq \frac{1}{2}|T_{R_n}|$ or $|T_{R_n} \cap R_1| \geq \frac{1}{2}|T_{R_n}|$. In the first case,

$$P^{\otimes n}(\delta(Z^n) = 2) \geq \frac{1}{2}P^{\otimes n}(T_{R_n}),$$

and in the second,

$$Q^{\otimes n}(\delta(Z^n) = 1) \geq \frac{1}{2}Q^{\otimes n}(T_{R_n}).$$

Thus, for every $n \geq n_0$,

$$\max\left\{P^{\otimes n}(\delta(Z^n) = 2), Q^{\otimes n}(\delta(Z^n) = 1)\right\} \geq \frac{\sqrt{2\pi}}{2e^m} e^{-L_0} n^{-(m-1)/2} e^{-nC(P,Q)}.$$

For $1 \leq n < n_0$, define

$$r_n^* := \inf_{\delta: \Omega^n \rightarrow \{1,2\}} \max\left\{P^{\otimes n}(\delta(Z^n) = 2), Q^{\otimes n}(\delta(Z^n) = 1)\right\}.$$

Because Ω^n is finite, there are only finitely many deterministic decision rules. Since P and Q have full support, no rule can have both errors equal to zero, so $r_n^* > 0$ for every n . Set

$$c_1 := \min_{1 \leq n < n_0} r_n^* n^{(m-1)/2} e^{nC(P,Q)} > 0$$

and

$$c_0 := \frac{\sqrt{2\pi}}{2e^m} e^{-L_0}.$$

Then $c_* := \min\{c_0, c_1\}$ works for every $n \geq 1$. □

Theorem 5.6 (Exact minimax exponent for reverse-KL uncertainty). *Assume Assumption 4.1 and let $0 \leq C_S < C_W(x, y)$. Then*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \inf_{\delta_n} \mathcal{E}_n(\delta_n) = -D^*(C_S).$$

Proof. Theorem 5.4 gives, for every $n \geq 1$,

$$\inf_{\delta_n} \mathcal{E}_n(\delta_n) \leq (n+1)^{|\mathcal{E}|} e^{-nD^*(C_S)}.$$

Fix $\varepsilon > 0$. By the definition of $D^*(C_S)$ there exist a cross pair $(f_\varepsilon, g_\varepsilon) \in F_x \times F_y$ and laws

$$P \in \mathcal{Q}_{f_\varepsilon}(C_S), \quad Q \in \mathcal{Q}_{g_\varepsilon}(C_S)$$

such that

$$C(P, Q) \leq D^*(C_S) + \frac{\varepsilon}{4}.$$

For $t \in (0, 1]$ define

$$P_t := (1-t)P + tp_{f_\varepsilon}, \quad Q_t := (1-t)Q + tp_{g_\varepsilon}.$$

Convexity of $r \mapsto D_{\text{KL}}(r \| p_{f_\varepsilon})$ and $r \mapsto D_{\text{KL}}(r \| p_{g_\varepsilon})$ gives

$$D_{\text{KL}}(P_t \| p_{f_\varepsilon}) \leq (1-t)D_{\text{KL}}(P \| p_{f_\varepsilon}) \leq C_S, \quad D_{\text{KL}}(Q_t \| p_{g_\varepsilon}) \leq (1-t)D_{\text{KL}}(Q \| p_{g_\varepsilon}) \leq C_S.$$

Thus $P_t \in \mathcal{Q}_{f_\varepsilon}(C_S)$ and $Q_t \in \mathcal{Q}_{g_\varepsilon}(C_S)$, and both laws have full support. As noted in the proof of Proposition 5.1(e), C is convex on $\Delta(\mathcal{E}) \times \Delta(\mathcal{E})$, so

$$C(P_t, Q_t) \leq (1-t)C(P, Q) + tC(p_{f_\varepsilon}, p_{g_\varepsilon}).$$

For t small enough,

$$C(P_t, Q_t) \leq D^*(C_S) + \frac{\varepsilon}{2}.$$

Fix such a t , write $\tilde{P} := P_t$ and $\tilde{Q} := Q_t$, and let $c_* := c_*(\tilde{P}, \tilde{Q}, |\mathcal{E}|)$ be the constant from Lemma 5.5. Because $\tilde{P} \in A_x(C_S)$ and $\tilde{Q} \in A_y(C_S)$, every decision rule δ_n satisfies

$$\mathcal{E}_n(\delta_n) \geq \max \left\{ \mathbb{P}_{\tilde{P} \otimes n}(\delta_n(E^{(n)}) = y), \mathbb{P}_{\tilde{Q} \otimes n}(\delta_n(E^{(n)}) = x) \right\}.$$

Lemma 5.5 therefore gives, for every $n \geq 1$,

$$\inf_{\delta_n} \mathcal{E}_n(\delta_n) \geq c_* n^{-|\mathcal{E}|/2} e^{-n(D^*(C_S) + \varepsilon)}.$$

Combining the upper and lower bounds and using $\lim_{n \rightarrow \infty} (\log n)/n = 0$ yields the limit. \square

Corollary 5.7 (Finite- n sandwich). *Assume Assumption 4.1 and let $0 \leq C_S < C_W(x, y)$. For every $\varepsilon > 0$ there exist constants $0 < c_\ell(\varepsilon) \leq c_u < \infty$ such that, for every $n \geq 1$,*

$$c_\ell(\varepsilon) n^{-|\mathcal{E}|/2} e^{-n(D^*(C_S) + \varepsilon)} \leq \inf_{\delta_n} \mathcal{E}_n(\delta_n) \leq c_u n^{|\mathcal{E}|} e^{-nD^*(C_S)}.$$

Proof. Take $c_u := 2$ and use Theorem 5.4, since $(n+1)^{|\mathcal{E}|} \leq 2n^{|\mathcal{E}|}$ for every $n \geq 1$. For the lower bound, use the preceding proof with the same perturbed pair (\tilde{P}, \tilde{Q}) and set $c_\ell(\varepsilon) := c_*(\tilde{P}, \tilde{Q}, |\mathcal{E}|)$. \square

6 Dynamic attacks and actor-side switching costs

Static aliasing is already enough to show that many natural frictions fail. The core point is that successful aliasing can be implemented by a constant strategy and therefore incurs no switching cost.

Definition 6.1 (Ergodically feasible dynamic strategy). Fix a horizon $n \geq 1$. A dynamic strategy is a sequence

$$\sigma = (q^{(1)}, \dots, q^{(n)}), \quad q^{(t)} = (q_f^{(t)})_{f \in \mathcal{F}} \in (\Delta(\mathcal{E}))^{\mathcal{F}}.$$

It is *ergodically feasible* relative to the family $(\mathcal{Q}_f)_{f \in \mathcal{F}}$ if the coordinatewise averages

$$\bar{q}_f := \frac{1}{n} \sum_{t=1}^n q_f^{(t)}$$

satisfy $\bar{q}_f \in \mathcal{Q}_f$ for every $f \in \mathcal{F}$.

For a dynamic strategy σ , let

$$Q_f^{(n)}(\sigma) := \bigotimes_{t=1}^n q_f^{(t)}$$

denote the product evidence law under frame f .

Proposition 6.1 (Dynamic lower bound). *Assume $A_x \cap A_y = \emptyset$ and define*

$$\beta_{xy} := \inf_{p \in A_x, q \in A_y} \left(-\log \sum_{e \in \mathcal{E}} \sqrt{p(e)q(e)} \right) > 0.$$

Then every ergodically feasible dynamic strategy satisfies

$$\frac{1}{n} \min_{f \in F_x, g \in F_y} C(Q_f^{(n)}(\sigma), Q_g^{(n)}(\sigma)) \geq \beta_{xy}.$$

In particular, time variation cannot create aliasing when the static uncertainty sets are separated.

Proof. Fix a cross pair (f, g) . Since Chernoff information dominates the Bhattacharyya exponent,

$$C(Q_f^{(n)}, Q_g^{(n)}) \geq -\log \text{BC}(Q_f^{(n)}, Q_g^{(n)}),$$

where

$$\text{BC}(P, Q) := \sum_z \sqrt{P(z)Q(z)}.$$

For product measures,

$$\text{BC}(Q_f^{(n)}, Q_g^{(n)}) = \prod_{t=1}^n \text{BC}(q_f^{(t)}, q_g^{(t)}),$$

so

$$\frac{1}{n} C(Q_f^{(n)}, Q_g^{(n)}) \geq \frac{1}{n} \sum_{t=1}^n \left(-\log \text{BC}(q_f^{(t)}, q_g^{(t)}) \right).$$

By convexity of $-\log$,

$$\frac{1}{n} \sum_{t=1}^n \left(-\log \text{BC}(q_f^{(t)}, q_g^{(t)}) \right) \geq -\log \left(\frac{1}{n} \sum_{t=1}^n \text{BC}(q_f^{(t)}, q_g^{(t)}) \right).$$

The Bhattacharyya coefficient is jointly concave, hence

$$\frac{1}{n} \sum_{t=1}^n \text{BC}(q_f^{(t)}, q_g^{(t)}) \leq \text{BC}(\bar{q}_f, \bar{q}_g).$$

Therefore

$$\frac{1}{n} C(Q_f^{(n)}, Q_g^{(n)}) \geq -\log \text{BC}(\bar{q}_f, \bar{q}_g).$$

Because $\bar{q}_f \in \mathcal{Q}_f \subseteq A_x$ and $\bar{q}_g \in \mathcal{Q}_g \subseteq A_y$, the right-hand side is at least β_{xy} . Taking the minimum over cross pairs proves the claim. \square

Corollary 6.2 (Actor-side switching costs are irrelevant for the existence boundary). *Suppose a dynamic strategy is additionally subject to any switching-cost budget of the form*

$$\sum_{t=2}^n \kappa_t \mathbf{1}\{q^{(t)} \neq q^{(t-1)}\} \leq B,$$

with arbitrary nonnegative coefficients κ_t and budget $B \geq 0$.

- (a) If $A_x \cap A_y \neq \emptyset$, the aliasing strategy can be chosen constant and therefore incurs zero switching cost.
- (b) If $A_x \cap A_y = \emptyset$, the switching constraint only shrinks the feasible strategy set, so Proposition 6.1 still applies.

Hence actor-side switching costs do not change the existence boundary between aliasing and recovery.

Proof. Part (a) is immediate: if $r \in A_x \cap A_y$, choose a cross pair (f, g) with $r \in \mathcal{Q}_f \cap \mathcal{Q}_g$ and repeat the same family at every time step. Part (b) is immediate because any switching-cost constraint defines a subset of the ergodically feasible strategies. \square

7 Mechanism design under common-preference misalignment

The original manuscript attempted to derive a mechanism-design impossibility from evidence aliasing. That claim is valid only under a preference model in which evidence is actually needed for implementation.

Definition 7.1 (Common-preference misalignment). All types share the same decision utility $u : \{x, y\} \rightarrow \mathbb{R}$ with

$$\Delta := u(x) - u(y) > 0.$$

The social choice function is

$$s(f) = \begin{cases} x, & f \in F_x, \\ y, & f \in F_y. \end{cases}$$

Thus every type prefers x , but the mechanism must still assign y to the F_y -types.

Under this preference model the mechanism must verify whether the realized type lies in F_x or in F_y .

Definition 7.2 (Strict robust implementation). A sequence of mechanisms strictly robustly implements s for the pair (x, y) if, for all sufficiently large n ,

- (i) under truthful reporting, the correct decision $s(f)$ is chosen with probability tending to 1 uniformly over all feasible post-distortion laws for each type; and
- (ii) truthful reporting is strictly better than every misreport, uniformly over all feasible post-distortion laws available to the true type.

Theorem 7.1 (Impossibility at overlap). *If $A_x \cap A_y \neq \emptyset$, then no mechanism sequence strictly robustly implements s .*

Proof. Choose $r \in A_x \cap A_y$ and corresponding types $f^* \in F_x, g^* \in F_y$ with $r \in \mathcal{Q}_{f^*} \cap \mathcal{Q}_{g^*}$. Because the two types have the same utility function over decisions and transfers, their expected utility from any report is identical when the evidence law is r . Therefore the set of optimal reports is the same for both types. Strict incentive compatibility forces the truthful report to be uniquely optimal, so the truthful reports of f^* and g^* must coincide. Under truthful play, the full observables are therefore identical under the two types: the same report and the same evidence law $r^{\otimes n}$. Any mechanism fed identical observables has the same decision distribution under the two types. But social correctness requires asymptotically choosing x under f^* and y under g^* . Contradiction. \square

Theorem 7.2 (Constructive implementation below the boundary). *If $A_x \cap A_y = \emptyset$, then there exists a mechanism sequence that strictly robustly implements s .*

Proof. Use the classifier ψ_n from (2.2), whose error satisfies

$$\varepsilon_n := (n+1)^{|\mathcal{E}|} \exp\left(-\frac{nd_{xy}^2}{8}\right) \rightarrow 0$$

by Theorem 3.1(B). Fix any penalty $P > \Delta$. Let the message space be $M = \{x, y\}$. Truthful reporting is

$$\rho(f) = x \text{ for } f \in F_x, \quad \rho(g) = y \text{ for } g \in F_y.$$

Define the mechanism as follows:

- (i) if the report is y , choose decision y and transfer 0;
- (ii) if the report is x , run the classifier ψ_n on the evidence, choose the classifier's output, and impose transfer $-P$ whenever ψ_n outputs y .

Social correctness. If the true type lies in F_x and reports truthfully, the decision is $\psi_n(E^{(n)})$, which equals x with probability at least $1 - \varepsilon_n$ uniformly over all feasible laws in A_x . If the true type lies in F_y and reports truthfully, the mechanism chooses y with probability 1. **Incentive compatibility for F_x -types.** Under truthful report x , an F_x -type gets expected utility at least

$$u(x) - \varepsilon_n(\Delta + P).$$

A misreport y yields utility exactly $u(y)$. Hence the truthful-report gain is at least

$$\Delta - \varepsilon_n(\Delta + P),$$

which is strictly positive for all sufficiently large n . **Incentive compatibility for F_y -types.** Under truthful report y , an F_y -type gets utility $u(y)$. If such a type misreports x , the classifier outputs x with probability at most ε_n and y with probability at least $1 - \varepsilon_n$, so the deviation utility is at most

$$\varepsilon_n u(x) + (1 - \varepsilon_n)(u(y) - P) = u(y) - P + \varepsilon_n(\Delta + P).$$

Because $P > \Delta$ and $\varepsilon_n \rightarrow 0$, truthful reporting is strictly better for all sufficiently large n . Thus both social correctness and strict robust incentive compatibility hold eventually. \square

Remark 7.1 (What the mechanism theorem does not say). The mechanism result is not a general revelation-principle theorem. It is a two-outcome construction under common-preference misalignment. Its purpose is to show that the verification boundary is not merely statistical; it is exactly the condition under which evidence can or cannot sustain implementation when preferences alone do not.

8 Limits and open problems

The repaired paper leaves three problems open.

Exact minimax exponents beyond KL geometry. Section 5 solves the reverse-KL model exactly. For arbitrary compact uncertainty classes, Theorem 3.1(B) still gives only a universal constructive exponent through the ℓ_1 gap. Determining the exact minimax exponent for general compact uncertainty sets remains open.

Adaptive and interactive adversaries. The static results assume independent and identically distributed evidence conditional on a chosen post-distortion law, and the dynamic corollary assumes ergodic feasibility of the average law. Fully interactive adversaries who adapt to the auditor’s intermediate actions require a game-theoretic extension that is not solved here.

Global aggregation beyond a fixed pair. The paper solves pairwise verification. It does not solve the global social-choice problem over all alternatives simultaneously. Extending pairwise verification boundaries to cycle-free multi-alternative aggregation is an open problem.

9 Conclusion

The paper’s core claim is now narrow and exact. Pairwise social choice under manipulable evidence reduces to robust verification between two compact uncertainty sets. Overlap implies aliasing and a worst-case error floor of $1/2$. Separation implies a universal exponentially consistent classifier. In the reverse-KL model the existence boundary is exactly the Chernoff witness capacity $C_W(x, y)$, and below that boundary the hidden-adversary problem has exact asymptotic exponent $D^*(C_S)$. Dynamic switching costs do not change the existence boundary, and the same criterion governs a two-outcome mechanism-design problem under common-preference misalignment. That is the full claim. It is structural, falsifiable, and useful. It is not a replacement for Arrow, Gibbard–Satterthwaite, robust statistics, or AI alignment.

References

- [1] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, 1952.
- [2] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2006.
- [3] Imre Csiszár and János Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, 1981.
- [4] Moritz Hardt, Megha Raghavan, and Manish Narasimhan. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122, 2016.
- [5] Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, 1964.
- [6] Peter J. Huber and Volker Strassen. Minimax tests and the Neyman–Pearson lemma for capacities. *Annals of Statistics*, 1(2):251–263, 1973.
- [7] Philippe Mongin. Spurious unanimity and the Pareto principle. *Economics and Philosophy*, 13(2):297–320, 1997.