OFF-POLICY LEARNING OF CAPACITY-CONSTRAINED ALLOCATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study off-policy learning of capacity-constrained allocations from individual logged rewards under a shared linear model and an unknown logger that may assign zero probability to available actions. Capacity constraints determine which unresolved reward directions matter, captured by a source-radius modulus. We characterize the exact infinite-data identification threshold for a fixed source law and propose a minimax-regret allocation rule with high-probability worst-batch guarantees without overlap or minimum-eigenvalue assumptions. Marginal and integral outputs can have fundamentally different sample complexity. A constructed family requires $\Theta(\log(1/\delta))$ samples for marginal outputs but $\Theta(\rho^{-1}\log(1/\delta))$ for integral outputs. Experiments recover the predicted finite-sample behavior, and a semi-synthetic study on real LLM query contexts shows lower empirical failure and regret under source ambiguity for MRA than regression and robust baselines.

1 INTRODUCTION

Allocating scarce resources from historical data requires joint decisions, with each assignment changing the options available to the others under limited capacity. In applications such as matching workers to jobs, routing patients to hospitals, or assigning computing requests to services, the learner must choose a joint allocation using logs generated by an unknown logging policy. The logs record rewards only at the individual context–action level, so evaluating alternative feasible assignments requires information about many pairs that were never observed. Offline contextual optimization and combinatorial bandits provide frameworks for learning structured decisions from partial feedback (Hu et al., 2024; Liu et al., 2025). Our central finding is that allowing the learner to randomize over feasible assignments can change what must be learned from the logged data, even when the hard capacity constraints are unchanged.

A natural approach is to estimate the reward model and optimize a feasible assignment under the estimate. We assume a shared linear reward model with known features. The historical logs may contain no information about some reward directions, so collecting more data in the same way need not recover the full reward model. Yet full recovery is not always necessary for good allocation decisions. Regret depends only on value differences between feasible assignments, and the capacity constraints determine which comparisons are possible. At the same time, an assignment optimized for one fitted model can perform poorly under another model with the same logged conditional means. The central question is therefore *which reward differences must be learned well enough for allocation, and which can remain uncertain while still permitting a good allocation?*

We show that the reward information needed for a good allocation can depend on whether the policy returns an integral assignment or a distribution over feasible assignments. Every realized assignment obeys the hard capacities, while performance is evaluated in expectation over the randomization. We construct a family in which frequently observed rewards suffice to choose a good marginal, whereas choosing a good integral assignment requires information available only from a rare source type. When that type is never observed, the target accuracy remains attainable with a marginal output, whereas no learner restricted to integral assignments can guarantee it. *Mixing can change the information required for allocation without relaxing feasibility.*

We develop Minimax-Regret Allocation (MRA), which constructs a regression confidence set and searches for a feasible marginal with small worst-case regret over the candidate models, each compared with its own optimal assignment. A double-oracle procedure provides certificates for

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.

the target regret and the optimization gap. Our analysis introduces a source-radius modulus whose zero-radius value captures irreducible source-fiber ambiguity and whose positive radii quantify how finite-sample uncertainty in source predictions translates into allocation regret. Together, these ingredients yield simultaneous worst-batch PAC control even when the source covariance is singular. Our general guarantees apply to the shared-feature model. We complement them with two explicit families that establish information separation and sharpness. We summarize our contributions below.

- **Allocation guarantees under partial identification.** For each fixed source law, we characterize the exact infinite-data identification threshold and prove worst-batch PAC guarantees for MRA. The source-radius modulus links finite-sample uncertainty to this limit, while finite incompatible model sets yield information lower bounds (Sections 4–6).
- **Sharp information separations.** Two families show complementary effects of unresolved source information. Marginal outputs remove a ρ^{-1} rare-observation factor unavoidable for integral assignments, while approaching a positive identification floor requires $\Theta(B^2(\rho\tau^2)^{-1}\log(1/\delta))$ samples. MRA attains both rates, and the second family requires a three-model witness despite pairwise compatibility (Theorems 9 and 10).
- **Empirical evaluation.** Synthetic experiments recover the predicted rare-observation separation and the $\rho^{-1}\tau^{-2}$ scaling above the identification floor. A controlled semi-synthetic study on real LLM query contexts further evaluates allocation under source ambiguity and compares MRA with regression, maximin, and reference-pessimistic baselines (Section 8).

2 Related Work

Off-policy and resource-constrained learning. Off-policy learning from logged bandit feedback uses doubly robust, counterfactual-risk, and balancing methods (Dudík et al., 2011; Swaminathan & Joachims, 2015; Kallus, 2018). Constraints appear in batch policy learning and contextual bandits with knapsacks (Le et al., 2019; Agrawal & Devanur, 2016). Offline contextual linear optimization and combinatorial bandits rely on conditional action-span and base-arm coverage (Hu et al., 2024; Liu et al., 2025), while support-aware and pessimistic methods address limited coverage (Sachdeva et al., 2020; Li et al., 2022; Jin et al., 2025). Beyond overlap, off-policy evaluation gives sharp partial-identification bounds under smoothness (Khan et al., 2024). We study hard within-batch capacities and worst-batch guarantees with unidentified allocation-relevant comparisons. MRA targets minimax regret against candidate optima rather than maximin value.

Decision-relevant information. Decision-focused learning trains predictors for downstream decision quality rather than prediction accuracy (Donti et al., 2017; Wilder et al., 2019). Recent work studies decision-sufficient linear measurements (Bennouna et al., 2025; Ye et al., 2026), while transductive linear bandits distinguish measurement directions from the target decision set (Fiez et al., 2019). For resource-constrained treatment allocation, ranking conditional average treatment effects can suffice without accurate magnitude estimates (Kamran et al., 2024). We quantify achievable allocation regret from logs under a fixed unknown source law, including positive identification floors.

Partial identification and multiple answers. Minimax-regret under partial identification has been studied for new treatments, asymmetric counterfactual utilities, and binary Gaussian decisions (Higbee, 2025; Ben-Michael et al., 2024; Yata, 2021). Confounding-robust off-policy learning under marginal sensitivity models gives sharp value bounds (Hess et al., 2026), while pure exploration with multiple correct answers studies information through alternative models (Degenne & Koolen, 2019). For a fixed logging law, our lower bounds use models with no common ϵ-good marginal. Allocation geometry bounds the witness size, while our coupled construction shows pairwise compatibility need not extend to three models.

Randomization and minimax regret. Minimax regret is classical in combinatorial optimization under uncertainty (Aissi et al., 2009). Randomized formulations use distributions over feasible solutions (Mastin et al., 2015) and can improve robust decisions under ambiguity (Delage et al., 2019). Mixed robust strategies can be learned from noisy evaluations (Sessa et al., 2020). Double-oracle methods use best responses and support interval-data minimax-regret lower bounds (McMahan et al., 2003; Gilbert & Spanjaard, 2017). With an offline confidence set and exact best responses, MRA achieves worst-batch PAC control and matching rates on two explicit families.

3 PROBLEM FORMULATION

For $n \geq 1$, let $[n] := \{1, \ldots, n\}$. All vectors are column vectors. For $\mathbf{A} \succeq 0$, write $\|\mathbf{x}\|_{\mathbf{A}} := \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$. For matrices, $\mathbf{A}^\dagger$ denotes the Moore–Penrose pseudoinverse. We use rank, range, $\ker$, span, conv, and $\langle \cdot, \cdot \rangle$ with their standard meanings. Vector inequalities are componentwise.

Offline data and reward model. The learner observes $\mathcal{D}_n := \{(X_t, A_t, Y_t)\}_{t=1}^n$, where $X_t \in \mathcal{X}$, $A_t \in \mathcal{A}(X_t)$, and Y_t is the observed reward. The known problem description specifies $\mathcal{X}$, available actions $\mathcal{A}(x) \subseteq [K]$, a bounded measurable feature map $\boldsymbol{\phi}(x, a) \in \mathbb{R}^d$, and a declared batch domain $\mathcal{I}$. Rewards lie in the known interval $[0, B]$ for some $B > 0$. We work on the effective feature span $\mathrm{span}\{\boldsymbol{\phi}(x, a) : x \in \mathcal{X},\ a \in \mathcal{A}(x)\} = \mathbb{R}^d$ and take the shared linear mean parameter from

$$\Theta := \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : 0 \leq \boldsymbol{\phi}(x, a)^\top \boldsymbol{\theta} \leq B \ \ \forall x \in \mathcal{X},\ a \in \mathcal{A}(x) \right\}. \tag{1}$$

Thus Θ is compact and convex without a Euclidean norm bound.

Assumption 1 (Offline sampling and deployment transport)**.** *The records in $\mathcal{D}_n$ are iid, with $(X_t, A_t) \sim \lambda$ for an unknown law λ. For some $\boldsymbol{\theta} \in \Theta$ fixed before sampling,*

$$0 \leq Y_t \leq B, \qquad \mathbb{E}[Y_t \mid X_t, A_t] = \boldsymbol{\phi}(X_t, A_t)^\top \boldsymbol{\theta},$$
$$0 \leq Y_i(a) \leq B, \qquad \mathbb{E}[Y_i(a) \mid I, \mathcal{D}_n] = \boldsymbol{\phi}(X_i, a)^\top \boldsymbol{\theta}.$$

for every $I \in \mathcal{I}$, unit i, and $a \in \mathcal{A}(X_i)$. Capacities and other assignments affect feasibility but not these conditional means. Otherwise, the source and deployment reward laws are unrestricted.

Set $\boldsymbol{\Sigma}_\lambda := \mathbb{E}_\lambda[\boldsymbol{\phi}(X, A)\boldsymbol{\phi}(X, A)^\top]$ and $r_\lambda := \mathrm{rank}(\boldsymbol{\Sigma}_\lambda)$. The source may omit context–action pairs, so $\boldsymbol{\Sigma}_\lambda$ may be singular. No overlap or minimum-eigenvalue condition is imposed.

Capacity-constrained allocations. A batch $I = (X_1, \ldots, X_m, \mathbf{b})$ has m contexts and integer capacities $\mathbf{b} \in \mathbb{Z}_+^K$. Its feasible assignments are

$$\mathcal{Z}(I) := \left\{ \mathbf{z} \in \{0,1\}^{m \times K} : \begin{array}{c} \displaystyle\sum_{a=1}^K z_{ia} = 1 \ (i \in [m]), \quad \sum_{i=1}^m z_{ia} \leq b_a \ (a \in [K]), \\ z_{ia} = 0 \ (i \in [m],\ a \notin \mathcal{A}(X_i)) \end{array} \right\}. \tag{2}$$

We assume $\mathcal{Z}(I) \neq \varnothing$ for every $I \in \mathcal{I}$ and fix measurable rules for nonunique selections. Let $\mathcal{P}(I) := \mathrm{conv}\, \mathcal{Z}(I)$. By network-flow integrality, replacing binarity in (2) by nonnegativity yields the same polytope (Schrijver, 2003). Thus every $\mathbf{Q} \in \mathcal{P}(I)$ is the marginal of a distribution over feasible integral assignments, with Q_{ia} the probability that unit i is assigned to action a. Sampling from such a distribution preserves feasibility, and linearity makes value and regret depend only on $\mathbf{Q}$. For $\mathbf{Q} \in \mathcal{P}(I)$ and $\boldsymbol{\theta} \in \Theta$, define

$$\boldsymbol{\Phi}(I, \mathbf{Q}) := \frac{1}{m} \sum_{i=1}^m \sum_{a=1}^K Q_{ia} \boldsymbol{\phi}(X_i, a), \qquad V_{\boldsymbol{\theta}}(I, \mathbf{Q}) := \boldsymbol{\Phi}(I, \mathbf{Q})^\top \boldsymbol{\theta}, \tag{3}$$

$$V_{\boldsymbol{\theta}}^*(I) := \max_{\mathbf{z} \in \mathcal{Z}(I)} V_{\boldsymbol{\theta}}(I, \mathbf{z}), \qquad \Delta_{\boldsymbol{\theta}}(I, \mathbf{Q}) := V_{\boldsymbol{\theta}}^*(I) - V_{\boldsymbol{\theta}}(I, \mathbf{Q}). \tag{4}$$

By construction, $0 \leq \Delta_{\boldsymbol{\theta}}(I, \mathbf{Q}) \leq B$.

Learner information and objective. The learner receives the known problem description, B, $\mathcal{D}_n$, the current batch I, and (ϵ, δ), but not λ or $\boldsymbol{\Sigma}_\lambda$, and no deployment samples in advance. It may randomize and, conditional on $\mathcal{D}_n$ and its randomness, induces a policy π with $\pi(I) \in \mathcal{P}(I)$. Randomness used only to realize a marginal is averaged in its value. Probabilities below are over the offline data and learner randomness. We assume all policy evaluations and suprema are measurable. Define the worst-batch regret

$$R_{\boldsymbol{\theta}}(\pi) := \sup_{I \in \mathcal{I}} \Delta_{\boldsymbol{\theta}}(I, \pi(I)). \tag{5}$$

For an environment class $\mathcal{E}$, where each $e \in \mathcal{E}$ specifies $\boldsymbol{\theta}_e$, λ_e, and admissible reward laws, define

$$N^*(\epsilon, \delta, \mathcal{E}) := \inf \left\{ n \geq 0 : \inf_{\widehat{\pi}_n} \sup_{e \in \mathcal{E}} \mathbb{P}_e\{R_{\boldsymbol{\theta}_e}(\widehat{\pi}_n) > \epsilon\} \leq \delta \right\}.$$

The guarantee is uniform over $\mathcal{E}$. Define $N^{(1)}(\epsilon, \delta; \mathcal{E})$ similarly, restricting to learners with $\widehat{\pi}_n(I) \in \mathcal{Z}(I)$ almost surely for every $I \in \mathcal{I}$. Learner-level randomization across integral outputs is allowed.

4 ALLOCATION GEOMETRY AND PARTIAL IDENTIFICATION

For a fixed batch, only differences between feasible allocation features affect decisions, while source data identify the reward parameter only up to source-null directions. For $I \in \mathcal{I}$, define

$$\mathbf{L}_I := \operatorname{span}\{\mathbf{\Phi}(I,\mathbf{z}) - \mathbf{\Phi}(I,\mathbf{z}') : \mathbf{z},\mathbf{z}' \in \mathcal{Z}(I)\}, \qquad p_I := \dim(\mathbf{L}_I) \le d.$$

The source fiber of $\boldsymbol{\theta} \in \Theta$ consists of parameters with the same source conditional means:

$$[\boldsymbol{\theta}]_\lambda = (\boldsymbol{\theta} + \ker \mathbf{\Sigma}_\lambda) \cap \Theta. \tag{6}$$

Indeed, for $\mathbf{d} = \boldsymbol{\nu} - \boldsymbol{\theta}$, $\mathbb{E}_\lambda[(\boldsymbol{\phi}^\top \mathbf{d})^2] = \mathbf{d}^\top \mathbf{\Sigma}_\lambda \mathbf{d}$. The fiber concerns conditional means only. Lower bounds requiring identical source laws specify the reward distributions explicitly. For nonempty compact $\mathcal{H} \subseteq \Theta$, define the robust allocation loss

$$v_I(\mathcal{H}) := \min_{\mathbf{Q} \in \mathcal{P}(I)} \max_{\boldsymbol{\nu} \in \mathcal{H}} \Delta_{\boldsymbol{\nu}}(I, \mathbf{Q}).$$

Thus $v_I(\mathcal{H}) \le \epsilon$ exactly when some $\mathbf{Q} \in \mathcal{P}(I)$ is ϵ-good for every model in $\mathcal{H}$.

Proposition 2 (Finite support and witnesses)**.** *Every feasible aggregate feature is realized by mixing at most $p_I + 1$ integral assignments. If $v_I(\mathcal{H}) > \epsilon$, some $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_k \in \mathcal{H}$ with $k \le p_I + 1$ satisfy*

$$v_I(\{\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_k\}) > \epsilon.$$

The proof uses Carathéodory's and Helly's theorems (Matoušek, 2002) and is given in Appendix A.1. For $t \ge 0$, let $\mathcal{N}_\lambda(\boldsymbol{\theta}, t) := \{\boldsymbol{\nu} \in \Theta : \|\boldsymbol{\nu} - \boldsymbol{\theta}\|_{\mathbf{\Sigma}_\lambda} \le Bt\}$ and define

$$\omega_{\boldsymbol{\theta}}(t) := B^{-1} \sup_{I \in \mathcal{I}} v_I(\mathcal{N}_\lambda(\boldsymbol{\theta}, t)), \qquad \omega(t) := \sup_{\boldsymbol{\theta} \in \Theta} \omega_{\boldsymbol{\theta}}(t).$$

Thus $B\omega_{\boldsymbol{\theta}}(t)$ is the worst robust allocation loss over source-norm neighborhoods of radius Bt, uniformly over $I \in \mathcal{I}$. Since $\mathcal{N}_\lambda(\boldsymbol{\theta}, 0) = [\boldsymbol{\theta}]_\lambda$,

$$B\omega(0) = \sup_{\boldsymbol{\theta} \in \Theta,\ I \in \mathcal{I}} v_I([\boldsymbol{\theta}]_\lambda),$$

the largest loss remaining when the source conditional means are known exactly. The learner does not know λ, so Section 5 constructs $\mathcal{C}_n$ from the logs and approximately solves $v_I(\mathcal{C}_n)$.

5 MINIMAX-REGRET LEARNING

The learner first computes a regression plug-in allocation, then searches for a marginal with low worst-case regret over a source confidence set. It returns the current marginal when either the worst-case regret is certified as $\le \epsilon$ or the minimax-regret gap as $\le \zeta_n$. The plug-in initializes the assignment pool and handles early returns and computation failures.

Source confidence set. For $n \ge 1$, let $\boldsymbol{\phi}_t := \boldsymbol{\phi}(X_t, A_t)$ and $\mathbf{y} := (Y_1, \dots, Y_n)^\top$, and define

$$\mathbf{F}_n := (\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_n)^\top, \qquad \mathbf{G}_n := \mathbf{F}_n^\top \mathbf{F}_n, \qquad \widehat{\boldsymbol{\theta}}_n := \mathbf{G}_n^\dagger \mathbf{F}_n^\top \mathbf{y}, \qquad r_n := \operatorname{rank}(\mathbf{F}_n).$$

For $r_n > 0$, let $\mathbf{E}_n \in \mathbb{R}^{d \times r_n}$ be an orthonormal basis of $\operatorname{range}(\mathbf{F}_n^\top)$ chosen by a fixed measurable rule from the logged design, and define

$$\beta_n := B\sqrt{\frac{1}{2}\log\frac{4r_n}{\delta}}, \qquad \mathcal{C}_n := \left\{\boldsymbol{\nu} \in \Theta : \left\|\mathbf{E}_n^\top \mathbf{G}_n^{1/2}(\boldsymbol{\nu} - \widehat{\boldsymbol{\theta}}_n)\right\|_\infty \le \beta_n\right\}.$$

If $r_n = 0$, set $\beta_n := 0$ and $\mathcal{C}_n := \Theta$. The estimate $\widehat{\boldsymbol{\theta}}_n$ need not lie in Θ and only centers the constraints. The set $\mathcal{C}_n$ restricts $\operatorname{range}(\mathbf{F}_n^\top)$ while leaving $\ker(\mathbf{F}_n)$ unrestricted except through Θ.

Minimax-regret game. The plug-in allocation is

$$\mathbf{z}_0(I) \in \operatorname*{argmax}_{\mathbf{z}\in\mathcal{Z}(I)} \mathbf{\Phi}(I,\mathbf{z})^\top \widehat{\boldsymbol{\theta}}_n.$$

For $\mathcal{C}_n \neq \varnothing$, MRA uses a zero-sum game in which the learner chooses $\mathbf{Q} \in \mathcal{P}(I)$, the adversary chooses $\boldsymbol{\nu} \in \mathcal{C}_n$, and the payoff is regret against the integral optimum under $\boldsymbol{\nu}$. Define

$$g_{n,I}(\mathbf{Q}) := \max_{\boldsymbol{\nu}\in\mathcal{C}_n} \Delta_{\boldsymbol{\nu}}(I,\mathbf{Q}), \qquad v_{n,I} := \min_{\mathbf{Q}\in\mathcal{P}(I)} g_{n,I}(\mathbf{Q}) = v_I(\mathcal{C}_n).$$

Thus $v_{n,I} \leq \epsilon$ exactly when a feasible marginal is ϵ-good for every model in $\mathcal{C}_n$, unlike the maximin-value objective $\max_{\mathbf{Q}\in\mathcal{P}(I)} \min_{\boldsymbol{\nu}\in\mathcal{C}_n} V_{\boldsymbol{\nu}}(I,\mathbf{Q})$. The double oracle maintains assignments $\mathbf{z}_1,\dots,\mathbf{z}_J$ and models $\boldsymbol{\nu}_1,\dots,\boldsymbol{\nu}_M$. Define

$$h_j := V^*_{\boldsymbol{\nu}_j}(I), \qquad D_{j\ell} := h_j - \mathbf{\Phi}(I,\mathbf{z}_\ell)^\top \boldsymbol{\nu}_j, \qquad \mathbf{D} := (D_{j\ell})_{j\in[M],\,\ell\in[J]}.$$

Here $D_{j\ell} = \Delta_{\boldsymbol{\nu}_j}(I,\mathbf{z}_\ell)$ is the restricted regret matrix, with rows indexing models and columns indexing assignments. Let $\mathcal{S}_J := \{\mathbf{p}\in\mathbb{R}^J_+ : \mathbf{1}^\top\mathbf{p} = 1\}$ and $\mathcal{S}_M := \{\mathbf{w}\in\mathbb{R}^M_+ : \mathbf{1}^\top\mathbf{w} = 1\}$. The restricted game has primal and dual programs

$$\begin{aligned} \min_{\mathbf{p}\in\mathcal{S}_J,\gamma} \quad & \gamma & \qquad \max_{\mathbf{w}\in\mathcal{S}_M,\xi} \quad & \xi \\ \text{s.t.} \quad & \mathbf{D}\mathbf{p} \leq \gamma\mathbf{1}, & \text{s.t.} \quad & \mathbf{D}^\top\mathbf{w} \geq \xi\mathbf{1}. \end{aligned}$$

The primal and dual mix assignments and models, respectively. For optimal $\mathbf{p}$ and $\mathbf{w}$, define

$$\mathbf{Q} := \sum_{\ell=1}^{J} p_\ell \mathbf{z}_\ell, \qquad \overline{\boldsymbol{\nu}} := \sum_{j=1}^{M} w_j \boldsymbol{\nu}_j.$$

The averaged model $\overline{\boldsymbol{\nu}}$ is an adversarial strategy, not an estimate. Its assignment best response and lower certificate are

$$\mathbf{z}' \in \operatorname*{argmax}_{\mathbf{z}\in\mathcal{Z}(I)} V_{\overline{\boldsymbol{\nu}}}(I,\mathbf{z}), \qquad L := \sum_{j=1}^{M} w_j h_j - V^*_{\overline{\boldsymbol{\nu}}}(I).$$

The model best response and upper certificate are

$$(\boldsymbol{\nu}',\mathbf{z}'') \in \operatorname*{argmax}_{(\boldsymbol{\nu},\mathbf{z})\in\mathcal{C}_n\times\mathcal{Z}(I)} [\mathbf{\Phi}(I,\mathbf{z}) - \mathbf{\Phi}(I,\mathbf{Q})]^\top\boldsymbol{\nu}, \qquad U := [\mathbf{\Phi}(I,\mathbf{z}'') - \mathbf{\Phi}(I,\mathbf{Q})]^\top\boldsymbol{\nu}' = g_{n,I}(\mathbf{Q}).$$

Here $\mathbf{Q}$ and $\overline{\boldsymbol{\nu}}$ are the current learner and adversarial mixtures. The assignment response $\mathbf{z}'$ tests the current assignment pool against $\overline{\boldsymbol{\nu}}$ and yields the lower certificate L, while $\boldsymbol{\nu}'$ finds the exact worst-case regret $U = g_{n,I}(\mathbf{Q})$, with $\mathbf{z}''$ only its optimal comparator. The responses $\mathbf{z}'$ and $\boldsymbol{\nu}'$ are added to their pools. This double-oracle construction (McMahan et al., 2003; Gilbert & Spanjaard, 2017) is related to randomized minmax-regret optimization (Mastin et al., 2015).

Algorithm 1 Minimax-Regret Allocation (MRA)

Require: $\Theta, \boldsymbol{\phi}, B,$ logs $\mathcal{D}_n$ with $n \geq 1$, batch I, $\epsilon > 0$, $\delta \in (0,1)$
Ensure: A feasible marginal $\mathbf{Q} \in \mathcal{P}(I)$ together with a finite mixture representation $\mathbf{Q} = \sum_{\ell=1}^{J} p_\ell \mathbf{z}_\ell$

Compute $\widehat{\boldsymbol{\theta}}_n$ and $\mathcal{C}_n$, and set $\zeta_n \leftarrow \epsilon/(n+1)$
$\mathbf{z}_0 \leftarrow \operatorname{argmax}_{\mathbf{z}\in\mathcal{Z}(I)} \mathbf{\Phi}(I,\mathbf{z})^\top \widehat{\boldsymbol{\theta}}_n$
if $\epsilon \geq B$ **or** $\mathcal{C}_n = \varnothing$ **then**
    return $\mathbf{z}_0$ with representation $\{(1,\mathbf{z}_0)\}$
$\mathcal{G} \leftarrow \{\mathbf{z}_0\}, \quad \mathcal{H} \leftarrow \{\boldsymbol{\nu}_1\}$ for fixed $\boldsymbol{\nu}_1 \in \mathcal{C}_n$
loop
    Form $\mathbf{D}$ from $\mathcal{G},\mathcal{H}$ and solve the restricted primal–dual game for $\mathbf{p},\mathbf{w}$
    $\mathbf{Q} \leftarrow \sum_{\ell=1}^{J} p_\ell \mathbf{z}_\ell, \quad \overline{\boldsymbol{\nu}} \leftarrow \sum_{j=1}^{M} w_j \boldsymbol{\nu}_j$
    $\mathbf{z}' \leftarrow \operatorname{argmax}_{\mathbf{z}\in\mathcal{Z}(I)} V_{\overline{\boldsymbol{\nu}}}(I,\mathbf{z}), \quad L \leftarrow \sum_{j=1}^{M} w_j h_j - V^*_{\overline{\boldsymbol{\nu}}}(I)$
    $\boldsymbol{\nu}' \leftarrow \operatorname{argmax}_{\boldsymbol{\nu}\in\mathcal{C}_n} \Delta_{\boldsymbol{\nu}}(I,\mathbf{Q}), \quad U \leftarrow \Delta_{\boldsymbol{\nu}'}(I,\mathbf{Q})$
    if $U \leq \epsilon$ **or** $U - L \leq \zeta_n$ **then**
        return $\mathbf{Q}$ with representation $\{(p_\ell,\mathbf{z}_\ell)\}_{\ell=1}^{J}$
    $\mathcal{G} \leftarrow \mathcal{G} \cup \{\mathbf{z}'\}, \quad \mathcal{H} \leftarrow \mathcal{H} \cup \{\boldsymbol{\nu}'\}$

Algorithm 1 assumes completed best-response optimization. If a required response cannot be certified under an external computation limit, it returns the plug-in allocation $\mathbf{z}_0$.

Proposition 3 (Game certificates). *At every exact iteration of Algorithm 1,*

$$L \le v_{n,I} \le g_{n,I}(\mathbf{Q}) = U.$$

Hence $U \le \epsilon$ certifies that $\mathbf{Q}$ is ϵ-good over $\mathcal{C}_n$, while $U - L \le \zeta_n$ implies $g_{n,I}(\mathbf{Q}) \le v_{n,I} + \zeta_n$. For every nonempty compact $\mathcal{C}_n$ and $\zeta_n > 0$, completed exact best-response execution terminates in finitely many iterations.

The proof and a sharper polyhedral iteration bound are in Appendix A.2.

6 Statistical guarantees and information limits

We first control uncertainty under the realized source design and relate it to the population source geometry. We then derive information lower bounds from models with no common good allocation.

Upper bounds. The confidence event below is shared across all batches and all data-dependent model responses. No union bound over the batch domain or the optimization path is required.

Lemma 4 (Simultaneous confidence event). *For every admissible environment,*

$$\mathbb{P}\{\boldsymbol{\theta} \notin \mathcal{C}_n\} \le \delta/2.$$

On the event $\{\boldsymbol{\theta} \in \mathcal{C}_n\}$, every $\boldsymbol{\nu} \in \mathcal{C}_n$ satisfies

$$\|\mathbf{F}_n(\boldsymbol{\nu} - \boldsymbol{\theta})\|_2 \le 2\beta_n\sqrt{r_n}. \tag{7}$$

For the realized source design, define

$$a_n(\boldsymbol{\theta}) := \sup_{I \in \mathcal{I}} v_I(\{\boldsymbol{\nu} \in \Theta : \|\mathbf{F}_n(\boldsymbol{\nu} - \boldsymbol{\theta})\|_2 \le 2\beta_n\sqrt{r_n}\}).$$

Thus $a_n(\boldsymbol{\theta})$ is the worst robust allocation loss over this empirical neighborhood, uniformly over $I \in \mathcal{I}$. On the event of Lemma 4, $v_{n,I} \le a_n(\boldsymbol{\theta})$ for all $I \in \mathcal{I}$.

Theorem 5 (Worst-batch PAC control). *For every admissible environment, Algorithm 1 with completed best-response optimization satisfies*

$$\mathbb{P}\{R_{\boldsymbol{\theta}}(\widehat{\pi}_n) > \epsilon\} \le \frac{\delta}{2} + \mathbb{P}\{a_n(\boldsymbol{\theta}) + \zeta_n > \epsilon\}. \tag{8}$$

Under an external computation limit, (8) gains the computation-failure probability described after Algorithm 1. Proofs are in Appendices A.3 and A.4. A relative random-design bound next converts (7) into population source distance without overlap or a minimum-eigenvalue condition.

Corollary 6 (Population-source sufficient sample size). *For a universal constant C and $r_\lambda > 0$, let*

$$t_n := \min\left\{1, \sqrt{\frac{C r_\lambda \log(4r_\lambda/\delta)}{n}}\right\},$$

and set $t_n := 0$ if $r_\lambda = 0$. Under completed best-response optimization,

$$B\omega_{\boldsymbol{\theta}}(t_n) + \zeta_n \le \epsilon \quad \Longrightarrow \quad \mathbb{P}\{R_{\boldsymbol{\theta}}(\widehat{\pi}_n) > \epsilon\} \le \delta.$$

For fixed λ, replacing $\omega_{\boldsymbol{\theta}}$ by ω makes the guarantee uniform over Θ. If $r_\lambda > 0$ and $B\omega(t) \le \epsilon - \xi$ for some $t, \xi > 0$, it suffices that

$$n \ge C\frac{r_\lambda \log(4r_\lambda/\delta)}{t^2}, \qquad n + 1 \ge \frac{\epsilon}{\xi}.$$

The learner need not know t, ω, or λ. Proofs are in Appendices A.5 and A.6.

Information lower bounds. Let $\mathcal{K}_\epsilon(\boldsymbol{\theta}) := \{\pi : R_{\boldsymbol{\theta}}(\pi) \le \epsilon\}$ be the set of ϵ-good policies at $\boldsymbol{\theta}$. For scaled-Bernoulli rewards under a common logger, define

$$D_\lambda(\boldsymbol{\theta}, \boldsymbol{\nu}) := \mathbb{E}_\lambda\left[\ell\left(\frac{\boldsymbol{\phi}^\top \boldsymbol{\theta}}{B}, \frac{\boldsymbol{\phi}^\top \boldsymbol{\nu}}{B}\right)\right], \qquad \ell(p, q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q},$$

with the usual extended-value conventions.

Proposition 7 (Information lower bound and identification limit)**.** *If $\mathcal{K}_\epsilon(\boldsymbol{\theta}) \cap \mathcal{K}_\epsilon(\boldsymbol{\nu}) = \varnothing$, then every learner with failure probability at most $\delta < 1/2$ under both models satisfies*

$$n D_\lambda(\boldsymbol{\theta}, \boldsymbol{\nu}) \ge (1 - 2\delta) \log \frac{1-\delta}{\delta}. \tag{9}$$

For a fixed source law λ, let $h := 1 + \sup_{I \in \mathcal{I}} p_I \le d + 1$. For every $\delta < 1/h$, no learner is uniformly δ-PAC at any $\epsilon < B\omega(0)$, whereas Algorithm 1 with completed best-response optimization is uniformly δ-PAC for all sufficiently large n at every $\epsilon > B\omega(0)$.

Thus $B\omega(0)$ is the exact infinite-data identification threshold for a fixed source law. The finite-witness geometry of Proposition 2 gives a corresponding finite-sample lower bound.

Proposition 8 (Finite-witness information lower bound)**.** *Fix $I \in \mathcal{I}$, $\epsilon \ge 0$, and $\boldsymbol{\theta}_0, \boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_k \in \Theta$ with $k \ge 2$. If $v_I(\{\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_k\}) > \epsilon$, every learner with failure probability at most $\delta < 1/k$ at each witness model satisfies*

$$n \max_{j \in [k]} D_\lambda(\boldsymbol{\theta}_0, \boldsymbol{\nu}_j) \ge \ell(1/k, \delta).$$

No success guarantee at $\boldsymbol{\theta}_0$ is required.

By Proposition 2, a witness with $k \le p_I + 1$ exists whenever the robust loss exceeds ϵ. If, for some $0 < \alpha \le 1/2$, $\boldsymbol{\phi}(X, A)^\top \boldsymbol{\nu}_j \in [\alpha B, (1-\alpha)B]$ holds λ-almost surely wherever $\boldsymbol{\phi}(X, A)^\top (\boldsymbol{\theta}_0 - \boldsymbol{\nu}_j) \ne 0$, then Appendix A.8 gives

$$D_\lambda(\boldsymbol{\theta}_0, \boldsymbol{\nu}_j) \le \frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\nu}_j\|^2_{\boldsymbol{\Sigma}_\lambda}}{\alpha(1-\alpha)B^2}.$$

Thus an incompatible witness within source radius Bt yields the lower scale $t^{-2} \log(1/\delta)$ for bounded witness size and interior constant. With Corollary 6, this matches the upper scale up to constants for bounded r_λ and comparable sufficient and incompatible radii. See Appendix A.10.

7 Sharpness and information separations

The general results leave open whether mixing can reduce the information needed for allocation and whether the source-radius bounds are sharp. Two fixed-dimensional instances in the original model answer both questions. Full constructions and proofs are in Appendix B.

7.1 A rare-observation factor avoided by mixing

Consider a batch $I = (X_1, X_2, X_3, \mathbf{b})$ with unit capacities and $\mathcal{A}(X_1) = \{1, 2\}$, $\mathcal{A}(X_2) = \{1, 2, 3\}$, and $\mathcal{A}(X_3) = \{2, 3\}$. Its three integral assignments are the identity and transpositions $(1, 2)$ and $(2, 3)$, denoted $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$. For $\mathbf{p} \in \mathcal{S}_3$, let $\mathbf{Q}(\mathbf{p}) := \sum_{j=1}^3 p_j \mathbf{z}_j$. Appendix B.1 constructs a shared-linear instance parameterized by $\mathbf{v} \ge 0$, $\mathbf{1}^\top \mathbf{v} \le B$, such that

$$\Delta_{\mathbf{v}}(I, \mathbf{Q}(\mathbf{p})) = \frac{1}{3}\left(\max_{j \in [3]} v_j - \mathbf{p}^\top \mathbf{v}\right).$$

Its three source types have conditional means $(\mu_1, \mu_2, \mu_3) = (v_1 + v_3/2, v_2 + v_3/2, v_3)$. Let λ_j denote their logging probabilities. For $0 \le \rho \le 1/2$, let $\mathcal{E}_\rho$ contain all bounded conditional reward laws and loggers with $\lambda_1, \lambda_2 \ge 1/4$ and $\lambda_3 \ge \rho$. The learner does not know ρ.

Theorem 9 (Mixed and integral sample complexity)**.** *For $0 < \rho \le 1/2$, $\epsilon_0 = B/8$, and $0 < \delta \le 1/16$,*

$$N^*(\epsilon_0, \delta; \mathcal{E}_\rho) \asymp \log(1/\delta), \qquad N^{(1)}(\epsilon_0, \delta; \mathcal{E}_\rho) \asymp \rho^{-1} \log(1/\delta). \tag{10}$$

Here $N^{(1)}$ restricts to integral marginals but allows outer randomization. Completed Algorithm 1 attains the first rate, while the integral plug-in rule $\mathbf{z}_0$ (OLS–flow) attains the second. At $\rho = 0$, the mixed upper persists, while every integral-output learner has worst-case failure at least $1/2$ for all n.

The key cancellation $d := v_1 - v_2 = \mu_1 - \mu_2$ makes the mixed decision learnable from source types 1 and 2 without identifying v_3, while the optimal integral matching can depend on v_3. In the integral lower bound, the frequent source means agree across instances and only the rare type reveals v_3, producing the rare-observation factor. A direct functional–mixing rule also attains the mixed order, so Theorem 9 separates the information required by mixed and integral outputs rather than asserting universal dominance of MRA. Appendix B.3 further proves $B\omega(t) \le B/9 + (4\sqrt{2}/3)Bt$, independently of the rare-type probability, so the modulus bound recovers the mixed rate.

7.2 A COUPLED POSITIVE FLOOR

Appendix B.4 constructs a fixed-dimensional family $\mathcal{E}_{H,\rho}$ with $\boldsymbol{\theta} = (c, u, v_1, v_2)^\top$, where $0 \le u \le c \le B$, $v_1, v_2 \ge 0$, $v_1 + v_2 \le c$, and $v_3 := c - v_1 - v_2$. The source observes means c and u with probabilities $1-\rho$ and ρ, while (v_1, v_2, v_3) remain latent and determine allocation scores. For each $1 \le L \le H$, the deployment domain has an L-fold lift with the same normalized three-point image, so H changes batch size but not reduced geometry. Even with exact (c, u), the latent allocation coordinates vary within the source fiber, so no marginal is uniformly optimal over it. The source-fiber minimax loss is at most $B/12$, with equality at $(c, u) = (B, B/2)$, hence $B\omega(0) = B/12$.

Theorem 10 (Coupled sharpness)**.** *There are universal $c_0, C_1, C_2 > 0$ such that, for every $H \ge 1$, $0 < \rho \le 1/2$, $0 < \delta \le 1/32$, and $0 < \tau \le c_0 B$,*

$$C_1 \frac{B^2}{\rho\tau^2} \log \frac{1}{\delta} \le N^*(B/12 + \tau, \delta; \mathcal{E}_{H,\rho}) \le C_2 \frac{B^2}{\rho\tau^2} \log \frac{1}{\delta}. \tag{11}$$

Algorithm 1 with exact responses attains the upper bound.

Every pair of models has a common policy with regret at most $5B/64 < B/12$, yet Appendix B.6 constructs three models with

$$v_I(\{\boldsymbol{\nu}_1(h), \boldsymbol{\nu}_2(h), \boldsymbol{\nu}_3(h)\}) \ge \frac{B}{12} + \frac{h}{36}.$$

Since $p_I = 2$ for every declared batch, the $p_I + 1 = 3$ bound in Proposition 2 is tight.

For $\boldsymbol{\theta}_0 = (B, B/2, B/3, B/3)^\top$ and $0 < t \le \sqrt{\rho}/4$, Appendix B.7 shows $B\omega_{\boldsymbol{\theta}_0}(t) - B/12 \asymp B\omega(t) - B/12 \asymp Bt/\sqrt{\rho}$. Thus τ corresponds to $t \asymp \tau\sqrt{\rho}/B$, and Corollary 6 with Proposition 8 recovers the rate in Theorem 10. Pairwise disjoint-good-policy bounds do not apply.

8 EXPERIMENTS

We evaluate the normalized regret in (4) at the returned marginal $\mathbf{Q}$, with $B = 1$, and define failure by the strict inequality $\Delta > \epsilon$. Comparison objectives are specified in Appendix D, while experimental protocols and uncertainty calculations are given in Appendix E.

Theory-aligned synthetic experiments. On the four Rare witnesses, Figure 1 reports sustained empirical 0.05 crossings using 500 logs per model and grid point with $\epsilon = 1/8$. MRA crosses at $n = 64$ for $\rho \le 0.1$ on the tested grid and at 32 for $\rho = 0.25$. OLS–flow and constrained-flow instead cross at $1{,}024$ when $\rho = 0.01$ and do not cross by $32{,}768$ when $\rho = 0$. The direct mixture crosses at 64 throughout, illustrating the mixed-versus-integral separation in Theorem 9. For MRA on the Coupled family with $\epsilon = 1/12 + \tau$, Figure 3 shows that all twelve (ρ, τ) settings share the tested-grid transition bracket $(10^{-3}, 10^{-2.5}]$ on the $n\rho\tau^2$ scale, consistent with Theorem 10.

Source-fiber stress test. We use LLMRouterBench contexts (Li et al., 2026) with synthetic linear Bernoulli rewards rather than observed benchmark scores. Each batch has five queries and five unit-capacity actions, one of which is never logged. We evaluate 20 independent logs for each $n \in \{2^{15}, 2^{16}, 2^{17}\}$ in four fixed environments. A log fails when its maximum regret over 32 held-out batches and six evaluator-only source-indistinguishable completions exceeds 0.1. We compare MRA with OLS–flow, constrained-flow, maximin, and reference pessimism. The latter two also allow mixtures and use the same $\mathcal{C}_n$. Figure 2 reports the prespecified primary Environment 1. At $n = 2^{17}$, MRA fails on $0/20$ logs in Environments 1–3 and $2/20$ in Environment 4, while every comparator fails on $20/20$ logs in every environment. In all 48 comparator–MRA cells, the comparator has higher mean log-level maximum regret by 0.0450–0.0730. Full environment-specific results are in Appendix E.5. These finite evaluations do not establish uniform PAC performance.

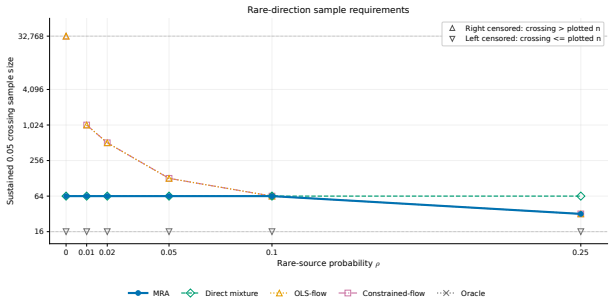


Figure 1: **Rare-direction sample requirements.** First tested sample size whose worst-model empirical failure remains at most 0.05 at every larger grid point. Triangles denote censored bounds beyond 32,768 or at or below 16.

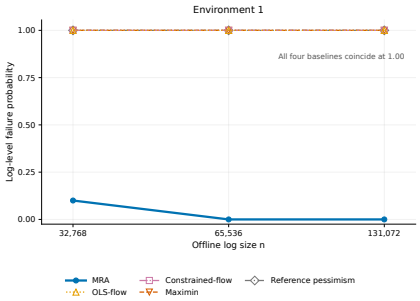


Figure 2: **Source-fiber failure.** Environment 1 with 20 independent logs per sample size. Full environment-specific results are in Appendix E.5.

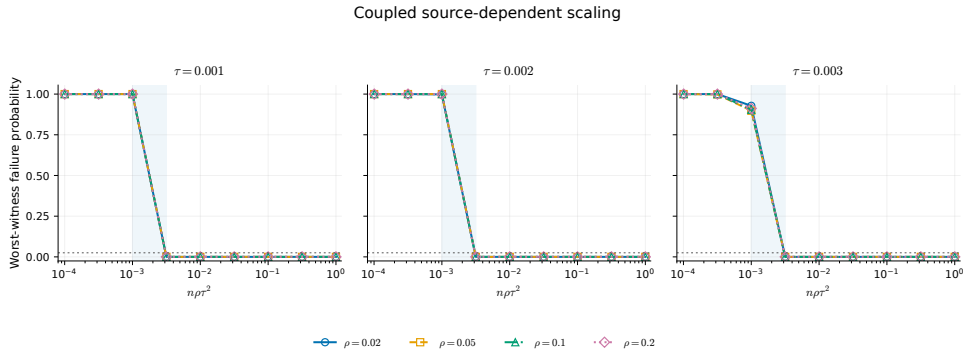


Figure 3: **Coupled source-dependent scaling.** Maximum empirical MRA failure over three witness models, using 500 logs per model and grid point. For all twelve (ρ, τ) settings, failure crosses $\delta = 0.025$ between the adjacent tested points 10^{-3} and $10^{-2.5}$ on the $n\rho\tau^2$ scale. The shaded region marks this interval, not a confidence interval.

9 Conclusion

We studied off-policy learning for allocations under capacity constraints when rewards follow a shared linear model and the logger is unknown and may assign zero probability to available actions. The source fiber formulation characterizes the allocation ambiguity that remains even with unlimited source data. Building on this characterization, MRA provides guarantees uniformly over deployment batches without requiring overlap or a lower bound on the eigenvalues of the source covariance. Our sharp examples show that allowing marginal rather than integral outputs can fundamentally reduce the information required for learning. They also quantify how limited source coverage increases the sample size needed to achieve accuracy above the ambiguity floor. The experiments recover these predicted phenomena and illustrate MRA in a controlled study using real query contexts.

More broadly, our results suggest that the relevant statistical target in off-policy allocation is not full recovery of the reward parameter, but the distinctions among source-compatible models that can change the optimal allocation. The admissible output class is part of this statistical problem because marginal and integral decisions can require different information.

A natural next step is to determine when this source geometry permits substantially cheaper computational certificates in large allocation problems. Another direction is sequential allocation under capacity constraints, where requests arrive over time and current decisions affect the feasible choices available later. Extending the source fiber perspective to such settings could clarify how ambiguity in off-policy data interacts with capacity depletion and the opportunity cost of using capacity early.

AI USE STATEMENT

Generative AI tools were used to assist with polishing the writing, coding, and reviewing mathematical proofs. The authors independently verified the final text, code, and mathematical arguments and take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

All assumptions and theoretical statements are specified in the main text, with complete proofs and sharp-instance constructions in Appendices A and B. Appendix C gives the computational formulations used by MRA, Appendix D specifies the comparison objectives, and Appendix E provides the experimental protocols, parameter settings, evaluation rules, uncertainty calculations, and additional results.

REFERENCES

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in neural information processing systems*, 29, 2016.

Hassene Aissi, Cristina Bazgan, and Daniel Vanderpooten. Min–max and min–max regret versions of combinatorial optimization problems: A survey. *European journal of operational research*, 197 (2):427–438, 2009.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4), August 2005. ISSN 0090-5364. doi: 10.1214/009053605000000282. URL http://dx.doi.org/10.1214/009053605000000282.

Eli Ben-Michael, Kosuke Imai, and Zhichao Jiang. Policy learning with asymmetric counterfactual utilities. *Journal of the American Statistical Association*, 119(548):3045–3058, 2024.

Omar Bennouna, Amine Bennouna, Saurabh Amin, and Asuman Ozdaglar. What data enables optimal decisions? an exact characterization for linear optimization. *Advances in Neural Information Processing Systems*, 38, 2025.

Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. On bayes risk lower bounds. *Journal of Machine Learning Research*, 17(218):1–58, 2016.

Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

Erick Delage, Daniel Kuhn, and Wolfram Wiesemann. "Dice"-sion–making under uncertainty: when can a random decision reduce risk? *Management Science*, 65(7):3282–3301, 2019.

Priya Donti, Brandon Amos, and J Zico Kolter. Task-based end-to-end model learning in stochastic optimization. *Advances in neural information processing systems*, 30, 2017.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011.

Tanner Fiez, Lalit Jain, Kevin Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

Hugo Gilbert and Olivier Spanjaard. A double oracle approach to minmax regret optimization problems with interval data. *European Journal of Operational Research*, 262(3):929–943, 2017.

Konstantin Hess, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Efficient and sharp off-policy learning under unobserved confounding. In *International Conference on Learning Representations*, 2026.

Samuel D Higbee. Policy learning with new treatments. *Quantitative Economics*, 16(4):1409–1456, 2025.

Yichun Hu, Nathan Kallus, Xiaojie Mao, and Yanchen Wu. Contextual linear optimization with bandit feedback. *Advances in Neural Information Processing Systems*, 37:26976–27004, 2024.

Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. Policy learning "without" overlap: Pessimism and generalized empirical bernstein's inequality. *The Annals of Statistics*, 53(4): 1483–1512, 2025. doi: 10.1214/25-AOS2511.

Nathan Kallus. Balanced policy evaluation and learning. *Advances in neural information processing systems*, 31, 2018.

Fahad Kamran, Maggie Makar, and Jenna Wiens. Learning to rank for optimal treatment allocation under resource constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 3727–3735. PMLR, 2024.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: Sharp partial identification under smoothness. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23734–23757. PMLR, 2024.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019.

Gene Li, Cong Ma, and Nati Srebro. Pessimism for offline linear contextual bandits using ℓ_p confidence sets. *Advances in Neural Information Processing Systems*, 35:20974–20987, 2022.

Hao Li, Yiqun Zhang, Zhaoyan Guo, Chenxu Wang, Shengji Tang, Qiaosheng Zhang, Yang Chen, Biqing Qi, Peng Ye, Lei Bai, et al. Llmrouterbench: A massive benchmark and unified framework for llm routing. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37733–37754, 2026.

Xutong Liu, Xiangxiang Dai, Jinhang Zuo, Siwei Wang, Carlee Joe-Wong, John C. S. Lui, and Wei Chen. Offline learning for combinatorial multi-armed bandits. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38251–38289. PMLR, 2025.

Andrew Mastin, Patrick Jaillet, and Sang Chin. Randomized minmax regret for combinatorial optimization under uncertainty. In *International symposium on algorithms and computation*, pp. 491–501. Springer, 2015.

Jiří Matoušek. *Lectures on discrete geometry*, volume 212. Springer, 2002.

H Brendan McMahan, Geoffrey J Gordon, and Avrim Blum. Planning in the presence of cost functions controlled by an adversary. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 536–543, 2003.

Noveen Sachdeva, Yi Su, and Thorsten Joachims. Off-policy bandits with deficient support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 965–975, 2020.

Alexander Schrijver. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer Science & Business Media, 2003.

Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, and Andreas Krause. Mixed strategies for robust optimization of unknown objectives. In *International Conference on Artificial Intelligence and Statistics*, pp. 2970–2980. PMLR, 2020.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International conference on machine learning*, pp. 814–823. PMLR, 2015.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 1658–1665, 2019.

Kohei Yata. Optimal decision rules under partial identification. *arXiv preprint arXiv:2111.04926*, 2021.

Yuhan Ye, Saurabh Amin, and Asuman Özdağlar. Learning decision-sufficient representations for linear optimization. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 6924–6975. PMLR, 2026.

APPENDIX

A MISSING PROOFS

A.1 PROOF OF PROPOSITION 2

Proposition 2 (restated)**.** *Every feasible aggregate feature is realizable by a mixture of at most $p_I + 1$ integral assignments. If $v_I(\mathcal{H}) > \epsilon$, then some $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_k \in \mathcal{H}$, with $k \leq p_I + 1$, satisfy*

$$v_I(\{\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_k\}) > \epsilon.$$

Proof. The first claim is Carathéodory's theorem on the p_I-dimensional aggregate-feature polytope. For the second, each $\boldsymbol{\nu} \in \mathcal{H}$ defines a closed halfspace of aggregate features with regret at most ϵ. If their intersection with the aggregate-feature polytope is empty, compactness of the polytope yields a finite subfamily with empty intersection. Applying Helly's theorem in its p_I-dimensional affine hull reduces this witness to at most $p_I + 1$ models (Matoušek, 2002). □

A.2 PROOF OF PROPOSITION 3

Proposition 3 (restated)**.** *At every exact iteration of Algorithm 1,*

$$L \leq v_{n,I} \leq g_{n,I}(\mathbf{Q}) = U.$$

Hence $U - L \leq \zeta_n$ implies $g_{n,I}(\mathbf{Q}) \leq v_{n,I} + \zeta_n$. For every nonempty compact $\mathcal{C}_n$ and $\zeta_n > 0$, completed exact best-response execution terminates after finitely many iterations.

If $\mathcal{C}_n$ is a polytope and model responses are selected at candidate vertices, exact execution terminates in at most

$$|\mathcal{Z}(I)| + |\operatorname{vert}(\mathcal{C}_n)| + 1$$

iterations.

Proof. For every feasible $\mathbf{Q}'$,

$$\max_{\boldsymbol{\nu} \in \mathcal{C}_n} \Delta_{\boldsymbol{\nu}}(I, \mathbf{Q}') \geq \sum_j w_j \Delta_{\boldsymbol{\nu}_j}(I, \mathbf{Q}') = \sum_j w_j h_j - \boldsymbol{\Phi}(I, \mathbf{Q}')^\top \overline{\boldsymbol{\nu}}.$$

Minimizing over $\mathbf{Q}'$ gives $L \leq v_{n,I}$. Since the current $\mathbf{Q}$ is feasible and (5) gives its exact worst-case regret,

$$v_{n,I} \leq g_{n,I}(\mathbf{Q}) = U.$$

The gap claim follows immediately.

Let γ be the common optimal value of the restricted primal and dual programs. If $\mathbf{z}'$ is already in the assignment pool, its weighted regret attains the minimum against the optimal dual mixture, so $L = \gamma$. If $\boldsymbol{\nu}'$ is already in the model pool, primal feasibility gives $U \leq \gamma$. Hence if both responses are already stored, the preceding certificate yields

$$L = v_{n,I} = U = \gamma.$$

Thus every nonterminating iteration adds a new assignment or a new model. For polyhedral $\mathcal{C}_n$, the assignment set and the selected candidate vertices are finite, which gives the stated iteration bound.

For a general compact candidate, let

$$C_{\boldsymbol{\phi}} := \sup_{x,a} \|\boldsymbol{\phi}(x, a)\|_2.$$

If $C_{\boldsymbol{\phi}} = 0$, every regret is zero. Otherwise,

$$|\Delta_{\boldsymbol{\nu}}(I, \mathbf{Q}) - \Delta_{\boldsymbol{\vartheta}}(I, \mathbf{Q})| \le 2C_{\boldsymbol{\phi}} \|\boldsymbol{\nu} - \boldsymbol{\vartheta}\|_2.$$

Consider a nonterminating iteration in which $\mathbf{z}'$ is already stored. Then $L = \gamma$ and every stored model satisfies $\Delta_{\boldsymbol{\nu}_j}(I, \mathbf{Q}) \le \gamma$, whereas

$$\Delta_{\boldsymbol{\nu}'}(I, \mathbf{Q}) = U > \gamma + \zeta_n.$$

Therefore

$$\|\boldsymbol{\nu}' - \boldsymbol{\nu}_j\|_2 > \frac{\zeta_n}{2C_{\boldsymbol{\phi}}}$$

for every stored model $\boldsymbol{\nu}_j$. Compactness permits only finitely many such separated model additions. All other nonterminating iterations add a new assignment, and $\mathcal{Z}(I)$ is finite. Hence exact execution terminates after finitely many iterations. □

A.3 Proof of Lemma 4

Lemma 4 (restated)**.** *For every admissible environment,*

$$\mathbb{P}\{\boldsymbol{\theta} \notin \mathcal{C}_n\} \le \delta/2.$$

On the event $\{\boldsymbol{\theta} \in \mathcal{C}_n\}$, every $\boldsymbol{\nu} \in \mathcal{C}_n$ satisfies

$$\|\mathbf{F}_n(\boldsymbol{\nu} - \boldsymbol{\theta})\|_2 \le 2\beta_n \sqrt{r_n}.$$

Proof. Condition on the logged design and write $\boldsymbol{\xi} = \mathbf{y} - \mathbf{F}_n\boldsymbol{\theta}$. By Hoeffding's lemma, for every $\mathbf{a} \in \mathbb{R}^n$,

$$\mathbb{E}[\exp(\mathbf{a}^\top \boldsymbol{\xi}) \mid \mathbf{F}_n] \le \exp\left(\frac{B^2\|\mathbf{a}\|_2^2}{8}\right). \tag{12}$$

For $r_n > 0$, put

$$\mathbf{L}_n := \mathbf{E}_n^\top \mathbf{G}_n^{\dagger/2} \mathbf{F}_n^\top, \qquad \mathbf{L}_n \mathbf{L}_n^\top = \mathbf{I}_{r_n}.$$

Using $\mathbf{y} = \mathbf{F}_n\boldsymbol{\theta} + \boldsymbol{\xi}$ and the fact that $\mathbf{G}_n^{1/2}$ annihilates $\ker(\mathbf{F}_n)$,

$$\mathbf{E}_n^\top \mathbf{G}_n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \mathbf{E}_n^\top \mathbf{G}_n^{\dagger/2} \mathbf{F}_n^\top \boldsymbol{\xi} = \mathbf{L}_n \boldsymbol{\xi}.$$

Each row of $\mathbf{L}_n$ has unit Euclidean norm, so (12) and the definition of β_n give coordinate failure probability at most $\delta/(2r_n)$. A union bound proves the first claim.

On this event, every $\boldsymbol{\nu} \in \mathcal{C}_n$ satisfies

$$\|\mathbf{E}_n^\top \mathbf{G}_n^{1/2}(\boldsymbol{\nu} - \boldsymbol{\theta})\|_\infty \le 2\beta_n.$$

Since $\mathbf{G}_n^{1/2}(\boldsymbol{\nu} - \boldsymbol{\theta})$ lies in $\mathrm{range}(\mathbf{F}_n^\top)$,

$$\|\mathbf{F}_n(\boldsymbol{\nu} - \boldsymbol{\theta})\|_2 = \|\mathbf{E}_n^\top \mathbf{G}_n^{1/2}(\boldsymbol{\nu} - \boldsymbol{\theta})\|_2 \le 2\beta_n \sqrt{r_n}.$$

If $r_n = 0$, then $\mathbf{F}_n = 0$ and both claims hold deterministically. □

A.4 Proof of Theorem 5

Theorem 5 (restated)**.** *For every admissible environment, Algorithm 1 with completed best-response optimization satisfies*

$$\mathbb{P}\{R_{\boldsymbol{\theta}}(\widehat{\pi}_n) > \epsilon\} \le \frac{\delta}{2} + \mathbb{P}\{a_n(\boldsymbol{\theta}) + \zeta_n > \epsilon\}.$$

Under an external computation limit, let $\mathcal{U}_n$ be the event that a required certified computation is not completed for at least one batch in the declared domain. Then

$$\mathbb{P}\{R_{\boldsymbol{\theta}}(\widehat{\pi}_n) > \epsilon\} \le \frac{\delta}{2} + \mathbb{P}\{a_n(\boldsymbol{\theta}) + \zeta_n > \epsilon\} + \mathbb{P}(\mathcal{U}_n).$$

Proof. If $\epsilon \geq B$, the claim is immediate. Assume $\epsilon < B$. On the event of Lemma 4, $v_{n,I} \leq a_n(\boldsymbol{\theta})$ simultaneously for all $I \in \mathcal{I}$.

If MRA stops with $U \leq \epsilon$, then

$$\Delta_{\boldsymbol{\theta}}(I, \mathbf{Q}) \leq g_{n,I}(\mathbf{Q}) = U \leq \epsilon.$$

If it stops with $U - L \leq \zeta_n$, Proposition 3 gives

$$\Delta_{\boldsymbol{\theta}}(I, \mathbf{Q}) \leq g_{n,I}(\mathbf{Q}) = U \leq v_{n,I} + \zeta_n \leq a_n(\boldsymbol{\theta}) + \zeta_n.$$

Hence, on the coverage event, $a_n(\boldsymbol{\theta}) + \zeta_n \leq \epsilon$ ensures that every batch receives an ϵ-good marginal. Combining this with confidence failure proves the claim. Under a computation limit, the same argument applies outside $\mathcal{U}_n$. □

A.5 Relative source norm

Lemma 11 (Relative source norm)**.** *There is a universal constant C such that, with probability at least $1 - \eta$ over the logged design,*

$$\|\boldsymbol{\nu} - \boldsymbol{\vartheta}\|_{\boldsymbol{\Sigma}_\lambda}^2 \leq \max\left\{\frac{2}{n}\|\mathbf{F}_n(\boldsymbol{\nu} - \boldsymbol{\vartheta})\|_2^2,\, CB^2\frac{r_\lambda + \log(1/\eta)}{n}\right\} \tag{13}$$

simultaneously for all $\boldsymbol{\nu}, \boldsymbol{\vartheta} \in \Theta$.

Proof. Let P and P_n denote source expectation and empirical averaging, and define

$$\mathcal{F} := \left\{(x, a) \mapsto B^{-1}\boldsymbol{\phi}(x, a)^\top(\boldsymbol{\nu} - \boldsymbol{\vartheta}) : \boldsymbol{\nu}, \boldsymbol{\vartheta} \in \Theta\right\}.$$

The class is separable, symmetric, star-shaped, and bounded by one in absolute value. On a source-full-measure set it lies in an r_λ-dimensional subspace of $L_2(P)$. The claim is immediate when $r_\lambda = 0$, so assume $r_\lambda > 0$ and let $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{r_\lambda})^\top$ be an orthonormal basis of this subspace.

For $u \in (0, 1]$, let

$$\mathcal{F}_u := \{f \in \mathcal{F} : Pf^2 \leq u\}, \qquad \mathcal{H}_u := \{f^2 : f \in \mathcal{F}_u\}.$$

If σ_i are independent Rademacher signs, then

$$\mathbb{E}\sup_{f \in \mathcal{F}_u}\left|\frac{1}{n}\sum_i \sigma_i f(X_i, A_i)\right| \leq \sqrt{\frac{r_\lambda u}{n}}.$$

Indeed, the coefficient vector of $f \in \mathcal{F}_u$ in the basis $\boldsymbol{\psi}$ has norm at most $\sqrt{u}$, while

$$\mathbb{E}\left\|\sum_i \sigma_i \boldsymbol{\psi}(X_i, A_i)\right\|_2^2 = nr_\lambda.$$

For $g = f^2 \in \mathcal{H}_u$, we have $0 \leq g \leq 1$ and $Pg^2 = Pf^4 \leq Pf^2 \leq u$. Since squaring is 2-Lipschitz on $[-1, 1]$, the contraction theorem gives

$$\mathbb{E}\sup_{g \in \mathcal{H}_u}\frac{1}{n}\sum_i \sigma_i g(X_i, A_i) \leq 2\sqrt{\frac{r_\lambda u}{n}}.$$

Applying Theorem 2.1 of Bartlett et al. (2005) with $x = \log(1/\eta)$ yields, with probability at least $1 - \eta$,

$$\sup_{f \in \mathcal{F}_u}(Pf^2 - P_n f^2) \leq 8\sqrt{\frac{r_\lambda u}{n}} + \sqrt{\frac{2u\log(1/\eta)}{n}} + \frac{4\log(1/\eta)}{3n}. \tag{14}$$

Choose

$$u := C\frac{r_\lambda + \log(1/\eta)}{n}$$

with sufficiently large universal C. If $u \geq 1$, boundedness proves the claim. Otherwise the right side of (14) is at most $u/2$. For $f \in \mathcal{F}$ with $Pf^2 > u$, star-shapedness gives

$$\widetilde{f} := \sqrt{\frac{u}{Pf^2}}\, f \in \mathcal{F}_u, \qquad P\widetilde{f}^2 = u.$$

Hence $P_n\widetilde{f}^2 \geq u/2$, so $P_n f^2 \geq Pf^2/2$. If $Pf^2 \leq u$, the second term in the maximum applies. Multiplying by B^2 proves (13). □

A.6 Proof of Corollary 6

Corollary 6 (restated). *There is a universal constant C. For $r_\lambda > 0$, define*

$$t_n := \min\left\{1, \sqrt{\frac{C r_\lambda \log(4r_\lambda/\delta)}{n}}\right\},$$

and set $t_n := 0$ when $r_\lambda = 0$. If

$$B\omega_{\boldsymbol{\theta}}(t_n) + \zeta_n \le \epsilon,$$

then Algorithm 1 with completed best-response optimization satisfies

$$\mathbb{P}\{R_{\boldsymbol{\theta}}(\widehat{\pi}_n) > \epsilon\} \le \delta.$$

For a fixed source law λ, replacing $\omega_{\boldsymbol{\theta}}$ by ω makes the condition uniform over Θ and all admissible bounded reward laws. In particular, if $B\omega(t) \le \epsilon - \xi$ for some $t, \xi > 0$ and $r_\lambda > 0$, it is sufficient that

$$n \ge C\frac{r_\lambda \log(4r_\lambda/\delta)}{t^2}, \qquad n+1 \ge \frac{\epsilon}{\xi}.$$

Proof. Almost surely, $r_n \le r_\lambda$. Apply Lemma 11 with $\eta = \delta/2$. If $r_n > 0$, every $\boldsymbol{\nu}$ in the set defining $a_n(\boldsymbol{\theta})$ satisfies

$$\|\mathbf{F}_n(\boldsymbol{\nu} - \boldsymbol{\theta})\|_2^2 \le 4r_n\beta_n^2 = 2B^2 r_n \log\frac{4r_n}{\delta}.$$

If $r_n = 0$, the empirical norm is zero. After enlarging the universal constant if necessary, Lemma 11 therefore gives

$$\|\boldsymbol{\nu} - \boldsymbol{\theta}\|_{\boldsymbol{\Sigma}_\lambda}^2 \le CB^2\frac{r_\lambda \log(4r_\lambda/\delta)}{n}$$

for every parameter in the set defining $a_n(\boldsymbol{\theta})$ whenever $r_\lambda > 0$. Also $\|\boldsymbol{\nu} - \boldsymbol{\theta}\|_{\boldsymbol{\Sigma}_\lambda} \le B$ for all $\boldsymbol{\nu}, \boldsymbol{\theta} \in \Theta$, since $\boldsymbol{\phi}(X, A)^\top(\boldsymbol{\nu} - \boldsymbol{\theta}) \in [-B, B]$ almost surely. Hence

$$a_n(\boldsymbol{\theta}) \le B\omega_{\boldsymbol{\theta}}(t_n).$$

The same conclusion holds with $t_n = 0$ when $r_\lambda = 0$. Theorem 5, together with the additional design failure probability $\delta/2$, proves the first claim.

For the uniform statement, $\omega_{\boldsymbol{\theta}}(t) \le \omega(t)$. The first sample-size condition gives $t_n \le t$, while the second gives $\zeta_n = \epsilon/(n+1) \le \xi$. Thus

$$B\omega_{\boldsymbol{\theta}}(t_n) + \zeta_n \le B\omega(t) + \xi \le \epsilon.$$

□

A.7 Proof of Proposition 7

Proposition 7 (restated). *If $\mathcal{K}_\epsilon(\boldsymbol{\theta}) \cap \mathcal{K}_\epsilon(\boldsymbol{\nu}) = \varnothing$, then every learner with failure probability at most $\delta < 1/2$ under both models satisfies*

$$nD_\lambda(\boldsymbol{\theta}, \boldsymbol{\nu}) \ge (1 - 2\delta)\log\frac{1-\delta}{\delta}.$$

For a fixed source law λ, let

$$h := 1 + \sup_{I \in \mathcal{I}} p_I \le d + 1.$$

For every $\delta < 1/h$, no learner is uniformly δ-PAC at any $\epsilon < B\omega(0)$, whereas Algorithm 1 with completed best-response optimization is uniformly δ-PAC for all sufficiently large n at every $\epsilon > B\omega(0)$.

Proof. Let $P_{\boldsymbol{\theta}}$ and $P_{\boldsymbol{\nu}}$ denote the one-record laws under the common logger and scaled-Bernoulli rewards. If the output is ϵ-good for $\boldsymbol{\theta}$, decide $\boldsymbol{\theta}$, and otherwise decide $\boldsymbol{\nu}$. Disjoint good-policy sets make both testing errors at most δ. KL data processing and iid additivity give

$$nD_\lambda(\boldsymbol{\theta}, \boldsymbol{\nu}) \ge (1-\delta)\log\frac{1-\delta}{\delta} + \delta\log\frac{\delta}{1-\delta},$$

which proves (9). Independent learner randomization adds no information.

The same two-law testing inequality also gives

$$\max\left\{\mathbb{P}_{\boldsymbol{\theta}}\{R_{\boldsymbol{\theta}}(\widehat{\pi}_n) > \epsilon\}, \mathbb{P}_{\boldsymbol{\nu}}\{R_{\boldsymbol{\nu}}(\widehat{\pi}_n) > \epsilon\}\right\} \geq \frac{1}{4}\exp\{-nD_\lambda(\boldsymbol{\theta}, \boldsymbol{\nu})\}. \tag{15}$$

For general testing-based lower bounds, see Chen et al. (2016).

It remains to establish the identification threshold. We first show that

$$\lim_{t\downarrow 0} \omega(t) = \omega(0).$$

Let

$$L := \sup_{x,a} \|\boldsymbol{\phi}(x,a)\|_2 < \infty.$$

Uniformly over batches and feasible marginals,

$$|\Delta_{\boldsymbol{\theta}}(I, \mathbf{Q}) - \Delta_{\boldsymbol{\nu}}(I, \mathbf{Q})| \leq 2L\|\boldsymbol{\theta} - \boldsymbol{\nu}\|_2. \tag{16}$$

Take any $t_j \downarrow 0$ and centers $\boldsymbol{\theta}_j$ approaching the supremum defining $\omega(t_j)$. By compactness, along a subsequence $\boldsymbol{\theta}_j \to \boldsymbol{\theta}_*$. The maximum Euclidean distance from $\mathcal{N}_\lambda(\boldsymbol{\theta}_j, t_j)$ to $[\boldsymbol{\theta}_*]_\lambda$ tends to zero. Otherwise a separated convergent sequence would have a limit $\boldsymbol{\nu}_*$ with $\|\boldsymbol{\nu}_* - \boldsymbol{\theta}_*\|_{\boldsymbol{\Sigma}_\lambda} = 0$, a contradiction. Comparing each batch with a minimax marginal for $[\boldsymbol{\theta}_*]_\lambda$ and using (16) yields

$$\limsup_{j\to\infty} \omega(t_j) \leq \omega(0).$$

Monotonicity gives the reverse inequality, so $\omega(t) \to \omega(0)$ as $t \downarrow 0$.

If $\epsilon < B\omega(0)$, some source fiber and batch have no common ϵ-good marginal. Proposition 2 supplies $k \leq h$ fixed models witnessing this failure. Their source means agree, so choose identical scaled-Bernoulli source laws. Under the common data and learner-randomization law, every policy fails at one of these models. Hence one model has failure probability at least $1/k \geq 1/h$ for every n.

If $\epsilon > B\omega(0)$, right continuity gives $t > 0$ with $B\omega(t) < \epsilon$. For sufficiently large n, Corollary 6 has $t_n \leq t$ and ζ_n smaller than the remaining slack. This proves eventual uniform δ-PAC achievability for the fixed source law. □

A.8 Source likelihood comparison

For fixed λ, define

$$\chi^2_\lambda(\boldsymbol{\theta}, \boldsymbol{\nu}) := \mathbb{E}_\lambda\left[\frac{\{\boldsymbol{\phi}(X,A)^\top(\boldsymbol{\theta} - \boldsymbol{\nu})\}^2}{\{\boldsymbol{\phi}(X,A)^\top\boldsymbol{\nu}\}\{B - \boldsymbol{\phi}(X,A)^\top\boldsymbol{\nu}\}}\right]. \tag{17}$$

Lemma 12 (Source likelihood comparison)**.** *For every $\boldsymbol{\theta}, \boldsymbol{\nu} \in \Theta$,*

$$\frac{2}{B^2}\|\boldsymbol{\theta} - \boldsymbol{\nu}\|^2_{\boldsymbol{\Sigma}_\lambda} \leq D_\lambda(\boldsymbol{\theta}, \boldsymbol{\nu}) \leq \chi^2_\lambda(\boldsymbol{\theta}, \boldsymbol{\nu}). \tag{18}$$

If, for some $0 < \alpha \leq 1/2$,

$$\alpha B \leq \boldsymbol{\phi}(X,A)^\top\boldsymbol{\nu} \leq (1-\alpha)B$$

holds λ-almost surely wherever $\boldsymbol{\phi}(X,A)^\top(\boldsymbol{\theta} - \boldsymbol{\nu}) \neq 0$, then

$$\chi^2_\lambda(\boldsymbol{\theta}, \boldsymbol{\nu}) \leq \frac{\|\boldsymbol{\theta} - \boldsymbol{\nu}\|^2_{\boldsymbol{\Sigma}_\lambda}}{\alpha(1-\alpha)B^2}. \tag{19}$$

Proof. At a fixed source context–action pair, put

$$p := \frac{\boldsymbol{\phi}(X,A)^\top\boldsymbol{\theta}}{B}, \qquad q := \frac{\boldsymbol{\phi}(X,A)^\top\boldsymbol{\nu}}{B}.$$

For $0 < q < 1$,

$$\frac{p^2}{q} + \frac{(1-p)^2}{1-q} - 1 = \frac{(p-q)^2}{q(1-q)}.$$

Also, $\log x \le x - 1$ gives

$$\ell(p, q) \le \frac{(p-q)^2}{q(1-q)}.$$

For fixed $q \in (0, 1)$, the function $\ell(p, q) - 2(p-q)^2$ has value and first derivative zero at $p = q$ and second derivative $1/[p(1-p)] - 4 \ge 0$. Hence

$$2(p-q)^2 \le \ell(p, q) \le \frac{(p-q)^2}{q(1-q)}.$$

The boundary cases follow under the usual extended-value conventions. Integrating and using

$$\mathbb{E}_\lambda[(p-q)^2] = B^{-2}\|\boldsymbol{\theta} - \boldsymbol{\nu}\|^2_{\boldsymbol{\Sigma}_\lambda}$$

proves (18). Under the additional assumption, $q(1-q) \ge \alpha(1-\alpha)$ wherever $p \ne q$, which proves (19). □

A.9 Proof of Proposition 8

Proposition 8 (restated). *Fix $I \in \mathcal{I}$, $\epsilon \ge 0$, and $\boldsymbol{\theta}_0, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_k \in \Theta$ with $k \ge 2$. If*

$$v_I(\{\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_k\}) > \epsilon,$$

then every learner with failure probability at most $\delta < 1/k$ at each witness model satisfies

$$n \max_{j \in [k]} D_\lambda(\boldsymbol{\theta}_0, \boldsymbol{\nu}_j) \ge \ell(1/k, \delta).$$

No success guarantee at $\boldsymbol{\theta}_0$ is required.

Proof. Let $\mathbb{P}_0$ be the joint data and learner-randomization law under $P_{\boldsymbol{\theta}_0}$, and let $\mathbb{P}_j$ be the corresponding law under $P_{\boldsymbol{\nu}_j}$. Define

$$\mathcal{B}_j := \left\{\Delta_{\boldsymbol{\nu}_j}\left(I, \widehat{\pi}_n(I)\right) > \epsilon\right\}.$$

Because the witness has no common ϵ-good marginal,

$$\sum_{j=1}^{k} \mathbb{P}_0(\mathcal{B}_j) \ge 1.$$

Choose j with $\mathbb{P}_0(\mathcal{B}_j) \ge 1/k$. If the learner has failure at most δ under every witness model, then $\mathbb{P}_j(\mathcal{B}_j) \le \delta$. KL data processing and iid additivity give

$$nD_\lambda(\boldsymbol{\theta}_0, \boldsymbol{\nu}_j) \ge \ell(\mathbb{P}_0(\mathcal{B}_j), \mathbb{P}_j(\mathcal{B}_j)) \ge \ell(1/k, \delta),$$

which proves the claim. □

Corollary 13 (Finite-witness chi-square bound). *Under the conditions above, let*

$$\gamma := \max_{j \in [k]} \chi^2_\lambda(\boldsymbol{\theta}_0, \boldsymbol{\nu}_j) < \infty.$$

Then every learner satisfies

$$\max_{j \in [k]} \mathbb{P}_{\boldsymbol{\nu}_j}\left\{R_{\boldsymbol{\nu}_j}(\widehat{\pi}_n) > \epsilon\right\} \ge \frac{1}{k^2(1+\gamma)^n}. \tag{20}$$

Proof. For the same j with $\mathbb{P}_0(\mathcal{B}_j) \ge 1/k$, absolute continuity and independence give

$$\mathbb{E}_j\left[\left(\frac{dP_0}{dP_j}\right)^2\right] = \left(1 + \chi^2_\lambda(\boldsymbol{\theta}_0, \boldsymbol{\nu}_j)\right)^n.$$

Cauchy–Schwarz yields

$$\frac{1}{k^2} \le \mathbb{P}_0(\mathcal{B}_j)^2 \le \mathbb{P}_j(\mathcal{B}_j)(1+\gamma)^n.$$

Since B_j implies worst-batch failure under $\boldsymbol{\nu}_j$, (20) follows. □

A.10 SOURCE-RADIUS COMPARISON

Corollary 14 (Source-radius comparison)**.** *Fix a source law λ with $r_\lambda > 0$ and $0 < \epsilon < B$. Suppose that for some $t_1, t_2, \xi, c > 0$, $\boldsymbol{\theta}_0 \in \Theta$, $I \in \mathcal{I}$, and $2 \le k \le p_I + 1$,*

$$B\omega(t_1) \le \epsilon - \xi, \qquad \boldsymbol{\nu}_j \in \mathcal{N}_\lambda(\boldsymbol{\theta}_0, t_2), \qquad v_I(\{\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_k\}) > \epsilon,$$

and

$$\max_{j \in [k]} \chi^2_\lambda(\boldsymbol{\theta}_0, \boldsymbol{\nu}_j) \le c t_2^2.$$

Then any learner that is δ-PAC at the witness models requires

$$n \ge \frac{\log(1/(k^2\delta))}{c t_2^2}, \tag{21}$$

whereas Algorithm 1 is uniformly δ-PAC for the fixed source law if

$$n \ge C \frac{r_\lambda \log(4 r_\lambda/\delta)}{t_1^2}, \qquad n + 1 \ge \frac{\epsilon}{\xi}. \tag{22}$$

Proof. Corollary 13 gives

$$\delta \ge \frac{1}{k^2(1 + c t_2^2)^n}.$$

Hence, using $\log(1 + x) \le x$,

$$n \ge \frac{\log(1/(k^2\delta))}{c t_2^2}.$$

For the upper bound, Corollary 6 applies once

$$n \ge C \frac{r_\lambda \log(4 r_\lambda/\delta)}{t_1^2}.$$

The second condition gives $\zeta_n = \epsilon/(n+1) \le \xi$, so

$$B\omega(t_1) + \zeta_n \le \epsilon.$$

This proves the stated PAC upper bound.

If $B\omega_{\boldsymbol{\theta}_0}(t_2) > \epsilon$, Proposition 2 supplies an incompatible witness of size at most $p_I + 1$. Lemma 12 gives $c = 1/[\alpha(1-\alpha)]$ whenever the selected witness means satisfy its interior condition. For bounded r_λ, k, c, and t_2/t_1, the lower and upper statistical terms have the same order as $\delta \downarrow 0$. □

B SHARP-INSTANCE CONSTRUCTIONS AND PROOFS

B.1 RARE-DIRECTION CONSTRUCTION

Definition 15 (Rare-direction allocation experiment)**.** Use the eligibility matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \tag{23}$$

with unit capacities. Its three matchings are the identity and the transpositions $(1,2)$ and $(2,3)$, denoted in this order by $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$. For $\mathbf{p} \in \mathcal{S}_3$, let

$$\mathbf{Q}(\mathbf{p}) := \sum_{j=1}^{3} p_j \mathbf{z}_j.$$

Take

$$\Theta = \{\mathbf{v} \in \mathbb{R}^3 : \mathbf{v} \ge 0,\ \mathbf{1}^\top \mathbf{v} \le B\}. \tag{24}$$

For the deployment batch, the eligible context–action mean rewards are the corresponding entries of

$$\begin{pmatrix} (v_1+v_2+v_3)/2 & v_2+v_3/2 & 0 \\ v_2+v_3/2 & (v_1+v_2+v_3)/2 & v_2/2+v_3 \\ 0 & v_2/2+v_3 & (v_1+v_2+v_3)/2 \end{pmatrix}. \tag{25}$$

The two displayed zeros lie outside eligibility and are not used.

The source law has three context–action types with feature vectors $(1,0,1/2)^\top$, $(0,1,1/2)^\top$, and $(0,0,1)^\top$. Their conditional means are

$$\mu_1 = v_1 + v_3/2, \qquad \mu_2 = v_2 + v_3/2, \qquad \mu_3 = v_3. \tag{26}$$

Let $\mathcal{E}_\rho$ contain this full model, all bounded conditional laws, and all logger probabilities satisfying $\lambda_1, \lambda_2 \geq 1/4$ and $\lambda_3 \geq \rho$, where $0 \leq \rho \leq 1/2$. The learner is not given ρ.

To make (24) coincide with the global parameter set in (1), the feature map also contains $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{1}$, outside the source support and the declared deployment batch.

A direct calculation gives

$$\mathbf{\Phi}(I, \mathbf{Q}(\mathbf{p})) = \begin{pmatrix} 1/6 \\ 1/2 \\ 1/2 \end{pmatrix} + \frac{1}{3}\mathbf{p}, \qquad \Delta_{\mathbf{v}}(I, \mathbf{Q}(\mathbf{p})) = \frac{1}{3}\left(\max_{j\in[3]} v_j - \mathbf{p}^\top \mathbf{v}\right). \tag{27}$$

Lemma 16 (A comparison mixture on the source fiber)**.** *For $t \in [-B, B]$, let $j = 1$ if $t \geq 0$ and $j = 2$ otherwise, and put $r = |t|$. If $r \geq B/2$, let $p_j(t) = 1$. Otherwise set*

$$p_j(t) = \frac{B+r}{3(B-r)}, \qquad p_3(t) = 1 - p_j(t), \qquad p_{3-j}(t) = 0. \tag{28}$$

For every original $\mathbf{v}$, with $d = v_1 - v_2$,

$$\Delta_{\mathbf{v}}(I, \mathbf{Q}(\mathbf{p}(t))) \leq \frac{B}{9} + \frac{2}{3}|d - t|. \tag{29}$$

Proof. First consider the face $\mathbf{1}^\top \mathbf{v} = B$ and let $d \geq 0$. The fixed-d slice has endpoints

$$((B+d)/2, (B-d)/2, 0), \qquad (d, 0, B-d).$$

Matching 2 is dominated. Mixing matching 1 with probability p and matching 3 otherwise gives worst regret

$$\frac{1}{3}\max\{(B+d)(1-p)/2, (B-2d)p\}.$$

For $d < B/2$, equalizing the terms gives (28) and

$$\frac{(B+d)(B-2d)}{9(B-d)} = \frac{B}{9} - \frac{2d^2}{9(B-d)} \leq B/9.$$

For $d \geq B/2$, matching 1 is optimal on the whole slice. Negative d exchanges indices 1 and 2.

For an arbitrary $\mathbf{v} \in \Theta$, adding

$$\frac{B - \mathbf{1}^\top \mathbf{v}}{3}\mathbf{1}$$

preserves d and regret and maps $\mathbf{v}$ to the face $\mathbf{1}^\top \mathbf{v} = B$. The two endpoints of the corresponding fixed-d slice, expressed with the positive and negative parts of d, are respectively 1- and 2-Lipschitz in ℓ_1. Matching convex coefficients therefore takes a point in a d-slice to a point in a t-slice at distance at most $2|d-t|$. For fixed $\mathbf{p}$, the function $\max_j(\mathbf{e}_j - \mathbf{p})^\top \mathbf{v}/3$ is $1/3$-Lipschitz in ℓ_1. This proves (29), including sign changes. □

B.2 Proof of Theorem 9

Theorem 9 (restated). *For $0 < \rho \le 1/2$, $\epsilon_0 = B/8$, and $0 < \delta \le 1/16$,*

$$N^*(\epsilon_0, \delta; \mathcal{E}_\rho) \asymp \log(1/\delta), \qquad N^{(1)}(\epsilon_0, \delta; \mathcal{E}_\rho) \asymp \rho^{-1} \log(1/\delta).$$

Here $N^{(1)}$ restricts outputs to integral marginals, allowing outer randomization. Completed Algorithm 1 attains the first order, while ordinary OLS–flow attains the second. At $\rho = 0$, the mixed upper remains valid, whereas every integral-output learner has worst-case failure at least $1/2$ for all n.

Proof. Upper bounds. Let N_j and $\overline{Y}_j$ be source counts and means. On $N_1, N_2 > 0$, put $\widehat{d} = \overline{Y}_1 - \overline{Y}_2$. The vector for the functional d is the difference of the first two source rows. Giving weight $1/N_1$ to type 1 and $-1/N_2$ to type 2 yields an unbiased source representation with squared norm $N_1^{-1} + N_2^{-1}$. The minimum-norm OLS representation is no larger. In this design OLS also fits every observed type mean exactly, so its fitted d is $\widehat{d}$, whether or not type 3 appears. Consequently every candidate obeys

$$|d_{\boldsymbol{\nu}} - \widehat{d}| \le \beta_n \sqrt{r_n} \sqrt{N_1^{-1} + N_2^{-1}} \le \sqrt{3}\, \beta_n \sqrt{N_1^{-1} + N_2^{-1}}. \tag{30}$$

Clipping $\widehat{d}$ to $[-B, B]$ only decreases its distance to every possible $d_{\boldsymbol{\nu}}$. Equation (29) supplies a comparison marginal proving

$$v_{n,I} \le B/9 + \frac{2\sqrt{3}}{3} \beta_n \sqrt{N_1^{-1} + N_2^{-1}}. \tag{31}$$

The learner does not evaluate the formula for this marginal. It solves its game. Conditional on both counts being at least $n/8$, the remainder is at most $(8\sqrt{3}/3)\beta_n/\sqrt{n}$. Those count failures have total probability at most $2e^{-n/32}$. Since $r_n \le 3$, a universal $n \ge C \log(1/\delta)$ makes this remainder and ζ_n sum to at most $B/72 = B/8 - B/9$, and makes the count failure at most $\delta/2$. On coverage and the stated count event, $v_{n,I} + \zeta_n \le \epsilon_0$. Proposition 3 and the stopping rules therefore ensure an ϵ_0-good output under completed optimization. This argument never requires $N_3 > 0$.

The same count concentration and conditional Hoeffding argument, applied directly to the induced marginal $\mathbf{Q}(\mathbf{p}(\widehat{d}))$, with a feasible fallback when a frequent type is missing, shows that the explicit functional–mixing rule also attains the order $\log(1/\delta)$.

Let $\pi_{0,n}$ denote the OLS–flow policy. When all counts are positive,

$$\widehat{v}_1 = \overline{Y}_1 - \overline{Y}_3/2, \qquad \widehat{v}_2 = \overline{Y}_2 - \overline{Y}_3/2, \qquad \widehat{v}_3 = \overline{Y}_3.$$

If each mean error is at most s, the coordinate errors are at most $3s/2$, and (25) plus score optimality bounds regret by s. Taking $s = B/8$ gives

$$\mathbb{P}\{R_{\mathbf{v}}(\pi_{0,n}) > B/8\} \le 2e^{-n/32} + e^{-n\rho/8} + 4e^{-n/256} + 2e^{-n\rho/64}. \tag{32}$$

This is uniform over the stated logger class and all bounded laws, and proves the claimed integral upper order.

Lower bounds. For the mixed lower bound choose

$$\mathbf{v}_1 = (15B/16, B/16, 0)^\top, \qquad \mathbf{v}_2 = (B/16, 15B/16, 0)^\top.$$

Goodness at $B/8$ implies $p_1 \ge 4/7$ and $p_2 \ge 4/7$, respectively, so the two good sets are disjoint. Choose $\lambda_1 = \lambda_2 = (1-\rho)/2$, $\lambda_3 = \rho$, and scaled-Bernoulli rewards. The rare laws agree. The one-record divergence is $(1-\rho)(7/8)\log 15$, bounded by a universal constant. Proposition 7 gives $n \gtrsim \log(1/\delta)$.

For the integral lower bound use

$$\mathbf{v}_3 = (23B/48, 23B/48, B/24)^\top, \qquad \mathbf{v}_4 = (B/8, B/8, 3B/4)^\top. \tag{33}$$

Both frequent means equal $B/2$. At $\mathbf{v}_3$, only matchings 1 and 2 are good, since matching 3 loses $7B/48$. At $\mathbf{v}_4$, only matching 3 is good, since either other matching loses $5B/24$. The rare type has means $B/24$ and $3B/4$. With its probability equal to ρ, the one-record divergence is less than 2ρ. Equation (15) gives maximum failure at least $\frac{1}{4}e^{-2n\rho}$ for every integral-output learner, including

outer randomization. At $\rho = 0$ the two source experiments are identical and maximum failure is at least $1/2$.

The pair (33) is not a mixed-policy conflict. The induced marginal $\mathbf{Q}((1/3, 0, 2/3)^\top)$ has regrets $7B/72$ and $5B/72$ under the two models, respectively. This explains why the integral lower bound does not apply to the mixed learner. All parameters and logger laws in both lower bounds are fixed before the random design. □

The distinction is in the information needed for the *output*, not in a better estimate of every reward. A direct functional–mixing rule also attains the mixed order. The proof constructs that rule only as a comparison marginal. Algorithm 1 finds its own mixture from the game. Thus the theorem does not assert superiority over every pre-existing mixed decision rule.

B.3 Rare-direction source radius

Lemma 17 (A source radius independent of rare observations)**.** *In the experiment of Definition 15, fix any logger with $\lambda_1, \lambda_2 \geq 1/4$. For every $t \geq 0$,*

$$B\omega(t) \leq \frac{B}{9} + \frac{4\sqrt{2}}{3} Bt. \tag{34}$$

The bound is independent of the probability of source type 3*.*

Proof. For a parameter $\mathbf{v}$, write $d_{\mathbf{v}} = v_1 - v_2$. Equation (26) gives $d_{\mathbf{v}} = \mu_1(\mathbf{v}) - \mu_2(\mathbf{v})$. If $\mathbf{v}' \in \mathcal{N}_\lambda(\mathbf{v}, t)$, weighted Cauchy–Schwarz yields

$$\begin{aligned} |d_{\mathbf{v}'} - d_{\mathbf{v}}| &\leq \sqrt{\lambda_1^{-1} + \lambda_2^{-1}} \left(\sum_{j=1}^{2} \lambda_j \{\mu_j(\mathbf{v}') - \mu_j(\mathbf{v})\}^2 \right)^{1/2} \\ &\leq 2\sqrt{2}\, \|\mathbf{v}' - \mathbf{v}\|_{\mathbf{\Sigma}_\lambda} \leq 2\sqrt{2}\, Bt. \end{aligned}$$

Use the induced marginal $\mathbf{Q}(\mathbf{p}(d_{\mathbf{v}}))$ from Lemma 16 as a comparison marginal. For every $\mathbf{v}'$ in the neighborhood,

$$\Delta_{\mathbf{v}'}(I, \mathbf{Q}(\mathbf{p}(d_{\mathbf{v}}))) \leq \frac{B}{9} + \frac{2}{3}|d_{\mathbf{v}'} - d_{\mathbf{v}}| \leq \frac{B}{9} + \frac{4\sqrt{2}}{3} Bt.$$

Taking the robust minimum and then the supremum over centers proves (34). □

B.4 Coupled construction and exact fiber geometry

Definition 18 (Coupled source model and lifted batches)**.** Let

$$\Theta = \{(c, u, v_1, v_2) : 0 \leq u \leq c \leq B,\ v_1, v_2 \geq 0,\ v_1 + v_2 \leq c\}, \qquad v_3 = c - v_1 - v_2. \tag{35}$$

Write $\mathbf{v} := (v_1, v_2, v_3)^\top$. The two source types have conditional means c and u and occur with probabilities $1 - \rho$ and ρ. Define

$$\mathbf{M} = \begin{pmatrix} 1/2 & 0 & 1/4 \\ 1/4 & 1/2 & 0 \\ 0 & 1/4 & 1/2 \end{pmatrix}, \qquad \mathbf{d} = (1, -1, 0)^\top, \qquad s = (u - c/2)/8, \qquad \mathbf{r} = \mathbf{M}\mathbf{v} + s\mathbf{d}. \tag{36}$$

Use eligibility (23), diagonal reward $c/2$, and the two swap rewards $c/2 + (r_2 - r_1)/2$ and $c/2 + (r_3 - r_1)/2$. For $1 \leq L \leq H$, replace each unit and action type by L distinguishable copies, with complete eligibility inside allowed type pairs and individual capacity one. Use a fixed global action set of size $3H$, with zero capacities for unused actions. Let $\mathcal{E}_{H,\rho}$ include the entire declared lift and relabeling domain and all bounded laws on (35). The global feature map also includes auxiliary context–action pairs with feature vectors $(1, -1, 0, 0)^\top$, $(0, 0, 1, 0)^\top$, $(0, 0, 0, 1)^\top$, and $(1, 0, -1, -1)^\top$. These pairs have zero logging probability and do not occur in the deployment domain.

Lemma 19 (The common allocation image of the lifted batches)**.** *The global feature rows generate the full model in Definition 18, and every eligible target reward is in* $[0, B]$*. At lift* L*, the feasible type tables are*

$$\mathbf{X}(a,b) = \begin{pmatrix} L-a & a & 0 \\ a & L-a-b & b \\ 0 & b & L-b \end{pmatrix}, \qquad a,b \geq 0, \quad a+b \leq L. \tag{37}$$

The normalized feature image is the triangle $\mathcal{S}_3$*, parameterized by*

$$\mathbf{p} = (1 - a/L - b/L, a/L, b/L)^\top.$$

For each $\mathbf{p} \in \mathcal{S}_3$*, fix a feasible marginal* $\mathbf{Q}_L(\mathbf{p}) \in \mathcal{P}(I_L)$ *realizing this normalized coordinate. Then, for every original model,*

$$\Delta_{\boldsymbol{\theta}}(I_L, \mathbf{Q}_L(\mathbf{p})) = \frac{1}{3}\left(\max_{j\in[3]} r_j - \mathbf{p}^\top \mathbf{r}\right). \tag{38}$$

For any relabeling $I \in \mathcal{I}$ *of* I_L*, let* $\mathbf{Q}_I(\mathbf{p})$ *denote the corresponding relabeling of* $\mathbf{Q}_L(\mathbf{p})$*. Each extreme point has an integral feasible lift.*

Proof. The global reward forms $c, u, c-u, v_1, v_2, c-v_1-v_2$ generate exactly (35). Each component of $\mathbf{Mv}$ lies in $[0, c/2]$, while $|s| \leq c/16$. Thus $|r_j - r_1| \leq 5c/8$, and every eligible target reward belongs to $[3c/16, 13c/16] \subseteq [0, B]$. No additional parameter restriction is introduced by the target.

The table follows by solving the three row and three column sum equations on the seven eligible type pairs. The two exchange amounts compete for the middle capacity. Its normalized image is the triangle with coordinates

$$\mathbf{p} = (1 - a/L - b/L,\ a/L,\ b/L)^\top \in \mathcal{S}_3.$$

Substitution in the reward sum proves (38). Each triangle vertex has an integral lift. □

Lemma 20 (Source-dependent minimax mixture)**.** *For any batch* $I \in \mathcal{I}$*, at exact source values* (c, u)*, the worst source-fiber loss of* $\mathbf{Q}_I(\mathbf{p})$ *is*

$$J(c, u, \mathbf{p}) = \frac{1}{3}\max_{j\in[3]}\left\{c/2 + sd_j - c(\mathbf{M}^\top \mathbf{p})_j - s\mathbf{p}^\top \mathbf{d}\right\}. \tag{39}$$

For $c > 0$*, a minimizing reduced coordinate and the optimal value are*

$$\mathbf{p}^*(c,u) = \begin{pmatrix} 1/6 + u/(3c) \\ 5/12 - u/(6c) \\ 5/12 - u/(6c) \end{pmatrix}, \qquad a(c,u) = c/12 - \frac{(u - c/2)^2}{48c}. \tag{40}$$

At $c = 0$*, necessarily* $u = 0$*, and every feasible marginal has zero loss. Set* $\mathbf{p}^*(0,0) := \mathbf{1}/3$ *and* $a(0,0) := 0$*.*

Proof. At fixed (c, u) the latent variable lies in $c\mathcal{S}_3$. Regret is convex in $\mathbf{v}$, so its maximum occurs at a latent vertex. At $\mathbf{v} = c\mathbf{e}_j$, matching j is optimal: its unperturbed gap is at least $c/4$, and the source-dependent perturbation changes a pairwise gap by at most $2|s| \leq c/8$. This proves the displayed expression for J.

The vector is feasible and equalizes the three terms in (39). For a matching lower bound, the feasible vertex weights

$$\mathbf{w}^* = \mathbf{1}/3 + (s/c)(-4/3, 8/3, -4/3)^\top$$

make the weighted average of those three losses equal to $a(c,u)$ for every $\mathbf{p} \in \mathcal{S}_3$. Both weight vectors are nonnegative since $|s| \leq c/16$. This proves $\min_{\mathbf{p}} J(c, u, \mathbf{p}) = a(c,u)$, including the zero-loss case $c = 0$ by continuity.

At $u = c/2$ the floor is $c/12$. Replacing the correlated score image by three independent intervals $[0, c/2]$ instead gives $c/9$. The original coupling is therefore essential to the claimed accuracy. □

B.5 Proof of Theorem 10

Theorem 10 (restated). *There are universal $c_0, C_1, C_2 > 0$ such that, for every integer $H \geq 1$, $0 < \rho \leq 1/2$, $0 < \delta \leq 1/32$, and $0 < \tau \leq c_0 B$,*

$$C_1 \frac{B^2}{\rho\tau^2} \log \frac{1}{\delta} \leq N^*(B/12 + \tau, \delta; \mathcal{E}_{H,\rho}) \leq C_2 \frac{B^2}{\rho\tau^2} \log \frac{1}{\delta}.$$

Completed Algorithm 1 attains the upper bound with completed exact responses.

Proof. Upper bound. The affine coefficients of the three branches in (39) give, for any fixed $\mathbf{p}$,

$$|J(c, u, \mathbf{p}) - J(c', u', \mathbf{p})| \leq \frac{5}{24}|c - c'| + \frac{1}{12}|u - u'|. \tag{41}$$

For example, the coefficient of u is $(d_j - \mathbf{p}^\top \mathbf{d})/24$, of magnitude at most $1/12$. The coefficient of c is obtained by substituting $s = u/8 - c/16$ in (39). Its extrema over the simplex occur at its three vertices and are bounded by $5/24$.

The source rank is at most two. On coverage and positive counts, every candidate has

$$|c' - c| \leq 2\sqrt{2}\,\beta_n/\sqrt{N_c}, \qquad |u' - u| \leq 2\sqrt{2}\,\beta_n/\sqrt{N_u}.$$

This follows either directly in source coordinates or by the minimum-norm influence representation. It does not require any latent estimate. Use the induced marginal $\mathbf{Q}_{I_L}(\mathbf{p}^*(c, u))$ from (40) only as a comparison policy. Since $a(c, u) \leq B/12$, (41) yields

$$v_{n,I_L} \leq \frac{B}{12} + 2\sqrt{2}\,\beta_n \left(\frac{5}{24\sqrt{N_c}} + \frac{1}{12\sqrt{N_u}} \right). \tag{42}$$

The event $N_c \geq n/4$, $N_u \geq n\rho/2$ fails with probability at most $2e^{-n\rho/8}$. The remainder in (42) is at most $CB\sqrt{\log(1/\delta)/(n\rho)}$ on this event. A sufficiently large constant in the theorem makes this remainder at most $\tau/2$, $\zeta_n \leq \tau/2$, and the count failure at most $\delta/2$. On coverage and the stated count event, $v_{n,I} + \zeta_n \leq B/12 + \tau$ uniformly over the batch domain. Proposition 3 and the stopping rules therefore prove the upper bound under completed optimization.

Equation (38) is the same normalized game at every declared lift and relabeling. Its confidence event is shared, so there is no union penalty in H.

Lower bound. Restrict to the smallest lift. Set $c = B$ and $u_1 = B/2 + h$, $u_2 = B/2 - h$, where $0 < h \leq B/4$. For source value u_1, use actual latent models $B\mathbf{e}_1$ and $B\mathbf{e}_3$. For u_2, use $B\mathbf{e}_2$ and $B\mathbf{e}_3$. Denote their losses by $\Delta_{11}, \Delta_{13}, \Delta_{22}, \Delta_{23}$. Directly expanding the affine branches in (39) gives, for every original marginal,

$$\tfrac{1}{3}\Delta_{11} + \tfrac{1}{6}\Delta_{13} + \tfrac{1}{3}\Delta_{22} + \tfrac{1}{6}\Delta_{23} = B/12 + h/36. \tag{43}$$

Take $h = 72\tau$ and $c_0 \leq 1/288$. No policy is $(B/12 + \tau)$-good at all four models. The sets good at both models with source u_1 and good at both with source u_2 are disjoint.

Give the c source deterministic reward B, and the u source scaled-Bernoulli reward with mean u_i. The two latent models within each source value have identical data laws. A learner with failure at most δ at each of the four fixed models induces a test between u_1, u_2 with each error at most 2δ, by the union bound within each pair and the preceding disjointness. The Bernoulli inequality

$$a \log(a/b) + (1 - a) \log((1 - a)/(1 - b)) \leq \frac{(a - b)^2}{b(1 - b)}$$

gives one-record divergence at most $C\rho h^2/B^2$, since both probabilities lie in $[1/4, 3/4]$. The two-law testing bound gives

$$4\delta \geq \tfrac{1}{2} \exp(-Cn\rho h^2/B^2),$$

which proves the lower order. The models and logger are fixed before sampling, and the argument permits arbitrary outer learner randomization. It is a lower bound for the same original marginal loss, not for parameter estimation. □

Computation on the coupled family. All declared lifts share the same three-point allocation image, so they induce the same reduced game. Exact response formulations are given in Appendix C.

B.6 Pairwise compatibility and a three-model witness

Lemma 21 (Pairwise compatibility and a three-model witness)**.** *In the experiment of Definition 18, every two parameters $\boldsymbol{\theta}, \boldsymbol{\nu} \in \Theta$ have a common policy with worst-batch regret at most $5B/64$. In particular, for every batch $I \in \mathcal{I}$,*

$$v_I(\{\boldsymbol{\theta}, \boldsymbol{\nu}\}) \le \frac{5B}{64} < \frac{B}{12}. \tag{44}$$

For $0 \le h \le B/4$, define the parameters in the original coordinates (c, u, v_1, v_2) by

$$\begin{aligned} \boldsymbol{\nu}_1(h) &= (B,\ B/2+h,\ B,\ 0)^\top, \\ \boldsymbol{\nu}_2(h) &= (B,\ B/2-h,\ 0,\ B)^\top, \\ \boldsymbol{\nu}_3(h) &= (B,\ B/2+h,\ 0,\ 0)^\top. \end{aligned} \tag{45}$$

For every batch $I \in \mathcal{I}$,

$$v_I(\{\boldsymbol{\nu}_1(h), \boldsymbol{\nu}_2(h), \boldsymbol{\nu}_3(h)\}) \ge \frac{(2B+h)^2}{16(3B+2h)} = \frac{B}{12} + \frac{h}{36} + \frac{h^2}{144(3B+2h)}. \tag{46}$$

Proof. By Lemma 19, it suffices to work with $\mathbf{p} \in \mathcal{S}_3$ and the regret in (38). For any parameter, $\mathbf{v} \ge 0$, $\mathbf{1}^\top \mathbf{v} = c \le B$, and $|s| \le c/16$. Equation (36) therefore gives

$$\begin{aligned} -5c/8 \le r_1 - r_2 \le 3c/8, \\ -5c/16 \le r_1 - r_3 \le 9c/16, \\ -9c/16 \le r_2 - r_3 \le 5c/16. \end{aligned} \tag{47}$$

For example, $r_1 - r_2 = v_1/4 - v_2/2 + v_3/4 + 2s$. The other two differences are $v_1/2 - v_2/4 - v_3/4 + s$ and $v_1/4 + v_2/4 - v_3/2 - s$. These identities prove all three ranges.

Choose an optimal index for each of the two parameters. If the indices agree, the corresponding integral anchor is optimal for both models. Otherwise use the appropriate mixture from

$$\mathbf{p}^{12} = (3/8, 5/8, 0)^\top, \qquad \mathbf{p}^{13} = (2/3, 0, 1/3)^\top, \qquad \mathbf{p}^{23} = (0, 1/3, 2/3)^\top.$$

When index 1 is optimal, the regret of $\mathbf{p}^{12}$ is $5(r_1 - r_2)/24 \le 5c/64$. When index 2 is optimal, it is $(r_2 - r_1)/8 \le 5c/64$. For $\mathbf{p}^{13}$, the bounds at optimal indices 1 and 3 are $c/16$ and $5c/72$. For $\mathbf{p}^{23}$, the bounds at optimal indices 2 and 3 are $5c/72$ and $c/16$. Since $c \le B$, each relevant bound is at most $5B/64$. The same mixture has a feasible lift at every batch. Using these lifts defines a common policy and proves (44).

The parameters in (45) belong to Θ. Their latent vectors are $B\mathbf{e}_1$, $B\mathbf{e}_2$, and $B\mathbf{e}_3$. The corresponding score vectors are

$$\mathbf{r}^{(1)} = \begin{pmatrix} B/2 + h/8 \\ B/4 - h/8 \\ 0 \end{pmatrix}, \qquad \mathbf{r}^{(2)} = \begin{pmatrix} -h/8 \\ B/2 + h/8 \\ B/4 \end{pmatrix}, \qquad \mathbf{r}^{(3)} = \begin{pmatrix} B/4 + h/8 \\ -h/8 \\ B/2 \end{pmatrix}.$$

Index j is optimal at $\boldsymbol{\nu}_j(h)$ for $j \in [3]$. Define

$$\mathbf{w}(h) = \frac{1}{2(3B+2h)} \begin{pmatrix} 2B+h \\ 2B+2h \\ 2B+h \end{pmatrix}.$$

This vector belongs to $\mathcal{S}_3$ and satisfies

$$\sum_{j=1}^{3} w_j(h) \mathbf{r}^{(j)} = \frac{B}{4} \mathbf{1}.$$

Consequently, for every $\mathbf{p} \in \mathcal{S}_3$,

$$\sum_{j=1}^{3} w_j(h)\Delta_{\boldsymbol{\nu}_{j(h)}}(I, \mathbf{Q}_I(\mathbf{p})) = \frac{1}{3}\left(\sum_{j=1}^{3} w_j(h) r_j^{(j)} - \frac{B}{4}\right) = \frac{(2B+h)^2}{16(3B+2h)}.$$

The maximum of the three regrets is at least this weighted average. Minimizing over $\mathbf{p}$ proves (46). □

Every batch in this experiment has $p_I = 2$. For $h > 0$, the three models in (45) have no common $B/12$-good marginal, although every pair has one. Thus the witness bound $p_I + 1$ in Proposition 2 can be attained. In particular, the pairwise condition in Proposition 7 cannot yield a lower bound at target $B/12 + \tau$ in this experiment.

B.7 Local modulus and source-radius comparison

Lemma 22 (Local modulus of the coupled experiment)**.** *Fix the source law of Definition 18 with $0 < \rho \le 1/2$, and let $\boldsymbol{\theta}_0 = (B, B/2, B/3, B/3)^\top$. Then $B\omega(0) = B/12$. For every $0 \le t \le \sqrt{\rho}/4$,*

$$\frac{B}{12} + \frac{Bt}{36\sqrt{\rho}} \le B\omega_{\boldsymbol{\theta}_0}(t) \le B\omega(t) \le \frac{B}{12} + \frac{\sqrt{29}\,Bt}{24\sqrt{\rho}}. \tag{48}$$

The constants do not depend on the lift bound H.

Proof. Only c and u are observed at the source. Hence

$$\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}_\lambda} = (1-\rho)(c'-c)^2 + \rho(u'-u)^2. \tag{49}$$

Lemma 20 shows that the exact fiber value is $a(c,u) \le B/12$, with equality at $(c,u) = (B, B/2)$. The normalized game is the same at every declared lift, so $B\omega(0) = B/12$.

For the lower bound, put $h = Bt/\sqrt{\rho}$. The three parameters in (45) lie in $\mathcal{N}_\lambda(\boldsymbol{\theta}_0, t)$ by (49). The condition on t ensures $h \le B/4$. Monotonicity of v_I under inclusion and (46) give

$$v_I(\mathcal{N}_\lambda(\boldsymbol{\theta}_0, t)) \ge \frac{B}{12} + \frac{Bt}{36\sqrt{\rho}}.$$

This proves the first inequality in (48).

For the upper bound, fix an arbitrary center $\boldsymbol{\theta}$ and batch $I \in \mathcal{I}$, with source values (c,u). Choose a minimizing reduced coordinate $\mathbf{p}_0 \in \mathcal{S}_3$ for its exact fiber. When $c = 0$, any $\mathbf{p}_0 \in \mathcal{S}_3$ has zero fiber loss. For every $\boldsymbol{\theta}' \in \mathcal{N}_\lambda(\boldsymbol{\theta}, t)$, with source values (c', u'), equation (41) gives

$$\begin{aligned}
\Delta_{\boldsymbol{\theta}'}(I, \mathbf{Q}_I(\mathbf{p}_0)) &\le J(c', u', \mathbf{p}_0) \\
&\le a(c,u) + \frac{5}{24}|c'-c| + \frac{1}{12}|u'-u| \\
&\le \frac{B}{12} + Bt\sqrt{\frac{25}{576(1-\rho)} + \frac{1}{144\rho}} \\
&\le \frac{B}{12} + \frac{\sqrt{29}\,Bt}{24\sqrt{\rho}}.
\end{aligned}$$

The penultimate inequality is weighted Cauchy–Schwarz applied to (49). The last uses $\rho/(1-\rho) \le 1$. Taking the robust minimum and the suprema over centers and batches proves the upper bound. □

The excess in Lemma 22 is of order $Bt/\sqrt{\rho}$. Combined with the three-model witness above, Corollary 6 and Proposition 8 recover the rate in Theorem 10 up to universal constants. Thus the general source-radius comparison is sharp on this family even though no pair is incompatible at the target accuracy.

C EXACT COMPUTATIONAL REDUCTION

This appendix records the exact computational reduction associated with the allocation-comparison geometry. The statistical guarantees do not require $p_I \ll d$ or efficient enumeration of the reduced allocation polytope.

C.1 LOSS-PRESERVING ALLOCATION COORDINATES

For each batch I, choose a feasible anchor $\mathbf{z}_I^0$, set

$$\mathbf{c}_I := \mathbf{\Phi}(I, \mathbf{z}_I^0),$$

and let $\mathbf{U}_I \in \mathbb{R}^{d \times p_I}$ have full column rank with $\mathrm{range}(\mathbf{U}_I) = \mathbf{L}_I$. Define

$$\mathbf{x}_I(\mathbf{Q}) := \mathbf{U}_I^\dagger[\mathbf{\Phi}(I, \mathbf{Q}) - \mathbf{c}_I], \qquad \mathcal{X}_I := \{\mathbf{x}_I(\mathbf{Q}) : \mathbf{Q} \in \mathcal{P}(I)\}, \qquad H_I(\mathbf{u}) := \max_{\mathbf{x} \in \mathcal{X}_I} \mathbf{x}^\top \mathbf{u}. \quad (50)$$

The set $\mathcal{X}_I$ is a finite polytope.

Theorem 23 (Exact allocation-feature reduction)**.** *For every $\boldsymbol{\theta} \in \Theta$, batch I, and $\mathbf{Q} \in \mathcal{P}(I)$,*

$$\mathbf{\Phi}(I, \mathbf{Q}) = \mathbf{c}_I + \mathbf{U}_I \mathbf{x}_I(\mathbf{Q}), \qquad \Delta_{\boldsymbol{\theta}}(I, \mathbf{Q}) = H_I(\mathbf{U}_I^\top \boldsymbol{\theta}) - \mathbf{x}_I(\mathbf{Q})^\top \mathbf{U}_I^\top \boldsymbol{\theta}. \quad (51)$$

Every $\mathbf{x} \in \mathcal{X}_I$ has a feasible lift in $\mathcal{P}(I)$. Hence this change of output coordinates preserves the regret of every feasible marginal at every $\boldsymbol{\theta} \in \Theta$, and therefore preserves robust allocation loss and worst-batch PAC sample complexity.

If

$$\mathcal{X}_I = \mathrm{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_{N_I}\},$$

then for every nonempty compact $\mathcal{C} \subseteq \Theta$,

$$\max_{\boldsymbol{\nu} \in \mathcal{C}} \Delta_{\boldsymbol{\nu}}(I, \mathbf{Q}) = \max_{j \in [N_I]} \max_{\boldsymbol{\nu} \in \mathcal{C}} [\mathbf{U}_I(\mathbf{x}_j - \mathbf{x}_I(\mathbf{Q}))]^\top \boldsymbol{\nu}. \quad (52)$$

Proof. For every $\mathbf{Q} \in \mathcal{P}(I)$, $\mathbf{\Phi}(I, \mathbf{Q}) - \mathbf{c}_I \in \mathbf{L}_I = \mathrm{range}(\mathbf{U}_I)$, so

$$\mathbf{U}_I \mathbf{U}_I^\dagger[\mathbf{\Phi}(I, \mathbf{Q}) - \mathbf{c}_I] = \mathbf{\Phi}(I, \mathbf{Q}) - \mathbf{c}_I.$$

This proves the first identity in (51). It also gives

$$V_{\boldsymbol{\theta}}(I, \mathbf{Q}) = \mathbf{c}_I^\top \boldsymbol{\theta} + \mathbf{x}_I(\mathbf{Q})^\top \mathbf{U}_I^\top \boldsymbol{\theta}, \qquad V_{\boldsymbol{\theta}}^*(I) = \mathbf{c}_I^\top \boldsymbol{\theta} + H_I(\mathbf{U}_I^\top \boldsymbol{\theta}),$$

and subtraction proves the regret identity. Every point of $\mathcal{X}_I$ has a feasible lift by definition. Under the fixed measurable-selection convention of Section 3, feasible lifts may be selected batchwise measurably, giving a loss-preserving correspondence between policies. Since the transformation preserves regret pointwise in $\boldsymbol{\theta}$, it therefore preserves minimax regret, common ϵ-good marginals, and worst-batch PAC sample complexity.

For the final claim, the finite comparator image gives

$$\max_{\boldsymbol{\nu} \in \mathcal{C}} \max_{j \in [N_I]} [\mathbf{U}_I(\mathbf{x}_j - \mathbf{x}_I(\mathbf{Q}))]^\top \boldsymbol{\nu} = \max_{j \in [N_I]} \max_{\boldsymbol{\nu} \in \mathcal{C}} [\mathbf{U}_I(\mathbf{x}_j - \mathbf{x}_I(\mathbf{Q}))]^\top \boldsymbol{\nu},$$

which is (52). □

C.2 CONSTRUCTING THE COMPARISON SPACE

Proposition 24 (Comparison-space construction)**.** *The comparison space L_I can be constructed using at most $2d + 1$ ordinary allocation optimizations.*

Proof. Start from one feasible anchor. Maintain subspaces $\mathbf{D} \subseteq L_I$ and $\mathbf{A} \subseteq L_I^\perp$, initially zero. While $\mathbf{D} + \mathbf{A} \neq \mathbb{R}^d$, choose a nonzero $\mathbf{w}$ orthogonal to $\mathbf{D} + \mathbf{A}$ and maximize and minimize $\mathbf{w}^\top \mathbf{\Phi}(I, \mathbf{z})$ over $\mathbf{z} \in \mathcal{Z}(I)$. If the two values differ, their feature difference adds an independent direction to $\mathbf{D}$. If they agree, $\mathbf{w} \in L_I^\perp$ and adds an independent direction to $\mathbf{A}$. Each pair of optimizations resolves one dimension, so after at most d rounds $\mathbf{D} = L_I$. Including the anchor gives at most $2d + 1$ ordinary allocation optimizations. □

This construction identifies the comparison space but does not by itself enumerate the vertices of $\mathcal{X}_I$.

C.3 Exact model responses under finite representations

For the computational statements in this subsection only, suppose the original parameter set has a supplied exact finite representation

$$\Theta = \{\boldsymbol{\nu} : \mathbf{A}\boldsymbol{\nu} \le \mathbf{b}\}, \qquad \mathcal{C}_n = \{\boldsymbol{\nu} : \mathbf{A}_n\boldsymbol{\nu} \le \mathbf{b}_n\}. \tag{53}$$

No global constraint is dropped.

For a batch I, let $\mathcal{E}_I := \{(i,a) : a \in \mathcal{A}(X_i)\}$ and write $\boldsymbol{\phi}_e := \boldsymbol{\phi}(X_i,a)$ and $Q_e := Q_{ia}$ for $e = (i,a)$.

Proposition 25 (Exact model responses)**.** *For fixed* $\mathbf{Q}$*, the model best-response value in* (5) *is the optimum of*

$$\begin{aligned} \max_{\boldsymbol{\nu},\mathbf{z},\mathbf{u}} \quad & \frac{1}{m}\sum_{e\in\mathcal{E}_I}\left(u_e - Q_e\boldsymbol{\phi}_e^\top\boldsymbol{\nu}\right) \\ \text{s.t.} \quad & \boldsymbol{\nu}\in\mathcal{C}_n, \qquad \mathbf{z}\in\mathcal{Z}(I), \\ & 0 \le u_e \le Bz_e, \qquad u_e \le \boldsymbol{\phi}_e^\top\boldsymbol{\nu}, \\ & u_e \ge \boldsymbol{\phi}_e^\top\boldsymbol{\nu} - B(1-z_e) \qquad (e\in\mathcal{E}_I). \end{aligned} \tag{54}$$

If a complete vertex representation $\mathcal{X}_I = \mathrm{conv}\{\mathbf{x}_1,\dots,\mathbf{x}_{N_I}\}$ *is available, the same response can instead be computed from the* N_I *support problems*

$$\max_{\mathbf{A}_n\boldsymbol{\nu}\le\mathbf{b}_n} \left[\mathbf{U}_I(\mathbf{x}_j - \mathbf{x}_I(\mathbf{Q}))\right]^\top\boldsymbol{\nu}, \qquad j\in[N_I]. \tag{55}$$

Proof. For $z_e = 0$, the constraints in (54) force $u_e = 0$. For $z_e = 1$, they force $u_e = \boldsymbol{\phi}_e^\top\boldsymbol{\nu}$. Hence $u_e = z_e\boldsymbol{\phi}_e^\top\boldsymbol{\nu}$ exactly, and the objective equals

$$\left[\boldsymbol{\Phi}(I,\mathbf{z}) - \boldsymbol{\Phi}(I,\mathbf{Q})\right]^\top\boldsymbol{\nu}$$

at one common model $\boldsymbol{\nu}\in\mathcal{C}_n$. This is precisely the model best-response problem.

When the complete allocation-image vertex list is available, (55) follows directly from (52). No enumeration of the vertices of $\mathcal{C}_n$ is required. □

These are exact formulations rather than a universal polynomial-time claim.

Certified responses. A certified target stop requires a verified upper bound on $g_{n,I}(\mathbf{Q})$. Certified gap stopping requires both a verified upper bound $U \ge g_{n,I}(\mathbf{Q})$ and a valid lower bound $L \le v_{n,I}$, so that $U - L \le \zeta_n$ certifies $g_{n,I}(\mathbf{Q}) \le v_{n,I} + \zeta_n$.

D Same-information baselines and objectives

This appendix specifies the comparison rules used under the same logged information as Algorithm 1.

Definition 26 (Constrained regression–flow)**.** The constrained-regression baseline first minimizes

$$\|\mathbf{y} - \mathbf{F}_n\boldsymbol{\nu}\|_2^2$$

over Θ, then selects the minimum-Euclidean-norm parameter among the minimizers, and finally applies the same integral flow tie rule as OLS–flow. Fitted source predictions are unique. No general dominance between constrained and unconstrained regression is assumed.

For the same nonempty candidate $\mathcal{C}_n$, consider

$$\max_{\mathbf{Q}\in\mathcal{P}(I)} \min_{\boldsymbol{\nu}\in\mathcal{C}_n} V_{\boldsymbol{\nu}}(I,\mathbf{Q}), \tag{56}$$

$$\max_{\mathbf{Q}\in\mathcal{P}(I)} \min_{\boldsymbol{\nu}\in\mathcal{C}_n} \left[\boldsymbol{\Phi}(I,\mathbf{Q}) - \boldsymbol{\Phi}(I,\mathbf{z}_0)\right]^\top\boldsymbol{\nu}, \tag{57}$$

$$\min_{\mathbf{Q}\in\mathcal{P}(I)} \max_{\boldsymbol{\nu}\in\mathcal{C}_n} \Delta_{\boldsymbol{\nu}}(I,\mathbf{Q}). \tag{58}$$

The first is maximin value, the second is reference-relative pessimism using the plug-in allocation $\mathbf{z}_0$, and the third is the minimax-regret objective targeted by Algorithm 1. Algorithm 1 may stop before optimizing the third objective to ζ_n accuracy when it certifies $U \le \epsilon$. Choosing $\mathbf{z}_0$ is feasible in (57) with objective zero, so a ζ-optimal solution protects value relative to $\mathbf{z}_0$ up to ζ on the coverage event. This is a different objective from minimax-regret protection.

Theorem 9 separates the information requirements of mixed and integral outputs. It does not imply statistical dominance of MRA over every mixed rule. In particular, the direct functional–mixing rule can attain the same order on the Rare experiment. The objectives in (56)–(58) are therefore used to distinguish alternative robust decision criteria, not to assert a universal sample-complexity ordering.

E Experimental details and additional results

E.1 Evaluation and comparison rules

All reported allocation values use the normalization in (3). We evaluate a randomized allocation through its marginal $\mathbf{Q}$, without sampling an integral realization for evaluation. Every failure comparison is strict. The same offline log and deployment batches are used by all methods within a replicate. A batch collection sharing one offline log is not treated as a collection of independent log replicates.

MRA uses the confidence set $\mathcal{C}_n$ and the responses in Algorithm 1. Both $U \le \epsilon$ and $U - L \le \zeta_n$, with $\zeta_n = \epsilon/(n+1)$, return the current marginal and its finite mixture. A gap stop with $U > \epsilon$ certifies approximate minimax optimality, not target success. The plug-in output is retained in the reward-bound early-return, empty or unresolved confidence-set, and computational-failure cases. Numerical certification, stopping reasons, final bounds, and mixture representations are recorded separately.

OLS–flow uses the plug-in allocation $\mathbf{z}_0$, while constrained-flow uses Definition 26. Maximin and reference pessimism optimize (56) and (57) over the same $\mathcal{C}_n$ and feasible marginal polytope as MRA. They are mixed-output objective comparators, not integral-only rules. For Rare, the direct mixture uses the frequent-source contrast in Lemma 16. The oracle maximizes value at the true parameter and is an evaluation reference, not a learner with the permitted offline information.

For each synthetic setting, let k be the number of prescribed true models and let $\mathcal{N}$ be its tested sample-size grid. Writing $\widehat{p}_{j,n}$ for the failure proportion from the 500 logs at model j, the plotted risk and sustained crossing at level η are

$$\begin{aligned} \widehat{p}_n &= \max_{j\in[k]} \widehat{p}_{j,n}, \\ \widehat{n}_\eta &= \inf\left\{ n \in \mathcal{N} : \widehat{p}_{n'} \le \eta \quad \forall n' \in \mathcal{N} \cap [n,\infty) \right\}. \end{aligned} \tag{59}$$

The maximum is taken after estimating each model's probability. It is neither a pooled risk nor a maximum over models inside a shared-log event. A crossing at the smallest grid point is left-censored. With $\inf \varnothing = \infty$, absence of a crossing is reported as right censoring beyond the largest grid point. Otherwise, the previous point and crossing form a tested-grid bracket. We do not interpolate a threshold or fit an asymptotic exponent.

E.2 Rare-direction protocol and full curves

We use the two mixed-output and two integral-output witness parameters from Appendix B.2. The source-type probabilities are $((1-\rho)/2, (1-\rho)/2, \rho)$, with $\rho \in \{0, 0.01, 0.02, 0.05, 0.1, 0.25\}$. The experiment fixes $B = 1$, $\epsilon = 1/8$, $\delta = 0.05$, and $n \in \{2^4, 2^5, \ldots, 2^{15}\}$. Each model and grid point uses 500 independent logs, giving 144,000 logs over the complete grid. The corresponding primary risk has $k = 4$ and $\eta = 0.05$ in (59). Table 1 lists the crossings used in Figure 1.

Figure 4 retains every sample size and all six source settings. The crossing comparison does not imply that MRA has lower failure at every sample size. For example, its worst-model failure is 0.904 at $\rho = 0.25$, $n = 16$, despite crossing at $n = 32$. The verified production results retain 12 exceptional empty or unresolved-confidence fallbacks across the full Rare grid. Their recorded outcomes are not removed or changed. The direct mixture's strong performance confirms the role of mixed outputs and the observable contrast, rather than a unique advantage of MRA over all mixed decision rules.

Table 1: **Rare sustained empirical crossings.** The criterion is (59) with $\eta = 0.05$. Censored bounds retain the limits of the tested grid and are not estimates of an unobserved crossing.

ρ	MRA	Direct mixture	OLS–flow	Constrained-flow	Oracle
0	64	64	$> 32{,}768$	$> 32{,}768$	≤ 16
0.01	64	64	1,024	1,024	≤ 16
0.02	64	64	512	512	≤ 16
0.05	64	64	128	128	≤ 16
0.1	64	64	64	64	≤ 16
0.25	32	64	32	32	≤ 16

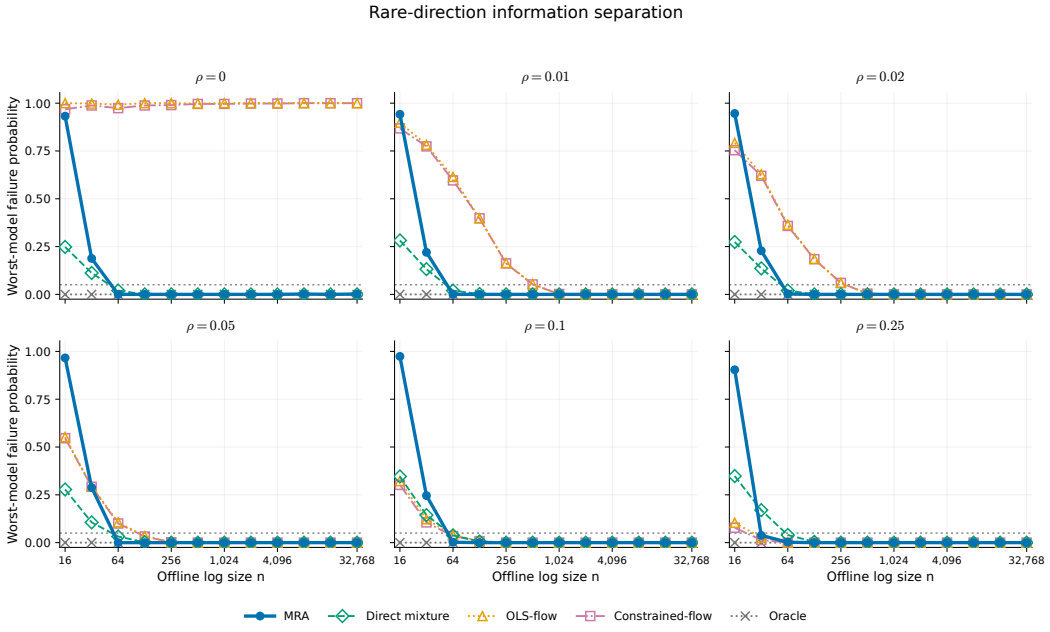


Figure 4: **Full Rare-direction failure curves.** Each panel reports the maximum of four separately estimated model failure probabilities for one ρ. Each model uses 500 independent logs per tested sample size, with $\epsilon = 1/8$ and $n \in \{2^4, \ldots, 2^{15}\}$. The dotted line is 0.05. All methods and small-sample outcomes are retained, including coincident curves and settings unfavorable to MRA.

Table 2: **Coupled tested-grid brackets.** Let n_- be the grid point immediately before the sustained crossing n_+. The bracket is $(n_-, n_+]$. Rescaled coordinates are rounded here to eight decimal places. All nominal brackets are $(10^{-3}, 10^{-2.5}]$.

τ	ρ	n_-	n_+	$n_-\rho\tau^2$	$n_+\rho\tau^2$
0.001	0.02	50,000	158,114	0.00100000	0.00316228
0.001	0.05	20,000	63,246	0.00100000	0.00316230
0.001	0.1	10,000	31,623	0.00100000	0.00316230
0.001	0.2	5,000	15,812	0.00100000	0.00316240
0.002	0.02	12,500	39,529	0.00100000	0.00316232
0.002	0.05	5,000	15,812	0.00100000	0.00316240
0.002	0.1	2,500	7,906	0.00100000	0.00316240
0.002	0.2	1,250	3,953	0.00100000	0.00316240
0.003	0.02	5,556	17,569	0.00100008	0.00316242
0.003	0.05	2,223	7,028	0.00100035	0.00316260
0.003	0.1	1,112	3,514	0.00100080	0.00316260
0.003	0.2	556	1,757	0.00100080	0.00316260

E.3 Coupled protocol and normalized transitions

We use the $L = 1$ lift of the construction in Appendix B.4 and the three witnesses $\boldsymbol{\nu}_1(h), \boldsymbol{\nu}_2(h), \boldsymbol{\nu}_3(h)$ in (45). The parameters are $B = 1$, $\delta = 0.025$, $\tau \in \{0.001, 0.002, 0.003\}$, $\rho \in \{0.02, 0.05, 0.1, 0.2\}$, $h = 72\tau$, and $\epsilon = 1/12 + \tau$. Every τ lies below $1/288$, as required by the chosen witness construction. Only MRA is evaluated on this theorem-aligned grid. For nominal points $x_j = 10^{-4+j/2}$, $j \in \{0, \dots, 8\}$, the sample sizes are the original saved integers $n_j = \lceil x_j/(\rho\tau^2) \rceil$. There are 500 independent logs per witness and grid point, for 162,000 logs in total. The primary risk has $k = 3$ and $\eta = 0.025$ in (59).

All twelve nominal transition brackets are $(10^{-3}, 10^{-2.5}]$. After the saved integer rounding, preceding normalized coordinates range from 0.001 to 0.0010008, and crossing coordinates range from 0.00316228 to 0.00316260. The pale band in Figure 3 represents the common nominal bracket, while the data retain their actual coordinates. It is not a confidence interval or a claim of identical continuous transition thresholds.

At nominal $x = 10^{-3}$, failure of the first two witnesses ranges from 0 to 0.006, whereas that of the third ranges from 0.898 to 1. Thus the maximum over all three remains high even when two witnesses are learned well. At $x = 10^{-2.5}$ and every larger tested point, all three witnesses have zero observed failures. These observations concern the fixed witnesses and do not estimate the supremum over the entire family in Theorem 10.

Table 2 records the actual brackets without changing the saved integer sample sizes.

E.4 LLM context construction and source-fiber evaluation

The prepared LLMRouterBench subset (Li et al., 2026) contains 3,000 complete queries from eight datasets and five models. Each environment fixes a calibration/source split of 1,950 queries and a disjoint set of 1,050 held-out queries, together with its projection and calibrated reward coefficients. Fixed hashing and a seeded bounded projection produce $\mathbf{u}(x) \in [-1, 1]^2$. With $\mathbf{e}_a$ the ath standard basis vector in $\mathbb{R}^5$, the shared features are

$$\mathbf{e}(x) = \frac{(1, \mathbf{u}(x)^\top)^\top}{\sqrt{3}}, \qquad \boldsymbol{\phi}(x, a) = \mathbf{e}(x) \otimes \mathbf{e}_a \in \mathbb{R}^{15}. \tag{60}$$

The fixed calibration rule uses calibration scores and targets median allocation spread 2ϵ subject to mean-reward bounds $[0.05, 0.95]$ on the feature cube. Learners use the original global parameter domain defined by reward bounds $[0, 1]$, not the tighter simulator margin.

Offline contexts are sampled with replacement from the source split, so n counts logged interactions rather than distinct queries. The logger chooses each of the first four actions with probability $1/4$ and never chooses action five. Rewards are Bernoulli with the fixed linear conditional mean. The four observed-action coefficient blocks remain fixed within an environment, and $\mathcal{H}$ contains the six

Table 3: **Source-fiber results with 20 logs per cell.** Failure counts are MRA's strict log-level failures. The remaining columns are means of $T_{n,r}$. OLS–flow, constrained-flow, and reference pessimism share the last column. All four comparators fail on 20/20 logs in each cell.

Environment	n	MRA failures	MRA	Maximin	OLS / constrained / reference
1	32,768	2/20	0.0909	0.1479	0.1507
1	65,536	0/20	0.0922	0.1431	0.1465
1	131,072	0/20	0.0912	0.1440	0.1488
2	32,768	1/20	0.0937	0.1612	0.1655
2	65,536	0/20	0.0936	0.1610	0.1664
2	131,072	0/20	0.0923	0.1581	0.1630
3	32,768	1/20	0.0900	0.1397	0.1498
3	65,536	0/20	0.0901	0.1374	0.1472
3	131,072	0/20	0.0906	0.1356	0.1441
4	32,768	7/20	0.0986	0.1650	0.1680
4	65,536	3/20	0.0959	0.1613	0.1675
4	131,072	2/20	0.0949	0.1658	0.1678

extreme admissible completions of the unobserved block. These completions induce identical source laws and belong to the same source fiber in (6). They are used only by the evaluator and do not replace the learner's finite-log confidence set $\mathcal{C}_n$.

Each held-out batch contains five distinct queries and five actions with capacities $\mathbf{b} = \mathbf{1}$. All $5! = 120$ assignments are feasible. In this small-instance implementation, model best responses explicitly evaluate all 120 comparators, while the restricted game is generated iteratively. These experiments therefore do not establish a computational advantage from avoiding enumeration. The fixed source-fiber settings retain a 30-iteration limit and the five-second limit for each response-support problem, with unchanged numerical tolerances and tie-breaking.

We index the four fixed environments as 1–4, corresponding respectively to seeds 1729, 2718, 3141, and 5772. Environment 1 is the original primary environment displayed in Figure 2. This ordering is unchanged. Each environment uses 20 independent logs at each of $n = 2^{15}$, 2^{16}, and 2^{17}. We use $\epsilon = 0.1$ and $\delta = 0.05$. All 150 earlier logs are retained. The remaining 90 are the predeclared additional replicates in Environments 2–4, with no replacement of existing outcomes. For a fixed environment and method, let $\widehat{\pi}_{n,r}$ be the policy learned from log r, and let $I_{r,1}, \ldots, I_{r,32}$ be its held-out batches. Define

$$T_{n,r} = \max_{j \in [32]} \max_{\boldsymbol{\nu} \in \mathcal{H}} \Delta_{\boldsymbol{\nu}}\big(I_{r,j}, \widehat{\pi}_{n,r}(I_{r,j})\big),$$
$$\widehat{p}_n = \frac{1}{20} \sum_{r=1}^{20} \mathbb{1}\{T_{n,r} > 0.1\}, \qquad \overline{T}_n = \frac{1}{20} \sum_{r=1}^{20} T_{n,r}. \tag{61}$$

Figure 5 plots $\widehat{p}_n$ and Figure 6 plots $\overline{T}_n$. The completion maximum is inside each log-level failure event. The displayed mean is not a mean over individual batches or completions before taking their maximum. Neither the 32 batches nor the six completions are additional independent log replicates. This finite collection does not exhaust the general worst-batch domain in (5).

E.5 Source-fiber numerical summaries and uncertainty

Table 3 gives the MRA failure counts and continuous means for all twelve environment–sample-size settings. Every comparator has 20 failed logs out of 20 in every cell. The reported mean log-level maxima of OLS–flow, constrained-flow, and reference pessimism coincide at the displayed precision, so the table groups those three means in one column. Their curves remain separately identified in the figures.

Binary uncertainty is summarized separately for each environment, sample size, and method by a two-sided exact 95% Clopper–Pearson interval. For example, $0/20$ has interval $[0, 0.1684]$, and $20/20$ has interval $[0.8316, 1]$. Thus zero observed failures do not certify population failure below

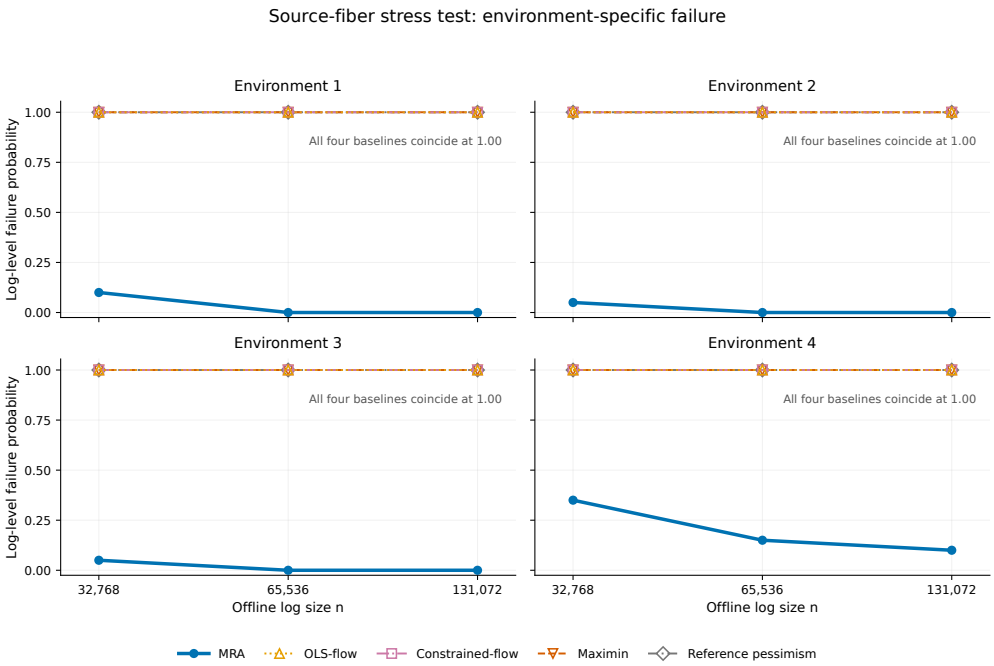


Figure 5: **Source-fiber failure in all four environments.** Each environment uses 20 independent logs per sample size. A log fails when its maximum regret across 32 held-out batches and six evaluator-only source-indistinguishable completions exceeds 0.1. All methods use the same logs and batches. The four comparator curves coincide at one. Environment 1 is also shown in Figure 2. The panels are separate environment-specific estimates.

$\delta = 0.05$. In Environment 4 at $n = 131{,}072$, MRA's $2/20$ interval is $[0.0123, 0.3170]$. These intervals are pointwise, not simultaneous over the grid.

For continuous loss, the machine-readable summaries include the mean, median, sample standard deviation, standard error, quartiles, and an approximate two-sided 95% Student-t interval for the mean. Writing s_n for the sample standard deviation of the 20 values $T_{n,r}$, the interval is $\overline{T}_n \pm t_{0.975,19} s_n/\sqrt{20}$, without clipping. This approximate interval concerns the mean of the log-level maximum, not its failure probability at $\epsilon = 0.1$. In particular, Environment 4 at $n = 32{,}768$ has $\overline{T}_n = 0.098604 < 0.1$ but seven failed logs. A mean below the reference line in Figure 6 does not imply that every log meets the target. The nonmonotone continuous means across some sample sizes are retained as observed.

All 48 differences between comparator and MRA means are positive. They range from 0.044966 to 0.072984. These are descriptive contrasts within paired cells, not 48 independent replications of a universal superiority claim. The source-fiber study concerns the four fixed context constructions and their finite batch collections. It does not establish performance on the original benchmark reward scores or on arbitrary deployment environments.

E.6 Implementation and result checks

The experiments use Python 3.12.5, NumPy 2.1.3, and SciPy 1.14.1 with the fixed solver settings and one numerical thread per worker. The final source-fiber extension preserves the original 150 logs and policies and adds exactly 90 predeclared logs. An independent saved-result verifier checks all 240 log inputs, 38,400 method–batch outputs, 1,200 log-level maxima, and 60 method–environment–sample-size summaries. It checks pairing, evaluator-only information, finite mixtures, strict failure indicators, and numerical aggregation without calling the learner or an optimization solver. Previously verified Rare and Coupled results are reused unchanged. The supplied result tables retain all settings and unfavorable outcomes, along with the distinct target-stop, gap-stop, and exceptional-fallback diagnostics.

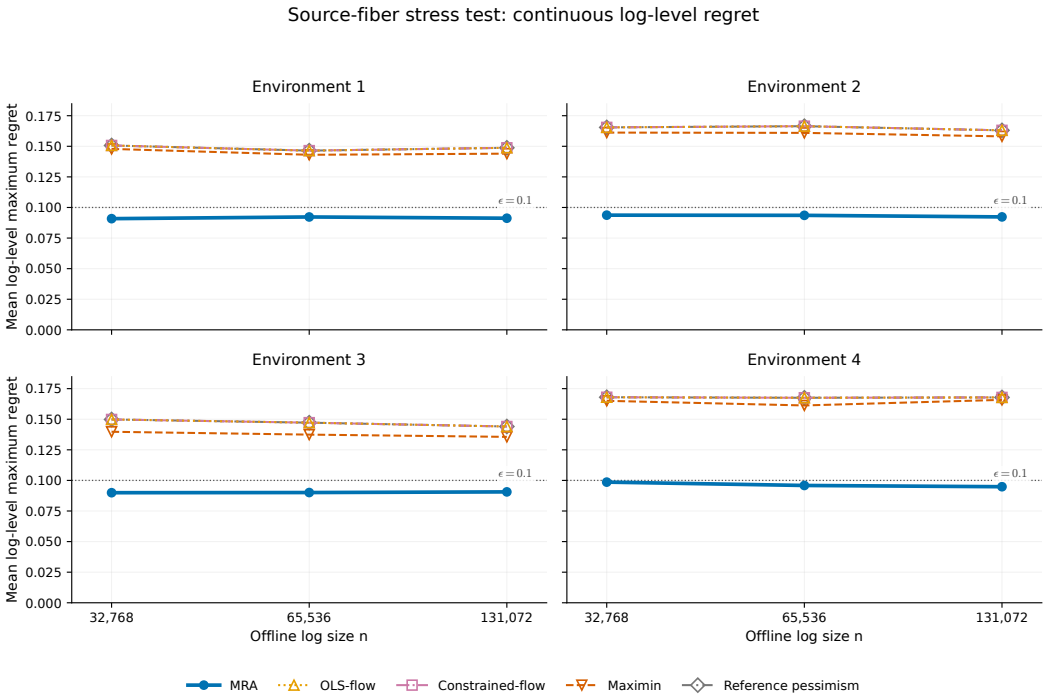


Figure 6: **Continuous source-fiber regret in all four environments.** Each point is the mean of 20 log-level maxima $T_{n,r}$ from (61), not a mean over batches before maximization. The same logs and batches are used as in Figure 5. The dotted line marks the individual-log failure threshold $\epsilon = 0.1$. A mean below the line does not imply that every log succeeds. Uncertainty calculations are described in Appendix E.5.