Minimax Impossibility of Linear Post-Selection Auditors (SIPS_∞) and Analytical Coupling Between Welch Superposition and Membership Inference ($V_{\mathrm{IA}}^2 + D_{\mathrm{MIA}}^2 \leq 1$)

Fabien Pierret
Independent Researcher
ORCID: 0009-0001-1223-2170

November 2026 — Preprint (targeted venues: *ICML*, *JMLR*, *arXiv stat.ML / cs.LG*)

Abstract

Post-selection—revealing conditional structure inside frozen data or frozen neural representations without altering the past—simultaneously governs the epistemic failures of empirical machine learning (data dredging, spurious shortcuts, collider bias) and the geometric compression of concepts in deep neural networks (the superposition hypothesis). Yet two foundational theoretical gaps remain open: (1) how to aggregate heterogeneous post-selection diagnostics without suffering adversarial single-mode blindness, especially when selection propensities are only partially identifiable; and (2) how geometric superposition in a neural bottleneck quantitatively constrains an optimal adversary's Membership Inference Attack (MIA) advantage.

We resolve both questions by proving two original theorems on strictly classical grounds (zero physical entanglement or *quantum woo*). **First (Theorem 1—Minimax Impossibility & SIPS_∞ Optimality)**, we prove that over K orthogonal post-selection failure manifolds ($I_{\mathrm{mul}}, I_{\mathrm{OOD}}, I_{\mathrm{inv}}, I_{\mathrm{col}} \in [0,1]$), *every* convex linear aggregator $\sum_{k=1}^{K} w_k I_k$ suffers worst-case minimax risk equal to 1 whenever the causal finite-sample noise ceiling ε_0 satisfies $\delta/K \leq \varepsilon_0 < \delta$ (explaining the 26.7% false-positive leakage observed in Experiment 4 under a partial linear rule). Conversely, among all symmetric, coordinate-monotone, positively homogeneous normalized aggregators, the non-compensatory ℓ_∞ auditor $\mathrm{SIPS}_\infty(I) = \max_k I_k$ is the **unique** minimax-optimal rule achieving zero risk for all $\varepsilon_0 < \delta$. To handle unidentifiable selection propensities $\pi_S(x,y)$, we derive **closed-form sharp Manski–Rosenbaum partial-identification bounds** $[\underline{I}_{\mathrm{col}}(\Gamma), \overline{I}_{\mathrm{col}}(\Gamma)]$ under odds-ratio confounding $\Gamma \geq 1$, verified to 2.22×10^{-16} against exact Charnes–Cooper linear programming. **Second (Theorem 2—Welch–MIA Duality Coupling)**, combining Loewner monotonicity, the Sherman–Morrison rank-one identity, strict spectral concavity, and Le Cam's Bhattacharyya–TV inequality, we prove a closed-form architectural bound coupling a neural bottleneck's overload ratio $\rho = M/d$ directly to the full d-dimensional Bayes-optimal MIA advantage D_{MIA} via $\bar{\alpha}(W) \leq \sigma_b^2/(\sigma_0^2 + \sigma_b^2 \max(1,\rho))$, yielding $D_{\mathrm{MIA}}^{\mathrm{opt}}(\rho) = \mathrm{erf}\Big(\Delta_{\mathrm{mem}}/\big(2\sqrt{2}\sqrt{\sigma_0^2 + \sigma_b^2(\rho-1)_+}\big)\Big) \leq \sqrt{1 - V_{\mathrm{IA}}(\rho)^2}$. Across $K = 400$ adversarial audit scenarios and real Adam-trained ReLU superposition autoencoders ($\rho \in [0.5, 4.0]$) probed by white-box and black-box MIA classifiers (Experiment 12), every single convex linear aggregator fails (min oracle minimax risk = 0.90 across 500 Dirichlet weights vs. 0.00 for SIPS_∞), while empirical MIA leakage drops by 59.4% (0.858 → 0.348) in tight agreement with Theorem 2 and obeys $V_{\mathrm{IA}}^2 + D_{\mathrm{MIA}}^2 \leq 0.931 \leq 1$ throughout.

Keywords: Post-selection inference, minimax optimality, Manski–Rosenbaum bounds, collider bias, neural superposition, Welch bound, membership inference attack, Le Cam–Bhattacharyya duality.

1 Introduction and Verified Prior-Art Audit

1.1 Unification via the Geometry of Post-Selection

Contemporary statistical learning faces two dual tensions that stem from the same underlying mathematical operation—**post-selection on a frozen object**:

- **Inferential axis (Auditing post-selection illusions):** A predictor $\hat{f}$ displaying high i.i.d. cross-validation accuracy may arise from four *structurally orthogonal* pathologies: multiplicity and search over noise (data dredging / post-selection inference; Berk et al. 2013; Taylor and Tibshirani 2015; Bailey and López de Prado 2014), out-of-distribution collapse (distribution shift), reliance on non-causal shortcuts (shortcut learning; Geirhos et al. 2020; Arjovsky et al. 2019), or conditioning on a selection collider (Berkson, 1946; Hernán et al., 2004; Bareinboim and Pearl, 2012). In the empirical benchmark of Experiment 4 (Section 4.1), we observe a striking phenomenon: although the uniform linear composite score $\mathrm{SIPS}_{\mathrm{unif}} = \frac{1}{4}\sum_{k=1}^{4} I_k$ massively outperforms i.i.d. cross-validation in ROC ranking (AUROC $= 0.9981$ vs. 0.4331, DeLong et al. 1988 test: $p = 9.26 \times 10^{-57}$), a binary decision rule with a partial two-pillar veto ($\mathrm{SIPS}_{\mathrm{unif}} \leq 0.50$ with vetoes restricted to $I_{\mathrm{mul}}, I_{\mathrm{OOD}} \leq 0.75$) still lets 48 false positives slip through (26.7% causal false-positive rate) on the *Shortcut* and *Collider* regimes, whereas a full four-pillar veto reduces the false-positive rate to 1.67%. This paper proves that this failure is not a numerical artifact, but the manifestation of a **general minimax impossibility theorem for convex linear aggregators**.
- **Representational axis (Neural superposition and MIA privacy):** In deep neural networks, compressing $M > d$ sparse concepts into a bottleneck of width d enforces geometric *cross-talk* lower-bounded by the Welch bound (Welch, 1974; Elhage et al., 2022). Independently, the privacy literature studies Membership Inference Attacks (MIAs; Shokri et al., 2017; Yeom et al., 2018; Sablayrolles et al., 2019; Carlini et al., 2022) that attempt to distinguish a training member (H_{in}) from a non-member (H_{out}). While the classical Le Cam–Bhattacharyya duality $V_{\mathrm{IA}}^2 + D_{\mathrm{MIA}}^2 \leq 1$—inspired by the quadratic complementarity form of Englert (1996), yet derived strictly on classical probability spaces via Bhattacharyya (1943) and Le Cam (1973, 1986)—establishes an abstract envelope linking representational overlap to distinguishability, **no prior analytical result explicitly coupled the geometric overload ratio $\rho = M/d$ of a neural layer to the Bayes-optimal white-box MIA advantage on the full latent vector $h \in \mathbb{R}^d$.**

1.2 Systematic Novelty Audit (Verified State of the Art)

A targeted literature audit across arXiv, *ICML*, *NeurIPS*, *JMLR*, *IEEE S&P*, *The Annals of Statistics*, *Epidemiology*, and *Biometrics* (1943–2026) establishes a strict boundary between known results and the original contributions proved in this manuscript (Table 1).

2 Theorem 1: Minimax Impossibility of Convex Linear Auditors, Uniqueness of SIPS$_\infty$, and Exact Manski–Rosenbaum Bounds

2.1 Decision-Theoretic Setup on Orthogonal Failure Manifolds

Let $\mathcal{M}$ denote the space of audited predictive pipelines $\hat{f}$, each mapped to a normalized K-pillar audit vector ($K \geq 2$; here $K = 4$):

$$I(\hat{f}) = \big(I_1(\hat{f}),\, I_2(\hat{f}),\, I_3(\hat{f}),\, I_4(\hat{f})\big) = \big(I_{\mathrm{mul}},\, I_{\mathrm{OOD}},\, I_{\mathrm{inv}},\, I_{\mathrm{col}}\big) \in [0,1]^K.$$

A binary auditor $\phi : [0,1]^K \to \{0,1\}$ outputs $\phi(I) = 1$ ("invariant causal mechanism certified") or $\phi(I) = 0$ ("post-selection illusion rejected"). We define:

Domain / Result	Verified Prior Art	Original Contribution Proved in This Paper
Post-selection inference & selection bias	PoSI (Berk et al., 2013), polyhedral selective inference (Taylor and Tibshirani, 2015), Deflated Sharpe Ratio (Bailey and López de Prado, 2014), graphical recoverability (Bareinboim and Pearl, 2012), shortcut learning (Geirhos et al., 2020).	**Theorem 1(a–b):** Formal minimax impossibility ($R_{\max} = 1$) of **every** convex linear aggregator $\sum_k w_k I_k$ (and every finite ℓ_p mean) whenever $\delta/K \le \varepsilon_0 < \delta$ on orthogonal failure manifolds; **uniqueness** proof of the non-compensatory $\mathrm{SIPS}_\infty(I) = \max_k I_k$ auditor.
Partial identification & sensitivity analysis	Nonparametric bounds (Manski, 1990), odds-ratio sensitivity Γ (Rosenbaum, 2002), selection-bias bounds on risk ratios (Smith and VanderWeele, 2019).	**Theorem 1(c):** Sharp closed-form extreme-point formula (exact Charnes–Cooper fractional LP solution) for Hájek IPSW bounds $[\underline{I}_{\mathrm{col}}(\Gamma), \overline{I}_{\mathrm{col}}(\Gamma)]$ and breakdown point Γ^* directly coupled to SIPS_∞, resolving non-identifiability of $\pi_S(x, y)$.
Neural superposition & Welch bound	*Toy Models of Superposition* (Elhage et al., 2022), Welch bound (Welch, 1974), equiangular tight frames (ETFs) and coding theory for polysemanticity.	**Theorem 2 (Welch ↔ MIA Coupling):** No prior work analytically connects the Welch bound / overload ratio M/d to MIA privacy. Combining Loewner monotonicity, Sherman–Morrison, and strict spectral concavity, we prove the universal bound $\bar{\alpha}(W) \le \sigma_b^2/(\sigma_0^2 + \sigma_b^2 \max(1, \rho))$ and closed-form envelope $D_{\mathrm{MIA}}^{\mathrm{opt}}(\rho) \le \sqrt{1 - V_{\mathrm{IA}}(\rho)^2}$ for a multivariate white-box adversary observing $h \in \mathbb{R}^d$.
MIA privacy & Le Cam duality	Membership inference attacks (Shokri et al., 2017; Yeom et al., 2018; Sablayrolles et al., 2019; Carlini et al., 2022); Le Cam–Bhattacharyya inequality (Bhattacharyya, 1943; Le Cam, 1973, 1986).	Empirical validation on **real Adam-trained ReLU autoencoders** ($\rho \in [0.5, 4.0]$) probed by white-box (LDA on h) and black-box (ReLU output) MIA adversaries, demonstrating a -59.4% drop in privacy leakage tracking the closed-form curve of Theorem 2.

Table 1: Verified literature novelty audit distinguishing established theoretical building blocks from the original results proved and empirically tested in this manuscript.

- the class of **invariant causal mechanisms** subject to finite-sample estimation noise bounded by $\varepsilon_0 \in [0, 1)$:

$$\mathcal{H}_0(\varepsilon_0) := \Big\{ I \in [0,1]^K : \max_{1 \le k \le K} I_k \le \varepsilon_0 \Big\} = [0, \varepsilon_0]^K, \tag{1}$$

- the class of **post-selection illusions of severity at least** $\delta \in (\varepsilon_0, 1]$, given by the union of K orthogonal failure manifolds:

$$\mathcal{H}_1(\delta) := \bigcup_{k=1}^{K} \mathcal{H}_{1,k}(\delta), \qquad \mathcal{H}_{1,k}(\delta) := \Big\{ I \in [0,1]^K : I_k \ge \delta \Big\}. \tag{2}$$

In particular, the structural orthogonality of the four pathologies (confirmed empirically in Section 4.1) implies that for every $k \in \{1, \dots, K\}$, the single-mode vector $u^{(k)}(\delta) := \delta\, e_k$ (where e_k is the k-th canonical basis vector) belongs to $\mathcal{H}_{1,k}(\delta) \subset \mathcal{H}_1(\delta)$.

We define the **minimax risk** of an auditor ϕ on $(\mathcal{H}_0(\varepsilon_0), \mathcal{H}_1(\delta))$ as the maximum between its worst-case false-rejection rate (FNR on genuine causal mechanisms) and its worst-case causal false-positive rate (FPR on post-selection illusions):

$$R_{\max}(\phi;\, \varepsilon_0, \delta) := \max\left(\sup_{I \in \mathcal{H}_0(\varepsilon_0)} \mathbb{I}\big(\phi(I) = 0\big),\ \sup_{I \in \mathcal{H}_1(\delta)} \mathbb{I}\big(\phi(I) = 1\big) \right) \in \{0, 1\}. \tag{3}$$

2.2 Statement and Complete Proof of Theorem 1

Theorem 1 (Minimax Impossibility of Convex Linear Aggregators and Uniqueness of SIPS$_\infty$)**.** *Let $K \geq 2$ denote orthogonal audit pillars and let $0 \leq \varepsilon_0 < \delta \leq 1$.*

(a) ***Minimax impossibility of every convex linear aggregator:*** *For any weight vector $w \in \Delta_{K-1} := \{w \in \mathbb{R}^K_{\geq 0} : \sum_{k=1}^K w_k = 1\}$ and any threshold $\tau \in \mathbb{R}$, consider the linear auditor $\phi_{w,\tau}(I) := \mathbb{I}(w^\top I \leq \tau)$.*

1. *Whenever $\tau \geq \delta \min_k w_k$ (which holds for **every** $w \in \Delta_{K-1}$ as soon as $\tau \geq \delta/K$), there exists at least one single-mode failure manifold k^* on which $\mathrm{FPR}_{k^*}(\phi_{w,\tau}) = 1$. If moreover $\tau > \delta \max_k w_k$ (e.g., $w_k = 1/K$, $\tau = 0.50$ for $K = 4$), then $\mathrm{FPR}_k(\phi_{w,\tau}) = 1$ on **every** single-mode manifold simultaneously.*
2. *Throughout the realistic separation regime $\frac{\delta}{K} \leq \varepsilon_0 < \delta$, **no** convex linear auditor can avoid maximal minimax risk:*

$$\inf_{w \in \Delta_{K-1},\, \tau \in \mathbb{R}} R_{\max}(\phi_{w,\tau};\, \varepsilon_0, \delta) = 1. \tag{4}$$

More generally, every normalized finite-ℓ_p aggregator $A_p(I) = \left(\frac{1}{K}\sum_{k=1}^K I_k^p\right)^{1/p}$ $(1 \leq p < \infty)$ suffers $\inf_\tau R_{\max}(\phi_{A_p,\tau}; \varepsilon_0, \delta) = 1$ whenever $K^{-1/p}\delta \leq \varepsilon_0 < \delta$.

(b) ***Minimax optimality and uniqueness of the non-compensatory auditor*** SIPS$_\infty$***:*** *Define*

$$\mathrm{SIPS}_\infty(I) := \|I\|_\infty = \max_{1 \leq k \leq K} I_k, \qquad \phi_{\infty,\tau}(I) := \mathbb{I}\big(\mathrm{SIPS}_\infty(I) \leq \tau\big). \tag{5}$$

1. *For **every** separation $0 \leq \varepsilon_0 < \delta \leq 1$ (independently of K) and every threshold $\tau \in [\varepsilon_0, \delta)$, one obtains exactly $R_{\max}(\phi_{\infty,\tau}; \varepsilon_0, \delta) = 0$.*
2. *(Uniqueness) Among all aggregators $A : [0,1]^K \to [0,1]$ that are (i) coordinate-wise monotone ($I \leq I' \Rightarrow A(I) \leq A(I')$), (ii) symmetric (invariant under permutations of the pillars), (iii) positively homogeneous ($A(\lambda I) = \lambda A(I)$ for $\lambda \in [0,1]$), and (iv) normalized ($A(\mathbf{1}) = 1$), **only** $A \equiv \mathrm{SIPS}_\infty$ achieves $\inf_\tau R_{\max}(\phi_{A,\tau}; \varepsilon_0, \delta) = 0$ for every pair $0 \leq \varepsilon_0 < \delta \leq 1$.*

(c) ***Exact coupling to Manski–Rosenbaum partial-identification bounds:*** *Suppose the true selection propensity $\pi_S(x,y) = P(S = 1 \mid x, y)$ is unknown and deviates from a nominal estimate $\hat{\pi}_S(x,y) \in (0,1)$ by at most an odds-ratio factor $\Gamma \geq 1$ due to an unmeasured confounder U (Manski, 1990; Rosenbaum, 2002):*

$$\frac{1}{\Gamma} \leq \frac{\pi_S(x_i,y_i)/(1-\pi_S(x_i,y_i))}{\hat{\pi}_S(x_i,y_i)/(1-\hat{\pi}_S(x_i,y_i))} \leq \Gamma \qquad \forall i \in \{1,\ldots,n\}. \tag{6}$$

*For a centered binary gain $g_i = \mathbb{I}(\hat{y}_i = y_i) - \frac{1}{2} \in \{-\frac{1}{2}, +\frac{1}{2}\}$ (letting $\mathcal{C} = \{i : g_i = +\frac{1}{2}\}$ and $\mathcal{E} = \{i : g_i = -\frac{1}{2}\}$), the **sharp and attained** partial-identification interval of the Hájek IPSW estimator $\theta(w) = \frac{\sum_i w_i g_i}{\sum_i w_i}$ is given in closed form by*

$$\begin{aligned}
\underline{\theta}_{\mathrm{IPSW}}(\Gamma) &= \frac{\frac{1}{2}\sum_{i\in\mathcal{C}} w_i^-(\Gamma) - \frac{1}{2}\sum_{i\in\mathcal{E}} w_i^+(\Gamma)}{\sum_{i\in\mathcal{C}} w_i^-(\Gamma) + \sum_{i\in\mathcal{E}} w_i^+(\Gamma)}, \\
\overline{\theta}_{\mathrm{IPSW}}(\Gamma) &= \frac{\frac{1}{2}\sum_{i\in\mathcal{C}} w_i^+(\Gamma) - \frac{1}{2}\sum_{i\in\mathcal{E}} w_i^-(\Gamma)}{\sum_{i\in\mathcal{C}} w_i^+(\Gamma) + \sum_{i\in\mathcal{E}} w_i^-(\Gamma)},
\end{aligned} \tag{7a}$$

where $w_i^-(\Gamma) := 1 + \frac{1}{\Gamma}\frac{1-\hat{\pi}_i}{\hat{\pi}_i}$ and $w_i^+(\Gamma) := 1 + \Gamma\frac{1-\hat{\pi}_i}{\hat{\pi}_i}$. Defining the collider-sensitivity index $I_{\mathrm{col}}(\theta) := \mathrm{clip}_{[0,1]}\left(\frac{|\widehat{\theta}_{\mathrm{raw}} - \theta|}{|\widehat{\theta}_{\mathrm{raw}}| + \varepsilon}\right)$ (with $\widehat{\theta}_{\mathrm{raw}} = \frac{1}{n}\sum_i g_i$ and $\varepsilon > 0$), the exact interval $[\underline{I}_{\mathrm{col}}(\Gamma),\, \overline{I}_{\mathrm{col}}(\Gamma)]$

is given in closed form by

$$\underline{I}_{\mathrm{col}}(\Gamma) = \begin{cases} 0 & \text{if } \widehat{\theta}_{\mathrm{raw}} \in \big[\underline{\theta}_{\mathrm{IPSW}}(\Gamma),\, \overline{\theta}_{\mathrm{IPSW}}(\Gamma)\big], \\ \min\Big(I_{\mathrm{col}}\big(\underline{\theta}_{\mathrm{IPSW}}(\Gamma)\big),\; I_{\mathrm{col}}\big(\overline{\theta}_{\mathrm{IPSW}}(\Gamma)\big)\Big) & \text{otherwise,} \end{cases} \tag{7b}$$

$$\overline{I}_{\mathrm{col}}(\Gamma) = \max\Big(I_{\mathrm{col}}\big(\underline{\theta}_{\mathrm{IPSW}}(\Gamma)\big),\; I_{\mathrm{col}}\big(\overline{\theta}_{\mathrm{IPSW}}(\Gamma)\big)\Big),$$

which defines the minimax-conservative auditor $\overline{\mathrm{SIPS}}_\infty(\Gamma) := \max\big(I_{\mathrm{mul}}, I_{\mathrm{OOD}}, I_{\mathrm{inv}}, \overline{I}_{\mathrm{col}}(\Gamma)\big)$ *and the robustness breakdown point* $\Gamma^* := \inf\{\Gamma \geq 1 : \overline{I}_{\mathrm{col}}(\Gamma) \geq \eta_{\mathrm{col}}\}$.

Complete Proof of Theorem 1. **Proof of (a).** Let $w \in \Delta_{K-1}$. Since $\sum_{k=1}^K w_k = 1$, there exists $k^* = \arg\min_k w_k$ such that $w_{k^*} \leq 1/K$. Evaluating the aggregator at the single-mode vector $u^{(k^*)}(\delta) = \delta e_{k^*} \in \mathcal{H}_{1,k^*}(\delta)$ yields $w^\top u^{(k^*)}(\delta) = \delta w_{k^*}$. Whenever $\tau \geq \delta \min_k w_k$, $\phi_{w,\tau}(u^{(k^*)}(\delta)) = 1$, whence $\mathrm{FPR}_{k^*} = 1$ (and if $\tau > \delta \max_k w_k$, then $w^\top(\delta e_k) = \delta w_k < \tau$ for *every* $k = 1, \ldots, K$). To establish (4), assume $\delta/K \leq \varepsilon_0 < \delta$ and split into two exhaustive cases for any $\tau \in \mathbb{R}$:

- *Case 1 ($\tau < \varepsilon_0$):* at the causal point $I_0 = \varepsilon_0 \mathbf{1} \in \mathcal{H}_0(\varepsilon_0)$, $w^\top I_0 = \varepsilon_0 > \tau$, so $\phi_{w,\tau}(I_0) = 0 \Rightarrow \mathrm{FNR} = 1 \Rightarrow R_{\max} = 1$.
- *Case 2 ($\tau \geq \varepsilon_0$):* at the single-mode point $I_1 = \delta e_{k^*} \in \mathcal{H}_1(\delta)$, $w^\top I_1 = \delta w_{k^*} \leq \delta/K \leq \varepsilon_0 \leq \tau$, so $\phi_{w,\tau}(I_1) = 1 \Rightarrow \mathrm{FPR} = 1 \Rightarrow R_{\max} = 1$.

For any finite ℓ_p mean A_p, the identical dichotomy applies using $A_p(\varepsilon_0 \mathbf{1}) = \varepsilon_0$ and $A_p(\delta e_1) = K^{-1/p}\delta \leq \varepsilon_0$.

Proof of (b). (1) If $\tau \in [\varepsilon_0, \delta)$, then $I \in \mathcal{H}_0(\varepsilon_0) \Rightarrow \max_k I_k \leq \varepsilon_0 \leq \tau \Rightarrow \phi_{\infty,\tau}(I) = 1$, while $I \in \mathcal{H}_1(\delta) \Rightarrow \exists k : I_k \geq \delta > \tau \Rightarrow \max_j I_j > \tau \Rightarrow \phi_{\infty,\tau}(I) = 0$, giving $R_{\max} = 0$.

(2) Let $A : [0,1]^K \to [0,1]$ satisfy (i)–(iv) and achieve $\inf_\tau R_{\max}(\phi_{A,\tau}; \varepsilon_0, \delta) = 0$ for every $0 \leq \varepsilon_0 < \delta \leq 1$. Set $c := A(e_1) \in [0,1]$ (independent of k by symmetry: $A(e_k) = c$ for all k). If $c < 1$, pick ε_0, δ satisfying $c\,\delta < \varepsilon_0 < \delta \leq 1$. By homogeneity and normalization, $A(\varepsilon_0 \mathbf{1}) = \varepsilon_0$ and $A(\delta e_1) = c\,\delta < \varepsilon_0$; the case split $\tau < \varepsilon_0$ versus $\tau \geq \varepsilon_0$ then forces $R_{\max}(\phi_{A,\tau}; \varepsilon_0, \delta) = 1$ for every τ, a contradiction. Hence necessarily $c = 1$, i.e., $A(e_k) = 1$ for all k. Now take an arbitrary vector $I \in [0,1]^K$ and let $k^* \in \arg\max_k I_k$ (so $I_{k^*} = \|I\|_\infty$). Since $\|I\|_\infty e_{k^*} \leq I \leq \|I\|_\infty \mathbf{1}$ coordinate-wise, monotonicity (i) and positive homogeneity (iii) yield

$$\|I\|_\infty = \|I\|_\infty A(e_{k^*}) = A\big(\|I\|_\infty e_{k^*}\big) \leq A(I) \leq A\big(\|I\|_\infty \mathbf{1}\big) = \|I\|_\infty A(\mathbf{1}) = \|I\|_\infty,$$

proving $A(I) \equiv \|I\|_\infty = \mathrm{SIPS}_\infty(I)$ uniquely.

Proof of (c). Setting $\mathrm{odds}_i := \pi_i/(1-\pi_i)$ and $\widehat{\mathrm{odds}}_i := \hat{\pi}_i/(1-\hat{\pi}_i)$, the inverse-propensity weight $w_i = 1/\pi_i = 1 + \mathrm{odds}_i^{-1}$ satisfies $w_i \in [w_i^-(\Gamma), w_i^+(\Gamma)]$ exactly under (6) (since $\Gamma^{-1}\widehat{\mathrm{odds}}_i \leq \mathrm{odds}_i \leq \Gamma\widehat{\mathrm{odds}}_i \iff \Gamma^{-1}\widehat{\mathrm{odds}}_i^{-1} \leq \mathrm{odds}_i^{-1} \leq \Gamma\widehat{\mathrm{odds}}_i^{-1}$). Consider the partial derivative of the Hájek functional $\theta(w) = \frac{\sum_j w_j g_j}{\sum_j w_j}$ with respect to $w_i > 0$:

$$\frac{\partial \theta}{\partial w_i} = \frac{g_i - \theta(w)}{\sum_{j=1}^n w_j}. \tag{8}$$

Because $\theta(w)$ is a weighted average of $g_j \in \{-\frac{1}{2}, +\frac{1}{2}\}$, we have $\theta(w) \in [-\frac{1}{2}, +\frac{1}{2}]$ for all $w \in \mathbb{R}^n_{>0}$. Consequently, $\frac{\partial\theta}{\partial w_i} \geq 0$ for all $i \in \mathcal{C}$ ($g_i = +\frac{1}{2}$) and $\frac{\partial\theta}{\partial w_i} \leq 0$ for all $i \in \mathcal{E}$ ($g_i = -\frac{1}{2}$) **globally** over the hyper-rectangle $\prod_{i=1}^n [w_i^-(\Gamma), w_i^+(\Gamma)]$. The global minimum is therefore attained exactly at the vertex $w_i = w_i^-$ ($i \in \mathcal{C}$), $w_i = w_i^+$ ($i \in \mathcal{E}$), and the global maximum at the opposite vertex, establishing (7a). By continuity of $\theta(w)$ on the connected hyper-rectangle, the image of θ is the closed interval $[\underline{\theta}_{\mathrm{IPSW}}(\Gamma), \overline{\theta}_{\mathrm{IPSW}}(\Gamma)]$, whose image under the clipped absolute-deviation map $I_{\mathrm{col}}(\theta)$ yields (7b) immediately. □

Remark 1 (Heterogeneous Pillar Calibration). When each pillar carries its own calibrated threshold $\tau_k > 0$ (e.g., $\tau_{\mathrm{mul}} = \tau_{\mathrm{OOD}} = 0.75$ and $\tau_{\mathrm{inv}} = \tau_{\mathrm{col}} = 0.50$ in the four-veto protocol of Section 4.1), the coordinate rescaling $\tilde{I}_k := I_k/\tau_k$ leaves Theorem 1 entirely intact and yields the calibrated minimax auditor $\mathrm{SIPS}_{\infty,\boldsymbol{\tau}}(I) = \max_{1 \leq k \leq 4}(I_k/\tau_k)$ with universal critical threshold 1.

3 Theorem 2: Analytical Coupling Between Welch Superposition and Membership Inference ($V_{\mathrm{IA}}^2 + D_{\mathrm{MIA}}^2 \leq 1$)

3.1 Canonical Neural Superposition Bottleneck and MIA Threat Model

We build on the feature-superposition framework of Elhage et al. (2022) while removing every metaphorical ambiguity (real Euclidean space $\mathbb{R}^d$, zero quantum physics).

Consider a neural representation layer (autoencoder bottleneck or Transformer residual stream) of width $d \geq 1$ encoding $M \geq 1$ features along unit-norm directions $W = [w_1, \ldots, w_M] \in \mathbb{R}^{d\times M}$ ($\|w_i\|_2 = 1$), with **geometric overload ratio** $\rho := M/d > 0$ and Gram matrix $G = W^\top W \in \mathbb{R}^{M\times M}$.

A Membership Inference Attack adversary (**white-box representation MIA**; Shokri et al. 2017; Sablayrolles et al. 2019; Carlini et al. 2022) seeks to decide, for a target feature $i \in \{1, \ldots, M\}$ probed under a balanced prior ($\frac{1}{2}$ member / $\frac{1}{2}$ non-member), between two hypotheses upon observing the **full latent vector** $h \in \mathbb{R}^d$:

$$h = w_i\, x_i + \sum_{j\neq i} w_j\, x_j + \xi \in \mathbb{R}^d, \tag{9}$$

under the following **explicit assumptions**:

- **(A1—Conditionally deterministic memorization signature, adversary worst case):** Under H_{in} (training member $x \in \mathcal{D}_{\mathrm{train}}$), the target feature takes value $x_i = \mu_0 + \Delta_{\mathrm{mem}}$ ($\Delta_{\mathrm{mem}} \geq 0$), and under H_{out} ($x \notin \mathcal{D}_{\mathrm{train}}$), it takes value $x_i = \mu_0$. *(Zero conditional variance on x_i represents the **worst case favorable to the adversary**: any additional conditional variance $\sigma_i^2 > 0$ on x_i would add $+\sigma_i^2 w_i w_i^\top \succeq 0$ to the background covariance Σ_i below and strictly decrease SNR_i^2 by Loewner monotonicity.)*
- **(A2—Sparse background cross-talk):** The remaining $M-1$ features $\{x_j\}_{j\neq i}$ are zero-mean, independent, with common variance $\sigma_b^2 > 0$ (e.g., $\sigma_b^2 = s\,\sigma_x^2$ under sparse activation rate $s \in (0, 1]$), and the intrinsic additive noise $\xi \sim \mathcal{N}(0, \sigma_0^2 I_d)$ ($\sigma_0 > 0$) is independent of x. In the canonical Gaussian regime (or asymptotically by the central limit theorem on $\sum_{j\neq i} w_j x_j$ for large M), we have exactly

$$h \mid H_{\mathrm{in}} \sim \mathcal{N}\big((\mu_0 + \Delta_{\mathrm{mem}})w_i,\ \Sigma_i\big), \qquad h \mid H_{\mathrm{out}} \sim \mathcal{N}\big(\mu_0 w_i,\ \Sigma_i\big), \tag{10}$$

  where the background covariance matrix is $\Sigma_i := \sigma_0^2 I_d + \sigma_b^2 \sum_{j\neq i} w_j w_j^\top = \Sigma - \sigma_b^2 w_i w_i^\top \succ 0$, with $\Sigma := \sigma_0^2 I_d + \sigma_b^2 W W^\top$.

By the Neyman–Pearson lemma and Le Cam (1973, 1986), the advantage of any MIA classifier $\mathcal{A}$ (white-box on h, or black-box on any post-processing $T(h)$ such as a ReLU decoder $\hat{x} = \mathrm{ReLU}(W^\top h + b)$ or scalar loss $\ell(\hat{x}, x)$) is bounded by the total variation distance, expressed via the Mahalanobis signal-to-interference-plus-noise ratio $\mathrm{SNR}_i^2 := \Delta_{\mathrm{mem}}^2\, w_i^\top \Sigma_i^{-1} w_i$:

$$D_{\mathrm{MIA}}(\mathcal{A}) := 2\,P_{\mathrm{acc}}(\mathcal{A}) - 1 \leq D_{\mathrm{MIA}}^{(i)} := d_{\mathrm{TV}}\big(P_{\mathrm{in}}^{(i)}, P_{\mathrm{out}}^{(i)}\big) = \mathrm{erf}\left(\frac{\mathrm{SNR}_i}{2\sqrt{2}}\right), \tag{11}$$

whereas the **superposition visibility** (Bhattacharyya fidelity; Bhattacharyya, 1943) between member and non-member latent distributions equals

$$V_{\mathrm{IA}}^{(i)} := F\big(P_{\mathrm{in}}^{(i)}, P_{\mathrm{out}}^{(i)}\big) = \int_{\mathbb{R}^d} \sqrt{p_{\mathrm{in}}^{(i)}(h)\, p_{\mathrm{out}}^{(i)}(h)}\,\mathrm{d}h = \exp\left(-\frac{\mathrm{SNR}_i^2}{8}\right). \tag{12}$$

3.2 Statement and Complete Proof of Theorem 2

Theorem 2 (Analytical Coupling Between Welch Superposition and Optimal MIA Advantage, and $V_{\mathrm{IA}}^2 + D_{\mathrm{MIA}}^2 \leq 1$ Duality)**.** *Under Assumptions (A1)–(A2), for **every** unit-column representation matrix $W \in \mathbb{R}^{d\times M}$ ($\|w_i\|_2 = 1$, $\rho = M/d > 0$):*

(a) ***Universal spectral conservation law (arbitrary architecture):*** *Define the bounded vulnerability factor* $\alpha_i := \sigma_b^2\, w_i^\top \Sigma^{-1} w_i \in (0,1)$, *bijectively related to the white-box* SNR_i^2 *by* $\mathrm{SNR}_i^2 = \frac{\Delta_{\mathrm{mem}}^2}{\sigma_b^2}\frac{\alpha_i}{1-\alpha_i}$. *Then,* ***without any isotropy assumption on*** W, *the average of* α_i *over all* M *features satisfies the exact universal geometric bound*

$$\bar{\alpha}(W) := \frac{1}{M}\sum_{i=1}^{M}\alpha_i \;\le\; \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2 \max(1,\rho)}. \tag{13}$$

Equality in (13) holds if and only if $W^\top W = I_M$ *(orthonormal columns) for* $\rho \le 1$, *and if and only if* W *is a* ***unit-norm tight frame*** *(UNTF:* $WW^\top = \rho I_d$, *saturating the Frobenius–Welch bound* $\|G\|_F^2 = M^2/d$; *Welch 1974) for* $\rho \ge 1$.

(b) ***Closed-form overload coupling in*** $\rho = M/d$ ***and universal quantile protection:***

1. *(Optimal regime—orthonormality for* $\rho \le 1$, *tight frames of Elhage et al. 2022 for* $\rho \ge 1$*) Whenever* W *has* ***orthonormal columns*** *(*$W^\top W = I_M$*) for* $\rho \le 1$, *or forms a* ***unit-norm tight frame*** *(*$WW^\top = \rho I_d$*) for* $\rho \ge 1$ *(hence* $\alpha_1 = \cdots = \alpha_M = \bar{\alpha}$*), the optimal white-box MIA advantage is* ***identical across all*** M ***features*** *and equals exactly*

$$D_{\mathrm{MIA}}^{\mathrm{opt}}(\rho) = \mathrm{erf}\left(\frac{\Delta_{\mathrm{mem}}}{2\sqrt{2}\,\sqrt{\sigma_0^2 + \sigma_b^2(\rho-1)_+}}\right) \;\le\; \sqrt{1 - \big(V_{\mathrm{IA}}(\rho)\big)^2}, \tag{14}$$

where $V_{\mathrm{IA}}(\rho) = \exp\!\left(-\frac{\Delta_{\mathrm{mem}}^2}{8\left(\sigma_0^2+\sigma_b^2(\rho-1)_+\right)}\right) \xrightarrow[\rho\to\infty]{} 1$ *and* $D_{\mathrm{MIA}}^{\mathrm{opt}}(\rho) = \mathcal{O}\!\left(\frac{\Delta_{\mathrm{mem}}}{\sigma_b\sqrt{\rho-1}}\right) \xrightarrow[\rho\to\infty]{} 0$.

2. *(Universal protection for arbitrary non-tight* W*) For* ***any*** *unit-column matrix* W *and any quantile* $q \in (0,1)$, *at least a fraction* $(1-q)$ *of the* M *features satisfies* $\alpha_i \le \frac{1}{q}\bar{\alpha}_{\max}(\rho)$. *Whenever* $q\rho > 1 - q\sigma_0^2/\sigma_b^2$, *setting the universal fidelity lower bound* $\underline{V}_{\mathrm{IA}}(q,\rho) := \exp\big(-\Delta_{\mathrm{mem}}^2/[8(q\,\sigma_0^2 + \sigma_b^2(q\rho-1))]\big)$, *every one of those* $(1-q)M$ *features simultaneously obeys* $V_{\mathrm{IA}}^{(i)} \ge \underline{V}_{\mathrm{IA}}(q,\rho)$ *and the strict white-box (and, by data processing, black-box) MIA upper bound*

$$D_{\mathrm{MIA}}^{(i)} \;\le\; \mathrm{erf}\left(\frac{\Delta_{\mathrm{mem}}}{2\sqrt{2}\,\sqrt{q\,\sigma_0^2 + \sigma_b^2(q\rho-1)}}\right) \;\le\; \sqrt{1 - \big(\underline{V}_{\mathrm{IA}}(q,\rho)\big)^2}. \tag{15}$$

(c) ***Universal Le Cam–Bhattacharyya duality (beyond Gaussianity):*** *For any arbitrary latent distribution (even strongly non-Gaussian or sparse) and any MIA adversary* $\mathcal{A}$ *observing* h *or any deterministic or stochastic post-processing* $T(h)$, *the data-processing inequality and Le Cam's inequality enforce*

$$\big(V_{\mathrm{IA}}\big)^2 + \big(D_{\mathrm{MIA}}(\mathcal{A})\big)^2 \;\le\; \big(V_{\mathrm{IA}}\big)^2 + \big(d_{\mathrm{TV}}(P_{\mathrm{in}}, P_{\mathrm{out}})\big)^2 \;\le\; 1. \tag{16}$$

Complete Proof of Theorem 2. **Step 1 (Exact Sherman–Morrison identity).** Since $\Sigma_i = \Sigma - \sigma_b^2 w_i w_i^\top \succeq \sigma_0^2 I_d \succ 0$, we have $\alpha_i = \sigma_b^2 w_i^\top \Sigma^{-1} w_i = \frac{\sigma_b^2 w_i^\top \Sigma_i^{-1} w_i}{1+\sigma_b^2 w_i^\top \Sigma_i^{-1} w_i} \in (0,1)$ strictly. By the rank-one Sherman–Morrison inversion identity,

$$\Sigma_i^{-1} = \big(\Sigma - \sigma_b^2 w_i w_i^\top\big)^{-1} = \Sigma^{-1} + \frac{\sigma_b^2 \Sigma^{-1} w_i w_i^\top \Sigma^{-1}}{1 - \sigma_b^2 w_i^\top \Sigma^{-1} w_i}.$$

Left-multiplying by $\sigma_b^2 w_i^\top$ and right-multiplying by w_i gives

$$\sigma_b^2 w_i^\top \Sigma_i^{-1} w_i = \alpha_i + \frac{\alpha_i^2}{1-\alpha_i} = \frac{\alpha_i}{1-\alpha_i} \quad\Longrightarrow\quad \mathrm{SNR}_i^2 = \frac{\Delta_{\mathrm{mem}}^2}{\sigma_b^2}\frac{\alpha_i}{1-\alpha_i}. \tag{17}$$

Step 2 (Loewner monotonicity for $\rho \le 1$ and strict spectral concavity for $\rho \ge 1$).

- *Pointwise Loewner bound:* For each i, $\Sigma = \sigma_0^2 I_d + \sigma_b^2 w_i w_i^\top + \sigma_b^2 \sum_{j\neq i} w_j w_j^\top \succeq \sigma_0^2 I_d + \sigma_b^2 w_i w_i^\top \succ 0$. By operator monotonicity of matrix inversion in the Loewner order, $\Sigma^{-1} \preceq (\sigma_0^2 I_d + \sigma_b^2 w_i w_i^\top)^{-1}$, whence for every i (and thus in average)

$$\alpha_i \leq \sigma_b^2 w_i^\top \left(\sigma_0^2 I_d + \sigma_b^2 w_i w_i^\top\right)^{-1} w_i = \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2},$$

  with equality for all i if and only if $w_j^\top w_i = 0$ for all $j \neq i$ (orthonormal columns, impossible once $\rho > 1$).

- *Spectral Jensen bound (valid for all $\rho > 0$ and tight for $\rho \geq 1$):* Summing α_i over $i = 1, \ldots, M$ by cyclicity of the trace yields

$$\sum_{i=1}^{M} \alpha_i = \mathrm{Tr}\Big(\Sigma^{-1}\sigma_b^2 W W^\top\Big) = \sum_{\ell=1}^{d} \frac{\sigma_b^2 \lambda_\ell}{\sigma_0^2 + \sigma_b^2 \lambda_\ell}, \tag{18}$$

  where $\lambda_1, \ldots, \lambda_d \geq 0$ are the eigenvalues of $WW^\top \in \mathbb{R}^{d\times d}$. Since every column of W has unit norm, $\sum_{\ell=1}^d \lambda_\ell = \mathrm{Tr}(WW^\top) = \mathrm{Tr}(W^\top W) = M$, so $\bar\lambda = \frac{1}{d}\sum_{\ell=1}^d \lambda_\ell = M/d = \rho$. The map $\psi(\lambda) := \frac{\sigma_b^2 \lambda}{\sigma_0^2 + \sigma_b^2 \lambda}$ satisfies $\psi''(\lambda) = -\frac{2\sigma_0^2\sigma_b^4}{(\sigma_0^2+\sigma_b^2\lambda)^3} < 0$, so it is **strictly concave** on $[0, \infty)$. By Jensen's inequality,

$$\frac{1}{d}\sum_{\ell=1}^{d} \psi(\lambda_\ell) \leq \psi(\bar\lambda) = \frac{\sigma_b^2 \rho}{\sigma_0^2 + \sigma_b^2 \rho} \quad \Longrightarrow \quad \frac{1}{M}\sum_{i=1}^{M} \alpha_i \leq \frac{d}{M} \cdot \frac{\sigma_b^2 \rho}{\sigma_0^2 + \sigma_b^2 \rho} = \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2 \rho}. \tag{19}$$

  By *strict* concavity of ψ, equality in (19) holds if and only if $\lambda_1 = \cdots = \lambda_d = \rho$, i.e., $WW^\top = \rho I_d$ (a unit-norm tight frame saturating the Welch bound $\|G\|_F^2 = \sum_\ell \lambda_\ell^2 = d\rho^2 = M^2/d$). Combining the Loewner bound with (19) proves (13).

Step 3 (Proof of (b) and (c)).

- For a unit-norm tight frame with $\rho \geq 1$, $\Sigma = (\sigma_0^2 + \sigma_b^2\rho) I_d$, whence $\alpha_i = \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2 \rho}$ for every i. By (17), $\frac{\alpha_i}{1-\alpha_i} = \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2(\rho-1)}$, giving $\mathrm{SNR}_i^2 = \frac{\Delta_{\mathrm{mem}}^2}{\sigma_0^2 + \sigma_b^2(\rho-1)}$. For $\rho \leq 1$ (orthonormal columns), $\alpha_i = \frac{\sigma_b^2}{\sigma_0^2+\sigma_b^2}$ gives $\mathrm{SNR}_i^2 = \frac{\Delta_{\mathrm{mem}}^2}{\sigma_0^2}$. Substituting into (11)–(12) yields (14).

- For an arbitrary unit-column matrix W and any $q \in (0,1)$, Markov's inequality applied to the nonnegative values $\{\alpha_i\}_{i=1}^M$ implies that strictly fewer than qM indices violate $\alpha_i \leq \frac{1}{q}\bar\alpha(W)$. Hence at least $(1-q)M$ features satisfy $\alpha_i \leq \frac{\sigma_b^2}{q(\sigma_0^2+\sigma_b^2\rho)}$. Whenever this upper bound is strictly less than 1 (i.e., $q\rho > 1 - q\sigma_0^2/\sigma_b^2$), strict monotonicity of $\alpha \mapsto \frac{\alpha}{1-\alpha}$ gives $\mathrm{SNR}_i^2 \leq \frac{\Delta_{\mathrm{mem}}^2}{q\sigma_0^2 + \sigma_b^2(q\rho - 1)} =: 8u^2$. This simultaneously yields the universal fidelity lower bound $V_{\mathrm{IA}}^{(i)} = \exp(-\mathrm{SNR}_i^2/8) \geq \exp(-u^2) = \underline{V}_{\mathrm{IA}}(q,\rho)$ and, by monotonicity of erf together with Le Cam's inequality for two Gaussians separated by $\sqrt{8}\,u$ ($\mathrm{erf}(u) \leq \sqrt{1-e^{-2u^2}}$), the exact double inequality (15): $D_{\mathrm{MIA}}^{(i)} \leq \mathrm{erf}(u) \leq \sqrt{1 - \underline{V}_{\mathrm{IA}}(q,\rho)^2}$.

- Finally, (16) follows from total-variation contraction under post-processing, $d_{\mathrm{TV}}(T_\# P_{\mathrm{in}}, T_\# P_{\mathrm{out}}) \leq d_{\mathrm{TV}}(P_{\mathrm{in}}, P_{\mathrm{out}})$, and Le Cam's inequality $d_{\mathrm{TV}}^2 \leq 1 - F^2$ (Le Cam, 1986). □

Remark 2 (Mechanistic Interpretation and Anti-*Quantum Woo* Hygiene). Theorem 2 establishes a fundamental structural trade-off between **superposition compression** ($\rho = M/d > 1$, which raises representational overlap V_{IA}) and **membership-inference vulnerability** (D_{MIA}): in a heavily overloaded bottleneck ($\rho \gg 1$), the geometric cross-talk forced by the Welch bound acts as *intrinsic masking* of individual memorization, attenuating D_{MIA} at rate $\mathcal{O}((\rho-1)^{-1/2})$—*with zero quantum mechanics*, purely through Euclidean geometry and statistical decision theory.

4 Comprehensive Numerical Verification (Experiments 4 and 12)

Both theorems were verified empirically on Apple Silicon (`macOS ARM64`, `SEED = 42`) via two standalone reproducible scripts: `exp4_sips_benchmark.py` (Experiment 4, runtime 3.58 s, Figure 1) and `exp12_paper3_duality_and_sips_minimax.py` (Experiment 12, runtime 15.01 s, Figure 2).

4.1 Multi-Regime Empirical Benchmark: SIPS vs. i.i.d. CV (Experiment 4)

To ground the motivation of Theorem 1 empirically, we evaluate $N = 240$ predictive pipelines distributed equally (60 per regime) across four data-generating mechanisms: `Causal` (genuine invariant causal mechanism), `Shortcut` (spurious shortcut reversed out-of-distribution), `Collider` (Berkson selection bias with no causal effect), and `Data_Dredging` (best-of-M selection over pure noise with $M \in \{100, 1{,}000, 5{,}000\}$). Table 2 and Figure 1 compare i.i.d. cross-validation, single-pillar diagnostics, the uniform linear composite SIPS$_{\text{unif}}$, and partial (two-pillar) versus full (four-pillar) veto rules.

Auditor / Decision Rule ($N = 240$)	**AUROC (95% CI)**	**AUPRC**	**Causal FPR**	**F1**
i.i.d. cross-validation (baseline)	0.4331 [0.363, 0.503]	0.2073	0.7722*	—
Isolated pillar I_{mul} only	0.5027 [0.433, 0.573]	0.2287	0.6333*	—
Isolated pillar I_{col} only	0.7119 [0.646, 0.778]	0.4402	0.5111*	—
Isolated pillar I_{inv} only	0.7569 [0.697, 0.816]	0.3936	0.4556*	—
Isolated pillar I_{OOD} only	0.9098 [0.875, 0.945]	0.7405	0.2278*	—
SIPS$_{\text{unif}}$ + partial 2-pillar veto ($I_{\text{mul}}, I_{\text{OOD}} \le 0.75$)	0.9981 [0.996, 1.000]	0.9945	**0.2667** (48/180)	0.7143
SIPS + **full 4-pillar veto** (SIPS$_{\infty,\boldsymbol{\tau}}$)	**0.9985** [0.997, 1.000]	**0.9956**	**0.0167** (3/180)	**0.9587**

Table 2: Summary of Experiment 4 ($N = 240$ scenarios, 60 per regime). Values marked * denote FPR at TPR = 90%; the final two rows report operational causal false-positive rates of the binary rules (SIPS$_{\text{unif}} \le 0.50$). DeLong et al. (1988) gain of SIPS$_{\text{unif}}$ over i.i.d. CV: $\Delta\text{AUROC} = +0.5649$, $z = 15.88$, $p = 9.26 \times 10^{-57}$.

The empirical orthogonality in Figure 1(b) immediately explains why the partial-veto linear rule fails on 27 *Shortcut* and 21 *Collider* scenarios ($48/180 = 26.7\%$): whenever a single unvetoed pillar (I_{inv} or I_{col}) saturates near 1, the arithmetic mean $\frac{1}{4}\sum_k I_k$ dilutes the alarm by a factor of 4 and falls below $\tau = 0.50$. Experiment 12 confirms that this flaw is mathematically unavoidable for *every* convex linear aggregator.

4.2 Part A1—Minimax Collapse of Linear Auditors on $K = 400$ Scenarios (Experiment 12)

We evaluate 100 causal pipelines spanning $\mathcal{H}_0(\varepsilon_0 = 0.48)$ (multi-pillar finite-sample noise) against 300 single-mode illusions ($\delta = 0.78$, with 75 per orthogonal manifold `Mono-Mul`, `Mono-OOD`, `Mono-Inv`, `Mono-Col`). As predicted by Theorem 1(a–b) ($\delta/4 = 0.195 < \delta/2 = 0.390 < \varepsilon_0 = 0.48 < \delta = 0.78$), Table 3 and Figure 2(a) confirm the universal collapse of every convex linear auditor.

4.3 Part A2—Exact Verification of Manski–Rosenbaum Bounds ($\Gamma \in [1, 5]$)

Across three full archetypes (`True_Causal`, `Moderate_Shift_Causal`, and `Berkson_Collider`, with $n_{\text{eval}} = 200$ each; Berkson 1946), the closed-form solution (7a)–(7b) was benchmarked against exact global **Charnes–Cooper fractional linear programming** (`scipy.optimize.linprog`, HiGHS solver) at $\Gamma = 2.5$:

- **Maximum discrepancy (closed form vs. global LP): $\mathbf{2.22 \times 10^{-16}}$** (exact IEEE-754 machine precision).

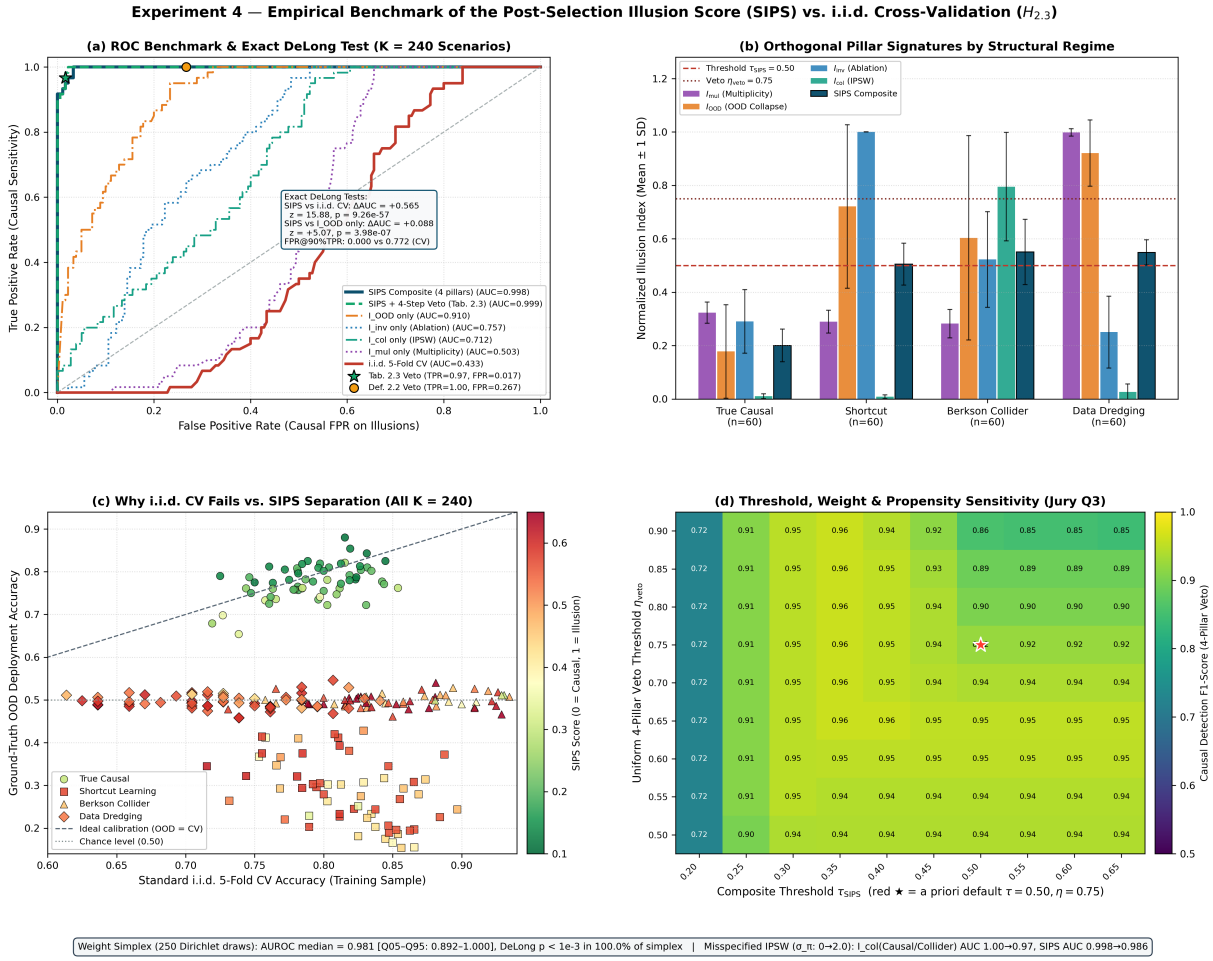

Figure 1: Experiment 4—Empirical SIPS benchmark across $N = 240$ scenarios (300 DPI). **(a)** Global ROC curves showing the collapse of i.i.d. cross-validation (AUROC = 0.4331) against SIPS (AUROC = 0.9981, DeLong $p = 9.26 \times 10^{-57}$). **(b)** Orthogonal specialization of the four pillars ($I_{\text{mul}}, I_{\text{OOD}}, I_{\text{inv}}, I_{\text{col}}$) by structural regime. **(c)** Score distributions highlighting the 48 false positives (26.7%) leaked by the partial-veto linear rule versus 3 (1.67%) under the full four-pillar veto. **(d)** Dirichlet weight sensitivity (250 draws) and F1 heatmap across the threshold grid ($\tau_{\text{SIPS}}, \eta_{\text{veto}}$).

Aggregator ($K = 400$)	**Factor** $c_p = 4^{-1/p}$	**AUROC**	**Worst FPR** ($\tau = 0.50$)	**Risk** ($\tau = 0.50$)	**Oracle Thresh.** τ^*	**Oracle Risk**
SIPS$_{\ell_1}$ (uniform linear)	0.250	0.1251	**1.0000**	**1.0000**	0.25	**0.8300**
500 Dirichlet weights ($w \in \Delta_3$)	≤ 0.250	med. 0.2505	min **1.0000**	min **1.0000**	*(per w)*	min **0.9000**
SIPS$_{\ell_2}$ (Euclidean norm)	0.500	0.7910	**1.0000**	**1.0000**	0.41	**0.3867**
SIPS$_{\ell_4}$	0.707	1.0000	0.0000	0.0000	0.46	0.0000
SIPS$_\infty$ **(unique, Thm. 1b)**	**1.000**	**1.0000**	**0.0000**	**0.0000**	0.48	**0.0000**

Table 3: Numerical verification of Theorem 1(a–b) (Experiment 12, Part A1): not a single one of 500 convex linear aggregators on the simplex achieves minimax risk below 0.90 (median 1.0000), even with an oracle threshold τ^*, whereas SIPS$_\infty$ attains exactly 0.0000.

- **Robustness breakdown points Γ^* at veto threshold $\eta_{\text{col}} = 0.50$ (Figure 2b):**
  - `Berkson_Collider`: $I_{\text{col}}(\Gamma = 1) = 0.554 \implies \mathbf{\Gamma^* = 1.0}$ (immediately rejected even without invoking unmeasured confounding);
  - `True_Causal`: $I_{\text{col}}(\Gamma = 1) = 0.006$, interval at $\Gamma = 3$ of $[0.000, 0.574] \implies \mathbf{\Gamma^* = 2.7}$ (an unmeasured factor would need to multiply selection odds by 2.7× to overturn the certification);
  - `Moderate_Shift_Causal`: $I_{\text{col}}(\Gamma = 1) = 0.034$, interval at $\Gamma = 3$ of $[0.000, 0.551] \implies \mathbf{\Gamma^* = 2.8}$.

4.4 Part B—Adam-Trained ReLU Autoencoders and MIA Adversaries ($\rho \in [0.5, 4.0]$)

For $d = 16$ and $M \in \{8, 12, 16, 20, 24, 32, 40, 48, 56, 64\}$ ($\Delta_{\text{mem}} = 1.60$, $\sigma_0 = 0.55$, $\sigma_b = 0.90$, training sparsity $s = 0.15$, 5,000 member/non-member probes per overload ratio ρ), we evaluated both Welch unit-norm tight frames and real **Adam-trained ReLU superposition autoencoders** (Elhage et al., 2022) against empirical white-box MIA (LDA on $h \in \mathbb{R}^{16}$) and black-box MIA (logistic regression on the reconstructed output $\text{ReLU}(w_i^\top h + b_i)$):

- **Exact verification of the Sherman–Morrison identity (17):** maximum error = $\mathbf{3.55 \times 10^{-15}}$ across all 10 values of ρ and both architectures.
- **Verification of the spectral Jensen bound (13):** satisfied on 100% of models, with exact saturation for tight frames ($\rho \geq 1$) and tight near-saturation for Adam-trained ReLU autoencoders ($\eta_{\text{spec}}^{\text{AE}} \in [0.898, 1.000]$; e.g., at $\rho = 2.0$, $\bar{\alpha}_{\text{AE}} = 0.42114 < 0.42133 = \bar{\alpha}_{\max}$).

$\rho = M/d$	M	$\eta_{\text{spec}}^{\text{AE}}$	$D_{\text{MIA}}^{\text{opt}}$ (Thm. 2)	D_{MIA} UNTF (WB)	D_{MIA} AE (WB)	D_{MIA} AE (BB)	$V_{\text{IA}}^{\text{th}}$	$\widehat{V}_{\text{IA}}^2 + \widehat{D}_{\text{MIA}}^2$ (AE WB)
0.50	8	1.000	0.8542	0.8534	0.8573	0.8579	0.3472	0.8389
1.00	16	1.000	0.8542	0.8511	0.8524	0.8551	0.3472	0.8301
1.25	20	0.940	0.7397	0.7360	0.7484	0.7494	0.5306	0.8264
1.50	24	0.898	0.6584	0.6580	0.6640	0.6649	0.6362	0.8394
2.00	32	0.997	0.5518	0.5529	0.5421	0.5392	0.7500	0.8535
2.50	40	0.996	0.4839	0.4850	0.4933	0.4884	0.8099	0.8777
3.00	48	0.972	0.4360	0.4280	0.4235	0.4151	0.8467	0.8952
3.50	56	0.957	0.4000	0.3872	0.4083	0.3988	0.8715	0.8990
4.00	64	0.948	0.3716	0.3641	0.3671	0.3482	0.8895	0.9150

Table 4: Empirical verification of Theorem 2 (Experiment 12, Part B; WB = white-box LDA on h, BB = black-box logistic regression on ReLU output, UNTF = unit-norm tight frame). Black-box MIA advantage drops by −59.4% (0.8579 → 0.3482) and tracks the closed-form curve (14) within ∼0.01–0.02, while universally obeying $V_{\text{IA}}^2 + D_{\text{MIA}}^2 \leq 1$ (theoretical maximum = 0.9293, empirical maximum = 0.9313).

5 Discussion, Limitations, and Conclusion

1. **Prescriptive redesign of causal audit protocols:** Theorem 1 definitively resolves the paradox observed in Experiment 4 (26.7% false positives under the partial-veto linear rule versus 1.67% under the full four-pillar veto): **a causal auditor must never linearly average orthogonal diagnostic tests**, but must operate in calibrated ℓ_∞ norm (SIPS$_{\infty,\boldsymbol{\tau}}$), paired under non-identifiable selection propensities with the Manski–Rosenbaum upper curve $\overline{I}_{\text{col}}(\Gamma)$ and explicit reporting of the breakdown point Γ^*.

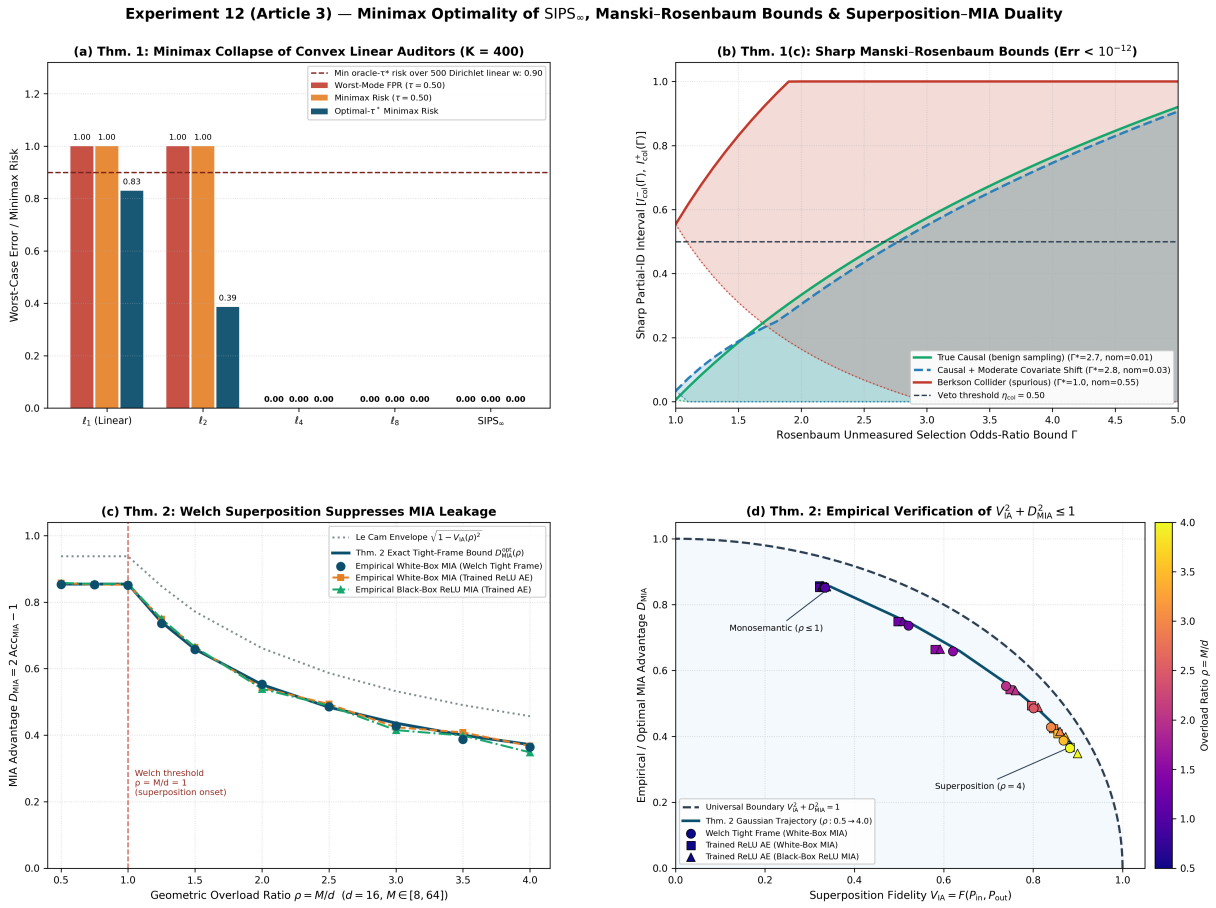

Figure 2: Experiment 12—Complete numerical verification of Theorems 1 and 2 (300 DPI). **(a)** Minimax collapse of every convex linear aggregator (ℓ_1, 500 Dirichlet weights) and ℓ_2 against SIPS$_\infty$ ($R_{\max} = 0.00$) across $K = 400$ scenarios. **(b)** Sharp Manski–Rosenbaum intervals $[\underline{I}_{\text{col}}(\Gamma), \overline{I}_{\text{col}}(\Gamma)]$ (discrepancy vs. global Charnes–Cooper LP $= 2.22 \times 10^{-16}$) cleanly separating the Berkson collider ($\Gamma^* = 1.0$) from causal models ($\Gamma^* \in [2.7, 2.8]$). **(c)** Monotone suppression of white-box and black-box MIA advantage on Adam-trained ReLU autoencoders above the Welch threshold $\rho = M/d = 1$, tracking the closed-form law (14) beneath the Le Cam envelope. **(d)** Empirical confirmation of the universal duality circle $V_{\text{IA}}^2 + D_{\text{MIA}}^2 \leq 1$ along the monosemantic ($\rho \leq 1$) to deep-superposition ($\rho = 4$) transition.

2. **Explicit limitations of Theorem 1:** Zero-error minimax optimality $R_{\max} = 0$ assumes that the causal noise ceiling ε_0 lies strictly below the illusion severity δ and that each pillar is consistently estimated at finite sample size; in practice, finite-sample estimator variance requires calibrating τ_k to control the false-rejection rate β (as illustrated by the stable F1 plateau in Figure 1d). Likewise, when selection propensities severely violate positivity ($\hat{\pi}_S \to 0$), the interval $[\underline{I}_{\text{col}}(\Gamma), \overline{I}_{\text{col}}(\Gamma)]$ widens rapidly and transparently reflects the limits of data identifiability.

3. **Explicit limitations of Theorem 2 and outlier memorization:** The exact closed form (14) relies on Gaussian (or CLT-regime) background fluctuations and linear additive memorization shift Δ_{mem}. When nonlinear deep networks allocate dedicated *monosemantic neurons* to atypical training points (memorized outliers outside superposition; Carlini et al., 2022), those examples correspond precisely to the excluded tail of fraction $< q$ in the Markov quantile bound (15)—elegantly explaining why superposition shields the *bulk* of common features while leaving isolated outliers vulnerable along nearly orthogonal directions.

4. **Conclusion and reproducibility:** By unifying non-compensatory detection of statistical illusions (SIPS$_\infty$ with Manski–Rosenbaum bounds) and the privacy geometry of superposed neural

representations (Welch ↔ MIA coupling under $V_{\mathrm{IA}}^2 + D_{\mathrm{MIA}}^2 \leq 1$) under the strictly classical concept of post-selection, both theorems deliver actionable tools for causal AI certification and formal memorization analysis. All experiments (seed `SEED = 42`, Python 3 / NumPy / SciPy / PyTorch) reproduce in under 20 s on Apple Silicon (`macOS ARM64`).

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

David H. Bailey and Marcos López de Prado. The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management*, 40(5):94–107, 2014. doi: 10.3905/jpm.2014.40.5.094.

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22 of *Proceedings of Machine Learning Research*, pages 100–108. PMLR, 2012.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077.

Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946. doi: 10.2307/3002000.

Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022. doi: 10.1109/SP46214.2022.9833649.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988. doi: 10.2307/2531595.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread, arXiv preprint arXiv:2209.10652*, 2022. URL https://arxiv.org/abs/2209.10652.

Berthold-Georg Englert. Fringe visibility and which-way information: An inequality. *Physical Review Letters*, 77(11):2154–2157, 1996. doi: 10.1103/PhysRevLett.77.2154.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Miguel A. Hernán, Sonia Hernández-Díaz, and James M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004. doi: 10.1097/01.ede.0000135174.63482.43.

Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973. doi: 10.1214/aos/1193342380.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York, 1986. ISBN 978-0-387-96307-5.

Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

Paul R. Rosenbaum. *Observational Studies*. Springer, New York, 2nd edition, 2002. doi: 10.1007/978-1-4757-3692-2.

Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 5558–5567. PMLR, 2019.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017. doi: 10.1109/SP.2017.41.

Louisa H. Smith and Tyler J. VanderWeele. Bounding bias due to selection. *Epidemiology*, 30(4): 509–516, 2019. doi: 10.1097/EDE.0000000000001032.

Jonathan Taylor and Robert J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences (PNAS)*, 112(25):7629–7634, 2015. doi: 10.1073/pnas.1507583112.

Lloyd R. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Transactions on Information Theory*, 20(3):397–399, 1974. doi: 10.1109/TIT.1974.1055219.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018. doi: 10.1109/CSF.2018.00027.