Sharp minimax rates for average treatment effects with discrete confounding under fixed overlap

CausalSmith*

August 27, 2026

Abstract

This paper studies minimax estimation of the average treatment effect in a finite-alphabet observational model with binary treatment and outcome, unrestricted categorywise nuisance functions, and fixed overlap. For each fixed interior overlap level $\epsilon \in (0, 1/2)$, there are ϵ-dependent constants $a_\epsilon, \rho_\epsilon, C_\epsilon > 0$ and a cutoff N_ϵ such that, for every sample size $n \geq N_\epsilon$ and positive covariate alphabet size $d \leq \rho_\epsilon n \log n$, the minimax mean-squared risk over the overlap-restricted iid experiment class is

$$\frac{1}{n} + \frac{d^2}{n^2(\log n)^2}$$

up to constants depending on ϵ. Over that same range $n \geq N_\epsilon$, $d > 0$, $d \leq \rho_\epsilon n \log n$, the upper bound is attained by a computable two-split hybrid estimator with universal numerical tuning: pilot-certified heavy categories are estimated by empirical treatment-control ratios, while pilot-certified light categories are estimated by a Chebyshev reciprocal polynomial whose monomials are lifted by factorial moments. Within the range $n \geq N_\epsilon$ and $d \leq \rho_\epsilon n \log n$, the rate implies a parametric regime $d = O(\sqrt{n} \log n)$ and a consistency frontier $d = o(n \log n)$. The paper also gives an overlap-phase upper envelope: a centered estimator has risk at most $n^{-1} + 4(1/2 - \epsilon)^2$ for every positive n, d, and at the randomized endpoint $\epsilon = 1/2$ the minimax risk is bracketed between $1/(100n)$ and $1/n$ for every positive n, d.

1 Introduction

Average treatment effect estimation under unconfoundedness is usually organized around outcome regression, propensity scores, and overlap. In regular low-dimensional models, this program leads to semiparametric efficiency theory and first-order robust estimators [Hahn, 1998, Hirano et al., 2003, Robins et al., 1994, Bang and Robins, 2005, van der Laan and Rubin, 2006]. Modern high-dimensional causal inference extends the same target with structured nuisance estimation, orthogonal moments, and debiasing [Belloni et al., 2017, Chernozhukov et al., 2018, 2022]. This paper studies a complementary benchmark in which the covariate is fully discrete, the alphabet size d may grow with the sample size n, and every category carries its own unrestricted treatment probability and conditional outcome means.

*Machine-generated by the CausalSmith research pipeline. The displayed formal statements and proofs are Lean-verified under the paper's theorem-local citation interface; the verification appendix records the exact scope, toolchain, and commit, including any statement that is conditional on a published input. Cited work otherwise supplies framing and positioning.

The benchmark isolates the cost of discrete confounding under fixed overlap. The observed data are iid triples $O_i = (X_i, A_i, Y_i)$, with $X_i \in \{1, \ldots, d\}$, binary treatment A_i, and binary outcome Y_i. Under the overlap level ϵ, every positive-mass category has treatment probability in $[\epsilon, 1-\epsilon]$. The target is the adjusted average treatment effect

$$\tau(P) = \sum_{k=1}^{d} p_k\{\mu_{1k} - \mu_{0k}\},$$

written in the body as a homogeneous four-cell functional so that categories with small empirical support can be analyzed directly. Under consistency and conditional exchangeability, this observed-law functional has the usual potential-outcome interpretation.

The main result gives the sharp fixed-interior minimax rate. For each fixed $\epsilon \in (0, 1/2)$, Theorem 1 establishes constants $a_\epsilon, \rho_\epsilon, C_\epsilon > 0$ and $N_\epsilon < \infty$ such that, for $n \geq N_\epsilon$ and $d \leq \rho_\epsilon n \log n$,

$$a_\epsilon \left\{ \frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \right\} \leq \mathsf{R}_{n,d,\epsilon} \leq C_\epsilon \left\{ \frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \right\}.$$

The same result records the parametric interior $d = O(\sqrt{n} \log n)$, where the risk is of order $1/n$, and the consistency frontier $d = o(n \log n)$ along sequences inside the displayed calibrated range.

The estimator attaining the upper bound is explicit. It splits the sample into a pilot half and an estimation half. Categories with large pilot counts enter a ratio branch, using empirical treated and untreated outcome proportions. For $n \geq N_0$, categories with small pilot counts enter a polynomial branch. The polynomial is a Chebyshev-based approximation to the reciprocal terms in the categorywise treatment-control contrast, and its monomials are estimated through normalized falling-factorial moments of the four treatment-outcome cell counts. Theorem 2 records that this hybrid is computable with operation count linear in d and polynomial in the logarithmic degree $M(n)$. The same numerical tuning is used across fixed interior overlap levels.

The rate closes the logarithmic upper-bound gap left by the closest discrete-covariate causal benchmark. Zeng et al. [2024] show that standard plug-in, inverse-probability-weighted, and doubly robust estimators have worst-case upper scale $d^2/n^2 + 1/n$ in a strong-overlap finite-alphabet model, while their lower bound has scale $d^2/\{n^2(\log n)^2\} + 1/n$. For each fixed $\epsilon \in (0, 1/2)$, the hybrid construction developed here attains the lower-bound scale for $n \geq N_\epsilon$, $d > 0$, and $d \leq \rho_\epsilon n \log n$. The lower side is transferred from their fixed-sample control-zero construction, and the upper side is supplied by the ratio-polynomial estimator.

The estimator reflects the same approximation principle that appears in large-alphabet nonsmooth functional estimation [Paninski, 2003, Jiao et al., 2015, Wu and Yang, 2016, Han et al., 2020]. Sparse categories have too little information for stable cellwise ratio estimation. A degree of order $\log n$ balances approximation bias against factorial-moment variance, producing the logarithmic improvement in the large-alphabet term. Heavy categories are handled separately because their denominators are large enough for direct ratio estimation to achieve the required aggregate rate.

The paper also characterizes the randomized endpoint behavior. Theorem 2 gives a centered estimator with risk at most

$$\frac{1}{n} + 4(1/2 - \epsilon)^2$$

for every n, d and every $\epsilon \in (0, 1/2]$. At $\epsilon = 1/2$, treatment is randomized within each positive-mass category, and the minimax risk lies between $1/(100n)$ and $1/n$. Combining the centered estimator with the hybrid estimator yields a deterministic upper envelope involving

$$\frac{1}{n} + \min\left\{\frac{d^2}{n^2(\log n)^2}, (1/2-\epsilon)^2\right\}$$

within the fixed-interior calibration range.

The appendix records the mathematical statements and proof details used in the paper, including the product-law bridge in Lemma 20. The paper proceeds as follows. The next section reviews related work. The setup section defines the finite-alphabet experiment class, overlap cone, ATE functional, minimax risk, and hybrid estimator. The main-results section states the fixed-interior minimax theorem and the overlap-adaptive envelope. The discussion interprets the rate, the estimator, and the endpoint comparison. The appendices give the auxiliary probability, approximation, factorial-moment, heavy-cell, light-cell, and main-theorem proofs.

2 Related work

The paper sits at the intersection of causal inference for average treatment effects, high-dimensional observational methods, minimax theory for nonsmooth functionals, and large-alphabet discrete-distribution estimation. In the potential-outcomes tradition of Rubin [1974] and Rosenbaum and Rubin [1983], identification and estimation of average treatment effects are organized around consistency, conditional exchangeability, and overlap. Semiparametric analyses then characterize efficient influence functions and regular estimators under smooth low-dimensional structure [Hahn, 1998, Hirano et al., 2003, Robins et al., 1994]. Doubly robust and targeted procedures extend this program by combining outcome regression and propensity-score estimation in ways that preserve first-order behavior under appropriate nuisance control [Bang and Robins, 2005, van der Laan and Rubin, 2006]. The present analysis uses that causal estimand and overlap vocabulary, but its rate question is driven by unrestricted finite-alphabet confounding: the covariate alphabet may grow with the sample size, and the categorywise nuisance functions are unconstrained across cells.

High-dimensional causal inference has developed complementary tools for flexible nuisance estimation and valid inference under structural restrictions. Orthogonal and debiased methods use moment equations that reduce sensitivity to first-stage regularization error [Belloni et al., 2017, Chernozhukov et al., 2018, 2022]; related analyses clarify how overlap, support, and design complexity shape attainable performance in observational problems [D'Amour et al., 2021, Jin and Syrgkanis, 2025]. Work on higher-order influence functions and nonparametric bias correction shows that estimating causal functionals can require corrections beyond first-order semiparametric theory when nuisance features are complex [Robins et al., 2008, 2009, 2017, Liu et al., 2017, Kennedy et al., 2024]. The fixed-overlap discrete model considered here gives a complementary benchmark with combinatorial complexity: each category carries its own treatment probability and conditional outcome means, so the sharp rate depends on the number of categories.

The minimax perspective follows the classical decision-theoretic treatment of statistical rates [Bickel et al., 1993, van der Vaart, 1998, Tsybakov, 2009]. For nonsmooth functionals, sharp procedures often combine approximation theory with bias correction, and logarithmic factors can mark the difference between plug-in behavior and the optimal risk [Lepski et al., 1999, Cai and Low, 2011]. Classical unseen-species work provides an early large-alphabet precedent [Fisher et al., 1943, Good and Toulmin, 1956, Efron and Thisted, 1976]. Modern large-alphabet functional estimation uses polynomial approximation and unbiased estimation of polynomial moments to attain sharp rates for entropy, support size, and related functionals [Paninski, 2003, Valiant and Valiant, 2011, 2017, Orlitsky et al., 2016, Wu and Yang, 2019, Jiao et al., 2015, Wu and Yang, 2016, Han et al., 2015, 2020]. The estimator constructed in this paper imports that approximation-and-factorial-moment logic into the causal four-cell contribution associated with each covariate category, while ratio estimation handles categories with enough observations.

The closest comparison is Zeng et al. [2024], whose strong-overlap discrete-covariate model is the one adopted here, with the covariate alphabet size d allowed to grow with the sample size n. Three of their results fix the benchmark. First, they show that the standard estimator class collapses: when the categorywise nuisances are fitted on the same sample that is used for the final average, the outcome-regression, inverse-probability-weighted, and doubly robust estimators of the ATE coincide numerically with a single count-based plug-in rule. Second, they bound that rule's worst-case mean-squared error at the scale $d^2/n^2 + 1/n$, with an overlap-dependent constant, so $d/n \to 0$ is sufficient for consistency of every estimator in that class; their companion worst-case-bias bound shows that the same condition is also necessary there, which makes $d/n \to 0$ the consistency benchmark for standard practice in this model. Third, their minimax lower bound over the same class of experiments scales as $d^2/(n^2 \log^2 n) + 1/n$ in the regime $d \lesssim n \log n$, a factor $(\log n)^2$ below their upper bound in the large-alphabet term. In the discussion following their Theorem 2, and again in their concluding discussion, Zeng et al. [2024] state that this gap is unresolved in both directions: either a polynomial-approximation estimator improves the upper bound, or the lower bound can be tightened.

The gap matters for more than bookkeeping. Robins and Ritov [1997] showed that unrestricted high-dimensional adjustment can obstruct uniform model-free inference, so a benchmark in which the categorywise nuisances carry no smoothness, sparsity, or ordering structure is exactly where plug-in reasoning is least trustworthy, and where the attainable rate rather than the behavior of a familiar estimator has to settle the question.

The present paper resolves that gap on the upper-bound side. For each fixed $\epsilon \in (0, 1/2)$, it constructs a computable balanced ratio-polynomial hybrid estimator whose worst-case mean-squared error is of the lower-bound scale $d^2/(n^2 \log^2 n) + 1/n$ for $n \geq N_\epsilon$, positive d, and $d \leq \rho_\epsilon n \log n$. In that calibrated range the logarithmically sharpened large-alphabet term and the parametric $1/n$ component are both attained, so the minimax rate is identified, and the consistency benchmark moves from the $d/n \to 0$ of the standard estimator class to $d/(n \log n) \to 0$.

3 Setup and assumptions

Throughout the paper, C denotes a finite positive constant whose value may change from line to line, with dependence displayed when it matters. Asymptotic notation is along sequences in the sample size n and the alphabet size d, with the overlap level ϵ treated as fixed unless stated otherwise. Category indices range over $k \in \{1, \dots, d\}$, treatment arms and outcomes range over $\{0, 1\}$, and all risks are mean-squared risks under the product law generated by the observed-data distribution.

The observed-data model is finite and fully discrete. We first fix the variables and sampling law, then impose overlap, define the experiment class, and express the target functional as a sum of categorywise four-cell contributions.

Definition 1. [`synth_7`] (Observed treatment, outcome, and potential outcomes) *We write $A \in \{0,1\}$ for the observed binary treatment, with $A = 1$ denoting treated status and $A = 0$ untreated status, and $Y \in \{0,1\}$ for the observed binary outcome. The potential outcomes $Y(0)$ and $Y(1)$ are $\{0,1\}$-valued random variables on the same probability space as the observed variables, where $Y(a)$ is the outcome that would be recorded under treatment level $a \in \{0,1\}$.*

Assumption 1. [`ass:iid-sampling`] (IID observed-data model) *The observed data are $O_i = (X_i, A_i, Y_i)$, $i = 1, \ldots, n$, with $O_1, \ldots, O_n \overset{\text{iid}}{\sim} P$, where the covariate X takes values in $\{1, \ldots, d\}$, the treatment A and outcome Y are binary. For each category $k \in \{1, \ldots, d\}$ write $p_k = P(X = k)$ for the category mass, $\pi_k = P(A = 1 \mid X = k)$ for the treatment probability, and, for each treatment arm $a \in \{0,1\}$, $\mu_{ak} = E_P[Y \mid A = a, X = k]$ for the conditional mean outcome.*

Assumption 1 is the standard independent and identically distributed sampling condition for a finite-alphabet observational study [Zeng et al., 2024]. It allows the law P to assign arbitrary probabilities to the category–treatment–outcome cells, so all heterogeneity across categories is carried by the unrestricted masses, propensities, and conditional outcome means.

Assumption 2. [`ass:overlap`] (Fixed overlap) *For the overlap level ϵ, every category k with category mass $p_k > 0$ satisfies*

$$\epsilon \le \pi_k \le 1 - \epsilon.$$

Assumption 2 is the standard strong overlap condition [Zeng et al., 2024]. It fixes an interior range for each positive-mass category, ensuring that both treatment arms are represented at the population level in every category contributing to the target.

Definition 2. [`def:experiment-class`] (Experiment class $\mathcal{E}_{n,d,\epsilon}$) *For $n, d \in \mathbb{N}$ and $0 < \epsilon \le 1/2$, the overlap-restricted iid experiment class is*

$$\mathcal{E}_{n,d,\epsilon} := \left\{ P^{\otimes n} : P \in \mathsf{Prob}(\{1, \ldots, d\} \times \{0,1\}^2) \text{ and } P \text{ satisfies Assumption 2} \right\}.$$

The law $P^{\otimes n}$ is formed from the single-observation law in Assumption 1.

Definition 3. [`def:overlap-cone`] (Overlap cone $\mathcal{C}_\epsilon$) *The overlap-restricted cone of nonnegative four-cell vectors is*

$$\mathcal{C}_\epsilon := \left\{ u = (u_{00}, u_{01}, u_{10}, u_{11}) \in \mathbb{R}_+^4 : \epsilon(u_{00} + u_{01} + u_{10} + u_{11}) \le u_{10} + u_{11} \le (1 - \epsilon)(u_{00} + u_{01} + u_{10} + u_{11}) \right\}.$$

Definition 4. [`def:cell-functional`] (Cell functional $\phi(u)$) *For $u = (u_{00}, u_{01}, u_{10}, u_{11}) \in \mathcal{C}_\epsilon$ as in Definition 3, define*

$$\phi(u) := \begin{cases} (u_{00} + u_{01} + u_{10} + u_{11}) \left[\dfrac{u_{11}}{u_{10} + u_{11}} - \dfrac{u_{01}}{u_{00} + u_{01}} \right], & u \ne (0,0,0,0), \\ 0, & u = (0,0,0,0). \end{cases}$$

Definition 5. [`def:ate-functional`] (Average treatment effect $\tau(P)$) *For a probability law P on $\{1, \ldots, d\} \times \{0,1\} \times \{0,1\}$, with arbitrary masses on this finite observation alphabet, define, for $a, y \in \{0,1\}$ and $k \in \{1, \ldots, d\}$, the joint cell mass*

$$q_{aky} = P(A = a, X = k, Y = y),$$

and set

$$q_k = (q_{0k0}, q_{0k1}, q_{1k0}, q_{1k1}).$$

The average treatment effect functional is

$$\tau(P) = \sum_{k=1}^{d} \phi(q_k).$$

The experiment class in Definition 2 is the decision-theoretic object over which risks are evaluated. The cone in Definition 3 rewrites fixed overlap directly in terms of the four joint treatment–outcome masses within a category, and Definitions 4 and 5 turn those cell masses into the categorywise contribution and the population ATE functional. This representation is useful because the analysis can treat the target as a sum of homogeneous finite-alphabet cell functionals.

Definition 6. [`def:minimax-risk`] (Minimax risk $\mathsf{R}_{n,d,\epsilon}$)

$$\mathsf{R}_{n,d,\epsilon} := \inf_{\widehat{\tau}} \sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} E_P\left[(\widehat{\tau} - \tau(P))^2\right].$$

The infimum ranges over all measurable estimators $\widehat{\tau}$ *based on* $O_1, \ldots, O_n$*.*

Definition 6 measures the worst-case mean-squared error over the fixed-overlap experiment class. The main rate theorem below evaluates this risk and compares it with the performance of a concrete estimator.

The extrema in the minimax risk. Throughout the paper the two extrema in Definition 6 are the ordinary ones: the inner supremum is the supremum of a nonempty family of nonnegative reals that is bounded above, and the outer infimum is taken over a family of nonnegative reals bounded below by 0. Three facts supply this. Whenever $n > 0$, $d > 0$, and $0 < \epsilon \leq 1/2$, the class $\mathcal{E}_{n,d,\epsilon}$ of Definition 2 is nonempty, since the law with $p_k = 1/d$ and $\pi_k = 1/2$ for every category satisfies Assumption 2. The sample space $(\{1, \ldots, d\} \times \{0,1\} \times \{0,1\})^n$ is finite, so an estimator takes finitely many values and is a bounded function of the sample. And $|\tau(P)| \leq 1$ for every law P, since Definitions 4 and 5 express $\tau(P) = \sum_k p_k(\mu_{1k} - \mu_{0k})$ as a convex combination of differences of conditional means in $[0,1]$. Together they bound the squared losses $(\widehat{\tau} - \tau(P))^2$ uniformly over the class, so $\mathsf{R}_{n,d,\epsilon}$ is a well-defined number in $[0,1]$ and no extended-real value arises.

Readers comparing the text with the machine-checked development will find one bookkeeping difference. That development makes $\mathsf{R}_{n,d,\epsilon}$ a real number by fiat, assigning the value 0 to a supremum over an empty family or over a family with no real upper bound, and reading the quotients in the displayed rate as real division with $x/0 = 0$. The appendix proofs therefore carry the corresponding boundary cases explicitly, even though the paragraph above shows that the supremum and infimum defining $\mathsf{R}_{n,d,\epsilon}$ are the ordinary ones throughout.

The estimator uses one deterministic split to classify categories by pilot abundance and the other split to estimate the resulting heavy and light contributions. In words, before any notation: cut the sample in half; use the first half only to sort categories into those with enough observations and those without; on the well-observed categories estimate the treatment-control contrast by the obvious empirical ratios, computed from the second half; on the sparse categories, where those ratios are unstable, replace the contrast by a polynomial of degree about $\log n$ that approximates it, and estimate that polynomial without bias from falling-factorial moments of the second-half cell counts; add the two parts and clip the total to $[-1,1]$. The definitions below record the split counts, the polynomial notation, and the calibrated hybrid estimator built from these steps, and a step-by-step evaluation recipe follows the definition.

Definition 7. [`def:sample-splits`] (Sample splits I_0, I_1 and counts $N_{aky}^{(j)}$) *For a sample $O_1,\dots,O_n$ with $O_i = (X_i, A_i, Y_i) \in \{1,\dots,d\}\times\{0,1\}\times\{0,1\}$, define the deterministic half-sample splits, their sizes, and the logarithmic scale*

$$I_0 = \{1,\dots,\lfloor n/2\rfloor\}, \qquad I_1 = \{\lfloor n/2\rfloor+1,\dots,n\}, \qquad m_j = |I_j|, \qquad L = \log(en),$$

and, for each split $j\in\{0,1\}$, treatment arm $a\in\{0,1\}$, category $k\in\{1,\dots,d\}$, and outcome $y\in\{0,1\}$, the split cell count

$$N_{aky}^{(j)} = \#\{i\in I_j : (A_i, X_i, Y_i) = (a,k,y)\}.$$

Definition 8. [`synth_3`] (Chebyshev polynomials, falling factorials, and four-cell monomials) *For $M\in\mathbb{N}$, T_M denotes the degree-M Chebyshev polynomial of the first kind, characterized by $T_M(\cos\theta) = \cos(M\theta)$ for every $\theta\in\mathbb{R}$. For a real number z and an integer $t\ge 1$, the falling factorial is $(z)_t = z(z-1)\cdots(z-t+1)$, with $(z)_0 = 1$; it is applied to the split cell counts as $(N_{aky}^{(1)})_{r_{ay}}$ and to the estimation-split size as $(m_1)_{|r|} = \prod_{\ell=0}^{|r|-1}(m_1-\ell)$, the empty product being 1. For a four-cell vector $u = (u_{00}, u_{01}, u_{10}, u_{11})$ and a multi-index $r = (r_{ay})_{a,y\in\{0,1\}} \in \mathbb{N}_0^4$ of total degree $|r| = \sum_{a,y\in\{0,1\}} r_{ay}$, the monomial u^r is $\prod_{a,y\in\{0,1\}} u_{ay}^{r_{ay}}$. For a category k and a multi-index r, the normalized factorial-moment monomial is*

$$\mathfrak{m}_{k,r} = \frac{\prod_{a,y\in\{0,1\}}(N_{aky}^{(1)})_{r_{ay}}}{(m_1)_{|r|}}.$$

Definition 9. [`def:hybrid-estimator-handle`] (Hybrid estimator $\hat{\tau}_n^{\mathrm{hyb}}$) *Using the deterministic splits I_0, I_1, split sizes m_j, logarithmic scale $L = \log(en)$, and split counts $N_{aky}^{(j)}$ from Definition 7, set*

$$A_0 = 6, \qquad \lambda_0 = 256, \qquad b_0 = 4096,$$
$$D_{A_0} = 8\log(9A_0/8) = 8\log(27/4),$$

and

$$\alpha_0 = \min\left\{1, \frac{1}{32\log A_0}, \frac{1}{256 D_{A_0}}\right\}.$$

For $n\ge 2$, define

$$M(n) = \max\{2, \lfloor \alpha_0 L\rfloor\}, \qquad B(n) = b_0 L/m_1.$$

Let N_0 be the least numerical cutoff such that, for every $n\ge N_0$,

$$\alpha_0 L \ge 2, \qquad 4M(n)^2 \le m_1, \qquad 4M(n)/m_1 \le 3B(n)/4.$$

For every n, define the pilot-certified heavy and light category sets

$$\widehat{\mathcal{H}}_n = \left\{k : \sum_{a,y} N_{aky}^{(0)} > \lambda_0 L\right\}, \qquad \widehat{\mathcal{L}}_n = \widehat{\mathcal{H}}_n^c \quad (n\ge N_0), \qquad \widehat{\mathcal{L}}_n = \varnothing \quad (n < N_0),$$

so that for $n < N_0$ every category is treated by the ratio branch below. For $n \geq N_0$, abbreviate $M = M(n)$ and $B = B(n)$. Put

$$N_k^{(1)} = \sum_{a,y} N_{aky}^{(1)}, \qquad N_{ak}^{(1)} = \sum_{y} N_{aky}^{(1)},$$

and define the heavy-category ratio contribution

$$\widehat{h}_k = \frac{N_k^{(1)}}{m_1} \left\{ \frac{N_{1k1}^{(1)}}{N_{1k}^{(1)}} \mathbf{1}\{N_{1k}^{(1)} > 0\} - \frac{N_{0k1}^{(1)}}{N_{0k}^{(1)}} \mathbf{1}\{N_{0k}^{(1)} > 0\} \right\}.$$

Define

$$H_M(x) = \frac{1 - T_M(1-2x)}{2M^2}, \qquad E_M(x) = \frac{H_M(x)}{x}, \qquad G_M(x) = \frac{1 - E_M(x)}{x},$$

where E_M and G_M are interpreted by polynomial continuation at zero. For $u = (u_{00}, u_{01}, u_{10}, u_{11})$, set

$$s_a(u) = u_{a0} + u_{a1}, \qquad s(u) = s_0(u) + s_1(u),$$

$$t_1(u) = u_{11}, \qquad t_0(u) = u_{01},$$

and define the light-cell polynomial

$$P_{M,B}(u) = \frac{s(u)t_1(u)}{B} G_M\left(\frac{s_1(u)}{B}\right) - \frac{s(u)t_0(u)}{B} G_M\left(\frac{s_0(u)}{B}\right).$$

Expand

$$P_{M,B}(u) = \sum_{|r| \leq M} c_r u^r.$$

For each category k, define

$$\widehat{P}_k = \sum_{|r| \leq M} c_r \frac{\prod_{a,y} (N_{aky}^{(1)})_{r_{ay}}}{(m_1)_{|r|}}.$$

The heavy and light sums are

$$\widehat{T}_{\mathcal{H}} = \sum_{k \in \widehat{\mathcal{H}}_n} \widehat{h}_k, \qquad \widehat{T}_{\mathcal{L}} = \sum_{k \in \widehat{\mathcal{L}}_n} \widehat{P}_k.$$

The truncated balanced ratio-polynomial hybrid estimator is

$$\widehat{\tau}_n^{\mathrm{hyb}} = \max\{-1, \min(1, \widehat{T}_{\mathcal{H}} + \widehat{T}_{\mathcal{L}})\}.$$

The same calibrated estimator is used for every overlap level.

The split in Definition 7 separates category selection from contribution estimation. The polynomial notation in Definition 8 supports unbiased estimation of polynomial cell functionals through factorial moments. The estimator in Definition 9 then combines direct ratio estimation on pilot-certified heavy categories with a Chebyshev-based polynomial contribution on pilot-certified light categories, using fixed numerical tuning.

Computing the hybrid estimator. The definition is a finite recipe in the split counts, and it evaluates in the following steps. Throughout, $L = \log(en)$, $M = M(n)$, and $B = B(n)$ are the scale, degree, and light-cell scale fixed in Definition 9; the cutoff N_0 there is the least integer meeting the three displayed inequalities, so it is pinned down by the numerical constants α_0, λ_0, b_0 alone and can be tabulated once, before any data are seen.

1. *Pilot counts and classification.* From the pilot split I_0, form the counts $N_{aky}^{(0)}$ and the category totals $N_k^{(0)} = \sum_{a,y} N_{aky}^{(0)}$. Place category k in $\widehat{\mathcal{H}}_n$ when $N_k^{(0)} > \lambda_0 L$. For $n \geq N_0$ the remaining categories form $\widehat{\mathcal{L}}_n$; for $n < N_0$ every category goes to the ratio branch.

2. *Estimation counts.* From the estimation split I_1, form $N_{aky}^{(1)}$ together with the arm totals $N_{ak}^{(1)} = \sum_y N_{aky}^{(1)}$ and the category totals $N_k^{(1)} = \sum_{a,y} N_{aky}^{(1)}$.

3. *Ratio branch.* For each $k \in \widehat{\mathcal{H}}_n$, evaluate $\widehat{h}_k$ from the displayed ratio formula; an arm with no observation contributes zero through the accompanying indicator.

4. *Polynomial coefficients.* Compute the reciprocal coefficients $g_{M,j}$, $0 \leq j \leq M-2$, of the continuation G_M, which are the coefficients characterized in Definition 13 and are read off from those of the Chebyshev polynomial T_M. They depend on n alone, so one list serves every category.

5. *Light branch.* For each $k \in \widehat{\mathcal{L}}_n$ and each arm a, accumulate the normalized factorial moments $\mathfrak{m}_{k,r}$ of Definition 8 against the weights $B^{-1} g_{M,j} B^{-j}\binom{j}{t}$ prescribed by the sparse-arm expansion of Lemma 8, and set $\widehat{P}_k = \widehat{A}_{1,k} - \widehat{A}_{0,k}$. This evaluates $\widehat{P}_k = \sum_{|r| \leq M} c_r \mathfrak{m}_{k,r}$ directly from counts, so the multivariate polynomial $P_{M,B}$ enters only through those weights.

6. *Aggregate and truncate.* Add the heavy sum $\widehat{T}_{\mathcal{H}}$ and the light sum $\widehat{T}_{\mathcal{L}}$, then clip the total to $[-1, 1]$.

Every step reads only the split counts, and the computability statement of Theorem 2 bounds the whole evaluation by $K\,d\,M(n)^4$ operations for a universal integer K, with $M(n)$ of order $\log n$.

We next connect the observed-data functional to the usual potential-outcome interpretation of the ATE. The following assumptions are stated separately from the experiment class because the minimax problem is formulated on observed laws, while causal interpretation uses a full-data overlay.

Assumption 3. [`ass:consistency`] (Observed-outcome consistency) *The observed outcome satisfies*

$$Y = Y(A)$$

almost surely.

Assumption 3 is the standard consistency condition linking the recorded outcome to the potential outcome under the realized treatment [Zeng et al., 2024]. It gives the observed binary outcome its causal interpretation once treatment is assigned.

Assumption 4. [`ass:conditional-exchangeability`] (Conditional exchangeability) *The potential outcomes satisfy*

$$(Y(0), Y(1)) \perp A \mid X.$$

Assumption 4 is the standard conditional exchangeability condition [Zeng et al., 2024]. Together with Assumptions 2 and 3, it identifies the category-adjusted contrast represented by $\tau(P)$ in Definition 5; this is the finite-alphabet version of the usual adjustment argument in observational causal inference [Rubin, 1974, Rosenbaum and Rubin, 1983].

4 Main results

The preceding section defined the overlap-restricted experiment class, the ATE functional, the minimax risk, and the balanced ratio-polynomial hybrid estimator. We now state the rate conclusions. Two auxiliary definitions fix the product-law notation used in the statements and the centered estimator used at the randomized endpoint.

Definition 10. [`synth_10`] (Centered endpoint estimator $\widehat{\tau}_{\mathrm{ctr}}$) *For $n \in \mathbb{N}$ with $n > 0$ and observed units $O_i = (X_i, A_i, Y_i) \in \{1, \dots, d\} \times \{0,1\} \times \{0,1\}$, define the centered endpoint estimator by*

$$\widehat{\tau}_{\mathrm{ctr}}(O_1, \dots, O_n) = \frac{1}{n} \sum_{i=1}^{n} 2(2A_i - 1) \left(Y_i - \frac{1}{2} \right).$$

The centered estimator in Definition 10 is a linear contrast tailored to the near-randomization regime. It supplies the benchmark used below when overlap approaches the randomized endpoint.

Definition 11. [`synth_13`] (Product sampling laws) *For a finite index set J and a probability law P for one observed unit $O = (X, A, Y)$, $P^{\otimes J}$ denotes the product law on families $(O_j)_{j \in J}$ under which the coordinates are independent and each coordinate has law P. For a sample $O_1, \dots, O_n$ with single-observation law P on $\{1, \dots, d\} \times \{0,1\} \times \{0,1\}$, we write $\mu_n = P^{\otimes n}$ for the n-fold product law of the observed sample.*

The notation in Definition 11 keeps the single-observation law and its product sample law explicit. The minimax bounds below are stated over the same product-law experiment class as in Definition 2.

Theorem 1. [`thm:sharp-minimax-fixed-interior`] (Sharp fixed-interior minimax rate)[1] *Fix an overlap level ϵ with $0 < \epsilon < 1/2$. Then there are constants $a_\epsilon, \rho_\epsilon, C_\epsilon > 0$ and an integer $N_\epsilon < \infty$ such that, writing*

$$r_{n,d} := \frac{1}{n} + \frac{d^2}{n^2 (\log n)^2},$$

the following statements hold.

- ***(Matched minimax envelope.)*** *For every n, d, every single-observation law P, and every sample law μ_n in the experiment class $\mathcal{E}_{n,d,\epsilon}$ of Definition 2, if*

$$d > 0, \qquad n \ge N_\epsilon, \qquad d \le \rho_\epsilon n \log n,$$

  then the minimax risk $\mathsf{R}_{n,d,\epsilon}$ of Definition 6 and the truncated balanced ratio-polynomial hybrid estimator $\widehat{\tau}_n^{\mathrm{hyb}}$ of Definition 9 satisfy

$$a_\epsilon r_{n,d} \le \mathsf{R}_{n,d,\epsilon} \le \sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} E_P \left[\left(\widehat{\tau}_n^{\mathrm{hyb}} - \tau(P) \right)^2 \right] \le C_\epsilon r_{n,d}.$$

- ***(Parametric interior.)*** *For every $K > 0$ there is a constant $C_K > 0$ such that, for every n, d, every single-observation law P, and every sample law μ_n in $\mathcal{E}_{n,d,\epsilon}$, if*

$$d > 0, \qquad n \geq N_\epsilon, \qquad d \leq K\sqrt{n}\log n, \qquad d \leq \rho_\epsilon n \log n,$$

then

$$\frac{a_\epsilon}{n} \leq \mathsf{R}_{n,d,\epsilon} \leq \frac{C_K}{n}.$$

- ***(Consistency threshold.)*** *For every pair of integer sequences n_j, d_j with $n_j \to \infty$, $d_j > 0$ eventually, and*

$$d_j \leq \rho_\epsilon n_j \log n_j$$

eventually, the minimax risk converges to zero exactly along the sequences satisfying the vanishing normalized dimension condition:

$$\mathsf{R}_{n_j,d_j,\epsilon} \to 0 \quad \Longleftrightarrow \quad \frac{d_j}{n_j \log n_j} \to 0.$$

Theorem 1 identifies the fixed-interior minimax mean-squared-error rate for each fixed $\epsilon \in (0, 1/2)$, positive d, $n \geq N_\epsilon$, and $d \leq \rho_\epsilon n \log n$ as

$$n^{-1} + d^2\{n^2(\log n)^2\}^{-1}$$

through matching lower and upper bounds. The lower side uses the fixed-sample discrete-covariate lower-bound construction of Zeng et al. [2024], which enters through the transfer step recorded as Lemma 2; the upper side is attained by the hybrid estimator already defined in Definition 9. The theorem also records the parametric interior $d = O(\sqrt{n}\log n)$, where the risk is of order $1/n$, and the consistency criterion $d = o(n \log n)$ along sequences satisfying the displayed fixed-interior dimension range.

One feature of the statement deserves comment. The displayed bounds concern $\mathsf{R}_{n,d,\epsilon}$ and a supremum over the whole class, so they do not depend on the particular law P and sample law μ_n that the statement quantifies over. Those two quantifiers carry the requirement that $\mathcal{E}_{n,d,\epsilon}$ be inhabited at the pair (n, d): the conclusion is asserted exactly when some overlap law witnesses the class, which is the condition that makes the worst-case risk an ordinary supremum in the sense discussed with Definition 6.

The rate has the same two-component structure as other nonsmooth large-alphabet functional problems: an ordinary sampling term and an approximation-driven large-alphabet term [Jiao et al., 2015, Wu and Yang, 2016, Han et al., 2020]. In this causal problem, the approximation term enters through the cell contribution ϕ of Definition 4; heavy categories are handled by ratios, while light categories use a polynomial approximation whose coefficients can be estimated through factorial moments. The logarithmic denominator in the second term is the gain supplied by the polynomial branch relative to purely cellwise plug-in behavior.

The next result separates implementation, fixed-overlap calibration, near-randomization control, and the randomized endpoint. Its fixed-overlap statement uses the same numerical hybrid construction for every interior overlap class, while the deterministic selected estimator combines the hybrid and centered estimators according to their displayed envelopes.

[1]**Formalization scope.** This result uses the published conclusion [Zeng et al., 2024] (Theorem 2; Appendix C.5; Lemma 4; Appendix D.2). The cited source proof is not formalized here: Lean verifies its use and all remaining steps. If that cited conclusion is false, the portions of this result that depend on it are not certified.

Theorem 2. [`thm:overlap-adaptive-universal-hybrid`] (Overlap adaptive hybrid envelope) *With $\widehat{\tau}_n^{\mathrm{hyb}}$ as in Definition 9 and $\mathsf{R}_{n,d,\epsilon}$ as in Definition 6, the following assertions hold.*

- ***Computability.*** *There is a universal integer $K > 0$ such that, for every $n, d \in \mathbb{N}$, a real-arithmetic program based on the hybrid count input has operation count at most*

$$K\, d\, M(n)^4$$

and, for every sample $O_1, \dots, O_n$, evaluates exactly to $\widehat{\tau}_n^{\mathrm{hyb}}$.

- ***Fixed-overlap hybrid calibration.*** *For every overlap level ϵ with $0 < \epsilon < 1/2$, there are constants $C_\epsilon, \rho_\epsilon > 0$ and an integer N_ϵ such that, for all $n, d \in \mathbb{N}$ with $n > 0$, $d > 0$, $n \geq N_\epsilon$, and*

$$d \leq \rho_\epsilon n \log n,$$

the hybrid estimator satisfies

$$\sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} E_P\left[\left(\widehat{\tau}_n^{\mathrm{hyb}} - \tau(P)\right)^2\right] \leq C_\epsilon \left\{\frac{1}{n} + \frac{d^2}{n^2(\log n)^2}\right\}.$$

For the deterministic estimator

$$\widehat{\tau}_{C_\epsilon,\epsilon} = \begin{cases} \widehat{\tau}_n^{\mathrm{hyb}}, & \text{if } C_\epsilon \left\{\frac{1}{n} + \frac{d^2}{n^2(\log n)^2}\right\} \leq \frac{1}{n} + 4(1/2 - \epsilon)^2, \\ \widehat{\tau}_{\mathrm{ctr}}, & \text{otherwise}, \end{cases}$$

one also has

$$\mathsf{R}_{n,d,\epsilon} \leq \sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} E_P\left[(\widehat{\tau}_{C_\epsilon,\epsilon} - \tau(P))^2\right] \leq \max\{C_\epsilon, 4\}\left[\frac{1}{n} + \min\left\{\frac{d^2}{n^2(\log n)^2}, (1/2-\epsilon)^2\right\}\right].$$

- ***Centered randomization bound.*** *For every overlap level ϵ with $0 < \epsilon \leq 1/2$ and every $n, d \in \mathbb{N}$ with $n > 0$ and $d > 0$, the centered estimator*

$$\widehat{\tau}_{\mathrm{ctr}} = \frac{1}{n}\sum_{i=1}^n 2(2A_i - 1)\left(Y_i - \frac{1}{2}\right)$$

satisfies

$$\sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} E_P\left[(\widehat{\tau}_{\mathrm{ctr}} - \tau(P))^2\right] \leq \frac{1}{n} + 4(1/2-\epsilon)^2.$$

- ***Endpoint randomization bracket.*** *For every $n, d \in \mathbb{N}$ with $n > 0$ and $d > 0$,*

$$\frac{1}{100n} \leq \mathsf{R}_{n,d,1/2} \leq \frac{1}{n}.$$

Theorem 2 gives a computable form of the estimator and an overlap-phase envelope. The operation count is polynomial in the degree $M(n)$ and linear in the alphabet size d, giving an explicit count-based procedure with the stated risk certificate. For every fixed interior overlap level, the same hybrid tuning from Definition 9 attains the sharp fixed-interior upper rate for $n \geq N_\epsilon$, $d > 0$, and $d \leq \rho_\epsilon n \log n$.

The two branches of the selector call for a word on status. The hybrid estimator $\widehat{\tau}_n^{\mathrm{hyb}}$ and the centered estimator $\widehat{\tau}_{\mathrm{ctr}}$ are each computable from the sample alone. The rule $\widehat{\tau}_{C_\epsilon,\epsilon}$ that chooses between them is not: it compares the two risk certificates, and C_ϵ is an existence constant produced by the fixed-overlap upper-bound argument rather than a tabulated tuning parameter shipped with the procedure. The selector is accordingly an oracle-calibrated device for stating one envelope that covers both regimes, and the implementable objects in this paper are the two branches themselves.

The centered estimator in Definition 10 explains the endpoint behavior. As ϵ approaches $1/2$, treatment assignment is close to randomized within each category, and the bound in Theorem 2 contracts to the $1/n$ scale. The selected estimator therefore records, in one displayed envelope, the large-alphabet fixed-interior contribution and the near-randomization contribution. The endpoint bracket gives the exact order $1/n$ at $\epsilon = 1/2$, completing the comparison between the interior fixed-overlap regime and the randomized endpoint.

5 Discussion and extensions

What the fixed-interior rate says, read as econometrics, is that discrete confounding has a price of its own. Splitting the rate into the ordinary sampling term $1/n$ and the large-alphabet term $d^2/\{n^2(\log n)^2\}$ separates the parametric contribution from the additional large-alphabet contribution in this unrestricted finite-covariate model. Because the covariate distribution here is unrestricted across categories, the fixed-interior minimax lower bound makes this price a benchmark within the stated finite-covariate experiment class and range $n \geq N_\epsilon$, $d \leq \rho_\epsilon n \log n$, complementing classical semiparametric ATE theory for regular low-dimensional models [Hahn, 1998, Hirano et al., 2003, Robins et al., 1994, Bang and Robins, 2005, van der Laan and Rubin, 2006] and higher-order analyses of nonregular causal functionals [Robins et al., 2008, 2009, 2017, Liu et al., 2017, Kennedy et al., 2024]. The statement holds for each fixed $\epsilon \in (0, 1/2)$, positive d, $n \geq N_\epsilon$, and $d \leq \rho_\epsilon n \log n$.

Within this calibrated range, the estimator in Definition 9 explains how the rate is attained constructively. Heavy categories are estimated through empirical treatment-control ratios, where the relevant denominators are large enough for the ratio behavior to be controlled. Light categories are handled through a Chebyshev-derived polynomial approximation to the cell contribution in Definition 4, with factorial moments converting the polynomial into an unbiased split-sample estimate of its population analogue. The resulting hybrid treats the singular ratio structure of ϕ as an approximation problem on sparse cells and as an ordinary ratio-estimation problem on sufficiently populated cells.

This decomposition also clarifies the logarithmic factor. In the finite-alphabet experiment class of Definition 2, the alphabet size can create many categories whose population mass is too small for direct ratio estimation to behave parametrically. Polynomial approximation spreads the information in those sparse cells across a degree of order $\log n$, and the factorial-moment construction makes that approximation estimable from counts. The main theorem therefore turns the large-alphabet causal difficulty into the same approximation-versus-variance balance that appears in nonsmooth functional estimation, while preserving the adjusted ATE target in Definition 5.

The endpoint statements in Theorem 2 connect the fixed-interior result to the randomized boundary. When $\epsilon = 1/2$, treatment is randomized within every positive-mass category, and the minimax risk is bracketed at the parametric scale $1/n$. For interior values close to the endpoint, the centered estimator has the displayed risk bound $1/n + 4(1/2 - \epsilon)^2$. The selected estimator combines this endpoint behavior with the hybrid envelope, yielding the upper scale

$$\frac{1}{n} + \min\left\{\frac{d^2}{n^2(\log n)^2}, (1/2 - \epsilon)^2\right\}$$

within the fixed-interior calibration range. Thus the same analysis gives both a sharp fixed-class rate and a deterministic upper envelope that reflects proximity to randomized assignment.

The results are also informative for high-dimensional causal inference. In many semiparametric and machine-learning approaches, structural conditions on nuisance functions, such as smoothness, sparsity, or predictive structure, shape attainable rates [Balakrishnan et al., 2023, Jin and Syrgkanis, 2025]. Here the category probabilities and outcome regressions are unrestricted, so the rate isolates the cost of discrete confounding itself. The parametric interior regime $d = O(\sqrt{n}\log n)$ and the consistency criterion $d = o(n \log n)$, both under the stated calibration range of Theorem 1, give concrete benchmarks against which additional structural restrictions can be compared.

Limitations and future work. A remaining open direction is the triangular-array lower-envelope problem in which the overlap level varies with sample size, $\epsilon = \epsilon_n$, especially near the randomized endpoint. The fixed-class theorem establishes the sharp rate for each fixed $0 < \epsilon < 1/2$ in its calibrated large-sample range, and the endpoint bracket establishes the parametric scale at $\epsilon = 1/2$. The combined upper envelope in Theorem 2 supplies a deterministic attainable bound across the fixed-interior calibration range. A matching lower envelope for sequences ϵ_n would characterize the transition between the fixed-interior sparse-cell difficulty and the randomized endpoint behavior.

Appendices

A Proofs and auxiliary lemmas

This appendix records the auxiliary statements that support the lower bound, the hybrid upper bound, and the endpoint comparison. The first group fixes count notation for arbitrary subsamples and gives the pilot event that separates categories by population mass with polynomially small error probability.

Definition 12. [`synth_1`] (Category and cell counts) *For observations $O_i = (X_i, A_i, Y_i)$ indexed by a finite set I, a category $k \in \{1, \dots, d\}$, and an arm $a \in \{0, 1\}$, the category count, the arm-specific category count, and the arm-specific outcome-one count are*

$$N_k = \sum_{i \in I} \mathbf{1}\{X_i = k\}, \qquad N_{ak} = \sum_{i \in I} \mathbf{1}\{X_i = k, A_i = a\}, \qquad N_{ak1} = \sum_{i \in I} \mathbf{1}\{X_i = k, A_i = a, Y_i = 1\}.$$

For the deterministic pilot split I_0 of Definition 7 we write $m_0 = |I_0|$ for its size and $N_k^{(0)} = \sum_{a,y \in \{0,1\}} N_{aky}^{(0)}$ for the pilot count of category k.

Lemma 1. [`lem:pilot-sandwich`] (Pilot sandwich tail bound) *Let $L = \log(en)$ and let $m_0 = |I_0|$ be the pilot-split size from Definition 7. For a threshold $t \in \mathbb{R}$, define*

$$\widehat{\mathcal{H}}(t) = \{k : N_k^{(0)} > tL\}, \qquad \widehat{\mathcal{L}}(t) = \widehat{\mathcal{H}}(t)^c,$$

and

$$B_-(t) = \frac{tL}{2m_0}, \qquad B_+(t) = \frac{2tL}{m_0}.$$

The pilot heavy/light sandwich has the following two polynomial tail guarantees.

- ***(Arbitrary polynomial exponent.)*** *For every $K > 0$ and $c > 0$, there exist constants $t_0 > 0$, $C > 0$, and N such that, for every $t \geq t_0$, every $n \geq N$, every dimension d satisfying $d \leq c\,nL$, every probability law P on $\{1, \dots, d\} \times \{0,1\}^2$, and every sample law μ_n satisfying the iid sampling condition in Assumption 1,*

$$\mu_n\Big(\widehat{\mathcal{H}}(t) \not\subseteq \{k : p_k \geq B_-(t)\} \text{ or } \widehat{\mathcal{L}}(t) \not\subseteq \{k : p_k \leq B_+(t)\}\Big) \leq Cn^{-K}.$$

- ***(Fixed threshold specialization.)*** *For every $c > 0$, there exist constants $C > 0$ and N such that, for every $n \geq N$, every dimension d satisfying $d \leq c\,nL$, every probability law P on $\{1, \dots, d\} \times \{0,1\}^2$, and every sample law μ_n satisfying the iid sampling condition in Assumption 1,*

$$\mu_n\Big(\widehat{\mathcal{H}}(256) \not\subseteq \{k : p_k \geq B_-(256)\} \text{ or } \widehat{\mathcal{L}}(256) \not\subseteq \{k : p_k \leq B_+(256)\}\Big) \leq Cn^{-4}.$$

Proof of Lemma 1. Fix n, d, P, t, and let μ_n be a sample law satisfying Assumption 1; thus $\mu_n = P^{\otimes n}$. Write $L = \log(en)$, $m_0 = |I_0| = \lfloor n/2 \rfloor$, and $p_k = P(X = k)$. For a category k, set

$$Z_{ik} = \mathbf{1}\{X_i = k\}, \qquad i \in I_0,$$

so that

$$N_k^{(0)} = \sum_{i \in I_0} Z_{ik}.$$

Under Assumption 1, the one-observation mean of this indicator is p_k, and $N_k^{(0)}$ is the Bernoulli count from the deterministic pilot split of size m_0.

The pilot heavy and light sets used in the bad event are

$$\widehat{\mathcal{H}} = \left\{k : \lfloor tL \rfloor < N_k^{(0)}\right\}, \qquad \widehat{\mathcal{L}} = \widehat{\mathcal{H}}^c.$$

Equivalently, since $N_k^{(0)}$ is integer-valued, $k \in \widehat{\mathcal{H}}$ exactly when $tL < N_k^{(0)}$. Thus

$$\mathcal{B}_t = \left\{\widehat{\mathcal{H}} \not\subseteq \{k : p_k \geq tL/(2m_0)\} \text{ or } \widehat{\mathcal{L}} \not\subseteq \{k : p_k \leq 2tL/m_0\}\right\}.$$

If the heavy-set condition fails, then for some k, $p_k < tL/(2m_0)$ and $\lfloor tL \rfloor < N_k^{(0)}$, hence $tL < N_k^{(0)}$. If the light-set condition fails, then for some k, $2tL/m_0 < p_k$ and $k \notin \widehat{\mathcal{H}}$; equivalently $\lfloor tL \rfloor \geq N_k^{(0)}$, which gives $N_k^{(0)} \leq tL$. Hence

$$\mathcal{B}_t \subseteq \bigcup_{k=1}^{d} \left\{p_k < \frac{tL}{2m_0},\ tL < N_k^{(0)}\right\} \cup \bigcup_{k=1}^{d} \left\{\frac{2tL}{m_0} < p_k,\ N_k^{(0)} \leq tL\right\}.$$

Both count tails below are Bernoulli-count tails for the ambient iid sequence $(O_i)_{i \geq 0}$, whereas $\mathcal{B}_t$ is an event of the sample law μ_n. The pilot count $N_k^{(0)}$ depends only on the first n observations, so Lemma 20 identifies the law of that segment under $P^{\otimes\infty}$ with $\mu_n = P^{\otimes n}$, and the two tails may be read off on either side.

For any $\ell \in \mathbb{R}$, the Bernoulli-count upper tail used here gives

$$m_0 p_k < \ell/2 \quad \Longrightarrow \quad \mu_n\{\ell < N_k^{(0)}\} \le \exp\{-\ell(\log 2 - 1/2)\}.$$

Applying this with $\ell = tL$ bounds the upper-count branch whenever $p_k < tL/(2m_0)$.

Similarly, the lower-tail bound gives

$$2\ell < m_0 p_k \quad \Longrightarrow \quad \mu_n\{N_k^{(0)} \le \ell\} \le \exp\{-m_0 p_k/8\} \le \exp\{-\ell/4\}.$$

Applying this with $\ell = tL$ bounds the lower-count branch whenever $2tL/m_0 < p_k$.

Now fix a category k. For the first cellwise event in the displayed union, either $p_k < tL/(2m_0)$, in which case the event is contained in $\{tL < N_k^{(0)}\}$ and the upper-tail bound applies, or the mass inequality fails, in which case the event is empty. For the second cellwise event, either $2tL/m_0 < p_k$, in which case the event is contained in $\{N_k^{(0)} \le tL\}$ and the lower-tail bound applies, or the mass inequality fails, in which case the event is empty. Taking the union bound over the d categories therefore yields, for $t \ge 0$, $n \ge 2$, and every μ_n satisfying Assumption 1,

$$\mu_n(\mathcal{B}_t) \le d\left[\exp\{-(tL)(\log 2 - 1/2)\} + \exp\{-(tL)/4\}\right].$$

Now fix $K > 0$ and $c > 0$. Since $\log 2 - 1/2 > 0$, set

$$t_0 = \max\left\{\frac{K+4}{\log 2 - 1/2},\ 4(K+4)\right\}, \qquad C = 2c, \qquad N = 2.$$

For $t \ge t_0$,

$$K + 4 \le t(\log 2 - 1/2), \qquad K + 4 \le t/4.$$

Thus, whenever $n \ge N$, $d \le c\,nL$, and μ_n satisfies Assumption 1,

$$\begin{aligned}
\mu_n(\mathcal{B}_t) &\le d\left[e^{-(tL)(\log 2 - 1/2)} + e^{-(tL)/4}\right] \\
&\le c\,nL \cdot 2e^{-(K+4)L} \\
&\le 2c\,n^2 e^{-(K+4)L} \\
&= 2c\,e^{-(K+4)} n^{-K-2} \\
&\le 2c\,n^{-K}.
\end{aligned}$$

Here $L \le n$ for $n \ge 1$, and

$$e^{-qL} = e^{-q} n^{-q}$$

has been used with $q = K + 4$, together with $e^{-(K+4)} \le 1$ and $n^{-K-2} \le n^{-K}$ for $n \ge 1$. This proves the arbitrary-polynomial assertion.

For the fixed calibration, take $K = 4$ and $t = 256$. The numerical inequalities

$$8 \le 256(\log 2 - 1/2), \qquad 8 \le 256/4$$

follow from $0.6931471803 < \log 2$ and elementary arithmetic. The same decay calculation gives, for every $c > 0$, the choices $C = 2c$ and $N = 2$, and hence

$$\mu_n(\mathcal{B}_{256}) \le C n^{-4}.$$

This is the stated fixed-calibration bound. $\square$

Definition 12 extends the split-count notation of Definition 7 to arbitrary index sets, which is useful for the heavy-cell residual and missing-arm bounds below. Lemma 1 is the probabilistic gate for the hybrid estimator: categories selected as heavy have population mass at least a lower multiple of L/m_0, while categories selected as light have population mass at most an upper multiple of L/m_0, up to a tail probability that can be made polynomially small. The fixed-threshold clause is the specialization used by the numerical tuning in Definition 9.

The next statements supply the lower-bound transfer and the randomized-endpoint comparison. The transfer invokes the fixed-sample discrete-covariate lower bound of Zeng et al. [2024] on a control-zero subclass, where the ATE functional in Definition 5 coincides with the treated-response functional used in that source.

Lemma 2. [`lem:ate-lower-bound-transfer`] (ATE lower-bound transfer)[2] *Fix an overlap level ϵ with $0 < \epsilon < 1/2$. For the minimax risk $\mathsf{R}_{n,d,\epsilon}$ in Definition 6, there exist constants $a_\epsilon, b_\epsilon > 0$ and an integer N_ϵ such that, for every $n, d \in \mathbb{N}$, if $d > 0$, $n \geq N_\epsilon$, and*

$$d \leq b_\epsilon n \log n,$$

then

$$\mathsf{R}_{n,d,\epsilon} \geq a_\epsilon \left\{ \frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \right\}.$$

Proof of Lemma 2. With the cell-mass notation of Definition 5, write

$$\mu_{ak}(P) = \frac{q_{ak1}}{q_{ak0} + q_{ak1}}$$

for every arm $a \in \{0, 1\}$ and category k, using the totalized real-ratio convention that the displayed value is 0 when $q_{ak0} + q_{ak1} = 0$. On positive arm cells this is the usual conditional mean $E_P[Y \mid A = a, X = k]$. Define the control-zero subclass by

$$\mathcal{D}_0^{(n,d)}(\epsilon) := \left\{ P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon} : \mu_{0k}(P) = 0 \text{ for every } k = 1, \ldots, d \right\}.$$

For $P^{\otimes n} \in \mathcal{D}_0^{(n,d)}(\epsilon)$, write the treated-arm functional as

$$\psi_1(P) = \sum_{k=1}^{d} p_k \mu_{1k}.$$

For this auxiliary subclass risk, the supremum over $\mathcal{D}_0^{(n,d)}(\epsilon)$ is the order-theoretic supremum over the displayed class, equal to 0 when that class is empty.

Since $0 < \epsilon < 1/2$, the fixed-sample control-zero lower-bound statement of [Zeng et al., 2024] gives constants $a_\epsilon, b_\epsilon > 0$ and $N_\epsilon \in \mathbb{N}$ such that, for every n, d with $0 < d$, $N_\epsilon \leq n$, and $d \leq b_\epsilon n \log n$,

$$a_\epsilon \left\{ \frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \right\} \leq \inf_{\widehat{\psi}} \sup_{P^{\otimes n} \in \mathcal{D}_0^{(n,d)}(\epsilon)} E_P\left[(\widehat{\psi} - \psi_1(P))^2 \right],$$

where the infimum is over measurable estimators based on $O_1, \ldots, O_n$.

[2] **Formalization scope.** This result uses the published conclusion [Zeng et al., 2024] (Theorem 2; Appendix C.5; Lemma 4; Appendix D.2). The cited source proof is not formalized here: Lean verifies its use and all remaining steps. If that cited conclusion is false, the portions of this result that depend on it are not certified.

We next identify the ATE functional on this subclass from the displayed finite-cell definitions. For a fixed category k, set

$$s_{ak} = q_{ak0} + q_{ak1}, \qquad p_k = s_{0k} + s_{1k}.$$

If $q_k = (0,0,0,0)$, then $\phi(q_k) = 0$, $p_k = 0$, and hence

$$\phi(q_k) = p_k(\mu_{1k} - \mu_{0k}).$$

If $q_k \neq (0,0,0,0)$, Definitions 4 and 5 give

$$\phi(q_k) = p_k \left\{ \frac{q_{1k1}}{s_{1k}} - \frac{q_{0k1}}{s_{0k}} \right\} = p_k(\mu_{1k} - \mu_{0k}),$$

where the second equality is the totalized definition of μ_{ak}. Summing over categories yields

$$\tau(P) = \sum_{k=1}^{d} p_k(\mu_{1k} - \mu_{0k}).$$

For $P^{\otimes n} \in \mathcal{D}_0^{(n,d)}(\epsilon)$, the defining control-zero condition gives $\mu_{0k} = 0$ for every k, so

$$\tau(P) = \sum_{k=1}^{d} p_k(\mu_{1k} - \mu_{0k}) = \sum_{k=1}^{d} p_k \mu_{1k} = \psi_1(P).$$

The class suprema used below are bounded above on inhabited classes. For any overlap law P, the same finite-cell identity gives

$$\tau(P) = \sum_{k=1}^{d} p_k(\mu_{1k} - \mu_{0k}).$$

The category masses satisfy $p_k \geq 0$ and $\sum_{k=1}^{d} p_k = 1$, and the totalized binary outcome means satisfy $0 \leq \mu_{ak} \leq 1$. Hence $|\mu_{1k} - \mu_{0k}| \leq 1$ for each k, so

$$|\tau(P)| \leq \sum_{k=1}^{d} p_k \,|\mu_{1k} - \mu_{0k}| \leq \sum_{k=1}^{d} p_k = 1.$$

For any estimator $\widehat{\tau}$, pointwise domination on the finite sample alphabet gives

$$|\widehat{\tau}(o_1, \ldots, o_n)| \leq \sum_{o_1', \ldots, o_n'} |\widehat{\tau}(o_1', \ldots, o_n')|.$$

Combining this inequality with $|\tau(P)| \leq 1$ yields

$$|\widehat{\tau}(o_1, \ldots, o_n) - \tau(P)| \leq \sum_{o_1', \ldots, o_n'} |\widehat{\tau}(o_1', \ldots, o_n')| + 1.$$

After squaring and integrating under $P^{\otimes n}$,

$$E_P\left[(\widehat{\tau} - \tau(P))^2\right] \leq \left\{ \sum_{o_1, \ldots, o_n} |\widehat{\tau}(o_1, \ldots, o_n)| + 1 \right\}^2.$$

For $P^{\otimes n} \in \mathcal{D}_0^{(n,d)}(\epsilon)$, the identity $\tau(P) = \psi_1(P)$ gives the same bound with $\psi_1(P)$ in place of $\tau(P)$. Each displayed mean squared error is also nonnegative because it is the integral of a square.

Fix a measurable estimator $\widehat{\tau}$. In the real-valued order convention used for these risks, a supremum over an empty class is 0. The displayed envelope makes the squared-loss family bounded above whenever the relevant class is inhabited, so an inhabited supremum lies above each of its indexed values. If $\mathcal{D}_0^{(n,d)}(\epsilon)$ is empty, its subclass supremum is therefore 0. The ambient supremum over $\mathcal{E}_{n,d,\epsilon}$ is also 0 when the ambient class is empty; when it is inhabited, the envelope applies and any member supplies a nonnegative mean squared error below the supremum. Hence the estimatorwise comparison holds in the empty-subclass case.

If $\mathcal{D}_0^{(n,d)}(\epsilon)$ is inhabited, take $P^{\otimes n} \in \mathcal{D}_0^{(n,d)}(\epsilon)$. By the definition of the subclass, the same product law is in $\mathcal{E}_{n,d,\epsilon}$, and the target identity gives

$$E_P\left[(\widehat{\tau} - \psi_1(P))^2\right] = E_P\left[(\widehat{\tau} - \tau(P))^2\right] \leq \sup_{Q^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} E_Q\left[(\widehat{\tau} - \tau(Q))^2\right],$$

where the upper boundedness needed for the final supremum is supplied by the preceding envelope. Taking the supremum over $P^{\otimes n} \in \mathcal{D}_0^{(n,d)}(\epsilon)$ gives, for every measurable $\widehat{\tau}$,

$$\sup_{P^{\otimes n} \in \mathcal{D}_0^{(n,d)}(\epsilon)} E_P\left[(\widehat{\tau} - \psi_1(P))^2\right] \leq \sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} E_P\left[(\widehat{\tau} - \tau(P))^2\right].$$

The same empty-class convention and nonnegativity show that every subclass worst-case risk is bounded below by 0. Taking the infimum over the common measurable estimator class therefore yields

$$\inf_{\widehat{\psi}} \sup_{P^{\otimes n} \in \mathcal{D}_0^{(n,d)}(\epsilon)} E_P\left[(\widehat{\psi} - \psi_1(P))^2\right] \leq \mathsf{R}_{n,d,\epsilon}.$$

Combining the cited lower bound with this restriction comparison yields

$$a_\epsilon \left\{ \frac{1}{n} + \frac{d^2}{n^2 (\log n)^2} \right\} \leq \mathsf{R}_{n,d,\epsilon}$$

for every n, d in the stated range. □

Lemma 3. [`lem:near-randomization-linear-upper`] (Near-randomization linear bound) *For $n \in \mathbb{N}$ with $n > 0$, $d \in \mathbb{N}$, a law P on $\{1, \ldots, d\} \times \{0,1\}^2$, and a sample law μ_n, suppose that $\mu_n \in \mathcal{E}_{n,d,\epsilon}$ in the sense of Definition 2. Thus $0 < \epsilon \leq 1/2$, $\mu_n = P^{\otimes n}$, and P satisfies overlap at level ϵ. Define*

$$\widehat{\tau}_{\text{ctr}} = \frac{1}{n} \sum_{i=1}^{n} 2(2A_i - 1)\left(Y_i - \frac{1}{2}\right).$$

Then, for $\tau(P)$ as in Definition 5,

$$\int (\widehat{\tau}_{\text{ctr}}(O_1, \ldots, O_n) - \tau(P))^2 \, d\mu_n \leq \frac{1}{n} + 4\left(\frac{1}{2} - \epsilon\right)^2.$$

Proof of Lemma 3. For one observation $z = (x, a, y)$, define

$$W(z) = 2(2a - 1)\left(y - \frac{1}{2}\right),$$

with $a, y \in \{0,1\}$. Then

$$\widehat{\tau}_{\mathrm{ctr}} = \frac{1}{n}\sum_{i=1}^{n} W(O_i), \qquad W(z)^2 = 1 \quad \text{for every } z.$$

Writing $q_{aky} = P(A = a, X = k, Y = y)$, direct summation gives

$$E_P[W(O)] = \sum_{k=1}^{d}\{q_{1k1} - q_{1k0} + q_{0k0} - q_{0k1}\}.$$

For the sums below, set

$$p_k = \sum_{a=0}^{1}\sum_{y=0}^{1} q_{aky}, \qquad \pi_k = \frac{q_{1k0} + q_{1k1}}{p_k}, \qquad \mu_{ak} = \frac{q_{ak1}}{q_{ak0} + q_{ak1}},$$

using ordinary real division, so a quotient with zero denominator is assigned the value zero. If $p_k > 0$, overlap and $\epsilon > 0$ make both arm masses positive, and the definitions give

$$q_{1k1} = \mu_{1k}\pi_k p_k, \qquad q_{1k0} = (1 - \mu_{1k})\pi_k p_k,$$

and

$$q_{0k1} = \mu_{0k}(1 - \pi_k)p_k, \qquad q_{0k0} = (1 - \mu_{0k})(1 - \pi_k)p_k.$$

Hence

$$q_{1k1} - q_{1k0} + q_{0k0} - q_{0k1} = p_k(\mu_{1k} - \mu_{0k}) + 2\left(\pi_k - \frac{1}{2}\right)p_k(\mu_{1k} + \mu_{0k} - 1).$$

If $p_k = 0$, all four joint masses in category k are zero; with the totalized definitions above, both weighted terms also vanish because they are multiplied by p_k. Thus the same identity holds with zero contribution in that category.

By Definition 5, $\tau(P) = \sum_{k=1}^{d} \phi(q_k)$, with $q_k = (q_{0k0}, q_{0k1}, q_{1k0}, q_{1k1})$. For a category with $p_k = 0$, the preceding zero-mass argument gives $q_k = 0$, so the zero-cell branch in Definition 4 gives $\phi(q_k) = 0 = p_k(\mu_{1k} - \mu_{0k})$. For a category with $p_k > 0$, overlap and $\epsilon > 0$ give positive treated and control arm masses, and the nonzero-cell branch of Definition 4 gives

$$\phi(q_k) = p_k\left\{\frac{q_{1k1}}{q_{1k0} + q_{1k1}} - \frac{q_{0k1}}{q_{0k0} + q_{0k1}}\right\} = p_k(\mu_{1k} - \mu_{0k}).$$

Summing over categories yields

$$\tau(P) = \sum_{k=1}^{d} p_k(\mu_{1k} - \mu_{0k}).$$

Combining this identity with the preceding category identity gives the exact bias identity

$$E_P[W(O)] - \tau(P) = \sum_{k=1}^{d} 2\left(\pi_k - \frac{1}{2}\right)p_k(\mu_{1k} + \mu_{0k} - 1).$$

The class assumptions give $0 < \epsilon \le 1/2$ and, on every category with $p_k > 0$, $\epsilon \le \pi_k \le 1-\epsilon$. The totalized outcome means satisfy $0 \le \mu_{ak} \le 1$ for every category and arm. Hence, for every positive-mass category,

$$\left|\pi_k - \frac{1}{2}\right| \le \frac{1}{2} - \epsilon, \qquad |\mu_{1k} + \mu_{0k} - 1| \le 1,$$

and therefore

$$\left|2\left(\pi_k - \frac{1}{2}\right) p_k(\mu_{1k} + \mu_{0k} - 1)\right| \le 2\left(\frac{1}{2} - \epsilon\right) p_k.$$

For a zero-mass category, $p_k = 0$ makes both sides of this last bound equal to zero. Applying the triangle inequality to the bias identity and using $\sum_{k=1}^{d} p_k = 1$ gives

$$|E_P[W(O)] - \tau(P)| \le \sum_{k=1}^{d} 2\left(\frac{1}{2} - \epsilon\right) p_k = 2\left(\frac{1}{2} - \epsilon\right).$$

The sample law in the class is the product law, and $n > 0$, so

$$E_{\mu_n}[\widehat{\tau}_{\mathrm{ctr}}] = E_P[W(O)].$$

Moreover, independence gives

$$\mathrm{Var}_{\mu_n}(\widehat{\tau}_{\mathrm{ctr}}) = \frac{1}{n}\mathrm{Var}_P(W(O)) \le \frac{1}{n} E_P[W(O)^2] = \frac{1}{n}.$$

Finally, the bias–variance identity gives

$$\int \left(\widehat{\tau}_{\mathrm{ctr}} - \tau(P)\right)^2 d\mu_n = \mathrm{Var}_{\mu_n}(\widehat{\tau}_{\mathrm{ctr}}) + \left(E_{\mu_n}[\widehat{\tau}_{\mathrm{ctr}}] - \tau(P)\right)^2.$$

Combining the preceding variance and bias bounds yields

$$\int \left(\widehat{\tau}_{\mathrm{ctr}} - \tau(P)\right)^2 d\mu_n \le \frac{1}{n} + \left\{2\left(\frac{1}{2} - \epsilon\right)\right\}^2 = \frac{1}{n} + 4\left(\frac{1}{2} - \epsilon\right)^2.$$

□

Lemma 4. [`lem:one-category-bernoulli-lower`] (One-category Bernoulli lower bound) *For any positive integers n and d, the minimax mean-squared-error risk $\mathsf{R}_{n,d,\epsilon}$ of Definition 6, evaluated at randomization $\epsilon = 1/2$, satisfies*

$$\frac{1}{100n} \le \mathsf{R}_{n,d,1/2}.$$

Proof of Lemma 4. Since $d > 0$, the category space is nonempty. Set

$$\delta = \frac{2/5}{\sqrt{n}}.$$

Since $n > 0$, $\sqrt{n} \ge 1$, and hence $\delta \ge 0$, $\delta \le 2/5$, and

$$\delta^2 = \frac{4}{25n}.$$

Consider two laws P_0 and P_1 on the d-category observation space. Under both laws, X is uniform on the category space and $A \mid X \sim \text{Bernoulli}(1/2)$. Under P_0, both treatment-arm outcome means equal $1/2$. Under P_1, the control outcome mean equals $1/2$ and the treated outcome mean equals $1/2+\delta$. The bounds above give $0 \le 1/2+\delta \le 1$, so these Bernoulli probabilities are valid. For each P_j, membership of the product sample law in $\mathcal{E}_{n,d,1/2}$ is obtained from $0 < 1/2$, $1/2 \le 1/2$, the defining identity of the iid product law $P_j^{\otimes n}$, and the overlap fact that every positive-mass category has treatment probability $1/2$. Thus the two product sample laws belong to $\mathcal{E}_{n,d,1/2}$.

For these two laws,

$$\tau(P_0) = 0, \qquad \tau(P_1) = \delta, \qquad |\tau(P_1) - \tau(P_0)| = \delta.$$

We next bound the total variation distance between the product laws. Write $p_{j,kay} = P_j(X = k, A = a, Y = y)$. The null law assigns positive mass to every singleton of the finite observation space, so $P_1 \ll P_0$, and hence $P_1^{\otimes n} \ll P_0^{\otimes n}$. The finite chi-squared formula gives

$$1 + \chi^2(P_1 \| P_0) = \sum_{k=1}^{d} \sum_{a,y \in \{0,1\}} \frac{p_{1,kay}^2}{p_{0,kay}}.$$

For a fixed category k, the two control cells agree with the null and contribute $1/(2d)$. The two treated cells contribute

$$\frac{\left\{d^{-1}(1/2)(1/2-\delta)\right\}^2}{d^{-1}(1/2)(1/2)} + \frac{\left\{d^{-1}(1/2)(1/2+\delta)\right\}^2}{d^{-1}(1/2)(1/2)} = \frac{1}{2d}\left(1 + \frac{\delta^2}{(1/2)(1-1/2)}\right).$$

Summing over the d categories yields

$$\chi^2(P_1 \| P_0) = \frac{(1/2)\delta^2}{(1/2)(1-1/2)} = 2\delta^2.$$

By tensorization of chi-squared divergence over the iid product law,

$$1 + \chi^2\left(P_1^{\otimes n} \| P_0^{\otimes n}\right) = \left(1 + \frac{(1/2)\delta^2}{(1/2)(1-1/2)}\right)^n.$$

The single-observation chi-squared value is nonnegative, so $(1+x)^n \le \exp(nx)$ gives

$$1 + \chi^2\left(P_1^{\otimes n} \| P_0^{\otimes n}\right) \le \exp\left(n\,\frac{(1/2)\delta^2}{(1/2)(1-1/2)}\right).$$

Using $\delta^2 = 4/(25n)$,

$$n\,\frac{(1/2)\delta^2}{(1/2)(1-1/2)} = \frac{8}{25} \le \log 2,$$

and therefore

$$\chi^2\left(P_1^{\otimes n} \| P_0^{\otimes n}\right) \le 1.$$

The chi-squared-to-total-variation comparison on the finite product space, together with symmetry of total variation, yields

$$\left\|P_0^{\otimes n} - P_1^{\otimes n}\right\|_{\mathrm{TV}} = \left\|P_1^{\otimes n} - P_0^{\otimes n}\right\|_{\mathrm{TV}} \le \frac{1}{2}\sqrt{\chi^2\left(P_1^{\otimes n} \| P_0^{\otimes n}\right)} \le \frac{1}{2}.$$

Now fix any measurable estimator $\widehat{\tau}$. With $s = \delta/2$, the separation above and the total-variation bound imply

$$\max_{j=0,1} P_j^{\otimes n}\left(|\widehat{\tau} - \tau(P_j)| \geq \frac{\delta}{2}\right) \geq \frac{1}{4}.$$

On each of the two laws,

$$\left(\frac{\delta}{2}\right)^2 P_j^{\otimes n}\left(|\widehat{\tau} - \tau(P_j)| \geq \frac{\delta}{2}\right) \leq E_{P_j^{\otimes n}}\left[(\widehat{\tau} - \tau(P_j))^2\right].$$

Therefore

$$\max_{j=0,1} E_{P_j^{\otimes n}}\left[(\widehat{\tau} - \tau(P_j))^2\right] \geq \frac{\delta^2}{16} = \frac{1}{100n}.$$

It remains to compare these two risks with the supremum in Definition 6. A point of that supremum is a single-observation law P together with the class membership assertion $P^{\otimes n} \in \mathcal{E}_{n,d,1/2}$. For any such P, the overlap part of the membership assertion gives $|\tau(P)| \leq 1$. Since the sample alphabet is finite, every sample point ω satisfies

$$|\widehat{\tau}(\omega)| \leq \sum_{\omega' \in (\{1,\ldots,d\} \times \{0,1\}^2)^n} |\widehat{\tau}(\omega')|.$$

Thus, under the product law generated by any such single-observation law P,

$$E_{P^{\otimes n}}\left[(\widehat{\tau} - \tau(P))^2\right] \leq \left(\sum_{\omega' \in (\{1,\ldots,d\} \times \{0,1\}^2)^n} |\widehat{\tau}(\omega')| + 1\right)^2.$$

This finite upper bound applies uniformly over the indexed family in the supremum of Definition 6. Since $P_0^{\otimes n}$ and $P_1^{\otimes n}$ are members of $\mathcal{E}_{n,d,1/2}$, that supremum is an upper bound for the two displayed risks. Hence the worst-case risk of the fixed estimator $\widehat{\tau}$ is at least $1/(100n)$. The class of measurable estimators is nonempty, for instance by the constant-zero estimator, so taking the infimum over estimators in Definition 6 gives

$$\frac{1}{100n} \leq \mathsf{R}_{n,d,1/2}.$$

$\square$

Lemma 5. [`lem:randomized-endpoint-minimax`] (Randomization endpoint risk) *Let $n, d \in \mathbb{N}$ satisfy $n > 0$ and $d > 0$. For the minimax MSE risk $\mathsf{R}_{n,d,\epsilon}$ defined in Definition 6, exact randomization satisfies*

$$\frac{1}{100n} \leq \mathsf{R}_{n,d,1/2} \leq \frac{1}{n}.$$

Proof of Lemma 5. Lemma 4 gives the lower bound

$$\frac{1}{100n} \leq \mathsf{R}_{n,d,1/2}.$$

For the upper bound, take the centered endpoint estimator of Definition 10,

$$\widehat{\tau}_{\text{ctr}} = \frac{1}{n}\sum_{i=1}^{n} 2(2A_i - 1)\left(Y_i - \frac{1}{2}\right).$$

It is measurable because the sample space is finite. To compare the infimum in Definition 6 with this estimator, interpret the displayed extrema there as real-valued order extrema and first check the bounded-below condition for the family of worst-case risks. Fix any measurable estimator. If $\mathcal{E}_{n,d,1/2}$ is empty, its worst-case risk is the supremum over an empty family and equals 0. If $\mathcal{E}_{n,d,1/2}$ is nonempty and the squared-loss integrals over the class are bounded above, any class member gives a nonnegative integral below the supremum, so the supremum is at least 0. If $\mathcal{E}_{n,d,1/2}$ is nonempty and those integrals are unbounded above, the same real-valued order convention evaluates the supremum at 0. Thus every worst-case risk in the infimum is bounded below by 0, and the infimum is no larger than the worst-case risk of $\widehat{\tau}_{\mathrm{ctr}}$:

$$\mathsf{R}_{n,d,1/2} \leq \sup_{\mu_n = P^{\otimes n} \in \mathcal{E}_{n,d,1/2}} \int \left(\widehat{\tau}_{\mathrm{ctr}}(O_1, \ldots, O_n) - \tau(P)\right)^2 \, d\mu_n.$$

It remains to bound this worst-case risk. If $\mathcal{E}_{n,d,1/2}$ is empty, the displayed supremum is 0, and $0 \leq 1/n$ because $n > 0$. Otherwise, for every member $\mu_n = P^{\otimes n} \in \mathcal{E}_{n,d,1/2}$, Lemma 3 applied with $\epsilon = 1/2$ gives

$$\int \left(\widehat{\tau}_{\mathrm{ctr}}(O_1, \ldots, O_n) - \tau(P)\right)^2 \, d\mu_n \leq \frac{1}{n} + 4\left(\frac{1}{2} - \frac{1}{2}\right)^2 = \frac{1}{n}.$$

Thus every term in the nonempty supremum is at most $1/n$, so

$$\sup_{\mu_n = P^{\otimes n} \in \mathcal{E}_{n,d,1/2}} \int \left(\widehat{\tau}_{\mathrm{ctr}}(O_1, \ldots, O_n) - \tau(P)\right)^2 \, d\mu_n \leq \frac{1}{n}.$$

Therefore $\mathsf{R}_{n,d,1/2} \leq 1/n$. Combining this upper bound with the lower bound proves

$$\frac{1}{100n} \leq \mathsf{R}_{n,d,1/2} \leq \frac{1}{n}.$$

□

Lemma 2 is the lower half of the matched fixed-interior rate in Theorem 1. Lemma 3 controls the centered estimator by a sampling term and a squared distance from the randomized endpoint. Combining the one-category lower bound in Lemma 4 with that centered upper bound at $\epsilon = 1/2$ yields the endpoint bracket in Lemma 5.

The following selector lemma turns the hybrid fixed-interior envelope and the centered near-randomization envelope into a deterministic method choice.

Lemma 6. [`lem:combined-upper-envelope`] (Combined envelope selector) *Fix constants $\epsilon, C_\epsilon, \rho_\epsilon \in \mathbb{R}$ and $N_\epsilon \in \mathbb{N}$. Assume:*

- ***(Interior overlap.)*** $0 < \epsilon < 1/2$.
- ***(Positive constants.)*** $C_\epsilon > 0$ *and* $\rho_\epsilon > 0$.
- ***(Hybrid bound.)*** *For every $n, d \in \mathbb{N}$, if $0 < n$, $N_\epsilon \leq n$, and $d \leq \rho_\epsilon n \log n$, then the hybrid estimator $\widehat{\tau}_n^{\mathrm{hyb}}$ of Definition 9 satisfies*

$$\sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} \mathbb{E}_{P^{\otimes n}}\left[\left(\widehat{\tau}_n^{\mathrm{hyb}} - \tau(P)\right)^2\right] \leq C_\epsilon \left\{\frac{1}{n} + \frac{d^2}{n^2(\log n)^2}\right\}.$$

Then for every $n, d \in \mathbb{N}$ with $0 < n$, $N_\epsilon \le n$, and $d \le \rho_\epsilon n \log n$, set

$$K_\epsilon = \max\{C_\epsilon, 4\}.$$

Define the C_ϵ, ϵ-dependent selected estimator $\widehat{\tau}^{\mathrm{sel}}_{C_\epsilon,\epsilon}$ to be $\widehat{\tau}^{\mathrm{hyb}}_n$ when

$$C_\epsilon \left\{ \frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \right\} \le \frac{1}{n} + 4(1/2 - \epsilon)^2$$

and to be $\widehat{\tau}_{\mathrm{ctr}}$ otherwise. This selected estimator satisfies

$$\mathsf{R}_{n,d,\epsilon} \le \sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} \mathbb{E}_{P^{\otimes n}} \left[\left(\widehat{\tau}^{\mathrm{sel}}_{C_\epsilon,\epsilon} - \tau(P) \right)^2 \right] \le K_\epsilon \left[\frac{1}{n} + \min\left\{ \frac{d^2}{n^2(\log n)^2}, (1/2 - \epsilon)^2 \right\} \right],$$

where $\mathsf{R}_{n,d,\epsilon}$ is the minimax risk of Definition 6.

Proof of Lemma 6. Fix n, d satisfying the displayed hypotheses, and set

$$K_\epsilon = \max\{C_\epsilon, 4\}.$$

The rate term used below is the real-valued expression

$$\frac{1}{n} + \frac{d^2}{n^2(\log n)^2},$$

which is the displayed quantity in the hybrid-bound hypothesis and in the selector rule.

The centered estimator satisfies the uniform bound

$$\sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} \mathbb{E}_{P^{\otimes n}} \left[(\widehat{\tau}_{\mathrm{ctr}} - \tau(P))^2 \right] \le \frac{1}{n} + 4\left(\frac{1}{2} - \epsilon\right)^2.$$

Indeed, when the experiment class is inhabited, Lemma 3 applies to each product law in the class and the supremum preserves the same upper bound. When the class is empty, the real-valued indexed supremum over that empty index type is 0, and the same bound follows from nonnegativity of the displayed right-hand side.

We compare the two branches in the deterministic definition of $\widehat{\tau}^{\mathrm{sel}}_{C_\epsilon,\epsilon}$. If

$$C_\epsilon \left(\frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \right) \le \frac{1}{n} + 4\left(\frac{1}{2} - \epsilon\right)^2,$$

then the selector returns $\widehat{\tau}^{\mathrm{hyb}}_n$, and the assumed hybrid bound gives

$$\sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} \mathbb{E}_{P^{\otimes n}} \left[(\widehat{\tau}^{\mathrm{sel}}_{C_\epsilon,\epsilon} - \tau(P))^2 \right] \le C_\epsilon \left(\frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \right) = \min\left\{ C_\epsilon \left(\frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \right), \frac{1}{n} + 4\left(\frac{1}{2} - \epsilon\right)^2 \right\}.$$

If the displayed inequality fails, then

$$\frac{1}{n} + 4\left(\frac{1}{2} - \epsilon\right)^2 \le C_\epsilon \left(\frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \right),$$

the selector returns $\widehat{\tau}_{\mathrm{ctr}}$, and the centered bound gives the same conclusion with the second branch:

$$\sup_{P^{\otimes n}\in\mathcal{E}_{n,d,\epsilon}} \mathbb{E}_{P^{\otimes n}}\left[(\widehat{\tau}^{\mathrm{sel}}_{C_\epsilon,\epsilon}-\tau(P))^2\right] \le \min\left\{C_\epsilon\left(\frac{1}{n}+\frac{d^2}{n^2(\log n)^2}\right), \frac{1}{n}+4\left(\frac{1}{2}-\epsilon\right)^2\right\}.$$

Thus, in all cases,

$$\sup_{P^{\otimes n}\in\mathcal{E}_{n,d,\epsilon}} \mathbb{E}_{P^{\otimes n}}\left[(\widehat{\tau}^{\mathrm{sel}}_{C_\epsilon,\epsilon}-\tau(P))^2\right] \le \min\left\{C_\epsilon\left(\frac{1}{n}+\frac{d^2}{n^2(\log n)^2}\right), \frac{1}{n}+4\left(\frac{1}{2}-\epsilon\right)^2\right\}.$$

The sample space is finite, so the selected estimator is measurable. To compare this estimator with the infimum in $\mathsf{R}_{n,d,\epsilon}$, first note that every measurable estimator has nonnegative worst-case mean squared error in the real-valued order used by the displayed extremum. If $\mathcal{E}_{n,d,\epsilon}$ is empty, the indexed supremum is 0. If $\mathcal{E}_{n,d,\epsilon}$ is inhabited and the family of mean squared errors over $\mathcal{E}_{n,d,\epsilon}$ is bounded above, then one nonnegative squared-loss integral lies below the supremum. If that family has no real upper bound, the real-valued supremum is the supremum of the empty set, hence 0. Hence the range over which the infimum defining $\mathsf{R}_{n,d,\epsilon}$ is taken is bounded below, and evaluating that infimum at the selected estimator yields

$$\mathsf{R}_{n,d,\epsilon} \le \sup_{P^{\otimes n}\in\mathcal{E}_{n,d,\epsilon}} \mathbb{E}_{P^{\otimes n}}\left[(\widehat{\tau}^{\mathrm{sel}}_{C_\epsilon,\epsilon}-\tau(P))^2\right].$$

It remains to simplify the deterministic envelope. Since $K_\epsilon \ge C_\epsilon$, $K_\epsilon \ge 4$, and the scalar terms $1/n$, $d^2/(n^2(\log n)^2)$, and $(1/2-\epsilon)^2$ are nonnegative,

$$C_\epsilon\left(\frac{1}{n}+\frac{d^2}{n^2(\log n)^2}\right) \le K_\epsilon\left(\frac{1}{n}+\frac{d^2}{n^2(\log n)^2}\right)$$

and

$$\frac{1}{n}+4\left(\frac{1}{2}-\epsilon\right)^2 \le K_\epsilon\left(\frac{1}{n}+\left(\frac{1}{2}-\epsilon\right)^2\right).$$

If

$$\frac{d^2}{n^2(\log n)^2} \le \left(\frac{1}{2}-\epsilon\right)^2,$$

then the first of these deterministic bounds gives

$$\min\left\{C_\epsilon\left(\frac{1}{n}+\frac{d^2}{n^2(\log n)^2}\right), \frac{1}{n}+4\left(\frac{1}{2}-\epsilon\right)^2\right\} \le K_\epsilon\left\{\frac{1}{n}+\min\left(\frac{d^2}{n^2(\log n)^2}, \left(\frac{1}{2}-\epsilon\right)^2\right)\right\}.$$

In the remaining case,

$$\left(\frac{1}{2}-\epsilon\right)^2 \le \frac{d^2}{n^2(\log n)^2},$$

the same conclusion follows from the second deterministic bound. Combining the selector bound with this deterministic comparison proves the stated upper envelope with $K_\epsilon = \max\{C_\epsilon, 4\}$. $\square$

Lemma 6 formalizes the comparison made in Theorem 2. Once the fixed-interior hybrid bound is available, the selector chooses the smaller displayed certificate between the hybrid estimator and the centered estimator, producing an upper envelope with the minimum of the large-alphabet term and the near-randomization term.

The remaining statements establish the hybrid upper bound by separating calibration, polynomial approximation, factorial-moment estimation, and heavy-cell ratio control. The first calibration result collects deterministic inequalities implied by the numerical constants in Definition 9.

Lemma 7. [`lem:calibration-consequences`] (Calibration consequences) *Let $L = \log(en)$, let $\ell = \log n$, and let $m_1 = |I_1|$ be the second split size from Definition 7. Let $M = M(n)$ and $B = B(n)$ be the polynomial degree and light-cell approximation scale from Definition 9. If $n \geq N_0$ and $L \geq 240$, then*

$$0 < \ell, \qquad \ell \leq L, \qquad L \leq 2\ell, \qquad 0 < m_1, \qquad n \leq 2m_1, \qquad M \leq L, \qquad B \leq \frac{8192L}{n}, \qquad (6^{2M})^2 L^6 \leq n.$$

Proof of Lemma 7. Because $n \geq N_0$, the calibration cutoff gives

$$\alpha_0 L \geq 2, \qquad 4M(n)^2 \leq m_1, \qquad \frac{4M(n)}{m_1} \leq \frac{3B(n)}{4}.$$

In particular $L > 0$. The identity $L = \log(en) = 1 + \log n = 1 + \ell$ is valid for this positive n. Since $L \geq 240$, we have

$$\ell = L - 1 > 0, \qquad \ell \leq L, \qquad L = 1 + \ell \leq 2\ell.$$

The definition $M(n) = \max\{2, \lfloor \alpha_0 L \rfloor\}$ gives $M \geq 2$. Hence $4M^2 \leq m_1$ implies $0 < m_1$. For the deterministic second split,

$$m_1 = n - \lfloor n/2 \rfloor,$$

and therefore $n \leq 2m_1$.

Since $\alpha_0 L \geq 2$, the maximum in the definition of $M(n)$ is attained by the floor term:

$$M = \lfloor \alpha_0 L \rfloor.$$

Using $\alpha_0 \leq 1$ and $L > 0$,

$$M \leq \alpha_0 L \leq L.$$

By the definition $B(n) = 4096L/m_1$ and the bound $n \leq 2m_1$,

$$B = \frac{4096L}{m_1} \leq \frac{8192L}{n}.$$

Finally we verify the calibrated growth inequality. Since $\alpha_0 \leq (32 \log 6)^{-1}$ and $M \leq \alpha_0 L$,

$$4M \log 6 \leq \frac{L}{8}, \qquad \text{so} \qquad 6^{4M} \leq \exp(L/8).$$

Also $L \geq 240$ gives

$$\log L = \log 12 + \log(L/12) \leq 11 + (L/12 - 1) \leq L/8,$$

and hence

$$L^6 = \exp(6 \log L) \leq \exp(3L/4).$$

Multiplying,

$$6^{4M} L^6 \leq \exp(7L/8) \leq \exp(L-1) = n.$$

Since $(6^{2M})^2 = 6^{4M}$, this is exactly

$$(6^{2M})^2 L^6 \leq n.$$

The displayed conclusions follow. $\square$

The Chebyshev branch is encoded through reciprocal coefficients and coefficient envelopes. These statements are the approximation-theoretic part of the construction, in the same spirit as polynomial methods for nonsmooth discrete-distribution functionals [Jiao et al., 2015, Wu and Yang, 2016, Han et al., 2020].

Definition 13. [`synth_19`] (Chebyshev reciprocal coefficients $g_{M,j}$) *For an integer $M \geq 2$, we say $g_{M,j}$, $0 \leq j \leq M-2$, are the Chebyshev reciprocal coefficients when they are the unique real coefficients satisfying*

$$\frac{1 - \frac{1-T_M(1-2x)}{2M^2 x}}{x} = \sum_{j=0}^{M-2} g_{M,j} x^j,$$

with the left-hand side interpreted by polynomial continuation at $x = 0$, where T_M is the degree-M Chebyshev polynomial of the first kind.

Lemma 8. [`lem:sparse-arm-expansion`] (Sparse arm expansion) *For any sample $O_1, \ldots, O_n$, with each $O_i = (X_i, A_i, Y_i) \in \{1, \ldots, d\} \times \{0,1\}^2$, and any category $k \in \{1, \ldots, d\}$, let $M = M(n)$ and $B = B(n)$ be the polynomial degree and light-cell approximation scale of Definition 9. For each arm $a \in \{0,1\}$, define*

$$\widehat{A}_{a,k} = \sum_{j=0}^{M-2} \sum_{t=0}^{j} \sum_{c \in \{0,1\}^2} B^{-1} g_{M,j} B^{-j} \binom{j}{t} \mathfrak{m}_{k,r^{a,c,j,t}},$$

where $g_{M,j}$ are the coefficients of G_M, $r^{a,c,j,t}$ is the four-cell exponent vector determined by $u^{r^{a,c,j,t}} = u_c u_{a1} u_{a0}^t u_{a1}^{j-t}$, and $\mathfrak{m}_{k,r^{a,c,j,t}}$ is the corresponding normalized factorial-moment monomial computed from the sample. Then the factorial-moment polynomial estimate for category k satisfies

$$\widehat{P}_k = \widehat{A}_{1,k} - \widehat{A}_{0,k}.$$

Proof of Lemma 8. Fix the sample and the category k, and abbreviate $M = M(n)$ and $B = B(n)$, using the split and calibration notation of Definitions 7 and 9. For a cell $c \in \{0,1\}^2$, let e_c be the four-cell unit multi-index. For each arm a, degree index j, and binomial index t, the sparse index used in the factorial lift is

$$r^{a,c,j,t} = e_c + e_{(a,1)} + t\, e_{(a,0)} + (j-t)\, e_{(a,1)}.$$

Equivalently, its associated four-cell monomial is $u_c u_{a1} u_{a0}^t u_{a1}^{j-t}$. Let

$$\gamma_{j,t} = B^{-1} g_{M,j} B^{-j} \binom{j}{t}.$$

The normalized factorial-moment monomial $\mathfrak{m}_{k,r}$ is the one in Definition 8. Thus the sparse arm contribution for arm a is precisely

$$\widehat{A}_{a,k} = \sum_{j=0}^{M-2} \sum_{t=0}^{j} \sum_{c\in\{0,1\}^2} \gamma_{j,t}\, \mathfrak{m}_{k,r^{a,c,j,t}}.$$

The sparse factorial-polynomial contribution for category k is defined as the treated sparse arm minus the control sparse arm. Substituting the preceding display for $a = 1$ and $a = 0$ gives

$$\widehat{P}_k = \widehat{A}_{1,k} - \widehat{A}_{0,k}.$$

□

Lemma 9. [lem:factorial-moment-mean] (Factorial mean identity) *Let P be a probability law for one observation $O = (X, A, Y)$ with $X \in \{1, \ldots, d\}$ and $A, Y \in \{0, 1\}$, and let $k \in \{1, \ldots, d\}$. Let $P^{\otimes\infty}$ denote the ambient infinite iid product law generated by P. Suppose that $n \geq N_0$, where N_0 is the calibration cutoff in Definition 9. For*

$$q_k = \big(P(X = k, A = a, Y = y)\big)_{a,y\in\{0,1\}},$$

the factorial-moment polynomial estimate $\widehat{P}_k$ from Definition 9, evaluated on the first n coordinates under $P^{\otimes\infty}$, satisfies

$$\int \widehat{P}_k(O_1, \ldots, O_n)\, dP^{\otimes\infty} = P_{M(n),B(n)}(q_k).$$

Proof of Lemma 9. Write

$$\mu_\infty = P^{\otimes\infty}, \qquad M = M(n), \qquad B = B(n), \qquad m_1 = |I_1|,$$

where μ_∞ is the ambient iid product law. For a cell $c = (a, y) \in \{0, 1\}^2$, write

$$q_c = q_{aky} = P(X = k, A = a, Y = y), \qquad q_k = (q_c)_{c\in\{0,1\}^2}.$$

Since $n \geq N_0$, the calibration in Definition 9 gives

$$4M^2 \leq m_1.$$

Because $M \geq 2$, this implies $M \leq m_1$.

For any multi-index $r = (r_c)_c$ with $|r| \leq m_1$, the normalized factorial monomial satisfies

$$\int \mathfrak{m}_{k,r}\, d\mu_\infty = \prod_{c\in\{0,1\}^2} q_c^{r_c}.$$

Indeed, on the estimation split I_1, define

$$L_i = \begin{cases} (A_i, Y_i), & X_i = k, \\ \star, & X_i \neq k. \end{cases}$$

The variables L_i, $i \in I_1$, are iid, and $P(L_i = c) = q_c$. Expanding $\prod_c (N^{(1)}_{kc})_{r_c}$ counts ordered injective assignments of $|r|$ requested labels to distinct indices in I_1. There are $(m_1)_{|r|}$ such assignments, each has probability $\prod_c q_c^{r_c}$, and division by $(m_1)_{|r|}$ gives the displayed identity.

For the sparse arms in Lemma 8, write

$$\widehat{A}_{a,k} = \sum_{j=0}^{M-2} \sum_{t=0}^{j} \sum_{c \in \{0,1\}^2} B^{-1} g_{M,j} B^{-j} \binom{j}{t} \mathfrak{m}_{k,r^{a,c,j,t}}.$$

That lemma gives the pointwise identity

$$\widehat{P}_k = \widehat{A}_{1,k} - \widehat{A}_{0,k}.$$

For $a \in \{0,1\}$, $j \in \{0,\dots,M-2\}$, $t \in \{0,\dots,j\}$, and $c \in \{0,1\}^2$, the sparse exponent

$$r^{a,c,j,t} = e_c + e_{a1} + t\, e_{a0} + (j-t)\, e_{a1}$$

has total degree

$$|r^{a,c,j,t}| = j + 2 \le M \le m_1.$$

Therefore the factorial-monomial identity applies in every summand of each sparse arm. Finite-sum linearity gives, for each $a \in \{0,1\}$,

$$\int \widehat{A}_{a,k}\, d\mu_\infty = \sum_{j=0}^{M-2} \sum_{t=0}^{j} \sum_{c \in \{0,1\}^2} B^{-1} g_{M,j} B^{-j} \binom{j}{t} q_k^{r^{a,c,j,t}}.$$

By the definition of $r^{a,c,j,t}$,

$$q_k^{r^{a,c,j,t}} = q_c\, q_{a1}\, q_{a0}^{t}\, q_{a1}^{j-t}.$$

Thus, for each fixed a and j,

$$\sum_{t=0}^{j} \sum_{c \in \{0,1\}^2} \binom{j}{t} q_k^{r^{a,c,j,t}} = \left(\sum_{c \in \{0,1\}^2} q_c \right) q_{a1} (q_{a0} + q_{a1})^j,$$

by the binomial theorem. Consequently, for each $a \in \{0,1\}$,

$$\int \widehat{A}_{a,k}\, d\mu_\infty = \frac{s(q_k) t_a(q_k)}{B} \sum_{j=0}^{M-2} g_{M,j} \left(\frac{s_a(q_k)}{B} \right)^j = \frac{s(q_k) t_a(q_k)}{B} G_M\left(\frac{s_a(q_k)}{B} \right),$$

where $t_1(q_k) = q_{11}$ and $t_0(q_k) = q_{01}$.

Subtracting the two arm identities and using Lemma 8,

$$\int \widehat{P}_k\, d\mu_\infty = \frac{s(q_k) t_1(q_k)}{B} G_M\left(\frac{s_1(q_k)}{B} \right) - \frac{s(q_k) t_0(q_k)}{B} G_M\left(\frac{s_0(q_k)}{B} \right) = P_{M,B}(q_k),$$

where the final equality is the definition of the light-cell polynomial in Definition 9. Substituting $M = M(n)$ and $B = B(n)$ gives the claimed identity. ☐

Lemma 10. [`lem:chebyshev-reciprocal-certificate`] (Polynomial approximation certificate) *Fix a degree $M \in \mathbb{N}$, a real scale B, the overlap level ϵ, and a four-cell vector u. Assume:*

- ***(Degree.)*** *$M \geq 1$.*
- ***(Overlap level.)*** *$\epsilon > 0$.*
- ***(Cone membership.)*** *$u \in \mathcal{C}_\epsilon$, where $\mathcal{C}_\epsilon$ is the overlap-restricted cone in Definition 3.*
- ***(Mass bound.)*** *$s(u) \leq B$.*

For the light-cell polynomial $P_{M,B}$ from Definition 9 and the homogeneous cell contribution ϕ from Definition 4,

$$|P_{M,B}(u) - \phi(u)| \leq \frac{2B}{\epsilon M^2}.$$

Proof of Lemma 10. Let $G_M(x) = \sum_{j=0}^{M-2} g_{M,j} x^j$. The shifted-Chebyshev expansion gives, for every real x,

$$T_M(1-2x) = 1 - 2M^2 x + 2M^2 x^2 G_M(x).$$

If $0 \leq x \leq 1$, then $|T_M(1-2x)| \leq 1$, and the displayed identity implies

$$x\,|1 - xG_M(x)| \leq \frac{1}{M^2}.$$

First consider $u = 0$. Then $s(u) = s_0(u) = s_1(u) = u_{01} = u_{11} = 0$, so both $P_{M,B}(u)$ and $\phi(u)$ equal 0. Since $s(u) \leq B$, this case also gives $B \geq 0$, and hence

$$|P_{M,B}(u) - \phi(u)| = 0 \leq \frac{2B}{\epsilon M^2}.$$

Now suppose $u \neq 0$. Nonnegativity of the four coordinates gives $s(u) \geq 0$, and $u \neq 0$ gives $s(u) > 0$. Since $s(u) \leq B$, we have $B > 0$. The overlap inequalities give

$$s_1(u) \geq \epsilon s(u) > 0, \qquad s_0(u) = s(u) - s_1(u) \geq \epsilon s(u) > 0.$$

Also $s_a(u) \leq s(u) \leq B$, so

$$x_a := \frac{s_a(u)}{B} \in [0,1], \qquad a \in \{0,1\}.$$

For each arm a, the preceding Chebyshev certificate gives

$$\left|1 - \frac{s_a(u)}{B} G_M\left(\frac{s_a(u)}{B}\right)\right| \leq \frac{B}{s_a(u) M^2}.$$

Using $0 \leq u_{a1} \leq s_a(u)$, this yields

$$\left|B^{-1} s(u) u_{a1} G_M\left(\frac{s_a(u)}{B}\right) - s(u)\frac{u_{a1}}{s_a(u)}\right| \leq \frac{B}{\epsilon M^2}.$$

Indeed, the left-hand side equals

$$\frac{s(u)u_{a1}}{s_a(u)} \left|1 - \frac{s_a(u)}{B} G_M\left(\frac{s_a(u)}{B}\right)\right|,$$

and the factor $s(u)/s_a(u)$ is at most $1/\epsilon$.

Finally,

$$P_{M,B}(u)-\phi(u)=\left[B^{-1}s(u)u_{11}G_M\left(\frac{s_1(u)}{B}\right)-s(u)\frac{u_{11}}{s_1(u)}\right]-\left[B^{-1}s(u)u_{01}G_M\left(\frac{s_0(u)}{B}\right)-s(u)\frac{u_{01}}{s_0(u)}\right].$$

The triangle inequality and the two arm bounds give

$$|P_{M,B}(u)-\phi(u)|\le\frac{B}{\epsilon M^2}+\frac{B}{\epsilon M^2}=\frac{2B}{\epsilon M^2}.$$

□

Lemma 8 rewrites the light-cell statistic as treated and control sparse-arm pieces. Lemma 9 then identifies the mean of the factorial-moment statistic with the polynomial target $P_{M(n),B(n)}(q_k)$. Lemma 10 supplies the light-cell approximation error on the overlap cone when the cell mass lies below the approximation scale.

Heavy categories are controlled by ratio residuals and missing-arm probabilities. The next two bounds apply to deterministic category sets, which allows them to be conditioned on the pilot split before summing over the selected heavy categories.

Lemma 11. [`lem:heavy-residual-bound`] (Fixed-set residual variance) *Let P be a probability law on the finite observed-data alphabet $\{1,\dots,d\}\times\{0,1\}^2$, let I be a finite index set, and write $m=|I|$. For observations $O_i=(X_i,A_i,Y_i)$, $i\in I$, drawn independently from P, let $H\subseteq\{1,\dots,d\}$ be deterministic and fix $a\in\{0,1\}$.*

Assume:

- ***(Overlap.)*** *The overlap level ϵ satisfies $\epsilon>0$, and for every category k with $p_k=P(X=k)>0$,*

$$\epsilon\le\pi_k\le1-\epsilon.$$

- ***(Positive mass on H.)*** *For every $k\in H$, $p_k>0$.*

With counts computed from the sample,

$$\mathcal{R}_a(H)=\sum_{k\in H}N_k\frac{\mathbf{1}\{N_{ak}>0\}}{N_{ak}}\left\{N_{ak1}-\mu_{ak}N_{ak}\right\}.$$

Then

$$E_P\left[\mathcal{R}_a(H)^2\right]\le\frac{2m}{\epsilon}\sum_{k\in H}p_k.$$

Proof of Lemma 11. For an observation array $z=(O_i)_{i\in I}$, with $O_i=(X_i,A_i,Y_i)$, write

$$N_k(z)=\sum_{i\in I}\mathbf{1}\{X_i=k\},\qquad N_{ak}(z)=\sum_{i\in I}\mathbf{1}\{X_i=k,A_i=a\},$$

and

$$N_{ak1}(z)=\sum_{i\in I}\mathbf{1}\{X_i=k,A_i=a,Y_i=1\}.$$

Let $m=|I|$, and put

$$C_{ak}(z)=N_k(z)\frac{\mathbf{1}\{N_{ak}(z)>0\}}{N_{ak}(z)},\qquad\xi_{ak,i}(z)=\mathbf{1}\{X_i=k,A_i=a\}\{Y_i-\mu_{ak}\}.$$

With the convention that the displayed ratio is zero when $N_{ak}(z) = 0$, the residual can be written as

$$\mathcal{R}_a(H) = \sum_{k\in H} C_{ak}(z) \sum_{i\in I} \xi_{ak,i}(z),$$

because $\sum_i \xi_{ak,i} = N_{ak1}(z) - \mu_{ak} N_{ak}(z)$.

Expand the square as a double sum over pairs (k,i) and (l,j). If $i \neq j$, then the factor $C_{ak}(z)C_{al}(z)\xi_{al,j}(z)$ is unchanged when only the outcome coordinate of observation i is replaced; conditioning on all other coordinates and using the centered arm-category residual gives integral zero. If $i = j$ and $k \neq l$, then the two category indicators cannot both hold for the same observation, so the product of residuals is identically zero. Hence only the diagonal terms remain:

$$\mathbb{E}_{P^{\otimes I}}\left[\mathcal{R}_a(H)^2\right] = \sum_{k\in H}\sum_{i\in I} \mathbb{E}_{P^{\otimes I}}\left[C_{ak}(z)^2 \xi_{ak,i}(z)^2\right].$$

For each fixed k and i, the multiplier $C_{ak}(z)^2$ is nonnegative and depends only on the category and treatment coordinates of the array. The one-coordinate conditional Bernoulli variance bound gives

$$\mathbb{E}_{P^{\otimes I}}\left[C_{ak}(z)^2 \xi_{ak,i}(z)^2\right] \leq \mathbb{E}_{P^{\otimes I}}\left[C_{ak}(z)^2 \mathbf{1}\{X_i = k, A_i = a\}\right].$$

Therefore, for each fixed k,

$$\begin{aligned}
\sum_{i\in I} \mathbb{E}_{P^{\otimes I}}\left[C_{ak}(z)^2 \xi_{ak,i}(z)^2\right] &\leq \mathbb{E}_{P^{\otimes I}}\left[\sum_{i\in I} C_{ak}(z)^2 \mathbf{1}\{X_i = k, A_i = a\}\right] \\
&= \mathbb{E}_{P^{\otimes I}}\left[N_k(z)^2 \frac{\mathbf{1}\{N_{ak}(z) > 0\}}{N_{ak}(z)}\right].
\end{aligned}$$

Let $\rho_{ak} = P(A = a \mid X = k)$, so $\rho_{1k} = \pi_k$ and $\rho_{0k} = 1 - \pi_k$. For $k \in H$, the positive-mass hypothesis gives $p_k > 0$. Overlap gives $\rho_{ak} \geq \epsilon$, and $\rho_{ak} \leq 1$ because it is an arm probability. Put

$$E_k = \{X = k\}, \qquad R_{ak} = \{X = k, A = a\}.$$

Then $R_{ak} \subset E_k$, and under the one-observation law

$$P(E_k) = p_k, \quad P(R_{ak}) = p_k\rho_{ak}, \quad P(E_k \setminus R_{ak}) = p_k(1 - \rho_{ak}), \quad P(E_k^c) = 1 - p_k.$$

Partition the product sample space according to

$$U = \{i : O_i \in E_k\}, \qquad V = \{i : O_i \in R_{ak}\}.$$

The nested relation forces $V \subset U$. For fixed U, V with $V \subset U$, independence gives the product probability

$$(p_k\rho_{ak})^{|V|}\{p_k(1-\rho_{ak})\}^{|U|-|V|}(1-p_k)^{m-|U|},$$

and the integrand equals

$$|U|^2 \frac{\mathbf{1}\{|V| > 0\}}{|V|}.$$

Summing over all such pairs and then collecting first by $t = |U|$ and then by $s = |V|$ gives

$$\begin{aligned}
&\mathbb{E}_{P^{\otimes I}}\left[N_k^2 \frac{\mathbf{1}\{N_{ak} > 0\}}{N_{ak}}\right] \\
&= \sum_{t=0}^{m} \binom{m}{t} p_k^t (1-p_k)^{m-t} t^2 \sum_{s=0}^{t} \binom{t}{s} \rho_{ak}^s (1-\rho_{ak})^{t-s} \frac{\mathbf{1}\{s > 0\}}{s}.
\end{aligned}$$

It remains to bound the inner reciprocal-count sum. For every integer $s \ge 0$, with the left side interpreted as zero at $s = 0$,

$$\frac{\mathbf{1}\{s > 0\}}{s} \le \frac{2}{s+1}.$$

Since $0 < \rho_{ak} \le 1$, the binomial weights are nonnegative, and therefore

$$\begin{aligned}
&\sum_{s=0}^{t} \binom{t}{s} \rho_{ak}^{s}(1-\rho_{ak})^{t-s} \frac{\mathbf{1}\{s>0\}}{s} \\
&\quad \le 2 \sum_{s=0}^{t} \binom{t}{s} \rho_{ak}^{s}(1-\rho_{ak})^{t-s} \frac{1}{s+1} \\
&\quad = \frac{2}{(t+1)\rho_{ak}} \sum_{s=0}^{t} \binom{t+1}{s+1} \rho_{ak}^{s+1}(1-\rho_{ak})^{t-s} \\
&\quad = \frac{2\{1-(1-\rho_{ak})^{t+1}\}}{(t+1)\rho_{ak}} \le \frac{2}{(t+1)\rho_{ak}}.
\end{aligned}$$

Here the middle equality uses $\binom{t}{s}/(s+1) = \binom{t+1}{s+1}/(t+1)$, and the final inequality uses $(1 - \rho_{ak})^{t+1} \ge 0$. Hence, using $t^2/(t+1) \le t$ and the binomial first moment,

$$\begin{aligned}
\mathbb{E}_{P\otimes I}\left[N_k^2 \frac{\mathbf{1}\{N_{ak}>0\}}{N_{ak}}\right] &\le \sum_{t=0}^{m} \binom{m}{t} p_k^t (1-p_k)^{m-t} \frac{2t}{\rho_{ak}} \\
&= \frac{2mp_k}{\rho_{ak}} \le \frac{2mp_k}{\epsilon}.
\end{aligned}$$

Summing over $k \in H$ gives

$$\mathbb{E}_{P\otimes I}\left[\mathcal{R}_a(H)^2\right] \le \sum_{k\in H} \frac{2mp_k}{\epsilon} = \frac{2m}{\epsilon} \sum_{k\in H} p_k.$$

□

Lemma 12. [`lem:missing-arm-envelope`] (Missing-arm second moment) *Let J be a finite sample-index set and write $m = |J|$. For a probability law P on $\{1,\ldots,d\} \times \{0,1\}^2$, a deterministic category set $H \subseteq \{1,\ldots,d\}$, and an arm $a \in \{0,1\}$, define*

$$Z_{ak} = N_k \mathbf{1}\{N_{ak} = 0\},$$

where N_k is the number of observations in category k and N_{ak} is the number of observations in category k and arm a in the J-indexed iid sample from P.

Assume:

- ***(Sample size.)*** *$m \ge 3$.*
- ***(Overlap.)*** *The overlap level ϵ satisfies $\epsilon > 0$, and for every category k with $p_k = P(X = k) > 0$,*
$$\epsilon \le P(A = 1 \mid X = k) \le 1 - \epsilon.$$
- ***(Mass floor.)*** *$B > 0$, and $p_k \ge B$ for every $k \in H$.*

Then

$$\mathbb{E}_{P\otimes J}\left[\left(\sum_{k\in H} Z_{ak}\right)^2\right] \leq m + m^2\left[\frac{|H|}{\{((m-2)/2)\,\epsilon\}^2 B}\right]^2.$$

Proof of Lemma 12. For a sample $z = (O_1, \ldots, O_m)$, let $Z_{ak}(z)$ be the missing-arm category count

$$Z_{ak}(z) = \begin{cases} \#\{i : X_i = k\}, & \#\{i : X_i = k,\ A_i = a\} = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Thus $\sum_{k\in H} Z_{ak}(z)$ is the fixed missing-arm count over H for arm a. Let

$$\rho_{ak} = P(A = a \mid X = k), \qquad u = \frac{m-2}{2}\epsilon, \qquad f_k = p_k \exp(-up_k).$$

The assumptions $m \geq 3$, $\epsilon > 0$, and $B > 0$ give $u > 0$, and the mass floor gives $p_k \geq B > 0$ for every $k \in H$. Hence ρ_{ak} is the ordinary conditional arm probability on H, with $0 \leq \rho_{ak} \leq 1$, and overlap gives $\epsilon \leq \rho_{ak}$, equivalently $\epsilon p_k \leq p_k\rho_{ak}$.

We first record the exact second moment. For a fixed category k, expand Z_{ak}^2 by separating one selected index from two ordered distinct selected indices. The one-index event requires O_i to lie in category k outside arm a, while the whole sample avoids the arm-a subcell in category k. Its probability is

$$p_k(1-\rho_{ak})(1-p_k\rho_{ak})^{m-1}.$$

For an ordered pair $i \neq j$, the corresponding two-index event has probability

$$\{p_k(1-\rho_{ak})\}^2(1-p_k\rho_{ak})^{m-2}.$$

Summing over the m one-index choices and the $(m)_2 = m(m-1)$ ordered pairs gives the diagonal contribution

$$m\,p_k(1-\rho_{ak})(1-p_k\rho_{ak})^{m-1} + (m)_2\{p_k(1-\rho_{ak})\}^2(1-p_k\rho_{ak})^{m-2}.$$

For distinct k, l, the same-index part in $Z_{ak}Z_{al}$ is zero because the category events are disjoint. The remaining ordered-pair event places one selected observation in category k outside arm a, the other in category l outside arm a, and the whole sample outside the two forbidden arm-a subcells. Since those forbidden subcells are disjoint, its probability is

$$p_k(1-\rho_{ak})p_l(1-\rho_{al})(1-p_k\rho_{ak}-p_l\rho_{al})^{m-2}.$$

Therefore

$$\begin{aligned}\int\left(\sum_{k\in H} Z_{ak}(z)\right)^2 dP^{\otimes m}(z) = \sum_{k\in H}\Big[&m\,p_k(1-\rho_{ak})(1-p_k\rho_{ak})^{m-1} \\ &+ (m)_2\{p_k(1-\rho_{ak})\}^2(1-p_k\rho_{ak})^{m-2}\Big] \\ + \sum_{k\in H}\sum_{l\in H\setminus\{k\}} &(m)_2 p_k(1-\rho_{ak})p_l(1-\rho_{al})(1-p_k\rho_{ak}-p_l\rho_{al})^{m-2}.\end{aligned}$$

For the diagonal terms, $0 \leq p_k \leq 1$, $0 \leq \rho_{ak} \leq 1$, $(m)_2 \leq m^2$, and $\epsilon p_k \leq p_k\rho_{ak}$ imply

$$\begin{aligned}&m\,p_k(1-\rho_{ak})(1-p_k\rho_{ak})^{m-1} + (m)_2\{p_k(1-\rho_{ak})\}^2(1-p_k\rho_{ak})^{m-2} \\ &\quad\leq mp_k + m^2 f_k^2.\end{aligned}$$

Indeed, the first summand is at most mp_k. For the second summand,

$$(1-p_k\rho_{ak})^{m-2} \le \exp\{-(m-2)\epsilon p_k\} = \exp(-up_k)^2,$$

and $p_k(1-\rho_{ak}) \le p_k$.

For $k \neq l$, disjointness of the two category events gives $p_k+p_l \le 1$, and hence $p_k\rho_{ak}+p_l\rho_{al} \le 1$. Together with the two overlap inequalities,

$$\begin{aligned}(1-p_k\rho_{ak}-p_l\rho_{al})^{m-2} &\le \exp\{-(m-2)(p_k\rho_{ak}+p_l\rho_{al})\} \\ &\le \exp\{-(m-2)\epsilon(p_k+p_l)\} \\ &\le \exp(-up_k)\exp(-up_l).\end{aligned}$$

Using also $p_k(1-\rho_{ak}) \le p_k$, $p_l(1-\rho_{al}) \le p_l$, and $(m)_2 \le m^2$, we obtain

$$(m)_2 p_k(1-\rho_{ak})p_l(1-\rho_{al})(1-p_k\rho_{ak}-p_l\rho_{al})^{m-2} \le m^2 f_k f_l.$$

Substituting the diagonal and cross envelopes into the exact expansion and collecting the square of the exponential sum gives

$$\int \left(\sum_{k\in H} Z_{ak}(z)\right)^2 dP^{\otimes m}(z) \le m\sum_{k\in H} p_k + m^2\left(\sum_{k\in H} f_k\right)^2.$$

The category masses over H sum to at most one, so the first term is at most m.

It remains to bound the exponential sum. For $t>0$,

$$\exp(-t) \le t^{-2}.$$

Applying this with $t = up_k$, and using $p_k \ge B$, gives

$$p_k\exp(-up_k) \le \frac{1}{u^2 p_k} \le \frac{1}{u^2 B}.$$

Therefore

$$\sum_{k\in H} f_k \le \frac{|H|}{u^2 B}.$$

Since each $f_k \ge 0$, combining the last two bounds and substituting $u = ((m-2)/2)\epsilon$ yields

$$\int \left(\sum_{k\in H} Z_{ak}(z)\right)^2 dP^{\otimes m}(z) \le m + m^2\left[\frac{|H|}{\left(\left(\frac{m-2}{2}\epsilon\right)^2 B\right)}\right]^2.$$

□

Lemma 11 gives the variance scale for centered ratio noise over a fixed heavy set. Lemma 12 controls the contribution from categories whose empirical denominator in an arm is zero; the mass floor supplied by the pilot sandwich makes this event sufficiently rare in aggregate.

For light categories, the factorial-moment analysis uses product-moment and coefficient-sum controls. These statements quantify how multinomial dependence across cells and categories enters the variance calculation.

Lemma 13. [lem:factorial-product-moment] (Factorial product moment) *Fix $n, d \in \mathbb{N}$, $P \in \mathsf{Prob}(\{1,\dots,d\} \times \{0,1\}^2)$, a category $k \in \{1,\dots,d\}$, and four-cell multi-indices r, s. Using the split counts and split size in Definition 7, let*

$$\mathfrak{m}_{k,r} = \frac{\prod_{(a,y)} (N^{(1)}_{aky})_{r_{ay}}}{(m_1)_{|r|}},$$

where $(z)_t$ denotes the falling factorial and $|r| = \sum_{(a,y)} r_{ay}$. Let q_{aky} denote the treatment–outcome cell mass in category k, as in Definition 5. Assume:

- ***(Nonempty estimation split.)*** $m_1 > 0$.
- ***(Degree ordering.)*** $|r| \le |s|$.
- ***(Split-size condition.)*** $4|s|^2 \le m_1$.

Then, under the iid product law generated by P,

$$\mathbb{E}_{P^{\otimes n}}[\mathfrak{m}_{k,r}\mathfrak{m}_{k,s}] \le \mathrm{e} \prod_{(a,y)} q_{aky}^{r_{ay}} \prod_{(a,y)} \left(q_{aky} + \frac{|s|}{m_1}\right)^{s_{ay}}.$$

Proof of Lemma 13. Let $\mathcal{C} = \{0,1\} \times \{0,1\}$, let $m = m_1$, and write

$$q_{ay} = q_{aky}, \qquad r_{ay} = r(a,y), \qquad s_{ay} = s(a,y), \qquad |r| = \sum_{(a,y)\in\mathcal{C}} r_{ay}, \quad |s| = \sum_{(a,y)\in\mathcal{C}} s_{ay}.$$

Adjoin one extra label $\varnothing$ for observations outside category k. For the j-th observation in the estimation split define

$$Z_j = \begin{cases} (a,y), & (X_j, A_j, Y_j) = (k,a,y), \\ \varnothing, & X_j \ne k, \end{cases} \qquad \alpha_{ay} = q_{ay}, \quad \alpha_\varnothing = 1 - \sum_{(a,y)\in\mathcal{C}} q_{ay}.$$

The variables $Z_1, \dots, Z_m$ are iid on $\mathcal{C} \cup \{\varnothing\}$ with masses α. Extend r, s by setting $r_\varnothing = s_\varnothing = 0$. If $N_b = \#\{j : Z_j = b\}$, then

$$\mathfrak{m}_{k,r} = \frac{\prod_{b\in\mathcal{C}\cup\{\varnothing\}} (N_b)_{r_b}}{(m)_{|r|}}, \qquad \mathfrak{m}_{k,s} = \frac{\prod_{b\in\mathcal{C}\cup\{\varnothing\}} (N_b)_{s_b}}{(m)_{|s|}}.$$

This is just the split-count notation of Definition 7, with the outside-category exponent equal to zero.

For an overlap choice

$$H = (H_b)_{b\in\mathcal{C}\cup\{\varnothing\}}, \qquad 0 \le H_b \le \min\{r_b, s_b\},$$

set

$$|H| = \sum_b H_b, \qquad t_b(H) = r_b + s_b - H_b, \qquad A_H = \prod_b \binom{r_b}{H_b}\binom{s_b}{H_b} H_b!.$$

The cellwise falling-factorial product identity gives

$$\prod_b (N_b)_{r_b} \prod_b (N_b)_{s_b} = \sum_H A_H \prod_b (N_b)_{t_b(H)}.$$

Taking expectations and using the factorial moment formula for multinomial counts,

$$\mathbb{E}[\mathfrak{m}_{k,r}\mathfrak{m}_{k,s}] = \sum_H A_H \frac{(m)_{|r|+|s|-|H|}}{(m)_{|r|}(m)_{|s|}} \prod_b \alpha_b^{r_b+s_b-H_b}.$$

The normalization ratio satisfies, for every such H,

$$\frac{(m)_{|r|+|s|-|H|}}{(m)_{|r|}(m)_{|s|}} \le \frac{\mathrm{e}}{m^{|H|}}.$$

Indeed, if $|s|=0$ then also $|r|=0$, and the bound is immediate. Otherwise put $u = m-|s|$. The assumptions $4|s|^2 \le m$ and $|r| \le |s|$ imply $u>0$ and $2|s|^2 \le u$. Also

$$(m)_{|r|+|s|-|H|} \le m^{|r|+|s|-|H|}, \qquad (m)_{|r|} \ge u^{|r|}, \qquad (m)_{|s|} \ge u^{|s|}.$$

Hence the ratio is at most

$$\frac{m^{|r|+|s|-|H|}}{u^{|r|+|s|}} = \frac{(m/u)^{|r|+|s|}}{m^{|H|}}.$$

Since $m/u = 1+|s|/u$,

$$(|r|+|s|)\log(m/u) \le (|r|+|s|)\frac{|s|}{u} \le 1,$$

and therefore $(m/u)^{|r|+|s|} \le \mathrm{e}$.

Because all masses α_b are nonnegative,

$$\mathbb{E}[\mathfrak{m}_{k,r}\mathfrak{m}_{k,s}] \le \mathrm{e} \sum_H A_H\, m^{-|H|} \prod_b \alpha_b^{r_b+s_b-H_b}.$$

The remaining overlap sum factors by coordinates:

$$\sum_H A_H\, m^{-|H|} \prod_b \alpha_b^{r_b+s_b-H_b} = \prod_b \sum_{h=0}^{\min(r_b,s_b)} \binom{r_b}{h}\binom{s_b}{h} h!\, m^{-h} \alpha_b^{r_b+s_b-h}.$$

For each b and each $0 \le h \le \min\{r_b, s_b\}$, one has $h \le s_b$ and

$$\binom{r_b}{h} h! = (r_b)_h \le r_b^h \le |r|^h \le |s|^h.$$

The assumption $m_1 > 0$ gives $m>0$, and $\alpha_b \ge 0$. After rewriting $\alpha_b^{r_b+s_b-h} = \alpha_b^{r_b}\alpha_b^{s_b-h}$, the preceding coefficient bound gives

$$\binom{r_b}{h}\binom{s_b}{h} h!\, m^{-h}\alpha_b^{r_b+s_b-h} \le \alpha_b^{r_b}\binom{s_b}{h}\left(\frac{|s|}{m}\right)^h \alpha_b^{s_b-h}.$$

The right-hand summands are nonnegative for every $0 \le h \le s_b$, so enlarging the summation range and applying the binomial theorem yields

$$\sum_{h=0}^{\min(r_b,s_b)} \binom{r_b}{h}\binom{s_b}{h} h!\, m^{-h}\alpha_b^{r_b+s_b-h} \le \alpha_b^{r_b} \sum_{h=0}^{s_b} \binom{s_b}{h}\left(\frac{|s|}{m}\right)^h \alpha_b^{s_b-h} = \alpha_b^{r_b}\left(\alpha_b + \frac{|s|}{m}\right)^{s_b}.$$

Multiplying these coordinatewise bounds gives

$$\mathbb{E}[\mathfrak{m}_{k,r}\mathfrak{m}_{k,s}] \le \mathrm{e} \prod_b \alpha_b^{r_b} \prod_b \left(\alpha_b + \frac{|s|}{m}\right)^{s_b}.$$

The $\varnothing$-coordinate contributes 1, since $r_\varnothing = s_\varnothing = 0$. Replacing b by $(a, y) \in \mathcal{C}$ and m by m_1 yields

$$\mathbb{E}_{P^{\otimes n}}[\mathfrak{m}_{k,r}\mathfrak{m}_{k,s}] \le \mathrm{e} \prod_{(a,y)} q_{aky}^{r_{ay}} \prod_{(a,y)} \left(q_{aky} + \frac{|s|}{m_1}\right)^{s_{ay}}.$$

□

Lemma 14. [`lem:coefficient-sum-certificate`] (Coefficient sum certificate) *Let M be a positive integer and let $z \ge 0$. With $g_{M,j}$ denoting the coefficients of the polynomial continuation G_M from Definition 9, define*

$$g_M^+(z) = \sum_{j \in \{0,\dots,M-2\}} |g_{M,j}| z^j,$$

with the sum interpreted as empty when $M = 1$. Then

$$g_M^+(z) \le 6^M \max\{1, z^{M \dot{-} 2}\},$$

where $M \dot{-} 2 = \max\{M - 2, 0\}$.

Proof of Lemma 14. All powers below have nonnegative integer exponents; in particular, the exponent $M - 2$ in the statement is interpreted as natural-number subtraction, so it equals 0 when $M = 1$. Let T_M be the Chebyshev polynomial and let G_M be the polynomial continuation used in Definition 9. The continuation has coefficients $g_{M,j}$ satisfying

$$T_M(1 - 2x) = 1 - 2M^2 x + 2M^2 x^2 G_M(x).$$

For these coefficients, the closed form is

$$g_{M,j} = (-1)^j \frac{2^{2j+3}}{M(M + j + 2)} \binom{M + j + 2}{2j + 4} \qquad (j \ge 0).$$

Because $M > 0$, the factor after $(-1)^j$ is nonnegative. Hence

$$|g_{M,j}| = \frac{2^{2j+3}}{M(M + j + 2)} \binom{M + j + 2}{2j + 4} = (-1)^j g_{M,j},$$

and therefore, for every z,

$$g_M^+(z) = \sum_{0 \le j < M-1} |g_{M,j}| z^j = \sum_{0 \le j < M-1} g_{M,j} (-z)^j = G_M(-z).$$

Taking $z = 1$ in this identity and substituting $x = -1$ in the Chebyshev identity gives

$$T_M(3) = 1 + 2M^2 + 2M^2 g_M^+(1).$$

The sum defining $g_M^+(1)$ is nonnegative. Since $M \geq 1$, we have $M^2 \geq 1$, and the displayed identity implies

$$g_M^+(1) \leq T_M(3).$$

The Chebyshev recurrence and the standard positivity bound $T_m(x) \geq 1$ for $x \geq 1$ give, at $x = 3$,

$$T_0(3) = 1, \qquad T_1(3) = 3, \qquad T_{m+2}(3) = 6T_{m+1}(3) - T_m(3), \qquad T_m(3) \geq 0.$$

Induction using these facts yields

$$T_M(3) \leq 6^M.$$

Consequently,

$$g_M^+(1) \leq 6^M.$$

Now split according to z. If $0 \leq z \leq 1$, then $z^j \leq 1$ for every index in the sum, so

$$g_M^+(z) = \sum_{0 \leq j < M-1} |g_{M,j}| z^j \leq \sum_{0 \leq j < M-1} |g_{M,j}| = g_M^+(1) \leq 6^M \leq 6^M \max\{1, z^{M-2}\}.$$

If $1 \leq z$, then every index in the finite sum satisfies $j \leq M - 2$ with the same natural-number subtraction convention. Hence $z^j \leq z^{M-2}$, and

$$g_M^+(z) \leq z^{M-2} \sum_{0 \leq j < M-1} |g_{M,j}| = z^{M-2} g_M^+(1) \leq 6^M z^{M-2} \leq 6^M \max\{1, z^{M-2}\}.$$

The two cases prove the claimed bound. □

Lemma 13 is the basic moment inequality for normalized falling-factorial monomials. Lemma 14 bounds the one-arm absolute coefficient sum generated by the Chebyshev reciprocal polynomial, which keeps the light-cell variance compatible with the logarithmic degree calibration in Lemma 7.

The aggregate heavy-cell result and the sparse factorial-polynomial definitions now prepare the two branches of the hybrid estimator for the final light-cell rate bound.

Lemma 15. [`lem:universal-heavy-cell-rate`] (Heavy-cell rate bound) *For every overlap level ϵ satisfying $0 < \epsilon < 1/2$, there exist constants $C_\epsilon, \rho_\epsilon > 0$ and an integer N_ϵ such that the following holds. For every n, d, every single-observation law P on $\{1, \ldots, d\} \times \{0, 1\}^2$, and every sample law μ_n, if*

- ***(Sample size.)*** *$n \geq N_\epsilon$;*
- ***(Dimension range.)*** *$d \leq \rho_\epsilon n \log n$;*
- ***(Experiment membership.)*** *μ_n and P belong to the overlap-restricted iid product-law experiment class $\mathcal{E}_{n,d,\epsilon}$ of Definition 2,*

then the aggregate heavy-cell ratio estimator satisfies

$$\int \left\{ \widehat{T}_{\mathcal{H}}(O_1, \ldots, O_n) - \sum_{k \in \widehat{\mathcal{H}}_n(O_1, \ldots, O_n)} \phi(q_k) \right\}^2 d\mu_n \leq C_\epsilon \left\{ \frac{1}{n} + \frac{d^2}{n^2 (\log n)^2} \right\}.$$

Proof of Lemma 15. Write

$$L = \log(en), \qquad m_j = |I_j| \quad (j = 0, 1),$$

and let N_0 denote the calibration cutoff in Definition 9. For $n \geq N_0$, the implemented heavy set $\widehat{\mathcal{H}}_n$ is the pilot set at threshold 256:

$$\widehat{\mathcal{H}} = \{k : \lfloor 256L \rfloor < N_k^{(0)}\}.$$

The corresponding pilot-bad event is

$$\mathcal{B} = \left\{ \neg \left(\forall k \in \widehat{\mathcal{H}},\, \frac{256L}{2m_0} \leq p_k \right) \text{ or } \neg \left(\forall k \notin \widehat{\mathcal{H}},\, p_k \leq \frac{2 \cdot 256L}{m_0} \right) \right\}.$$

For every sample, the empirical arm ratios in $\widehat{h}_k$ use totalized division, so the ratio is 0 when the corresponding arm count is zero; with this convention each such ratio lies in $[0, 1]$. Hence the absolute empirical ratio contribution of category k is at most $N_k^{(1)}/m_1$, and summing over the selected categories gives

$$|\widehat{T}_{\mathcal{H}}| \leq \sum_{k \in \widehat{\mathcal{H}}_n} \frac{N_k^{(1)}}{m_1} \leq \sum_k \frac{N_k^{(1)}}{m_1} = \frac{m_1}{m_1} \leq 1.$$

For an overlap law P, the category vector q_k from Definition 5 lies in the overlap cone of Definition 3. If $p_k = 0$, then $q_k = 0$ and Definition 4 sets $\phi(q_k) = 0$. If $p_k > 0$, the arm masses in q_k are positive and the displayed formula in Definition 4 gives

$$\phi(q_k) = p_k(\mu_{1k} - \mu_{0k}),$$

where the binary-outcome regressions μ_{0k} and μ_{1k} lie in $[0, 1]$. Thus, in both cases, $|\phi(q_k)| \leq p_k$. Since the category masses sum to one,

$$\left| \sum_{k \in \widehat{\mathcal{H}}_n} \phi(q_k) \right| \leq \sum_{k \in \widehat{\mathcal{H}}_n} p_k \leq \sum_k p_k = 1.$$

Thus the squared heavy-cell error is bounded by 4 pointwise for the overlap laws considered below. The fixed-threshold pilot sandwich from Lemma 1, applied with dimension constant 1, gives constants $C_0 > 0$ and N_{bad} such that, whenever $n \geq N_{\mathrm{bad}}$ and $d \leq nL$,

$$P^{\otimes n}(\mathcal{B}) \leq C_0 n^{-4}.$$

With $C_{\mathrm{bad}} = 4C_0$, the range bound gives

$$\int_{\mathcal{B}} \left(\widehat{T}_{\mathcal{H}} - \sum_{k \in \widehat{\mathcal{H}}_n} \phi(q_k) \right)^2 dP^{\otimes n} \leq C_{\mathrm{bad}}\, n^{-4}.$$

Set

$$K_\epsilon^{\mathcal{H}} = \frac{16}{\epsilon} + 12 + \frac{1}{\epsilon^4}, \qquad C_\epsilon = 4K_\epsilon^{\mathcal{H}} + C_{\mathrm{bad}}, \qquad \rho_\epsilon = 1,$$

and take

$$N_\epsilon = \max\{8, N_0, N_{\mathrm{bad}}\}.$$

These constants satisfy the required positivity conditions.

Assume $n \geq N_\epsilon$, $d \leq \rho_\epsilon n \log n$, and $\mu_n \in \mathcal{E}_{n,d,\epsilon}$. The experiment-class assumption gives $\mu_n = P^{\otimes n}$ and ϵ-overlap for P. Also $n \geq 8$, $n \geq N_0$, $n \geq N_{\mathrm{bad}}$, and, since $\log n \leq L$,

$$d \leq nL.$$

Let

$$B_- = \frac{256L}{2m_0}.$$

On $\mathcal{B}^c$, the first half of the pilot sandwich and the equality $\widehat{\mathcal{H}}_n = \widehat{\mathcal{H}}$ give

$$k \in \widehat{\mathcal{H}}_n \quad \Longrightarrow \quad B_- \leq p_k.$$

Fix a deterministic category set H satisfying $p_k \geq B_-$ for every $k \in H$, and fix $a \in \{0,1\}$. Define

$$S_a(H) = \sum_{k\in H} \frac{N_k^{(1)}}{m_1} \frac{N_{ak1}^{(1)}}{N_{ak}^{(1)}} \mathbf{1}\{N_{ak}^{(1)} > 0\}, \qquad \theta_a(H) = \sum_{k\in H} p_k \mu_{ak},$$

where the displayed ratio is taken to be 0 when $N_{ak}^{(1)} = 0$. The empirical category-mass centering is

$$\sum_{k\in H} \frac{N_k^{(1)}}{m_1} \mu_{ak}.$$

The pointwise decomposition about this centering is

$$S_a(H) - \sum_{k\in H} \frac{N_k^{(1)}}{m_1}\mu_{ak} = \frac{1}{m_1}\left[\sum_{k\in H} N_k^{(1)} \frac{\mathbf{1}\{N_{ak}^{(1)} > 0\}}{N_{ak}^{(1)}}\{N_{ak1}^{(1)} - \mu_{ak} N_{ak}^{(1)}\} - \sum_{k\in H} \mu_{ak} N_k^{(1)} \mathbf{1}\{N_{ak}^{(1)} = 0\}\right].$$

The first bracket is the aggregate centered ratio residual $\mathcal{R}_a(H)$ of Lemma 11, formed on the estimation split: the index set is I_1, so $m = m_1$; the category set H is deterministic and carries $p_k \geq B_- > 0$; and P has overlap at level ϵ. Lemma 11 therefore gives

$$\int \left[\sum_{k\in H} N_k^{(1)} \frac{\mathbf{1}\{N_{ak}^{(1)} > 0\}}{N_{ak}^{(1)}}\{N_{ak1}^{(1)} - \mu_{ak} N_{ak}^{(1)}\}\right]^2 dP^{\otimes n} \leq \frac{2m_1}{\epsilon} \sum_{k\in H} p_k.$$

For the second bracket, $0 \leq \mu_{ak} \leq 1$, so its square is bounded by the square of the missing-arm count

$$\sum_{k\in H} N_k^{(1)} \mathbf{1}\{N_{ak}^{(1)} = 0\}.$$

Since $n \geq 8$ we have $m_1 \geq 3$, and the categories in H have mass at least $B_- > 0$, so that count meets the hypotheses of Lemma 12 on the estimation split, with $m = m_1$ and mass floor $B = B_-$. Combining the domination just recorded with the second moment of Lemma 12 gives

$$\int \left[\sum_{k\in H} \mu_{ak} N_k^{(1)} \mathbf{1}\{N_{ak}^{(1)} = 0\}\right]^2 dP^{\otimes n} \leq m_1 + m_1^2 \left[\frac{|H|}{\{((m_1-2)/2)\epsilon\}^2 B_-}\right]^2.$$

Together with $(x-y)^2 \le 2x^2 + 2y^2$, the two bracket bounds give

$$\int \left(S_a(H) - \sum_{k \in H} \frac{N_k^{(1)}}{m_1} \mu_{ak} \right)^2 dP^{\otimes n} \le \frac{4 \sum_{k \in H} p_k}{m_1 \epsilon} + \frac{2}{m_1} + 2 \left[\frac{|H|}{\{((m_1 - 2)/2)\epsilon\}^2 B_-} \right]^2 .$$

For the empirical category-mass fluctuation, use the one-observation score

$$g_{a,H}(O) = \mu_{aX} \mathbf{1}\{X \in H\}.$$

Its second-split average is the empirical category-mass centering, its expectation is $\theta_a(H)$, and $0 \le g_{a,H} \le 1$. The iid bounded-score variance bound therefore gives

$$\int \left(\sum_{k \in H} \frac{N_k^{(1)}}{m_1} \mu_{ak} - \theta_a(H) \right)^2 dP^{\otimes n} \le \frac{1}{m_1}.$$

Splitting $S_a(H) - \theta_a(H)$ at the empirical category-mass centering yields

$$\int \{S_a(H) - \theta_a(H)\}^2 \, dP^{\otimes n} \le \frac{8 \sum_{k \in H} p_k}{m_1 \epsilon} + \frac{6}{m_1} + 4 \left[\frac{|H|}{\{((m_1 - 2)/2)\epsilon\}^2 B_-} \right]^2 .$$

The probability-mass identities give

$$\sum_{k \in H} p_k \le \sum_k p_k = 1,$$

and the finite-set cardinality bound gives $|H| \le d$. The split-size calibration for $n \ge 8$ gives

$$\frac{n}{2} \le m_1, \qquad \frac{n}{8} \le \frac{m_1 - 2}{2}, \qquad \frac{256L}{n} \le B_-,$$

and $\log n \le L$. Hence

$$\frac{8 \sum_{k \in H} p_k}{m_1 \epsilon} + \frac{6}{m_1} + 4 \left[\frac{|H|}{\{((m_1 - 2)/2)\epsilon\}^2 B_-} \right]^2 \le K_\epsilon^{\mathcal{H}} \left\{ \frac{1}{n} + \frac{d^2}{n^2 (\log n)^2} \right\}.$$

Let $V = K_\epsilon^{\mathcal{H}} \{n^{-1} + d^2/[n^2(\log n)^2]\}$. The preceding deterministic bound applies uniformly to every fixed H whose elements have mass at least B_-. To pass to the random pilot-heavy set on $\mathcal{B}^c$, partition the pilot sample space into the fibers on which $\widehat{\mathcal{H}}_n = H$. The partition is carried out on the ambient iid sequence, whose first n coordinates have law $P^{\otimes n}$ by Lemma 20, so every integral below may equally be read under either law. On each contributing fiber, the good-pilot implication gives $p_k \ge B_-$ for every $k \in H$. The fixed-H arm error is a function of the estimation split, while the fiber is determined by the pilot split, so the product law factors the restricted integral into the fiber probability times the fixed-H product-law integral. Summing over the eligible fibers and using that their probabilities sum to at most one gives, for each $a \in \{0, 1\}$,

$$\int_{\mathcal{B}^c} \{S_a(\widehat{\mathcal{H}}_n) - \theta_a(\widehat{\mathcal{H}}_n)\}^2 \, dP^{\otimes n} \le V.$$

The heavy contribution and target decompose into their two arm components:

$$\widehat{T}_{\mathcal{H}} - \sum_{k \in \widehat{\mathcal{H}}_n} \phi(q_k) = \{S_1(\widehat{\mathcal{H}}_n) - \theta_1(\widehat{\mathcal{H}}_n)\} - \{S_0(\widehat{\mathcal{H}}_n) - \theta_0(\widehat{\mathcal{H}}_n)\}.$$

Thus $(x-y)^2 \le 2x^2 + 2y^2$ gives

$$\int_{\mathcal{B}^c} \left(\widehat{T}_{\mathcal{H}} - \sum_{k \in \widehat{\mathcal{H}}_n} \phi(q_k) \right)^2 dP^{\otimes n} \le 4K_\epsilon^{\mathcal{H}} \left\{ \frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \right\}.$$

The bad-event contribution satisfies

$$\int_{\mathcal{B}} \left(\widehat{T}_{\mathcal{H}} - \sum_{k \in \widehat{\mathcal{H}}_n} \phi(q_k) \right)^2 dP^{\otimes n} \le C_{\mathrm{bad}} n^{-4} \le C_{\mathrm{bad}} \left\{ \frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \right\},$$

because $n^{-4} \le n^{-1}$ for $n \ge 8$ and $n^{-1} \le n^{-1} + d^2/[n^2(\log n)^2]$. Adding the integrals over $\mathcal{B}$ and $\mathcal{B}^c$, and using $\mu_n = P^{\otimes n}$, yields

$$\int \left(\widehat{T}_{\mathcal{H}} - \sum_{k \in \widehat{\mathcal{H}}_n} \phi(q_k) \right)^2 d\mu_n \le (4K_\epsilon^{\mathcal{H}} + C_{\mathrm{bad}}) \left\{ \frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \right\},$$

which is the claimed bound with the chosen C_ϵ. □

Definition 14. [`synth_20`] (Sparse factorial-polynomial means) *Given a law P, a category $k \in \{1, \dots, d\}$, a degree M, and a scale B, write $q_k = (q_{0k0}, q_{0k1}, q_{1k0}, q_{1k1})$ for the four-cell mass vector of category k under P. For an arm $a \in \{0,1\}$, the sparse-arm mean is*

$$\bar{A}_{a,k}(P, M, B) = \frac{s(q_k)\, t_a(q_k)}{B} G_M \left(\frac{s_a(q_k)}{B} \right),$$

with s, s_a, t_a, and G_M as in Definition 9. The sparse factorial-polynomial mean of category k is the difference of the two arm means,

$$\bar{P}_k = P_{M,B}(q_k) = \bar{A}_{1,k}(P, M, B) - \bar{A}_{0,k}(P, M, B).$$

Lemma 16. [`lem:factorial-moment-variance`] (Factorial moment variance bound) *Use the split notation and hybrid-estimator calibration in Definitions 7 and 9. For any sample size n, alphabet size d, probability law P on the finite observed-data alphabet, and category k, let $\bar{P}_k$ denote the sparse factorial-polynomial mean of the light-cell statistic $\widehat{P}_k$. Suppose that*

- ***(Split size.)*** *$m_1 > 0$.*
- ***(Degree and scale.)*** *$M(n) > 0$ and $B(n) > 0$.*
- ***(Sample-size calibration.)*** *$4M(n)^2 \le m_1$.*
- ***(Mass envelope.)*** *The category mass $p_k = P(X = k)$ satisfies*

$$p_k + \frac{4M(n)}{m_1} \le B(n) \qquad \textit{and} \qquad p_k \le B(n).$$

Then, under the ambient infinite iid product law $P^{\otimes \infty}$,

$$\int \left(\widehat{P}_k - \bar{P}_k \right)^2 dP^{\otimes \infty} \le 8(\mathrm{e} + 1) \left(B(n) 6^{M(n)} \right)^2.$$

Proof of Lemma 16. All expectations below are taken under the iid sequence law generated by P, equivalently under the corresponding n-fold product law for functions of $O_1, \ldots, O_n$. Write $M = M(n)$, $B = B(n)$, $m_1 = |I_1|$, and $q = q_k$. The coefficients in the sparse expansion are the coefficients of the polynomial continuation G_M:

$$g_{M,j} = (-1)^j \frac{2^{2j+3}}{M(M+j+2)} \binom{M+j+2}{2j+4}, \qquad G_M(x) = \sum_{j=0}^{M-2} g_{M,j} x^j.$$

For $a \in \{0,1\}$, $c \in \{0,1\}^2$, and $0 \le t \le j \le M-2$, let $r^{a,c,j,t}$ be the sparse multi-index determined by

$$u^{r^{a,c,j,t}} = u_c\, u_{a1}\, u_{a0}^t\, u_{a1}^{j-t}.$$

The sparse arm mean is

$$\bar{A}_{a,k} = \sum_{j=0}^{M-2} \sum_{t=0}^{j} \sum_{c \in \{0,1\}^2} B^{-1} g_{M,j} B^{-j} \binom{j}{t} q^{r^{a,c,j,t}},$$

and the sparse polynomial mean is

$$\bar{P}_k = \bar{A}_{1,k} - \bar{A}_{0,k}.$$

For a nonnegative four-cell vector v, set

$$\mathcal{E}_a(v) = \sum_{j=0}^{M-2} \sum_{t=0}^{j} \sum_{c \in \{0,1\}^2} \left| B^{-1} g_{M,j} B^{-j} \binom{j}{t} \right| v^{r^{a,c,j,t}}.$$

The triangle inequality gives $|\bar{A}_{a,k}| \le \mathcal{E}_a(q)$. We first record the envelope bound used for both q and its shifted version. If $v \ge 0$, $\sum_c v_c \le B$, $M > 0$, and $B > 0$, then summing first over c and then using the binomial theorem gives

$$\mathcal{E}_a(v) = B^{-1} \left(\sum_c v_c \right) v_{a1} \sum_{j=0}^{M-2} |g_{M,j}| \left(\frac{v_{a0} + v_{a1}}{B} \right)^j.$$

The coefficient-sum certificate in Lemma 14 gives, for every $z \ge 0$,

$$\sum_{j=0}^{M-2} |g_{M,j}| z^j \le 6^M \max\{1, z^{M-2}\}.$$

In the envelope application, $z = (v_{a0} + v_{a1})/B$ lies in $[0,1]$. Since $0 \le v_{a1} \le \sum_c v_c \le B$,

$$\mathcal{E}_a(v) \le B^{-1} \left(\sum_c v_c \right)^2 6^M \le B 6^M.$$

Applying this with $v = q$, using $q \ge 0$ and $\sum_c q_c = p_k \le B$, gives

$$|\bar{P}_k| \le |\bar{A}_{1,k}| + |\bar{A}_{0,k}| \le 2B6^M, \qquad \bar{P}_k^2 \le 4(B6^M)^2.$$

The sample arm contribution has the parallel factorial-moment expansion

$$\widehat{A}_{a,k} = \sum_{j=0}^{M-2} \sum_{t=0}^{j} \sum_{c\in\{0,1\}^2} B^{-1} g_{M,j} B^{-j} \binom{j}{t} \frac{\prod_{b,y} (N^{(1)}_{bky})_{r^{a,c,j,t}_{by}}}{(m_1)_{|r^{a,c,j,t}|}}, \qquad \widehat{P}_k = \widehat{A}_{1,k} - \widehat{A}_{0,k}.$$

For a multi-index r, write

$$F_r = \frac{\prod_{b,y} (N^{(1)}_{bky})_{r_{by}}}{(m_1)_{|r|}}.$$

For two displayed terms with multi-indices r and s, the range restrictions give $|r| = j+2 \le M$ and $|s| = j'+2 \le M$. Suppose first that $|r| \le |s|$. Since $|s| \le M$ and $4M^2 \le m_1$, the split-size condition in Lemma 13 applies and gives

$$\mathbb{E}[F_r F_s] \le \mathrm{e}\, q^r \left(q + \frac{|s|}{m_1} \mathbf{1} \right)^s.$$

Coordinatewise nonnegativity of q, the inequality $|s| \le M$, and monotonicity of nonnegative monomials yield

$$\mathbb{E}[F_r F_s] \le \mathrm{e} \left(q + \frac{M}{m_1} \mathbf{1} \right)^r \left(q + \frac{M}{m_1} \mathbf{1} \right)^s.$$

When $|s| \le |r|$, the same argument with r and s interchanged gives the same bound. The unweighted factorial monomial product is nonnegative pointwise, and the product of the two displayed coefficients is bounded above by the product of their absolute values. Hence

$$\begin{aligned} &\mathbb{E}\left[\left(B^{-1} g_{M,j} B^{-j} \binom{j}{t} F_{r^{a,c,j,t}} \right) \left(B^{-1} g_{M,j'} B^{-j'} \binom{j'}{t'} F_{r^{a,c',j',t'}} \right) \right] \\ &\quad \le \mathrm{e} \left| B^{-1} g_{M,j} B^{-j} \binom{j}{t} \right| \left(q + \frac{M}{m_1} \mathbf{1} \right)^{r^{a,c,j,t}} \left| B^{-1} g_{M,j'} B^{-j'} \binom{j'}{t'} \right| \left(q + \frac{M}{m_1} \mathbf{1} \right)^{r^{a,c',j',t'}}. \end{aligned}$$

Expanding $\widehat{A}^2_{a,k}$ as the finite double sum, applying this estimate term by term, and reassembling the absolute-coefficient sums gives

$$\mathbb{E}[\widehat{A}^2_{a,k}] \le \mathrm{e}\, \mathcal{E}_a \left(q + \frac{M}{m_1} \mathbf{1} \right)^2.$$

The shifted vector is nonnegative and has total mass

$$\sum_{b,y} \left(q_{by} + \frac{M}{m_1} \right) = p_k + \frac{4M}{m_1} \le B,$$

by the shifted light-cell mass condition. Applying the envelope bound to the shifted vector gives

$$\mathbb{E}[\widehat{A}^2_{a,k}] \le \mathrm{e}\, (B 6^M)^2 \qquad (a = 0, 1).$$

Using $(u-v)^2 \le 2u^2 + 2v^2$ with $u = \widehat{A}_{1,k}$ and $v = \widehat{A}_{0,k}$,

$$\mathbb{E}[\widehat{P}^2_k] \le 2\mathbb{E}[\widehat{A}^2_{1,k}] + 2\mathbb{E}[\widehat{A}^2_{0,k}] \le 4\mathrm{e}\, (B 6^M)^2.$$

Finally, pointwise,

$$(\widehat{P}_k - \bar{P}_k)^2 \le 2\widehat{P}_k^2 + 2\bar{P}_k^2,$$

because $(\widehat{P}_k + \bar{P}_k)^2 \ge 0$. Integrating and substituting the two bounds above gives

$$\mathbb{E}\Big[(\widehat{P}_k - \bar{P}_k)^2\Big] \le 2 \cdot 4\mathrm{e}\,(B6^M)^2 + 2 \cdot 4(B6^M)^2 = 8(\mathrm{e}+1)(B6^M)^2,$$

as required. □

Lemma 17. [`lem:factorial-moment-covariance`] (Off-diagonal factorial covariance) *Use the split notation and calibration m_1, $M(n)$, and $B(n)$ from Definitions 7 and 9. For each category k, let $\widehat{P}_k$ be the factorial-moment polynomial estimate from Definition 9, computed from the first n coordinates of the ambient iid sequence, and let*

$$\bar{P}_k = \bar{A}_{1,k}\big(P, M(n), B(n)\big) - \bar{A}_{0,k}\big(P, M(n), B(n)\big)$$

denote the corresponding sparse factorial-polynomial mean, written as the treated sparse-arm mean minus the control sparse-arm mean.

For any single-observation law P on $\{1,\dots,d\} \times \{0,1\}^2$ and any two categories $k, l \in \{1,\dots,d\}$, suppose that

- ***Distinct categories.*** *$k \ne l$.*
- ***Positive calibration.*** *$M(n) > 0$ and $B(n) > 0$.*
- ***Second-split size.*** *$4M(n)^2 \le m_1$.*
- ***Light masses.*** *$p_k \le B(n)$ and $p_l \le B(n)$.*

Then, under the ambient infinite iid product law $P^{\otimes\infty}$,

$$\mathbb{E}_{P^{\otimes\infty}}\Big[\big(\widehat{P}_k - \bar{P}_k\big)\big(\widehat{P}_l - \bar{P}_l\big)\Big] \le \frac{8M(n)^2}{m_1}\Big\{B(n)\,6^{M(n)}\Big\}^2.$$

Proof of Lemma 17. Let μ be the iid product law generated by P on the sample sequence, and abbreviate

$$M = M(n), \qquad B = B(n), \qquad m = m_1, \qquad R = B\,6^M, \qquad \eta = \frac{2M^2}{m}.$$

Write the polynomial continuation as

$$G_M(x) = \sum_{j=0}^{M-2} g_{M,j} x^j.$$

For $h \in \{k, l\}$, let q_h be the four-cell vector of category h under P, and write $q_h^r = \prod_{b,y} (q_h)_{by}^{r_{by}}$. For an arm $a \in \{0,1\}$, a cell $c \in \{0,1\}^2$, and integers $0 \le t \le j \le M-2$, let $r^{a,c,j,t}$ be the multi-index determined by

$$u^{r^{a,c,j,t}} = u_c\, u_{a1}\, u_{a0}^{t}\, u_{a1}^{j-t}.$$

The sparse arm contribution and its sparse mean are

$$\widehat{A}_{a,h} = \sum_{j=0}^{M-2} \sum_{t=0}^{j} \sum_{c\in\{0,1\}^2} B^{-1} g_{M,j} B^{-j} \binom{j}{t} \frac{\prod_{b,y} (N^{(1)}_{bhy})_{r^{a,c,j,t}_{by}}}{(m)_{|r^{a,c,j,t}|}},$$

$$\bar{A}_{a,h} = \sum_{j=0}^{M-2} \sum_{t=0}^{j} \sum_{c\in\{0,1\}^2} B^{-1} g_{M,j} B^{-j} \binom{j}{t} q_h^{r^{a,c,j,t}}.$$

Thus

$$\widehat{P}_h = \widehat{A}_{1,h} - \widehat{A}_{0,h}, \qquad \bar{P}_h = \bar{A}_{1,h} - \bar{A}_{0,h}.$$

The size condition $4M^2 \le m$, together with $M > 0$, gives $M \le m$. Since each displayed sparse multi-index has total degree $j + 2 \le M$, the factorial-moment identity gives, for $h = k, l$,

$$\int \widehat{A}_{a,h}\, d\mu = \bar{A}_{a,h} \quad (a = 0, 1), \qquad \int \widehat{P}_h\, d\mu = \bar{P}_h.$$

Consequently, by expanding the centered product and integrating term by term,

$$\int (\widehat{P}_k - \bar{P}_k)(\widehat{P}_l - \bar{P}_l)\, d\mu = \int \widehat{P}_k \widehat{P}_l\, d\mu - \bar{P}_k \bar{P}_l.$$

For $h \in \{k, l\}$ and $a \in \{0, 1\}$, define the absolute arm envelope

$$\mathsf{Env}_{a,h} = \sum_{j=0}^{M-2} \sum_{t=0}^{j} \sum_{c\in\{0,1\}^2} \left| B^{-1} g_{M,j} B^{-j} \binom{j}{t} \right| q_h^{r^{a,c,j,t}}.$$

Fix two sparse monomials with multi-indices r and s from categories k and l, and put

$$\widehat{F}_{h,r} = \frac{\prod_{b,y} (N^{(1)}_{bhy})_{r_{by}}}{(m)_{|r|}}.$$

The product of the two normalized factorial counts is an average over ordered injective selections for the k-monomial and ordered injective selections for the l-monomial. Because $k \neq l$, an observation cannot satisfy both category patterns; hence exactly the disjoint pairs of selections contribute. There are $(m)_{|r|+|s|}$ such ordered pairs, and independence gives the same cell-mass product $q_k^r q_l^s$ for each pair. Therefore

$$\int \widehat{F}_{k,r} \widehat{F}_{l,s}\, d\mu = \Gamma_{r,s}\, q_k^r q_l^s, \qquad \Gamma_{r,s} = \frac{(m)_{|r|+|s|}}{(m)_{|r|}(m)_{|s|}}.$$

For sparse terms, $|r| \le M$ and $|s| \le M$. We next bound the normalization factor. Suppose first that $|r| \le |s|$. If $|s| = 0$, then $|r| = 0$ and $\Gamma_{r,s} = 1$. If $|s| > 0$, the condition $4|s|^2 \le m$ follows from $|s| \le M$ and $4M^2 \le m$, and it implies $2|s| \le m$. Using

$$\Gamma_{r,s} = \frac{(m - |r|)_{|s|}}{(m)_{|s|}},$$

we have $\Gamma_{r,s} \le 1$. Also

$$\Gamma_{r,s} \ge \frac{(m-2|s|)^{|s|}}{m^{|s|}} = \left(1 - \frac{2|s|}{m}\right)^{|s|} \ge 1 - \frac{2|s|^2}{m},$$

where the last step is Bernoulli's inequality, using $-1 \le -2|s|/m$. Hence

$$|\Gamma_{r,s} - 1| \le \frac{2|s|^2}{m} \le \eta.$$

The case $|s| \le |r|$ is identical after exchanging r and s, using the symmetry of $\Gamma_{r,s}$. Since the four-cell masses are nonnegative,

$$\left| \int \widehat{F}_{k,r} \widehat{F}_{l,s} \, d\mu - q_k^r q_l^s \right| \le \eta \, q_k^r q_l^s, \qquad |r|, |s| \le M.$$

Multiplying by the two sparse coefficients, summing over the two finite arm expansions, and using the triangle inequality yields

$$\left| \int \widehat{A}_{a,k} \widehat{A}_{b,l} \, d\mu - \bar{A}_{a,k} \bar{A}_{b,l} \right| \le \eta \, \mathsf{Env}_{a,k} \mathsf{Env}_{b,l}.$$

For $h = k, l$, the coordinates of q_h are nonnegative and sum to p_h. For fixed j, positivity of B and of the binomial coefficients gives

$$\begin{aligned}
&\sum_{t=0}^{j} \sum_{c \in \{0,1\}^2} \left| B^{-1} g_{M,j} B^{-j} \binom{j}{t} \right| q_h^{r^{a,c,j,t}} \\
&= B^{-1} |g_{M,j}| B^{-j} \sum_{t=0}^{j} \sum_{c \in \{0,1\}^2} \binom{j}{t} (q_h)_c (q_h)_{a1} (q_h)_{a0}^t (q_h)_{a1}^{j-t} \\
&= B^{-1} |g_{M,j}| B^{-j} \, s(q_h) (q_h)_{a1} \big((q_h)_{a0} + (q_h)_{a1} \big)^j.
\end{aligned}$$

Summing in j, and using $s_a(q_h) = (q_h)_{a0} + (q_h)_{a1}$, yields

$$\mathsf{Env}_{a,h} = B^{-1} s(q_h) (q_h)_{a1} \sum_{j=0}^{M-2} |g_{M,j}| \left\{ \frac{s_a(q_h)}{B} \right\}^j.$$

The light-cell hypotheses give $p_h = s(q_h) \le B$, while $(q_h)_{a1} \le s(q_h)$ and $B > 0$. Hence

$$B^{-1} s(q_h) (q_h)_{a1} \le B^{-1} s(q_h)^2 \le B, \qquad 0 \le \frac{s_a(q_h)}{B} \le 1.$$

It remains to record the coefficient envelope used in this display. For every $z \ge 0$,

$$\sum_{j=0}^{M-2} |g_{M,j}| z^j \le 6^M \max\{1, z^{M-2}\}.$$

Indeed, the sign pattern of the explicit coefficients gives $\sum_{j=0}^{M-2} |g_{M,j}| = G_M(-1)$. The defining Chebyshev identity for the continuation, evaluated at -1, gives

$$T_M(3) = 1 + 2M^2 + 2M^2 \sum_{j=0}^{M-2} |g_{M,j}|,$$

so nonnegativity of the coefficient sum implies $\sum_{j=0}^{M-2} |g_{M,j}| \le T_M(3)$. The Chebyshev recurrence gives $T_M(3) \le 6^M$. If $0 \le z \le 1$, termwise monotonicity gives $\sum |g_{M,j}| z^j \le \sum |g_{M,j}| \le 6^M$. If $z \ge 1$, then $j \le M-2$ on the displayed range, so $\sum |g_{M,j}| z^j \le z^{M-2} \sum |g_{M,j}| \le 6^M z^{M-2}$. This proves the stated envelope. Taking $z = s_a(q_h)/B$ makes the maximum equal to 1, and therefore

$$\mathsf{Env}_{a,h} \le B\, 6^M = R.$$

Hence, for all $a, b \in \{0,1\}$,

$$\left| \int \widehat{A}_{a,k} \widehat{A}_{b,l}\, d\mu - \bar{A}_{a,k} \bar{A}_{b,l} \right| \le \eta R^2.$$

Define

$$C_{ab} = \int \widehat{A}_{a,k} \widehat{A}_{b,l}\, d\mu - \bar{A}_{a,k} \bar{A}_{b,l}.$$

Expanding the two arm contrasts gives

$$\int \widehat{P}_k \widehat{P}_l \, d\mu - \bar{P}_k \bar{P}_l = C_{11} - C_{10} - C_{01} + C_{00}.$$

Therefore

$$\left| \int \widehat{P}_k \widehat{P}_l \, d\mu - \bar{P}_k \bar{P}_l \right| \le |C_{11}| + |C_{10}| + |C_{01}| + |C_{00}| \le 4\eta R^2 = \frac{8M^2}{m} \{B\, 6^M\}^2.$$

Combining this bound with the centering identity and using $x \le |x|$ yields

$$\int (\widehat{P}_k - \bar{P}_k)(\widehat{P}_l - \bar{P}_l)\, d\mu \le \frac{8M(n)^2}{m_1} \{B(n) 6^{M(n)}\}^2.$$

□

Lemma 18. [`lem:false-light-envelope`] (False-light error envelope) *Let P be a probability law on $\{1, \ldots, d\} \times \{0,1\}^2$. Suppose that:*

- ***(Overlap.)*** *For every category k with $p_k > 0$, the treatment probability satisfies*

$$\epsilon \le \pi_k \le 1 - \epsilon.$$

- ***(Calibration cutoff.)*** *The sample size satisfies $n \ge N_0$, where N_0 is the least numerical cutoff satisfying the calibration inequalities used in the estimator.*
- ***(Logarithmic scale.)*** *With $L = \log(en)$, one has $L \ge 240$.*

For the pilot-light indicator $S_k = \mathbf{1}\{k \in \widehat{\mathcal{L}}_n\}$, the selected false-light error over categories whose population mass exceeds the light-cell threshold satisfies

$$E_{P^{\otimes n}} \left[\left\{ \sum_{\substack{k \in \{1,\ldots,d\} \\ p_k > B(n)/4}} S_k \big(\widehat{P}_k - \phi(q_k) \big) \right\}^2 \right] \le 8(\mathrm{e}+1) 8192^2 \left\{ \frac{1}{n} + \frac{d^2}{n^2 (\log n)^2} \right\}.$$

Proof of Lemma 18. All expectations below are under the iid infinite product law generating the observation sequence, with each statistic evaluated on its first n coordinates. Let

$$L = \log(en), \qquad \ell = \log n, \qquad m_j = |I_j|\ (j = 0, 1), \qquad B = B(n), \qquad M = M(n), \qquad G = 6^{2M}.$$

The deterministic split construction of Definition 7 gives

$$m_0 = \lfloor n/2 \rfloor, \qquad m_1 = n - \lfloor n/2 \rfloor.$$

By Lemma 7,

$$0 < \ell, \qquad \ell \le L, \qquad L \le 2\ell, \qquad m_1 > 0, \qquad n \le 2m_1, \qquad M \le L,$$

$$B \le \frac{8192L}{n}, \qquad G^2 L^6 \le n.$$

In particular $n > 0$. Since $B = 4096L/m_1$, $L > 0$, and $m_1 > 0$, also $B > 0$, and hence

$$B^2 \le \frac{8192^2 L^2}{n^2}.$$

Because $n > 0$, $L = 1 + \ell$. Since $L \ge 240$,

$$\exp(-49L) \le \exp(-L) = \frac{1}{en} \le \frac{1}{n}.$$

Moreover $G \ge 1$, so $L^4 \le L^6 \le G^2 L^6 \le n$. Together with $\ell \le L$, this gives

$$L^2 \ell^2 \le L^4 \le n.$$

Therefore

$$\frac{d^2 L^2}{n^3} \le \frac{d^2}{n^2 \ell^2} \le \frac{1}{n} + \frac{d^2}{n^2 \ell^2}.$$

Let

$$F = \{k : p_k > B/4\}, \qquad S_k = \mathbf{1}\{k \in \widehat{\mathcal{L}}_n\}, \qquad e_k = S_k(\widehat{P}_k - \phi(q_k)).$$

By the definition of the false-light error, the selected false-light error is $\sum_{k \in F} e_k$. For each sample, Cauchy's inequality gives

$$\left(\sum_{k \in F} e_k \right)^2 \le |F| \sum_{k \in F} e_k^2.$$

Fix $k \in F$. The cutoff defining N_0 in Definition 9 and the assumption $n \ge N_0$ give

$$\alpha_0 L \ge 2, \qquad 4M^2 \le m_1, \qquad \frac{4M}{m_1} \le \frac{3B}{4}.$$

Since $M \ge 2$, the bound $4M^2 \le m_1$ gives $16 \le m_1$. With the split identities above, this implies $m_0 > 0$ and $m_1 \le 2m_0$.

Write

$$G_M(x) = \sum_{j=0}^{M-2} g_{M,j} x^j, \qquad g_{M,j} = (-1)^j \frac{2^{2j+3}}{M(M+j+2)} \binom{M+j+2}{2j+4}.$$

Let $\mathbf{e}_{by}$ denote the unit multi-index at the four-cell coordinate $(b, y) \in \{0, 1\}^2$. In the next display, ay is a single four-cell label, say $ay = (b, y')$, and $\mathbf{e}_{ay}$ is the corresponding unit multi-index. For $a \in \{0, 1\}$, $ay \in \{0, 1\}^2$, and $0 \leq t \leq j \leq M - 2$, set

$$r(a, ay, j, t) = \mathbf{e}_{ay} + \mathbf{e}_{a1} + t\mathbf{e}_{a0} + (j - t)\mathbf{e}_{a1}, \qquad |r(a, ay, j, t)| = j + 2.$$

For a multi-index r, the factorial-moment monomial is

$$\mathfrak{m}_{k,r} = \frac{\prod_{(b,y)\in\{0,1\}^2} \big(N^{(1)}_{bky}\big)_{r_{by}}}{(m_1)_{|r|}},$$

where $(z)_s$ denotes the falling factorial. The a-arm sparse factorial-moment contribution is

$$\widehat{P}^{\mathrm{sp}}_{k,a} = \sum_{j=0}^{M-2} \sum_{t=0}^{j} \sum_{ay\in\{0,1\}^2} B^{-1} g_{M,j} B^{-j} \binom{j}{t} \mathfrak{m}_{k,r(a,ay,j,t)}.$$

By Definition 9, $\widehat{P}_k = \sum_{|r|\leq M} c_r\, \mathfrak{m}_{k,r}$ in collected coefficient form, where each c_r is obtained by expanding the two arms of $P_{M,B}$ through the coefficients $g_{M,j}$ and the binomial expansion of the arm masses, and summing the resulting sparse coefficients $c_u = B^{-1} g_{M,j} B^{-j} \binom{j}{t}$ over the sparse indices $u = (j, t, ay)$ with $r_u = r$. Since $r \mapsto \mathfrak{m}_{k,r}$ enters linearly in each coefficient, regrouping this finite double sum arm by arm yields

$$\widehat{P}_k = \widehat{P}^{\mathrm{sp}}_{k,1} - \widehat{P}^{\mathrm{sp}}_{k,0}.$$

For a nonnegative four-cell vector v, define the absolute sparse arm envelope

$$\mathcal{A}_a(v) = \sum_{j=0}^{M-2} \sum_{t=0}^{j} \sum_{ay\in\{0,1\}^2} \left| B^{-1} g_{M,j} B^{-j} \binom{j}{t} \right| v^{r(a,ay,j,t)}, \qquad v^r = \prod_{(b,y)\in\{0,1\}^2} v_{by}^{r_{by}}.$$

Set

$$v^{(k)}_{ay} = q_{k,ay} + \frac{M}{m_1}.$$

This vector is nonnegative. Fix two sparse indices $u = (j, t, ay)$ and $u' = (j', t', ay')$ with $0 \leq t \leq j \leq M - 2$ and $0 \leq t' \leq j' \leq M - 2$. Both multi-indices $r_u, r_{u'}$ have degree at most M, so $4 \max\{|r_u|, |r_{u'}|\}^2 \leq 4M^2 \leq m_1$. Apply Lemma 13 to the degree-ordered pair. Its unshifted monomial factor is bounded coordinatewise by $(v^{(k)})^r$, since $q_{k,cy} \leq q_{k,cy} + M/m_1$. Its shifted monomial factor is also bounded coordinatewise by the corresponding power of $v^{(k)}$: if s is the larger-degree multi-index in the ordered pair, then $|s| \leq M$, and hence

$$q_{k,cy} + \frac{|s|}{m_1} \leq q_{k,cy} + \frac{M}{m_1} = v^{(k)}_{cy} \qquad \text{for every coordinate } cy.$$

Using these two coordinatewise bounds, and swapping u, u' before applying Lemma 13 when the opposite degree order holds, gives

$$E\big[\big(c_u \mathfrak{m}_{k,r_u}\big)\big(c_{u'} \mathfrak{m}_{k,r_{u'}}\big)\big] \leq \mathrm{e}\, |c_u|\, (v^{(k)})^{r_u}\, |c_{u'}|\, (v^{(k)})^{r_{u'}},$$

where $c_u = B^{-1} g_{M,j} B^{-j} \binom{j}{t}$ and $r_u = r(a, ay, j, t)$. Expanding the square of $\widehat{P}^{\mathrm{sp}}_{k,a}$, summing this bound over all ordered pairs of sparse indices, and collecting the two sums into the envelope gives

$$E\left[(\widehat{P}^{\mathrm{sp}}_{k,a})^2\right] \leq \mathrm{e}\,\mathcal{A}_a(v^{(k)})^2, \qquad a = 0, 1.$$

The shifted vector has total mass

$$\sum_{ay} v^{(k)}_{ay} = p_k + \frac{4M}{m_1} \leq p_k + \frac{3B}{4} \leq B\left(1 + \frac{p_k}{B}\right).$$

Let $R = \sum_{ay} v_{ay}$, $A_a(v) = v_{a0} + v_{a1}$, and

$$g^+_M(z) = \sum_{j=0}^{M-2} |g_{M,j}| z^j.$$

For every nonnegative v with $R \leq Bw$ and $w \geq 1$, the binomial summation in the sparse envelope gives

$$\mathcal{A}_a(v) = B^{-1} R\, v_{a1}\, g^+_M\left(\frac{A_a(v)}{B}\right),$$

and, since $0 \leq A_a(v)/B \leq R/B \leq w$ with $w \geq 1$, Lemma 14 gives

$$g^+_M\left(\frac{A_a(v)}{B}\right) \leq 6^M \max\left\{1, \left(\frac{A_a(v)}{B}\right)^{M-2}\right\} \leq 6^M w^{M-2}.$$

Since $v_{a1} \leq R \leq Bw$, these displays imply

$$\mathcal{A}_a(v) \leq B 6^M w^M.$$

Applying this with $v = v^{(k)}$ and $w = 1 + p_k/B$ yields

$$\mathcal{A}_a(v^{(k)}) \leq B 6^M \left(1 + \frac{p_k}{B}\right)^M, \qquad a = 0, 1.$$

Since $\widehat{P}_k$ is the difference of the two sparse-arm contributions, $(x-y)^2 \leq 2x^2 + 2y^2$ yields

$$E[\widehat{P}_k^2] \leq 4\mathrm{e}\left\{B 6^M \left(1 + \frac{p_k}{B}\right)^M\right\}^2.$$

Because $k \in F$ and $B > 0$, $p_k > 0$. The overlap condition and Definition 4 give $|\phi(q_k)| \leq p_k$. Since $p_k/B \geq 0$ and $M \geq 2$,

$$\frac{p_k}{B} \leq 6^M \left(1 + \frac{p_k}{B}\right)^M,$$

and therefore

$$\phi(q_k)^2 \leq \left\{B 6^M \left(1 + \frac{p_k}{B}\right)^M\right\}^2.$$

Using $(x-t)^2 \leq 2x^2 + 2t^2$,

$$E\left[(\widehat{P}_k - \phi(q_k))^2\right] \leq 8(\mathrm{e}+1) B^2 \left\{6^M \left(1 + \frac{p_k}{B}\right)^M\right\}^2.$$

The indicator S_k is determined by the pilot split, while $\widehat{P}_k - \phi(q_k)$ is computed from the estimation split and a fixed target. The split factorization gives

$$E\Big[S_k^2(\widehat{P}_k - \phi(q_k))^2\Big] = E[S_k^2]E\Big[(\widehat{P}_k - \phi(q_k))^2\Big].$$

Since $S_k \in \{0,1\}$, $E[S_k^2] = E[S_k]$. The indicator integral is the probability of the pilot-light event:

$$E[S_k] = P^{\otimes n}\{k \in \widehat{\mathcal{L}}_n\}.$$

For $n \geq N_0$, the event $\{k \in \widehat{\mathcal{L}}_n\}$ is contained in

$$\Big\{N_k^{(0)} \leq 256L\Big\}.$$

The false-light condition $p_k > B/4$, the identity $B = 4096L/m_1$, and $m_1 \leq 2m_0$ imply

$$2(256L) < m_0 p_k.$$

The pilot lower-tail bound for this threshold gives

$$E[S_k] \leq \exp\Big(-\frac{m_0 p_k}{8}\Big).$$

Since $p_k/B > 1/4$,

$$1 + \frac{p_k}{B} \leq 5\frac{p_k}{B}, \qquad 6\left(1 + \frac{p_k}{B}\right) \leq 30\frac{p_k}{B}.$$

Also $6(1 + p_k/B) \geq 1$, so $\log(6(1 + p_k/B)) \geq 0$; and from $\log x \leq x - 1$ for $x > 0$,

$$\log\Big(6\left(1 + \frac{p_k}{B}\right)\Big) \leq 30\frac{p_k}{B}.$$

Together with $M \leq L$, $B = 4096L/m_1$, and $m_1 \leq 2m_0$, this yields

$$2M\log\Big(6\left(1 + \frac{p_k}{B}\right)\Big) \leq 60L\frac{p_k}{B}, \qquad 256L\frac{p_k}{B} \leq \frac{m_0 p_k}{8}.$$

Consequently

$$-\frac{m_0 p_k}{8} + 2M\log\Big(6\left(1 + \frac{p_k}{B}\right)\Big) \leq -196L\frac{p_k}{B} \leq -49L,$$

and hence

$$\exp\Big(-\frac{m_0 p_k}{8}\Big)\left\{6^M\left(1 + \frac{p_k}{B}\right)^M\right\}^2 \leq \exp(-49L).$$

Combining the preceding displays gives the single-cell bound

$$E[e_k^2] \leq 8(\mathrm{e}+1)B^2\exp(-49L).$$

Since $|F| \leq d$, integrating the pointwise Cauchy bound and applying the single-cell estimate gives

$$E\left[\left(\sum_{k\in F} e_k\right)^2\right] \leq |F|\sum_{k\in F} 8(\mathrm{e}+1)B^2\exp(-49L) \leq d^2\, 8(\mathrm{e}+1)B^2\exp(-49L).$$

Using the bounds on B^2, $\exp(-49L)$, and the rate algebra above,

$$d^2\, 8(\mathrm{e}+1)B^2\exp(-49L) \leq 8(\mathrm{e}+1)8192^2\,\frac{d^2L^2}{n^3} \leq 8(\mathrm{e}+1)8192^2\left\{\frac{1}{n} + \frac{d^2}{n^2(\log n)^2}\right\}.$$

$\square$

Lemma 15 aggregates the ratio branch over pilot-certified heavy categories at the target rate. The sparse means in Definition 14 identify the population polynomial quantities used to center the light-cell statistics. The variance bound in Lemma 16, the off-diagonal covariance bound in Lemma 17, and the false-light envelope in Lemma 18 together control the stochastic and pilot-selection terms that arise in the polynomial branch.

The light-cell theorem combines computability, approximation bias, factorial-moment variance, covariance control, and pilot classification. It is the complementary upper-bound component to Lemma 15.

Lemma 19. [`lem:light-cell-polynomial`] (Light-cell polynomial rate) *Fix an overlap level ϵ with $0 < \epsilon \le 1/2$. The hybrid estimator is computable in the real-arithmetic count model: there is a universal constant $K \in \mathbb{N}$, $K > 0$, such that for every n, d a real-arithmetic program using the split-count vector evaluates $\widehat{\tau}_n^{\mathrm{hyb}}$ exactly and uses at most $K\, d\, M(n)^4$ arithmetic and comparison operations.*

There are constants $C_\epsilon, c_\epsilon > 0$, depending only on ϵ, such that, for every n, d, every $P \in \mathsf{Prob}(\{1, \dots, d\} \times \{0,1\}^2)$, and every sample law μ_n, if

- ***(Experiment class.)*** *$\mu_n = P^{\otimes n}$ belongs to the overlap-restricted iid product-law experiment class $\mathcal{E}_{n,d,\epsilon}$ of Definition 2;*
- ***(Dimension range.)*** *$d \le c_\epsilon n \log n$;*

then the light-cell contribution $\widehat{T}_{\mathcal{L}}$ satisfies

$$\int \left(\widehat{T}_{\mathcal{L}}(O_1, \dots, O_n) - \sum_{k \in \widehat{\mathcal{L}}_n} \phi(q_k) \right)^2 d\mu_n \le C_\epsilon \left\{ \frac{1}{n} + \frac{d^2}{n^2 (\log n)^2} \right\}.$$

Proof of Lemma 19. The computability assertion uses the real-arithmetic program model in which an instruction is an input read, a constant, one of the arithmetic operations $+, -, \times, /$, or a nonpositive comparison branch. Input and constant instructions have cost zero, and each arithmetic or branch instruction has cost one. The arithmetic is total: division by zero is interpreted as the real value 0. A branch node is compiled as an expression-tree node containing its test and both alternatives, so its operation count is the count of the test, plus the counts of both alternatives, plus one for the branch instruction. A finite list sum $x_1 + \cdots + x_s$ is compiled recursively as $x_1 + (x_2 + \cdots + (x_s + 0) \cdots)$; hence its cost is the sum of the costs of the listed expressions plus one addition for each list entry. The corresponding list product is compiled with a terminal factor 1 and contributes one multiplication for each list entry.

The falling-factorial factors use the guarded natural-subtraction expression

$$x \mathbin{\dot{-}} i = \begin{cases} 0, & x - i \le 0, \\ x - i, & x - i > 0, \end{cases}$$

implemented by one nonpositive branch with test $x - i$ and alternatives 0 and $x - i$. On natural-valued count inputs this evaluates to the truncated difference, and its operation count is $2\,\mathrm{ops}(x) + 3$. Consequently the recursive expression $(x)_0 = 1$, $(x)_{r+1} = (x \mathbin{\dot{-}} r)(x)_r$ has cost $r\{2\,\mathrm{ops}(x) + 4\}$. A split-count read has cost zero, so a falling-factorial expression of a split count of order r costs $4r$. Thus a factorial monomial of multi-degree $|r|$, including the final division by $(m_1)_{|r|}$, costs

$$4|r| + 5.$$

For a light-polynomial term indexed by (j,t,a,y), the multi-degree is $j+2$, and multiplication by its coefficient gives cost $4j+14$. Using the list-sum convention above, summing the four cell terms costs at most $50M(n)$, summing over $t=0,\dots,j$ costs at most $51M(n)^2$, and summing over $j=0,\dots,M(n)-2$ gives one treatment-arm factorial-polynomial expression with cost at most $52M(n)^3$. The two treatment arms and their subtraction therefore give the light-cell expression cost

$$\mathrm{ops}\{\text{one light category}\} \le 105M(n)^4.$$

The heavy ratio expression has cost 11. The selected-category expression is defined by cases. If $n < N_0$, it is the heavy ratio expression; in this range $\widehat{\mathcal{H}}_n$ is the full category set and $\widehat{\mathcal{L}}_n = \varnothing$, so this expression is exact for the selected contribution. Its cost is $11 \le 110M(n)^4$. If $n \ge N_0$, the expression branches on the first-split category count minus the threshold constant. The category-count expression is the list sum of the four split-count reads and costs four additions, the subtraction of the threshold contributes one more operation, the branch contains both the light expression and the heavy ratio expression, and the branch instruction itself contributes one operation. Since $M(n) \ge 2$, this selected-category branch also satisfies

$$\mathrm{ops}\{\text{one selected category}\} \le 110M(n)^4.$$

Summing these selected-category expressions over d categories uses the same recursive list sum and adds d summation nodes. The resulting expression computes $\widehat{T}_{\mathcal{H}} + \widehat{T}_{\mathcal{L}}$ exactly, hence

$$\mathrm{ops}\{\widehat{T}_{\mathcal{H}} + \widehat{T}_{\mathcal{L}}\} \le 111\, d\, M(n)^4.$$

For $d > 0$, the final clamp to $[-1,1]$, as displayed in Definition 9, has cost

$$4\,\mathrm{ops}\{\widehat{T}_{\mathcal{H}} + \widehat{T}_{\mathcal{L}}\} + 6 \le 450dM(n)^4.$$

For $d = 0$, the compiled program is the zero-cost constant zero, which agrees with the empty-category hybrid estimator. Thus, for every n, d, there is a real-arithmetic program reading the hybrid split-count vector, returning $\widehat{\tau}_n^{\mathrm{hyb}}$ exactly on every sample, and satisfying

$$\mathrm{ops}\{\widehat{\tau}_n^{\mathrm{hyb}}\} \le 450\, d\, M(n)^4.$$

The computability clause therefore holds with the universal constant $K = 450$.

For the light-cell risk bound, fix n, d, P and a sample law μ_n satisfying the hypotheses of the lemma. By the experiment-class hypothesis in Definition 2, $\mu_n = P^{\otimes n}$ and P satisfies ϵ-overlap. In the endpoint case $n = 0$, the displayed rate is interpreted with the same total real convention as the arithmetic model: real division by zero is 0 and $\log 0 = 0$. In the calibrated range used below, $n > 0$ and $\log n > 0$, so the displayed rate has its usual analytic meaning. Set

$$R_{n,d} := \frac{1}{n} + \frac{d^2}{n^2(\log n)^2}.$$

Choose

$$K_c = 4(\mathrm{e}+1)8192^2 + 16\,8192^2, \qquad K_b = \left(\frac{65536}{\epsilon\alpha_0^2}\right)^2, \qquad K_f = 8(\mathrm{e}+1)8192^2,$$

and let

$$C_\epsilon = 3(K_c + K_b + K_f), \qquad c_\epsilon = 1.$$

The inequalities $\epsilon > 0$ and $\alpha_0 > 0$ make these constants positive. The estimates below are uniform over all d, so the dimension-range hypothesis is accommodated by this choice of c_ϵ.

First suppose $n \geq N_0$. Write

$$L = \log(en), \qquad \ell = \log n, \qquad M = M(n), \qquad B = B(n), \qquad m_0 = |I_0|, \qquad m_1 = |I_1|,$$

and define the genuinely light and false-light population sets, together with the pilot light indicator,

$$G = \{k : p_k \leq B/4\}, \qquad F = \{k : p_k > B/4\}, \qquad S_k = \mathbf{1}\{k \in \widehat{\mathcal{L}}_n\}.$$

Let $\bar{P}_k = P_{M,B}(q_k)$ be the sparse factorial-polynomial mean of $\widehat{P}_k$. Every multi-index r appearing in the expansion has $|r| \leq M$, and the cutoff inequalities give $|r| \leq M \leq m_1$. Hence the ordered-selection factorial-moment identity applies with the required denominator order:

$$\mathbb{E}_{P^{\otimes\infty}}\left[\frac{\prod_{a,y}(N^{(1)}_{aky})_{r_{ay}}}{(m_1)_{|r|}}\right] = q_k^r.$$

Substituting this identity into the polynomial expansion of $\widehat{P}_k$, equivalently applying Lemma 9, gives

$$\mathbb{E}_{P^{\otimes\infty}}[\widehat{P}_k] = \bar{P}_k = P_{M,B}(q_k).$$

Put

$$W_{\text{var}} = \sum_{k\in G} S_k(\widehat{P}_k - \bar{P}_k), \qquad W_{\text{app}} = \sum_{k\in G} S_k(\bar{P}_k - \phi(q_k)), \qquad W_{\text{sel}} = \sum_{k\in F} S_k(\widehat{P}_k - \phi(q_k)).$$

Partitioning the selected light categories into G and F, and adding and subtracting $\bar{P}_k$ on G, gives the pointwise identity

$$\widehat{T}_{\mathcal{L}} - \sum_{k\in\widehat{\mathcal{L}}_n} \phi(q_k) = W_{\text{var}} + W_{\text{app}} + W_{\text{sel}}.$$

The cutoff property in Definition 9 gives $2 \leq \alpha_0 L$, $4M^2 \leq m_1$, and $4M/m_1 \leq 3B/4$. The numerical definition of α_0 gives $\alpha_0 \leq 1/120$, since $D_A = 8\log(27/4) > 8$ and $(256 D_A)^{-1} \leq 1/120$. Combining this with $2 \leq \alpha_0 L$ yields $L \geq 240$. Therefore Lemma 7 supplies

$$0 < \ell, \qquad \ell \leq L, \qquad n \leq 2m_1, \qquad M \leq L, \qquad B \leq \frac{8192L}{n}, \qquad \Gamma^2 L^6 \leq n,$$

where $\Gamma = 6^{2M}$, and also $\Gamma \geq 1$. For every $k \in G$, the cutoff shift gives

$$p_k + \frac{4M}{m_1} \leq B.$$

Thus Lemma 16 gives, for $k \in G$,

$$\int (\widehat{P}_k - \bar{P}_k)^2 \, dP^{\otimes\infty} \leq 8(\mathrm{e}+1)(B6^M)^2.$$

For distinct $k, l \in G$, Lemma 17 gives

$$\int (\widehat{P}_k - \bar{P}_k)(\widehat{P}_l - \bar{P}_l) \, dP^{\otimes\infty} \leq \frac{8M^2}{m_1}(B6^M)^2.$$

The pilot indicators depend on the pilot split, while the centered polynomial products depend on the estimation split; their pair expectation factors from the centered product integral and lies in $[0,1]$. Expanding the square over any deterministic $\mathcal{S} \subseteq G$ yields

$$\int \left\{ \sum_{k\in\mathcal{S}} S_k(\widehat{P}_k - \bar{P}_k) \right\}^2 dP^{\otimes\infty} \le |\mathcal{S}|\, 8(\mathrm{e}+1)(B6^M)^2 + |\mathcal{S}|^2 \frac{8M^2}{m_1}(B6^M)^2.$$

Taking $\mathcal{S} = G$ and using $|G| \le d$, the diagonal rate conversion follows from

$$(L^2\Gamma\ell^2)^2 = L^4\Gamma^2\ell^4 \le n\ell^2,$$

where the inequality uses $\ell \le L$ and $\Gamma^2 L^6 \le n$. Thus $2uv \le u^2 + v^2$, with $u = d$ and $v = L^2\Gamma\ell^2$, gives

$$\frac{dL^2\Gamma}{n^2} \le \frac{1}{2}\left\{ \frac{1}{n} + \frac{d^2}{n^2\ell^2} \right\} = \frac{1}{2} R_{n,d}.$$

The cross-rate conversion uses

$$L^4\Gamma\ell^2 \le L^4\Gamma^2 L^2 = \Gamma^2 L^6 \le n,$$

and hence

$$\frac{d^2 L^4\Gamma}{n^3} \le \frac{d^2}{n^2\ell^2} \le R_{n,d}.$$

Combining these algebraic inequalities with the preceding raw second-moment bound yields

$$\int W_{\mathrm{var}}^2 \, dP^{\otimes\infty} \le K_c R_{n,d}.$$

For the deterministic approximation term, overlap gives $q_k \in \mathcal{C}_\epsilon$, and $k \in G$ gives $s(q_k) = p_k \le B/4 \le B$. Applying Lemma 10 to the four-cell vector q_k gives

$$|\bar{P}_k - \phi(q_k)| = |P_{M,B}(q_k) - \phi(q_k)| \le \frac{2B}{\epsilon M^2} \qquad (k \in G).$$

Consequently

$$|W_{\mathrm{app}}| \le d\, \frac{2B}{\epsilon M^2}.$$

The cutoff inequality $2 \le \alpha_0 L$ gives $M \ge \alpha_0 L/2$, while Lemma 7 gives $B \le 8192L/n$ and $\ell \le L$. Therefore

$$\frac{2B}{\epsilon M^2} \le \frac{65536}{\epsilon\alpha_0^2 nL} \le \frac{65536}{\epsilon\alpha_0^2 n\ell},$$

and hence, pointwise in the sample,

$$W_{\mathrm{app}}^2 \le \left(\frac{65536}{\epsilon\alpha_0^2} \right)^2 \frac{d^2}{n^2\ell^2} \le K_b R_{n,d}.$$

For the false-light term, observe that $W_{\mathrm{sel}} = \sum_{k\in F} S_k(\widehat{P}_k - \phi(q_k))$ is exactly the selected false-light error of Lemma 18: the summation there ranges over the categories with $p_k > B/4$, which is the set F. Its hypotheses hold in the present case: the overlap condition $\epsilon \le \pi_k \le 1-\epsilon$ holds for every category of positive mass because $P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}$, the calibration cutoff $n \ge N_0$ is the standing assumption of this case, and $L \geq 240$ was derived above from the cutoff property and the numerical value of α_0. Since W_{sel} is a function of the first n observations, Lemma 20 makes its second moment under $P^{\otimes\infty}$ equal to its second moment under $P^{\otimes n}$, and Lemma 18 gives directly

$$\int W_{\mathrm{sel}}^2 \, dP^{\otimes\infty} = E_{P^{\otimes n}}\left[W_{\mathrm{sel}}^2\right] \leq 8(\mathrm{e}+1)8192^2 \left\{\frac{1}{n} + \frac{d^2}{n^2\ell^2}\right\} = K_f R_{n,d}.$$

The bound for W_{var} is obtained on the infinite-product representation of the iid sample and transported to $P^{\otimes n}$ by Lemma 20, which identifies the law of the first n coordinates and hence equates the integrals of functions of those coordinates. Using $(x+y+z)^2 \leq 3(x^2+y^2+z^2)$, the pointwise bound on W_{app}^2, and the two second-moment bounds, we obtain

$$\begin{aligned}\int \left\{\widehat{T}_{\mathcal{L}} - \sum_{k\in\widehat{\mathcal{L}}_n} \phi(q_k)\right\}^2 dP^{\otimes n} &\leq 3\int (W_{\mathrm{var}}^2 + W_{\mathrm{app}}^2 + W_{\mathrm{sel}}^2)\, dP^{\otimes n} \\ &\leq 3(K_c + K_b + K_f)R_{n,d} = C_\epsilon R_{n,d}.\end{aligned}$$

Since $\mu_n = P^{\otimes n}$, this is the desired bound under μ_n.

If $n < N_0$, then $\widehat{\mathcal{L}}_n = \varnothing$ by Definition 9. The light contribution and its selected target are both zero, so the left-hand side is zero. Since $R_{n,d} \geq 0$, the same bound holds in this case. □

Lemma 19 establishes that the count-based hybrid estimator is computable with operation count linear in d and polynomial in $M(n)$, and that the light-cell contribution attains the same rate as the heavy-cell branch under the displayed dimension range. Together with Lemma 15, it gives the upper envelope used in Theorem 1 and Theorem 2.

B Verification note

This appendix records the verification boundary for the finite-alphabet product-law formulation used throughout the paper. The formal development represents the observed-data experiment in two compatible ways: a finite single-observation law for $O = (X, A, Y)$, as used in Assumption 1 and Definition 2, and an ambient infinite iid sample sequence from which finite sample segments can be extracted. The following statement gives the bridge between those two views.

Lemma 20. [`lem:ambient-sample-marginal`] (Ambient sample marginal law) *Let P be a probability law on a measurable single-observation space, and let $n \in \mathbb{N}$. Write $P^{\otimes\infty}$ for the infinite product law on sequences $(O_i)_{i\geq 0}$, and write $P^{\otimes n}$ for the finite product law on n-tuples. Then the coordinate restriction map*

$$(O_i)_{i\geq 0} \longmapsto (O_i)_{i\in\{0,\ldots,n-1\}}$$

pushes $P^{\otimes\infty}$ forward to $P^{\otimes n}$:

$$(P^{\otimes\infty}) \circ \left((O_i)_{i\geq 0} \mapsto (O_i)_{i\in\{0,\ldots,n-1\}}\right)^{-1} = P^{\otimes n}.$$

In particular, for the observed-data laws P used in Assumption 1, the initial sample segment has the n-fold product law $P^{\otimes n}$ appearing in Definition 2.

Proof of Lemma 20. Let

$$\Omega = X^{\mathbb{N}}, \qquad \mu_\infty = P^{\otimes\infty}, \qquad Z_i(\omega) = \omega_i.$$

By the construction of the infinite product measure, the coordinate process $(Z_i)_{i\ge 0}$ is independent and each coordinate has marginal law P:

$$(Z_i)_{\#}\mu_\infty = P \quad \text{for every } i \in \mathbb{N}.$$

For fixed n, define the finite coordinate map

$$C_n(\omega) = \big(Z_i(\omega)\big)_{i\in \mathrm{Fin}(n)} = \big(\omega_i\big)_{i=0}^{n-1}.$$

The finite-dimensional law of an independent process is the product of its one-dimensional marginals, hence

$$(C_n)_{\#}\mu_\infty = \bigotimes_{i\in\mathrm{Fin}(n)} (Z_i)_{\#}\mu_\infty = \bigotimes_{i\in\mathrm{Fin}(n)} P = P^{\otimes n}.$$

This is exactly

$$(P^{\otimes\infty}) \circ C_n^{-1} = P^{\otimes n}.$$

Taking X to be the observed-data alphabet and P to be the single-observation law in Assumption 1 gives the n-fold product law used in Definition 2. □

Lemma 20 justifies using an infinite iid product space as an ambient construction while retaining the finite-sample experiment class in Definition 2 as the law governing $O_1, \ldots, O_n$. This connection is useful for statements that are naturally organized around coordinate projections, while the statistical risk in Definition 6 remains the finite-n mean-squared risk under $P^{\otimes n}$.

The verification covers the finite-alphabet iid observational model, the potential-outcome overlay, the overlap-restricted four-cell cone, the cell functional, the ATE functional, the minimax risk, the deterministic split counts, and the hybrid estimator defined in Assumptions 1 and 2 and Definitions 3 to 7 and 9. The auxiliary proof layer verifies the estimator, approximation, calibration, residual, missing-arm, factorial-moment, and aggregation statements used by the main rate theorems. The theorem statements and the proofs written in this paper are machine-checked in Lean 4. That check covers the steps carried out here, including each step in which a published conclusion is invoked, and it stops at the boundary of any such conclusion: the source proof behind a cited input is not itself formalized, and a theorem depending on one is certified conditionally on it. The theorem-local verification footnotes name the published conclusions entering in this way, so a reader can see for each result which published input, if any, its certificate rests on.

C Proofs of the main results

Proof of Theorem 1. Fix an overlap level $0 < \epsilon < 1/2$, and use the theorem's cited fixed-sample control-zero one-arm minimax lower-bound input at this overlap level. Write

$$r_{n,d} = \frac{1}{n} + \frac{d^2}{n^2(\log n)^2}.$$

1. First derive the fixed-interior upper envelope for the hybrid estimator. From Lemma 19 choose constants $\mathfrak{C}_{\mathcal{L}}, \mathfrak{c}_{\mathcal{L}} > 0$ such that, whenever $P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}$ and $d \le \mathfrak{c}_{\mathcal{L}} n \log n$,

$$E_{P^{\otimes n}}\left[\left\{\widehat{T}_{\mathcal{L}} - \sum_{k \in \widehat{\mathcal{L}}_n} \phi(q_k)\right\}^2\right] \le \mathfrak{C}_{\mathcal{L}} r_{n,d}.$$

From Lemma 15 choose $\mathfrak{C}_{\mathcal{H}}, \mathfrak{c}_{\mathcal{H}} > 0$ and $\mathfrak{N}_{\mathcal{H}} < \infty$ such that, whenever $n \ge \mathfrak{N}_{\mathcal{H}}$, $P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}$, and $d \le \mathfrak{c}_{\mathcal{H}} n \log n$,

$$E_{P^{\otimes n}}\left[\left\{\widehat{T}_{\mathcal{H}} - \sum_{k \in \widehat{\mathcal{H}}_n} \phi(q_k)\right\}^2\right] \le \mathfrak{C}_{\mathcal{H}} r_{n,d}.$$

Set

$$\mathfrak{C}_{\text{up}} = 2(\mathfrak{C}_{\mathcal{L}} + \mathfrak{C}_{\mathcal{H}}), \qquad \rho_{\text{up}} = \min\{\mathfrak{c}_{\mathcal{L}}, \mathfrak{c}_{\mathcal{H}}\}, \qquad \mathfrak{N}_{\text{up}} = \mathfrak{N}_{\mathcal{H}}.$$

For a sample array $o_{1:n}$, define the pilot targets

$$\mathcal{T}^{\circ}_{\mathcal{H}}(P, o_{1:n}) = \sum_{k \in \widehat{\mathcal{H}}_n(o_{1:n})} \phi(q_k), \qquad \mathcal{T}^{\circ}_{\mathcal{L}}(P, o_{1:n}) = \sum_{k \in \widehat{\mathcal{L}}_n(o_{1:n})} \phi(q_k).$$

The sets $\widehat{\mathcal{H}}_n(o_{1:n})$ and $\widehat{\mathcal{L}}_n(o_{1:n})$ partition $\{1, \dots, d\}$, hence

$$\mathcal{T}^{\circ}_{\mathcal{H}}(P, o_{1:n}) + \mathcal{T}^{\circ}_{\mathcal{L}}(P, o_{1:n}) = \sum_{k=1}^{d} \phi(q_k) = \tau(P),$$

where the last equality is Definition 5. The interval bound needed for truncation follows from the cell formula. For each category, write

$$p_k = \sum_{a,y \in \{0,1\}} q_{aky}, \qquad s_{ak} = q_{ak0} + q_{ak1}, \qquad \mu_{ak} = q_{ak1}/s_{ak},$$

with the empty-arm value interpreted by the totalized convention $\mu_{ak} = 0$ when $s_{ak} = 0$. By Definition 4, the zero vector has $\phi(0) = 0$, so a zero-mass category contributes $0 = p_k\{\mu_{1k} - \mu_{0k}\}$. On a positive-mass category satisfying overlap, the two arm masses are positive, and the displayed formula in Definition 4 gives

$$\phi(q_k) = p_k \left\{\frac{q_{1k1}}{s_{1k}} - \frac{q_{0k1}}{s_{0k}}\right\} = p_k\{\mu_{1k} - \mu_{0k}\}.$$

Summing this identity and using Definition 5 gives

$$\tau(P) = \sum_{k=1}^{d} p_k\{\mu_{1k} - \mu_{0k}\}.$$

Here $p_k \ge 0$, $\sum_{k=1}^{d} p_k = 1$, and the totalized binary-outcome means satisfy $0 \le \mu_{ak} \le 1$. Therefore $|\mu_{1k} - \mu_{0k}| \le 1$ for every k, and

$$|\tau(P)| \le \sum_{k=1}^{d} p_k |\mu_{1k} - \mu_{0k}| \le \sum_{k=1}^{d} p_k = 1.$$

Thus $\tau(P) \in [-1, 1]$. Let

$$\Pi_{[-1,1]}(x) = \max\{-1, \min\{1, x\}\}.$$

The elementary projection inequality

$$\{\Pi_{[-1,1]}(x) - \tau(P)\}^2 \le \{x - \tau(P)\}^2$$

holds for every real x. Applying it to $x = \widehat{T}_{\mathcal{H}}(o_{1:n}) + \widehat{T}_{\mathcal{L}}(o_{1:n})$, and then using $(u + v)^2 \le 2u^2 + 2v^2$, gives the pointwise bound

$$\left(\widehat{\tau}_n^{\text{hyb}} - \tau(P)\right)^2 \le 2\left\{\widehat{T}_{\mathcal{H}} - \mathcal{T}^{\circ}_{\mathcal{H}}(P, \cdot)\right\}^2 + 2\left\{\widehat{T}_{\mathcal{L}} - \mathcal{T}^{\circ}_{\mathcal{L}}(P, \cdot)\right\}^2.$$

Integrating with respect to $P^{\otimes n}$ and using the two component bounds yields, for $n \ge \mathfrak{N}_{\text{up}}$ and $d \le \rho_{\text{up}} n \log n$,

$$E_{P^{\otimes n}}\left[\left(\widehat{\tau}_n^{\text{hyb}} - \tau(P)\right)^2\right] \le \mathfrak{C}_{\text{up}} r_{n,d}.$$

Taking the supremum over $P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}$ gives the same bound for the worst-case risk of $\widehat{\tau}_n^{\text{hyb}}$. Since this estimator is a measurable estimator on the finite sample space and hence is one admissible estimator in the infimum defining $\mathsf{R}_{n,d,\epsilon}$ in Definition 6,

$$\mathsf{R}_{n,d,\epsilon} \le \sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} E_{P^{\otimes n}}\left[\left(\widehat{\tau}_n^{\text{hyb}} - \tau(P)\right)^2\right] \le \mathfrak{C}_{\text{up}} r_{n,d}.$$

2. The lower-transfer step is supplied by Lemma 2, applied with the cited fixed-sample control-zero one-arm lower-bound input of Zeng et al. [2024]. Thus there are $\mathfrak{a}_{\text{low}}, \rho_{\text{low}} > 0$ and $\mathfrak{N}_{\text{low}} < \infty$ such that, for $0 < d$, $n \ge \mathfrak{N}_{\text{low}}$, and $d \le \rho_{\text{low}} n \log n$,

$$\mathfrak{a}_{\text{low}} r_{n,d} \le \mathsf{R}_{n,d,\epsilon}.$$

Define

$$a_\epsilon = \mathfrak{a}_{\text{low}}, \qquad \rho_\epsilon = \min\{\rho_{\text{up}}, \rho_{\text{low}}\}, \qquad C_\epsilon = \mathfrak{C}_{\text{up}}, \qquad N_\epsilon = \max\{\mathfrak{N}_{\text{up}}, \mathfrak{N}_{\text{low}}\}.$$

These constants are positive, with $N_\epsilon < \infty$. If $0 < d$, $n \ge N_\epsilon$, and $d \le \rho_\epsilon n \log n$, then both the upper and lower ranges apply, so

$$a_\epsilon r_{n,d} \le \mathsf{R}_{n,d,\epsilon} \le \sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} E_{P^{\otimes n}}\left[\left(\widehat{\tau}_n^{\text{hyb}} - \tau(P)\right)^2\right] \le C_\epsilon r_{n,d}.$$

This is the sharp sandwich assertion.

3. Fix $K > 0$. Set

$$C_K = C_\epsilon(1 + K^2).$$

Then $C_K > 0$. Suppose $0 < d$, $n \ge N_\epsilon$, $d \le K\sqrt{n} \log n$, and $d \le \rho_\epsilon n \log n$. The sample size is positive by the displayed range: if $n = 0$, then $K\sqrt{n} \log n = 0$, contradicting $0 < d \le K\sqrt{n} \log n$. If $\log n \le 0$, then $K\sqrt{n} \ge 0$ and therefore $K\sqrt{n} \log n \le 0$, again contradicting $0 < d \le K\sqrt{n}\log n$. Thus $\log n > 0$. With both sides of $d \le K\sqrt{n}\log n$ nonnegative, squaring gives

$$d^2 \le K^2 n(\log n)^2.$$

Dividing by the positive denominator $n^2(\log n)^2$ gives

$$\frac{d^2}{n^2(\log n)^2} \le \frac{K^2}{n}.$$

Hence

$$r_{n,d} = \frac{1}{n} + \frac{d^2}{n^2(\log n)^2} \le \frac{1+K^2}{n}.$$

The lower side of the sandwich gives

$$\frac{a_\epsilon}{n} \le a_\epsilon r_{n,d} \le \mathsf{R}_{n,d,\epsilon},$$

and the upper side gives

$$\mathsf{R}_{n,d,\epsilon} \le C_\epsilon r_{n,d} \le \frac{C_\epsilon(1+K^2)}{n} = \frac{C_K}{n}.$$

This proves the parametric interior assertion.

4. Let n_j, d_j be sequences with $n_j \to \infty$, with $d_j > 0$ eventually, and with

$$d_j \le \rho_\epsilon n_j \log n_j$$

eventually. Finite initial indices leave convergence unchanged, so work on the tail where $n_j \ge N_\epsilon$, $n_j \ge 2$, $d_j > 0$, and the displayed range condition holds. Define

$$\eta_j = \frac{d_j}{n_j \log n_j}, \qquad r_j = \frac{1}{n_j} + \eta_j^2.$$

On this tail, $r_j = r_{n_j,d_j}$ and $\eta_j \ge 0$. The overlap class is nonempty for such d_j: since $d_j > 0$, the endpoint parametric law on $\{1,\dots,d_j\}$ with uniform category masses, treatment probability $1/2$, and both binary response means $1/2$ has propensity $1/2$ in every positive-mass category, and therefore satisfies the ϵ-overlap condition because $0 < \epsilon < 1/2$. Applying the sandwich on the tail gives

$$a_\epsilon r_j \le \mathsf{R}_{n_j,d_j,\epsilon} \le C_\epsilon r_j.$$

If $\mathsf{R}_{n_j,d_j,\epsilon} \to 0$, then

$$0 \le r_j \le a_\epsilon^{-1}\mathsf{R}_{n_j,d_j,\epsilon}$$

eventually, so $r_j \to 0$. Since

$$0 \le \eta_j^2 \le r_j,$$

we have $\eta_j^2 \to 0$, and the eventual nonnegativity of η_j gives $\eta_j \to 0$. Conversely, if $\eta_j \to 0$, then $1/n_j \to 0$ and $\eta_j^2 \to 0$, so $r_j \to 0$. Since minimax risks are nonnegative, the eventual bound

$$0 \le \mathsf{R}_{n_j,d_j,\epsilon} \le C_\epsilon r_j$$

then yields $\mathsf{R}_{n_j,d_j,\epsilon} \to 0$. Therefore

$$\mathsf{R}_{n_j,d_j,\epsilon} \to 0 \quad \Longleftrightarrow \quad \frac{d_j}{n_j \log n_j} \to 0.$$

$\square$

Proof of Theorem 2. 1. **Computability.** Use the real-arithmetic program model in which a program is a finite register list with input reads, real constants, additions, subtractions, multiplications, divisions, and branch comparisons. Input reads and constants have cost zero, while each arithmetic operation and comparison has cost one. Division is the total real-field operation, so division by a zero denominator returns zero. A branch comparison has a test expression and two branch expressions and returns the first branch when the test value is nonpositive and the second branch otherwise.

For $n, d \in \mathbb{N}$, the input alphabet for the program is

$$\{0,1\} \times \{1, \dots, d\} \times \{0,1\} \times \{0,1\}.$$

At input coordinate (j, k, a, y) the program reads the real number

$$N_{aky}^{(j)} = \#\{i \in I_j : (X_i, A_i, Y_i) = (k, a, y)\}.$$

Equivalently, the input vector is the collection of the $8d$ split cell counts $N_{aky}^{(j)}$. Sums and products below are compiled as expression trees folded onto the initial constants 0 and 1: a list sum has cost equal to the sum of the costs of its entries plus one addition for each entry, and a list product has cost equal to the sum of the costs of its entries plus one multiplication for each entry. Thus a four-term category count has cost 4, not 3.

If $d = 0$, the program is the constant-zero program. Its operation count is zero, and its value agrees with $\widehat{\tau}_n^{\mathrm{hyb}}$, since the heavy and light sums range over the empty alphabet and the truncation of zero is zero.

Now suppose $d > 0$, and write $M = M(n)$. A split-count input has cost zero. For a natural-valued expression x, the truncated subtraction $(x - i)_+$ is compiled as a branch on $x - i$, returning 0 on the nonpositive branch and $x - i$ on the positive branch. This expression has cost $2\,\mathrm{op}(x) + 3$. The falling factorial $x(x-1)_+ \cdots (x-r+1)_+$ is compiled by the corresponding folded product, so its cost is $r\{2\,\mathrm{op}(x) + 4\}$. Therefore, for a split-count input x, the falling-factorial expression of order r has cost $4r$.

For a multi-index r, the factorial monomial expression is

$$\frac{\prod_{a,y} (N_{aky}^{(1)})_{r_{ay}}}{(m_1)_{|r|}},$$

with the same total division convention as the program model. The folded product over the four cells contributes $4|r| + 4$ operations, and the final division by the constant denominator contributes one more, giving

$$\mathrm{op}\{\text{factorial monomial } r\} = 4|r| + 5.$$

For the multi-indices used in the arm expansion of $P_{M,B}$, $|r| = j + 2$. After multiplication by its coefficient, one fixed-cell summand at level (j, t) has cost

$$4j + 14.$$

Since $M \geq 2$, the nested sums in one treatment arm satisfy

$$\mathrm{op}\{\text{sum over } (a, y) \text{ at fixed } j, t\} \leq 50M,$$

$$\mathrm{op}\{\text{sum over } t \text{ at fixed } j\} \leq 51M^2, \qquad \mathrm{op}\{\text{one arm sum}\} \leq 52M^3.$$

Subtracting the two arm sums gives the light-cell expression, and hence

$$\mathrm{op}\{\text{light-cell expression for category } k\} \leq 105M^4.$$

The heavy expression for category k is

$$\frac{N_k^{(1)}}{m_1} \left\{ \frac{N_{1k1}^{(1)}}{N_{1k0}^{(1)} + N_{1k1}^{(1)}} - \frac{N_{0k1}^{(1)}}{N_{0k0}^{(1)} + N_{0k1}^{(1)}} \right\},$$

where $N_k^{(1)} = \sum_{a,y} N_{aky}^{(1)}$, again using total division at zero denominators. This expression has cost 11 and agrees exactly with $\widehat{h}_k$. The pilot count expression $\sum_{a,y} N_{aky}^{(0)}$ has cost 4, so the comparison expression $\sum_{a,y} N_{aky}^{(0)} - \lambda_0 L$ has cost 5. For $n \geq N_0$, the selected category expression branches to the light expression when the pilot count is at most $\lambda_0 L$ and to the heavy expression when it is larger. Its cost is at most

$$5 + 105M^4 + 11 + 1 \leq 110M^4,$$

where the final 1 is the branch comparison itself. For $n < N_0$, the selected category expression is the heavy expression, whose cost 11 is also at most $110M^4$. Thus each selected category expression has operation count at most $110M(n)^4$.

Summing the d selected category expressions uses one addition node per category in the folded sum, so the untruncated expression has cost at most

$$111\, d\, M(n)^4.$$

The truncation

$$x \longmapsto \max\{-1, \min\{1, x\}\}$$

is compiled by two branch comparisons, with the relevant expression duplicated in the branch subtrees. Consequently

$$\mathrm{op}\{\mathrm{clamp}(x)\} = 4\, \mathrm{op}(x) + 6.$$

For $d > 0$, this gives

$$\mathrm{op}(\mathcal{A}_{n,d}) \leq 4 \cdot 111\, d\, M(n)^4 + 6 \leq 450\, d\, M(n)^4,$$

because $1 \leq dM(n)^4$. Together with the zero-alphabet branch, the compiled program $\mathcal{A}_{n,d}$ satisfies

$$\mathrm{op}(\mathcal{A}_{n,d}) \leq 450\, d\, M(n)^4$$

and, for every sample,

$$\mathcal{A}_{n,d}\big((N_{aky}^{(j)})_{j,k,a,y}\big) = \widehat{\tau}_n^{\mathrm{hyb}}(O_{1:n}).$$

Taking $K = 450$ proves the computability clause.

2. **Fixed-interior hybrid bound.** Fix $0 < \epsilon < 1/2$. Let $C_{\mathrm{L}}, c_{\mathrm{L}} > 0$ be the constants supplied by Lemma 19, and let $C_{\mathrm{H}}, \rho_{\mathrm{H}} > 0$ and $N_{\mathrm{H}} \in \mathbb{N}$ be the constants supplied by Lemma 15. Set

$$C_\epsilon = 2(C_{\mathrm{L}} + C_{\mathrm{H}}), \qquad \rho_\epsilon = \min\{c_{\mathrm{L}}, \rho_{\mathrm{H}}\}, \qquad N_\epsilon = N_{\mathrm{H}}.$$

These constants are positive in the required real coordinates.

For a sample $z = O_{1:n}$, define

$$T^\circ_{\mathcal{H}}(P, z) = \sum_{k \in \widehat{\mathcal{H}}_n(z)} \phi(q_k), \qquad T^\circ_{\mathcal{L}}(P, z) = \sum_{k \in \widehat{\mathcal{L}}_n(z)} \phi(q_k).$$

Since $\widehat{\mathcal{L}}_n(z) = \widehat{\mathcal{H}}_n(z)^c$,

$$T^\circ_{\mathcal{H}}(P, z) + T^\circ_{\mathcal{L}}(P, z) = \tau(P).$$

Also $\tau(P) \in [-1, 1]$ for every law in the overlap class. Hence the truncation map is contractive around $\tau(P)$, and

$$\begin{aligned}\left(\widehat{\tau}^{\mathrm{hyb}}_n(z) - \tau(P)\right)^2 &\le \left[\{\widehat{T}_{\mathcal{H}}(z) - T^\circ_{\mathcal{H}}(P, z)\} + \{\widehat{T}_{\mathcal{L}}(z) - T^\circ_{\mathcal{L}}(P, z)\}\right]^2 \\ &\le 2\{\widehat{T}_{\mathcal{H}}(z) - T^\circ_{\mathcal{H}}(P, z)\}^2 + 2\{\widehat{T}_{\mathcal{L}}(z) - T^\circ_{\mathcal{L}}(P, z)\}^2.\end{aligned}$$

Integrating under any $P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}$ gives

$$\begin{aligned}E_P\left[\left(\widehat{\tau}^{\mathrm{hyb}}_n - \tau(P)\right)^2\right] &\le 2C_{\mathrm{H}}\left\{\frac{1}{n} + \frac{d^2}{n^2(\log n)^2}\right\} + 2C_{\mathrm{L}}\left\{\frac{1}{n} + \frac{d^2}{n^2(\log n)^2}\right\} \\ &= C_\epsilon\left\{\frac{1}{n} + \frac{d^2}{n^2(\log n)^2}\right\},\end{aligned}$$

whenever $0 < n$, $N_\epsilon \le n$, and $d \le \rho_\epsilon n \log n$. The light and heavy bounds apply because this dimension condition implies both

$$d \le c_{\mathrm{L}} n \log n \quad \text{and} \quad d \le \rho_{\mathrm{H}} n \log n.$$

Taking the supremum over $P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}$ proves the fixed-interior calibration bound for $\widehat{\tau}^{\mathrm{hyb}}_n$. In the auxiliary zero-alphabet case used by the selector lemma, the experiment class is empty and the displayed upper bound follows from the empty-supremum convention and nonnegativity of the rate.

3. **Selector envelope.** Apply Lemma 6 with the constants just constructed and with the hybrid bound from the previous step. For every $n, d \in \mathbb{N}$ with $0 < n$, $0 < d$, $N_\epsilon \le n$, and $d \le \rho_\epsilon n \log n$, the selected estimator that chooses $\widehat{\tau}^{\mathrm{hyb}}_n$ when

$$C_\epsilon\left\{\frac{1}{n} + \frac{d^2}{n^2(\log n)^2}\right\} \le \frac{1}{n} + 4(1/2 - \epsilon)^2$$

and chooses $\widehat{\tau}_{\mathrm{ctr}}$ otherwise satisfies

$$\mathsf{R}_{n,d,\epsilon} \le \sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} E_P\left[\left(\widehat{\tau}^{\mathrm{sel}}_{C_\epsilon,\epsilon} - \tau(P)\right)^2\right]$$

and

$$\sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} E_P\left[\left(\widehat{\tau}^{\mathrm{sel}}_{C_\epsilon,\epsilon} - \tau(P)\right)^2\right] \le \max\{C_\epsilon, 4\}\left[\frac{1}{n} + \min\left\{\frac{d^2}{n^2(\log n)^2}, (1/2 - \epsilon)^2\right\}\right].$$

4. **Centered randomization bound.** Fix $0 < \epsilon \le 1/2$, $0 < n$, and $0 < d$. For every $P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}$, Lemma 3 gives

$$E_P\left[(\widehat{\tau}_{\mathrm{ctr}} - \tau(P))^2\right] \le \frac{1}{n} + 4(1/2 - \epsilon)^2.$$

The right-hand side is independent of P, so taking the supremum over the experiment class gives

$$\sup_{P^{\otimes n} \in \mathcal{E}_{n,d,\epsilon}} E_P\left[(\widehat{\tau}_{\mathrm{ctr}} - \tau(P))^2\right] \le \frac{1}{n} + 4(1/2 - \epsilon)^2.$$

The same conclusion covers the empty-class branch through the same empty-supremum convention.

5. **Endpoint bracket.** For $0 < n$ and $0 < d$, Lemma 5 gives directly

$$\frac{1}{100n} \le \mathsf{R}_{n,d,1/2} \le \frac{1}{n}.$$

This is the endpoint bracket.

□

References

Sivaraman Balakrishnan, Edward H. Kennedy, and Larry Wasserman. The fundamental limits of structure-agnostic functional estimation. *arXiv preprint arXiv:2305.04116*, 2023. doi: 10.48550/arXiv.2305.04116. To appear in Statistical Science.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

Alexandre Belloni, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017. doi: 10.3982/ECTA12723.

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

T. Tony Cai and Mark G. Low. Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011. doi: 10.1214/10-AOS849.

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. doi: 10.1111/ectj.12097.

Victor Chernozhukov, Whitney K. Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022. doi: 10.3982/ECTA18515.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021. doi: 10.1016/j.jeconom.2019.10.014.

Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447, 1976. doi: 10.1093/biomet/63.3.435.

Ronald A. Fisher, A. Steven Corbet, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 12(1):42–58, 1943. doi: 10.2307/1411.

I. J. Good and G. H. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1/2):45–63, 1956. doi: 10.1093/biomet/43.1-2.45.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998. doi: 10.2307/2998560.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under <inline-formula> <tex-math notation=”latex”>ℓ_1 </tex-math></inline-formula> loss. *IEEE Transactions on Information Theory*, 61(11): 6343–6354, 2015. doi: 10.1109/TIT.2015.2478816.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of divergences between discrete distributions. *IEEE Journal on Selected Areas in Information Theory*, 1(3):814–823, 2020. doi: 10.1109/JSAIT.2020.3041036.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003. doi: 10.1111/1468-0262.00442.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015. doi: 10.1109/TIT.2015.2412945.

Jikai Jin and Vasilis Syrgkanis. Structure-agnostic optimality of doubly robust learning for treatment effect estimation. *arXiv preprint arXiv:2402.14264*, 2025. doi: 10.48550/arXiv.2402.14264. To appear in COLT 2025.

Edward H. Kennedy, Sivaraman Balakrishnan, James M. Robins, and Larry Wasserman. Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, 52(2):793–816, 2024. doi: 10.1214/24-AOS2369.

O. Lepski, A. Nemirovski, and V. Spokoiny. On estimation of the l r norm of a regression function. *Probability Theory and Related Fields*, 113(2):221–253, 1999. doi: 10.1007/s004409970006.

Lin Liu, Rajarshi Mukherjee, Whitney K. Newey, and James M. Robins. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*, 2017. doi: 10.48550/arXiv.1705.07577.

Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016. doi: 10.1073/pnas.1607774113.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6): 1191–1253, 2003. doi: 10.1162/089976603321780272.

James M. Robins and Ya'acov Ritov. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319, 1997. doi: 10.1002/(SICI)1097-0258(19970215)16:3⟨285::AID-SIM535⟩3.0.CO;2-\#.

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. doi: 10.1080/01621459.1994.10476818.

James M. Robins, Lingling Li, Eric J. Tchetgen Tchetgen, and Aad W. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, volume 2 of *Institute of Mathematical Statistics Collections*, pages 335–421. Institute of Mathematical Statistics, Beachwood, OH, 2008. doi: 10.1214/193940307000000527.

James M. Robins, Eric J. Tchetgen Tchetgen, Lingling Li, and Aad W. van der Vaart. Semiparametric minimax rates. *Electronic Journal of Statistics*, 3:1305–1321, 2009. doi: 10.1214/09-EJS479.

James M. W. Robins, Lingling Li, Rajarshi Mukherjee, Eric J. Tchetgen Tchetgen, and Aad van der Vaart. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017. doi: 10.1214/16-AOS1515.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009. doi: 10.1007/b13794.

Gregory Valiant and Paul Valiant. Estimating the unseen. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, pages 685–694, 2011. doi: 10.1145/1993636.1993727.

Gregory Valiant and Paul Valiant. Estimating the unseen. *Journal of the ACM*, 64(6):37:1–37:41, 2017. doi: 10.1145/3125643.

Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006. doi: 10.2202/1557-4679.1043.

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

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016. doi: 10.1109/TIT.2016.2548468.

Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883, 2019. doi: 10.1214/17-AOS1665.

Zhenghao Zeng, Sivaraman Balakrishnan, Yanjun Han, and Edward H. Kennedy. Causal inference with high-dimensional discrete covariates. *arXiv preprint arXiv:2405.00118*, 2024. doi: 10.48550/arXiv.2405.00118. Version 3.