PRESERVE THE TARGET, RELAX THE SHARPNESS: CERTIFIED COMPRESSION FOR SCALABLE INSTRUMENTAL VARIABLE PARTIAL IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sharp partial identification protects against false certainty, but exact causal representations can become large as outcome support grows. Naive coarsening reduces computation by changing the observed representation and can silently change the estimand or exclude data-compatible causal effects. We introduce a different computational primitive for finite instrumental-variable (IV) models: *certified target-preserving stochastic representation reduction*. A column-stochastic outcome map is lifted independently to every potential-outcome coordinate while treatment-response types are left unchanged. If the map preserves outcome values in expectation, observation commutes with the lift, every potential-outcome mean and affine mean contrast is preserved model-by-model, treatment-response restrictions are retained, and the original identified set is contained in the compressed set for arbitrary finite instrument and treatment cardinalities. Within the shared product-kernel class we characterize necessity and local uniqueness; with endpoint retention, nontrivial deterministic support reduction cannot satisfy the same exact mean-preservation certificate, so stochasticity is structural rather than cosmetic. We prove conditional sharpness bounds and a negative theorem excluding any support-uniform mesh-only rate, then turn nested support refinement into an anytime certified algorithm whose exact surrogate sets tighten monotonically to the original finite-support problem. A running-intersection rule preserves this monotonicity when the downstream causal optimizer is itself only outer-certified. Empirically, containment holds in all 13,660 generalized validation cells and all 7,692 nested refinement states. Adaptive allocation gives modest gains on nonadversarial cases and larger gains on an adversarial family, where observable-displacement-guided refinement reduces median width inflation at retained support size $M = 32$ from 7.04% to 5.90%. Fresh-process one-thread timing shows representative multivalued-IV speedups of 8.0–$18.7\times$, while an original outcome support of size $n = 512$ compressed to $M = 32$ gives a median $236\times$ speedup against the fastest certified exact backend. A nonasymptotic finite-sample extension and a public Job Training Partnership Act (JTPA) benchmark retain the same one-sided validity. We do not evade the exponential barrier for exact sharp analytical IV representations; we change the objective so that computation is traded against *sharpness*, not against the scientific target or validity.

1 THE PROBLEM: WHAT MAY BE COMPRESSED WITHOUT CHANGING THE CONCLUSION?

Partial identification reports the set of causal effects compatible with the observed data and maintained assumptions when the data do not justify a unique causal effect (Manski, 1990; 2003). An identified set is *sharp* when it contains exactly all such compatible target values, rather than merely a valid outer set; this avoids replacing genuine causal ambiguity by an unjustified point estimate. Instrumental-variable (IV) models with imperfect compliance are a canonical example (Imbens & Angrist, 1994; Balke & Pearl, 1997). Yet the representation required to preserve *sharpness* can itself become the bottleneck when outcomes have many ordered support points, as happens after fine discretization

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

of continuous measurements, with rich ordinal outcomes, or when the same causal model must be queried repeatedly at multiple resolutions.

There are two distinct computational obstacles. Pointwise sharp bounds can require large latent-response linear programs (LPs). Separately, exact *analytical* sharp bounds may have unavoidable output-size growth. For a binary instrument Z, binary treatment D, and outcome support of size n, Boushehrian et al. show that the exact sharp average treatment effect (ATE) formula contains

$$N(n) = 5\,4^{n-1} - 2^{n+2} + 4 \tag{1}$$

linear pieces and establish exponential lower bounds for exact analytical representations (Boushehrian et al., 2026). This does not say that every pointwise LP takes exponential time. It says that an exact sharp symbolic representation cannot generally be made small.

The natural computational response is to reduce outcome resolution. The scientific danger is that ordinary binning or nearest-representative rounding changes the observed law and typically changes outcome means—and therefore mean-based causal estimands—themselves; a smaller problem can then answer a *different* causal question. The key question is therefore not simply whether the support can be compressed, but:

> *Can we reduce the causal representation while certifying that no original target value is lost and the target itself is unchanged model-by-model?*

We answer yes for finite IV models. Our asymmetry is deliberate: preserve the estimand exactly, preserve validity through outer identified-set inclusion, and spend approximation error only on sharpness. This converts compression from an uncontrolled preprocessing choice into a causal certificate.

Why the gap is not ordinary quantization or causal abstraction. Barycentric/dual quantization already provides randomized expectation-preserving interpolation (Pagès & Wilbertz, 2012); we do not claim that numerical primitive as new. Causal abstraction already studies exact and lossy transformations between causal models (Rubenstein et al., 2017; Beckers & Halpern, 2019; Geiger et al., 2025; Xia & Bareinboim, 2025; Buchholz et al., 2024). Those frameworks principally ask whether a higher-level model reproduces a class of low-level causal queries. Partial-identification methods and confidence-set procedures, meanwhile, typically take the representation as fixed (Chernozhukov et al., 2007; Rosen, 2008; Song et al., 2024). The missing link is a *target-specific representation theorem*: a map on observed outcomes together with an explicit latent causal lift that preserves a designated target, preserves structural restrictions, and certifies one-sided identified-set validity while exposing exactly where sharpness can be lost. To our knowledge, this combination has not been established for scalable finite-IV partial identification.

Contributions. Our contributions are: **(i) Finite-IV causal lift.** For arbitrary finite $|Z| = K$ and $|D| = J$, we construct a product lift proving observation commutation, exact preservation of every $\mathbb{E}[Y_d]$ and affine mean contrast, preservation of treatment-response restrictions, and identified-set inclusion; the same argument conditionally preserves any finite vector of outcome features satisfying a linear moment-preservation equation. **(ii) Why stochastic barycentric compression is structural.** Within the shared product-kernel class, we characterize all kernels that preserve an affine contrast for every latent law: for the ATE, only a common mean shift is possible; endpoint retention forces zero shift, local two-endpoint support uniquely gives barycentric probabilities, and a deterministic kernel satisfying exact mean preservation cannot reduce support. **(iii) Sharpness theory and anytime certification.** We derive instance-dependent dual interpolation bounds, prove a negative theorem excluding any support-uniform mesh-only rate, and turn nested refinement into an anytime algorithm whose certified outer sets tighten monotonically and recover the exact finite-support problem under full refinement; a running intersection preserves monotone validity when the downstream optimizer is itself only outer-certified. **(iv) Finite-sample validity and composability.** A simultaneous Clopper–Pearson/Bonferroni confidence box propagates through the positive compression map, and any certified outer optimizer applied after compression remains outer-valid for the original target by transitivity. **(v) Evidence across theory, computation, and data.** We validate finite-K, finite-J lifts and treatment-response restrictions, benchmark adaptive anytime refinement, demonstrate clean runtime gains beyond binary IV, directly instantiate the exact analytical piece-count barrier, compare against the fastest certified exact backend in contention-free fresh processes, and demonstrate finite-sample validity on a public randomized-encouragement job-training dataset.

Operational significance. The method is a representation layer rather than a causal solver: one may start from a cheap certified surrogate and spend additional computation by refining its retained support. Exact or outer-certified downstream optimizers remain valid for the original target, while the running-intersection rule preserves monotone tightening under approximate certified solves.

2 Related work and novelty boundary

Partial identification and IV bounds. Manski developed the general partial-identification perspective (Manski, 1990; 2003); Balke–Pearl derived sharp nonparametric bounds under imperfect compliance and Bonet studied IV geometry (Balke & Pearl, 1997; Bonet, 2001). Tight symbolic and automated optimization methods now cover broad discrete causal models (Sachs et al., 2023; Duarte et al., 2024). In particular, Duarte et al. use certified global optimization to progressively tighten valid outer bounds while keeping the original causal representation fixed. Our representation-level certificate is complementary: compression can precede such an optimizer, because an outer bound on a compressed identified set is automatically an outer bound on the original target set. Proposition 7 makes this composition explicit and separates representation refinement from optimizer refinement. Song et al. study categorical IV characterization and inference (Song et al., 2024). Chernozhukov et al. and Rosen develop general confidence-set machinery for partially identified models and moment inequalities (Chernozhukov et al., 2007; Rosen, 2008). These methods address identification or inference for a specified representation. Our question is orthogonal: when the representation itself is large, which reduction operations preserve the target and a valid outer set by construction?

Coarsening, discretization, and quantization. Coarsening can change causal targets and identified sets. Gabriel et al. study partial identification when a coarsened exposure is itself the target, while Ou–Nabi analyze discretization bias for causal functionals (Gabriel et al., 2025; Ou & Nabi, 2026). Dual quantization uses randomized barycentric splitting with intrinsic stationarity (Pagès & Wilbertz, 2012). We use that interpolation property as an ingredient, then prove a causal product lift, a target-set certificate, and a necessity result specific to partial identification.

Causal abstraction. Rubenstein et al. formalize exact transformations across structural causal models (Rubenstein et al., 2017); Beckers–Halpern refine abstraction notions (Beckers & Halpern, 2019); Geiger et al. give a broad causal-abstraction theory for mechanistic interpretability (Geiger et al., 2025); Xia–Bareinboim explicitly study inference under lossy projected abstractions (Xia & Bareinboim, 2025); and Buchholz et al. define stochastic transformations between causal spaces (Buchholz et al., 2024). Our objective is narrower in semantic scope but different in its guarantee. We do not require every surrogate model to have a low-level preimage and do not claim preservation of all causal queries. We require that *every original feasible model* have a surrogate image with the same designated target. Extra surrogate models are allowed and are precisely the source of relaxed sharpness. This one-sided asymmetry is what turns lossy representation reduction into a valid outer partial-identification method.

3 Certified compression for finite IV models

Let $Z \in \{0, \dots, K-1\}$ denote the instrument and $D \in \{0, \dots, J-1\}$ the treatment. Outcomes take ordered numerical values

$$g = (g_0 < \cdots < g_{n-1}).$$

For each treatment value d, let Y_d denote the potential outcome under treatment d; for each instrument value z, let D_z denote the treatment that would be taken under $Z = z$. Under consistency ($Y = Y_D$), exclusion (the instrument affects the outcome only through treatment), and random assignment $Z \perp (\mathbf{Y}, \mathbf{D})$, a latent response type is

$$(\mathbf{Y}, \mathbf{D}) = ((Y_0, \dots, Y_{J-1}), (D_0, \dots, D_{K-1})).$$

Let $\mathcal{R} \subseteq \{0, \dots, J-1\}^K$ denote the allowed class of treatment-response vectors. Write $q_{\mathbf{y},\mathbf{d}}$ for the latent probability of outcome-index vector $\mathbf{y} = (y_0, \dots, y_{J-1})$ and treatment-response vector $\mathbf{d} \in \mathcal{R}$. Let $p_{ydz} := \Pr(Y = g_y, D = d \mid Z = z)$ denote the observed conditional cell probability. Then

$$p_{ydz} = \sum_{\mathbf{y},\mathbf{d}'} q_{\mathbf{y},\mathbf{d}'} \mathbf{1}\{d'_z = d\} \mathbf{1}\{y_d = y\}. \tag{2}$$

Equivalently, write $Aq = p$, where A is the linear response-type observation operator induced by Eq. (2).

For $\alpha \in \mathbb{R}^J$ and intercept γ, define the affine mean target

$$\theta(q) = \gamma + \sum_{d=0}^{J-1} \alpha_d \mathbb{E}_q[Y_d],$$

and its sharp identified set

$$\mathcal{I}_\theta(p) = \{\theta(q) : q \text{ is model-feasible and } Aq = p\}.$$

For the binary average treatment effect (ATE), $J = 2$, $\alpha = (-1, 1)$, and $\gamma = 0$.

Let $r = (r_0 < \cdots < r_{m-1})$ be the retained outcome support and $T \in \mathbb{R}^{m \times n}$ a column-stochastic map, meaning

$$T_{ay} \geq 0, \qquad \sum_{a=0}^{m-1} T_{ay} = 1 \quad \text{for every } y.$$

We use m for a generic surrogate support cardinality in the theory and M for the chosen retained-support budget in experiments; in fixed-size experiments, $m = M$.

The mean-preservation condition $r^T T = g^T$ is equivalent to

$$\sum_{a=0}^{m-1} r_a T_{ay} = g_y \quad \text{for every original support point } g_y.$$

Thus each original outcome value is preserved in expectation.

Apply the same T independently to every potential-outcome coordinate while leaving the treatment-response vector unchanged:

$$\bar{q}_{\mathbf{a},\mathbf{d}} = \sum_{\mathbf{y}} q_{\mathbf{y},\mathbf{d}} \prod_{h=0}^{J-1} T_{a_h y_h}. \tag{3}$$

Let $\bar{Y}_d$ denote the resulting surrogate potential outcome on support r, let $\bar{p}$ denote the observed law induced by $\bar{q}$, and let $\mathcal{I}_{\theta,m}(\bar{p})$ denote the corresponding surrogate identified set.

Theorem 1 (Finite-IV target-preserving lift)**.** *Assume $T \geq 0$, $\mathbf{1}_m^T T = \mathbf{1}_n^T$, and $r^T T = g^T$, where $\mathbf{1}_k$ denotes the k-dimensional all-ones vector. Then for every finite J, K and every treatment-response class $\mathcal{R}$: (i) observation commutes, $\bar{p}_{adz} = \sum_y T_{ay} p_{ydz}$; (ii) the full marginal distribution of $\mathbf{D}$ is unchanged; (iii) $\mathbb{E}_{\bar{q}}[\bar{Y}_d] = \mathbb{E}_q[Y_d]$ for every d, hence every affine mean contrast θ is preserved; and (iv)*

$$\boxed{\mathcal{I}_\theta(p) \subseteq \mathcal{I}_{\theta,m}(Tp)}.$$

Any restriction expressed only through treatment-response types or their marginal distribution is preserved. If $R = |\mathcal{R}|$, the original latent LP has $n^J R$ variables, nJK observation rows, and $Kn^J R$ observation-matrix nonzeros; after compression these become $m^J R$, mJK, and $Km^J R$.

Corollary 2 (Finite feature vectors)**.** *Let $\phi_1, \ldots, \phi_H$ be numerical outcome features and define $\Phi_g \in \mathbb{R}^{n \times H}$ and $\Phi_r \in \mathbb{R}^{m \times H}$ by*

$$\Phi_g(y, h) = \phi_h(g_y), \qquad \Phi_r(a, h) = \phi_h(r_a).$$

If T is column stochastic and $\Phi_r^T T = \Phi_g^T$, then the product lift preserves

$$\mathbb{E}_{\bar{q}}[\phi_h(\bar{Y}_d)] = \mathbb{E}_q[\phi_h(Y_d)]$$

for every feature h and treatment d, hence any affine target built from these finitely many moments. Identified-set inclusion and treatment-response preservation follow exactly as in Theorem 1. This statement is conditional on the existence of such a kernel; it does not claim that arbitrary collections of moments admit nontrivial compression.

Why the theorem is true. Fix (a, d, z) and substitute Eq. (3) into Eq. (2). Summing every surrogate potential-outcome coordinate except a_d gives one by column stochasticity, so

$$\bar{p}_{adz} = \sum_{\mathbf{y},\mathbf{d}'} q_{\mathbf{y},\mathbf{d}'} \mathbf{1}\{d'_z = d\} T_{a,y_d} = \sum_y T_{ay} p_{ydz}.$$

Summing over all $\mathbf{a}$ instead leaves the complete $\mathbf{D}$ marginal unchanged. Finally,

$$\mathbb{E}_{\bar{q}}[\bar{Y}_d] = \sum_{\mathbf{y},\mathbf{d}} q_{\mathbf{y},\mathbf{d}} \sum_a r_a T_{a,y_d} = \sum_{\mathbf{y},\mathbf{d}} q_{\mathbf{y},\mathbf{d}} g_{y_d} = \mathbb{E}_q[Y_d].$$

Thus every original feasible latent model maps to a surrogate feasible model with the same target. We never require the converse; additional surrogate models explain exactly why sharpness can be lost. Appendix A.2 gives the full index proof.

3.1 Why barycentric weights are not arbitrary

For a column-stochastic T, define

$$f_y = \sum_a r_a T_{ay}, \qquad e_y = f_y - g_y.$$

Theorem 3 (Necessity and local uniqueness)**.** *Fix the shared product kernel in Eq.* (3)*. For a nonconstant affine target $\gamma + \sum_d \alpha_d \mathbb{E}[Y_d]$, preservation for every latent distribution is equivalent to*

$$\sum_{d=0}^{J-1} \alpha_d e_{y_d} = 0 \quad \textit{for every } (y_0, \ldots, y_{J-1}).$$

If $\sum_d \alpha_d \neq 0$, then $f_y = g_y$ for every y. If $\sum_d \alpha_d = 0$, then $f_y = g_y + \kappa$ for one common κ. In particular, ATE preservation permits only a common shift. If retained values lie in $[g_0, g_{n-1}]$ and include both global endpoints, then $\kappa = 0$. If column y is supported only on retained endpoints $l_y < h_y$, its probabilities are uniquely

$$T_{l(y),y} = \frac{h_y - g_y}{h_y - l_y}, \qquad T_{h(y),y} = \frac{g_y - l_y}{h_y - l_y}. \tag{4}$$

Corollary 4 (Nontrivial exact mean-preserving reduction requires stochasticity)**.** *Assume the stronger pointwise mean-preservation condition $r^T T = g^T$ with strictly distinct original values $g_0, \ldots, g_{n-1}$. If T is deterministic, so each column contains one unit entry, then the image of column y must be a retained value equal to g_y. Consequently every original outcome value must be represented distinctly and $m \geq n$; a strict support reduction $m < n$ is impossible. Thus stochastic splitting is necessary for nontrivial support reduction under the mean-preserving certificate of Theorem 1.*

Proof mechanism. Test preservation on point-mass latent laws. Their target error is $\sum_d \alpha_d e_{y_d}$. If $s = \sum_d \alpha_d \neq 0$, setting every coordinate to the same y yields $s e_y = 0$. If $s = 0$, comparing two tuples that differ in a coordinate with nonzero coefficient yields $e_y = e_{y'}$, hence a common κ. At the global minimum, convexity gives $f_0 \geq g_0$ and therefore $\kappa \geq 0$; at the maximum, $f_{n-1} \leq g_{n-1}$ gives $\kappa \leq 0$. Thus endpoint retention forces $\kappa = 0$. With only l_y, h_y available, column stochasticity leaves one free weight and the equation $(1 - w) l_y + w h_y = g_y$ has the unique solution in Eq. (4). The claim is intentionally scoped to a shared coordinatewise product kernel; arbitrary joint latent transformations can create cancellations not available in this class.

Complexity separation. With contiguous bins and retained endpoints, each original outcome column has at most two nonzeros. Constructing Tp costs $O(nJK)$. For fixed J, K, m, the expensive latent representation depends on m, not n: $m^J R$ variables versus $n^J R$ originally. In the unrestricted binary case this is $4m^2$ versus $4n^2$, and the exact analytical formula changes from $N(n)$ to $N(m)$. Thus a fixed $m = 8$ surrogate has 80,900 analytical pieces even when the original support grows, whereas Eq. (1) gives 5.37×10^9 pieces at $n = 16$ and 2.31×10^{19} at $n = 32$. This creates a smaller *certified surrogate problem*; it does not evade the lower bound for the exact original sharp problem and says nothing about universal pointwise LP runtime.

4 WHAT IS LOST: SHARPNESS, AND HOW FAR VALIDITY EXTENDS

Outer validity alone does not control tightness. In the binary-IV ATE specialization, write the original sharp interval as $\mathcal{I}(p) = [L(p), U(p)]$ and the compressed interval as $\mathcal{I}_m(Tp) = [L_m(Tp), U_m(Tp)]$.

Let $E \in \{0,1\}^{n\times m}$ embed retained support points back into the original support and define $\widetilde{p} = ETp$. Because retained endpoints are original outcomes, $\mathcal{I}_m(Tp) = \mathcal{I}(\widetilde{p})$.

Let λ and η denote optimal dual coefficient vectors for the original lower- and upper-bound LPs, respectively. Weak duality gives

$$0 \le L(p) - L_m(Tp) \le \lambda^T(p - \widetilde{p}), \tag{5}$$

$$0 \le U_m(Tp) - U(p) \le \eta^T(\widetilde{p} - p). \tag{6}$$

Define the probability-weighted expected displacement

$$\rho_T(p) = \sum_{y,d,z} p_{ydz} \sum_a T_{ay}|g_y - r_a|.$$

If the optimal dual coefficients are Lipschitz across retained intervals with constants K_L and K_U, then the two endpoint losses are bounded by $K_L\rho_T(p)$ and $K_U\rho_T(p)$. Let Δ denote the mesh size, i.e. the largest numerical diameter of any retained-support interval. Two-endpoint barycentric splitting gives $\rho_T(p) \le \Delta$, and therefore

$$d_H(\mathcal{I}(p), \mathcal{I}_m(Tp)) \le \max(K_L, K_U)\Delta, \qquad \mathrm{width}(\mathcal{I}_m(Tp)) - \mathrm{width}(\mathcal{I}(p)) \le (K_L + K_U)\Delta,$$

where d_H is Hausdorff distance between intervals. The constants are instance-dependent; this is not a support-uniform rate.

Theorem 5 (No support-uniform mesh-only rate)**.** *For every $0 < \varepsilon < 1/2$, there is a feasible four-outcome binary-IV law on $(0, \varepsilon/2, \varepsilon, 1)$ whose original ATE is point identified, while barycentrically compressing the first three outcomes to $(0, \varepsilon)$ produces a surrogate interval of width exactly one. Therefore no support-uniform $f(\Delta) \to 0$ can bound sharpness loss using mesh alone.*

The construction and endpoint attainability are proved in Appendix A.6. The positive counterpart is an anytime certificate.

Proposition 6 (Anytime certified nested refinement)**.** *Let $T_1, \ldots, T_L$ be valid target-preserving compression maps such that $T_\ell = R_\ell T_{\ell+1}$ for column-stochastic target-preserving refinement maps R_ℓ. Then*

$$\mathcal{I}_1(T_1p) \supseteq \mathcal{I}_2(T_2p) \supseteq \cdots \supseteq \mathcal{I}_L(T_Lp) \supseteq \mathcal{I}(p).$$

For ordered finite support, take retained endpoint knots and repeatedly add an unretained original support point, using barycentric interpolation between adjacent knots. Every step satisfies the factorization above. After at most $n-2$ additions all support points are retained, $T_L = I$, and $\mathcal{I}_L = \mathcal{I}(p)$.

Proposition 7 (Composition with certified outer optimization)**.** *For refinement level ℓ, let $\mathcal{B}_\ell$ be any certified outer interval satisfying $\mathcal{I}_\ell(T_\ell p) \subseteq \mathcal{B}_\ell$; $\mathcal{B}_\ell$ may come from an incomplete anytime optimizer. Define $\mathcal{C}_1 = \mathcal{B}_1$ and $\mathcal{C}_\ell = \mathcal{C}_{\ell-1} \cap \mathcal{B}_\ell$. Then $\mathcal{I}(p) \subseteq \mathcal{C}_L \subseteq \cdots \subseteq \mathcal{C}_2 \subseteq \mathcal{C}_1$. If every surrogate is solved exactly, $\mathcal{C}_\ell = \mathcal{I}_\ell(T_\ell p)$. If the final map is the identity and its solve is exact, $\mathcal{C}_L = \mathcal{I}(p)$.*

Anytime procedure. Propositions 6–7 yield Algorithm 1. For a retained index set S, let $T(S)$ denote its barycentric map on retained support g_S. $\mathsf{SolveOuter}(\bar{p}, S)$ denotes any routine returning an interval $\mathcal{B}$ such that $\mathcal{I}_S(\bar{p}) \subseteq \mathcal{B}$; the routine may be exact or approximate provided this outer-validity certificate holds; π selects the next unretained support index, and Stop is any user-specified computational stopping criterion, such as a support-size or time budget. The refinement policy affects only where resolution is allocated; validity does not depend on that choice.

Refinement policies. Algorithm 1 separates the *certificate* from the *allocation heuristic*. In all experiments, adding $y^\star$ only splits an existing retained-support interval, so the resulting barycentric maps remain genuinely nested. We consider geometric, mass-weighted, observable-displacement (ρ), and compressed-dual sensitivity rules for selecting the next support index $y^\star$. These policies may differ substantially in sharpness at a fixed support budget, but every intermediate $\mathcal{C}_\ell$ remains certified regardless of the policy. With exact surrogate solves, $\mathcal{B}_\ell = \mathcal{I}_\ell(T_\ell p)$ and nesting implies $\mathcal{C}_\ell = \mathcal{B}_\ell$; with an incomplete certified optimizer, the running-intersection update restores monotone tightening. Appendix A.7 gives the policy definitions and proofs.

Algorithm 1 Anytime Certified Target-Preserving Refinement

Require: Observed law p on $g_0 < \cdots < g_{n-1}$; certified outer-set routine (SolveOuter); refinement policy (π); stopping rule (Stop)
Ensure: Nested certified outer intervals $\mathcal{C}_1 \supseteq \mathcal{C}_2 \supseteq \cdots \supseteq \mathcal{I}(p)$

$S_1 \leftarrow \{0, n-1\}$ ▷ indices of the retained global endpoints
$\mathcal{C}_0 \leftarrow (-\infty, +\infty)$
$\ell \leftarrow 1$
while true do
  Construct the barycentric map $T_\ell \leftarrow T(S_\ell)$
  $\bar{p}_\ell \leftarrow T_\ell p$
  $\mathcal{B}_\ell \leftarrow \mathsf{SolveOuter}(\bar{p}_\ell, S_\ell)$ ▷ $\mathcal{I}_\ell(\bar{p}_\ell) \subseteq \mathcal{B}_\ell$
  $\mathcal{C}_\ell \leftarrow \mathcal{C}_{\ell-1} \cap \mathcal{B}_\ell$
  emit $(S_\ell, \mathcal{C}_\ell)$
  if $|S_\ell| = n$ **or** $\mathsf{Stop}(\ell, S_\ell, \mathcal{C}_\ell)$ **then**
    break
  end if
  $y^\star \leftarrow \pi(p, S_\ell, \mathcal{B}_\ell)$ ▷ $y^\star \in \{0, \ldots, n-1\} \setminus S_\ell$
  $S_{\ell+1} \leftarrow S_\ell \cup \{y^\star\}$ ▷ split one existing support interval
  $\ell \leftarrow \ell + 1$
end while
return $\{(S_j, \mathcal{C}_j)\}_{j=1}^{\ell}$

4.1 Nonasymptotic finite-sample outer inference

Let C_{ydz} be observed counts and $N_z = \sum_{y,d} C_{ydz}$. Fix a confidence error level $\delta \in (0, 1)$. For every cell, construct a two-sided Clopper–Pearson interval $[\ell_{ydz}, u_{ydz}]$ with error probability $\delta/(nJK)$ (Clopper & Pearson, 1934). Conditional on $(N_z)_z$, a union bound gives simultaneous coverage of every population cell with probability at least $1 - \delta$. Define the sample-robust target interval by optimizing over latent laws whose implied observed probabilities lie inside this box.

Theorem 8 (Finite-sample outer validity after compression)**.** *Let $[\ell, u]$ be the simultaneous cell-probability box above, and let $[L_N, U_N]$ be the target range over all admissible latent laws satisfying $\ell \le Aq \le u$. For any T satisfying Theorem 1, define $\bar{\ell} = T\ell$ and $\bar{u} = Tu$ cellwise and let $[\bar{L}_N, \bar{U}_N]$ be the analogous compressed robust interval. Then, conditional on $(N_z)_z$,*

$$\Pr(\mathcal{I}_\theta(p) \subseteq [L_N, U_N]) \ge 1 - \delta, \qquad \Pr\big(\mathcal{I}_\theta(p) \subseteq [\bar{L}_N, \bar{U}_N]\big) \ge 1 - \delta.$$

Proof. On the simultaneous event $\ell \le p \le u$, every population latent law with $Aq = p$ is feasible for the original robust LP, so the entire population identified set is contained in $[L_N, U_N]$. Because $T \ge 0$, the same event implies $T\ell \le Tp \le Tu$. Theorem 1 maps every population-feasible q to a compressed latent law with observed distribution Tp and the same target, so the same target values lie in $[\bar{L}_N, \bar{U}_N]$. The simultaneous event has probability at least $1 - \delta$. This procedure is intentionally conservative; it supplies a simple finite-sample certificate, not an efficiency-optimal confidence region.

5 Evidence: validity, anytime refinement, computation, and data

Correctness/generalization experiments use case-level parallelism only; wall-clock comparisons are strictly serial, one solver thread, with a fresh process for each method/repetition. All LP endpoints must pass independent primal, dual, and gap checks.

Validity and sharpness stress tests. Across 2,740 feasible finite-IV populations we solve the exact original and compressed LPs, explicitly form the product-lifted latent law, and independently verify observation commutation, every $\mathbb{E}[Y_d]$, affine-target preservation, treatment-response marginals, containment, and solver certificates. The resulting 13,660 compressed intervals contain the exact original interval in all 13,660 cells. This includes $(K, J, n) = (3, 2, 48), (4, 2, 40), (3, 3, 12)$ and a

Table 1: Contention-free fresh-process binary-IV timing at $M = 32$; 30 paired measurements per row. Sharpness uses the same paired cases. IQR denotes the interquartile range (25th–75th percentile), and p90 denotes the 90th percentile.

n	median speedup	speedup IQR	median inflation	p90 inflation
128	12.94×	11.97–14.72	0.77%	8.70%
256	70.99×	65.73–80.19	0.67%	7.01%
512	236.17×	201.31–570.74	0.50%	4.96%

Table 2: Representative multivalued-IV clean timing. Variable reduction is exact latent-variable count reduction $(n/m)^J$; inflation is relative to the exact original interval.

setting	M	variable reduction	median speedup	median inflation
$(3, 2, 48)$	8	36.0×	12.8×	4.36%
$(3, 3, 12)$	6	8.0×	8.0×	0.47%
$(4, 2, 40)$	8	25.0×	18.7×	4.08%

monotone treatment-response $(3, 3, 16)$ suite. The adversarial $(3, 2, 48)$ family is designed to find poor sharpness rather than validity failures; it reaches 63.65% relative width inflation, and at $M = 32$ its worst case remains 20.48% wider. Full configuration-level results are in Appendix D.

Clean binary exact-vs-compressed timing. Fixtures are built before timing. Each $(\text{case}, \text{method}, \text{repeat})$ runs in a fresh subprocess with Basic Linear Algebra Subprograms (BLAS) libraries and solver threads fixed to one and no case running concurrently. For each case/repetition, the exact reference is the fastest successfully certified backend among full and reduced primal formulations using the HiGHS linear-programming solver, a reduced HiGHS dual formulation, and one-thread Gurobi dual-simplex reoptimization. Six cases (three nonadversarial, three strong-IV) are repeated five times, giving 30 paired measurements per (n, M). Table 1 reports $M = 32$. A separate fixed-resolution stress test reaches $n = 10^7, M = 8$: compression plus the certified surrogate LP takes 0.879 s and 759 MB peak resident-set size (RSS). The corresponding original binary latent formulation would have 4×10^{14} variables; no exact solve at that scale is attempted or used to claim an empirical speedup. Binary validation additionally gives 972/972 barycentric containment, while nearest endpoint rounding contains only 4/162 strong-IV intervals.

Anytime and adaptive certified refinement. We evaluate four genuinely nested policies—geometric, mass-weighted, observable-displacement (ρ), and compressed-dual guided—on 36 nonadversarial randomly generated binary-IV populations, 24 adversarial $(K, J, n) = (3, 2, 48)$ populations, and the empirical job-training benchmark. The 244 nested trajectories contain 7,692 certified refinement states and 7,448 interlevel transitions, with no numerical monotonicity violation; the maximum numerical error in the interlevel factorization $T_\ell = R_\ell T_{\ell+1}$ is 3.33×10^{-16}. Figure 1 shows the runtime–sharpness trajectory. At $M = 32$, nonadversarial median inflation is 2.13% for geometric refinement and 1.99% for dual-guided refinement. On the adversarial family, geometric refinement has 7.04% median / 9.37% p90 inflation, versus 6.06% / 8.54% for dual-guided and 5.90% / 8.23% for ρ-guided refinement. An independently recomputed partition chosen to minimize the maximum bin diameter (a minimax-diameter partition) gives 7.45% / 10.13% at the same M but is not nested. The gain is therefore instance-dependent rather than universal: adaptivity changes *where* representation budget is spent, while certification is unchanged.

Clean multivalued-IV timing. The variable-count reduction predicted by Theorem 1 also produces real gains beyond binary IV. We run a separate serial fresh-process, one-thread timing phase on three cases with three repeats each (nine paired measurements per row). Table 2 reports the most compressed tested resolution for each setting; the full resolution sweep is in Appendix C.4. All 108 clean timing jobs pass solver certification, with maximum primal/dual/gap diagnostic below 10^{-15}.

Direct connection to the exact analytical barrier. We independently materialize the exact analytical construction for $n = 2, \ldots, 8$ and recover Eq. (1) exactly at every size. At $n = 8$ the construction

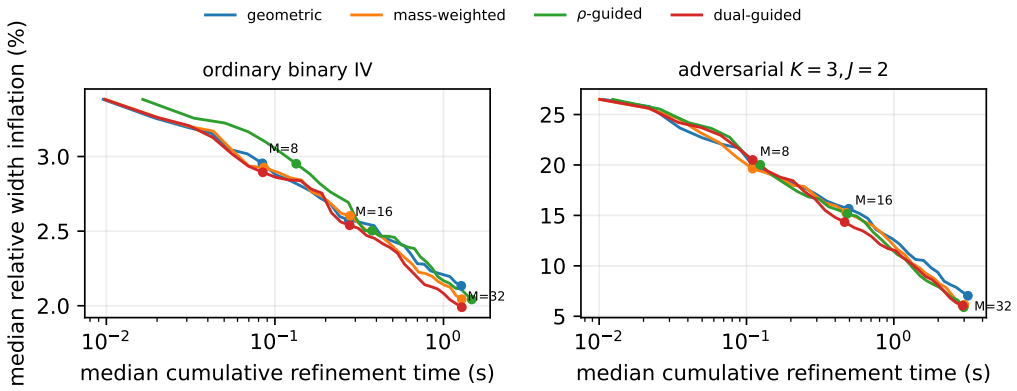

Figure 1: Anytime certified refinement. Each point is the median across matched instances at one retained support size; labeled points are $M \in \{8, 16, 32\}$. All four curves are nested certified trajectories, so additional computation cannot widen the exact surrogate interval. Adaptive policies give modest gains on nonadversarial (ordinary) cases and larger gains on the adversarial family.

contains 80,900 pieces, takes 2.78 s, and reaches 199 MB on the test server. The formula then grows to 5.37×10^9 pieces at $n = 16$ and 2.31×10^{19} at $n = 32$, while a fixed $m = 8$ surrogate stays at 80,900 pieces. This experiment illustrates the representation-size barrier directly; it is not interpreted as evidence that every pointwise original LP has exponential runtime.

Finite-sample inference and JTPA. We evaluate Theorem 8 on 12 feasible $(K, J, n) = (3, 2, 16)$ populations, 100 independent samples per population, and $N \in \{500, 1000, 5000\}$, giving 3,600 sampling tasks and 10,800 intervals. Every reported configuration attains 1200/1200 empirical coverage at nominal 95%; the 100% rate reflects conservatism, not calibration to 100%. For a public randomized-encouragement application, we use the Job Training Partnership Act (JTPA) data(Mathematical Research Data Initiative & Han, 2026), with randomized assignment as the instrument, training receipt as the treatment, and 30-month earnings as the outcome. We use 9,223 complete observations and a deterministic finite-support earnings transformation with 113 numerical values. The uncompressed 95% robust interval is $[-42{,}726, 58{,}042]$; a one-shot $M = 32$ surrogate is only 0.14% wider while measured bound-computation time falls from 0.949 s to 0.028 s ($33.8\times$). In the nested JTPA experiment, a mass-weighted trajectory reaches 0.117% inflation by $M = 32$ while remaining certified at every preceding stage. The estimand is explicitly the finite-support earnings-scale ATE, not the exact continuous-earnings ATE. for more details see the appendix.

6 Discussion and Conclusion

Certified target-preserving compression separates three requirements that are often conflated: target fidelity, validity, and sharpness. For finite IV models, the proposed product lift preserves the designated mean target and treatment-response structure model-by-model while guaranteeing that every original compatible target value remains feasible after compression. The resulting computational gain is therefore purchased through possible loss of sharpness, not by changing the scientific estimand or sacrificing one-sided validity; nested refinement makes this tradeoff anytime and recovers the exact finite-support problem under full refinement. The scope is nevertheless important. Outcomes are finite and ordered, and the surrogate latent space still scales as $m^J|\mathcal{R}|$, so large treatment-response spaces can remain difficult. No support-uniform mesh-only sharpness rate exists, adaptive policies need not dominate universally, and the Clopper–Pearson/Bonferroni construction is intentionally conservative. The uniqueness result is restricted to the shared coordinatewise product-kernel class, and composability with certified outer optimizers does not imply that every such optimizer is empirically accelerated. The JTPA benchmark likewise targets a prespecified finite-support earnings transformation rather than the exact continuous-earnings ATE. Within these boundaries, the framework provides a practical middle ground between exact-but-large sharp representations and uncontrolled coarsening: *preserve the target, certify validity, and relax only sharpness.*

REPRODUCIBILITY STATEMENT

The supplementary material contains complete proofs, deterministic seeds, source code for all validation, timing, nested-refinement, analytical-scaling, finite-sample, and JTPA analyses, raw experimental outputs, solver diagnostics, deterministic configurations, and reproducibility metadata. Wall-clock experiments use serial one-thread fresh processes; case-level parallelism is used only for correctness and throughput studies. The public JTPA data source, license, and deterministic preprocessing are documented in the appendix and supplement.

ETHICS STATEMENT

The simulation studies use synthetic population laws. The empirical benchmark uses an openly released JTPA dataset under its stated CC BY-NC 4.0 license; no new data were collected and no individual-level decisions are made. The empirical analysis is methodological: it demonstrates the finite-support IV certificate and should not be interpreted as a new substantive policy evaluation of JTPA. Causal interpretation still depends on the stated IV assumptions.

AI USE STATEMENT

AI tools were used solely for literature search and language editing. The author verified the cited sources and mathematical statements. The author takes full responsibility for mathematical formulations and proofs, code development, inspection and checking of executable analyses, and the reported numerical results for the submission.

REFERENCES

Alberto Abadie, Joshua Angrist, and Guido Imbens. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117, 2002. doi: 10.1111/1468-0262.00270.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997. doi: 10.1080/01621459.1997.10474074.

Sander Beckers and Joseph Y. Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2678–2685, 2019. doi: 10.1609/aaai.v33i01.33012678.

Blai Bonet. Instrumentality tests revisited. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 48–55. Morgan Kaufmann, 2001.

Arefe Boushehrian, Mohammad Reza Badri, Sina Akbari, and Negar Kiyavash. Fundamental limits and optimal methods for sharp analytical causal bounds in instrumental variable models. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 689–744. PMLR, 2026. URL `https://proceedings.mlr.press/v337/boushehrian26a.html`.

Simon Buchholz, Junhyung Park, and Bernhard Schölkopf. Products, abstractions and inclusions of causal spaces. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 430–449. PMLR, 2024. URL `https://proceedings.mlr.press/v244/buchholz24a.html`.

Victor Chernozhukov, Han Hong, and Elie Tamer. Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284, 2007. doi: 10.1111/j.1468-0262.2007.00794.x.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, 119(547):1778–1793, 2024. doi: 10.1080/01621459.2023.2216909.

Erin E. Gabriel, Michael C. Sachs, and Arvid Sjölander. The impact of coarsening an exposure on partial identifiability in instrumental variable settings. *Biostatistics*, 26(1):kxae042, 2025. doi: 10.1093/biostatistics/kxae042.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://jmlr.org/papers/v26/23-0058.html.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. doi: 10.2307/2951620.

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

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003. doi: 10.1007/b97478.

Mathematical Research Data Initiative and Sukjin Han. Job training partnership act, 2026. Version v1, CC BY-NC 4.0.

Xiaxian Ou and Razieh Nabi. Coarsening bias from variable discretization in causal functionals. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 5138–5167. PMLR, 2026. URL https://proceedings.mlr.press/v337/ou26a.html.

Gilles Pagès and Benedikt Wilbertz. Intrinsic stationarity for vector quantization: Foundation of dual quantization. *SIAM Journal on Numerical Analysis*, 50(2):747–780, 2012. doi: 10.1137/110827041.

Adam M. Rosen. Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. *Journal of Econometrics*, 146(1):107–117, 2008. doi: 10.1016/j.jeconom.2008.08.001.

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017. URL https://auai.org/uai2017/proceedings/papers/11.pdf.

Michael C. Sachs, Gustav Jonzon, Arvid Sjölander, and Erin E. Gabriel. A general method for deriving tight symbolic bounds on causal effects. *Journal of Computational and Graphical Statistics*, 32(2):567–576, 2023. doi: 10.1080/10618600.2022.2071905.

Yilin Song, F. Richard Guo, K. C. Gary Chan, and Thomas S. Richardson. The categorical instrumental variable model: Characterization, partial identification, and statistical inference, 2024. arXiv:2405.09510.

Kevin Muyuan Xia and Elias Bareinboim. Causal abstraction inference under lossy representations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68225–68235. PMLR, 2025. URL https://proceedings.mlr.press/v267/xia25a.html.

A Complete proofs, generalization, and claim boundaries

A.1 Generic linear lifting principle

Let $\mathcal{Q}$ and $\bar{\mathcal{Q}}$ be latent model classes, A and $\bar{A}$ their observation maps, T a linear map between observed distributions, and $c^T q, \bar{c}^T \bar{q}$ scalar linear targets. Let $S : q \mapsto \bar{q}$ be a linear map between latent representations. If

$$S\mathcal{Q} \subseteq \bar{\mathcal{Q}}, \qquad \bar{A}S = TA, \qquad \bar{c}^T S = c^T,$$

then every $q \in \mathcal{Q}$ with $Aq = p$ maps to $Sq \in \bar{\mathcal{Q}}$ with observation Tp and identical target. Therefore

$$\{c^T q : q \in \mathcal{Q}, Aq = p\} \subseteq \{\bar{c}^T \bar{q} : \bar{q} \in \bar{\mathcal{Q}}, \bar{A}\bar{q} = Tp\}.$$

Convexity is unnecessary for inclusion; finite IV response-type models make these target sets compact intervals. This is a sufficient certificate, not a statement that every stochastic observation map has a compatible causal latent lift.

A.2 Finite-IV product lift: complete proof

Let $\mathcal{Y} = \{0, \dots, n-1\}$ index numerical outcomes g_y, $\mathcal{D} = \{0, \dots, J-1\}$, and $\mathcal{Z} = \{0, \dots, K-1\}$. Let $\mathcal{R} \subseteq \mathcal{D}^K$ be an arbitrary allowed treatment-response class. A latent type is $(\mathbf{y}, \mathbf{d}) \in \mathcal{Y}^J \times \mathcal{R}$, with probability $q_{\mathbf{y},\mathbf{d}}$. Consistency, exclusion, and random assignment of Z imply

$$p_{ydz} = \sum_{\mathbf{y}',\mathbf{d}'} q_{\mathbf{y}',\mathbf{d}'} \mathbf{1}\{d'_z = d\}\mathbf{1}\{y'_d = y\}. \tag{A.1}$$

Let $T \in \mathbb{R}^{m \times n}$ be column stochastic and let $r \in \mathbb{R}^m$. Define

$$\bar{q}_{\mathbf{a},\mathbf{d}} = \sum_{\mathbf{y}} q_{\mathbf{y},\mathbf{d}} \prod_{h=0}^{J-1} T_{a_h y_h}. \tag{A.2}$$

Nonnegativity is immediate. Summing over all $\mathbf{a}$ gives

$$\sum_{\mathbf{a}} \bar{q}_{\mathbf{a},\mathbf{d}} = \sum_{\mathbf{y}} q_{\mathbf{y},\mathbf{d}} \prod_h \sum_{a_h} T_{a_h y_h} = \sum_{\mathbf{y}} q_{\mathbf{y},\mathbf{d}}. \tag{A.3}$$

Hence total mass and the complete treatment-response marginal are preserved. In particular, if the original model assigns mass only to $\mathcal{R}$ (e.g. a monotone/admissible compliance class), the surrogate does too; more generally any constraint expressed solely in that marginal remains exactly true.

For observation commutation, fix (a, d, z) and use (A.1)–(A.2):

$$\begin{aligned}
\bar{p}_{adz} &= \sum_{\mathbf{a}',\mathbf{d}'} \bar{q}_{\mathbf{a}',\mathbf{d}'} \mathbf{1}\{d'_z = d\}\mathbf{1}\{a'_d = a\} \\
&= \sum_{\mathbf{y},\mathbf{d}'} q_{\mathbf{y},\mathbf{d}'} \mathbf{1}\{d'_z = d\} \sum_{\mathbf{a}'} \mathbf{1}\{a'_d = a\} \prod_h T_{a'_h y_h} \\
&= \sum_{\mathbf{y},\mathbf{d}'} q_{\mathbf{y},\mathbf{d}'} \mathbf{1}\{d'_z = d\} T_{a,y_d} \prod_{h \neq d} \sum_{a_h} T_{a_h y_h} \\
&= \sum_y T_{ay} \sum_{\mathbf{y},\mathbf{d}'} q_{\mathbf{y},\mathbf{d}'} \mathbf{1}\{d'_z = d\}\mathbf{1}\{y_d = y\} = \sum_y T_{ay} p_{ydz}.
\end{aligned}$$

Thus $\bar{A}S = TA$ index by index.

Now assume $r^T T = g^T$. For every potential-outcome coordinate d,

$$\begin{aligned}
\mathbb{E}_{\bar{q}}[\bar{Y}_d] &= \sum_{\mathbf{a},\mathbf{d}'} r_{a_d} \bar{q}_{\mathbf{a},\mathbf{d}'} \\
&= \sum_{\mathbf{y},\mathbf{d}'} q_{\mathbf{y},\mathbf{d}'} \left(\sum_{a_d} r_{a_d} T_{a_d y_d} \right) \prod_{h \neq d} \sum_{a_h} T_{a_h y_h} \\
&= \sum_{\mathbf{y},\mathbf{d}'} q_{\mathbf{y},\mathbf{d}'} g_{y_d} = \mathbb{E}_q[Y_d].
\end{aligned}$$

Therefore every affine target $\gamma + \sum_d \alpha_d \mathbb{E}[Y_d]$ is unchanged. Combining commutation with target preservation proves the identified-set inclusion in Theorem 1.

If $R = |\mathcal{R}|$, the original latent vector has $n^J R$ entries. There are nJK observed cells. For each instrument value z, one latent type contributes to exactly one observed (y, d) cell, so each latent column has exactly K nonzeros and the observation matrix has $Kn^J R$ nonzeros. The surrogate counts replace n by m. The unrestricted case has $R = J^K$. The binary unrestricted specialization yields $4n^2$ variables, $4n$ rows, and $8n^2$ nonzeros.

Finite vectors of preserved outcome features. Corollary 2 follows by replacing the scalar numerical identity $r^T T = g^T$ with a finite feature identity. Write the yth original feature vector as $\Phi_g(y) \in \mathbb{R}^H$ and the ath retained feature vector as $\Phi_r(a) \in \mathbb{R}^H$. If

$$\Phi_r^T T = \Phi_g^T,$$

then for every feature coordinate h and treatment coordinate d,

$$\begin{aligned}\mathbb{E}_{\bar{q}}[\phi_h(\bar{Y}_d)] &= \sum_{\mathbf{y},\mathbf{d}'} q_{\mathbf{y},\mathbf{d}'} \sum_a T_{a,y_d} \phi_h(r_a) \\ &= \sum_{\mathbf{y},\mathbf{d}'} q_{\mathbf{y},\mathbf{d}'} \phi_h(g_{y_d}) = \mathbb{E}_q[\phi_h(Y_d)].\end{aligned}$$

Observation commutation and the treatment-response marginal proof do not use the numerical outcome values, so they are unchanged. Hence every affine functional of the finitely many preserved feature moments has the same model-by-model value and inherits identified-set inclusion. Existence is a geometric condition: for each original state y, $\Phi_g(y)$ must lie in the convex hull of retained feature vectors used by column y. We do not claim that arbitrary feature collections admit a strict support reduction.

Composition with certified optimization. Suppose a downstream procedure returns a certified outer set $C(Tp)$ satisfying $\mathcal{I}_{\theta,m}(Tp) \subseteq C(Tp)$. Theorem 1 immediately gives

$$\mathcal{I}_\theta(p) \subseteq \mathcal{I}_{\theta,m}(Tp) \subseteq C(Tp).$$

Thus certified representation reduction and certified approximate optimization compose without a new causal argument. This establishes validity of a compression-then-certified-optimizer pipeline; it does not imply that such a pipeline is always faster or tighter, which is an empirical question.

A.3 Necessity and uniqueness within a shared product outcome kernel

For a column-stochastic T, define

$$f_y = \sum_a r_a T_{ay}, \qquad e_y = f_y - g_y.$$

Under the product lift, the difference between the surrogate and original affine mean target at a point-mass latent law with potential-outcome indices $(y_0, \ldots, y_{J-1})$ is exactly

$$\sum_{d=0}^{J-1} \alpha_d e_{y_d}. \tag{A.4}$$

Because preservation is required for every latent distribution, point masses show necessity of (A.4)$= 0$ for every tuple; linearity shows sufficiency.

Assume α is not identically zero. If $s = \sum_d \alpha_d \neq 0$, choose $y_0 = \cdots = y_{J-1} = y$. Then $se_y = 0$ for every y, so $e_y = 0$ and $f_y = g_y$. If $s = 0$, select a coordinate $d_\star$ with $\alpha_{d_\star} \neq 0$. Compare two tuples that differ only at coordinate $d_\star$, taking values y and y'. Subtracting their two instances of (A.4) yields $\alpha_{d_\star}(e_y - e_{y'}) = 0$; hence $e_y = \kappa$ for all y. Conversely any common κ makes (A.4) equal $\kappa \sum_d \alpha_d = 0$. This proves the complete characterization. If all $\alpha_d = 0$, the target is constant and imposes no condition on T.

For the binary ATE, $\alpha = (-1, 1)$, so preservation for every latent distribution is equivalent to $f_y = g_y + \kappa$. Suppose every retained value lies in $[g_0, g_{n-1}]$ and the retained support contains

both global endpoints. Since f_0 is a convex combination of retained values, $f_0 \geq g_0$, hence $\kappa \geq 0$. Likewise $f_{n-1} \leq g_{n-1}$, hence $\kappa \leq 0$. Thus $\kappa = 0$.

Finally suppose column y is supported only on two retained numerical values $l_y < h_y$. Let w be its mass on h_y. Column stochasticity gives mass $1 - w$ on l_y, and exact mean preservation requires

$$(1 - w)l_y + wh_y = g_y.$$

The unique solution is

$$w = \frac{g_y - l_y}{h_y - l_y}, \qquad 1 - w = \frac{h_y - g_y}{h_y - l_y}.$$

For a singleton $l_y = h_y = g_y$, unit mass is forced. Hence the barycentric rule is unique under local two-endpoint support.

Proof of Corollary 4. If T is deterministic and column stochastic, every column y equals one basis vector $e_{a(y)}$. The strong mean-preservation equation $r^T T = g^T$ therefore reads

$$r_{a(y)} = g_y \qquad \text{for every original support index } y.$$

Because the original values g_y are distinct, two different original indices cannot share one retained state: if $a(y) = a(y')$, then $g_y = r_{a(y)} = r_{a(y')} = g_{y'}$, a contradiction. Hence $y \mapsto a(y)$ is injective and $m \geq n$. If retained values are selected from the original support, every g_y itself must be retained. Therefore deterministic strict support reduction is incompatible with the exact first-moment certificate. This claim uses the strong equation $r^T T = g^T$; ATE preservation alone, before endpoint anchoring, permits the common-shift freedom characterized above.

This is not universal necessity among arbitrary causal compressions. If one allows a joint latent transformation that does not factor as one shared outcome kernel applied coordinatewise, different potential-outcome coordinates can be coupled so target errors cancel. Likewise coordinate-specific kernels can have additional degrees of freedom. The theorem therefore answers "why barycentric?" only inside the explicitly stated shared product-kernel/local-support class.

A.4 Binary-IV latent program

For the sharpness theory below specialize to $J = K = 2$, unrestricted treatment-response types, and target $\mathbb{E}[Y_1 - Y_0]$. Write $q_{ijuv} = \Pr(Y_0 = g_i, Y_1 = g_j, D_0 = u, D_1 = v)$. The exact interval is

$$L(p) = \min_{q \geq 0 : Aq = p} c^T q, \qquad U(p) = \max_{q \geq 0 : Aq = p} c^T q,$$

with $c_{ijuv} = g_j - g_i$. Finite-dimensional LP duality gives

$$L(p) = \max_{A^T \lambda \leq c} \lambda^T p, \qquad U(p) = \min_{A^T \eta \geq c} \eta^T p.$$

For the contiguous endpoint construction, $T \geq 0$, $\mathbf{1}^T T = \mathbf{1}^T$, and $r^T T = g^T$, so Theorem 1 gives the unpadded outer interval.

A.5 Dual interpolation certificate and conditional rate

Let $E \in \{0, 1\}^{n \times m}$ embed retained values back into the original support, and let $\widetilde{p} = ETp$ cellwise. Then

$$L_m(Tp) = L(\widetilde{p}), \qquad U_m(Tp) = U(\widetilde{p}).$$

Indeed every surrogate model embeds in the original support with law $\widetilde{p}$. Conversely T fixes every retained outcome, so compressing any model feasible for $\widetilde{p}$ returns law Tp and preserves its target. Both target sets therefore coincide. This step uses that retained endpoints belong to the original support.

Choose any original optimal lower dual λ and optimal upper dual η. Weak duality at $\widetilde{p}$, followed by the outer inclusion above, gives

$$0 \leq L(p) - L_m(Tp) \leq \lambda^T (p - \widetilde{p}), \tag{7}$$

$$0 \leq U_m(Tp) - U(p) \leq \eta^T (\widetilde{p} - p). \tag{8}$$

Thus the right sides are nonnegative in exact arithmetic. Defining $\mathcal{J}_T v(y,d,z) = \sum_a T_{ay} v(i(a),d,z)$, where $i(a)$ is the original index of retained state a, the first right side is exactly $\sum_{ydz} p_{ydz}(\lambda_{ydz} - \mathcal{J}_T\lambda(y,d,z))$, and the second is $\sum_{ydz} p_{ydz}(\mathcal{J}_T\eta(y,d,z) - \eta_{ydz})$. They are bounded by the corresponding weighted absolute interpolation errors.

Suppose there are optimal duals with

$$|\lambda_{ydz} - \lambda_{i(a)dz}| \le K_L|g_y - r_a|, \quad |\eta_{ydz} - \eta_{i(a)dz}| \le K_U|g_y - r_a|$$

for every positive T_{ay} and every d, z. Define

$$\rho_T(p) = \sum_{ydz} p_{ydz} \sum_a T_{ay}|g_y - r_a|, \quad \Delta = \max_{\text{bins } b}(h_b - l_b).$$

For a point in a nonsingleton bin, its expected absolute displacement is $2(g_y - l_y)(h_y - g_y)/(h_y - l_y) \le (h_y - l_y)/2$. Because each conditional law $p(\cdot,\cdot \mid z)$ sums to one and $K = 2$ in this binary-IV subsection,

$$\sum_{y,d,z} p_{ydz} = 2.$$

Therefore $\rho_T(p) \le \Delta$. Equations (7)–(8) imply

$$d_H(\mathcal{I}(p), \mathcal{I}_m(Tp)) \le \max(K_L, K_U)\rho_T(p) \le \max(K_L, K_U)\Delta,$$

$$\text{width}(\mathcal{I}_m(Tp)) - \text{width}(\mathcal{I}(p)) \le (K_L + K_U)\rho_T(p) \le (K_L + K_U)\Delta.$$

Here d_H is the Hausdorff distance between compact intervals. The constants may depend on p, g, n and the chosen dual solutions. Uniform rates for a family require uniformly bounded constants. Computing these certificates generally requires the original LP, so they are diagnostics and conditional theory, not a cheap surrogate-only stopping rule. Singleton bins contribute zero. A singleton original support has identically zero ATE and is handled separately.

A.6 Counterexample to every uniform mesh-only vanishing rate

For any $0 < \varepsilon < 1/2$, set $g = (0, \varepsilon/2, \varepsilon, 1)$. The only nonzero observed cells are

$$p_{\varepsilon/2,0,0} = 1/2, \quad p_{1,1,0} = 1/2, \quad p_{0,0,1} = p_{\varepsilon,0,1} = 1/4, \quad p_{1,1,1} = 1/2.$$

A feasible latent distribution assigns probability $1/2$ to $(Y_0, Y_1, D_0, D_1) = (\varepsilon/2, 1, 0, 1)$ and probabilities $1/4$ each to $(0,1,1,0)$ and $(\varepsilon,1,1,0)$. The untreated supports for $z = 0$ and $z = 1$ are disjoint, so there can be no never-takers. Because each instrument arm has untreated mass $1/2$, the two switching strata exhaust the population. There can be no always-takers either. Everyone's Y_1 is observed as one and $\mathbb{E}Y_0 = \varepsilon/2$. Hence $\mathcal{I}(p) = \{1 - \varepsilon/2\}$.

Use bins $\{0, \varepsilon/2, \varepsilon\}$ and $\{1\}$. There are three retained values and $\Delta = \varepsilon$. Compression makes both instrument arms identical: untreated mass $1/4$ at each of $0, \varepsilon$ and treated mass $1/2$ at one. For either arm, observed contributions imply $\mathbb{E}Y_1 \in [1/2, 1]$ and $\mathbb{E}Y_0 \in [\varepsilon/4, \varepsilon/4 + 1/2]$, hence

$$\mathcal{I}_m(Tp) = [-\varepsilon/4, 1 - \varepsilon/4].$$

Both extrema are attainable: use never-takers of masses $1/4$ at each untreated value and always-takers of mass $1/2$ with $Y_1 = 1$; set the unobserved outcomes to zero/one as appropriate. The original interval has width zero; the compressed interval has width exactly one. Its Hausdorff error is $1 - \varepsilon/4$. Thus there is no bound $f(\Delta) \to 0$ uniform over all ordered supports and feasible laws, even for $n = 4$, $m = 3$, and outcome span one. In particular an unconditional $C\Delta$ statement with support-independent C is false. This varying-support counterexample does not contradict eventual exactness on one fixed finite support.

A.7 Anytime nested refinement and finite-support convergence

Let the retained original support indices be a knot set $S \subseteq \{0, \dots, n-1\}$ containing 0 and $n - 1$. Between adjacent knots $a < b$, define T_S by the unique barycentric interpolation from each g_y, $a \le y \le b$, to g_a and g_b. If $S \subset S'$ and S' is obtained by adding knots, apply the coarse rule T_S to each fine retained value in S' to obtain a stochastic matrix $R_{S \leftarrow S'}$. On every coarse interval, the

two coarse barycentric coordinate functions are affine in the numerical outcome value. Taking their expectation under the fine barycentric representation therefore reproduces their value at the original point, giving the exact factorization

$$T_S = R_{S\leftarrow S'} T_{S'}, \qquad r_S^T R_{S\leftarrow S'} = r_{S'}^T.$$

The second identity means the interlevel map is itself target preserving. Applying Theorem 1 to the fine surrogate as the "original" model yields

$$\mathcal{I}(p) \subseteq \mathcal{I}_{S'}(T_{S'}p) \subseteq \mathcal{I}_S(T_S p).$$

Induction proves Proposition 6 for any chain $S_1 \subset S_2 \subset \cdots \subset S_L$. The rule used to select the next knot never enters this proof. On a fixed finite support, adding one previously unretained point at each stage reaches $S = \{0, \ldots, n-1\}$ after at most $n-2$ additions; then $T_S = I$ and the exact original problem is recovered. This is eventual finite-support exactness, not a support-uniform rate.

Proof of Proposition 7. For every level ℓ, Proposition 6 gives $\mathcal{I}(p) \subseteq \mathcal{I}_\ell(T_\ell p)$, while certification of the inner optimizer gives $\mathcal{I}_\ell(T_\ell p) \subseteq \mathcal{B}_\ell$. Hence $\mathcal{I}(p) \subseteq \mathcal{B}_\ell$ for every ℓ. Intersecting sets that all contain $\mathcal{I}(p)$ preserves that containment, and the recursive definition $\mathcal{C}_\ell = \mathcal{C}_{\ell-1} \cap \mathcal{B}_\ell$ makes $\mathcal{C}_\ell \subseteq \mathcal{C}_{\ell-1}$ immediate. If $\mathcal{B}_\ell = \mathcal{I}_\ell(T_\ell p)$ at every level, nestedness implies $\bigcap_{j\le\ell} \mathcal{B}_j = \mathcal{I}_\ell(T_\ell p)$. If moreover $T_L = I$ and level L is solved exactly, this equals $\mathcal{I}(p)$. This argument requires no convergence rate and applies to any optimizer whose reported interval is certified to contain the exact compressed identified set. It therefore supplies a safe composition rule for representation-level refinement and optimization-level anytime methods.

The same nesting logic applies to the finite-sample robust construction. If the fine feasible observation lies in $[T_{S'}\ell, T_{S'}u]$, positivity of $R_{S\leftarrow S'}$ maps it into $[T_S\ell, T_S u]$, while the interlevel product lift preserves the target. Hence robust outer intervals also tighten monotonically under the same nested chain.

Adaptive allocation policies. The implementation considers four policy families; all return an unretained original support point and therefore preserve the theorem above.

1. *Largest-diameter/geometric*: split the current knot interval with the largest numerical diameter near its midpoint.
2. *Mass-weighted*: prioritize an interval by observed outcome mass times numerical diameter and split near its mass-weighted median.
3. *Observable-displacement (ρ) guided*: choose the support point that maximally reduces the current probability-weighted expected barycentric displacement $\rho_T(\hat{p})$. This is an allocation diagnostic, not a universal error certificate.
4. *Compressed-dual sensitivity*: use endpoint dual multipliers from the current compressed equality LP to prioritize displaced observed mass in intervals with large estimated dual variation. These are surrogate duals used only as a heuristic; the conditional bound in Appendix A.5 is stated using optimal duals of the original problem and should not be conflated with this policy.

For comparison, an independently recomputed knot set chosen approximately to minimize the maximum interval diameter is evaluated at each matched M. Each such point is individually certified, but because the sets are not constrained to be nested their sequence is not promised to tighten monotonically. No policy is claimed to possess a universal approximation-error rate; Theorem 5 rules out such a claim without additional regularity.

B Finite-sample outer inference: complete argument

Fix a confidence error level $\delta \in (0,1)$. For each instrument stratum z, condition on its sample size N_z. The cell count C_{ydz} is marginally binomial with parameters (N_z, p_{ydz}), even though the full vector of cells is multinomial. Let $B = nJK$ be the number of observed cells, and for each cell let $[\ell_{ydz}, u_{ydz}]$ be a two-sided Clopper–Pearson interval with total cellwise error probability δ/B. Exact binomial coverage and the union bound imply

$$\Pr(\ell_{ydz} \le p_{ydz} \le u_{ydz} \text{ for all } y, d, z \mid (N_z)_z) \ge 1 - \delta. \tag{A.5}$$

Define

$$\mathcal{Q}_N = \{q \geq 0 : \mathbf{1}^T q = 1,\ \ell \leq Aq \leq u\}, \qquad [L_N, U_N] = \{c^T q : q \in \mathcal{Q}_N\}.$$

On the event in (A.5), if q belongs to the population feasible set $\{q : Aq = p\}$, then $q \in \mathcal{Q}_N$. Therefore the complete population identified set is contained in $[L_N, U_N]$.

For the compressed construction, apply the positive observation map cellwise. Since $T \geq 0$, (A.5) implies

$$T\ell \leq Tp \leq Tu.$$

For every population-feasible latent law q, Theorem 1 constructs a compressed latent law $\bar{q}$ with observation Tp and identical target. Thus $\bar{q}$ is feasible for the compressed robust LP with box $[T\ell, Tu]$, and the same original target value belongs to $[\bar{L}_N, \bar{U}_N]$. The probability statement follows from (A.5). No asymptotic approximation is used. The price is conservatism from cellwise exact intervals and the union bound.

This construction separates three quantities that should not be conflated: the population sharp identified set $\mathcal{I}_\theta(p)$; the population compressed outer set $\mathcal{I}_{\theta,m}(Tp)$; and the sample-robust outer confidence interval. The first two differ only through representation/sharpness. The third additionally reflects sampling uncertainty.

C Algorithms and implementation details

C.1 Barycentric endpoint map

Partition the ordered support into contiguous bins. For each nonsingleton bin with retained numerical endpoints $l < h$, an original value $g_y \in [l, h]$ is mapped to l and h with Eq. (4). A singleton maps to itself. Each column of T has at most two nonzeros, is stochastic, and satisfies $r^T T = g^T$ up to floating-point roundoff. The population observed law is compressed by applying T separately to the outcome coordinate for every treatment and instrument cell.

The implementation checks expectation preservation with a scale-aware floating-point tolerance

$$\max\{2 \times 10^{-12},\ 64\,\epsilon_{\text{mach}} \max(1, \|g\|_\infty)\},$$

where ϵ_{mach} denotes machine epsilon. This scale-aware tolerance is needed when numerical outcomes are measured in large physical units such as dollars. This arithmetic check is separate from the LP feasibility/certificate tolerance.

C.2 LP certification

Population validation and finite-sample experiments use SciPy/HiGHS with requested primal and dual feasibility tolerances 10^{-9}. Returned endpoint solutions are independently checked for primal feasibility, dual feasibility, and duality gap, with a common acceptance threshold 2×10^{-8}. The clean timing study additionally evaluates multiple exact backends and selects the fastest backend that passes the same certification gate for each paired case/repetition. Gurobi is restricted to one solver thread.

C.3 Anytime/adaptive refinement results

We evaluate the four nested policies from Appendix A.7 on 36 nonadversarial randomly generated binary-IV populations (24 with $n = 128$, 12 with $n = 256$), 24 adversarial $(K, J, n) = (3, 2, 48)$ populations, and JTPA. The resulting 244 nested policy trajectories contain 7,692 certified states and 7,448 interlevel transitions. Every trajectory is numerically monotone. Across all states, the largest primal residual is 6.87×10^{-10}, dual residual 1.53×10^{-11}, duality gap 8.01×10^{-11}, and interlevel factorization error 3.33×10^{-16}. These diagnostics are comfortably inside the common 2×10^{-8} acceptance threshold; Table 3 reports the sharpness results and Figure 2 shows the JTPA refinement trajectory.

The nonadversarial-family gains from adaptivity are intentionally modest. At $M = 32$, median relative inflation is 2.13% for geometric refinement, 1.99% for dual-guided refinement, and 2.04%

Table 3: Nested-policy and independent-partition sharpness at selected retained support sizes. Generated directly from the raw per-instance trajectories. Independent minimax partitions are individually certified but not constrained to be nested.

family	policy	M	cases	median infl.	p90 infl.
adversarial k3	dual	8	24	20.50%	28.40%
adversarial k3	dual	16	24	14.34%	19.43%
adversarial k3	dual	32	24	6.06%	8.54%
adversarial k3	geometric	8	24	19.86%	28.29%
adversarial k3	geometric	16	24	15.66%	20.42%
adversarial k3	geometric	32	24	7.04%	9.37%
adversarial k3	independent minimax	8	24	20.29%	27.29%
adversarial k3	independent minimax	16	24	16.15%	19.65%
adversarial k3	independent minimax	32	24	7.45%	10.13%
adversarial k3	mass	8	24	19.64%	27.32%
adversarial k3	mass	16	24	15.26%	18.39%
adversarial k3	mass	32	24	6.14%	8.60%
adversarial k3	rho	8	24	20.02%	26.25%
adversarial k3	rho	16	24	15.16%	18.42%
adversarial k3	rho	32	24	5.90%	8.23%
ordinary binary	dual	8	36	2.89%	5.15%
ordinary binary	dual	16	36	2.54%	4.54%
ordinary binary	dual	32	36	1.99%	3.48%
ordinary binary	geometric	8	36	2.95%	5.10%
ordinary binary	geometric	16	36	2.57%	4.60%
ordinary binary	geometric	32	36	2.13%	3.90%
ordinary binary	independent minimax	8	36	2.90%	5.10%
ordinary binary	independent minimax	16	36	2.64%	4.66%
ordinary binary	independent minimax	32	36	2.04%	3.96%
ordinary binary	mass	8	36	2.93%	5.17%
ordinary binary	mass	16	36	2.60%	4.55%
ordinary binary	mass	32	36	2.04%	3.73%
ordinary binary	rho	8	36	2.95%	5.18%
ordinary binary	rho	16	36	2.50%	4.57%
ordinary binary	rho	32	36	2.04%	3.74%

for ρ-guided refinement. The hard family separates the policies more clearly: geometric refinement gives 7.04% median / 9.37% p90 inflation, dual-guided 6.06% / 8.54%, and ρ-guided 5.90% / 8.23%. No policy is claimed to dominate instancewise or asymptotically.

C.4 MULTIVALUED-IV CLEAN TIMING

The multivalued-IV benchmark has two phases. A broad case-parallel phase validates all requested settings, while timing claims use a separate serial fresh-process phase with one BLAS/solver thread and no concurrent case. Clean timing uses three cases and three repetitions for each setting, yielding nine paired measurements per retained support size. Every one of the 108 clean jobs passes certification; the largest primal residual is 2.78×10^{-16}, dual residual 9.99×10^{-16}, and duality gap 6.67×10^{-16}. Table 4 reports the complete sweep.

The strongest representative settings are $(K, J, n, M) = (3, 2, 48, 8)$ with 36× fewer latent variables, 12.8× median wall-time speedup and 4.36% median inflation; $(4, 2, 40, 8)$ with 25× fewer variables, 18.7× speedup and 4.08% inflation; and $(3, 3, 12, 6)$ with 8× fewer variables, 8.0× speedup and

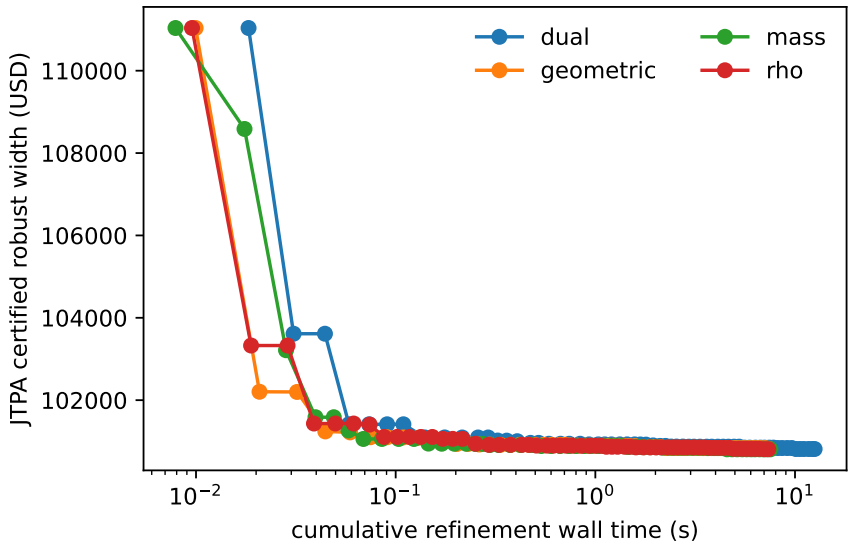

Figure 2: JTPA nested refinement. Every displayed policy is a certified nested chain; the exact robust interval width is the limiting reference. Cumulative time includes all preceding refinement levels and therefore should not be compared to a one-shot solve as a speedup.

Table 4: Complete multivalued-IV clean timing sweep. Variable reduction follows exactly from $(n/m)^J$.

setting	M	pairs	median speedup	median infl.
$(3, 2, 48)$	8	9	12.8×	4.36%
$(3, 2, 48)$	16	9	7.7×	3.30%
$(3, 2, 48)$	24	9	3.3×	2.35%
$(3, 3, 12)$	6	9	8.0×	0.47%
$(3, 3, 12)$	8	9	3.6×	0.35%
$(3, 3, 12)$	10	9	2.6×	0.20%
$(4, 2, 40)$	8	9	18.7×	4.08%
$(4, 2, 40)$	16	9	6.7×	2.94%
$(4, 2, 40)$	24	9	2.9×	2.16%

0.47% inflation. This supports the computational consequence of the finite-J, K theorem beyond binary IV without claiming that speedup is determined by variable count alone.

C.5 Direct analytical-representation scaling

The archived exact analytical construction is materialized for $n = 2, \ldots, 8$. Its piece count agrees exactly with Eq. (1) at every size: $8, 52, 260, 1156, 4868, 19972, 80900$. On the test server, construction time rises from 3.0×10^{-4} s at $n = 2$ to 2.78 s at $n = 8$, while peak RSS reaches 199 MB. Formula evaluation gives $N(16) = 5{,}368{,}446{,}980$ and $N(32) = 23{,}058{,}430{,}074{,}957{,}070{,}340$. A fixed $m = 8$ surrogate remains at 80,900 pieces regardless of original n. Figure 3 visualizes this growth. These measurements illustrate output-size growth of the exact analytical representation; they are not extrapolated into a lower bound on pointwise LP runtime.

C.6 Certified-optimizer complementarity and AutoBounds compatibility

Proposition 7 is implementation-independent: if a downstream optimizer returns any certified outer set for the compressed problem, transitivity makes it outer-valid for the original target, and the running intersection is monotone. We additionally attempted a direct comparison with the public AutoBounds

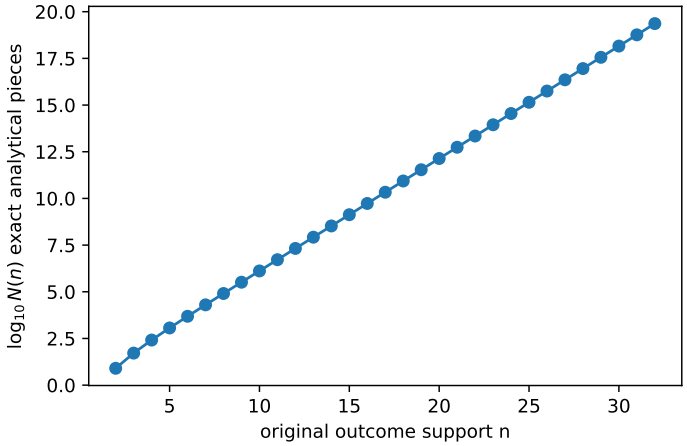

Figure 3: Exact analytical piece count $N(n)$ from Eq. (1). The plotted values are formula-exact; the construction itself is explicitly materialized and count-checked through $n = 8$.

implementation associated with Duarte et al. (2024). The package and SCIP bindings installed successfully at git commit `4e6d75886a732d46e46a0e549b987670beac2e4b`, but every planned original/compressed call terminated before optimization because its SCIP worker launcher attempted to transfer non-picklable PySCIPOpt model objects under the active multiprocessing context. The recorded error is: *"Failed to launch SCIP workers ... pyscipopt models are not picklable."* We therefore report no AutoBounds performance numbers and make no empirical claim that compression accelerates AutoBounds. This is a software-compatibility limitation, not a failure of the outer-set composition theorem.

C.7 Reproducibility of the experiments

Each independent job records its deterministic seed, experimental configuration, solver status, objective values, certification diagnostics, problem dimensions, timing, and software metadata. Rerunning skips matching successful jobs. A separate aggregation program rejects containment or monotonicity violations, reports missing or unsupported external-method runs, and generates CSV, JSON, TeX tables, and figures directly from raw files.

D Population validation and sharpness details

D.1 Binary validation

For binary Z, D, support sizes $n \in \{128, 256, 512\}$ and retained budgets $M \in \{8, 16, 32\}$ are evaluated over 90 ordinary and 18 strong-instrument cases per support size. Barycentric containment holds in 972/972 cells, with target and observed-mean distortions at numerical precision. A nearest-endpoint rule, which does not preserve outcome means, contains only 4/162 strong-instrument intervals. At $n = 512$, nonadversarial-family p90 normalized excess widths are 0.0639, 0.0539, and 0.0501 for $M = 8, 16, 32$; Figure 4 summarizes the validity comparison.

D.2 Finite-K, finite-J validation

Each case samples a feasible latent population, constructs its observed law, solves the original LP, forms the explicit product-lifted latent law, solves the compressed LP, and checks observation commutation, each potential-outcome mean, the affine target, the complete treatment-response marginal, interval containment, and solver diagnostics. Table 5 reports the complete results.

The adversarial suite uses budgets $M \in \{4, 6, 8, 12, 16, 24, 32\}$. Its median relative width inflations are 24.60%, 22.41%, 20.53%, 17.87%, 15.18%, 11.03%, and 7.56%, respectively; p90 values are

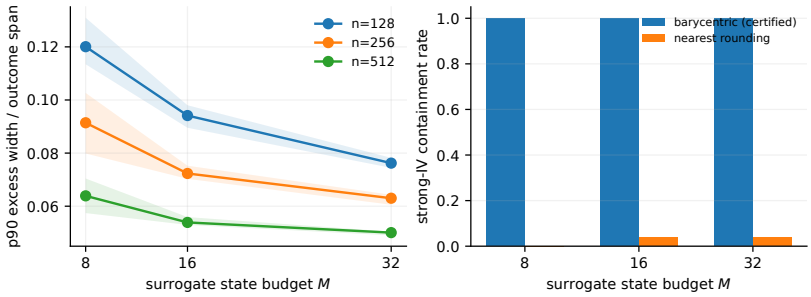

Figure 4: Binary-IV population validation. Barycentric compression preserves outer validity across the validation grid, while nearest deterministic rounding fails on most strong-instrument cases.

Table 5: Detailed generalized population validation.

Configuration	cases	cells	coverage	median	p90	max
$K = 3, J = 2, n = 48$, random	600	2400	2400/2400	2.73%	6.01%	13.86%
$K = 4, J = 2, n = 40$, random	600	2400	2400/2400	2.21%	5.34%	11.82%
$K = 3, J = 3, n = 12$, random	240	960	960/960	0.33%	0.75%	1.84%
$K = 3, J = 3, n = 16$, monotone	300	900	900/900	0.00%	$< 10^{-12}$	0.29%
$K = 3, J = 2, n = 48$, adversarial	1000	7000	7000/7000	17.39%	26.89%	63.65%

33.37%, 29.80%, 27.19%, 23.58%, 20.02%, 14.90%, and 10.72%. The worst observed $M = 32$ case remains 20.48% wider than the exact original interval. The purpose of this suite is to quantify conservatism, not to search only for favorable instances.

E CONTENTION-FREE TIMING PROTOCOL AND COMPLETE SUMMARY

Fixtures are generated before measurement. No cases execute concurrently. Each $(\text{case}, \text{backend}, \text{repeat})$ executes in a fresh Python subprocess. OpenMP, OpenBLAS, MKL, NumExpr, and the LP solver are restricted to one thread. For each case/repetition, the oracle exact time is the smallest successfully certified time among: a full sparse HiGHS primal formulation, a reduced sparse HiGHS primal dual-simplex formulation, a reduced exact dual dual-simplex formulation, and a Gurobi reduced primal using one-thread dual-simplex reoptimization. The grid uses $n \in \{128, 256, 512\}$, trials $\{0, 4, 8, 90, 93, 96\}$, budgets $M \in \{8, 16, 32, 64\}$, and five repetitions, yielding 30 paired speedups per (n, M). Across the 90 exact case/repetition reference selections, the chosen backends are reduced HiGHS dual (45), Gurobi (26), reduced HiGHS primal (15), and full HiGHS primal (4). Table 6 reports the complete timing grid.

Table 6: Complete clean timing summary. IQR is the 25th–75th percentile of paired speedups.

n	M	median speedup	IQR	median compressed time (s)
128	8	38.76	33.02–46.58	0.0160
128	16	28.90	24.55–38.47	0.0221
128	32	12.94	11.97–14.72	0.0512
128	64	4.28	4.05–5.36	0.1484
256	8	199.27	179.09–218.06	0.0178
256	16	152.63	140.09–179.58	0.0236
256	32	70.99	65.73–80.19	0.0529
256	64	25.13	22.44–27.19	0.1523
512	8	576.67	472.36–1239.69	0.0223
512	16	482.31	398.29–1141.73	0.0275
512	32	236.17	201.31–570.74	0.0552
512	64	88.11	74.53–194.26	0.1484

E.1 LARGE-SUPPORT SCALING

A fixed-resolution implementation is evaluated through $n = 2048$. For $n = 10^7$, $M = 8$, a streaming specialization for uniform support takes 0.863 s for compression and 0.016 s for the certified surrogate LP, 0.879 s total, with 759 MB peak RSS. Analytical enumeration over the 80,900 surrogate vertices agrees with the LP to 1.11×10^{-16}. The original binary latent LP would contain 4×10^{14} variables. This experiment demonstrates surrogate scalability only; no exact original solve is performed at $n = 10^7$. Figure 5 summarizes the fixed-resolution scaling.

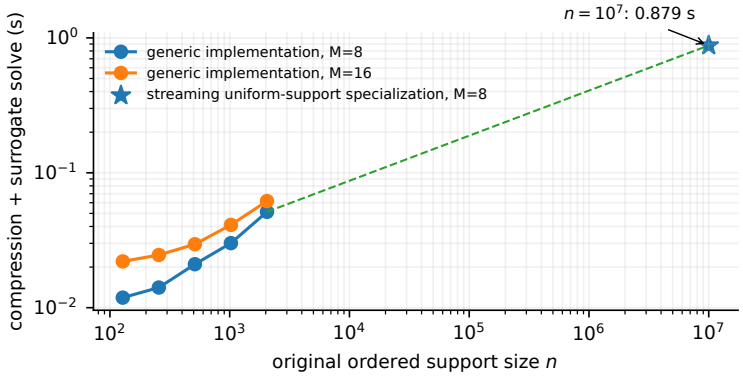

Figure 5: Fixed-resolution scaling. The largest point is a streaming implementation of the same barycentric map on uniform support and is not used as an empirical speedup against an unrun exact solve.

F FINITE-SAMPLE EXPERIMENT DETAILS

The finite-sample study uses $K = 3, J = 2, n = 16$, 12 independently generated feasible population laws, 100 repeated samples from each law, and total sample sizes $N \in \{500, 1000, 5000\}$. Instrument strata are sampled according to the experiment generator and the robust LP conditions on the resulting stratum sizes. Every sample produces three intervals: uncompressed, compressed at $M = 8$, and compressed at $M = 16$. Thus there are 3,600 independent sampling tasks and 10,800 reported intervals; Table 7 reports coverage and conservatism.

Table 7: Finite-sample coverage and conservatism. Width ratio is interval width divided by the population identified-set width.

N	M	method	runs	coverage	median ratio	p90 ratio
500	–	original	1200	1.000	1.883	1.970
500	8	compressed	1200	1.000	1.944	2.031
500	16	compressed	1200	1.000	1.906	1.993
1000	–	original	1200	1.000	1.699	1.779
1000	8	compressed	1200	1.000	1.750	1.827
1000	16	compressed	1200	1.000	1.725	1.803
5000	–	original	1200	1.000	1.384	1.433
5000	8	compressed	1200	1.000	1.418	1.479
5000	16	compressed	1200	1.000	1.402	1.459

All endpoint solutions satisfy primal, dual, and gap checks below 9×10^{-16} in this experiment. The observed 100% coverage is compatible with, but stronger than, the nominal lower guarantee because the simultaneous construction is conservative.

G JTPA EMPIRICAL BENCHMARK

The public JTPA dataset is distributed by the Mathematical Research Data Initiative and Sukjin Han under CC BY-NC 4.0 (Mathematical Research Data Initiative & Han, 2026). We use ‘Z2‘ as randomized assignment, ‘D2‘ as actual training receipt, and ‘earnings‘ as 30-month earnings, consistent with the documented variables and the established JTPA IV application (Abadie et al., 2002). After removing rows missing the required variables, 9,223 observations remain: 3,090 with $Z = 0$ and 6,133 with $Z = 1$. The observed training-receipt difference between assignment arms is 0.6407.

Before applying the certified stochastic compression, we deterministically define the finite-support empirical estimand used in this benchmark. Earnings are sorted and partitioned into approximately equal-frequency bins while keeping ties together; each bin is represented by its sample mean earnings. A requested 128 bins produces 113 distinct numerical support values because ties are not split. Raw earnings range from 0 to 155,760 dollars. This preprocessing step defines the empirical finite-grid outcome and is *not* claimed to preserve the exact continuous-earnings ATE; the certified stochastic compression is applied only after this finite-support target has been defined.

For each instrument arm, the 95% simultaneous Clopper–Pearson/Bonferroni cell box is constructed on the 113-point outcome-by-treatment table. The original and compressed robust LPs use the same confidence construction and certification tolerance. Table 8 reports the resulting intervals and bound-computation timings.

Table 8: JTPA finite-support robust intervals. Inflation is relative to the original robust interval; speedup uses measured bound-computation wall time.

method	interval (USD)	width	inflation	speedup
original	$[-42{,}726, 58{,}042]$	100,768	–	$1.0\times$
$M = 8$	$[-43{,}116, 58{,}085]$	101,201	0.430%	$107.9\times$
$M = 16$	$[-42{,}957, 58{,}085]$	101,042	0.272%	$78.1\times$
$M = 32$	$[-42{,}855, 58{,}055]$	100,909	0.140%	$33.8\times$
$M = 64$	$[-42{,}819, 58{,}046]$	100,865	0.096%	$12.4\times$

The original method wall time is 0.949 s; compressed wall times are 0.0088, 0.0122, 0.0281, and 0.0766 s for $M = 8, 16, 32, 64$. All compressed robust intervals contain the original robust interval. The largest JTPA LP certificate diagnostic is below 1.2×10^{-10}, comfortably inside the 2×10^{-8} acceptance threshold.

H DETAILED RELATED-WORK POSITIONING

Table 9 summarizes the closest conceptual lineages and the claim boundary.

Table 9: Closest conceptual lineages and the claim boundary.

Lineage	Representative work	Typical guarantee	Relation to this work
Sharp IV bounds	Balke–Pearl (Balke & Pearl, 1997); Boushehrian et al. (Boushehrian et al., 2026)	Exact/sharp bounds or exact representation limits	Motivates controlled relaxation; we do not claim to beat the lower bound for exact sharp formulas.
Discrete causal optimization	Duarte et al. (Duarte et al., 2024); Sachs et al. (Sachs et al., 2023)	Optimization or symbolic bounds on a fixed representation	Our layer changes representation before the same class of causal optimization.
Partial-ID inference	Chernozhukov et al. (Chernozhukov et al., 2007); Rosen (Rosen, 2008); Song et al. (Song et al., 2024)	Confidence regions/sets under sampling uncertainty	We add a simple nonasymptotic outer box that remains valid after target-preserving compression.
Coarsening	Gabriel et al. (Gabriel et al., 2025); Ou–Nabi (Ou & Nabi, 2026)	Effects or bias induced by coarsening	We preserve the designated finite-support target exactly model-by-model.
Dual quantization	Pagès–Wilbertz (Pagès & Wilbertz, 2012)	Mean-preserving stochastic splitting	Supplies the interpolation primitive; no causal identified-set certificate by itself.
Causal abstraction	Rubenstein et al. (Rubenstein et al., 2017); Beckers–Halpern (Beckers & Halpern, 2019); Geiger et al. (Geiger et al., 2025); Xia–Bareinboim (Xia & Bareinboim, 2025); Buchholz et al. (Buchholz et al., 2024)	Cross-level causal-query or mechanism correspondence	Our guarantee is target-specific, one-sided, and permits extra surrogate latent models.