SAFE PROBING OF CENSORED SHOWDOWN RECORDS: WHAT AN AGENT CAN LEARN AND HOW TO ACT ON IT

Anonymous authors
Paper under double-blind review

ABSTRACT

In poker-like imperfect-information games every hand leaves a public record of betting and payoffs, yet an opponent's private cards appear only at showdown. Weak hands fold, so showdown records systematically under-represent the hands an opponent model needs. Safe exploitation rests on the sequence-form exploitability certificate, which bounds what a response can lose against any opponent. It certifies safety but does not quantify observation: how much a probe may collect inside a safety budget, or what the censored records identify. Even the full terminal record can leave the opponent plan only partially identified. This paper couples the two questions and deploys against the residual ambiguity. A linear observation-rate objective composed with the certificate yields a *capacity linear program* (LP-C), whose optimum is the maximum target-showdown mass extractable inside a bounded exploitability budget under a nominal opponent model. The sharp identified set of opponent plans is a moment polytope: the opponent's *plan* may remain partially identified, while the probe's own *value* is point-identified from observable payoffs. Deploying robustly over the identified set is a single LP whose optimum certifies a one-sided finite-sample lower bound on deployed value at nominal coverage. Stacking probes refines the identified set and its guarantees. Under a fixed total sample budget, a dual-certificate analysis bounds the width of any deployment whose payoff vector lies in the certificate row space. It decomposes the effect of adding a probe into a weight term and a radius term. On the exhaustively enumerated Leduc Hold'em tree, the cell-polytope certificate over two probes attains a mean margin over the game value of 0.0895 against 0.0000 for a certificate built only from the two probes' mean payoffs. Over three probes, empirical–Bernstein intervals lift it to 0.0822 from 0.0062 under a Hoeffding box. The constructions apply to any finite perfect-recall two-player zero-sum game with hero-observable terminal cells.

1 INTRODUCTION

An agent that learns to play poker against an unknown opponent sees a public record of every hand: the betting history, the public cards, and the payoff. A showdown is the only event that exposes the opponent's private cards, and the hands that reach showdown are not a random sample. Weak hands fold when facing a bet, and a folded hand never shows its cards. Showdown records therefore under-represent the hands an opponent model needs most, and the filter is controlled by the opponent's own play. In the language of missing-data theory (Rubin, 1976; Little & Rubin, 2019), the observations are *missing not at random* (MNAR), and the ensuing bias is structural: whenever the censoring deletes the weak hands in the limit, the population observed-hand distribution differs from the prior by at least the prior mass of the censored hands (Theorem 4.1).

Opponent exploitation is usually formulated as a deployment problem (Nashed & Zilberstein, 2022). Given an opponent model, the agent fits a best response and adds a safety margin, so that the response is not itself exploited. Restricted Nash responses (Johanson et al., 2007), safe opponent exploitation (Ganzfried & Sandholm, 2015), and safety–exploitability trade-offs (Wang et al., 2011) all assume that a model is available. Data-driven variants fit one from game records (Johanson & Bowling, 2009), but those records are what the censoring above distorts. The certificate machinery that makes safe play practical (Zhang & Sandholm, 2020; 2021) bounds what a strategy can lose; it does not say how much a probe can observe inside the same budget. Two obstacles follow. The MNAR selection

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.

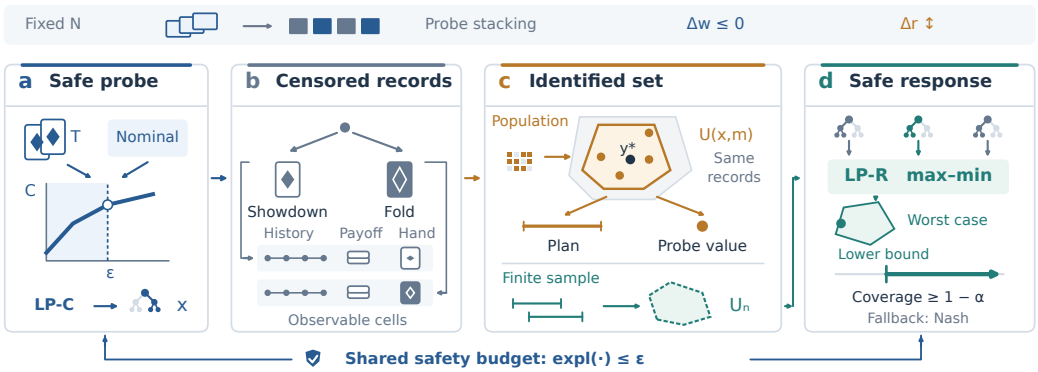

Figure 1: Safe probing under showdown censoring (schematic). **a**, LP-C maximizes target-showdown mass under a nominal opponent model within safety budget ε; its capacity curve is concave. **b**, the opponent's play selects which private hands are revealed (MNAR). Both showdown and fold records retain public history and hero payoff; only folded private hands stay hidden. **c**, population moments define $U(x, m)$, containing y^* and all observationally equivalent plans. Plan coordinates may span intervals while the probe's own value is a point. Finite-sample intervals instead define U_n. **d**, LP-R optimizes the worst-case value over U_n, yielding a value lower bound with coverage $\geq 1 - \alpha$; it falls back to Nash if no finite optimum exists. Bottom: probing and deployment share the deterministic exploitability budget. Top: with fixed total sample size, stacking weakly lowers the width bound's weight term ($\Delta_{\mathrm{w}} \leq 0$); its radius term is nonnegative when old-cell radii grow coordinatewise and has no general sign otherwise. The constructions apply to arbitrary two-player zero-sum extensive-form games with hero-observable terminal cells; poker illustrates this general scope.

biases precisely the quantity that a fitted model encodes. And even with the full terminal record in hand, the opponent plan may remain only partially identified: it becomes point-identified exactly when the observable cell moments pin a single plan.

This paper studies both questions together: how much a safe probe can observe, and what those observations identify. The coupling matters, because a probe that collects more labels can also remove the variation that identifies the opponent plan, so a deployment must stay safe against whatever the censored records still leave open. Both questions reduce to linear programs on the sequence-form exploitability certificate (Theorem 3.3). The target-showdown rate is a linear function of the hero plan, and composing it with the certificate yields LP-C: its optimum is the largest target-showdown mass extractable at safety level ε under a nominal opponent model, and its value is a concave, piecewise-affine capacity curve in ε (Theorems 5.1 and 5.2). A showdown record constrains the opponent only through the terminal-cell distribution it induces, so the sharp identified set of opponent plans is a moment polytope. The opponent's plan may remain only partially identified while the probe's own value, a moment of the hero-observable payoff, is point-identified (Theorems 6.4 and 6.7). Deploying robustly over that set under the same exploitability budget is a second LP, LP-R, whose optimum certifies a one-sided finite-sample lower bound on deployed value at nominal coverage (Theorem 6.8).

On Leduc Hold'em, whose game tree is enumerated exhaustively, each construction is verified numerically (Section 8). Inside an exploitability budget of $\varepsilon = 0.1$ the capacity probe collects $2.4\times$ the target showdowns of a Nash probe. With two probes, a certificate built on the cell polytope certifies a mean margin of 0.0895 over the game value, where a certificate built only from the probes' mean payoffs certifies 0.0000. The naive estimator's total-variation error grows as the censoring strengthens, from 0.126 at $\lambda = 1$ to 0.184 at $\lambda = 0$; propensity reweighting reduces it to 0.011–0.015 against 0.078–0.176 for the naive estimator.

Three constructions carry the analysis (Figure 1).

- **We propose** two linear programs that couple what a safe probe may observe with what it may safely deploy: LP-C maximizes the target-showdown mass extractable inside an exploitability budget under a nominal opponent model, and LP-R certifies a one-sided, finite-sample lower bound on deployed value over the identified set.

- **We characterize** what the censored records identify. The sharp identified set of opponent plans is a moment polytope; the opponent's plan may stay partially identified while the probe's own value is point-identified; and a probe that keeps positive weight on every reachable cell identifies the value of every deployment (Lemma 6.2 and Corollary 6.3).
- **We find** that stacking probes under a fixed total sample budget is not a free refinement. For a fixed deployment whose payoff vector lies in the certificate row space, the width bound splits exactly into a weight term, which weakly decreases, and a radius term that moves with the realized sampling radii; an added probe can tighten, loosen, or leave the bound unchanged (Theorem 7.2).

All three share one backbone, the sequence-form exploitability certificate, and one scope: an arbitrary two-player zero-sum extensive-form game with hero-observable terminal cells.

2 Related Work

The closest work, CS-RNR (Li & Huang, 2026), certifies an exploit from an anytime-valid confidence sequence on pooled action frequencies before restricting the response. Our certificates sit one step earlier, on what a safe probe can observe and what the censored records identify. The opponent evidence available to in-context opponent models (Jing et al., 2024a;b) and to LLM poker agents (Li et al., 2026; Zhuang et al., 2025; Wang et al., 2026) is the same censored record. This paper studies how that censoring couples what a safe probe may collect (Section 5) with what the records leave identified. It then deploys robustly against the residual ambiguity (Section 6).

Showdown records are missing data, and the sharp set of Section 6 follows the partial-identification (Manski, 2003; Beresteanu et al., 2009) and inverse-propensity (Horvitz & Thompson, 1952) lines. Appendix C collects the sequence-form duality, counterfactual-regret, and weighted-evaluation lines that the constructions reuse. It compares this work with CS-RNR and data-driven restricted responses on inputs, safety object, and guarantee form (Table 10).

3 Preliminaries

A two-player zero-sum extensive-form game Γ with perfect recall admits a *sequence form* (Zinkevich et al., 2007; Von Stengel, 1996). Player 1, the hero, plays an implementation plan $x \in X = \{x \geq 0 : Ex = e\}$, and player 2, the opponent, plays an implementation plan $y \in Y = \{y \geq 0 : Fy = f\}$. The plans are indexed by the sequence sets S_1 and S_2, with one flow constraint per information set. Each constraint system enforces flow conservation: a unit of realization mass enters at the root, and the mass that reaches an information set equals the mass that leaves through its actions. The payoff matrix A pairs the two plans, and $x^T Ay$ is the hero's expected payoff under (x, y). Throughout, Γ is finite and known, both players have perfect recall, the opponent plays a fixed plan $y^* \in Y$ across hands, hands are i.i.d., and payoffs are constant on cells (Definition 6.1).

The game value is $v^* = \max_{x \in X} \min_{y \in Y} x^T Ay$, and the *exploitability* of a hero plan is $\mathrm{expl}(x) = v^* - \min_{y \in Y} x^T Ay$: the value the hero loses against a best-responding opponent. Leduc Hold'em (Dahl, 2001) is the running example, with three ranks (Jack J, Queen Q, King K) and two suits.

The opponent is drawn from a *censoring family* $\{\mu_\lambda\}_{\lambda \in [0,1]}$. A baseline strategy σ_2, such as Nash, is perturbed by a parameter λ that scales the probability that weak hands call toward showdown; the removed mass is reweighted onto folding.

Two visibilities stay distinct throughout. A *cell* groups the terminals that the hero cannot tell apart, as Definition 6.1 makes precise. For every hand the hero records the complete cell: the betting history, the board, the payoff, and the hero's own hand. This holds for fold and showdown terminals alike, because the hero knows its own cards even when they are not publicized. The opponent's private hand is included only in showdown cells, and its *private-hand label* h_{opp} is therefore visible only at showdown. The prior P has full support over the finite hand set $\mathcal{H}$. The probability that a private hand h_{opp} reaches showdown under the behavioral strategy σ_x induced by the hero plan x is $S_\lambda(x, h_{\mathrm{opp}}) = P_{(\sigma_x, \mu_\lambda)}(\mathrm{showdown} \mid h_{\mathrm{opp}})$.

Assumption 3.1 (Deleted weak hands avoid showdown). For a target set $T \subseteq \mathcal{H}$ (the weak hands), $S_\lambda(x, h_{\text{opp}}) \to 0$ as $\lambda \to 0$ for every $h_{\text{opp}} \in T$, and there are $\lambda_0, \underline{s} > 0$ with $S_\lambda(x, h_{\text{opp}}) \geq \underline{s}$ for all $0 < \lambda < \lambda_0$ and all $h_{\text{opp}} \notin T$.

Definition 3.2 (Observation distribution)**.** Under a fixed hero plan x and opponent μ_λ with $\sum_{h'} P(h')S_\lambda(x, h') > 0$, the *observation distribution* is $q_\lambda(h_{\text{opp}}) = P(h_{\text{opp}})S_\lambda(x, h_{\text{opp}}) / \sum_{h'} P(h')S_\lambda(x, h')$, the conditional distribution of the opponent hand given that a showdown occurred.

Theorem 3.3 (Sequence-form exploitability certificate)**.** *For any hero plan $x \in X$, with I_2 the information sets of the opponent, q a dual certificate with one coordinate per row of F (the root constraint and each set in I_2), and $\varepsilon \geq 0$ the exploitability budget,*

$$\text{expl}(x) \leq \varepsilon \iff \exists q \in \mathbb{R}^{|I_2|+1} : \; F^T q \leq A^T x, \quad f^T q \geq v^* - \varepsilon.$$

Strong LP duality identifies $\min_{y \in Y} x^T A y$ with the maximum of $f^T q$ over dual certificates q with $F^T q \leq A^T x$. This duality is the standard safety mechanism behind safe exploitation (Ganzfried & Sandholm, 2015; Johanson et al., 2007) (Appendix A). Section 5 composes the certificate with an observation objective, and Section 6 reuses it as the safety constraint of a robust deployment.

4 Showdown Records Are Missing Not at Random

The observation distribution of Definition 3.2 is the raw material of any learner that works from showdown records; within the censoring family it is a distorted version of the prior. Total variation (TV) compares the two hand distributions as probability laws, $\text{TV}(q, P) = \frac{1}{2} \sum_h |q(h) - P(h)|$, and it admits a clean lower bound here.

Theorem 4.1 (MNAR bias lower bound)**.** *Fix a hero plan x and a target set with $0 < P(T) < 1$. Under Assumption 3.1, $\liminf_{\lambda \to 0} \text{TV}(q_\lambda, P) \geq P(T)$; in particular, for any $\delta > 0$ there is a censoring level such that $\text{TV}(q_\lambda, P) \geq P(T) - \delta$: asymptotically, the population law of naive showdown frequencies lies at least $P(T)$ from the prior.*

The filter removes the weak hands and reweights the rest, so the observed distribution stays at least the removed mass from the prior; the proof is in Appendix A.

The bound is tight under balanced selection within T^c (Corollary D.1). The event bound $\text{TV}(q_\lambda, P) \geq |q_\lambda(T) - P(T)|$ holds at every censoring level; Appendix D records its Leduc values.

The bias is correctable in principle. The selection probability is a known function of the hero behavior and of the *nominal* censoring model $\bar{\mu}$ that the agent chooses. Inverse-propensity weighting (IPW) (Horvitz & Thompson, 1952) divides each observed indicator by its showdown propensity. The true opponent is a separate object, treated in Section 6; a misspecified nominal model biases the weights.

Theorem 4.2 (IPW unbiasedness and Hoeffding certificate)**.** *Let $\tau = P(h_{\text{opp}} \in T)$ denote the private-hand marginal, that is, the prior mass of the target hands. Fix the probing strategy σ_{probe} and the opponent instance μ, and let $p(h_{\text{opp}}) = P(\text{showdown} \mid h_{\text{opp}})$ be the showdown propensity of a private hand, with $p_{\min} = \min_{h_{\text{opp}}} p(h_{\text{opp}}) > 0$. Write $S_i \in \{0, 1\}$ for the showdown indicator of hand i. For n independent hands define $\hat{\tau}_{\text{IPW}} = n^{-1} \sum_i \mathbf{1}\{h_{\text{opp}}^{(i)} \in T, S_i = 1\}/p(h_{\text{opp}}^{(i)})$. Then $\mathbb{E}[\hat{\tau}_{\text{IPW}}] = \tau$ and, for every $\delta > 0$ and $\eta \in (0, 1)$,*

$$P\big(|\hat{\tau}_{\text{IPW}} - \tau| \geq \delta\big) \leq 2 \exp\big(-2n\delta^2 p_{\min}^2\big).$$

Thus error at most δ holds with probability at least $1 - \eta$ whenever $n \geq p_{\min}^{-2} \ln(2/\eta)/(2\delta^2)$.

Conditioning on the private hand cancels the propensity. The propensity is a function of the hero plan, so an adaptive probe changes it over time. A *predictable* probe, whose hand-t behavior is fixed before hand t, keeps the reweighted average unbiased (Theorem D.2, Appendix D).

Reweighting restores the hand marginals; it does not recover opponent behavior, since a fold never reveals its private hand. The opponent plan becomes point-identified exactly when the observed cell moments pin a single plan (Section 6), so marginal error and deployed value are reported separately.

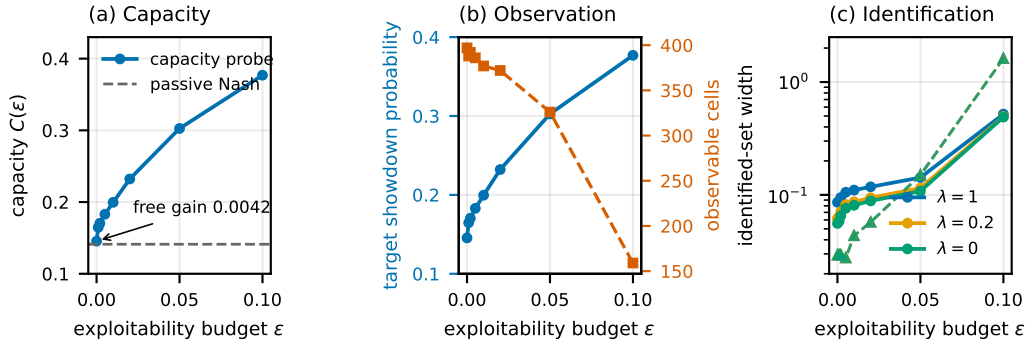

Figure 2: Population quantities on Leduc, $T = \{\mathrm{J}, \mathrm{Q}\}$. (a) Capacity $C(\varepsilon; T)$ versus the exploitability budget at $\lambda = 0$: $C(0) = 0.145$, $C(0.1) = 0.377$, slope 12.50 over $[0, 0.002]$, and a free gain of 0.0042 over Nash at $\varepsilon = 0$. (b) The target-showdown rate rises with ε while the number of observable cells falls from 397 to 159. (c) Identification: the mean plan width of the sharp identified set (solid) grows from 0.062 to 0.495 at $\lambda = 0.2$, and the largest cross-deployment payoff width (dashed; nearly identical across λ) from 0.029 to 1.616.

5 The Reveal Budget

The observation resource is the target-showdown rate: the probability mass of hands in T that reach showdown. Under a fixed censoring parameter the opponent's behavior is exogenous, so this rate is a linear function of the hero plan. The opponent behavior entering that rate is a *nominal* censoring model $\bar{\mu} = (\sigma_{\text{base}}, \lambda)$ fixed by the analyst, while the true opponent y^* stays arbitrary.

Theorem 5.1 (Capacity LP correctness)**.** *Fix the censoring parameter λ. The observation rate $\theta(x) = P_{(\sigma_x, \mu_\lambda)}(\text{showdown} \wedge h_{\text{opp}} \in T)$ is a linear function of the hero plan, $\theta(x) = c^T x$, where $c[a] = \sum_{t:\, h_{\text{opp}}(t) \in T,\ a_t = a} y^{\mu_\lambda}[b_t]\, p(t)$ accumulates the censored opponent's reach and the chance weight over the showdown terminals t reached through hero sequence a. Here $y^{\mu_\lambda}[b_t]$ is the opponent's sequence-form realization weight on its own sequence b_t at terminal t, and $p(t)$ is the chance weight of the deal. Hence $C(\varepsilon; T) = \max\{\theta(x) : x \in X,\ \text{expl}(x) \le \varepsilon\}$ is the optimal value of a linear program.*

$$\begin{aligned} \text{(LP-C)} \qquad & \max\, c^T x \\ & \text{s.t. } Ex = e, \quad x \ge 0, \\ & \qquad F^T q \le A^T x, \quad f^T q \ge v^* - \varepsilon. \end{aligned}$$

Theorem 5.2 (Structure of $C(\varepsilon; T)$)**.** *$C(\cdot; T)$ is nondecreasing, concave, and piecewise affine on $[0, \infty)$. The right-derivative at $\varepsilon = 0$, $\rho = C'_+(0; T)$, exists and $C(\varepsilon; T) = C(0; T) + \rho\, \varepsilon$ on an initial interval $[0, \varepsilon_0]$, with $\varepsilon_0 > 0$ depending on the instance $(\Gamma, T, \bar{\mu})$.*

Figure 2 shows the curve; Appendix A gives the proof. On Leduc the realized slope 12.50 over $[0, 0.002]$ lower-bounds ρ, the per-unit price of target showdowns in exploitability.

Corollary 5.3 (Capacity at zero is equilibrium collection)**.** *Let $\mathcal{E} = \{x \in X : \text{expl}(x) = 0\}$, the equilibrium polytope of hero plans (nonempty, compact, convex). Then $C(0; T) = \max_{x \in \mathcal{E}} \theta(x)$, and $C(0; T) = 0$ iff $\theta \equiv 0$ on $\mathcal{E}$, i.e. iff no equilibrium hero plan places positive mass on a hero sequence from which the nominal opponent $\bar{\mu}$ reaches a target-hand showdown with positive probability.*

Lemma 5.4 (Exploitability–capacity lever)**.** *For every safe probe x with $\text{expl}(x) \le \varepsilon$, $\theta(x) \le C(0; T) + \rho\, \varepsilon$, where $\rho = C'_+(0; T)$ is the one-sided derivative of Theorem 5.2. Equivalently, ρ is the shadow price of the exploitability certificate $f^T q \ge v^*$ at $\varepsilon = 0$: the marginal target-showdown mass an exploitability budget buys, never exceeding ρ per unit.*

6 SHARP PARTIAL IDENTIFICATION AND ROBUST-SAFE DEPLOYMENT

Section 5 asked how much target-showdown mass a safe probe can collect. The censored records identify the opponent only through the terminal *cells* they realize; this section characterizes that identification and derives a deployment safe against the residual ambiguity.

Definition 6.1 (Observable terminal cells and moment map)**.** Each terminal t is hero-observable up to its *cell* $\kappa(t)$: a showdown terminal records the full profile (both private hands, the board, and the betting history); a fold terminal records everything except the opponent's private hand, which is marginalized out. The hero payoff is constant on cells, $u(t) = u_\kappa$ for all $t \in \kappa$, which holds in poker-like games. For a probe x, the moment matrix $\Phi_x \in \mathbb{R}^{K \times |S_2|}$ has entries $\Phi_x[\kappa, j] = \sum_{t \in \kappa:\, b_t = j} p(t)\, x[a_t]$, with K the number of cells and a_t, b_t the hero and opponent sequences at terminal t. Hence $(\Phi_x y)_\kappa = P_{(x,y)}(\text{terminal cell} = \kappa)$ is the cell distribution induced by (x, y); the cells partition the terminals, hence $\sum_\kappa (\Phi_x y)_\kappa = 1$. The cell is the complete record of a hand as observed by the agent, for every terminal, fold or showdown. It carries the hero's own payoff, while the opponent's private hand is a label that enters only the showdown cells.

Lemma 6.2 (Row scaling)**.** *Let a_κ be the hero sequence of cell κ and let $B[\kappa, j] = \sum_{t \in \kappa:\, b_t = j} p(t)$, which does not depend on the probe. Then $\Phi_x = D_x B$ with $D_x = \operatorname{diag}(x[a_\kappa])_\kappa$, and $A^T z = B^T D_z u$ for every $z \in X$, with u the cell payoff vector of Lemma 6.6. Hence $A^T z \in \operatorname{rowspace}(B)$ for every deployment z, and $\operatorname{rowspace}([F; \Phi_x])$ depends on x only through the support $\{a_\kappa : x[a_\kappa] > 0\}$.*

Corollary 6.3 (Full-support probes cover every deployment)**.** *If $x[a_\kappa] > 0$ on every cell with $B[\kappa, \cdot] \neq 0$, then $\operatorname{rowspace}([F; \Phi_x]) \supseteq \operatorname{rowspace}(B)$, so $A^T z \in \operatorname{rowspace}([F; \Phi_x])$ for every deployment z: every deployment value is point-identified on $U(x, m)$ (Theorem 6.7) and the width certificate of Theorem 7.2 is finite for every z.*

Theorem 6.4 (Sharp identified set)**.** *Fix a probe x and let $m = \Phi_x y^*$ be the population cell moments. Define $U(x, m) = \{y \geq 0 : Fy = f,\ \Phi_x y = m\}$. Then:*

- *(i) $U(x, m)$ is a nonempty compact convex polytope containing y^*;*
- *(ii) $U(x, m)$ equals the observational equivalence class $\{y \in Y : \Phi_x y = \Phi_x y^*\}$, $\Phi_x y$ being the law of the observable record under (x, y): every $y \in U(x, m)$ induces the same observable-record distribution as y^*, so no estimator on observable data can distinguish them, and every $y \in Y \setminus U(x, m)$ is refuted by the population moments. Sharpness is relative to the observation model of Definition 6.1, which records the* entire *cell distribution;*
- *(iii) $U(x, m) \subseteq y^* + \mathcal{N}$ with $\mathcal{N} = \{d : Fd = 0,\ \Phi_x d = 0\}$, hence $\dim U(x, m) \leq \dim \ker[F; \Phi_x]$;*
- *(iv) if $\ell^T = a^T F + b^T \Phi_x$ for some (a, b), equivalently $\ell \in \operatorname{rowspace}([F; \Phi_x])$, then $\ell^T y = a^T f + b^T m$ is constant on $U(x, m)$: the functional is point-identified. The converse can fail at the boundary of the positive orthant, so per-coordinate widths are measured by linear programming over the identified set (Appendix E).*

Proposition 6.5 (Rowspace governs identification)**.** *Fix ε, the nominal model $\bar{\mu}$ with opponent plan $\bar{y}$, the true opponent y^* with moments $m = \Phi_x y^*$, $m' = \Phi_{x'} y^*$, and $P_\varepsilon = \{x \in X : \operatorname{expl}(x) \leq \varepsilon\}$; write $R(x) = \operatorname{rowspace}([F; \Phi_x])$. (i) With $\chi_T \in \{0, 1\}^K$ the indicator of the showdown cells recording $h_{\text{opp}} \in T$, the capacity $\theta(x) = \chi_T^T \Phi_x \bar{y}$ is one linear functional of the* nominal *cell moments, whereas the identified set $U(x, m)$ (Theorem 6.4) is cut by all the affine constraints $Fy = f$, $\Phi_x y = m$ on the* true *opponent; every functional in $R(x)$ is point-identified on it, and only these when $U(x, m)$ contains a coordinatewise strictly positive point. (ii) If $x, x' \in P_\varepsilon$ satisfy $R(x) \not\subseteq R(x')$ and $U(x', m')$ contains a coordinatewise strictly positive point, then some $\ell \in R(x) \setminus R(x')$ is point-identified on $U(x, m)$ yet has positive LP-measured width on $U(x', m')$.*

Appendix B instantiates this contrast on Leduc, where two equal-budget probes differ in the dimension of the kernel of $[F; \Phi_x]$ and in mean certificate margin (Table 9). Because the hero payoff is constant on cells, the probe's own value is itself a cell moment, which the next lemma records.

Lemma 6.6 (Value factorization)**.** *Under cell-measurable payoffs, $A^T x = \Phi_x^T u$ for every probe x; hence $x^T A y = u^T \Phi_x y$ for every y.*

Theorem 6.7 (The probe's own value is point-identified)**.** *On $U(x, m)$ the probe's expected payoff is constant, $x^T Ay = u^T m$ for every $y \in U(x, m)$. More generally, any deployment z with $A^T z \in \mathrm{rowspace}([F; \Phi_x])$ has point-identified value on $U(x, m)$.*

Partial identification of the opponent's *plan* (Manski, 2003; Beresteanu et al., 2009) therefore coexists with exact identification of the probe's own *value*.

Theorem 6.8 (Single-LP robust-safe response)**.** *Let $U = \{y \geq 0 : \ Fy = f, \ Gy \leq g\}$ be nonempty (hence compact, as a subset of Y). Two instances matter: the population set of Theorem 6.4 with $G = [\Phi_x; -\Phi_x]$, $g = [m; -m]$, and a finite-sample set U_n with $g = [u_n; -l_n]$, where u_n and l_n are the upper and lower simultaneous confidence bounds on the cell moments. Then:*

(a) for every hero plan x, strong LP duality gives

$$\min_{y \in U} x^T Ay = \max\{\, f^T w + g^T s : \ F^T w + G^T s \leq A^T x, \ s \leq 0 \,\};$$

(b) the robust-safe value $R(\varepsilon) = \max\{\min_{y \in U} x^T Ay : \ x \in X, \ \mathrm{expl}(x) \leq \varepsilon\}$ is attained by the single LP (LP-R) over (x, w, s, q):

$$\begin{aligned} \max \quad & f^T w + g^T s \\ \text{s.t. } & Ex = e, \ x \geq 0, \qquad F^T w + G^T s \leq A^T x, \ s \leq 0, \\ & F^T q \leq A^T x, \quad f^T q \geq v^* - \varepsilon; \end{aligned}$$

(c) $R(\varepsilon) \geq v^*$ for every $\varepsilon \geq 0$ and every nonempty $U \subseteq Y$;

*(d) **(one-sided finite-sample certificate)** if $U = U_n$ is the data-dependent set above with $P(y^* \in U_n) \geq 1 - \alpha$, and $(\hat{x}, \hat{w}, \hat{s}, \hat{q})$ solves LP-R, falling back to a Nash plan on paths where LP-R has no finite optimum (LP-R is always feasible, and those are exactly the paths with $U_n = \emptyset$, where it is unbounded; they lie in $\{y^* \notin U_n\}$), then*

$$P\big(\hat{x}^T Ay^* \ \geq \ f^T \hat{w} + g^T \hat{s}\big) \ \geq \ 1 - \alpha,$$

while $\mathrm{expl}(\hat{x}) \leq \varepsilon$ holds deterministically on every path. On the fallback branch the plan is Nash and the certificate is its equilibrium dual value v^*, which satisfies $x^T Ay^* \geq v^*$ for every opponent y^*.

Part (d) is pathwise: on the coverage event the bound holds for the realized $\hat{x}$ and any dual-feasible (w, s), with no sample splitting (Appendix A). Adding probes refines both the identified set and its guarantees monotonically.

Theorem 6.9 (Monotone multi-probe refinement)**.** *Let probes $x_1, \ldots, x_k$ generate sets $U_1, \ldots, U_k$ (population moment sets or confidence sets), and let $U^{(k)} = \bigcap_{i \leq k} U_i$.*

(a) in the population case $y^* \in U^{(k)}$, so $U^{(k)} \neq \emptyset$, and the stacked set is again of the form treated by Theorem 6.8 with $\Phi = [\Phi_{x_1}; \ldots; \Phi_{x_k}]$;

(b) $U^{(k+1)} \subseteq U^{(k)}$;

(c) whenever $U^{(k+1)} \neq \emptyset$ (automatic in the population case; the coverage event for confidence sets), for every deployment z the identified value interval $I_k(z) = [\min_{y \in U^{(k)}} z^T Ay, \ \max_{y \in U^{(k)}} z^T Ay]$ nests, $I_{k+1}(z) \subseteq I_k(z)$, and the robust-safe value $R_k(\varepsilon)$ is nondecreasing in k;

*(d) **(finite sample)** if $P(y^* \in U_i) \geq 1 - \alpha_i$ with $\sum_i \alpha_i = \alpha$, then $P(y^* \in U^{(k)}) \geq 1 - \alpha$ (Bonferroni). Each U_i may be built from $n_i \geq 2$ i.i.d. hands, at level $\alpha_i \in (0, 1)$, as a simultaneous empirical–Bernstein box (Maurer & Pontil, 2009; Freedman, 1975): for a cell frequency $\hat{m}$ the radius is $\sqrt{2\hat{m}(1 - \hat{m})L_i/(n_i - 1)} + \frac{7}{3} L_i/(n_i - 1)$ with $L_i = \ln(8K_i/\alpha_i)$ and K_i the number of cells of probe i, fixed before sampling, blended with a Hoeffding radius (Hoeffding, 1963) so the realized radius never exceeds either candidate. The $n_i - 1$ denominator is the Maurer–Pontil convention: their unbiased variance V_n equals $\frac{n_i}{n_i - 1} \hat{m}(1 - \hat{m})$ for Bernoulli observations. The blend does not disturb the pathwise arguments of Theorem 6.8, which run on the realized set U_i.*

Figure 2(b,c) shows the population tradeoff of Proposition 6.5: concentrating showdowns into fewer cells loses plan identification even as it gains hand labels. Lemma 6.2 locates that tradeoff in the support of the probe. Concentration rescales the rows of Φ_x and can drop them entirely, so a support-limited probe can fail the rowspace hypothesis of Theorem 7.2 for deployments it does not cover. A probe with positive weight on every reachable cell covers all of them (Corollary 6.3). Exploitability is convex and capacity is linear in the probe plan, so the mixture $(1-s)x_{\mathrm{cap}} + sx_{\mathrm{fs}}$, with x_{cap} capacity-optimal and x_{fs} safe full-support, stays safe and retains $1-s$ of the capacity.

The tradeoff is finite-sample. A rare cell enters the certificate with weight proportional to $1/x[a_\kappa]$, a weight that follows from the row-scaled representation. Its empirical–Bernstein radius has a variance part that grows like $\sqrt{\widehat{m}_\kappa}$ and an additive $7L_i/[3(n_i-1)]$ part that does not vanish on rare or empty cells. Section 8 measures both.

7 Fixed-Budget Probe Stacking

Population stacking refines the identified set monotonically by Theorem 6.9. Under a fixed total budget, an additional probe is paid for by shrinking every other probe's sample. The logarithmic factor that sets the per-cell radii, $\ln(8K_i/\alpha_i)$, grows with the number of probes when α is split equally, while the realized radii need not grow coordinatewise. The two channels of the width bound separate exactly.

Lemma 7.1 (Difference inclusion). *For the stacked box $U^{(k)}$ of Theorem 6.9 with realized per-cell radii $r_{i,\kappa} = (u_{i,\kappa} - \ell_{i,\kappa})/2$, every difference $d = y - y' \in U^{(k)} - U^{(k)}$ satisfies $Fd = 0$ and $|\Phi^{(k)} d| \le 2r$, where $\Phi^{(k)} = [\Phi_{x_1}; \ldots; \Phi_{x_k}]$ and r stacks the $r_{i,\kappa}$.*

Theorem 7.2 (L1-dual width bound). *Let $U^{(k)} \neq \emptyset$ and let $W_k(z) = \max_{y,y' \in U^{(k)}} z^T A(y-y')$ be the width of deployment z over the stacked set, and let r be the realized radius vector of Lemma 7.1. If $A^T z \in \mathrm{rowspace}([F; \Phi^{(k)}])$,*

$$W_k(z) \;\le\; 2 \cdot \min\Big\{ \textstyle\sum_{i,\kappa} r_{i,\kappa} |w_{i,\kappa}| :\; F^T\lambda + \Phi^{(k)T} w = A^T z \Big\},$$

where the dual variable λ runs over the rows of F and is unrelated to the censoring parameter of the same name. The certificate value is $+\infty$ otherwise (the width itself stays finite, $U^{(k)}$ being compact). The bound is deterministic on every realized nonempty set and is computed from the realized $(F, \Phi^{(k)}, r)$ alone, without y^*.

Corollary 7.3 (Rank redundancy: no benefit at the population level). *If $\mathrm{rowspace}([F; \Phi^{(k+1)}]) = \mathrm{rowspace}([F; \Phi^{(k)}])$, then the exact-moment sets coincide, hence every population width is unchanged: $W_k(z) = W_{k+1}(z)$ for all z.*

Corollary 7.4 (Fixed-budget criterion). *Under a fixed total budget ($n_i = N/k$ per probe), fix a deployment z with $A^T z \in \mathrm{rowspace}([F; \Phi^{(k)}])$ and let $S^{(k)}$ be the affine set $\{(\lambda, w) :\; F^T\lambda + \Phi^{(k)T} w = A^T z\}$. Define the* weight term Δ_{w} *(the change in the minimum with the radii held at their level-k values, the new cells receiving coefficient 0) and the* radius term Δ_{r} *(the change over the same feasible set $S^{(k+1)}$ when the radii move from level k to level $k+1$):*

$$\Delta_{\mathrm{w}} = \min_{S^{(k+1)}} \|r^{(k)} \cdot w\|_1 - \min_{S^{(k)}} \|r^{(k)} \cdot w\|_1, \qquad \Delta_{\mathrm{r}} = \min_{S^{(k+1)}} \|r^{(k+1)} \cdot w\|_1 - \min_{S^{(k+1)}} \|r^{(k)} \cdot w\|_1.$$

Then the change of the ℓ_1 objective of Theorem 7.2, half the change of its bound, telescopes exactly, $\min_{S^{(k+1)}} \|r^{(k+1)} \cdot w\|_1 - \min_{S^{(k)}} \|r^{(k)} \cdot w\|_1 = \Delta_{\mathrm{w}} + \Delta_{\mathrm{r}}$, with $\Delta_{\mathrm{w}} \le 0$; and $\Delta_{\mathrm{r}} \ge 0$ provided the realized radii of the pre-existing cells grow coordinatewise, $r^{(k+1)}_{i,\kappa} \ge r^{(k)}_{i,\kappa}$ for every old cell. Stacking strictly tightens the bound iff $\Delta_{\mathrm{r}} < -\Delta_{\mathrm{w}}$.

Zero-padding the old certificate gives $\Delta_{\mathrm{w}} \le 0$. When the added probe is the deployment itself, Lemma 6.6 supplies a certificate supported only on the new cells, whose level-k radii are zero. The intermediate optimum then vanishes and the criterion reduces to comparing the two bounds. Corollary 6.3 covers the other extreme: one full-support probe makes the rowspace hypothesis of Theorem 7.2 vacuous for every deployment. The stacking audit has a tightening family F_A and a loosening family F_B (Table 6).

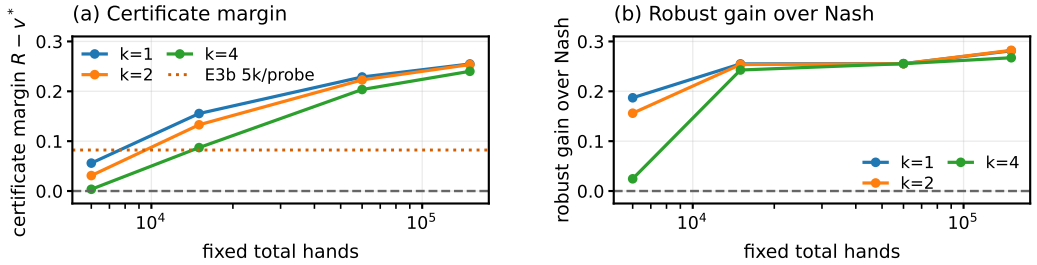

Figure 3: Certified value under a fixed total budget. (a) Mean certificate margin $R - v^*$ against the fixed total number of hands on a logarithmic axis for $k = 1, 2, 4$ probes; the dotted line marks the E3b baseline of $5{,}000$ hands per probe. (b) Mean robust true-value gain over Nash.

Table 1: Headline quantities on Leduc Hold'em.

Block	Headline quantity
E1–E2	naive TV 0.126 (λ=1), 0.184 (λ=0); IPW TV 0.011–0.015 against naive 0.078–0.176
E3	rank $302/308$; own value $\leq 10^{-15}$; coverage $50/50$
E3d–E3e	rank $202 \to 337$; cell polytope 0.0895 against payoff-only 0.0000
E4–E5	bias -0.0014 against naive -0.074; F_A tightens $40/40$
E6–E8	capacity rate 0.388; LP-R margin 0.093 against Nash 0.030; 0.0806 over three probes

8 Experiments

All experiments are on Leduc Hold'em with three card ranks and two suits, target $T = \{\mathrm{J}, \mathrm{Q}\}$ unless stated otherwise. The tree is exhaustively enumerated, sequence-form LPs are solved by a general-purpose solver, and every stated number is cross-checked to explicit tolerances (Appendix E). Table 1 collects each block's headline quantity. The per-block tables for blocks E1–E8, the theorem each verifies, and the full configuration are in Appendix B.

E1–E2: Observation bias and its correction. The naive estimator is biased toward the hands that reach showdown, and its TV error grows with the censoring strength; the event bound of Section 4 is attained at $\lambda \leq 0.2$. Leduc's card marginal is known in advance, so the recovery target is known too and an IPW failure cannot hide inside model error. Rank-conditioned reweighting reduces marginal estimation error in all six censoring configurations, and a one-sided paired Wilcoxon signed-rank test over the 50 seeds favors it in every seed, with exact null floor 2^{-50}.

E3: Identification, certificates, and probe design. Concentrating showdown mass into fewer cells costs identification rank, so a deployment carried over to another probe's moments is ambiguous while every probe's own value is point-identified to machine precision (Figure 2c). Mixing the capacity probe with a safe full-support probe restores full rank at a weight that costs almost no capacity, the population prediction of Corollary 6.3 (Table 4). Replacing the cell constraints by one mean-payoff interval per probe collapses the margin, and restoring the sampling radius to zero recovers only part of it, so the loss is representational and not a sampling artifact (Table 5). The empirical–Bernstein box dominates Hoeffding by a factor of 13.4 in margin, and the total-showdown-rate probe improves the margin furthest among six equal-budget designs (Figure 3, Table 9).

E4–E7: Budget, adaptation, and baselines. The adaptive probe alternates between two plans: the low-showdown-rate Nash plan and the high-showdown-rate capacity-optimal probe at $\varepsilon = 0.1$. It is fixed within a block of 200 hands and updates only at block boundaries from hands already observed, so it is predictable in the sense of Theorem D.2. Under it the IPW estimator stays near-unbiased while the naive and stale-propensity estimators carry the MNAR bias (Table 11). The width bound tightens exactly when the radius term stays below the weight gain. The two stacking families separate those terms: a rank-redundant addition tightens the bound, while a rank-increasing one can loosen it even as the width falls (Table 6). The uniform probe, which randomizes at every hero information set, has a lower target-showdown rate than Nash at far higher exploitability, and the capacity probe at $\varepsilon = 0.1$ gains $2.4\times$ the Nash rate inside the safety budget (Lemma 5.4).

AI USE STATEMENT

In this work, we used generative AI tools to implement and run the experiments, analyze their results, and edit the manuscript; we did not use them for the theoretical model, the claims, or the proofs. We reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The proofs of all theoretical results are given in Section A. The experiments use fixed random seeds. The code behind the reported experiments is provided in the anonymous supplementary material: one entry point per experiment, each stating its configuration and the values it should reproduce, with the bundled Leduc game engine, the implementations as compiled modules, and the frozen logs of the recorded runs.

REFERENCES

Kenshi Abe and Yusuke Kaneko. Off-policy exploitability-evaluation in two-player zero-sum markov games. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 78–87, 2021.

Jacob D. Abernethy, Kareem Amin, and Ruihao Zhu. Threshold bandits, with and without censored feedback. In *NIPS*, pp. 4889–4897, 2016. URL http://papers.nips.cc/paper/6149-threshold-bandits-with-and-without-censored-feedback.

András Antos, Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games. *Theoretical Computer Science*, 473:77–99, 2013.

Arie Beresteanu, Ilya Molchanov, and Francesca Molinari. Sharp identification regions in models with convex predictions: games, individual choice, and incomplete data. Technical report, cemmap working paper, 2009.

Martino Bernasconi-de Luca, Federico Cacciamani, Simone Fioravanti, Nicola Gatti, Alberto Marchesi, and Francesco Trovò. Exploiting opponents under utility constraints in sequential games. *Advances in Neural Information Processing Systems*, 34:13177–13188, 2021.

Darse Billings, Denis Papp, Jonathan Schaeffer, and Duane Szafron. Opponent modeling in poker. *Aaai/iaai*, 493(499):105, 1998.

Michael Bowling, Michael Johanson, Neil Burch, and Duane Szafron. Strategy evaluation in extensive games with importance sampling. In *Proceedings of the 25th international conference on Machine learning*, pp. 72–79, 2008.

Noam Brown and Tuomas Sandholm. Safe and nested subgame solving for imperfect-information games. *Advances in neural information processing systems*, 30, 2017.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456): 885–890, 2019.

Neil Burch, Martin Schmid, Matej Moravcik, Dustin Morill, and Michael Bowling. Aivat: A new variance reduction technique for agent evaluation in imperfect information games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 885–889, 2018.

Fredrik A Dahl. A reinforcement learning algorithm applied to simplified two-player texas hold'em poker. In *European Conference on Machine Learning*, pp. 85–96. Springer, 2001.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Haobo Fu, Ye Tian, Hongxiang Yu, Weiming Liu, Shuang Wu, Jiechao Xiong, Ying Wen, Kai Li, Junliang Xing, Qiang Fu, and Wei Yang. Greedy when sure and conservative when uncertain about the opponents. In *International Conference on Machine Learning*, volume 162, pp. 6829–6848. PMLR, 2022.

Sam Ganzfried and Tuomas Sandholm. Safe opponent exploitation. *ACM Transactions on Economics and Computation (TEAC)*, 3(2):1–28, 2015.

Sam Ganzfried and Qingyun Sun. Bayesian opponent exploitation in imperfect-information games. In *CIG*, pp. 1–8. IEEE, 2018.

Zhenxing Ge, Zheng Xu, Tianyu Ding, Linjian Meng, Bo An, Wenbin Li, and Yang Gao. Safe and robust subgame exploitation in imperfect information games. In *Forty-first International Conference on Machine Learning*, 2024.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Yuheng Jing, Kai Li, Bingyun Liu, Yifan Zang, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. Towards offline opponent modeling with in-context learning. In *The Twelfth International Conference on Learning Representations*, 2024a.

Yuheng Jing, Bingyun Liu, Kai Li, Yifan Zang, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. Opponent modeling with in-context search. *Advances in Neural Information Processing Systems*, 37:61549–61591, 2024b.

Michael Johanson and Michael Bowling. Data biased robust counter strategies. In *Artificial Intelligence and Statistics*, pp. 264–271. PMLR, 2009.

Michael Johanson, Martin Zinkevich, and Michael Bowling. Computing robust counter-strategies. *Advances in neural information processing systems*, 20, 2007.

Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte carlo sampling for regret minimization in extensive games. *Advances in neural information processing systems*, 22, 2009.

Boning Li and Longbo Huang. Efficient online pruning and abstraction for imperfect information extensive-form games. In *The Thirteenth International Conference on Learning Representations*, 2025.

Boning Li and Longbo Huang. Agents that certify their own exploits: Confidence-scheduled restricted responses for safe opponent exploitation. *arXiv preprint arXiv:2607.28520*, 2026.

Boning Li, Zhixuan Fang, and Longbo Huang. Rl-cfr: improving action abstraction for imperfect information extensive-form games with reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 27752–27770, 2024.

Boning Li, Baoxiang Wang, and Longbo Huang. Pokerskill: Llms can play expert-level poker without training or solvers. *arXiv preprint arXiv:2605.30094*, 2026.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

Mingyang Liu, Chengjie Wu, Qihan Liu, Yansen Jing, Jun Yang, Pingzhong Tang, and Chongjie Zhang. Safe opponent-exploitation subgame refinement. *Advances in Neural Information Processing Systems*, 35:27610–27622, 2022.

Charles F Manski. *Partial identification of probability distributions*. Springer, 2003.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Peter McCracken and Michael Bowling. Safe strategies for agent modelling in games. In *AAAI Technical Report (2)*, pp. 103–110, 2004.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Samer Nashed and Shlomo Zilberstein. A survey of opponent modeling in adversarial domains. *Journal of Artificial Intelligence Research*, 73:277–327, 2022.

Arkadi Nemirovski, Shmuel Onn, and Uriel G Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, et al. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *International Conference on Machine Learning*, pp. 8525–8535. PMLR, 2021.

Marc Ponsen, Marc Lanctot, and Steven De Jong. Mcrnr: fast computing of restricted nash responses by means of sampling. In *Proceedings of the 3rd AAAI Conference on Interactive Decision Theory and Game Theory*, pp. 43–49, 2010.

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.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pp. 550–558, 2005.

Finbarr Timbers, Nolan Bard, Edward Lockhart, Marc Lanctot, Martin Schmid, Neil Burch, Julian Schrittwieser, Thomas Hubert, and Michael Bowling. Approximate exploitability: Learning a best response. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3487–3493. ijcai.org, 2022.

Bernhard Von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996.

Han Wang, Philippe Beardsell, Boning Li, Aaron Sasmita, Shuai Li, Hongyuan Zha, and Baoxiang Wang. Solver-guided reasoning for mixed-equilibrium strategies, 2026. URL `https://arxiv.org/abs/2608.06741`.

Shijia Wang, Jiao Wang, and Bangyan Song. Horse-cfr: Hierarchical opponent reasoning for safe exploitation counterfactual regret minimization. *Expert Systems with Applications*, 263:125697, 2025.

Zhikun Wang, Abdeslam Boularias, Katharina Mülling, and Jan Peters. Balancing safety and exploitability in opponent modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pp. 1515–1520, 2011.

Brian Hu Zhang and Tuomas Sandholm. Small nash equilibrium certificates in very large games. *Advances in Neural Information Processing Systems*, 33:7161–7172, 2020.

Brian Hu Zhang and Tuomas Sandholm. Finding and certifying (near-) optimal strategies in black-box extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5779–5788, 2021.

Richard Zhuang, Akshat Gupta, Richard Yang, Aniket Rahane, Zhengyu Li, and Gopala Anumanchipalli. Pokerbench: Training large language models to become professional poker players. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26175–26182, 2025.

Martin Zinkevich, Michael Bowling, Nolan Bard, Morgan Kan, and Darse Billings. Optimal unbiased estimators for evaluating agent performance. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, pp. 573. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20, 2007.

A PROOFS

Proof of Theorem 4.1. Since $0 < P(T) < 1$, both conditional averages $S_\lambda(T) = \mathbb{E}[S_\lambda(x, h_{\text{opp}}) \mid h_{\text{opp}} \in T]$ and $S_\lambda(T^c)$ (defined analogously) exist, and $q_\lambda(T) = \frac{P(T)S_\lambda(T)}{P(T)S_\lambda(T)+P(T^c)S_\lambda(T^c)}$. For $\lambda < \lambda_0$ Assumption 3.1 gives $S_\lambda(T^c) \geq \underline{s} > 0$, so the denominator is at least $P(T^c)\underline{s} > 0$ and q_λ is well defined; the hand set is finite, so the pointwise limits on T give $S_\lambda(T) \to 0$, hence $q_\lambda(T) \to 0$. By the event characterization of total variation,

$$\mathrm{TV}(q_\lambda, P) = \sup_A \big|q_\lambda(A) - P(A)\big| \geq \big|q_\lambda(T) - P(T)\big| \to P(T).$$

The second claim follows by choosing λ small enough. □

Proof of Corollary D.1. Fix a reference hand $\bar{h} \in T^c$ ($T^c \neq \emptyset$ because $P(T) < 1$, and $P(\bar{h}) > 0$ because P has full support). The ratio assumption gives $S_\lambda(h)/S_\lambda(\bar{h}) \to 1$ for every $h \in T^c$, so $q_\lambda(h)/q_\lambda(\bar{h}) \to P(h)/P(\bar{h})$ on T^c, while $q_\lambda(h) \to 0$ on T (proof of Theorem 4.1). Normalizing the distribution by its T^c-mass gives $q_\lambda(h) \to P(h)/P(T^c)$ for $h \in T^c$. Substituting into the finite-space identity $\mathrm{TV}(q, P) = \frac{1}{2}\sum_h |q(h) - P(h)|$, split over T and T^c, gives

$$\mathrm{TV}(q_\lambda, P) \to \tfrac{1}{2}P(T) + \tfrac{1}{2}\sum_{h \notin T} P(h)\Big(\tfrac{1}{P(T^c)} - 1\Big) = \tfrac{1}{2}\big(P(T) + 1 - P(T^c)\big) = P(T).$$

□

Proof of Theorem 5.1. Under fixed μ_λ the opponent's behavior and chance are exogenous, so the reach probability of a terminal t is $p(t)\, x[a_t]\, y^{\mu_\lambda}[b_t]$. Grouping the target-showdown terminals by $a_t = a$ yields $\theta(x) = \sum_a c[a]x[a]$. The objective and the constraints ($Ex = e$, $\mathrm{expl}(x) \leq \varepsilon$; Theorem 3.3) are linear in (x, q), so $C(\varepsilon; T)$ is a linear-programming optimum. □

Proof of Theorem 3.3. By definition $\mathrm{expl}(x) \leq \varepsilon$ iff $\min_{y \in Y} x^T A y \geq v^* - \varepsilon$. The inner program $\min_{y \geq 0,\, Fy = f} (A^T x)^T y$ has a nonempty bounded feasible set and a feasible dual; strong LP duality gives $\min_{y \in Y} x^T A y = \max_{q:\, F^T q \leq A^T x} f^T q$. Substituting proves the equivalence, and LP-C follows by substituting the dual certificate into the capacity program of Theorem 5.1. □

Proof of Theorem 5.2. $C(\varepsilon)$ is the value function of LP-C with right-hand side $[\,\mathbf{0};\, -(v^* - \varepsilon)\,]$ entering the inequality constraints. The parametric-LP sensitivity theorem applies: the constraint matrices are fixed and finite-dimensional, LP-C is feasible for every $\varepsilon \geq 0$ (a Nash hero plan with its equilibrium dual certificate is feasible at $\varepsilon = 0$, hence at every larger budget), and the objective is bounded, $0 \leq c^T x \leq 1$ on the compact X since $\theta(x)$ is a probability. Hence the value function is finite, continuous, nondecreasing under right-hand-side relaxation, and concave. A convex combination of plans feasible at budgets $\varepsilon_1, \varepsilon_2$ is feasible at the mixed budget, by linearity of the certificate constraints. The value function is piecewise affine with finitely many pieces, since there are finitely many optimal bases; in particular it is affine on some initial interval $[0, \varepsilon_0]$ with $\varepsilon_0 > 0$, so $C(\varepsilon) = C(0) + \rho\,\varepsilon$ there, and by concavity $\rho = \sup_{\varepsilon > 0}[C(\varepsilon) - C(0)]/\varepsilon$. The right-derivative $\rho \geq 0$ is the one-sided sensitivity of the optimum to the right-hand side of the certificate constraint. Precisely, let $\pi \geq 0$ be the multiplier of $f^T q \geq v^* - \varepsilon$ in the dual of LP-C. The perturbation moves only a right-hand side, so every dual solution that is optimal at $\varepsilon = 0$ stays dual feasible for $\varepsilon > 0$, and weak duality gives $C(\varepsilon) \leq C(0) + \pi\varepsilon$ for its multiplier π; conversely, a single dual solution is optimal for all $\varepsilon \in (0, \varepsilon_0)$ (finitely many bases), it is optimal at $\varepsilon = 0$ by continuity of the dual

objective in ε, and its multiplier attains ρ. Hence ρ is the smallest certificate multiplier over the optimal dual solutions of LP-C at $\varepsilon = 0$: the shadow price of the certificate when the dual optimum is unique, and the minimum over the optimal dual face otherwise. □

Proof of Corollary 5.3. By Theorem 3.3, x is feasible in LP-C at $\varepsilon = 0$ iff there is q with $F^T q \le A^T x$ and $f^T q \ge v^*$, i.e. iff $\mathrm{expl}(x) \le 0$; since $\mathrm{expl}(x) \ge 0$ always (v^* is the max–min value), feasibility is equivalent to $\mathrm{expl}(x) = 0$, that is, to $x \in \mathcal{E}$. The projection of the LP-C(0) feasible set onto x is therefore exactly $\mathcal{E}$, and $\mathcal{E}$ is nonempty (a finite zero-sum game has a Nash equilibrium, whose hero plan lies in X). The values coincide, so $C(0;T) = \max_{x\in\mathcal{E}} \theta(x)$. The coefficient c is nonnegative, so $\theta \equiv 0$ on $\mathcal{E}$ iff every $x \in \mathcal{E}$ is supported on sequences with $c[a] = 0$; by construction $c[a] > 0$ precisely when the nominal opponent $\bar{\mu}$ reaches a target-hand showdown terminal on hero sequence a with positive probability, giving the zero-criterion. □

Proof of Lemma 5.4. C is concave and right-differentiable with $\rho = C'_+(0)$ (Theorem 5.2), so the difference quotients $[C(\varepsilon)-C(0)]/\varepsilon$ are nonincreasing and converge to ρ: $C(\varepsilon) \le C(0)+\rho\varepsilon$ for every $\varepsilon \ge 0$. Every x with $\mathrm{expl}(x) \le \varepsilon$ satisfies $\theta(x) \le C(\mathrm{expl}(x)) \le C(0) + \rho\,\mathrm{expl}(x) \le C(0) + \rho\varepsilon$. The shadow-price claim restates the proof of Theorem 5.2: ρ is the smallest multiplier of the certificate constraint $f^T q \ge v^*$ over the optimal dual solutions of LP-C at $\varepsilon = 0$, hence the shadow price of that constraint whenever the dual optimum is unique. □

Proof of Proposition 6.5. (i) By Definition 6.1 a showdown cell records the opponent's private hand, so membership $h_{\mathrm{opp}} \in T$ is constant on each showdown cell and χ_T is well defined. Summing $(\Phi_x \bar{y})_\kappa = \sum_{t\in\kappa} p(t)\, x[a_t]\, \bar{y}[b_t]$ over the cells with $(\chi_T)_\kappa = 1$ gives $\chi_T^T \Phi_x \bar{y} = \sum_{t \text{ showdown: } h_{\mathrm{opp}}(t)\in T} p(t)\, x[a_t]\, \bar{y}[b_t] = \sum_a c[a]x[a] = \theta(x)$, with c the coefficient of Theorem 5.1. The identified set is $U(x,m) = \{y \ge 0 : Fy = f,\ \Phi_x y = m\}$ with $m = \Phi_x y^*$ (Theorem 6.4), and part (iv) of that theorem identifies every $\ell \in R(x)$. Conversely, if $U(x,m)$ contains $y^\circ > 0$ coordinatewise and $\ell \notin R(x)$, the argument of part (ii) below with $x' = x$ produces two points of $U(x,m)$ on which ℓ differs, so ℓ is not point-identified. (ii) Take $\ell \in R(x) \setminus R(x')$, nonempty by hypothesis. By Theorem 6.4(iv), $\ell^T y = a^T f + b^T m$ is constant on $U(x,m)$. Since $R(x')^\perp = N(x') = \{d : Fd = 0,\ \Phi_{x'} d = 0\}$ and $\ell \notin R(x')$, the projection of ℓ onto $N(x')$ is nonzero, so there is $d \in N(x')$ with $\ell^T d \ne 0$. Let $y^\circ \in U(x',m')$ with $y^\circ > 0$ coordinatewise; for $t > 0$ small enough that $y^\circ \pm td \ge 0$, both $y^\circ \pm td$ satisfy $F(y^\circ \pm td) = f$ and $\Phi_{x'}(y^\circ \pm td) = m'$, hence lie in $U(x',m')$, and $\ell^T(y^\circ + td) - \ell^T(y^\circ - td) = 2t\,\ell^T d \ne 0$: the LP-measured interval of ℓ over $U(x',m')$ has positive width. Without a strictly positive point no converse is claimed. □

Proof of Theorem 4.2. Conditioning on h_{opp} cancels $p(h_{\mathrm{opp}})$, so each summand has expectation $P(h_{\mathrm{opp}} \in T) = \tau$. Each summand lies in $[0, p_{\min}^{-1}]$; Hoeffding's inequality gives the bound. Only target hands contribute a nonzero summand, so for $T \ne \emptyset$ the range is in fact $[0, 1/\min_{h\in T} p(h)]$ and the bound holds with $p_{\min}$ replaced by $\min_{h\in T} p(h)$. The same conditioning gives the exact variance $\mathrm{Var}(\hat{\tau}_{\mathrm{IPW}}) = n^{-1}\big(\sum_{h\in T} P(h)/p(h) - \tau^2\big)$. □

Proof of Theorem D.2. Condition first on $\mathcal{F}_{t-1}$ and $h_{\mathrm{opp}}^{(t)}$. The conditional expectation of $S_t/p_t(h_{\mathrm{opp}}^{(t)})$ equals one. Averaging over the fresh private hand, whose marginal is P by the independent-dealing assumption of the theorem, gives τ; the tower property proves the result. For the tail bound, the centered terms $Z_t - \tau$ are martingale differences with respect to the hand filtration, and given $\mathcal{F}_{t-1}$ each lies in an interval of length $p_{\min}^{-1}$, since $Z_t \in [0, p_{\min}^{-1}]$; the Azuma–Hoeffding inequality gives $P(|\sum_{t\le n}(Z_t - \tau)| \ge n\delta) \le 2\exp\big(-2n^2\delta^2/(n\,p_{\min}^{-2})\big) = 2\exp(-2n\delta^2 p_{\min}^2)$. For stale weights $\tilde{Z}_t = \mathbf{1}\{h_{\mathrm{opp}}^{(t)} \in T, S_t = 1\}/\tilde{p}_t(h_{\mathrm{opp}}^{(t)})$ with $\tilde{p}_t(h)$ $\mathcal{F}_{t-1}$-measurable for every h, the factor $1/\tilde{p}_t(h)$ is fixed given $(\mathcal{F}_{t-1}, h_{\mathrm{opp}}^{(t)} = h)$, so the same conditioning gives $\mathbb{E}[\tilde{Z}_t \mid \mathcal{F}_{t-1}] = \sum_{h\in T} P(h)\, p_t(h)/\tilde{p}_t(h)$, so

$$\big|\mathbb{E}[\tilde{Z}_t \mid \mathcal{F}_{t-1}] - \tau\big| = \Big|\sum_{h\in T} P(h)\Big(\frac{p_t(h)}{\tilde{p}_t(h)} - 1\Big)\Big| \le \tau \sup_{h\in T}\Big|\frac{p_t(h)}{\tilde{p}_t(h)} - 1\Big|,$$

finite because $\tilde{p}_t \ge p_{\min} > 0$. □

Proof of Theorem 6.4. (i) Y is a compact convex polytope (sequence-form realization constraints), and $U(x, m)$ is its intersection with the affine subspace $\{\Phi_x y = m\}$; it contains y^* by construction of m. (ii) The observable record of a hand is exactly its cell, and by Definition 6.1 the cell distribution under (x, y) is $\Phi_x y$; hence identical observable distributions iff $\Phi_x y = \Phi_x y^* = m$. (iii) Any two elements of $U(x, m)$ differ by d with $Fd = 0$ and $\Phi_x d = 0$, so $U(x, m) - y^* \subseteq \mathcal{N}$; a polytope contained in an affine translate of a linear space has dimension at most that of the space. (iv) On $U(x, m)$, $\ell^T y = a^T(Fy) + b^T(\Phi_x y) = a^T f + b^T m$, a constant. □

Exact identification criterion. The converse of Theorem 6.4(iv) holds once the implicit equalities of $U(x, m)$ are added. Let $J = \{j : \ y_j > 0 \text{ for some } y \in U(x, m)\}$ and let H stack the coordinate rows e_j^T, $j \notin J$. Then $\ell^T y$ is constant on $U(x, m)$ iff $\ell \in \mathrm{rowspace}([F; \Phi_x; H])$, and $\dim U(x, m) = \dim \ker[F; \Phi_x; H]$. Indeed, a positive convex combination of points of $U(x, m)$, one positive at each $j \in J$, is a point $y^\circ \in U(x, m)$ with $y^\circ_j > 0$ for all $j \in J$; for $d \in \ker[F; \Phi_x; H]$ and small $t > 0$ both $y^\circ \pm td$ lie in $U(x, m)$, while every difference of two points of $U(x, m)$ lies in that kernel. The affine hull of $U(x, m)$ is therefore $y^\circ + \ker[F; \Phi_x; H]$, and the functionals constant on it are exactly the orthogonal complement of the kernel, the row space. Proposition 6.5(i) is the case $J = S_2$; the per-coordinate LP maxima of Appendix E compute J, since $j \in J$ iff $\max_{y \in U(x,m)} y_j > 0$ (a zero width alone only says that the coordinate is pinned, not that it is pinned at zero).

Proof of Lemma 6.2. Every terminal t in a cell κ shares the hero sequence $a_t = a_\kappa$, because the cell records the full betting history and the hero's private hand. Hence $\Phi_x[\kappa, j] = \sum_{t \in \kappa : b_t = j} p(t) x[a_t] = x[a_\kappa] \sum_{t \in \kappa : b_t = j} p(t) = x[a_\kappa] B[\kappa, j]$, which is the first claim. For the second, expand the payoff matrix by cell,

$$(A^T z)_j = \sum_a A[a, j] z[a] = \sum_\kappa B[\kappa, j]\, u_\kappa\, z[a_\kappa] = (B^T D_z u)_j,$$

using $A[a, j] = \sum_{t : a_t = a,\, b_t = j} p(t) u_\kappa$ and the cell-measurability of u. The identity is linear in z, so it holds for every $z \in \mathbb{R}^{|S_1|}$ and in particular for every deployment. The vector $B^T D_z u = \sum_\kappa z[a_\kappa] u_\kappa\, B[\kappa, \cdot]^T$ is a linear combination of the rows of B, so $A^T z \in \mathrm{rowspace}(B)$. Finally, $\Phi_x = D_x B$ with D_x diagonal nonnegative; dropping the rows with $x[a_\kappa] = 0$ and rescaling the rest by the positive weights $x[a_\kappa]$ leaves the row space unchanged, so it depends on x only through the support. □

Proof of Corollary 6.3. If $x[a_\kappa] > 0$ on every cell with $B[\kappa, \cdot] \neq 0$, the rows of $\Phi_x = D_x B$ that are dropped by a zero weight are zero rows of B, and the remaining rows are positive multiples of the corresponding rows of B; hence $\mathrm{rowspace}(\Phi_x) = \mathrm{rowspace}(B)$ and $\mathrm{rowspace}([F; \Phi_x]) \supseteq \mathrm{rowspace}(B)$. By Lemma 6.2, $A^T z = B^T D_z u \in \mathrm{rowspace}(B)$ for every deployment z; explicitly $(A^T z)^T = b^T \Phi_x$ with $b_\kappa = z[a_\kappa] u_\kappa / x[a_\kappa]$ on the cells with $x[a_\kappa] > 0$ and $b_\kappa = 0$ otherwise, the dropped cells having $B[\kappa, \cdot] = 0$. Theorem 6.4(iv) with $(a, b) = (0, b)$ makes $z^T A y$ constant on $U(x, m)$. The same membership is the hypothesis $A^T z \in \mathrm{rowspace}([F; \Phi^{(k)}])$ of Theorem 7.2 for any stack containing x, so the certificate there is finite for every z. □

Proof of Lemma 6.6. Group the sequence-form payoff decomposition by cell and use cell-measurability:

$$x^T A y = \sum_\kappa \sum_{t \in \kappa} p(t)\, u(t)\, x[a_t]\, y[b_t] = \sum_\kappa u_\kappa \sum_{t \in \kappa} p(t)\, x[a_t]\, y[b_t] = \sum_\kappa u_\kappa (\Phi_x y)_\kappa = u^T \Phi_x y.$$

Since this holds for all $y \in \mathbb{R}^{|S_2|}$ (the identity is linear in y), $A^T x = \Phi_x^T u$. □

Proof of Theorem 6.7. Lemma 6.6 gives $A^T x = \Phi_x^T u \in \mathrm{rowspace}(\Phi_x)$; apply Theorem 6.4(iv) with $a = 0$, $b = u$. The general statement is Theorem 6.4(iv) verbatim. □

Proof of Theorem 6.8. (a) The inner problem $\min\{c^T y : \ Fy = f,\ Gy \leq g,\ y \geq 0\}$ with $c = A^T x$ is feasible ($U \neq \emptyset$) and bounded (U compact), so strong duality holds. Rewriting $Gy \leq g$ as $-Gy \geq -g$ with multiplier $t \geq 0$ and substituting $s = -t \leq 0$ yields the stated dual. (b) By

Theorem 3.3, $\mathrm{expl}(x) \le \varepsilon$ iff $\exists q:\ F^T q \le A^T x,\ f^T q \ge v^* - \varepsilon$. By (a), for each feasible x the inner value is the maximum of $f^T w + g^T s$ over dual-feasible (w, s). Since the outer problem also maximizes, $R(\varepsilon)$ equals the joint supremum of $f^T w + g^T s$ over the LP-R feasible set. For feasibility take $x = x_{\mathrm{Nash}}$, q its equilibrium dual, and (w, s) the inner-dual optimizer for x_{Nash}; boundedness follows from weak duality and the finiteness of the game. (c) With $x = x_{\mathrm{Nash}}$: $\mathrm{expl}(x_{\mathrm{Nash}}) = 0 \le \varepsilon$ and $\min_{y \in U} x_{\mathrm{Nash}}^T A y \ge \min_{y \in Y} x_{\mathrm{Nash}}^T A y = v^*$ since $U \subseteq Y$. (d) The argument is pathwise: fix any realized path on which $y^* \in U_n$. For *that* path, and for *any* hero plan x, in particular the data-dependent $\hat{x}$, the chain

$$x^T A y^* \ \ge\ \min_{y \in U_n} x^T A y \ \ge\ f^T w + g^T s$$

holds for any dual-feasible (w, s) at x: the first inequality because $y^* \in U_n$ is feasible for the inner program, the second by weak duality. No uniformity over x, no sample splitting, and no independence between $\hat{x}$ and U_n is needed. On paths where LP-R has no finite optimum, the fallback is a Nash plan x^* with its equilibrium dual certificate $(w^*, 0)$, satisfying $F^T w^* \le A^T x^*$ and $f^T w^* = v^*$. Then $x^{*T} A y^* \ge v^* = f^T w^*$ for every opponent y^*, so the certificate holds with probability one on the fallback branch as well. LP-R is feasible on every path: $(x^*, w^*, 0, w^*)$ satisfies all its constraints whatever (G, g) are. So it has no finite optimum exactly when it is unbounded, and that happens exactly when $U_n = \emptyset$: if $U_n \neq \emptyset$, weak duality bounds the objective by $\min_{y \in U_n} x^T A y \le \max_{x,y} x^T A y$; if $U_n = \emptyset$, the inner program at x^* is infeasible while its dual is feasible, so that dual, and with it LP-R, is unbounded. Since $y^* \in U_n$ forces $U_n \neq \emptyset$, the fallback paths form a subset of $\{y^* \notin U_n\}$. The safety claim is a hard constraint of LP-R and involves no randomness. □

Lemma A.1 (Maurer–Pontil empirical Bernstein)**.** *Let $X_1, \dots, X_n$, $n \ge 2$, be i.i.d.* Bernoulli(μ) *with empirical mean $\hat{m} = n^{-1} \sum_i X_i$. For every $\delta \in (0, 1)$,*

$$P\Big(|\hat{m} - \mu| \ \ge\ \sqrt{\tfrac{2\hat{m}(1-\hat{m})\ln(2/\delta)}{n-1}} + \tfrac{7}{3} \cdot \tfrac{\ln(2/\delta)}{n-1}\Big) \ \le\ 2\delta.$$

Proof of Lemma A.1. Maurer & Pontil (2009) state the one-sided empirical Bernstein bound $\mu - \hat{m} \le \sqrt{2V_n \ln(2/\delta)/n} + \frac{7}{3}\ln(2/\delta)/(n-1)$ with probability at least $1 - \delta$; here $V_n = (n-1)^{-1} \sum_i (X_i - \hat{m})^2$ denotes the unbiased sample variance. Applying it to X_i and to $1 - X_i$, which have the same V_n, and taking a union bound gives the two-sided event with probability at most 2δ. For Bernoulli observations $V_n = \frac{n}{n-1}\hat{m}(1-\hat{m})$, and substituting into $\sqrt{2V_n \ln(2/\delta)/n}$ gives $\sqrt{2\hat{m}(1-\hat{m})\ln(2/\delta)/(n-1)}$. □

Proof of Theorem 6.9. (a) Each $U_i \ni y^*$ by Theorem 6.4(i); intersecting preserves membership, and stacking the equality (or band) constraints is set intersection. (b) Immediate. (c) On $U^{(k+1)} \neq \emptyset$ a minimum over a smaller nonempty set is no smaller and a maximum no larger, giving the interval nesting. In the population case $y^* \in U^{(k+1)}$ by (a); for confidence sets this is the coverage event. For R_k, each fixed x's inner value $\min_{y \in U^{(k)}} x^T A y$ is nondecreasing in k, hence so is the maximum over the unchanged outer feasible set. (d) Bonferroni: $P(y^* \notin \bigcap_i U_i) \le \sum_i P(y^* \notin U_i) \le \sum_i \alpha_i$. For one probe with $n_i \ge 2$ hands, each cell frequency $\hat{m}_{i,\kappa}$ averages i.i.d. Bernoulli indicators with expectation $(\Phi_{x_i} y^*)_\kappa$. With $L_i = \ln(8K_i/\alpha_i)$ and radius

$$r_{i,\kappa} = \sqrt{\frac{2\,\hat{m}_{i,\kappa}(1 - \hat{m}_{i,\kappa})\,L_i}{n_i - 1}} + \frac{7}{3} \cdot \frac{L_i}{n_i - 1},$$

Lemma A.1 with $\delta = \alpha_i/(4K_i)$ (so $\ln(2/\delta) = L_i$) gives, per cell, $P(|\hat{m}_{i,\kappa} - (\Phi_{x_i} y^*)_\kappa| \ge r_{i,\kappa}) \le 2\delta = \alpha_i/(2K_i)$. The blend also keeps a Hoeffding radius $r^H_{i,\kappa} = \sqrt{\ln(4K_i/\alpha_i)/(2n_i)}$, which has per-cell failure $\le \alpha_i/(2K_i)$ by Hoeffding's inequality, and clips to $\min(r_{i,\kappa}, r^H_{i,\kappa})$. A cell fails the blended box only if it fails at least one candidate radius, so its realized failure probability is at most $\alpha_i/(2K_i) + \alpha_i/(2K_i) = \alpha_i/K_i$; a union bound over the K_i cells yields the claim. □

Proof of Lemma 7.1. $Fy = Fy' = f$ gives $Fd = 0$; each realized band $\ell_{i,\kappa} \le (\Phi_{x_i} y)_\kappa \le u_{i,\kappa}$ pins $(\Phi_{x_i} d)_\kappa$ to $[-(u_{i,\kappa} - \ell_{i,\kappa}),\ u_{i,\kappa} - \ell_{i,\kappa}]$, and stacking gives $|\Phi^{(k)} d| \le 2r$. □

Proof of Theorem 7.2. By Lemma 7.1, $W_k(z) \le \max\{z^T A d : \ Fd = 0,\ |\Phi^{(k)} d| \le 2r\}$. Write the right-hand LP with variables d (free) and constraints $Fd = 0$, $\Phi^{(k)} d \le 2r$, $-\Phi^{(k)} d \le 2r$. Its dual has variables λ (free) and $u^+, u^- \ge 0$, with value $\min 2\sum_{i,\kappa} r_{i,\kappa}(u^+_{i,\kappa} + u^-_{i,\kappa})$ subject to $F^T\lambda + \Phi^{(k)T}(u^+ - u^-) = A^T z$; minimizing the split $w = u^+ - u^-$ (each component written $|w| = u^+ + u^-$ with one of $u^\pm$ zero at the optimum) gives $2 \cdot \min\{\sum r|w| : \ \exists\lambda,\ F^T\lambda + \Phi^{(k)T} w = A^T z\}$. Weak duality yields the claim. If $A^T z \notin \mathrm{rowspace}([F; \Phi^{(k)}])$ no such (λ, w) exists, the dual is infeasible, and the primal relaxation is unbounded. □

Proof of Corollary 7.3. Each row of $\Phi^{(k+1)}$ lies in $\mathrm{rowspace}([F; \Phi^{(k)}])$, so on the exact affine set $Fy = f$, $\Phi^{(k)} y = m^{(k)}$ the additional moments are pinned by the same linear combination and equal $m^{(k+1)}$. □

Proof of Corollary 7.4. The identity is a telescoping about the common middle term $\min_{S^{(k+1)}} \|r^{(k)} \cdot w\|_1$. For Δ_{w}: zero-padding maps any feasible (λ, w) of the k-problem to a feasible point of $S^{(k+1)}$, and with the new-cell coefficients 0 the embedded objective equals the k-problem objective, so the minimum over the superset is no larger, $\Delta_{\mathrm{w}} \le 0$. For Δ_{r}: under coordinatewise radius growth the weighted norm $\|r^{(k+1)} \cdot w\|_1 \ge \|r^{(k)} \cdot w\|_1$ holds pointwise on $S^{(k+1)}$ (weights are nonnegative, old weights are no smaller, new weights are nonnegative), so its minimum over the same feasible set is no smaller, $\Delta_{\mathrm{r}} \ge 0$. Monotonicity is therefore asserted only conditionally and is audited in Appendix E. □

Pathwise reading of the finite-sample certificate. On the coverage event the bound of Theorem 6.8(d) holds for the realized $\hat{x}$ and any dual-feasible (w, s), with no sample splitting. Empty confidence sets make the inner program infeasible, and there the Nash fallback certifies v^*.

Remarks on the certificate and the capacity coefficient. The certificate encodes "within ε of game value" as a linear constraint on (x, q).

The capacity coefficient is built from a *nominal* censoring model $\bar{\mu} = (\sigma_{\mathrm{base}}, \lambda)$ fixed by the analyst; only the probe plan x is controlled. The model makes c and the propensities of Theorem 4.2 computable, so LP-C counts the target-showdown labels a safe probe collects under that model. The true opponent y^* remains arbitrary, and a misspecified nominal model biases the propensities.

Remarks on the fixed-budget criterion. A probe that adds no new rowspace pins no new population moment; a finite-sample certificate can still change, because a redundant map supplies another representation of the same rowspace at fixed radii. The intermediate objective measures a change in certificate representation: even a duplicate probe can make it zero for its own deployment by placing all payoff weights on the new, zero-cost cells. Thus a negative Δ_{w} need not add a population-identifying direction. The full width-bound change is $2(\Delta_{\mathrm{w}} + \Delta_{\mathrm{r}})$.

B Detailed Experimental Tables

This appendix holds the block overview (Table 2) and the per-block tables of Section 8: the fixed-total-budget study (Table 3), the full-support mixtures of E3d (Table 4), the payoff-only ablation of E3e (Table 5), the stacking contrast behind the fixed-budget criterion (Table 6), its extension to all four budgets (Table 7), and the equal-budget probe-design study (Table 9).

In Table 9, θ_e is the population target-showdown rate under the evaluation opponent ($\lambda = 0.2$). The designs use the nominal opponent ($\lambda = 0$). LP-R margin is the certified lower bound minus v^*; true value is the learned deployment's $\hat{x}^T A y^*$, without subtracting v^*. Payoffs are in units of the ante, the forced initial bet. Entries with $\pm$ give the sample standard deviation across seeds; true values and widths are seed means. The mixture is the equal-weight mixture of the Nash and capacity-0.1 implementation plans. The MI-selected design maximizes nominal MI within the 22-candidate pool specified in Appendix E. E6 optimizes capacity separately at each evaluation λ and E7 fixes the design at $\lambda = 0$, which gives the two capacity observation rates 0.388 and 0.383 at evaluation $\lambda = 0.2$.

Three tables report a certificate margin for the capacity-ε=0.1 probe under different designs: 0.101 (Table 4, one probe, 5,000 hands, unsplit α), 0.0895 (Table 5, two probes with Nash, 10,000 hands in

Table 2: Numerical verification on Leduc Hold'em; $\lambda = 0.2$ for the identification, baseline, and fixed-budget blocks.

Block	Theorems	Key results
E1 MNAR bias	event bound, §4	naive TV 0.126 (λ=1), 0.184 (λ=0); bound attained at $\lambda \leq 0.2$
E2 IPW	4.2, D.2	IPW TV 0.011–0.015 vs. naive 0.078–0.176, p=8.9×10^{-16}
E3a Sharp ID	6.4, 6.7	rank 302/308; own value $\leq 10^{-15}$; width 1.616 at λ=0.2
E3b Robust-safe	6.8, 6.9	cover $50/50$; EB margin 0.0822 ± 0.0173 vs. Hoeffding 0.0062 ± 0.0015
E4a Adaptive probes	D.2	HT–IPW bias -0.0014; naive -0.074; stale $+0.263$ at n=5,000
E4b Capacity	5.1, 5.2	$C(0)$=0.145, $C(0.1)$=0.377 at λ=0; concave in ε
E3d Full support	6.2, 6.3	rank 202→337; population width → 0; margin 0.101→0.079
E3e Payoff-only LP-R	6.8	cell polytope 0.0895 vs. payoff-only 0.0000; zero radius 0.0396
E5 Stacking	7.2, 7.4	F_A tightens $40/40$, 10.13→2.74; F_B loosens to 19.02
E6 Probe baselines	5.4	rate/expl.: cap. $0.388/0.1$; Nash $0.162/0$; uniform $0.155/2.574$; call-station $0.667/1.381$
E7 Probe design	6.5	LP-R margin 0.093 vs. Nash 0.030; total-showdown (MI-sel.) 0.142
E8 Direct payoff	6.7, 6.8	direct selects Nash; LP-R margin 0.0806 with three probes

Table 3: Fixed-total-budget study: splitting a small budget hurts, and the k-gap narrows as data dominate. Margin is the certificate margin $R - v^*$; width is the maximal candidate payoff width. All twelve cells covered the true plan $10/10$.

	k=1	k=2	k=4
margin at N=6,000	0.0560	0.0310	0.0035
width at N=6,000	1.19	1.44	1.74
margin at N=150,000	0.2548	0.2536	0.2401
width at N=150,000	0.233	0.307	0.439

total, $\alpha_i = \alpha/2$, a different uncertainty set), and 0.093 ± 0.026 (Table 9, one probe, a separate draw). Probe count, total sample, confidence split, and uncertainty set all change the optimized certificate, so compare within a table.

Equal-budget probe contrast on Leduc. The equal-budget probes (E7) realize the contrast of Proposition 6.5. The capacity probe at $\varepsilon = 0.1$ and the total-showdown-rate probe, the *MI-selected* candidate, reach nearly equal target-showdown rates ($\theta_e = 0.383$), both inside P_ε at $\varepsilon = 0.1$. The MI-selected probe is the candidate of highest nominal-model mutual information (MI) among a finite pool of 22 safe probes (Appendix E). Their moment maps differ in rank: the nullity (the dimension of the kernel) is $\dim \ker[F; \Phi_x] = 135$ for the capacity probe and 71 for the MI-selected probe. The two also differ in mean certificate margin, 0.093 for the capacity probe and 0.142 for the MI-selected probe (Table 9). Identification cost depends on the deployment. Capacity identifies its own population value but widens the Nash deployment's interval. The MI-selected probe narrows both finite-sample widths relative to capacity, and relative to Nash it narrows only its own-deployment width.

C Additional Related Work

Showdown records are missing data. The filter is ignorable for the known game-rule prior but not for the opponent behavior generating the selection; the distinction is standard in missing-data theory (Rubin, 1976; Little & Rubin, 2019; Robins et al., 1994). The sharp set of Section 6 follows partial identification (Manski, 2003; Beresteanu et al., 2009), and inverse-propensity weighting (IPW) descends from Horvitz–Thompson (HT) (Horvitz & Thompson, 1952). Censored-feedback bandits and partial-monitoring games (Abernethy et al., 2016; Antos et al., 2013) describe feedback through action-dependent signals.

Opponent modeling in poker began with hand-frequency statistics and Bayesian hand estimation (Billings et al., 1998; Southey et al., 2005); Bayesian exploitation of hand estimates followed

Table 4: E3d: capacity-optimal probe mixed with a safe full-support probe at weight s, 50 seeds, $n = 5{,}000$ hands, $\lambda = 0.2$. Population width is the width of the Nash deployment over the exact-moment set; the last two columns are finite-sample widths under the empirical–Bernstein box. Each margin carries its seed standard deviation, and the seed standard deviation of every width is at most 0.024 (Appendix E). Every mixture satisfies $\mathrm{expl}(x) \le 0.1$.

s	capacity	rank	pop. width (Nash)	margin	width (Nash)	width (cap-0.1)
0	0.377	202	1.616	0.101±0.021	1.884	0.632
0.01	0.375	337	$< 10^{-15}$	0.079±0.020	1.904	0.678
0.05	0.365	337	$< 10^{-15}$	0.070±0.019	1.904	0.688
0.1	0.353	337	$< 10^{-15}$	0.072±0.024	1.908	0.699
0.5	0.259	337	$< 10^{-15}$	0.049±0.020	1.502	0.835

Table 5: E3e: same two probes (Nash and capacity-ε=0.1), same 50 seeds and hands, $v^* = -0.0856$. The payoff-only sets constrain only the two mean payoffs, one interval per probe; the exact-means arm uses zero radius and isolates representational error from sampling error. Margin is the minimum over the set minus v^*; true value is $\hat{x}^T A y^*$.

Uncertainty set	mean margin	worst seed	true value	exploitability
cell polytope, EB box	0.0895	0.0499	0.2043	0.100
payoff-only, EB intervals	0.0000	0.0000	−0.0460	0.000
payoff-only, exact means	0.0396	0.0396	−0.0460	0.000
Nash	0.0000	0.0000	−0.0460	0.000

(Ganzfried & Sun, 2018). The deployment side spans safe opponent exploitation (Ganzfried & Sandholm, 2015), restricted Nash and data-biased responses (Johanson et al., 2007; Ponsen et al., 2010; Johanson & Bowling, 2009), safe agent modelling (McCracken & Bowling, 2004), safety–exploitability balancing (Wang et al., 2011), safe subgame solving and its exploitative refinements (Brown & Sandholm, 2017; Liu et al., 2022; Ge et al., 2024), greedy-when-sure exploitation (Fu et al., 2022), hierarchical opponent reasoning (Wang et al., 2025), utility-constrained exploitation (Bernasconi-de Luca et al., 2021), and regularized equilibrium learning (Perolat et al., 2021); Nashed & Zilberstein (2022) survey the area.

Safe-play certificates use sequence-form duality (Zhang & Sandholm, 2020; 2021; Nemirovski et al., 2010). Counterfactual regret minimization (CFR) and the sampled variants built on it (Zinkevich et al., 2007; Lanctot et al., 2009) are used by the superhuman poker agents (Moravčík et al., 2017; Brown & Sandholm, 2018; 2019), which reach full-scale games through action and information abstraction (Li et al., 2024; Li & Huang, 2025) and real-time search. The LPs here reuse that duality for the observation and identification objectives alike. Their dimensions follow the sequence form: LP-C has $|S_1| + |I_2|$ variables, one flow equality per hero information set, and $|S_2| + 1$ inequality rows. Weighted estimators have a long line in game evaluation (Zinkevich et al., 2006; Bowling et al., 2008; Burch et al., 2018; Abe & Kaneko, 2021); approximate exploitability (Timbers et al., 2022) estimates the certificate computed here by linear programming on the enumerated game. The tails are empirical–Bernstein (Maurer & Pontil, 2009), martingale (Freedman, 1975), and Hoeffding (Hoeffding, 1963) bounds, the last anchoring Theorem 4.2.

D Balanced Selection and Predictable Adaptive Probes

Corollary D.1 (Equality under balanced selection within T^c)**.** *Under the hypotheses of Theorem 4.1, if $S_\lambda(x, h_{\mathrm{opp}})/S_\lambda(x, h'_{\mathrm{opp}}) \to 1$ for all $h_{\mathrm{opp}}, h'_{\mathrm{opp}} \notin T$, then $\mathrm{TV}(q_\lambda, P) \to P(T)$ as $\lambda \to 0$. In particular the ratio condition is vacuous whenever T^c is a singleton.*

When the uncensored hands are observed with equal relative frequency the bound is tight. On Leduc the audited censoring suppresses only the calls that weak hands make when facing a bet in round 1, so Assumption 3.1 fails and the limit $P(T) = 2/3$ is not reached. The event bound $\mathrm{TV}(q_\lambda, P) \ge |q_\lambda(T) - P(T)|$ is attained at the strong-censoring levels, where $q_0(T) = 0.483$ and $q_{0.2}(T) = 0.517$ against $P(T) = 2/3$.

Table 6: E5: effect of adding one probe at N=6,000 on the bound and width of the $\varepsilon = 0.1$ deployment, mean over 10 seeds. The intermediate term *mid* is the minimum over the refined feasible set with new-cell radii set to zero. It vanishes whenever the deployment's payoff vector lies in the rowspace of the added probe's moment map together with F: that holds for the added probes of F_A and F_C and for the capacity-ε=0 step of F_B, and it fails for its ε=0.02 step. The bound overestimates the width, by a factor between 3 and 20 in these rows, and the two can move in opposite directions (the second F_B step and the F_C step). Both reported deployments satisfy the rowspace condition of Theorem 7.2 in every row. The deployment that LP-R learns satisfies the condition in 2 of 160 stacking steps and is certified by its own one-sided value bound (Table 7).

Family	rank	Δ_{w}	Δ_{r}	mid	bound	width
F_A, k=1	$302 \to 302$	-5.06	1.37	0.00	$10.13 \to 2.74$	$0.967 \to 0.826$
F_B, k=1	$302 \to 304$	-5.06	9.51	0.00	$10.13 \to 19.02$	$0.967 \to 1.212$
F_B, k=2	$304 \to 308$	-3.66	5.46	5.85	$19.02 \to 22.62$	$1.212 \to 1.142$
F_C, k=1	$308 \to 308$	-11.31	2.28	0.00	$22.62 \to 4.56$	$1.142 \to 1.145$

Table 7: E5 across all four total budgets, fresh run, 10 seeds per row. Bound and width are for the capacity-ε=0.1 deployment before and after the step; bound/width is the ratio after the step; tightens is the fraction of seeds in which the bound falls. The last column is the one-sided value certificate of Theorem 6.8 for the deployment that LP-R learns at level k+1, minus v^*, mean and seed standard deviation; it is finite in all 160 steps. The $126/160$ radius-growth count and the $2/160$ rowspace count in Appendix E are computed on these same 160 steps.

N	step	bound	width	bound/width	tightens	LP-R margin
6,000	F_A, k=1	$10.13 \to 2.74$	$0.967 \to 0.826$	3.3	1.0	0.023±0.010
	F_B, k=1	$10.13 \to 19.02$	$0.967 \to 1.212$	15.7	0.0	0.005±0.001
	F_B, k=2	$19.02 \to 22.62$	$1.212 \to 1.142$	19.8	0.0	0.020±0.009
	F_C, k=1	$22.62 \to 4.56$	$1.142 \to 1.145$	4.0	1.0	0.003±0.002
15,000	F_A, k=1	$4.52 \to 1.37$	$0.679 \to 0.538$	2.5	1.0	0.136±0.020
	F_B, k=1	$4.52 \to 8.08$	$0.679 \to 0.898$	9.0	0.0	0.085±0.024
	F_B, k=2	$8.08 \to 9.36$	$0.898 \to 0.787$	11.9	0.0	0.132±0.010
	F_C, k=1	$9.36 \to 2.19$	$0.787 \to 0.747$	2.9	1.0	0.103±0.014
60,000	F_A, k=1	$1.39 \to 0.49$	$0.367 \to 0.250$	2.0	1.0	0.224±0.007
	F_B, k=1	$1.39 \to 2.35$	$0.367 \to 0.509$	4.6	0.0	0.204±0.015
	F_B, k=2	$2.35 \to 2.55$	$0.509 \to 0.421$	6.0	0.0	0.220±0.007
	F_C, k=1	$2.55 \to 0.77$	$0.421 \to 0.371$	2.1	1.0	0.205±0.008
150,000	F_A, k=1	$0.64 \to 0.26$	$0.206 \to 0.144$	1.8	1.0	0.253±0.007
	F_B, k=1	$0.64 \to 1.05$	$0.206 \to 0.316$	3.3	0.0	0.238±0.010
	F_B, k=2	$1.05 \to 1.10$	$0.316 \to 0.238$	4.6	0.1	0.245±0.007
	F_C, k=1	$1.10 \to 0.39$	$0.238 \to 0.210$	1.9	1.0	0.238±0.006

The propensity is a function of the hero plan, so an adaptive probe changes the propensities over time. As long as it is *predictable* (its hand-t behavior is fixed before hand t and the realized propensities are recorded), reweighting by the realized propensities remains unbiased, with variance still depending on them.

Theorem D.2 (Predictable adaptive probes preserve IPW unbiasedness)**.** *Let $\mathcal{F}_{t-1}$ be the history before hand t. Suppose hands are dealt freshly and independently of $\mathcal{F}_{t-1}$, so $P(h^{(t)}_{\mathrm{opp}} = h \mid \mathcal{F}_{t-1}) = P(h)$; the adaptive probe σ_t is $\mathcal{F}_{t-1}$-measurable and its realized propensity $p_t(h) = P(S_t = 1 \mid h_{\mathrm{opp}} = h, \mathcal{F}_{t-1})$ is known with $p_t(h) \geq p_{\min} > 0$. Then*

$$Z_t = \mathbf{1}\{h^{(t)}_{\mathrm{opp}} \in T,\, S_t = 1\}/p_t(h^{(t)}_{\mathrm{opp}})$$

satisfies $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = \tau$, and the adaptive IPW average is unbiased; moreover, for a deterministic number of hands n and every $\delta > 0$, $P\big(|n^{-1}\sum_{t\leq n} Z_t - \tau| \geq \delta\big) \leq 2\exp(-2n\delta^2 p^2_{\min})$, the bound of Theorem 4.2. If $T \neq \emptyset$ and the weights instead use a stale but predictable ($\mathcal{F}_{t-1}$-measurable) $\tilde{p}_t \geq p_{\min}$, the reweighted term $\tilde{Z}_t$ has conditional bias $\big|\mathbb{E}[\tilde{Z}_t \mid \mathcal{F}_{t-1}] - \tau\big| \leq \tau \sup_{h\in T} \big|p_t(h)/\tilde{p}_t(h) - 1\big|$.

E4a: Adaptive probes. The block update rule is:

Table 8: The second F_B step, the only step in Table 7 whose intermediate objective is nonzero, decomposed at all four total budgets on the same fresh run (10 seeds per row). The $N = 6{,}000$ row is the F_B, k=2 row of Table 6. With m_k and m_{k+1} the minimum before and after the step, $\Delta_{\mathrm{w}} = \mathrm{mid} - m_k$ and $\Delta_{\mathrm{r}} = m_{k+1} - \mathrm{mid}$, and the bound is $2m$, so $\mathrm{bound}_{k+1} - \mathrm{bound}_k = 2(\Delta_{\mathrm{w}} + \Delta_{\mathrm{r}})$. The bound rises although the width falls at every budget, because the radius term exceeds the weight term.

N	Δ_{w}	Δ_{r}	mid	bound	width
6,000	-3.658	5.457	5.851	$19.02 \to 22.62$	$1.212 \to 1.142$
15,000	-1.700	2.339	2.340	$8.08 \to 9.36$	$0.898 \to 0.787$
60,000	-0.590	0.688	0.585	$2.35 \to 2.55$	$0.509 \to 0.421$
150,000	-0.285	0.311	0.240	$1.05 \to 1.10$	$0.316 \to 0.238$

Table 9: E7, N=5,000 hands, $\lambda = 0.2$, $\alpha = 0.05$, $\varepsilon = 0.1$, 50 seeds: equal-budget single-probe designs on the downstream LP-R certificate. Capacity-0.1 minus Nash has paired mean margin difference 0.0624 (95% t interval 0.0534 to 0.0714); MI-selected minus capacity-0.1 has difference 0.0497 (0.0416 to 0.0577). Call-station, a probe that always calls and never folds, violates the safety budget with exploitability 1.381 and never enters a safe selection set.

Probe	θ_e	LP-R margin	true value	widths (Nash / cap-0.1)
Nash	0.162	0.030 ± 0.018	0.082	1.276 / 1.046
capacity ε=0	0.166	0.031 ± 0.019	0.083	1.274 / 1.045
capacity ε=0.1	0.383	0.093 ± 0.026	0.202	1.884 / 0.633
total-showdown (MI-selected)	0.383	0.142 ± 0.019	0.208	1.849 / 0.626
mixture 0.5	0.273	0.054 ± 0.023	0.147	1.463 / 0.813
call-station	0.667	0.000	-0.046	1.920 / 0.752

1. Set the current probe to the Nash plan.
2. Sample one full block of 200 hands with the current probe.
3. Compute the block's target-showdown rate $\hat{r}$.
4. Switch to the capacity plan if $\hat{r} < 0.15$, switch back to the Nash plan if $\hat{r} > 0.35$, and keep the current plan otherwise.
5. Repeat from step 2. A trailing partial block is sampled but never updates the plan.

Realized propensities stay bounded away from zero, so HT–IPW is near-unbiased while the naive and stale-propensity estimators carry the MNAR bias, Table 11.

E AUDITS

E.1 FIXED-BUDGET GATE AUDIT (E5, COROLLARY 7.4)

The radius-term sign $\Delta_{\mathrm{r}} \geq 0$ in Corollary 7.4 is asserted conditionally on coordinatewise growth of the realized radii. A pre-existing cell's plug-in variance $\hat{m}(1 - \hat{m})$ can fall with the new data allocation, enough to offset the smaller per-probe sample and the larger logarithmic factor. The audit checks the premise directly on Leduc across the 10 seeds and all 160 stacking steps of the fresh run behind Tables 6, 7, and 8. The premise $r_{\mathrm{old}}^{(k+1)} \geq r^{(k)}$ holds coordinatewise-exactly on $126/160 = 78.75\%$ of stacking steps; at the coordinate level 99.97% of pre-existing-cell radii satisfy it, but a single violating coordinate defeats the step-level premise. Violations occur exactly when the plug-in variance drops.

Re-solving the old, intermediate, and new half-width objectives from the stored cell radii reproduces all 40 F_A transitions, with maximum old/new objective discrepancy 2.2×10^{-14}. The intermediate minimum is zero on every transition. At $N = 6{,}000$, the seed-mean terms are $\Delta_{\mathrm{w}} = -5.0639$ and $\Delta_{\mathrm{r}} = 1.3693$, giving the full-bound change $2(-5.0639 + 1.3693) \approx -7.3891$ shown in Table 6. Thus this tightening occurs with a positive radius term. The stored logs for F_B and F_C omit cell identities, so all three families are recomputed from a fresh sample. Per-probe cell identities are

Table 10: The deployment-side lines compared on inputs, safety object, and the form of the guarantee.

	CS-RNR (Li & Huang, 2026)	Data-driven restricted responses (Johanson et al., 2007; Ponsen et al., 2010)	This work
Input statistic	anytime-valid confidence sequence on pooled action frequencies	opponent model fitted to observed play	cell-moment confidence box from censored records
Safety object	the deployed strategy itself	the deployed mixture, through its Nash weight p	the deployed plan against every $y \in Y$: $\mathrm{expl}(x) \le \varepsilon$
Guarantee form	full-tree best-response certificate against a budget $\varepsilon_{\max}$	worst-case loss controlled by p, which floors value against every opponent including the actual one	exploitability certificate over all of Y, plus a one-sided finite-sample lower bound on value over the confidence set U_n at coverage $1-\alpha$
Bounds deployed value	bounds loss against a reference, not value against the actual opponent	a worst-case floor, with no data-dependent statistical refinement of it	yes: a data-dependent lower bound with $1-\alpha$ coverage over the sampling distribution

Table 11: E4a: mean bias of the hand-marginal estimators under the predictable adaptive probe. Realized propensities stay above $p_{\min} = 0.203$ and HT weights below 4.93; the payoff widths match, 1.876 for the adaptive probe and 1.875 for the fixed mean.

Estimator	bias at n=5,000	bias at n=200,000
naive	-0.074	-0.073
HT–IPW (adaptive)	-0.0014	$+0.0016$
stale-propensity	$+0.263$	$+0.264$

frozen at the same budget splits, and Table 6 reports those transitions. The intermediate minimum vanishes on the steps whose added probe spans the deployment's payoff vector together with F.

E.2 Identification audit (E3a, Theorems 6.4 and 6.7)

The population audit on Leduc agrees with the sharp identified set to the tolerances below. The Nash probe induces 306 showdown and 96 fold cells with $\mathrm{rank}([F;\Phi_x]) = 302$ and nullity 35. The capacity-maximizing probe at $\varepsilon = 0.1$ induces 129 showdown and 30 fold cells with rank 202 and nullity 135. Every probe's own value has identified width $\le 10^{-15}$ (Theorem 6.7). Cross-probe deployments are partially identified: at $\lambda = 0.2$ the Nash plan against the moments of the $\varepsilon = 0.1$ probe leaves a payoff interval of width 1.616, and the $\varepsilon = 0.02$ plan against Nash moments one of width 0.015. At $\lambda = 1$ the same cross-deployment has width 1.618, so the ambiguity is stable across censoring levels. Stacking all five audited probes gives rank 308, nullity 29, and pins all five candidate-deployment values to width $\le 10^{-15}$. At $\lambda = 0.2$, 48 of 337 realization coordinates remain unpinned under the stacked moments, of which 28 lie on the opponent's support.

Here "nullity" means $\dim\ker[F;\Phi_x]$, the dimension of the kernel of the moment constraints. The identified set lies in a translate of that kernel, so nullity is an upper bound on the affine dimension of the identified set (Theorem 6.4); it does not count the coordinates with positive LP width. Unpinned-coordinate counts can exceed it: 56 free coordinates against nullity 35 for the Nash probe, 215 against nullity 135 for the $\varepsilon = 0.1$ probe, and 48 against nullity 29 for the stack. A low-dimensional polytope can still project non-trivially onto many coordinates. These widths come from linear programming over the identified set.

E.3 EQUAL-BUDGET PROBE DESIGN (E7)

E7 compares six single-probe designs on the downstream LP-R certificate, each with N=5,000 hands, α=0.05, safety budget ε=0.1, and censoring level λ=0.2, over the same 50 seeds (Table 9): Nash, the capacity probes at ε=0 and ε=0.1, the MI-selected probe, a 0.5 mixture of the ε=0.1 capacity probe and Nash, and a call-station probe. The probe selected by nominal-model mutual information, written MI-selected below, is the best among a finite pool of 22 safe LP candidates, all solved at the same budget ε=0.1 under the *nominal* censoring model with λ_{design}=0. The pool holds the target-capacity probe, the total-showdown-rate probe, and 20 probes maximizing independent Gaussian-random linear objectives generated with seed 1234. Candidates are scored by their nominal-model mutual information $\mathrm{MI}(\mathrm{rank}_2; \mathrm{cell})$, the mutual information between the opponent's private-hand rank rank_2 and the recorded terminal cell.

The winner is fixed once before sampling and reused across all seeds, and it is the analytical total-showdown probe, with nominal-model MI 0.835, which outranks all Gaussian-objective candidates. Samples are then drawn at λ_{eval}=0.2 from a separate evaluation opponent that is used only for sampling and truth evaluation, never for probe design. Its moment-map nullity is $\dim\ker[F; \Phi_x]$=71, versus 135 for the capacity probe at ε=0.1. The paired mean difference in true deployed value, MI-selected minus capacity-0.1, is 0.0057, with a 95% t interval from -0.0001 to 0.0114 that contains zero; the stronger separation is in certificate margin (Table 9).

E.4 FULL-SUPPORT MIXTURES AND PAYOFF-ONLY SETS (E3D, E3E)

E3d mixes the capacity-optimal probe x_{cap} at $\varepsilon = 0.1$ with one fixed safe full-support plan $x_{\text{fs}} = (1 - t^*)\, x_{\text{Nash}} + t^*\, x_{\text{unif}}$, where x_{unif} is the implementation plan of uniform behavior at every hero information set and $t^* = 0.0445$ is the largest weight with $\mathrm{expl}(x_{\text{fs}}) \le 0.1$, found by bisection on the convex exploitability. The probe at weight s is $(1 - s)x_{\text{cap}} + s\, x_{\text{fs}}$, with the same x_{fs} for every s, $n = 5{,}000$ hands, and unsplit $\alpha = 0.05$. Every row of Table 4 is a mean over 50 seeds; the seed standard deviation of the margin is 0.019–0.024 and that of every width is at most 0.024.

E3e replaces the cell box of each probe by one two-sided interval on its mean observed payoff. Payoffs lie in $[-13, 13]$, in units of the ante, the forced initial bet; this is the range of the hero's terminal payoff on the enumerated tree. With $n = 5{,}000$ hands per probe and $\alpha_i = \alpha/2$, the radius is the smaller of the Maurer–Pontil empirical–Bernstein radius, $26\big(\sqrt{2\hat{v}\log(8/\alpha_i)/(n-1)} + \frac{7}{3}\log(8/\alpha_i)/(n-1)\big)$ with $\hat{v}$ the sample variance of the range-normalized payoffs, and the Hoeffding radius $26\sqrt{\log(4/\alpha_i)/(2n)}$. The constraint set is $\{y \in Y :\ \hat{\mu}_i - r_i \le (A^T x_i)^T y \le \hat{\mu}_i + r_i,\ i = 1, 2\}$, and LP-R re-optimizes the deployment over it as over the cell box. The exact-means arm sets $r_i = 0$ at the population mean payoff. The comparison is between full-cell constraints and two on-policy mean-payoff constraints from the same records.

E.5 FIXED-BUDGET STACKING PROTOCOL (E5)

E5 fixes a total budget across k probes and splits it equally, applying Bonferroni over probes at every stacking step ($\alpha_i = \alpha/k$, α=0.05, evaluation opponent λ=0.2, target T={J, Q}, 10 seeds), for totals $N \in \{6{,}000, 15{,}000, 60{,}000, 150{,}000\}$. It compares three probe families and reports, per step, the L1-dual bound and realized width of the covered ε=0.1 deployment (and of Nash where noted; Table 6): F_A adds the ε=0.1 capacity probe to Nash (rank-redundant, 302 → 302); F_B adds the ε=0 probe and then the ε=0.02 probe to Nash (rank-increasing, 302 → 304 → 308); F_C starts from the three-probe base {Nash, ε=0, ε=0.02} and adds ε=0.1 (308 → 308).

The bound of Theorem 7.2 is $2 \cdot \min\{\sum_{i,\kappa} r_{i,\kappa}|w_{i,\kappa}|\}$ over $F^T\lambda + \Phi^{(k)T} w = A^T z$. The ℓ_1 minimum is a half-width target and the factor 2 sits in the width bound. The decomposition evaluates the added probe's moments at zero radius (the zero-padded level-k certificate) to isolate the weight change. When the added probe raises the rank, as in both F_B steps, the refined set does contain new moment directions; the zero-radius intermediate then measures the representational change and the new directions together.

Table 12: Experimental configuration and numerical discipline common to all blocks of Section 8.

Setting	Value
Game	Leduc Hold'em (three card ranks, two suits); target $T = \{\mathrm{J}, \mathrm{Q}\}$
LP solver	HiGHS (presolve off; primal/dual feasibility tolerances 10^{-9}); LP-R check $< 10^{-7}$
Censoring	$\lambda = 0.2$ (identification and fixed-budget blocks)
Sampling	$n_i = 5{,}000$ hands, $\alpha = 0.05$; 50 seeds (E2, E3b, E3d, E3e, E7, E8), 10 (E3c, E5)
E5 budget	totals 6,000–150,000 across probes
Width tolerance	free/pinned if width $> 10^{-6}$; rank via SVD, $\approx 10^{-13}$

E.6 Solver discipline

All LP constructs (LP-C, LP-R, the L1-dual width bound, and every per-coordinate width) are solved with the HiGHS LP solver (presolve disabled; primal and dual feasibility tolerances 10^{-9}). Each LP-R optimum is validated by re-solving the inner minimization over the realized set and checking agreement below 10^{-7}; the exploitability budget $\mathrm{expl}(\hat{x}) \leq \varepsilon$ is verified directly from the certificate. Numerical ranks are computed by singular-value thresholding on the stacked constraint matrices $[F; \Phi^{(k)}]$ with threshold $\sigma_1 \cdot \max(\text{shape}) \cdot \varepsilon_{\text{float64}}$ ($\approx 10^{-13}$ on Leduc).

Confidence boxes are built over all terminal cells reachable by the probe with positive probability, enumerated before sampling; zero-frequency cells are retained and counted in the fixed K. The failure budget is split by Bonferroni over probes, $\alpha_i = \alpha/k$ per probe ($k = 1$ for a single probe). Within a probe the per-cell budget and its split between the empirical–Bernstein and Hoeffding candidate radii are those of the simultaneous box in Theorem 6.9(d). The 10^{-6} tolerance is a width-classification tolerance: a realization coordinate or deployment is reported free (or ambiguous) when its LP-measured width exceeds 10^{-6}, and pinned otherwise. All cell partitions are enumerated on the full Leduc state space. Seeds and configurations are fixed and reported in the experiment logs.

E.7 Censoring family and plug-in response

The censoring family is generated from the opponent's equilibrium behavior. Writing σ_2 for its Nash behavioral strategy, μ_λ agrees with σ_2 at every information set except the round-1 sets where the opponent holds a weak hand, J or Q, and faces a bet or a raise. At those sets the call probability is scaled by λ and the removed mass is added to the fold probability; every other action probability is unchanged. At $\lambda = 0$ a weak hand never calls a round-1 bet, which lowers the target showdown rate and concentrates showdown records on strong hands.

The plug-in response of E3b is built from the unique minimum-norm point of the confidence set: the plan y that satisfies the cell-moment box and the flow constraints and minimizes $\|y\|_2$. A safety-constrained best response to that single plan gives the deployed probe, and the reported number is the worst-case payoff of that probe over the same confidence set, averaged over the 50 seeds: -0.0249 under the empirical–Bernstein box against $v^* = -0.0856$, and -0.1703 under the Hoeffding box. The minimum-norm point is one canonical representative of the confidence set. The robust response maximizes over the whole set and attains a mean margin of 0.0822.

F Direct Payoff Certification Baseline

The probe's own value can be estimated from its observed payoffs without constructing an opponent confidence set. We compare this direct route with LP-R on the same data.

Observed payoffs and fixed ranges. For each probe x_j, the recorded terminal cell determines the hero's payoff $G_{j,t}$. At a fold, this payoff depends on the betting history and on which player folds; the opponent's hidden card plays no role. At showdown, all required cards are observed. Before sampling, we enumerate every cell reachable by x_j against any opponent and compute its payoff range $[a_j, b_j]$, so the range carries no sampling error. For $a_j < b_j$, let $Z_{j,t} = (G_{j,t} - a_j)/(b_j - a_j) \in [0, 1]$, with empirical mean $\bar{Z}_j$ and unbiased sample variance S_j^2.

Table 13: E8: matched direct-payoff versus LP-R deployment, mean ± standard deviation over 50 seeds, in units of the ante. Direct certification selects Nash in all $50/50$ runs of every row: its gain is 0, value -0.0460, and gap 0.0396, each with zero standard deviation. LP-R selects a non-Nash deployment in every run. Both methods are safe and their selected-value certificates cover in $50/50$ runs of every row; no numerical failures occur. LP-R's MI-selected value standard deviation is 7.69×10^{-6}, rounded to 0.0000 in the table. Two-sided 95% t intervals at 49 degrees of freedom for the LP-R gain, in row order, are 0.0253 to 0.0354, 0.0255 to 0.0363, 0.0855 to 0.1000, 0.0476 to 0.0608, 0.1369 to 0.1480, and 0.0738 to 0.0875.

Records	LP-R gain	LP-R value	LP-R gap
Single probe: $N = 5{,}000$			
Nash	0.0303 ± 0.0178	0.0815 ± 0.0743	0.1368 ± 0.0578
Capacity-0	0.0309 ± 0.0189	0.0833 ± 0.0736	0.1380 ± 0.0564
Capacity-0.1	0.0928 ± 0.0255	0.2020 ± 0.0203	0.1948 ± 0.0213
Mixture	0.0542 ± 0.0233	0.1465 ± 0.0615	0.1780 ± 0.0439
total-showdown (MI-selected)	0.1424 ± 0.0194	0.2077 ± 0.0000	0.1508 ± 0.0194
Three probes: $N = 15{,}000$			
Nash + capacity-0/0.1	0.0806 ± 0.0241	0.1973 ± 0.0260	0.2022 ± 0.0210

Scalar lower bounds. With failure budget α_j for a probe and $n_j \geq 2$, define

$$r_j^{\mathrm{H}} = \sqrt{\frac{\log(2/\alpha_j)}{2n_j}}, \qquad r_j^{\mathrm{EB}} = \sqrt{\frac{2S_j^2 \log(4/\alpha_j)}{n_j}} + \frac{7\log(4/\alpha_j)}{3(n_j - 1)}.$$

The one-sided Hoeffding and empirical–Bernstein inequalities (Hoeffding, 1963; Maurer & Pontil, 2009) each have failure probability at most $\alpha_j/2$. Thus

$$L_j = a_j + (b_j - a_j)\big(\bar{Z}_j - \min\{r_j^{\mathrm{H}}, r_j^{\mathrm{EB}}\}\big)$$

is a lower bound on $x_j^T A y^*$ with failure probability at most α_j. If $a_j = b_j$, its value is known exactly. We also allow the known-game worst-case floor $\min_{y \in Y} x_j^T A y$, and include a Nash strategy with deterministic certificate v^*. The direct method selects the candidate with the largest certificate using a fixed tie-breaking order. For k sampled probes we use $\alpha_j = \alpha/k$, so all scalar bounds hold simultaneously with probability at least $1 - \alpha$, including after deployment selection.

Matched comparisons. Both methods use $\alpha = 0.05$ and deployment safety budget $\varepsilon = 0.1$. The five safe E7 probes are compared individually using their respective $5{,}000$-hand batches. A separate three-probe comparison uses Nash, capacity-0, and capacity-0.1, with $5{,}000$ hands per probe and $15{,}000$ hands in total. In each comparison, direct certification and LP-R receive the same records, and LP-R builds its cell-frequency boxes with the same per-probe failure budgets.

The designs use $\lambda_{\mathrm{design}} = 0$ and samples use $\lambda_{\mathrm{eval}} = 0.2$. Seeds are $s = 0, \ldots, 49$, with sampling seed $1000s + 7$ for each probe, matching E7. The three-probe comparison reuses these batches, whereas the original E3b replication uses offsets $0, 1, 2$; these are separate draws. We report certified gain above v^*, actual deployment value, their certificate gap, safety, empirical coverage, and Nash fallback frequency. Means and standard deviations are over seeds, with standard deviations using divisor 49. Paired mean differences use two-sided t intervals with 49 degrees of freedom. Numerical failures remain in the record and are reported separately from a deliberate Nash fallback.

The direct bounds use the fixed range $[-13, 13]$ for all five safe probes. None exceeds the Nash certificate in these samples. LP-R uses the full cell distribution to select a different deployment, producing positive certified gains in Table 13. The direct method selects Nash in every run, so its certified gain is zero and each tabulated LP-R gain is itself the paired difference. A two-sided t interval at 49 degrees of freedom adds 2.01 standard errors to the tabulated mean, and all six intervals lie strictly above zero. The direct method has the smaller certificate gap on its selected Nash deployment, while LP-R reaches higher values by certifying different responses. The three-probe LP-R result differs from E3b because it reuses the E7 batches specified above.

G CONCLUSION

Safe probing under showdown censoring reduces to three linear programs. LP-C maximizes the target-showdown mass a safe probe can collect. Robust deployment over the identified moment polytope certifies a one-sided finite-sample value guarantee. A dual-certificate analysis separates a weight term, which can only decrease, from a radius term that moves with the realized sampling radii. The radius term is nonnegative when the old-cell radii grow coordinatewise (Section 7). The three constructions share one mechanism, the sequence-form exploitability certificate, and one scope, an arbitrary finite perfect-recall two-player zero-sum game with hero-observable terminal cells. On Leduc the capacity probe collects $2.4\times$ the target showdowns of Nash at the same exploitability budget. The cell polytope certifies a positive margin where a payoff-only certificate certifies none, and a rank-increasing probe can loosen the width bound even as the width falls (Appendix E). Pre-data probe selection and probes for rank or robust value remain open.