FROM IDENTIFIABILITY TO SAMPLE COMPLEXITY FOR DISCRETE LATENT VARIABLES SEEN THROUGH KNOWN CHANNELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal representation learning seeks to recover meaningful hidden variables from observations. When observations merge hidden states, which properties remain recoverable, and how much data is needed to estimate them? We study this question in finite discrete models. A matrix built from the observation process gives an exact test for whether an average hidden property can be recovered. For every recoverable target, we construct an estimator whose mean squared error has an explicit leading term proportional to the inverse of sample size, plus a smaller correction. A matching lower bound shows that no method improves this leading error in the worst case, and the estimator has a normal large sample limit with optimal variance. When the observation process is unknown but belongs to a finite set, indistinguishable processes with different target values make consistent estimation impossible. With sufficient separation, an adaptive estimator matches the same leading error as if the process were known. Under a common support condition, candidate processes whose observed laws are close, at the scale of one over the square root of the sample size, but whose targets differ leave an error that does not vanish. Simulations and a semi-synthetic census study with a randomized-response channel illustrate these predictions.

1 INTRODUCTION

A hidden variable Z takes one of d values. We never see Z itself. We see two things generated from it: a coarse label $X = f(Z)$, where the map f may send several hidden values to the same label, and a noisy side signal S whose distribution depends on Z. From n independent draws of the observed pair (X, S) we want to learn about the distribution p of Z. This is the simplest form of a question that runs through causal representation learning (Schölkopf et al., 2021): the data are a lossy view of the causes we care about, and we want to know what the view still determines.

Three questions arise in this order.

1. **What is recoverable?** An average hidden property, say the fraction of the population in a hidden state, is recoverable if any two hidden distributions that produce the same observed law agree on it. Which averages have this property?
2. **How much data?** For a recoverable average, how does the estimation error fall with n, with what constant, and can any method do better?
3. **What if the observation process is uncertain?** Suppose the analyst knows the process only up to a finite list of candidates. When can the data pick the right one, and when does the uncertainty leave an error that no sample size removes?

We answer all three in the finite discrete setting, with explicit constants. The setting is narrow by choice. It contains several classical models as special cases: randomized response for sensitive survey questions (Warner, 1965), prevalence estimation from an imperfect diagnostic test (Rogan & Gladen, 1978), coarsened or censored survey categories, finite-alphabet communication channels, and local differential privacy mechanisms for categorical data (Kairouz et al., 2016). It is a finite abstraction motivated by the auxiliary-variable models that drive identifiability theory in nonlinear ICA and causal representation learning (Hyvärinen et al., 2019; Khemakhem et al., 2020): there, an observed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

side variable makes latent structure visible through an unknown decoder that is learned; here, the side signal S plays the role of the side variable, but the decoder and the channel are known and only the latent distribution is unknown. We do not claim finite-sample guarantees for neural representation learners. What the finite setting buys is that every question above has a finite-dimensional, computable answer, with explicit constants.

Contributions. The observation process is summarized by a known matrix A whose columns are the conditional laws of (X, S) given each hidden value, so that the observed law is $q = Ap$.

1. *An exact recoverability test* (Section 3). An average $\langle v, p\rangle$ is recoverable if and only if v lies in the column space of $A^\top$. The test is a rank computation, and the same space is the exact boundary of consistent estimation.
2. *An explicit finite-sample risk with a matching lower bound* (Section 4). A plug-in estimator that estimates its own weights from the sample has mean squared error $V_A/n + O(1/n^2)$ uniformly over an interior slab of priors, where V_A is an explicit variance, and no estimator beats V_A/n to first order in the worst case. The first-order theory here is classical, and we say precisely which parts are (Table 1). What is new is the uniform second-order control for the feasible estimator, a dichotomy separating it from every fixed weighting, a two-sided bound showing where dimension and conditioning enter, and an expansion for the price of estimating the channel from calibration data.
3. *Three separated statements for an unknown channel* (Section 5). If two candidate channels can produce the same observed law with different targets, no estimator is consistent. If the candidate images are separated, a selector plus plug-in matches the oracle to first order, with the exact selection exponent (the polytope Chernoff information of Brandão et al. (2020)). If two candidates are close in observed law but far in target, a two-point bound gives an error that stays bounded away from zero when the closeness is of order $n^{-1/2}$. The two exponents match in rate but not in constant, and we state the open band.
4. *Evidence* (Section 6). Simulations track V_A over three orders of magnitude of conditioning. A semi-synthetic study applies the theory to census microdata passed through a randomized-response channel, where the channel is known by design and the ground truth is available.

2 The model: a hidden variable seen through a channel

Objects. The hidden variable Z takes values in a finite set $\mathcal{Z} = \{z_1, \dots, z_d\}$ with distribution $p = (p_1, \dots, p_d)$, all $p_j > 0$. Two observed quantities are generated from Z. The first is a label $X = f(Z)$, where $f : \mathcal{Z} \to \mathcal{X}$ is a known deterministic map onto a finite set. The map is usually many-to-one: several hidden values share one label, and the set $F_x := f^{-1}(x)$ of hidden values behind label x is called the *fiber* of x. The second is a side signal $S \in \mathcal{S}$ drawn from a known conditional law $\kappa(z, \cdot)$ given $Z = z$; we call κ the *auxiliary channel*. The data are n independent copies of (X, S).

The observation matrix. Everything about the observation process is encoded in one matrix. Index rows by observable pairs $y = (x, s) \in \mathcal{Y} := \mathcal{X} \times \mathcal{S}$, columns by hidden values, and set

$$A_{(x,s),j} := \mathbf{1}\{f(z_j) = x\}\, \kappa(z_j, s). \tag{1}$$

Column j is the law of (X, S) given $Z = z_j$, so the observed law is the vector $q := Ap$ with $q_{(x,s)} = P(X = x, S = s)$. Let $K := |\mathcal{Y}|$ and let $\Omega_+ := \{y : \sum_j A_{y,j} > 0\}$ be the cells that can occur. Throughout, f and κ (hence A) are known, and only p is unknown; Section 5 relaxes this to A known up to a finite list.

Running example. Three hidden states, a two-outcome label, a two-valued side signal. Let $\mathcal{Z} = \{1, 2, 3\}$, $f(1) = f(2) = a$, $f(3) = b$, and let $S \in \{0, 1\}$ be a noisy detector with $\kappa(1, 1) = 0.9$, $\kappa(2, 1) = 0.3$, $\kappa(3, 1) = 0.5$. The label alone cannot separate states 1 and 2; they share the fiber F_a. The side signal can, because its law differs across the fiber. This is the general pattern: labels resolve fibers, and the auxiliary channel buys resolution inside fibers.

Classical instances. In Warner's randomized response, Z is a sensitive attribute, there is no informative label, and κ is the randomization device the interviewer designed, so A is known exactly. In prevalence estimation from a screening test, Z is true disease status, S the test result, and κ is built from the published sensitivity and specificity. In a locally private release of a categorical attribute, κ is the privacy mechanism. In all three, an analyst who ignores the channel and reads S as if it were Z gets a biased answer; the classical corrections invert the channel.

Targets. Most population summaries of a finite hidden variable are averages: a prevalence $P(Z \in B) = \langle \mathbf{1}_B, p\rangle$, a contrast between two groups, an expected cost. All of them have the form $\theta(p) = \langle v, p\rangle = \sum_j v_j p_j$ for a vector $v \in \mathbb{R}^d$, and since Z is finite these are exactly the linear functionals of p. We therefore ask which $\langle v, p\rangle$ can be recovered and how well.

3 What is recoverable

Two hidden distributions p, p' are *observationally equivalent* if $Ap = Ap'$: they produce the same data law and nothing can tell them apart. A target v is *identifiable* if $\langle v, p\rangle = \langle v, p'\rangle$ whenever $Ap = Ap'$. Write V^A for the set of identifiable v.

Proposition 1 (Recoverability test)**.** $V^A = \mathrm{col}(A^\top)$. *Moreover the failure is witnessed: for every $v \notin \mathrm{col}(A^\top)$ there are $p \neq p'$ with all entries positive, $Ap = Ap'$, and $\langle v, p\rangle \neq \langle v, p'\rangle$.*

The inclusion $\mathrm{col}(A^\top) \subseteq V^A$ is one line: if $v = A^\top a$ then $\langle v, p\rangle = \langle a, q\rangle$ depends on p only through q. The converse perturbs an interior p along a kernel direction of A and is given in Appendix A. The proposition is elementary, and we present it as the preliminary it is; its value is in what it organizes. The identifiable space equals the row space of the observation matrix, so recoverability is a rank computation. If the side signal is dropped, A collapses to the fiber-indicator matrix and V^A becomes the functions constant on fibers: the fibers are the floor, and everything beyond that floor is bought by κ. In the running example $\mathrm{rank}(A) = 3 = d$, so every average is recoverable; if instead $\kappa(1, \cdot) = \kappa(2, \cdot)$, the rank drops to 2 and $p_1 - p_2$ is lost for good. Section 4 shows that the same space is the exact boundary of consistent estimation, and Section 6 shows the boundary as a bias floor that no sample size removes.

4 How much data: known channel

4.1 The estimator

Fix an identifiable target $v \in \mathrm{col}(A^\top)$ and write $\theta := \langle v, p\rangle$. A natural estimator is a weighted count of observed cells, $\langle a, \hat{q}_n\rangle$, where $\hat{q}_n$ is the empirical law on $\mathcal{Y}$. It is unbiased for θ under *every* p exactly when the weights are *feasible*: $A^\top a = v$. When $\mathrm{rank}(A) < K_+$ there are many feasible weightings, and their variances differ. With $D_q := \mathrm{diag}(q)$ on Ω_+, the minimum-variance feasible weighting is

$$a^*(v; q) := D_q^{-1} A \, (A^\top D_q^{-1} A)^+ v, \qquad V_A(v; q) := v^\top (A^\top D_q^{-1} A)^+ v - \theta^2, \tag{2}$$

where $(\cdot)^+$ is the Moore–Penrose pseudoinverse; $A^\top a^* = v$ holds for every positive q, and V_A/n is the variance of the oracle estimator $\langle a^*(v; q), \hat{q}_n\rangle$ that knows q (Appendix B). The vector $a^* - \theta\mathbf{1}$ is also the efficient influence function of θ in the model $\{Ap\}$, so V_A bounds every regular estimator, not only linear ones. This is the classical Gauss–Markov and semiparametric efficiency theory (Bickel et al., 1993; van der Vaart, 1998), and Table 1 records which statements below are classical and which are ours.

The feasible estimator replaces q by a regularized empirical law $\tilde{q}_n := (1 - 1/n)\hat{q}_n + (1/n)u$, with u uniform on Ω_+, so that the weights are always defined:

$$\hat{T}_n := \langle \hat{a}_n, \hat{q}_n\rangle, \qquad \hat{a}_n := a^*(v; \tilde{q}_n). \tag{3}$$

Because $A^\top \hat{a}_n = v$ holds identically, the first-order bias of $\hat{T}_n$ cancels for every sample; this is a strong form of Neyman orthogonality (Neyman, 1979; Chernozhukov et al., 2018), and it is what drives the second-order rate below.

Table 1: What is classical and what is new in Sections 4 and 5.

Classical (cited, not claimed)	**This paper**
$\mathrm{col}(A^\top)$ as the identifiable space; Warner's and Rogan–Gladen's corrections as instances of $q = Ap$	Proposition 1 as the organizing preliminary; the same space as the exact boundary of consistent estimation (Theorem 2(d))
Efficient weights a^*, the bound V_A, the efficient influence function, first-order efficiency and the CLT (Bickel et al., 1993; van der Vaart, 1998; Hansen, 1982)	Uniform $O(1/n^2)$ oracle regret of the feasible plug-in on the interior slab, rank-deficient A included (Theorem 2(a))
Projection identity for feasible weightings (Rao & Mitra, 1972)	Its pairing with (a): every fixed weighting pays first-order regret somewhere while the plug-in pays second order everywhere (Proposition 3)
van Trees inequality (Gill & Levit, 1995)	Together with (a), the global asymptotic minimax constant on the slab, rank-deficient A included (Theorem 2(b))
Variance inflation from estimated misclassification rates (Rogan & Gladen, 1978; Greenland, 1988; Buonaccorsi, 2010)	The overidentified-channel expansion with constant $C_{\mathrm{cal}} = \sum_j p_j^2 \mathrm{Var}_{A_{\cdot j}}(a^*)$ (Theorem 5)
Exact binary exponent for convex hypothesis classes (Brandão et al., 2020)	Finite-n multiclass form $(M-1)e^{-nC^*}$ with a computable C^*, and the estimation-risk consequences (Proposition 8)

4.2 The rate theorem

Let $\Delta_{\geq\varepsilon}(\mathcal{Z}) := \{p : \min_j p_j \geq \varepsilon\}$ be the interior slab.

Theorem 2 (Finite-sample rate, lower bound, normality, cliff)**.** *Fix $0 < \varepsilon < 1/d$ and write $\sup_p$ for the supremum over $p \in \Delta_{\geq\varepsilon}(\mathcal{Z})$. Let $v \in \mathrm{col}(A^\top)$ in (a) to (c).*

(a) ***Upper bound.*** *$\mathbb{E}_p[(\hat{T}_n - \theta)^2] = V_A(v; Ap)/n + O(1/n^2)$ uniformly over $p \in \Delta_{\geq\varepsilon}(\mathcal{Z})$, with a constant depending only on ε, A, v.*

(b) ***Lower bound.*** *For every sequence of estimators $\hat{\theta}_n$, $\liminf_n n \sup_p \mathbb{E}_p[(\hat{\theta}_n - \theta)^2] \geq \sup_p V_A(v; Ap)$. With (a), $\inf_{\hat{\theta}_n} \sup_p n\,\mathbb{E}_p[(\hat{\theta}_n - \theta)^2] \to \sup_p V_A(v; Ap)$.*

(c) ***Normality.*** *For each $p \in \Delta_{\geq\varepsilon}(\mathcal{Z})$, $\sqrt{n}(\hat{T}_n - \theta) \to \mathcal{N}(0, V_A(v; Ap))$ in distribution, and V_A is the semiparametric efficiency bound.*

(d) ***Cliff.*** *If instead $v \notin \mathrm{col}(A^\top)$, then $\sup_p \mathbb{E}_p[(\hat{\theta}_n - \theta)^2] \not\to 0$ for every estimator.*

In words: for a recoverable target the leading error is V_A/n, nothing does better in the worst case, and the error is normal with that variance; for an unrecoverable target no sample size helps. The proof (Appendix B) has three ingredients that are specific to the feasible estimator. First, the identity $A^\top a^*(v; q') = v$ for all positive q' gives the exact decomposition $\hat{T}_n - \theta = \langle a^*, \hat{q}_n - q\rangle + \langle \hat{a}_n - a^*, \hat{q}_n - q\rangle$ with no cross term at the population level. Second, the map $q' \mapsto a^*(v; q')$ is smooth with bounded derivatives on a compact neighborhood of the slab image, because $\ker(A^\top D_{q'}^{-1} A) = \ker A$ for every positive q', so the pseudoinverse has constant rank there (Golub & Pereyra, 1973). Third, the cross moment between the linear term and the quadratic remainder is a signed third central moment of a multinomial, which is $O(1/n^2)$, not $O(n^{-3/2})$. The lower bound is a van Trees argument along the least favorable direction $(A^\top D_q^{-1} A)^+(v - \theta\mathbf{1})$ (Gill & Levit, 1995).

4.3 What the second-order statement adds

Once q is fixed, the first-order theory is complete and classical. The next three results are about what changes when the weights must be estimated, when the channel is ill conditioned, and when the channel itself is estimated.

Proposition 3 (Projection identity and regret dichotomy)**.** *For every feasible a and every q on the slab image, $n\mathrm{Var}_q\langle a, \hat{q}_n\rangle - V_A(v; q) = \|a - a^*(v; q)\|^2_{D_q}$ (Rao & Mitra, 1972). Consequently, if*

$\mathrm{rank}(A) < K_+$ *and* $a^*(v;\cdot)$ *is not constant on the slab image, with* $\|a^*(v;q_1) - a^*(v;q_2)\|_2 = \Delta > 0$*, then every fixed feasible* a *has oracle regret at least* $c_A\varepsilon\Delta^2/(4n)$ *at* q_1 *or at* q_2*, where* $c_A := \min_{y\in\Omega_+}\sum_j A_{y,j} > 0$*, while the plug-in* (3) *has oracle regret* $O(1/n^2)$ *uniformly by Theorem 2(a).*

The identity is classical generalized Gauss–Markov theory. The point is the pairing: no fixed weighting, however well tuned, closes the $1/n$ versus $1/n^2$ gap when the efficient weighting moves with q, which is the generic overidentified case. (The plug-in's gap to the oracle can be negative at a given q, since $\hat{T}_n$ is not unbiased; "regret" here means the gap in absolute value.) In a sharp-kernel configuration with $\mathrm{rank}(A) = 4 < K = 6$ (Appendix E), the plug-in beats even a competitor tuned at the slab center from $n = 400$ onward at each of three tested slab points, and does not yet win at $n = 200$, consistent with a second-order theory.

Lemma 4 (Where dimension enters)**.** *Let* $v \in \mathrm{col}(A^\top)$*,* $q = Ap$ *with* $q_{\min} := \min_{y\in\Omega_+} q_y > 0$*, and let* s_+ *be the smallest nonzero singular value of* A*. Then*

$$q_{\min}\, v^\top (A^\top A)^+ v - \theta^2 \;\le\; V_A(v;q) \;\le\; v^\top (A^\top A)^+ v - \theta^2 \;\le\; \|v\|_2^2/s_+^2.$$

The displays in Theorem 2 contain neither d nor K. The lemma says where they hide: in the conditioning pair $(s_+(A), q_{\min})$ through V_A. A large alphabet is harmless when A is well conditioned, and a small alphabet is hopeless when two columns nearly coincide in a direction that v sees; for a target $\theta \neq 0$ the sample size for relative error δ is $n \approx V_A/(\delta\theta)^2$ and carries the entire burden. Section 6 traces V_A from 0.43 to 443 at fixed $d = 4$ and fixed rate.

Theorem 5 (Price of an estimated channel)**.** *Suppose* $\mathrm{rank}(A) = d$*,* $p \in \Delta_{\ge\varepsilon}(\mathcal{Z})$*, and for each hidden value* j *the column* $A_{\cdot j}$ *is replaced by the empirical law* $\hat{A}_{\cdot j}$ *of* m *independent calibration draws from it, independent of the main sample. Let* $\hat{\theta}$ *be the plug-in* (3) *built with* $\hat{A}$ *in place of* A*, clipped to* $[-\|v\|_\infty, \|v\|_\infty]$*. Then, as* $n, m \to \infty$*, uniformly over* $p \in \Delta_{\ge\varepsilon}(\mathcal{Z})$*,*

$$\mathbb{E}(\hat{\theta} - \theta)^2 = \frac{V_A(v;q)}{n} + \frac{C_{\mathrm{cal}}}{m} + O\Big(\frac{1}{n^2} + \frac{1}{m^2} + \frac{1}{nm}\Big), \qquad C_{\mathrm{cal}} = \sum_{j=1}^{d} p_j^2\, \mathrm{Var}_{Y\sim A_{\cdot j}}\big(a^*(v;q)_Y\big),$$

with a constant depending only on ε*,* A*,* v*.*

Calibration noise in hidden category j costs p_j^2 times the variance of the efficient weights under that category's column. The key step is exact: feasibility gives $a^*(A';r)^\top q = \theta + p^\top(A - A')^\top a^*(A';r)$ for every channel A' of rank d and interior r, and admissible channel errors have zero column sums, so the channel derivative collapses to $-a^{*\top}(\mathrm{d}A)p$ (Appendix C). The binary invertible case recovers the Rogan–Gladen variance with estimated sensitivity and specificity; the general validation-sampling literature (Greenland, 1988; Buonaccorsi, 2010) covers matrix corrections, and the overidentified pseudoinverse form with its per-column constant is the addition here. Two consequences: a calibration budget is best allocated as $m_j \propto p_j\sqrt{\mathrm{Var}_{A_{\cdot j}}(a^*)}$, and the constant is checkable. For the configuration of our estimated-channel experiment the theorem gives $C_{\mathrm{cal}} = 0.1006$; the experiment, run before the theorem existed, had measured an inflation of about $0.10/m$. At $\mathrm{rank}(A) < d$ the expansion fails, because finite-m calibration makes $\hat{A}$ full rank and the pseudoinverse jumps; we state the theorem at full column rank only.

5 How much data: unknown channel

Now the analyst knows only that A is one of M candidates $A^{(1)}, \ldots, A^{(M)}$ on a common alphabet $\mathcal{Y}$, with target vectors $v^{(m)} \in \mathrm{col}(A^{(m)\top})$. Think of a diagnostic test with one of several published calibrations, or a privacy mechanism deployed at one of several strengths. *Adaptation* means estimating $\theta_m(p) := \langle v^{(m)}, p\rangle$ without being told m, and its difficulty is the gap between the achievable risk and the oracle risk $V_{\max}/n$ of an analyst who is told m. Throughout this section the prior ranges over the slab $\Delta_{\ge\varepsilon}(\mathcal{Z})$ for a fixed $0 < \varepsilon < 1/d$, every supremum $\sup_{(m,p)}$ is over $m \in \{1, \ldots, M\}$ and $p \in \Delta_{\ge\varepsilon}(\mathcal{Z})$, and $V_{\max} := \sup_{(m,p)} V_{A^{(m)}}(v^{(m)}; A^{(m)}p)$. Write $\mathcal{Q}_m := \{A^{(m)}p : p \in \Delta(\mathcal{Z})\}$ for the image polytope of candidate m, $B := \max_m \|v^{(m)}\|_\infty$, and $\gamma := \min_{m\neq m'} \inf\{\|q - q'\|_1 : q \in \mathcal{Q}_m, q' \in \mathcal{Q}_{m'}\}$ for the ℓ_1 separation, computable by linear programming. Three statements follow, in increasing order of what they assume.

Proposition 6 (Overlap makes estimation impossible). *If $A^{(m)}p = A^{(m')}p'$ for some $m \neq m'$ and priors p, p' in the slab, while $\Delta := |\theta_m(p) - \theta_{m'}(p')| > 0$, then for every estimator and every n, $\sup_{(m,p)} \mathbb{E}[(\hat{\theta}_n - \theta_m(p))^2] \geq \Delta^2/4$.*

Theorem 7 (Separation makes adaptation free to first order). *Assume $\gamma > 0$ and that each per-model clipped plug-in $\bar{T}_n^{(m)}$ has risk at most C_m/n over the slab, $C := \max_m C_m$ (the separation γ may be computed over the full image polytopes, as in our experiments: the slab images are subsets, so their separation is at least γ). The estimator that selects $\hat{m} := \arg\min_m \mathrm{dist}_1(\hat{q}_n, \mathcal{Q}_m)$ and returns $\bar{T}_n^{(\hat{m})}$ satisfies*

$$\sup_{(m,p)} \mathbb{E}[(\hat{\theta}_n - \theta_m(p))^2] \leq \frac{C}{n} + 8B^2 K\, e^{-n\gamma^2/(8K^2)}.$$

By Theorem 2(a), C/n may be replaced by $V_{\max}/n + O(1/n^2)$.

The ℓ_1 selector is simple but its exponent carries an alphabet factor $1/(8K^2)$ from a union bound. Replacing it by pairwise likelihood-ratio tests between the least favorable pair of each two polytopes removes the factor and identifies the exponent exactly. For $s \in (0,1)$ let $\rho_s(q,q') := \sum_y q_y^s (q'_y)^{1-s}$, write $\mathcal{Q}_m^\varepsilon := \{A^{(m)}p : p \in \Delta_{\geq\varepsilon}\}$ for the slab image of candidate m ($\varepsilon = 0$ is allowed when every entry of every candidate is positive), let $\rho_s(\mathcal{Q}_m^\varepsilon, \mathcal{Q}_{m'}^\varepsilon)$ be the supremum of ρ_s over the two slab images, and define the *polytope Chernoff information* $C^* := \min_{m\neq m'} \sup_{s\in(0,1)} [-\log \rho_s(\mathcal{Q}_m^\varepsilon, \mathcal{Q}_{m'}^\varepsilon)]$.

Proposition 8 (Exact selection exponent). *Assume all candidates share the active support Ω_+ and the slab images $\mathcal{Q}_m^\varepsilon$ are pairwise disjoint; every $q \in \mathcal{Q}_m^\varepsilon$ then has $q_y \geq \beta > 0$ on Ω_+. Run, for every pair $m \neq m'$, the tilted likelihood-ratio test between the two extremal points of the pair (defined in the proof), and let $\hat{m}$ be the candidate that wins all of its $M-1$ comparisons, with an arbitrary choice if no candidate does. Then $\sup_{(m,p)} P(\hat{m} \neq m) \leq (M-1)e^{-nC^*}$, and no selector achieves a better exponent. Moreover $C^* = \min_{m\neq m'} \inf_{q\in\mathcal{Q}_m^\varepsilon, q'\in\mathcal{Q}_{m'}^\varepsilon} C(q,q')$ with C the pairwise Chernoff information, $C^* \geq -\log(1-H_*^2) \geq \gamma^2/8$ where H_* is the Hellinger separation, and γ alone does not determine C^*: two pairs with the same $\gamma = 0.2$ can have Chernoff informations 0.0051 and 0.0296.*

The binary exactness statement, converse included, is the adversarial Chernoff theorem of Brandão et al. (2020); we add the finite-n multiclass prefactor form, a computation of the exponent by one-dimensional search over concave maximizations, and the estimation consequence: risk at most $V_{\max}/n + O(1/n^2) + 4B^2(M-1)e^{-nC^*}$, whose selection term is $o(1/n)$ if and only if $nC^* - \log n \to \infty$.

For the lower bound we need a witness: two candidates whose observed laws are close but whose targets differ. Define the *adaptation modulus* $\omega(\gamma) := \sup\{|\theta_m(p) - \theta_{m'}(p')| : m \neq m',\ p, p' \in \Delta_{\geq\varepsilon}(\mathcal{Z}),\ \|A^{(m)}p - A^{(m')}p'\|_1 \leq \gamma\}$, the largest target gap between candidates whose observed laws are within γ. A witness at scale γ_n with target gap Δ_0 exists exactly when $\omega(C_H\gamma_n) \geq \Delta_0$; if $\omega(\gamma) \to 0$ as $\gamma \to 0$, there is no such witness and the lower bound below carries no information at small scales.

Theorem 9 (Finite-sample lower bound for nearby channels). *Suppose there are candidates $m_1 \neq m_2$ and priors p_1, p_2 in the slab with $\|A^{(m_1)}p_1 - A^{(m_2)}p_2\|_1 \leq C_H\gamma_n$, $|\theta_{m_1}(p_1) - \theta_{m_2}(p_2)| \geq \Delta_0 > 0$, and both observed laws at least $q_{\min}$ on a common support. Then for all n with $C_H^2\gamma_n^2 \leq 4q_{\min}$, every estimator has $\sup_{(m,p)} \mathbb{E}[(\hat{\theta}_n - \theta_m(p))^2] \geq \frac{\Delta_0^2}{16} \exp\big(-\frac{C_H^2}{2q_{\min}} n\gamma_n^2\big)$.*

Corollary 10 (Regimes). *Let γ_n be the separation of a sequence of candidate families with fixed K, B, $q_{\min}$, and assume the constants in Theorem 2(a) and the convergence in Theorem 2(b) are uniform along the sequence (for instance, all candidates range over a compact set of channels of constant rank, on which the constants of Appendix B are bounded). If $n\gamma_n^2 \geq 8K^2\log n + \omega(1)$ (or $nC_n^* - \log n \to \infty$ with the selector of Proposition 8), the minimax risk is $V_{\max}/n + o(1/n)$: adaptation is free to first order. If the witness of Theorem 9 exists and $n\gamma_n^2 \leq (2q_{\min}/C_H^2)\log n - \omega(1)$, then n times the minimax risk tends to infinity: adaptation is impossible to first order; if moreover $\gamma_n\sqrt{n}$ stays bounded, the risk is bounded below by a constant. In the band $(2q_{\min}/C_H^2)\log n + \omega(1) \leq n\gamma_n^2 \leq 8K^2\log n - \omega(1)$ our bounds do not meet, and the rate there is open.*

The two exponents, $1/(8K^2)$ on the upper side and $C_H^2/(2q_{\min})$ on the lower side, match in the variable $n\gamma_n^2$ and not in constant, for the transparent reason that they are proved with different

Table 2: Never-married prevalence ($\theta = 0.328$) from randomized-response reports of $N = 32{,}561$ census records; 300 replications. Bias and standard deviation (SD) are empirical; $\sqrt{V_A/N}$ is the theoretical prediction; coverage is of the 95% interval built from the plug-in V_A.

	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$
naive bias	-0.170	-0.149	-0.097
corrected bias (single report)	< 0.001	< 0.001	< 0.001
corrected SD (single report), measured / $\sqrt{V_A/N}$	0.0244 / 0.0239	0.0101 / 0.0108	0.0048 / 0.0049
coverage of the 95% V_A interval	0.96	0.96	0.96

witnesses. We therefore describe the picture as transition *scales*, not a sharp threshold. For a fixed separation $\gamma > 0$ the lower bound of Theorem 9 decays exponentially in n and says nothing about first-order risk; it bites when the separation shrinks like $n^{-1/2}$, the scale at which channel uncertainty and sampling error are comparable. Matching estimation-risk constants remain open.

6 Experiments

All experiments run on one CPU in minutes; configurations, seeds, and Monte Carlo standard errors are in Appendix E, and code is in the supplement.

Census microdata through a randomized-response channel. We take the UCI Adult census extract (Kohavi, 1996) ($N = 32{,}561$ records) and treat marital status ($d = 7$ categories) as the hidden variable, so the true p is the sample distribution and the ground truth is known exactly. Each record is passed through k-ary randomized response (Warner, 1965; Kairouz et al., 2016): the true category is kept with probability $e^\varepsilon/(e^\varepsilon + 6)$ and otherwise replaced by a uniformly random other category. The channel is known by design, which is the theory's standing assumption, and the study is semi-synthetic: real labels, an author-imposed channel, no field nonresponse. We repeat the randomization and resample records 300 times for $\varepsilon \in \{0.5, 1, 2\}$, and estimate the never-married prevalence ($\theta = 0.328$) and a divorced-versus-married contrast. Table 2 reports the prevalence target. The naive estimator that reads the private report as the truth is biased by more than 40 standard deviations. The corrected estimator is unbiased with standard deviation matching $\sqrt{V_A/N}$, and the 95% intervals built from the plug-in V_A cover at 0.93 to 0.98 across all twelve cells (three ε, two targets, two estimators). Dropping ε from 2 to 0.5 raises the condition number of the two-report channel from 1.35 to 8.26 and V_A roughly twentyfold at fixed rate: the privacy sweep is the conditioning path of Lemma 4 on real labels. Because V_A is computable before any data are collected, it also prices designs. Two independent reports per record cut the standard deviation by a factor 1.27 to 1.49 (predicted 1.30 to 1.41); the efficient joint weighting adds a further 0.2 to 4.3% over the fair baseline of averaging two corrected single-report estimates; splitting a privacy budget ε into two $\varepsilon/2$ reports is strictly worse (1.4 to 1.7 times the standard deviation); and with asymmetric reports ($\varepsilon = 2$ and 0.5) the plug-in weights reproduce the covariance-aware inverse-variance combination to three significant figures (0.00462 versus 0.00462) without being told which report is more informative, while the plain average gives 0.0114.

Adaptation on the same data. Each record supplies two independent reports, and the analyst knows only that the mechanism is one of two two-report channels ($\varepsilon = 1$ or 2 per report). The two 49-cell image polytopes are disjoint by linear programming ($\gamma = 0.324$; polytope Chernoff information $C^* = 0.0177$; $\gamma^2/8 = 0.0131 < H_*^2 = 0.0175 < C^*$, so the chain of Proposition 8 is visible on real channels), whereas single-report images are nested and selection would be ill posed. Selection error is $1/400$ per truth arm at $n = 250$ and $0/400$ at $n \geq 1000$, and the adaptive-to-oracle mean squared error ratio is 0.995 to 1.007 across tested n: the separated regime of Corollary 10 end to end.

Conditioning and scaling in simulation. Figure 1 (left) follows a one-parameter family in which two columns of A merge: V_A rises three orders of magnitude while $n\cdot$MSE stays within 4.5% of it (Monte Carlo standard error about 2.6%), and at the rank drop the plug-in lands on its hand-computable bias floor $((p_3 - p_4)/2)^2 = 0.0025$. This is Lemma 4 and Theorem 2(d) in one picture.

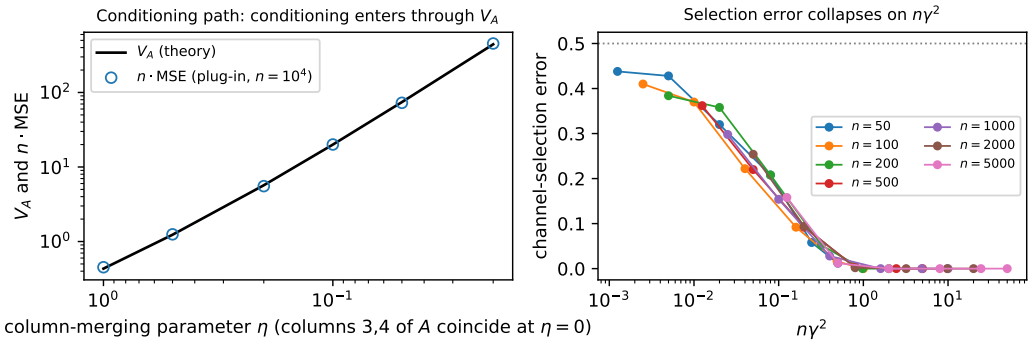


Figure 1: Left: as two columns of A merge ($\eta \to 0$, $d = 4$, $n = 10^4$, 3,000 replications), V_A grows from 0.43 to 443 and n·MSE tracks it within 4.5% throughout; at $\eta = 0$ (not shown) the rank drops and the error settles at the predicted bias floor 0.0025 (measured 0.00252). Right: channel-selection error in a two-candidate coarsening design plotted against $n\gamma^2$ for seven sample sizes; the curves collapse, as the rate statements predict, and the constant $8K^2$ is not claimed.

Figure 1 (right) uses a genuinely coarsening two-candidate design (many-to-one fibers, the alternative channel a within-fiber translate, so the polytopes are provably disjoint with $\gamma \propto \alpha$) and plots selection error against $n\gamma^2$ over 7 sample sizes and 6 separations: cells with the same $n\gamma^2$ but different (n, γ) agree within Monte Carlo noise, error is at most 0.012 for $n\gamma^2 \in (0.5, 1]$ and 0 of 500 trials beyond, and the adaptive-to-oracle ratio is at most 1.056 in the worst bin and 1.000 once $n\gamma^2 > 1$.

Robustness. Across 8 configurations (three designed, five random), 3 sample sizes and 20 seeds, $n \cdot \mathrm{MSE}/V_A$ lies in $[0.98, 1.02]$ in 23 of 24 cells (the remaining cell at 1.022, within one per-seed Monte Carlo standard error, about three standard errors of the twenty-seed mean); the seed-to-seed spread is comparable to the Monte Carlo error (ratio 0.6 to 1.2), so we report Monte Carlo standard errors throughout. A $d = 60$, $K = 80$ configuration holds the ratio in $[0.94, 0.99]$ up to $n = 10^5$: a large alphabet with a well-conditioned channel behaves as Lemma 4 predicts. Two stress tests sit outside the theory and are labeled as such. Misspecifying the channel by a fraction δ produces a bias linear in δ and a floor quadratic in it. Estimating the channel from m calibration draws per category adds $\approx 0.10/m$ to the risk at $m \in \{50, 200, 1000, 5000\}$; Theorem 5 predicts $C_{\mathrm{cal}} = 0.1006$, and the measured ratios to the prediction are 1.06, 1.01, 1.05, 0.97.

7 Related Work

Identifiability results in nonlinear ICA and causal representation learning are population-level statements about what is recoverable up to a symmetry (Hyvärinen & Morioka, 2016; Hyvärinen et al., 2019; Khemakhem et al., 2020; von Kügelgen et al., 2021; Zimmermann et al., 2021; Lippe et al., 2022; Ahuja et al., 2023), with impossibility results delimiting what unsupervised methods can do (Locatello et al., 2019; Horan et al., 2021). Finite-sample statements exist in two settings we know of: Lyu & Fu (2022) give finite-sample identifiability bounds for contrastive nonlinear ICA with function learners, and Acartürk et al. (2024) give PAC-style sample complexity for interventional CRL under linear mixing, without lower bounds. Neither covers the discrete coarsened setting, where exact constants are available, and we do not claim that our tools transfer to theirs. On the statistics side, our model is the multinomial model under a known linear coarsening. The efficient-score and influence-function theory is that of Bickel et al. (1993) and van der Vaart (1998, Ch. 25); coarsened and missing data are treated semiparametrically by Robins et al. (1994), Tsiatis (2006) and, under nonignorable missingness, Miao & Tchetgen Tchetgen (2016). Randomized response (Warner, 1965), screening-test correction (Rogan & Gladen, 1978), and local differential privacy for categorical data (Kairouz et al., 2016) are exact instances of $q = Ap$, and our V_A recovers Warner's variance at $d = 2$. Those literatures treat particular channels, mostly to first order and with the channel known; what they do not jointly provide, and what this paper assembles for a general known channel, is the uniform

second-order control of the feasible plug-in, the global minimax constant on the slab for rank-deficient channels, the calibration expansion in its overidentified form, and the operator-uncertainty theory.

8 LIMITATIONS AND CONCLUSION

This work does not provide finite-sample guarantees for modern causal representation learning systems with neural encoders and continuous latents. It covers finite latent and observable alphabets with a channel known exactly or up to a finite list; unknown supports, estimated channels of deficient rank, and alphabets growing with n are outside the theory. In the adaptive setting the upper and lower exponents agree in rate but not in constant, and one band of separations is unresolved. Within its scope the theory is explicit and checkable: a rank test says what is recoverable, an explicit variance says how much data a target needs and where dimension and conditioning enter, a matching lower bound says that no method does better, and three separated statements say when an uncertain channel is harmless, costly, or fatal. The census study shows the same objects computing prices for real survey designs before any data are collected.

REPRODUCIBILITY STATEMENT

All theoretical statements are proved in Appendices A to D with their assumptions stated in full. Appendix E gives the main experimental configurations (alphabets, kernels, priors, targets, sample sizes, replications) with Monte Carlo standard errors; the remaining details (random seeds, the randomized configurations, the calibration and coarsening designs in full) are fixed inside the scripts. The census extract is public; the randomization code, the linear programs for γ, and the concave maximizations for C^* are included in the anonymized supplementary code, which reproduces every table and figure on a single CPU.

ETHICS STATEMENT

The work is theoretical. The only data used are a public census extract with an author-imposed randomization; no individual is identifiable and no new data were collected. The methods are designed to estimate population averages under privacy mechanisms and imperfect measurement, and we see no direct negative societal consequences.

AI USE STATEMENT

Large language models were used substantially in preparing this manuscript: drafting and editing text, preparing LaTeX, searching and checking the literature, writing and reviewing experimental code, and cross-checking derivations, proofs, and numerical results. The authors posed the research questions, chose the statistical models and assumptions, decided every claim, derived and verified the mathematical statements and proofs and the experimental results, and take full responsibility for the content. Every bibliography entry was checked against its primary record.

REFERENCES

Emre Acartürk, Burak Varıcı, Karthikeyan Shanmugam, and Ali Tajer. Sample complexity of interventional causal representation learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 372–407. PMLR, 2023.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

Fernando G. S. L. Brandão, Aram W. Harrow, James R. Lee, and Yuval Peres. Adversarial hypothesis testing and a quantum Stein's lemma for restricted measurements. *IEEE Transactions on Information Theory*, 66(8):5037–5054, 2020.

John P. Buonaccorsi. *Measurement Error: Models, Methods, and Applications*. Chapman & Hall/CRC, 2010.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Richard D. Gill and Boris Y. Levit. Applications of the van Trees inequality: A Bayesian Cramér–Rao bound. *Bernoulli*, 1(1/2):59–79, 1995.

Gene H. Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10(2):413–432, 1973.

Sander Greenland. Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*, 7:745–757, 1988.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

Daniella Horan, Eitan Richardson, and Yair Weiss. When is unsupervised disentanglement possible? In *Advances in Neural Information Processing Systems*, volume 34, pp. 5150–5161, 2021.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2436–2444. PMLR, 2016.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Ron Kohavi. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207, 1996.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 13557–13603. PMLR, 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4114–4124. PMLR, 2019.

Qi Lyu and Xiao Fu. On finite-sample identifiability of contrastive learning-based nonlinear independent component analysis. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 14582–14600. PMLR, 2022.

Wang Miao and Eric J. Tchetgen Tchetgen. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016.

Jerzy Neyman. $c(\alpha)$ tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, 41:1–21, 1979.

C. Radhakrishna Rao and Sujit Kumar Mitra. Generalized inverse of a matrix and its applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, pp. 601–620. University of California Press, 1972.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

Walter J. Rogan and Beth Gladen. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1):71–76, 1978.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

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

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12979–12990. PMLR, 2021.

A Proof of Proposition 1

Proof. ($\supseteq$) If $v = A^\top a$ and $Ap = Ap'$ then $\langle v, p\rangle = a^\top Ap = a^\top Ap' = \langle v, p'\rangle$. ($\subseteq$) Let $v \notin \mathrm{col}(A^\top)$ and decompose $v = v_\parallel + v_\perp$ with $v_\parallel \in \mathrm{col}(A^\top)$ and $v_\perp \in \ker(A) \setminus \{0\}$. Since every column of A sums to one, $\mathbf{1}_K^\top A = \mathbf{1}_d^\top$, hence $\mathbf{1}_d^\top v_\perp = \mathbf{1}_K^\top A v_\perp = 0$. Choose an interior p_0 and $\delta > 0$ so small that $p_\pm := p_0 \pm \delta v_\perp / \|v_\perp\|$ have positive entries; they sum to one because $\mathbf{1}^\top v_\perp = 0$. Then $Ap_+ = Ap_-$ while $\langle v, p_+ - p_-\rangle = 2\delta\|v_\perp\| \neq 0$. The pair (p_+, p_-) is the witness. If a slab $\Delta_{\geq\varepsilon}$ with $\varepsilon < 1/d$ is prescribed, take $p_0 = \mathbf{1}/d$ and δ small enough that $p_\pm \in \Delta_{\geq\varepsilon}$. □

B Proofs for Section 4

Throughout, $K_+ := |\Omega_+|$, $\Delta := \hat{q}_n - q$ on Ω_+, and $M(q') := A^\top D_{q'}^{-1} A$.

Lemma 11 (Feasibility and oracle variance)**.** *Let $q' > 0$ on Ω_+. (a) $\ker M(q') = \ker A$ and $\mathrm{col}M(q') = \mathrm{col}(A^\top)$. (b) For $v \in \mathrm{col}(A^\top)$, $a^*(v; q') := D_{q'}^{-1} A\, M(q')^+ v$ satisfies $A^\top a^* = v$. (c) For $q = Ap$, $\mathrm{Var}_q\langle a^*(v; q), \hat{q}_n\rangle = V_A(v; q)/n$ with V_A as in* (2)*.*

Proof. (a) $A\delta = 0$ implies $M\delta = 0$; conversely $M\delta = 0$ gives $\|D_{q'}^{-1/2} A\delta\|^2 = \delta^\top M\delta = 0$. (b) $A^\top a^* = MM^+ v = v$ since $v \in \mathrm{col}M$. (c) $\mathrm{Cov}(\hat{q}_n) = (D_q - qq^\top)/n$, so $n\mathrm{Var} = a^{*\top} D_q a^* - (a^{*\top} q)^2 = v^\top M^+ v - \theta^2$, using $a^{*\top} D_q a^* = v^\top M^+ A^\top D_q^{-1} A M^+ v = v^\top M^+ v$ and $a^{*\top} q = a^{*\top} Ap = v^\top p = \theta$. □

Lemma 12 (Multinomial moments)**.** *For $\Delta = \hat{q}_n - q$: $\mathbb{E}[\Delta_i \Delta_j \Delta_k] = \mu_{ijk}(q)/n^2$ with $\mu_{ijk}(q) = \delta_{ijk} q_i - \delta_{ij} q_i q_k - \delta_{ik} q_i q_j - \delta_{jk} q_j q_i + 2 q_i q_j q_k$, so $|\mu_{ijk}| \leq 5$; and $\mathbb{E}\|\Delta\|_2^4 \leq 3K_+^2/n^2$.*

Proof. Write $\Delta_a = \frac{1}{n}\sum_\ell (Z_a^{(\ell)} - q_a)$ with $Z_a^{(\ell)} = \mathbf{1}\{Y_\ell = a\}$. In the triple sum only $\ell_1 = \ell_2 = \ell_3$ survives by centering and independence; expanding $\mathbb{E}[(Z_i - q_i)(Z_j - q_j)(Z_k - q_k)]$ with $Z_a Z_b = \delta_{ab} Z_a$ gives μ_{ijk}, and there are n such terms out of n^3. Each summand of μ_{ijk} is a product of probabilities. For the fourth moment, $n^4 \mathbb{E}\Delta_k^4 = n q_k(1 - q_k)[1 + 3(n-2)q_k(1-q_k)] \leq 3n^2/16 + n/4 \leq 3n^2$ so $\mathbb{E}\Delta_k^4 \leq 3/n^2$; Cauchy–Schwarz handles $k \neq \ell$, and there are K_+^2 terms. □

Lemma 13 (Smoothness of the weights)**.** *Fix $\varepsilon > 0$, $v \in \mathrm{col}(A^\top)$. The slab image $\mathcal{Q}_\varepsilon := \{Ap : p \in \Delta_{\geq\varepsilon}\}$ is compact with all coordinates on Ω_+ bounded below by $r_0 > 0$. On the closed $r_0/2$-neighborhood $\mathcal{K}_0$ of $\mathcal{Q}_\varepsilon$, the map $q' \mapsto a^*(v; q')$ is real analytic, with $\sup_{\mathcal{K}_0} \|\nabla a^*\| \leq C_J$, $\sup_{\mathcal{K}_0} \|\nabla^2 a^*\|_{\mathrm{op}} \leq C_S$ and $\sup_{\mathcal{K}_0} \|\nabla^3 a^*\|_{\mathrm{op}} \leq C_T$ for finite constants depending only on ε, A, v.*

Proof. On $\mathcal{K}_0$ each $q'_y \in [r_0/2, 1 + r_0]$, so $D_{q'}^{-1}$ is analytic. By Lemma 11(a), $\ker M(q') = \ker A$ and $\mathrm{col}M(q') = \mathrm{col}(A^\top) =: W$ for every $q' \in \mathcal{K}_0$. Hence $M(q')$ restricted to W is an analytic family of invertible operators $W \to W$, and $M(q')^+$ is the inverse of that family on W composed with the fixed orthogonal projection onto W: $M(q')^+ = P_W\,(M(q')|_W)^{-1} P_W$. Inversion of an analytic invertible family is analytic (Cramer's rule), so $M(q')^+$ and $a^*(v; q') = D_{q'}^{-1} A M(q')^+ v$ are analytic on the compact set $\mathcal{K}_0$; continuity of the first three derivatives yields C_J, C_S, C_T. (For general constant-rank families this is the differentiability theorem of Golub & Pereyra (1973); here the fixed range and kernel make it elementary.) □

Proof of Theorem 2(a). Write $a := a^*(v; q)$, $h := \tilde{q}_n - q = (1 - 1/n)\Delta + (u - q)/n$.

Step 1 (exact decomposition). Since $A^\top a^*(v; q') = v$ for every positive q' (Lemma 11(b)), $q^\top a^*(v; q') = p^\top A^\top a^*(v; q') = \theta$ for every such q'; hence $q^\top(\hat{a}_n - a) = 0$ exactly and $\hat{T}_n - \theta = \underbrace{a^\top \Delta}_{L_n} + \underbrace{(\hat{a}_n - a)^\top \Delta}_{R_n}$.

Step 2. $\mathbb{E}L_n = 0$ and $\mathbb{E}L_n^2 = V_A/n$ by Lemma 11(c).

Step 3 (good event). Let $G_n := \{\|\Delta\|_\infty \leq \sqrt{\log n/n}\}$. Hoeffding and a union bound give $P(G_n^c) \leq 2K_+/n^2$. On G_n^c, $|\hat{T}_n - \theta| \leq \|\hat{a}_n\|_\infty + |\theta|$, and $\|\hat{a}_n\|_\infty \leq \|v\|_2\, \|M(\tilde{q}_n)^+\|\, K_+ / \min_y \tilde{q}_{n,y}$ is at

most a polynomial in n by the regularization $\tilde{q}_n \geq 1/(nK_+)$ and $\|M(\tilde{q})^+\| \leq 1/\sigma^+_{\min}(M(\tilde{q})) \leq \max_y \tilde{q}_y/s_+(A)^2 \leq 1/s_+(A)^2$ (Lemma 4 argument). Since $P(G_n^c)$ decays as n^{-2} and can be made n^{-r} for any r by enlarging the constant in G_n, the bad-event contribution is $O(1/n^2)$.

Step 4 (Taylor on the good event). For n large, G_n implies $\tilde{q}_n \in \mathcal{K}_0$, and Lemma 13 gives $\hat{a}_n - a = Jh + S(h)$ with $J := \nabla a^*(v; q)$, $\|J\| \leq C_J$, $\|S(h)\| \leq C_S\|h\|^2$. Hence $R_n = \Delta^\top Jh + S(h)^\top \Delta$, and on G_n, $|R_n| \leq C_J\|\Delta\|^2 + C_J\|u-q\|\|\Delta\|/n + C_S\|h\|^2\|\Delta\|$. Squaring and applying Lemma 12, $\mathbb{E}[R_n^2 \mathbf{1}_{G_n}] = O(C_J^2 K_+^2/n^2)$.

Step 5 (cross term). The leading part of $\mathbb{E}[L_n R_n \mathbf{1}_{G_n}]$ is $(1-1/n)\sum_{i,j,k} a_i J_{jk} \mathbb{E}[\Delta_i \Delta_j \Delta_k]$, which is bounded by $5\|a\|_1\|J\|_{1\to 1} K_+^3/n^2$ by Lemma 12; the remaining pieces are $O(\mathbb{E}\|\Delta\|^2/n) = O(1/n^2)$ and $O(\mathbb{E}\|\Delta\|^2\|h\|^2) = O(1/n^2)$. Removing the indicator changes each term by $O(P(G_n^c)^{1/2} \cdot \text{poly}) = O(1/n^2)$ after enlarging the good-event constant as in Step 3.

Step 6. Adding $\mathbb{E}L_n^2 = V_A/n$, the cross and square terms, and the bad-event contribution gives $\mathbb{E}(\hat{T}_n - \theta)^2 = V_A/n + O(1/n^2)$, with all constants depending only on ε, A, v. □

Proof of Theorem 2(b). Let $V_* := \sup_p V_A(v; Ap)$ over the slab, fix $\eta > 0$, and choose p_0 in the relative interior of the slab ($\min_j p_{0,j} > \varepsilon$) with $V(p_0) := V_A(v; Ap_0) \geq V_* - \eta$, which is possible because $p \mapsto V_A(v; Ap)$ is continuous on the slab and the relative interior is dense in it. With $I_0 := M(q_0)$, $q_0 = Ap_0$, set $\delta_0 := I_0^+(v - \theta_0 \mathbf{1})$, $\theta_0 := \langle v, p_0 \rangle$. Then $\mathbf{1}^\top \delta_0 = 0$ (because $\mathbf{1} \in \text{col}(A^\top)$ and I_0^+ maps $\text{col}(A^\top)$ to itself while $\mathbf{1}^\top I_0^+ w = q_0^\top D_{q_0}^{-1} A I_0^+ w = p_0^\top I_0 I_0^+ w = p_0^\top w$ for $w \in \text{col}(A^\top)$, which vanishes at $w = v - \theta_0 \mathbf{1}$), the target derivative is $v^\top \delta_0 = V(p_0)$, and $\delta_0^\top I_0 \delta_0 = V(p_0)$. Consider the one-dimensional submodel $p_t := p_0 + t\delta_0$, $|t| \leq h_n := n^{-1/2+\alpha}$ with $\alpha \in (0, 1/2)$, which lies in the slab for all large n because p_0 is in its relative interior, and the prior $\pi_n(t) = h_n^{-1} \cos^2(\pi t/(2h_n))$ with Fisher information $\pi^2/h_n^2 = o(n)$. The van Trees inequality (Gill & Levit, 1995) gives $\sup_p \mathbb{E}_p(\hat{\theta}_n - \theta)^2 \geq V(p_0)^2/(n(V(p_0) + o(1)) + o(n))$, so $\liminf_n n \sup_p \text{MSE} \geq V(p_0) \geq V_* - \eta$. □

Proof of Theorem 2(c). By Steps 1 to 5, $R_n = O_p(1/n)$, so $\sqrt{n}(\hat{T}_n - \theta) = a^\top \sqrt{n}\Delta + o_p(1) \to \mathcal{N}(0, a^\top(D_q - qq^\top)a) = \mathcal{N}(0, V_A)$ by the multinomial CLT and Slutsky. For efficiency: the score of the submodel $p + th$ ($\mathbf{1}^\top h = 0$) at y is $s_h(y) = (Ah)_y/q_y$, and $\mathbb{E}_q[\phi_q(Y)s_h(Y)] = a^\top Ah - \theta \mathbf{1}^\top Ah = h^\top v = \frac{d}{dt}\theta(p + th)|_0$ with $\phi_q(y) := a_y - \theta$, so ϕ_q is an influence function. It lies in the tangent space: with $h^* := M(q)^+ v$ and $h_0 := h^* - \theta p$, we have $\mathbf{1}^\top h^* = q^\top a = \theta$ (columns of A sum to one), so $\mathbf{1}^\top h_0 = 0$ and $s_{h_0}(y) = (Ah^*)_y/q_y - \theta = a_y - \theta = \phi_q(y)$. An influence function in the tangent space is the efficient one, and V_A is the efficiency bound (van der Vaart, 1998, Theorem 25.20). □

Proof of Theorem 2(d). Take the witness pair (p_+, p_-) of Proposition 1 inside the slab (Appendix A, last sentence): the two laws of the data coincide, so any estimator has the same distribution under both, and $\max_\pm \mathbb{E}_{p_\pm}(\hat{\theta}_n - \langle v, p_\pm \rangle)^2 \geq \langle v, p_+ - p_- \rangle^2/4 > 0$ for every n. □

Proof of Proposition 3. $a - a^* \in \ker(A^\top)$ and $D_q a^* = AM(q)^+ v \in \text{col}(A)$, so $\langle D_q a^*, a - a^* \rangle = \langle M(q)^+ v, A^\top(a - a^*) \rangle = 0$ and $a^\top D_q a = a^{*\top} D_q a^* + (a - a^*)^\top D_q (a - a^*)$; subtract θ^2 on both sides (both weightings are unbiased). For the dichotomy, $\max_i \|a - a^*(q_i)\|_2 \geq \Delta/2$ by the triangle inequality, and $\|w\|^2_{D_{q_i}} \geq c_A \varepsilon \|w\|_2^2$ on the slab image, because $q_y = \sum_j A_{y,j} p_j \geq \varepsilon \sum_j A_{y,j} \geq c_A \varepsilon$ for every $y \in \Omega_+$ when $p \geq \varepsilon$ entrywise. □

Proof of Lemma 4. Since $q_y \leq 1$, $D_q^{-1} \succeq I$ and $M(q) \succeq A^\top A$; since $q_y \geq q_{\min}$, $M(q) \preceq q_{\min}^{-1} A^\top A$. The three matrices share kernel $\ker A$ and range $\text{col}(A^\top)$, so the Loewner order inverts under the pseudoinverse on that range: $q_{\min}(A^\top A)^+ \preceq M(q)^+ \preceq (A^\top A)^+$. Evaluate at $v \in \text{col}(A^\top)$ and subtract θ^2. The last inequality is $v^\top (A^\top A)^+ v \leq \|v\|^2/s_+^2$. □

C Proof of Theorem 5

Throughout, $\operatorname{rank}(A) = d$, so $G(A', q') := A'^{\top} D_{q'}^{-1} A'$ is invertible for every channel A' near A and every positive q', and the weights are $a^*(A'; q') := D_{q'}^{-1} A' G(A', q')^{-1} v$. Write $q = Ap$, $\Delta := \hat{q}_n - q$, $E := \hat{A} - A$, $h := \tilde{q}_n - q = (1 - 1/n)\Delta + (u - q)/n$, and $\hat{\theta}_0 := \langle a^*(\hat{A}; \tilde{q}_n), \hat{q}_n \rangle$ for the estimator before clipping. All constants below depend only on ε, A, v.

Step 1 (exact decomposition). For every column-stochastic A' of rank d and every positive r, feasibility $A'^{\top} a^*(A'; r) = v$ gives $a^*(A'; r)^{\top} q = a^*(A'; r)^{\top} A' p + a^*(A'; r)^{\top} (A - A') p = \theta + p^{\top} (A - A')^{\top} a^*(A'; r)$. Hence, with $\hat{a} := a^*(\hat{A}; \tilde{q}_n)$,

$$\hat{\theta}_0 - \theta = \hat{a}^{\top} \Delta - p^{\top} E^{\top} \hat{a}. \tag{4}$$

Step 2 (calibration noise). The columns $E_{\cdot j}$ are independent, centered, and supported on the support of $A_{\cdot j}$ (an empirical law of m draws never puts mass outside the true support), with $\operatorname{Cov}(E_{\cdot j}) = (\operatorname{diag}(A_{\cdot j}) - A_{\cdot j} A_{\cdot j}^{\top})/m$. Lemma 12 applies to each column with m in place of n: $\mathbb{E}[E_{ij} E_{kj} E_{lj}] = O(m^{-2})$ and $\mathbb{E}\|E_{\cdot j}\|^4 = O(m^{-2})$. More generally, for a centered average of m bounded vectors, $\mathbb{E}\|E_{\cdot j}\|^{2k} \le C_k K^k m^{-k}$ for every fixed k (integrate the Hoeffding tail), and likewise $\mathbb{E}\|\Delta\|^{2k} \le C_k K_+^k n^{-k}$.

Step 3 (good event and smoothness). Let $G := \{\|\Delta\|_\infty \le c_0 \sqrt{\log n / n}\} \cap \{\|E\|_\infty \le c_0 \sqrt{\log m / m}\}$. By Hoeffding and a union bound over the K_+ cells and the $K_+ d$ channel entries, $P(G^c) \le 2K_+ n^{-2c_0^2} + 2K_+ d\, m^{-2c_0^2}$, and we take c_0 large enough that $P(G^c) \le n^{-4} + m^{-4}$ for all $n, m \ge 2$. On G, for n, m large, the pair $(\hat{A}, \tilde{q}_n)$ lies in a compact neighborhood $\mathcal{K}_1$ of (A, q) on which every channel has rank d (Weyl's inequality: $s_{\min}(\hat{A}) \ge s_{\min}(A) - \|E\|$), all supported entries and all coordinates of q' are bounded below by a positive constant, and $(A', q') \mapsto a^*(A'; q')$ is real analytic (the entries of $G(A', q')^{-1}$ are rational functions with nonvanishing denominator). Its derivatives up to order three are therefore bounded on $\mathcal{K}_1$ by constants C_1, C_2, C_3.

Step 4 (channel derivative). Let $\phi(A') := a^*(A'; q)^{\top} q$ with q fixed. Using $D_q^{-1} q = \mathbf{1}_K$ and $A'^{\top} \mathbf{1}_K = \mathbf{1}_d$, $\phi(A') = v^{\top} G(A', q)^{-1} \mathbf{1}_d$. With $\mathrm{d}G = \mathrm{d}A^{\top} D_q^{-1} A + A^{\top} D_q^{-1} \mathrm{d}A$ and the identities $G(A, q)^{-1} \mathbf{1}_d = p$ (because $G(A, q) p = A^{\top} D_q^{-1} q = \mathbf{1}_d$) and $D_q^{-1} A p = \mathbf{1}_K$,

$$\mathrm{d}\phi = -v^{\top} G^{-1} \mathrm{d}A^{\top} \mathbf{1}_K - v^{\top} G^{-1} A^{\top} D_q^{-1}\, \mathrm{d}A\, p = -a^{*\top} \mathrm{d}A\, p,$$

the first term vanishing because admissible perturbations have zero column sums.

Step 5 (Taylor expansion on the good event). On G, expand $\hat{a} = a^*(A + E; q + h)$ to second order in (E, h) with a third-order remainder: $\hat{a} = a^* + J_q h + J_A E + \frac{1}{2} \nabla^2 a^*[(h, E), (h, E)] + r_3$, where J_q, J_A are the partial derivatives at (A, q) and $\|r_3\| \le C_3 (\|h\| + \|E\|)^3$. Substituting into (4) and collecting terms by order gives

$$\hat{\theta}_0 - \theta = L + Q + R, \qquad L := a^{*\top} \Delta - a^{*\top} E p,$$

where Q collects the terms of exact second order in (Δ, E) together with the terms of first order in (Δ, E) that carry a factor $1/n$ from the deterministic part $(u - q)/n$ of h, namely

$$\begin{aligned} Q = {} & (1 - \tfrac{1}{n}) \Delta^{\top} J_q \Delta + \Delta^{\top} J_A E - (1 - \tfrac{1}{n}) p^{\top} E^{\top} J_q \Delta - p^{\top} E^{\top} J_A E \\ & + \tfrac{1}{n} \Delta^{\top} J_q (u - q) - \tfrac{1}{n}\, p^{\top} E^{\top} J_q (u - q), \end{aligned}$$

and R collects everything else: terms of third order in (Δ, E), terms of second order carrying a factor $1/n$, and the products of r_3 with Δ or E. On G, $|R| \le C(\|\Delta\| + \|E\|)^3 + Cn^{-1}(\|\Delta\| + \|E\|)^2 + Cn^{-2}(\|\Delta\| + \|E\|)$.

Step 6 (the linear term). The two pieces of L are independent and centered, so $\mathbb{E}L^2 = a^{*\top} \operatorname{Cov}(\hat{q}_n) a^* + \sum_j p_j^2 a^{*\top} \operatorname{Cov}(E_{\cdot j}) a^* = V_A(v; q)/n + C_{\mathrm{cal}}/m$ by Lemma 11(c) and Step 2. Moreover $\mathbb{E}L^4 = O(n^{-2} + m^{-2})$ by Step 2.

Step 7 (cross terms). We bound $\mathbb{E}[LQ]$ term by term, using independence of Δ and E and that both are centered. (i) $\mathbb{E}[(a^{*\top}\Delta)(\Delta^\top J_q\Delta)] = \sum_{i,j,k} a^*_i (J_q)_{jk}\mathbb{E}[\Delta_i\Delta_j\Delta_k] = O(n^{-2})$ by Lemma 12; $\mathbb{E}[(a^{*\top}Ep)(\Delta^\top J_q\Delta)] = 0$. (ii) $\mathbb{E}[(a^{*\top}\Delta)(\Delta^\top J_A E)]$ and $\mathbb{E}[(a^{*\top}Ep)(\Delta^\top J_A E)]$ both vanish, each containing a lone centered factor from one of the two independent samples. (iii) $\mathbb{E}[(a^{*\top}\Delta)(p^\top E^\top J_q\Delta)] = 0$ and $\mathbb{E}[(a^{*\top}Ep)(p^\top E^\top J_q\Delta)] = 0$ for the same reason. (iv) $\mathbb{E}[(a^{*\top}\Delta)(p^\top E^\top J_A E)] = 0$; $\mathbb{E}[(a^{*\top}Ep)(p^\top E^\top J_A E)]$ is a sum of third moments of the columns of E, hence $O(m^{-2})$. (v) $\frac{1}{n}\mathbb{E}[(a^{*\top}\Delta)(\Delta^\top J_q(u-q))] = O(n^{-2})$ and $\frac{1}{n}\mathbb{E}[(a^{*\top}Ep)(p^\top E^\top J_q(u-q))] = O((nm)^{-1})$; the two remaining mixed products vanish. Hence $\mathbb{E}[LQ] = O(n^{-2} + m^{-2} + (nm)^{-1})$. For the remainder, Cauchy–Schwarz and the moment bounds of Step 2 give $|\mathbb{E}[LR\mathbf{1}_G]| \le (\mathbb{E}L^2)^{1/2}(\mathbb{E}[R^2\mathbf{1}_G])^{1/2} = O\big((n^{-1/2}+m^{-1/2})(n^{-3/2}+m^{-3/2}+n^{-2}+n^{-3})\big)$, and every product here is $O(n^{-2}+m^{-2})$: the mixed ones because $n^{-1/2}m^{-3/2} \le \max(n^{-2}, m^{-2})$ and $m^{-1/2}n^{-3/2} \le \max(n^{-2}, m^{-2})$ (compare the two cases $n \ge m$ and $n < m$).

Step 8 (square terms). $\mathbb{E}[Q^2\mathbf{1}_G] \le C(\mathbb{E}\|\Delta\|^4 + \mathbb{E}\|\Delta\|^2\|E\|^2 + \mathbb{E}\|E\|^4 + n^{-2}\mathbb{E}\|\Delta\|^2 + n^{-2}\mathbb{E}\|E\|^2) = O(n^{-2} + (nm)^{-1} + m^{-2})$ and $\mathbb{E}[R^2\mathbf{1}_G] = O(n^{-3} + m^{-3} + n^{-2}(n^{-2}+m^{-2}))$, by Step 2 and independence.

Step 9 (assembling). Because $\hat\theta$ is clipped to $[-\|v\|_\infty, \|v\|_\infty]$ and $|\theta| \le \|v\|_\infty$, the bad event contributes at most $4\|v\|_\infty^2 P(G^c) = O(n^{-4} + m^{-4})$. On G, for n, m large, $|\hat\theta_0 - \theta| \le |L| + |Q| + |R| < \|v\|_\infty$, so the clipping is inactive and $\hat\theta = \hat\theta_0$ there. Removing the indicator from $\mathbb{E}[L^2\mathbf{1}_G]$ and from the cross terms costs at most $(\mathbb{E}L^4)^{1/2}P(G^c)^{1/2} + (\mathbb{E}Q^4 + \mathbb{E}R^4)^{1/2}P(G^c)^{1/2} = O(n^{-3} + m^{-3})$. Adding Steps 6 to 8,

$$\mathbb{E}(\hat\theta - \theta)^2 = \frac{V_A(v;q)}{n} + \frac{C_{\mathrm{cal}}}{m} + O\Big(\frac{1}{n^2} + \frac{1}{m^2} + \frac{1}{nm}\Big),$$

with constants depending only on ε, A, v, uniformly over the slab.

Exactly identified check. If $K_+ = d$ and A is invertible on Ω_+, then $a^* = A^{-\top}v$ and $\hat\theta_0 = v^\top \hat A^{-1}\hat q_n$; the delta method gives the same derivative $-v^\top A^{-1}\mathrm{d}A\,A^{-1}q = -a^{*\top}\mathrm{d}A\,p$, and for $d = 2$ the constant is the Rogan–Gladen variance with estimated sensitivity and specificity.

D Proofs for Section 5

Proof of Proposition 6. The two parameter points produce the same data law, so any estimator has the same distribution under both, and $\max_i \mathbb{E}_i(\hat\theta_n - t_i)^2 \ge (t_1 - t_2)^2/4$ for $t_1 = \theta_m(p)$, $t_2 = \theta_{m'}(p')$. □

Proof of Theorem 7. Fix the truth (m, p), $q = A^{(m)}p$. On $\{\|\hat q_n - q\|_1 < \gamma/2\}$ the nearest-polytope selector returns m, because $\mathrm{dist}_1(\hat q_n, \mathcal{Q}_m) < \gamma/2 < \mathrm{dist}_1(\hat q_n, \mathcal{Q}_{m'})$ for $m' \ne m$. Hoeffding with a union bound over K cells gives $P(\hat m \ne m) \le P(\|\hat q_n - q\|_\infty \ge \gamma/(2K)) \le 2K\exp(-n\gamma^2/(2K^2)) \le 2K\exp(-n\gamma^2/(8K^2))$ (a factor 4 in the exponent is given away to keep the statement simple). On the good event the risk is at most C_m/n; on its complement the clipped error is at most $2B$, contributing at most $4B^2 \cdot 2Ke^{-n\gamma^2/(8K^2)}$. □

Proof of Proposition 8. Fix a pair $\mathcal{P} := \mathcal{Q}^\varepsilon_m$, $\mathcal{Q} := \mathcal{Q}^\varepsilon_{m'}$ and $s \in (0,1)$. The map $(q, q') \mapsto \rho_s(q, q')$ is continuous and jointly concave (each summand is a weighted geometric mean), so its supremum over the compact convex product is attained at some (μ, ν) with entries at least $\beta > 0$ on Ω_+ (cells outside Ω_+ have probability zero under every candidate and are dropped from the alphabet). First-order optimality along $\mu_t = (1-t)\mu + tq$ gives $\sum_y q_y(\nu_y/\mu_y)^{1-s} \le \rho_s(\mathcal{P}, \mathcal{Q})$ for all $q \in \mathcal{P}$, and symmetrically $\sum_y q'_y(\mu_y/\nu_y)^s \le \rho_s(\mathcal{P}, \mathcal{Q})$ for all $q' \in \mathcal{Q}$. Let $T := \prod_i \nu(Y_i)/\mu(Y_i)$ and vote m' when $T > 1$. Under $q \in \mathcal{P}$, Markov's inequality on T^{1-s} and independence give $P_q(T > 1) \le (\sum_y q_y(\nu_y/\mu_y)^{1-s})^n \le \rho_s(\mathcal{P}, \mathcal{Q})^n$; under $q' \in \mathcal{Q}$, the same on T^{-s}. Both errors are at most $\rho_s(\mathcal{P}, \mathcal{Q})^n$ uniformly over the polytopes. By Berge's maximum theorem $s \mapsto \rho_s(\mathcal{P}, \mathcal{Q})$ is continuous on $[0, 1]$ with endpoint values 1, and disjointness gives $\rho_s < 1$ inside, so the supremum of

$-\log \rho_s$ is attained at an interior s^*; the extremal pair (μ, ν) of each candidate pair is computed at its own s^*. The selector runs this test for every unordered pair and returns the candidate that wins all $M-1$ of its comparisons, choosing arbitrarily if there is none. If m is the truth, then $\hat{m} \neq m$ only if m loses at least one of its comparisons, and each loss has probability at most e^{-nC^*}; the union bound gives $(M-1)e^{-nC^*}$ uniformly over p in the slab.

For the identification $C^*(\mathcal{P}, \mathcal{Q}) = \inf_{(q,q')} C(q, q')$: $f(s,(q,q')) := -\log \rho_s(q,q')$ is concave in s (log-convexity of $s \mapsto \rho_s$ by Hölder) and convex in (q, q') (concavity of ρ_s), continuous on $[\sigma, 1-\sigma] \times (\mathcal{P} \times \mathcal{Q})$; Sion's minimax theorem exchanges $\sup_s$ and $\inf_{(q,q')}$ on that set, and uniform continuity on the compact $[0,1] \times (\mathcal{P} \times \mathcal{Q})$ with $f(0,\cdot) = f(1,\cdot) = 0$ lets $\sigma \downarrow 0$.

Converse: $C(\cdot,\cdot)$ is continuous on the compact product, so the infimum is attained at $(\bar{q}, \bar{q}')$. Any selector induces a binary test between $\bar{q}^{\otimes n}$ and $\bar{q}'^{\otimes n}$ whose errors are bounded by the multiclass errors; the Chernoff converse (Cover & Thomas, 2006, Sec. 11.9) gives $\alpha_n + \beta_n \geq e^{-n(C(\bar{q},\bar{q}')+o(1))}$, hence the multiclass minimax error is at least $\frac{1}{2}e^{-n(C^*+o(1))}$. The binary statement, converse included and in a stronger adaptive-adversary model, is Theorem 7 of Brandão et al. (2020).

Finally, $s = \frac{1}{2}$ gives the Bhattacharyya affinity $1 - H^2$, $-\log(1-x) \geq x$, and $H^2 \geq \|q - q'\|_1^2/8$ by Cauchy–Schwarz, whence $C^* \geq \gamma^2/8$. The example: on a binary alphabet, $\mathrm{Bern}(0.5)$ versus $\mathrm{Bern}(0.6)$ and $\mathrm{Bern}(0.01)$ versus $\mathrm{Bern}(0.11)$ both have $\gamma = 0.2$ with Chernoff informations 0.00508 and 0.02956. The estimation consequence follows from the risk decomposition of Theorem 7 with the new selector. □

Proof of Theorem 9. Write $q_k := A^{(m_k)}p_k$, $\theta_k := \theta_{m_k}(p_k)$, $\rho := \rho_{1/2}(q_1, q_2)$. Le Cam's two-point bound gives $\max_k \mathbb{E}_k(\hat{\theta}_n - \theta_k)^2 \geq (\Delta_0^2/8)(1 - \|P_{q_1}^{\otimes n} - P_{q_2}^{\otimes n}\|_{TV})$, and $\|P^{\otimes n} - Q^{\otimes n}\|_{TV} \leq \sqrt{1-\rho^{2n}}$ (Tsybakov, 2009, Lemma 2.3), so with $1 - \sqrt{1-x} \geq x/2$ the risk is at least $(\Delta_0^2/16)\rho^{2n}$. On a common support with all coordinates at least $q_{\min}$, $H^2(q_1,q_2) = \frac{1}{2}\sum_y (q_{1,y} - q_{2,y})^2/(\sqrt{q_{1,y}} + \sqrt{q_{2,y}})^2 \leq \|q_1 - q_2\|_1^2/(8q_{\min})$, so $\rho = 1 - H^2 \geq 1 - x$ with $x := \|q_1 - q_2\|_1^2/(8q_{\min}) \leq C_H^2\gamma_n^2/(8q_{\min}) \leq 1/2$. Using $\log(1-x) \geq -x/(1-x) \geq -2x$ for $x \leq 1/2$, $\rho^{2n} \geq \exp(-4nx) \geq \exp(-C_H^2 n\gamma_n^2/(2q_{\min}))$. □

Proof of Corollary 10. If $n\gamma_n^2 \geq 8K^2 \log n + \omega(1)$ the selection term of Theorem 7 is $o(1/n)$, so the upper bound is $V_{\max}/n + o(1/n)$ by the assumed uniformity of Theorem 2(a) along the sequence, and the per-model van Trees bound (Theorem 2(b), uniform along the sequence by assumption, applied to the true model and maximized over m) gives the matching lower bound; the same holds with Proposition 8 when $nC_n^* - \log n \to \infty$. If the witness exists and $n\gamma_n^2 \leq (2q_{\min}/C_H^2)\log n - \omega(1)$, Theorem 9 gives a lower bound of at least $(\Delta_0^2/16)\exp(-C_H^2 n\gamma_n^2/(2q_{\min})) \geq (\Delta_0^2/16)\, n^{-1} e^{\omega(1)}$, so n times the risk tends to infinity; if $\gamma_n\sqrt{n} \to c < \infty$ the same bound converges to $(\Delta_0^2/16)e^{-C_H^2 c^2/(2q_{\min})} > 0$. In the intermediate band the upper bound decays like $e^{-n\gamma_n^2/(8K^2)}$, slower than $1/n$, while the two-point lower bound is $o(1/n)$; neither side determines the rate. □

E Experimental details and additional results

Estimator and defaults. All simulations use the regularized plug-in (3) exactly as stated; oracle curves use $a^*(v;q)$ with the true q. Monte Carlo standard errors (MC-SE) are reported for every rescaled quantity; the standard error of $n\cdot$MSE is $\sqrt{2}\, V_A n/\sqrt{N_{\mathrm{MC}}} \cdot (1/n)$ to first order.

Configurations for the rate experiments. C1: $d = 4$, $|\mathcal{X}| = 2$ (two-to-one f), $|\mathcal{S}| = 3$, $\mathrm{rank}A = 4$, $v = e_1$. C2: $d = 3$, $|\mathcal{X}| = 1$, $|\mathcal{S}| = 3$, $\mathrm{rank}A = 3$, $v = e_1$. C3: $d = 5$, $|\mathcal{X}| = 2$, $|\mathcal{S}| = 2$, $\mathrm{rank}A = 4 < d$, v the normalized first column of $A^\top$. Five further configurations draw a random surjective f, Dirichlet kernel rows, an interior Dirichlet p, and $v = A^\top a$ normalized. Sample sizes $n \in \{100, 1000, 10000\}$, 20 seeds, 2,000 replications per seed. Result: mean $n \cdot \mathrm{MSE}/V_A \in [0.983, 1.022]$ over the 24 cells, and the across-seed standard deviation is comparable to the MC-SE in every cell (ratio 0.6 to 1.2; for C1 at $n = 10^4$: ratio 1.003, seed SD 0.021, MC-SE 0.019).

Conditioning path (Figure 1, left). $d = 4$, fibers $f = (a, a, b, b)$, kernel rows $(.7, .2, .1)$, $(.2, .5, .3)$, $(.6, .3, .1)$ and $(1-\eta)(.6, .3, .1) + \eta(.1, .3, .6)$; $p = (.3, .2, .3, .2)$,

$v = e_3$, $n = 10^4$, 3,000 replications. Values of $(\eta, \mathrm{rank}A, \mathrm{cond}(A), V_A, n \cdot \mathrm{MSE})$: $(1.0, 4, 2.0, 0.433, 0.452)$, $(0.5, 4, 3.6, 1.233, 1.249)$, $(0.2, 4, 10.2, 5.690, 5.535)$, $(0.1, 4, 21.3, 19.98, 20.04)$, $(0.05, 4, 43.8, 74.26, 72.63)$, $(0.02, 4, 111.2, 442.8, 454.0)$; at $\eta = 0$ the rank is 3, and the MSE is 2.52×10^{-3} against the theoretical floor $((p_3 - p_4)/2)^2 = 2.5 \times 10^{-3}$.

Large alphabet. $d = 60$, $|\mathcal{X}| = 20$ (three-to-one fibers), $|\mathcal{S}| = 4$, $K = 80$, random interior p, $v = A^\top a$ normalized, 800 replications: $n \cdot \mathrm{MSE}/V_A = 0.985, 0.981, 0.947$ at $n = 10^3, 10^4, 10^5$ (the last within about one MC-SE of one); $V_A = 0.035$.

Regret dichotomy (Proposition 3). $d = 4$, fibers (a, a, b, b), $|\mathcal{S}| = 3$, kernel rows $(.9, .05, .05)$, $(.05, .9, .05)$, $(.05, .05, .9)$, $(.34, .33, .33)$, so $K = 6$, $\mathrm{rank}A = 4$; $v = e_1 - e_2$. Over the tested slab points $\max_q \|a^*(q) - a^*(q_0)\|_2 = 0.87$, where q_0 is the slab center. Fixed-weight risks are computed exactly, plug-in risks by Monte Carlo with at least 10^5 replications. At $n = 200$ the plug-in does not yet win: it ties at the first grid point (0.003071 ± 0.000010 versus 0.003082 for the center-tuned fixed weighting) and is about 2.3 to 2.5 Monte Carlo standard errors worse at the other two (0.002973 versus 0.002948; 0.003470 versus 0.003449). From $n = 400$ the plug-in has strictly smaller risk at every grid point (0.001466 versus 0.001541 at the first point), and the advantage persists through $n = 6400$.

Misspecified channel (beyond theory). The analyst uses $\kappa_{\mathrm{wrong}} = (1 - \delta)\kappa + \delta\kappa_{\mathrm{alt}}$ while data follow κ; $n = 10^4$. Asymptotic bias grows linearly (0.0010 at $\delta = 0.01$ to 0.0125 at $\delta = 0.10$) and the MSE floor quadratically, dominating V_A/n once $\mathrm{bias}^2 > V_A/n$ (around $\delta \approx 0.07$ here).

Estimated channel (Theorem 5). Each kernel row is estimated from m calibration draws per hidden value (support-aware), $n = 10^4$, 1,500 replications: MSE $= 2.2 \times 10^{-3}$, 5.7×10^{-4}, 1.7×10^{-4}, 8×10^{-5} at $m = 50, 200, 1000, 5000$ against oracle 6×10^{-5}, that is $V_A/n + c/m$ with $c \approx 0.10$ over three decades; the theorem's constant for this configuration is $C_{\mathrm{cal}} = 0.1006$, and the ratios of measured inflation to prediction are 1.06, 1.01, 1.05, 0.97, each within three MC-SE.

Scaling collapse (Figure 1, right). $d = 4$, many-to-one fibers $f = (a, a, b, b)$, $|\mathcal{S}| = 3$; the alternative kernel is a within-fiber parallel translate of κ by α, which keeps the two polytopes disjoint for every $\alpha > 0$ with $\gamma \propto \alpha$ (checked by linear programming inside the script); target the prevalence of a pooled category; $n \in \{50, 100, 200, 500, 1000, 2000, 5000\}$, six values of α, 500 trials per cell, ℓ_1-projection selector with random tie-breaking. Cells at the same $n\gamma^2$ but different (α, n) agree within MC noise (for example 0.254, 0.220, 0.222 at $n\gamma^2 \approx 0.05$); error is at most 0.012 for $n\gamma^2 \in (0.5, 1]$ and 0 of 500 trials for $n\gamma^2 > 1$; the adaptive-to-oracle MSE ratio is at most 1.056 in the worst bin and 1.000 for $n\gamma^2 > 1$. The bound $(M - 1)(\rho^*)^n$ with $s = 1/2$, computed on the full-simplex images (an upper bound on the slab-image value), exceeds the empirical selection error at every cell and is conservative at interior truths (at $\alpha = 0.2$, $n = 2000$: bound 0.836, empirical 0.002), because it tests the worst pair of polytope points while the truth sits inside; this design has many-to-one labels, and the comparison is an empirical check.

Census study. UCI Adult, $N = 32{,}561$, hidden variable marital status ($d = 7$), k-ary randomized response with keep probability $e^\varepsilon/(e^\varepsilon + 6)$, $\varepsilon \in \{0.5, 1, 2\}$, 300 replications with records resampled and the mechanism re-randomized (the sampling model that V_A prices). Estimators: naive, single-report corrected (unique feasible weights, since the single-report channel is square and invertible), and the efficient two-report plug-in on the 49-cell alphabet. Coverage of the 95% V_A interval across the twelve (ε, target, estimator) cells: 0.927 to 0.98. Contrast target (divorced minus married-civilian, $\theta = -0.323$): same pattern; the efficient two-report gain over the average of two corrected reports is 0.9%, 1.3%, 4.3% at $\varepsilon = 0.5, 1, 2$. Equal total privacy: two $\varepsilon/2$ reports give 1.38, 1.67, 1.62 times the standard deviation of one ε report. Asymmetric reports ($\varepsilon = 2$ and 0.5), prevalence target: corrected report 1 only 0.00470, plain average of corrections 0.01141, oracle inverse-variance average 0.00462, efficient joint plug-in 0.00462. Adaptation: candidates are the two-report channels at $\varepsilon \in \{1, 2\}$; selection error $1/400$ per truth arm at $n = 250$ (per-arm 95% interval $[0.00006, 0.0139]$) and $0/400$ at $n \geq 1000$; adaptive-to-oracle MSE ratio 0.995 to 1.007.