WHEN ARE FEATURE CROSSES DISCOVERABLE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Why can explicit feature crosses help models that can represent interactions? Expressive capacity does not ensure that finite data suffice to fit and certify a useful additive correction to a trained predictor. We formalize this gap as *checkpoint-local discoverability*, separating residual fitting from post-selection certification. A local risk expansion connects one-step utility to curvature-weighted residual prediction. In a fixed-design Gaussian model, we derive a sufficient exposure–cost frontier for linear smoothers that balances accessible signal, noise fitting, and search complexity. For nested Gaussian projection learners that fully expose the same signal, refinement strictly lowers the probability of positive gain by adding noise-only directions. We introduce CROSSAUDIT, a sample-split audit that selects the pre-validation update with the largest positive simultaneous lower confidence bound and otherwise returns a null update. For bounded logit updates under logistic loss, every positive certificate guarantees exact conditional population-risk improvement at the prescribed confidence level despite candidate selection. Controlled synthetic and semi-synthetic CTR experiments reproduce the refinement effect: atomic representations can lose utility while preserving the target function, whereas hierarchical sharing retains the coarse signal. Real-label audits show that pair-indexed utility need not be interaction-only: main-effect repairs can be stronger, while the tested post-fit residualized pairs receive no positive certificates.

1 INTRODUCTION

Explicit feature crosses remain common in tabular prediction even when the base model can represent interactions. Architectures such as Wide & Deep, DeepFM, DCN, xDeepFM, AutoInt, and DCN-V2 learn interactions during end-to-end training (Cheng et al., 2016; Guo et al., 2017; Wang et al., 2017; Lian et al., 2018; Song et al., 2019; Wang et al., 2021). Their representational capacity leaves a finite-data question: which useful directions can a fitted model still fail to recover?

We study this question at a fixed trained predictor f. Given candidate representations and finite fitting and validation budgets, can an additive update z be estimated and certified to reduce risk? We call this *checkpoint-local discoverability*. It depends on the residual signal left by the checkpoint, the update representation and fitting rule, and the data available for estimation and validation. This perspective treats an explicit cross as a candidate repair to a particular predictor.

For categorical fields $C_1, \dots, C_F$, a cross indexed by $T \subseteq [F]$ is the joint variable $C_T = (C_j)_{j \in T}$. For example, crossing device type and country creates cells such as `(mobile, US)`. An atomic table assigns one coefficient to each joint state. A strict refinement $T' \supsetneq T$, such as additionally crossing browser, can represent every function of C_T, but divides coarse-cell support among more cells. Here support is the number of observations available for fitting a cell's update. Our atomic learner fits cells observed in D_2 and assigns zero to unseen cells; a hierarchical learner can reuse a coarse effect at an unseen refined state. Figure 1 shows the resulting contrast on real CTR covariates: atomic gains decline under refinement, while a shared coarse component preserves utility.

Discoverability has two stages. *Fitting* asks whether checkpoint residual signal can be extracted without incurring excessive estimation cost. *Certification* asks whether an independently evaluated update improves population risk after accounting for candidate search. We study fitting through linear residual learners and certification through independent validation.

We formulate feature-cross utility as checkpoint-local residual prediction and derive an exposure–minus–noise identity and, under Gaussian noise, a sufficient exposure–cost frontier for fixed-design

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

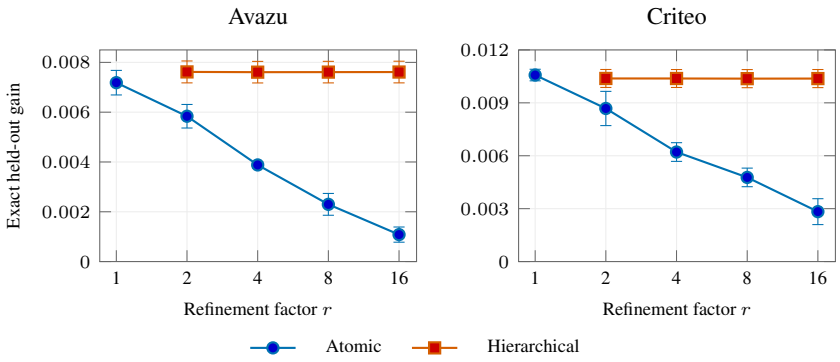

Figure 1: Held-out log-loss gains under refinement of a planted pair function. An independently generated categorical variable subdivides source-pair cells without changing the signal. Atomic refinement reduces utility, whereas hierarchical sharing preserves the coarse component. For each source pair and representation–refinement setting, the plotted point uses the non-null hyperparameter tuple with the largest D_3 LCB, computed under the common 318-tuple simultaneous correction. The independent reporting split D_4 provides its exact empirical gain even when that LCB is nonpositive. Error bars are sample standard deviations over three seed-level means, each averaging three source pairs. Certification counts and hyperparameter grids appear in Appendix E.

linear smoothers. We then prove that nested Gaussian projection learners exposing the same signal are strictly ordered by their probabilities of positive gain, and test the refinement mechanism with atomic and hierarchical estimators. Finally, CROSSAUDIT uses simultaneous lower confidence bounds and an explicit null decision to certify pre-validation updates with bounded exact gains. Joint real-label audits show that useful permissive pair repairs can reflect lower-order residual structure.

Related work. The checkpoint-local residual view connects to stagewise gradient boosting (Friedman, 2001), while the fitting analysis uses classical linear-smoother theory and functional ANOVA (Hastie et al., 2009; Hoeffding, 1948). Interaction screening and hierarchical regularization study related questions of candidate discovery and structured interaction modeling (Tsang et al., 2018; Bien et al., 2013). Recent tabular methods likewise emphasize the roles of data regime, inductive bias, and training design (Borisov et al., 2024; Gorishniy et al., 2025; Hollmann et al., 2025). For certification, empirical-Bernstein concentration and held-out risk control provide the statistical foundation (Maurer & Pontil, 2009; Angelopoulos et al., 2025; 2024). CROSSAUDIT specializes these tools to checkpoint-relative loss gains of updates fixed before validation, while our fitting results characterize how representation granularity changes residual-estimation cost.

2 A CHECKPOINT VIEW OF FEATURE CROSSES

Let $(X, Y) \sim P$, let $f : \mathcal{X} \to \mathbb{R}$ be fixed, and define $R(f) = \mathbb{E}[\ell(f(X), Y)]$. The checkpoint's negative score, conditional score, curvature, and Newton residual are

$$s_f(X,Y) = -\,\partial_u \ell(u,Y)\big|_{u=f(X)}, \qquad g_f(x) = \mathbb{E}[s_f(X,Y) \mid X = x],$$
$$a_f(x) = \mathbb{E}\Big[\partial_u^2 \ell(u,Y)\big|_{u=f(x)} \,\Big|\, X = x\Big], \quad r_f(x) = \frac{g_f(x)}{a_f(x)}.$$

Assume $a_f(x) > 0$ almost surely, and write $\langle u, v\rangle_f = \mathbb{E}[a_f(X)u(X)v(X)]$ and $\|u\|_f^2 = \langle u, u\rangle_f$.

Lemma 2.1 (Local residual identity)**.** *Assume that $\ell(u, y)$ is three times differentiable in u, all displayed expectations are finite, and*

$$|\partial_u^3 \ell(u, Y)| \le L_3$$

almost surely along the line segment between $f(X)$ and $f(X) + z(X)$. Then every fixed measurable update z satisfies

$$R(f) - R(f+z) = \langle r_f, z\rangle_f - \frac{1}{2}\|z\|_f^2 + \mathrm{Rem}_3(z),$$

where

$$|\mathrm{Rem}_3(z)| \le \frac{L_3}{6}\mathbb{E}|z(X)|^3.$$

The local gain is therefore alignment with the checkpoint residual minus curvature-weighted step cost, $\langle r_f, z\rangle_f - \frac{1}{2}\|z\|_f^2$, up to a cubic remainder. For squared loss the identity is exact and $r_f(x) = \mathbb{E}[Y \mid X = x] - f(x)$; for logistic loss it yields the usual Newton response and curvature weights. Thus the relevant object is not an interaction contained abstractly in the model class, but residual structure that a finite-sample learner can estimate strongly enough to offset noise and step cost.

3 Discoverability balances exposure and cost

Condition on $X_{1:n}$ and write the residual-response model as $y = \mu + \varepsilon$, where $\mathbb{E}[\varepsilon \mid X_{1:n}] = 0$ and $\mathrm{Cov}(\varepsilon \mid X_{1:n}) = \Sigma$. A linear residual learner returns $u = By$. For fixed $B, C, \Sigma \in \mathbb{R}^{n\times n}$ with $C, \Sigma \succeq 0$, define

$$\mathcal{G}_C(u;\mu) = \frac{1}{n}\left(\mu^\top u - \frac{1}{2}u^\top C u\right).$$

Proposition 3.1 (Exposure-minus-noise identity)**.** *Conditional on $X_{1:n}$, suppose B, C, μ, Σ are fixed. Then*

$$\mathbb{E}[\mathcal{G}_C(By;\mu) \mid X_{1:n}] = \frac{1}{2n}\left[\mu^\top H_{B,C}\mu - \mathrm{tr}(CB\Sigma B^\top)\right],$$

where

$$H_{B,C} = B + B^\top - B^\top C B.$$

The identity separates exposed signal, $\mu^\top H_{B,C}\mu$, from absorbed noise, $\mathrm{tr}(CB\Sigma B^\top)$. Expressiveness can increase both; its finite-sample value depends on their difference rather than on function-class inclusion alone.

3.1 A prior-coded exposure–cost frontier

Fix $0 < \delta \le 1/2$. Let

$$y = \mu + \sigma\xi, \qquad \sigma > 0, \qquad \xi \sim \mathcal{N}(0, I_n),$$

and let $\{A_T\}$ be a finite or countable family of fixed symmetric contractions,

$$A_T = A_T^\top, \qquad 0 \preceq A_T \preceq I_n.$$

Assign weights $\pi(T) > 0$, fixed independently of ξ, such that $\sum_T \pi(T) \le 1$, and define

$$L_T = \log\frac{1}{\pi(T)\delta}, \qquad v_T = \mathrm{tr}(A_T^2), \qquad e_T^{\mathrm{gain}}(\mu) = \frac{\mu^\top(2A_T - A_T^2)\mu}{\|\mu\|^2},$$

with $e_T^{\mathrm{gain}}(0) = 0$. Writing $\rho^2 = \|\mu\|^2$, define the unnormalized gain

$$G_T(y;\mu) = n\mathcal{G}_I(A_T y;\mu) = \mu^\top A_T y - \frac{1}{2}y^\top A_T^2 y.$$

Its expectation is

$$\mathbb{E}[G_T(y;\mu)] = \frac{1}{2}\left(\rho^2 e_T^{\mathrm{gain}}(\mu) - \sigma^2 v_T\right),$$

where v_T is the learner's noise-fitting dimension.

Theorem 3.2 (Sufficient exposure–cost frontier)**.** *Let*

$$\Theta_T = v_T + \sqrt{v_T L_T} + L_T.$$

There is a universal constant $C_0 > 0$ such that, with probability at least $1 - \delta$, simultaneously for every T,

$$G_T(y;\mu) \ge \frac{1}{4}\rho^2 e_T^{\mathrm{gain}}(\mu) - C_0\sigma^2\Theta_T.$$

Consequently, for every $C_1 > 4C_0$,

$$\frac{\rho^2}{\sigma^2}e_T^{\mathrm{gain}}(\mu) \ge C_1\Theta_T \quad\Longrightarrow\quad G_T(y;\mu) > 0$$

on the same event.

The sufficient condition compares accessible signal-to-noise with effective dimension and search cost. Equivalently, define the oracle index

$$\mathsf{Disc}_T(\mu) = \frac{(\rho^2/\sigma^2)e_T^{\mathrm{gain}}(\mu)}{v_T + \sqrt{v_T L_T} + L_T}.$$

A candidate becomes easier to discover when it exposes more residual signal, fits fewer noise directions, or receives a smaller search penalty. This index is an analysis quantity because it depends on the unknown conditional signal μ. The fixed-design fitting result is separate from population-risk certification, which Section 5 handles using independent validation; Appendix B.6 connects the stages through an achieved population-gain margin.

3.2 Incremental exposure

Full exposure can overstate the contribution of a cross when its directions overlap with an intercept, main effects, or other checkpoint repairs. Let $A_{T|0}$ be the candidate smoother after residualization against a reference space $\mathcal{V}_0$. When it is a symmetric contraction in whitened coordinates, define

$$E_{T|0}(\mu) = \mu^\top(2A_{T|0} - A_{T|0}^2)\mu, \qquad v_{T|0} = \mathrm{tr}(A_{T|0}^2),$$

and

$$\mathsf{Disc}_{T|0}(\mu) = \frac{E_{T|0}(\mu)/\sigma^2}{v_{T|0} + \sqrt{v_{T|0}L_T} + L_T}.$$

These oracle quantities remain graded under dependent covariates. In the product-ANOVA setting below, exact projection yields a sharper zero–one exposure law. CrossAudit does not estimate these indices; it directly evaluates fitted updates on independent data.

4 Why atomic refinement can hurt

A strict refinement can represent every function of its parent while allocating observations across more independently fitted cells. The following model isolates this estimation cost. Its ANOVA assumption holds for independent fields with constant curvature, and more generally when the curvature-weighted measure preserves the product-space decomposition.

Assumption 4.1 (Product local ANOVA setting)**.** *The categorical fields admit an orthogonal functional-ANOVA decomposition in the local inner product:*

$$L^2(C_1, \dots, C_F) = \bigoplus_{S \subseteq [F]} U_S, \qquad \mathcal{V}_T = \bigoplus_{S \subseteq T} U_S,$$

where U_S is the pure interaction subspace indexed by S.

All spaces in this section use the curvature-weighted L^2 inner product. Let $\mathcal{V}_0$ be a closed reference subspace, let P_0 be its orthogonal projection, and let $P_{0,T}$ project onto $\mathcal{V}_0 + \mathcal{V}_T$. For $(I - P_0)h \neq 0$, define deterministic incremental exposure

$$e_{T|0}^{\mathrm{det}}(h) = \frac{\|(P_{0,T} - P_0)h\|_f^2}{\|(I - P_0)h\|_f^2}.$$

We write $u \perp_f \mathcal{V}$ when $\langle u, v\rangle_f = 0$ for every $v \in \mathcal{V}$.

Theorem 4.2 (Representation-relative projection comparison)**.** *Under Assumption 4.1, let $0 \neq h_S \in U_S$ and $h_S \perp_f \mathcal{V}_0$. Then*

$$e_{T|0}^{\mathrm{det}}(h_S) = \mathbf{1}\{S \subseteq T\}.$$

Thus the source cross and every strict refinement fully expose the same pure interaction.

In the Gaussian fixed-design model, let $P_T \preceq P_{T'}$ be nested orthogonal projections of ranks $d < d'$ satisfying $P_T\mu = P_{T'}\mu = \mu \neq 0$. With $\kappa = \|\mu\|^2/\sigma^2$, their gains can be coupled as

$$G_T = \frac{\sigma^2}{2}(\kappa - U), \qquad G_{T'} = \frac{\sigma^2}{2}(\kappa - U - V),$$

where

$$U \sim \chi^2_d, \qquad V \sim \chi^2_{d'-d}, \qquad U \perp V.$$

Therefore, for every finite $\kappa > 0$,

$$\Pr(G_T > 0) = F_{\chi^2_d}(\kappa) > F_{\chi^2_{d'}}(\kappa) = \Pr(G_{T'} > 0).$$

For these projection learners, refinement preserves signal exposure and adds the independent noise term V, reducing the probability of positive gain at every finite signal-to-noise level. Hierarchical sharing changes this tradeoff by retaining the coarse direction while shrinking fine-cell deviations. Representation cost therefore depends on how the learner uses the refined input. Quantile comparisons and a coarsening-adaptive complexity benchmark appear in Appendices B.4 and C.1.

5 CrossAudit: certifying local improvement

CrossAudit evaluates updates on independent validation. Simultaneous confidence bounds account for candidate selection, and a null option preserves the checkpoint when no update is certified.

5.1 Protocol

We use five splits: D_0 trains the checkpoint, D_1 generates candidate representations, D_2 fits updates, D_3 selects and certifies, and D_4 evaluates the selected update. Preprocessing, support rules, fitted updates, steps, bounds, and prior weights are fixed using D_0, D_1, D_2 before D_3 is evaluated. The selection is then locked before reporting on D_4. Let $\mathcal{F}_{\mathrm{pre}}$ be the sigma-field generated by D_0, D_1, D_2, the checkpoint, all pre-validation randomness, and the resulting candidates, updates, bounds, and prior weights. Conditional on $\mathcal{F}_{\mathrm{pre}}$, the audited update functions are fixed and D_3 remains independent validation data. Index an audited update by $j = (T, \lambda, \gamma)$, and let $\widetilde{z}_j$ denote its actual stepped and bounded form. For an independent draw $(X, Y) \sim P$, define the exact gain

$$D_j = \ell(f(X), Y) - \ell(f(X) + \widetilde{z}_j(X), Y).$$

The certificate gain function is chosen before validation. Its value D_j^{cert} equals D_j when the exact gain is bounded. Its conditional target is

$$\Delta_j^{\mathrm{cert}} = \mathbb{E}[D_j^{\mathrm{cert}} \mid \mathcal{F}_{\mathrm{pre}}].$$

Let $\widehat{\Delta}_j^{\mathrm{cert}}$ and $\widehat{V}_j$ denote its sample mean and unbiased sample variance on D_3. Algorithm 1 gives the full implementation.

Theorem 5.1 (Selected one-step certificate)**.** *Fix* $0 < \delta < 1$ *and condition on* $\mathcal{F}_{\mathrm{pre}}$*. Let* $\mathcal{J}$ *be a finite or countable* $\mathcal{F}_{\mathrm{pre}}$*-measurable tuple family. Suppose its weights and bounds are also* $\mathcal{F}_{\mathrm{pre}}$*-measurable and satisfy*

$$\pi_j > 0, \qquad \sum_{j \in \mathcal{J}} \pi_j \le 1.$$

Suppose D_3 *contains* $n_{\mathrm{val}} \ge 2$ *i.i.d. observations from* P*, independent of* $\mathcal{F}_{\mathrm{pre}}$*, and*

$$|D_j^{\mathrm{cert}}| \le B_j \qquad \textit{almost surely}$$

for every $j \in \mathcal{J}$*. Define*

$$b_j(L) = \sqrt{\frac{2\widehat{V}_j L}{n_{\mathrm{val}}}} + \frac{14 B_j L}{3(n_{\mathrm{val}} - 1)}, \qquad L_j = \log \frac{2}{\pi_j \delta}.$$

Then, with conditional probability at least $1 - \delta$*, simultaneously for all* $j \in \mathcal{J}$*,*

$$\Delta_j^{\mathrm{cert}} \ge \mathrm{LCB}_j := \widehat{\Delta}_j^{\mathrm{cert}} - b_j(L_j).$$

Replacing L_j *by*

$$L_j = \log \frac{4}{\pi_j \delta}$$

yields the simultaneous two-sided event

$$\left| \widehat{\Delta}_j^{\mathrm{cert}} - \Delta_j^{\mathrm{cert}} \right| \le b_j(L_j)$$

with conditional probability at least $1 - \delta$*.*

For the finite family evaluated in an audit, CROSSAUDIT selects the update with the largest positive LCB. If no LCB is positive, it returns the null update, whose gain and penalty are zero. The simultaneous event over the family in Theorem 5.1 covers the selected tuple. For logistic loss,

$$\ell(u, y) = \log(1 + e^u) - yu, \qquad y \in \{0, 1\},$$

the loss is 1-Lipschitz in u. Hence

$$\|\widetilde{z}_j\|_\infty \le B_j \quad \Longrightarrow \quad |D_j| \le B_j,$$

so a positive LCB certifies exact, unclipped conditional population-risk improvement. If D_j^{cert} is obtained by clipping the gain, the certificate applies only to that clipped population target and does not certify the original risk difference. The fitting and validation stages meet through a population-gain margin. For H uniformly weighted exact-gain updates, write $\Delta_j = \mathbb{E}[D_j \mid \mathcal{F}_{\mathrm{pre}}]$ and let

$$L = \log \frac{4H}{\delta}, \qquad q_j = B_j \left[\sqrt{\frac{2L}{n_{\mathrm{val}} - 1}} + \frac{14L}{3(n_{\mathrm{val}} - 1)} \right].$$

If fitting produces some candidate with $\Delta_j > 2q_j$ with probability at least p_{fit}, then CROSSAUDIT returns a positive certificate with probability at least $p_{\mathrm{fit}} - \delta$. Thus fitting must create a population margin, while validation must resolve that margin against the update range and search size. Up to universal constants, a sufficient validation-size scaling is

$$n_{\mathrm{val}} \gtrsim \frac{B_j^2 \log(H/\delta)}{\Delta_j^2} + \frac{B_j \log(H/\delta)}{\Delta_j}.$$

Atomic tables assign zero to cells unseen in D_2, and all caps or global shrinkage factors are fixed before validation. The formal fit-to-certificate statement, full bounding formulas, and penalized oracle inequality appear in Appendix B.6, Appendix C.5, and Corollary B.3, respectively.

5.2 WHAT EXACTLY IS BEING AUDITED?

An *incremental audit* residualizes the candidate basis against a fixed reference space before fitting. Exact weighted projection gives an orthogonalized design on the fitting sample; ridge residualization provides an approximate removal of the reference directions. A *permissive audit* uses the original candidate basis. A pair table can then fit constituent main effects, intercept-like shifts, interactions, and correlated recalibration directions. Its certificate measures the utility of a pair-indexed repair. Section 6.5 compares intercept, main-effect, permissive-pair, post-fit residualized-pair, and shuffled updates within a joint audit. Repair type and checkpoint identity are included in the tuple index whenever they share a simultaneous budget.

6 EXPERIMENTS

We examine how exposure and fitting cost relate to update gain, how validation behaves as the candidate family grows, how refinement affects atomic and hierarchical estimators, and which repairs remain at real CTR checkpoints. For any update z, the empirical gain on split D is

$$\widehat{\Delta}_D(z) = \frac{1}{|D|} \sum_{(X_i, Y_i) \in D} \left[\ell(f(X_i), Y_i) - \ell(f(X_i) + z(X_i), Y_i) \right].$$

Split D_3 selects and certifies; D_4 is inspected only after selection is locked and is used for descriptive held-out evaluation.

6.1 CONTROLLED MECHANISM CHECKS

Table 1 compares oracle exposure diagnostics using the known planted signal. Accounting for fitting cost improves discrimination of positive test gain, while residualization isolates the direction left beyond the reference space. The source pair and its refinement have equal incremental exposure but different estimation costs. The logistic study also shows that exact and quadratic gains can disagree at large steps. CROSSAUDIT therefore evaluates exact held-out gains for bounded updates. Configurations and complete results appear in Appendix D.

Table 1: Controlled mechanism checks. Exposure quantities are oracle diagnostics computed from the known planted signal. AUC measures candidate-level discrimination between positive and nonpositive untouched-test update gains; it is not click-label AUC. The test split is used only for final evaluation. Values with ± are means and sample standard deviations over the corresponding seed sweeps.

Mechanism	Evaluation criterion	Quantitative evidence
Cost-aware exposure	Discrimination of positive untouched-test gain (AUC; higher is better)	**Exposure–cost: $\mathbf{0.940 \pm 0.018}$**; raw signal: 0.757 ± 0.043. **Improvement: $\mathbf{\Delta AUC = +0.183}$.**
Incremental exposure	Association with oracle residual gain (Spearman's ρ; higher is better)	**Incremental exposure: $\mathbf{1.000 \pm 0.000}$**; full exposure: 0.600 ± 0.062. **Improvement: $\mathbf{\Delta\rho = +0.400}$.**
Support fragmentation	Comparison of the source pair, its strict atomic refinement, and a wrong candidate	Exposure–cost: $\mathbf{379.5} \rightarrow 196.4$ under refinement; wrong candidate: 3.17. Untouched-test gain: $0.396/0.392/-3.71\times10^{-4}$, respectively. **Refinement reduces support efficiency by 48.2%.**
Local quadratic accuracy	Exact versus quadratic logistic gain (values reported as exact/quadratic)	At $\gamma = 1$: $0.1128/0.1077$. At $\gamma = 3$: $\mathbf{0.0060/-0.2345}$. **Large steps yield 60% sign disagreement.**

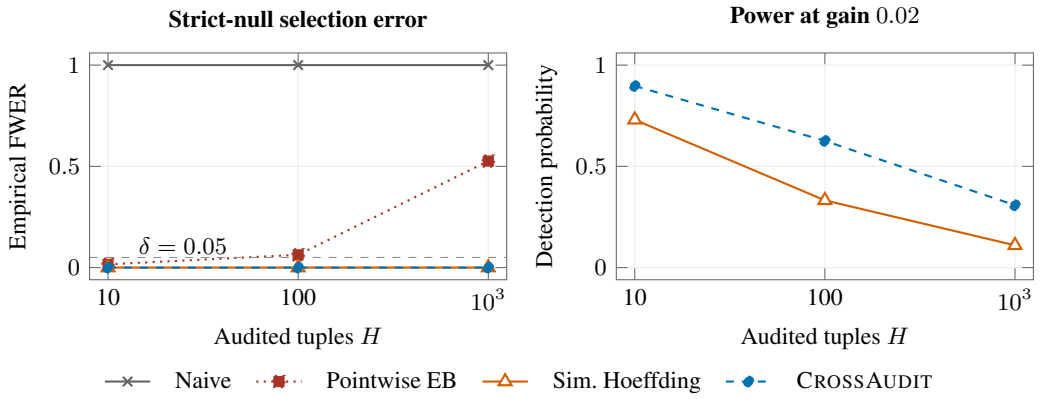

Figure 2: Family-wise error and detection power under candidate search. Each point uses 1000 repetitions with $n_{\text{val}} = 3000$ and $\delta = 0.05$. Neither simultaneous method produces a false certificate in these null trials. Under the tested alternative, empirical-Bernstein bounds yield higher power than simultaneous Hoeffding bounds.

6.2 Does validation survive candidate search?

We isolate the validation layer using pre-specified gains in $[-0.3, 0.3]$. Under the strict null all $H \in \{10, 100, 1000\}$ candidates have zero population gain; under the alternative one candidate has gain 0.02. Figure 2 reports 1000 repetitions per setting with $n_{\text{val}} = 3000$ and $\delta = 0.05$. Naive maximization fails in essentially every null repetition, and the pointwise empirical-Bernstein family-wise error rate (FWER) rises from 0.016 at $H = 10$ to 0.526 at $H = 1000$. Neither simultaneous method produces a false certificate in the tested repetitions (the one-sided 95% exact-binomial upper bound is 0.0030). At $H = 1000$, CrossAudit attains power 0.308, versus 0.110 for simultaneous Hoeffding bounds (Hoeffding, 1963). Full results, including $\delta = 0.1$, appear in Appendix D.3.

6.3 How do atomic and hierarchical estimators respond to fixed-signal refinement?

We use fixed 200K-row Avazu and Criteo samples (Avazu, 2014; Criteo Labs, 2014), split as 60K/30K/40K/30K/40K across $D_0, \ldots, D_4$. For each of seeds 3, 4, and 5, three source pairs are selected using only pre-validation support statistics. A random pair table is residualized against a weighted constituent-field additive fit, standardized, and planted at strength $\beta = 0.5$ on top of a real-label main-effect background. An independent $U \sim \text{Uniform}\{0, \ldots, 15\}$ defines the nested

Table 2: Common-candidate real-label audits over three seeds. Gains are exact empirical D_4 log-loss reductions of locked selections, reported as mean ± sample standard deviation; null selections contribute zero.

Dataset	Checkpoint	Certified	D_4 gain
Avazu	Main logistic	3/3	$2.48\times10^{-4} \pm 4.93\times10^{-5}$
	FM	3/3	$5.02\times10^{-4} \pm 1.89\times10^{-4}$
	Embedding MLP	3/3	$1.30\times10^{-3} \pm 6.56\times10^{-4}$
Criteo	Main logistic	1/3	$6.00\times10^{-5} \pm 1.04\times10^{-4}$
	FM	1/3	$4.91\times10^{-5} \pm 8.51\times10^{-5}$
	Embedding MLP	3/3	$8.02\times10^{-4} \pm 5.26\times10^{-4}$

refinements $R_r = U \bmod r$, $r \in \{1, 2, 4, 8, 16\}$. Thus every representation (A, B, R_r) contains the same planted function $h(A, B)$; refinement changes support rather than representability.

We compare atomic tables with the hierarchical update $z = u(A, B) + v(A, B, R_r)$. For checkpoint probability $p_i = \mathrm{sigmoid}(f(X_i))$, define

$$w_i = \max\{p_i(1 - p_i), 10^{-4}\}, \qquad r_i = \frac{Y_i^{(\beta)} - p_i}{w_i}.$$

The hierarchical update is fitted on D_2 by

$$\min_{u,v} \sum_{i \in D_2} w_i \left(r_i - u_{A_i B_i} - v_{A_i, B_i, R_r(i)}\right)^2 + \lambda_u \|u\|_2^2 + \lambda_v \|v\|_2^2 .$$

Within each source-pair setting, all 318 representation–regularization–step and control tuples share one simultaneous family with $\delta = 0.1$. Complete signal construction, grids, and support rules appear in Appendix E. Figure 1 shows declining atomic gains as the representation is refined. From $r = 1$ to $r = 16$, mean D_4 gain falls from 0.00718 to 0.00108 on Avazu and from 0.01058 to 0.00283 on Criteo. Over the same range, the mean number of retained refined cells rises from 1,917 to 12,194 and from 2,099 to 13,299, respectively. Across $r \in \{2, 4, 8, 16\}$, hierarchical gain remains near 0.00761 on Avazu and 0.01038 on Criteo. At $r = 16$, the paired hierarchical-minus-atomic gain differences are 0.00653 ± 0.00074 and 0.00754 ± 0.00037, respectively.

At $r = 16$, hierarchical updates receive positive certificates in all 18 dataset–seed–source-pair settings. Atomic updates are certified in $7/9$ Avazu settings and $9/9$ Criteo settings. No shuffled or support-matched wrong control is certified in any of the 18 source-pair settings.

For a fitted component g, define

$$\|g\|_W^2 = \sum_{i \in D_2} w_i g(X_i)^2.$$

For the D_3-selected hierarchical updates at $r = 16$, the mean fitted energy ratio

$$\frac{\|v\|_W^2}{\|u\|_W^2 + \|v\|_W^2}$$

is below 0.2%, so the fitted fine component has little weighted energy relative to the coarse component in this parameterization. Because u and v need not be orthogonal, the ratio is not a decomposition of prediction variance or gain. Complete curves appear in Table 8.

6.4 Which repairs remain at real checkpoints?

We audit real CTR labels at main-effect logistic, factorization-machine (Rendle, 2010), and embedding-MLP checkpoints using locked common sets of 42–54 field pairs generated from D_1. All bounded pair updates and shuffled controls are fixed before D_3 and share a simultaneous budget with $\delta = 0.1$. Because the primary audit is permissive, a positive certificate establishes a pair-indexed repair rather than an interaction-only effect. Across 18 dataset–checkpoint–seed audits, CrossAudit certifies 14 updates, all with positive empirical D_4 gain, while every shuffled control returns null. All Avazu checkpoints are corrected; on Criteo, main-effect and FM checkpoints are each certified in one of three seeds, whereas the MLP is certified in all three. MLP checkpoints have higher base log-loss and larger remaining repair gains, consistent with more residual structure at these fitted predictors. Full configurations, base metrics, and paired comparisons appear in Appendix F and Tables 10–11.

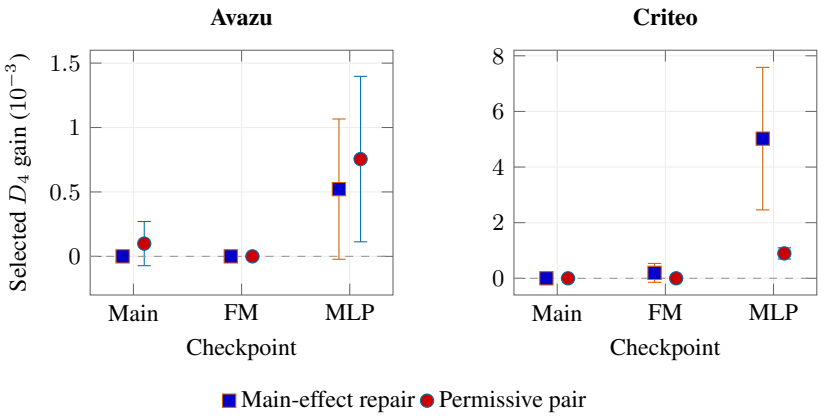

Figure 3: Family-specific selections for main-effect and permissive-pair repairs under the joint audit. Within each repair family, selection uses its largest positive validation LCB or the null, with the common simultaneous correction. Points are mean null-inclusive D_4 gains over report seeds 101, 503, and 4099; error bars show sample standard deviations. These are family-specific selections, not a decomposition of the global selection across repair families. The panels use different vertical ranges. The other repair families and certification counts are reported in Appendix H.

6.5 Pair repair or lower-order repair?

A permissive pair table can fit lower-order residuals as well as interactions. We compare intercept, main-effect, permissive-pair, post-fit residualized-pair, and shuffled updates using report seeds $101, 503, 4099$. For each fitted permissive pair, the residualized update subtracts its weighted projection onto the all-main-effect space on D_2. All repair types and checkpoints share one two-sided simultaneous budget within each dataset–seed setting, with $\delta = 0.1$. Figure 3 compares the main-effect and permissive-pair selections. The global policy selects four main-effect repairs, four permissive pairs, and ten null updates; every non-null selection has positive empirical D_4 gain. At the Criteo MLP, mean null-inclusive gain is $(5.022 \pm 2.562) \times 10^{-3}$ for main effects and $(0.894 \pm 0.201) \times 10^{-3}$ for permissive pairs. For seed 503, the simultaneous intervals place a lower bound of 9.08×10^{-4} on the selected main-effect repair's population-gain advantage over the selected pair.

Post-fit residualized pairs and shuffled controls select null in all 18 audits. Matched comparisons also show that removing the fitted main-effect component eliminates most of the selected pairs' held-out utility. These results locate much of the observed pair-repair gain in lower-order residual structure. Certification counts, family-level gains, and matched comparisons appear in Tables 12, 13, and 14.

7 Conclusion

Feature-cross utility is checkpoint-local, depending on residual signal and update-estimation cost. Our prior-coded exposure–cost frontier makes accessible signal, noise-fitting dimension, and candidate-search cost explicit. For nested Gaussian projections exposing the same signal, refinement strictly reduces the probability of positive gain. Semi-synthetic CTR experiments show the corresponding finite-sample pattern: atomic refinement can lose utility as support fragments, while hierarchical sharing retains the coarse signal. CrossAudit uses simultaneous held-out bounds and a null option; for bounded logit updates, positive certificates guarantee conditional population-risk improvement at the prescribed confidence level despite candidate selection. In controlled search experiments, it controls selection-induced error while showing higher power than the tested simultaneous Hoeffding baseline. Across real-label checkpoint audits, certified pair-indexed repairs achieve positive empirical held-out gains, illustrating the value of auditing a fitted predictor rather than relying on expressivity alone. Real-label audits also indicate that permissive pair repairs can mainly capture lower-order residual structure, while the tested post-fit residualized pairs remain uncertified. Feature-cross evaluation should therefore ask, in order, what residual signal remains, what representation can estimate it at the available support, and what independent evidence certifies the resulting update.

AI USE STATEMENT

Generative-AI tools assisted manuscript editing, development of experimental code, and preparation of LaTeX figures. The authors are responsible for the scientific claims, implementation, and reported results.

ETHICS STATEMENT

A certified cross may still encode sensitive attributes or proxies and may fail under distribution shift. Statistical certification complements, but does not replace, governance and deployment review.

REPRODUCIBILITY STATEMENT

Appendices B–H contain proofs, implementation details, configurations, preprocessing, seeds, and additional results; the supplementary material provides scripts, locked candidate manifests, and configurations.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Avazu. Click-through rate prediction. Kaggle competition dataset, 2014. URL https://www.kaggle.com/c/avazu-ctr-prediction.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013. doi: 10.1214/13-AOS1096. URL https://doi.org/10.1214/13-AOS1096.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, 2024. doi: 10.1109/TNNLS.2022.3229161. URL https://doi.org/10.1109/TNNLS.2022.3229161.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pp. 7–10. Association for Computing Machinery, 2016. doi: 10.1145/2988450.2988454. URL https://doi.org/10.1145/2988450.2988454.

Criteo Labs. Display advertising challenge. Kaggle competition dataset, 2014. URL https://www.kaggle.com/c/criteo-display-ad-challenge.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451. URL https://doi.org/10.1214/aos/1013203451.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *The Thirteenth International Conference on Learning Representations*, 2025.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. DeepFM: A factorization-machine based neural network for CTR prediction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 1725–1731, 2017. doi: 10.24963/ijcai.2017/239. URL https://doi.org/10.24963/ijcai.2017/239.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2 edition, 2009. doi: 10.1007/978-0-387-84858-7. URL https://doi.org/10.1007/978-0-387-84858-7.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6. URL https://doi.org/10.1038/s41586-024-08328-6.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics—Simulation and Computation*, 18(3):1059–1076, 1989. doi: 10.1080/03610918908812806. URL https://doi.org/10.1080/03610918908812806.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395. URL https://doi.org/10.1214/aos/1015957395.

Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. xDeepFM: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1754–1763. Association for Computing Machinery, 2018. doi: 10.1145/3219819.3220023. URL https://doi.org/10.1145/3219819.3220023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pp. 115–124, 2009. URL https://www.learningtheory.org/colt2009/papers/012.pdf.

Christopher C. Paige and Michael A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software*, 8(1):43–71, 1982. doi: 10.1145/355984.355989. URL https://doi.org/10.1145/355984.355989.

Steffen Rendle. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pp. 995–1000. IEEE, 2010. doi: 10.1109/ICDM.2010.127. URL https://doi.org/10.1109/ICDM.2010.127.

Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. AutoInt: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1161–1170. Association for Computing Machinery, 2019. doi: 10.1145/3357384.3357925. URL https://doi.org/10.1145/3357384.3357925.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByOfBggRZ.

Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*. Association for Computing Machinery, 2017. doi: 10.1145/3124749.3124754. URL https://doi.org/10.1145/3124749.3124754.

Ruoxi Wang, Rakesh Shivanna, Derek Z. Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed H. Chi. DCN V2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the Web Conference 2021*, pp. 1785–1797. Association for Computing Machinery, 2021. doi: 10.1145/3442381.3450078. URL https://doi.org/10.1145/3442381.3450078.

A Appendix overview

Appendix B contains proofs and theoretical extensions; Appendix C gives implementation details and CrossAudit pseudocode; Appendices D and E report controlled and real-covariate refinement experiments; Appendix F describes CTR data, checkpoints, and common-candidate audits; Appendix G reports compute and reproducibility details; Appendix H gives joint-repair and residualization analyses; and Appendix I states scope and limitations.

B Proofs

B.1 Proof of Lemma 2.1

Proof. For fixed (X, Y), Taylor's theorem gives

$$\ell(f(X)+z(X),Y) = \ell(f(X),Y) + \partial_u \ell(f(X),Y)z(X) + \frac{1}{2}\partial_u^2 \ell(f(X),Y)z(X)^2 + \mathrm{rem}_3(X,Y),$$

where

$$|\,\mathrm{rem}_3(X,Y)| \le \frac{L_3}{6}|z(X)|^3.$$

Taking expectations and subtracting from $R(f)$ yields

$$R(f) - R(f+z) = \mathbb{E}[-\partial_u \ell(f(X),Y)z(X)] - \frac{1}{2}\mathbb{E}[\partial_u^2 \ell(f(X),Y)z(X)^2] - \mathbb{E}[\mathrm{rem}_3(X,Y)].$$

Since $z(X)$ is measurable with respect to X,

$$\mathbb{E}[-\partial_u \ell(f(X),Y)z(X)] = \mathbb{E}[g_f(X)z(X)] = \langle r_f, z\rangle_f,$$

and similarly

$$\mathbb{E}[\partial_u^2 \ell(f(X),Y)z(X)^2] = \|z\|_f^2.$$

Setting $\mathrm{Rem}_3(z) = -\mathbb{E}[\mathrm{rem}_3(X,Y)]$ proves the identity and the stated remainder bound. □

B.2 Proof of Proposition 3.1

Proof. Using $u = By$ and $y = \mu + \varepsilon$,

$$\mathbb{E}[\mu^\top By \mid X_{1:n}] = \mu^\top B\mu.$$

Moreover,

$$\mathbb{E}[y^\top B^\top CBy \mid X_{1:n}] = \mu^\top B^\top CB\mu + \mathrm{tr}(B^\top CB\Sigma).$$

Therefore,

$$\mathbb{E}[\mathcal{G}_C(By;\mu) \mid X_{1:n}] = \frac{1}{n}\left[\mu^\top B\mu - \frac{1}{2}\mu^\top B^\top CB\mu - \frac{1}{2}\,\mathrm{tr}(CB\Sigma B^\top)\right].$$

Using $\mu^\top B\mu = \frac{1}{2}\mu^\top(B+B^\top)\mu$ gives the result. □

B.3 Proof of Theorem 3.2

Proof. Fix T and suppress its subscript where unambiguous. Substituting $y = \mu + \sigma\xi$ gives

$$G_T(y;\mu) = \mu^\top A(\mu+\sigma\xi) - \frac{1}{2}(\mu+\sigma\xi)^\top A^2(\mu+\sigma\xi)$$
$$= \frac{1}{2}\mu^\top(2A-A^2)\mu - \frac{\sigma^2}{2}\xi^\top A^2\xi + \sigma\mu^\top(A-A^2)\xi.$$

Adding and subtracting $\frac{\sigma^2}{2}\operatorname{tr}(A^2)$ yields

$$G_T = \frac{1}{2}\rho^2 e_T^{\text{gain}}(\mu) - \frac{\sigma^2}{2}v_T + Q_T + Z_T,$$

where

$$Q_T = -\frac{\sigma^2}{2}\left(\xi^\top A^2\xi - \operatorname{tr}(A^2)\right), \qquad Z_T = \sigma\mu^\top(A-A^2)\xi.$$

Because $0 \preceq A \preceq I_n$,

$$\left\|A^2\right\|_{\text{op}} \le 1, \qquad \operatorname{tr}(A^4) \le \operatorname{tr}(A^2) = v_T.$$

The one-sided Laurent–Massart inequality (Laurent & Massart, 2000) gives, for every $x > 0$,

$$Q_T \ge -\sigma^2\left(\sqrt{v_T x} + x\right)$$

with probability at least $1 - e^{-x}$.

The variable Z_T is centered Gaussian with variance

$$\operatorname{Var}(Z_T) = \sigma^2\mu^\top(A-A^2)^2\mu.$$

For every $a \in [0,1]$, $(a-a^2)^2 \le 2a - a^2$, so functional calculus gives

$$\operatorname{Var}(Z_T) \le \sigma^2\mu^\top(2A-A^2)\mu = \sigma^2\rho^2 e_T^{\text{gain}}(\mu).$$

The one-sided Gaussian tail bound and $2ab \le a^2 + b^2$ imply

$$Z_T \ge -\frac{1}{4}\rho^2 e_T^{\text{gain}}(\mu) - 2\sigma^2 x$$

with probability at least $1 - e^{-x}$. Combining the two events gives

$$G_T \ge \frac{1}{4}\rho^2 e_T^{\text{gain}}(\mu) - \sigma^2\left(\frac{1}{2}v_T + \sqrt{v_T x} + 3x\right)$$

with failure probability at most $2e^{-x}$.

Choose

$$x_T = \log\frac{2}{\pi(T)\delta} = L_T + \log 2.$$

Since $\delta \le 1/2$ and $\pi(T) \le 1$, we have $L_T \ge \log 2$ and hence $x_T \le 2L_T$. Thus a universal $C_0 > 0$ satisfies

$$G_T \ge \frac{1}{4}\rho^2 e_T^{\text{gain}}(\mu) - C_0\sigma^2\left(v_T + \sqrt{v_T L_T} + L_T\right).$$

Finally,

$$\sum_T 2e^{-x_T} = \delta\sum_T \pi(T) \le \delta,$$

so a union bound proves the simultaneous statement. If

$$\frac{\rho^2}{\sigma^2}e_T^{\text{gain}}(\mu) \ge C_1\Theta_T$$

for $C_1 > 4C_0$, then

$$G_T \ge \left(\frac{C_1}{4} - C_0\right)\sigma^2\Theta_T > 0.$$

□

B.4 Proof and quantitative extensions of Theorem 4.2

Corollary B.1 (Quantitative projection separation). *Under the Gaussian projection setting of Theorem 4.2, let $q_k(p)$ denote the p-quantile of χ^2_k. If $\alpha, \beta \in (0,1)$ satisfy*

$$q_d(1-\alpha) < q_{d'}(\beta),$$

then every

$$\kappa \in [q_d(1-\alpha), q_{d'}(\beta)]$$

obeys

$$\Pr(G_T > 0) \geq 1-\alpha, \qquad \Pr(G_{T'} > 0) \leq \beta.$$

Moreover, for any $C_1 > 4C_0$, if $\Theta_{T'} > \Theta_T$, then every

$$\kappa \in [C_1\Theta_T, C_1\Theta_{T'})$$

satisfies the sufficient frontier condition for T but not for T'. The latter comparison concerns sufficient conditions only.

Proof of Theorem 4.2 and Corollary B.1. Assumption 4.1 gives

$$\mathcal{V}_T = \bigoplus_{J \subseteq T} U_J.$$

If $S \subseteq T$, then $h_S \in U_S \subseteq \mathcal{V}_T$. Since $h_S \perp_f \mathcal{V}_0$,

$$P_0 h_S = 0, \qquad P_{0,T} h_S = h_S,$$

and therefore $e^{\det}_{T|0}(h_S) = 1$. If $S \not\subseteq T$, orthogonality of the ANOVA decomposition gives $h_S \perp_f \mathcal{V}_T$. Together with $h_S \perp_f \mathcal{V}_0$, this implies

$$P_{0,T} h_S = P_0 h_S = 0,$$

and hence $e^{\det}_{T|0}(h_S) = 0$.

For the Gaussian comparison, let P be an orthogonal projection of rank d satisfying $P\mu = \mu$. Since $P^2 = P$,

$$G_P(y;\mu) = \mu^\top P y - \frac{1}{2} y^\top P y.$$

Substituting $y = \mu + \sigma\xi$ and using $P\mu = \mu$ gives

$$G_P(y;\mu) = \frac{1}{2}\|\mu\|^2 - \frac{\sigma^2}{2}\|P\xi\|^2.$$

Because $\xi \sim \mathcal{N}(0, I_n)$, $\|P\xi\|^2 \sim \chi^2_d$, and thus

$$\Pr(G_P > 0) = F_{\chi^2_d}(\kappa).$$

For $P_T \preceq P_{T'}$, write $P_{T'} = P_T + Q$, where Q is an orthogonal projection of rank $d' - d$ whose range is orthogonal to that of P_T. Gaussian orthogonality gives

$$\|P_{T'}\xi\|^2 = \|P_T\xi\|^2 + \|Q\xi\|^2$$

with independent

$$U := \|P_T\xi\|^2 \sim \chi^2_d, \qquad V := \|Q\xi\|^2 \sim \chi^2_{d'-d}.$$

This yields

$$G_T = \frac{\sigma^2}{2}(\kappa - U), \qquad G_{T'} = \frac{\sigma^2}{2}(\kappa - U - V).$$

Both chi-square variables have positive densities on $(0,\infty)$, so for every finite $\kappa > 0$,

$$\Pr(U < \kappa \leq U + V) > 0.$$

Consequently,

$$F_{\chi^2_d}(\kappa) > F_{\chi^2_{d'}}(\kappa).$$

If

$$q_d(1-\alpha) \le \kappa \le q_{d'}(\beta),$$

continuity gives

$$F_{\chi^2_d}(\kappa) \ge 1-\alpha, \qquad F_{\chi^2_{d'}}(\kappa) \le \beta.$$

The strict quantile separation guarantees a nonempty interval.

Finally, for $A_T = P_T$ and $P_T\mu = \mu$,

$$e_T^{\mathrm{gain}}(\mu) = 1, \qquad v_T = \mathrm{tr}(P_T^2) = d.$$

Hence every

$$\kappa \in [C_1\Theta_T, C_1\Theta_{T'})$$

satisfies the sufficient frontier condition for T and fails it for T'. □

B.5 Proof of Theorem 5.1

Proof. Condition on $\mathcal{F}_{\mathrm{pre}}$. The tuple family, fitted updates, bounds, and weights are now fixed, while observations in D_3 remain i.i.d.

For a fixed j, the empirical-Bernstein inequality for a variable in $[-B_j, B_j]$ gives

$$\Delta_j^{\mathrm{cert}} \ge \widehat{\Delta}_j^{\mathrm{cert}} - \sqrt{\frac{2\widehat{V}_j L_j}{n_{\mathrm{val}}}} - \frac{14 B_j L_j}{3(n_{\mathrm{val}}-1)}$$

except on an event of probability at most $2e^{-L_j}$ (Maurer & Pontil, 2009). The factor $14B_j$ is seven times the range width $2B_j$.

For the one-sided result, choose

$$L_j = \log \frac{2}{\pi_j \delta}.$$

Then

$$\sum_{j\in\mathcal{J}} 2e^{-L_j} = \delta \sum_{j\in\mathcal{J}} \pi_j \le \delta,$$

so a union bound proves the simultaneous lower bounds.

For the two-sided result, apply the same inequality to D_j^{cert} and $-D_j^{\mathrm{cert}}$. The failure probability for a fixed j is at most $4e^{-L_j}$. Choosing

$$L_j = \log \frac{4}{\pi_j \delta}$$

and applying another union bound proves the two-sided event. Since the event is simultaneous, selecting a tuple after observing all validation statistics does not alter its validity. □

B.6 Fit-to-certificate bridge

Corollary B.2 (Fit-to-certificate bridge)**.** *Consider H exact-gain updates with uniform weights $1/H$. Let*

$$\Delta_j = \mathbb{E}[D_j \mid \mathcal{F}_{\mathrm{pre}}], \qquad |D_j| \le B_j,$$

and define

$$L = \log \frac{4H}{\delta}, \qquad q_j = B_j \left[\sqrt{\frac{2L}{n_{\mathrm{val}}-1}} + \frac{14L}{3(n_{\mathrm{val}}-1)} \right].$$

If the fitting stage produces some j satisfying

$$\Delta_j > 2q_j$$

with probability at least p_{fit}, then CrossAudit *returns a positive certificate with probability at least*

$$p_{\mathrm{fit}} - \delta.$$

Under Lemma 2.1, it is sufficient that

$$\langle r_f, \widetilde{z}_j \rangle_f - \frac{1}{2} \|\widetilde{z}_j\|_f^2 - \frac{L_3}{6} \mathbb{E}|\widetilde{z}_j(X)|^3 > 2q_j.$$

For $\Delta_j > 0$, a sufficient validation-size scaling is

$$n_{\text{val}} \gtrsim \frac{B_j^2 L}{\Delta_j^2} + \frac{B_j L}{\Delta_j}.$$

Proof. For observations in $[-B_j, B_j]$, the unbiased sample variance obeys

$$\widehat{V}_j \leq \frac{n_{\text{val}}}{n_{\text{val}} - 1} B_j^2.$$

Therefore,

$$b_j(L) \leq B_j \left[\sqrt{\frac{2L}{n_{\text{val}} - 1}} + \frac{14L}{3(n_{\text{val}} - 1)} \right] = q_j.$$

On the simultaneous two-sided event of Theorem 5.1,

$$\widehat{\Delta}_j \geq \Delta_j - b_j(L),$$

and hence

$$\widehat{\Delta}_j - b_j(L) \geq \Delta_j - 2b_j(L) \geq \Delta_j - 2q_j.$$

If fitting produces a tuple with $\Delta_j > 2q_j$, that tuple has a positive LCB, so maximization over the candidate family and the null must return a positive certificate.

Let $\mathcal{A}$ be the pre-validation event that fitting produces such a tuple. Conditional validity of Theorem 5.1 gives

$$\Pr(\text{CrossAudit certifies}) \geq \Pr(\mathcal{A}) - \delta \geq p_{\text{fit}} - \delta.$$

Lemma 2.1 gives

$$\Delta_j = \langle r_f, \widetilde{z}_j \rangle_f - \frac{1}{2} \|\widetilde{z}_j\|_f^2 + \text{Rem}_3(\widetilde{z}_j),$$

and the lower remainder bound proves the displayed sufficient condition. Solving the two terms of q_j for n_{val}, up to universal constants, gives the stated scaling. □

B.7 Penalized oracle inequality

Corollary B.3 (Penalized oracle inequality for CrossAudit)**.** *Let $\mathcal{J}$ be a finite audited family. Work on the simultaneous two-sided event of Theorem 5.1, and set*

$$\text{pen}_j = b_j(L_j)$$

for every non-null tuple. Let the null tuple $\varnothing$ have gain and penalty zero, and suppose

$$\widehat{j} \in \arg\max_{j \in \mathcal{J} \cup \{\varnothing\}} \left\{ \widehat{\Delta}_j^{\text{cert}} - \text{pen}_j \right\}.$$

Then

$$\Delta_{\widehat{j}}^{\text{cert}} \geq \max_{j \in \mathcal{J} \cup \{\varnothing\}} \left\{ \Delta_j^{\text{cert}} - 2\,\text{pen}_j \right\}.$$

Proof. On the simultaneous event,

$$\Delta_j^{\text{cert}} \geq \widehat{\Delta}_j^{\text{cert}} - \text{pen}_j, \qquad \widehat{\Delta}_j^{\text{cert}} \geq \Delta_j^{\text{cert}} - \text{pen}_j$$

for every tuple, including the null. By the definition of $\widehat{j}$,

$$\widehat{\Delta}_{\widehat{j}}^{\text{cert}} - \text{pen}_{\widehat{j}} \geq \widehat{\Delta}_j^{\text{cert}} - \text{pen}_j\,.$$

Combining these inequalities and maximizing over j proves the claim. □

C IMPLEMENTATION DETAILS

C.1 COARSENING-ADAPTIVE REPRESENTATION COST

A refined input need not force a high-dimensional estimator. For a refined cross T', let $\mathcal{R}(T')$ be a pre-specified collection of structured representations and define

$$\mathfrak{C}_{\mathcal{R}(T')}(\mu;\epsilon) = \inf_{\substack{m\in\mathcal{R}(T'):\\ E_{m|0}(\mu)\geq(1-\epsilon)E_{T'|0}(\mu)}} \left[v_{m|0} + \log\frac{1}{\pi(m)}\right], \qquad 0\leq\epsilon<1.$$

Each m specifies a residual learner with exposure $E_{m|0}$, noise-fitting dimension $v_{m|0}$, and pre-specified weight $\pi(m)>0$, where $\sum_m \pi(m)\leq 1$. An empty feasible set has value $+\infty$. This oracle benchmark separates input refinement from estimator complexity: a refined input remains inexpensive when a learner can find a low-complexity coarsening or sharing pattern, although the search still costs $\log(1/\pi(m))$.

C.2 RESIDUAL AND NEWTON FITTING

For squared loss, the residual response is

$$y_i = Y_i - f(X_i).$$

For logistic loss, write

$$f_i = f(X_i), \qquad p_i = \mathrm{sigmoid}(f_i), \qquad s_i = Y_i - p_i, \qquad a_i = p_i(1-p_i).$$

For numerical stability, let $\epsilon_{\mathrm{curv}}>0$ be a pre-specified curvature floor and use

$$w_i = \max\{a_i, \epsilon_{\mathrm{curv}}\}, \qquad r_i = \frac{s_i}{w_i}.$$

Let $Q_T(x)$ be a candidate basis. Ridge Newton fitting solves

$$\widehat{\beta}_{T,\lambda} = \arg\min_{\beta}\left\{\sum_{i\in D_2} w_i\big(r_i - Q_T(X_i)^\top\beta\big)^2 + \lambda\,\|\beta\|_2^2\right\},$$

and returns

$$\widehat{z}_{T,\lambda}(x) = Q_T(x)^\top\widehat{\beta}_{T,\lambda}.$$

For an atomic one-hot table, the objective separates across observed cells. If $\mathcal{I}_c$ indexes the D_2 observations in cell c, then

$$\widehat{\beta}_c = \frac{\sum_{i\in\mathcal{I}_c} s_i}{\sum_{i\in\mathcal{I}_c} w_i + \lambda}.$$

Cells absent from D_2 receive zero update. The curvature floor is 10^{-4} in the planted hierarchical experiment and 10^{-5} in the real-label audits.

C.3 RESIDUALIZATION FOR INCREMENTAL AUDITS

Let Q_0 be the reference basis and $W=\mathrm{diag}(w_i)$. For $\lambda_0\geq 0$, define

$$S_{0,\lambda_0} = Q_0\left(Q_0^\top W Q_0 + \lambda_0 I\right)^{-1} Q_0^\top W,$$

using a Moore–Penrose pseudoinverse when needed. The residualized candidate design is

$$\widetilde{Q}_T = (I - S_{0,\lambda_0})Q_T.$$

When $\lambda_0=0$, $S_{0,0}$ is the weighted orthogonal projection onto the column space of Q_0, and

$$Q_0^\top W\widetilde{Q}_T = 0.$$

For $\lambda_0>0$, it is a ridge smoother and the residualization is only approximate. The controlled incremental diagnostics use the unregularized pseudoinverse; the permissive real-label audits make no orthogonality claim.

C.4 Effective dimension

For the ridge smoother

$$A_T = W^{1/2} Q_T (Q_T^\top W Q_T + \lambda I)^{-1} Q_T^\top W^{1/2},$$

the effective noise-fitting dimension is

$$v_T = \operatorname{tr}(A_T^2).$$

When exact evaluation is expensive, we use Hutchinson estimation (Hutchinson, 1989). For a vector η with independent Rademacher coordinates,

$$\operatorname{tr}(A_T^2) = \mathbb{E}_\eta \left\| A_T \eta \right\|_2^2 .$$

C.5 Certificate variables and bounds

Exact and clipped targets. When exact gains are bounded by construction, we set

$$D_j^{\mathrm{cert}} = D_j.$$

Otherwise, a bound fixed before validation may define

$$D_j^{\mathrm{cert}} = \operatorname{clip}(D_j, -B_j, B_j).$$

The latter construction certifies only the clipped population quantity. Every reported logistic exact-gain certificate instead bounds the logit update itself.

Atomic updates. Let $\widehat{\beta}_{j,\mathrm{cell}}$ be fitted on D_2. Unseen cells receive zero update. Given nominal step γ and pre-specified pointwise cap c, define

$$\widetilde{z}_j(x) = \operatorname{clip}\left(\gamma \widehat{\beta}_{j,\mathrm{cell}(x)}, -c, c\right).$$

A pre-validation exact-gain bound is

$$B_j = \min\left\{ c,\ |\gamma| \max_{\text{cell observed in } D_2} |\widehat{\beta}_{j,\mathrm{cell}}| \right\},$$

where the maximum is zero if no cell is retained.

Implemented empirical-Bernstein penalty. For validation gains satisfying $|D_{ij}^{\mathrm{cert}}| \le B_j$, define

$$\widehat{\Delta}_j^{\mathrm{cert}} = \frac{1}{n_{\mathrm{val}}} \sum_{i \in D_3} D_{ij}^{\mathrm{cert}}, \qquad \widehat{V}_j = \frac{1}{n_{\mathrm{val}} - 1} \sum_{i \in D_3} \left(D_{ij}^{\mathrm{cert}} - \widehat{\Delta}_j^{\mathrm{cert}} \right)^2 .$$

For a uniform family of H tuples,

$$\mathrm{pen}_j = \sqrt{\frac{2\widehat{V}_j L}{n_{\mathrm{val}}}} + \frac{14 B_j L}{3(n_{\mathrm{val}} - 1)}.$$

One-sided experiments use

$$L = \log \frac{2H}{\delta},$$

while joint comparisons use

$$L = \log \frac{4H}{\delta}.$$

Tuple identities, prior weights, bounds, and δ are fixed before D_3 is inspected.

Structure-preserving shrinkage. If a pre-validation update z_j satisfies $\|z_j\|_\infty \le M_j$, global scalar shrinkage with $0 \le \eta_j \le 1$,

$$\widetilde{z}_j = \eta_j z_j,$$

gives the exact logistic gain bound

$$B_j = \eta_j M_j.$$

Unlike pointwise clipping, this construction preserves properties such as additivity and residualization.

Table 3: Configurations for the controlled mechanism studies.

Study	Data and mechanism	Learners	Key settings	Seeds
Exposure–cost scaling	Independent categorical fields with a centered order-3 residual	Ridge tables over candidate triples	$d = 10,\ q = 4$; $n_{\mathrm{tr}} = 4000$; $\lambda = 2,\ \sigma = 1$	10
Incremental exposure	Main effects plus a centered pair residual	Main-effect reference and candidate pairs	$d = 8,\ q = 4$; $\beta = 1.25,\ \sigma = 0.5$	10
Atomic-refinement cost	Pair residual with pair and triple candidates	Residualized ridge tables	Signal amplitudes: $0.4,\ 0.7,\ 1,\ 1.25$	10
Audit locality	Logistic labels with a planted interaction	Exact and quadratic validation gains	Step sizes: $0.1,\ 0.25,\ 0.5,\ 1,\ 2,\ 3$	5

C.6 Externally generated candidate lists

The certificate is conditional on the pre-validation generator and imposes no requirement on how it proposes or ranks candidates. In the CTR experiments, each checkpoint-specific generator runs on D_1. Its output is canonicalized into unordered field pairs and stored before update fitting. Every checkpoint in a common-candidate comparison audits the same union of these stored lists. The supplementary material contains the locked lists, generation scripts, support summaries, and deterministic tie-breaking configuration.

C.7 CrossAudit pseudocode

Algorithm 1 CrossAudit: sample-split residual cross audit

Require: Dataset D, base learner, candidate generator, prior weights π, ridge grid Λ, step grid Γ, and level δ

Split D into D_0, D_1, D_2, D_3, D_4
Train checkpoint f on D_0
Generate candidates $\mathcal{C}$ on D_1
Apply only pre-specified support rules based on D_1, D_2
for $T \in \mathcal{C}$ **do**
    Build Q_T, optionally residualized against a fixed reference basis
    for $\lambda \in \Lambda$ **do**
        Fit $\widehat{z}_{T,\lambda}$ on D_2
        for $\gamma \in \Gamma$ **do**
            Construct the stepped, bounded update $\widetilde{z}_{T,\lambda,\gamma}$
            Fix its gain bound using only pre-validation information
            Compute certificate gains and the simultaneous penalty on D_3
        end for
    end for
end for
Select the largest positive validation LCB, or the null if none is positive
Lock the selection
Evaluate the locked selection on D_4 for descriptive reporting

D Extended controlled experiments

D.1 Configurations

The synthetic experiments use d independent categorical fields with q uniformly sampled levels. The symbol k denotes planted interaction order, β signal amplitude, and σ Gaussian-noise standard deviation. Planted tables are centered against the lower-order reference space used by each diagnostic.

Table 4: Exposure–cost diagnostic on controlled data.

Metric	Mean ± sample standard deviation
AUC: exposure–cost ratio	0.940 ± 0.018
AUC: naive signal strength	0.757 ± 0.043
Spearman: ratio vs. test gain	0.425 ± 0.181
Spearman: validation gain vs. test gain	0.748 ± 0.095
Selected test gain	0.319 ± 0.009

Table 5: Post-selection stress test over 1000 repetitions. FWER is the frequency of at least one false certificate under the strict null; power is the detection frequency of the non-null candidate.

δ	H	Naive FWER	Pointwise FWER	Hoeffding FWER	CrossAudit FWER	Hoeffding power	CrossAudit power
0.05	10	1.000	0.016	0.000	0.000	0.730	0.897
0.05	100	1.000	0.063	0.000	0.000	0.332	0.626
0.05	1000	1.000	0.526	0.000	0.000	0.110	0.308
0.10	10	0.999	0.018	0.000	0.000	0.829	0.933
0.10	100	1.000	0.200	0.000	0.000	0.437	0.710
0.10	1000	1.000	0.918	0.000	0.000	0.175	0.404

D.2 Exposure–cost diagnostics

The exposure–cost ratio is markedly better than raw signal strength at discriminating candidates with positive untouched-test gain. Validation gain is more directly associated with test gain, as expected because it estimates the same target on independent data.

D.3 Full post-selection stress test

For candidate j and validation observation i, simulate

$$D_{ij} = \mu_j + 0.2\varepsilon_{ij}, \qquad \Pr(\varepsilon_{ij} = 1) = \Pr(\varepsilon_{ij} = -1) = \frac{1}{2},$$

independently across i and j. Under the strict null, $\mu_j = 0$ for every j; under the alternative, $\mu_1 = 0.02$ and $\mu_j = 0$ for $j > 1$. We use

$$n_{\mathrm{val}} = 3000, \qquad B = 0.3.$$

The bound B is conservative for the simulated support. Candidate independence is used only to construct this power experiment; Theorem 5.1 permits arbitrary dependence among candidate gains evaluated on the same validation observations.

For variables in $[-B, B]$, the simultaneous one-sided Hoeffding penalty is

$$\mathrm{pen}_{\mathrm{H}} = B\sqrt{\frac{2\log(H/\delta)}{n_{\mathrm{val}}}}.$$

Pointwise empirical Bernstein uses $L = \log(2/\delta)$ without correction for H and is included as a deliberately invalid post-selection baseline.

Zero false certificates in 1000 repetitions do not imply zero population FWER; the corresponding one-sided 95% exact-binomial upper confidence bound is approximately 0.0030.

D.4 Incremental exposure and support cost

When main effects are already available, incremental exposure has Spearman correlation 1.000 ± 0.000 with oracle residual gain, compared with 0.600 ± 0.062 for full exposure. The true pair and its strict atomic refinement both have unit incremental exposure, but the refinement has a smaller exposure–cost ratio because it fits more directions.

Table 6: Atomic-refinement support-cost diagnostic. Exposure and ratio entries are mean ± sample standard deviation; the final column is mean untouched-test gain.

Candidate	Incremental exposure	Exposure–cost ratio	Test gain
Wrong pair or triple	0.00128 ± 0.00078	3.17 ± 2.82	-3.71×10^{-4}
Strict atomic refinement	1.000 ± 0.000	196.4 ± 195.4	0.392
True pair	1.000 ± 0.000	379.5 ± 381.3	0.396

Table 7: Exact and quadratic logistic validation gains over five seeds. Absolute gap is the mean absolute discrepancy; sign disagreement is the fraction of seeds with different signs.

Step γ	Exact gain	Quadratic gain	Absolute gap	Sign disagreement
0.10	0.0191 ± 0.0010	0.0191 ± 0.0010	5.8×10^{-7}	0.0
0.25	0.0444 ± 0.0023	0.0443 ± 0.0023	2.3×10^{-5}	0.0
0.50	0.0774 ± 0.0042	0.0771 ± 0.0041	3.5×10^{-4}	0.0
1.00	0.1128 ± 0.0070	0.1077 ± 0.0067	5.2×10^{-3}	0.0
2.00	0.0934 ± 0.0103	0.0295 ± 0.0085	6.4×10^{-2}	0.0
3.00	0.0060 ± 0.0127	-0.2345 ± 0.0121	0.2405	0.6

D.5 How local is the quadratic approximation?

For logistic loss, exact and quadratic held-out gains agree closely at small steps. At larger steps the approximation drifts, and at $\gamma = 3$ it often predicts the wrong sign. Exact validation is therefore preferable whenever the update can be bounded.

E Real-covariate hierarchical recovery

E.1 Pair-indexed planted signal

For each seed, a pre-specified support rule based on statistics computed on $D_1 \cup D_2$ selects three source pairs. For a selected pair (A, B), draw

$$h_0(A, B) = \theta_{A,B}, \qquad \theta_{a,b} \sim \mathcal{N}(0, 1).$$

Categorical vocabularies are fixed using pre-validation information; levels outside them map to an `OTHER` state.

Let $D_{\text{pre}} = D_0 \cup D_1 \cup D_2$. We train a main-effect background model on the real labels in D_0 and denote its logit by $\eta_{\text{bg}}(X)$. Using stabilized curvature weights from this model, we fit $h_0(A, B)$ in

$$\mathcal{V}^{\text{add}}_{A,B} = \{c + u(A) + v(B)\}$$

over D_{pre}. The additive fit uses LSQR (Paige & Saunders, 1982), ridge parameter 10^{-8}, and at most 300 iterations. If $\widehat{m}(A, B)$ is the fitted additive function, define

$$h_{\text{raw}}(A, B) = h_0(A, B) - \widehat{m}(A, B),$$

and standardize it to weighted mean zero and weighted variance one on D_{pre}, producing $h(A, B)$. Under dependent covariates, this construction removes the targeted empirical constituent-field component but need not make h orthogonal to functions of every other field.

Conditional on the real covariates, generate labels independently as

$$Y^{(\beta)} \sim \text{Bernoulli}\left[\text{sigmoid}\left(\eta_{\text{bg}}(X) + \beta h(A, B)\right)\right], \qquad \beta = 0.5.$$

The audited checkpoint is a main-effect logistic model trained on D_0 using these semi-synthetic labels.

E.2 Nested refinements and fitted representations

Independently of the labels and planted table, sample

$$U_i \sim \text{Uniform}\{0, \ldots, 15\}$$

using a fixed pre-validation seed, and define

$$R_r(i) = U_i \bmod r, \qquad r \in \{1, 2, 4, 8, 16\}.$$

Because the refinement factors are powers of two, the partitions are nested. Every representation (A, B, R_r) contains the same planted function $h(A, B)$.

The atomic learner assigns an independent coefficient to each refined cell observed in D_2; unseen cells receive zero update. The hierarchical learner uses

$$z(A, B, R_r) = u(A, B) + v(A, B, R_r)$$

and solves

$$\min_{u,v} \sum_{i \in D_2} w_i \left(r_i - u_{A_i B_i} - v_{A_i, B_i, R_r(i)}\right)^2 + \lambda_u \|u\|_2^2 + \lambda_v \|v\|_2^2 .$$

The atomic grid is

$$\lambda \in \{0.1, 1, 3, 10, 30, 100\},$$

and the hierarchical grid is

$$\lambda_u \in \{0.1, 1, 10, 100\}, \qquad \frac{\lambda_v}{\lambda_u} \in \{0.1, 1, 10, 100\}.$$

Both use

$$\gamma \in \{0.25, 0.5, 1\}.$$

The five atomic refinement levels contribute $5 \times 6 \times 3 = 90$ tuples; four nontrivial hierarchical levels contribute $4 \times 4 \times 4 \times 3 = 192$; and support-matched wrong and shuffled controls contribute $2 \times 6 \times 3 = 36$. Thus

$$H = 90 + 192 + 36 = 318.$$

With uniform weights and $\delta = 0.1$,

$$L = \log \frac{2H}{0.1} = \log \frac{2 \cdot 318}{0.1} \approx 8.758.$$

Every unavailable component has a deterministic zero fallback. Before validation, we compute the exact supremum of each fitted update over the retained encoded domain; after stepping, this supremum is the logistic gain bound. No reported certificate clips the gain variable.

E.3 Complete refinement curves

No shuffled or support-matched wrong control is certified in any of the 18 source-pair settings. At $r = 16$, hierarchical sharing improves over atomic refinement by 0.00653 ± 0.00074 on Avazu and 0.00754 ± 0.00037 on Criteo, where $\pm$ denotes the sample standard deviation across three seed-level paired means, each averaging the three source-pair differences.

F CTR experimental details

F.1 Datasets, fixed samples, and preprocessing

We use the public Criteo Display Advertising Challenge and Avazu CTR datasets (Criteo Labs, 2014; Avazu, 2014). Every experiment operates on a fixed 200K-row sample. Deterministic sampling scripts, split seeds, and compact row-index metadata are included in the supplementary material; raw data are not redistributed.

Avazu contains 22 raw fields. The real-label audit replaces raw `hour` with day and hour-of-day, producing 23 audit fields; the planted experiment retains the original 22-field representation. Vocabularies and numerical bin edges use only D_0, D_1, D_2. Infrequent and unseen levels map to `OTHER`. The planted experiment treats every field as categorical and retains the 80 most frequent pre-validation levels. The real-label audit retains at most 1000 levels per categorical field, quantile-bins numerical Criteo fields into at most 32 bins, and decomposes Avazu `hour`. No threshold, candidate list, or support rule uses D_3 or D_4.

Table 8: Atomic and hierarchical refinement curves at $\beta = 0.5$. Within each representation–refinement subfamily, D_3 selects the non-null tuple with the largest LCB under the common 318-tuple penalty. Its D_4 gain is reported even when the LCB is nonpositive; certification counts are shown separately. Values are mean $\pm$ sample standard deviation over three seed-level means, each averaging three source pairs.

Dataset	Representation	r	D_4 gain	Validation LCB	Certified
Avazu	Atomic	1	0.00718 ± 0.00049	0.00400 ± 0.00026	$9/9$
	Atomic	2	0.00584 ± 0.00047	0.00281 ± 0.00027	$9/9$
	Atomic	4	0.00388 ± 0.00012	0.00162 ± 0.00024	$9/9$
	Atomic	8	0.00230 ± 0.00044	0.00075 ± 0.00021	$9/9$
	Atomic	16	0.00108 ± 0.00031	0.00027 ± 0.00016	$7/9$
	Hierarchical	2	0.00762 ± 0.00044	0.00392 ± 0.00023	$9/9$
	Hierarchical	4	0.00761 ± 0.00043	0.00391 ± 0.00023	$9/9$
	Hierarchical	8	0.00761 ± 0.00043	0.00391 ± 0.00024	$9/9$
	Hierarchical	16	0.00761 ± 0.00043	0.00392 ± 0.00023	$9/9$
Criteo	Atomic	1	0.01058 ± 0.00033	0.00605 ± 0.00063	$9/9$
	Atomic	2	0.00868 ± 0.00097	0.00466 ± 0.00066	$9/9$
	Atomic	4	0.00621 ± 0.00053	0.00327 ± 0.00064	$9/9$
	Atomic	8	0.00477 ± 0.00053	0.00205 ± 0.00050	$9/9$
	Atomic	16	0.00283 ± 0.00073	0.00097 ± 0.00047	$9/9$
	Hierarchical	2	0.01038 ± 0.00051	0.00592 ± 0.00076	$9/9$
	Hierarchical	4	0.01038 ± 0.00051	0.00591 ± 0.00076	$9/9$
	Hierarchical	8	0.01037 ± 0.00051	0.00592 ± 0.00076	$9/9$
	Hierarchical	16	0.01038 ± 0.00051	0.00591 ± 0.00076	$9/9$

Table 9: CTR samples and five-way split sizes. Raw/audit fields are the numbers of fields before and after experiment-specific encoding.

Dataset	Rows	Raw/audit fields	Click rate	$\lvert D_0\rvert$	$\lvert D_1\rvert$	$\lvert D_2\rvert/\lvert D_3\rvert/\lvert D_4\rvert$
Criteo	200K	39/39	0.250	60K	30K	40K/30K/40K
Avazu	200K	22/23	0.175	60K	30K	40K/30K/40K

F.2 Checkpoint training and common-candidate audits

The common-candidate experiment uses seeds $17, 271, 2027$. Every checkpoint is trained only on D_0. An internal 80/20 split of D_0 selects the epoch count by early stopping, after which the model is retrained for that number of epochs on all of D_0. Output biases are initialized to the empirical D_0 click-rate logit.

Each checkpoint-specific generator runs on D_1 and returns at most 40 canonicalized unordered field pairs. Their locked union is audited at every checkpoint and contains 42–54 source pairs across the reported runs. For each pair, the audit uses

$$\lambda \in \{0.001, 0.01, 0.1, 1, 10\}, \qquad \gamma \in \{0.125, 0.25, 0.5, 1\}.$$

Fitted logit updates are pointwise bounded by 2, with a tighter candidate-specific bound computed from the D_2-fitted coefficients when available. Original and shuffled-control tuples share a uniform simultaneous budget with $\delta = 0.1$ within each dataset–checkpoint–seed audit.

The shuffled control first removes the intercept-only Newton direction:

$$s_i^{\mathrm{ctrl}} = s_i - w_i \frac{\sum_{k \in D_2} s_k}{\sum_{k \in D_2} w_k}, \qquad w_i = \max\{p_i(1-p_i), 10^{-5}\}.$$

It then permutes s_i^{ctrl} over D_2 before fitting the pair table. Because $\sum_i s_i^{\mathrm{ctrl}} = 0$, permutation does not restore the removed global score direction.

F.3 Paired checkpoint diagnostics

Within each dataset and seed, every checkpoint audits the same candidate union. This permits candidate-level comparisons of D_3 LCB rankings and certification decisions; candidate-level D_4 outcomes play no role in ranking.

Table 10: Base checkpoint quality on untouched D_4, reported as mean $\pm$ sample standard deviation over seeds 17, 271, and 2027.

Dataset	Checkpoint	Log-loss	AUC
Avazu	Main logistic	0.40591 ± 0.00182	0.74139 ± 0.00233
	FM	0.40681 ± 0.00149	0.74172 ± 0.00200
	Embedding MLP	0.40967 ± 0.00197	0.73831 ± 0.00273
Criteo	Main logistic	0.47353 ± 0.00235	0.76559 ± 0.00052
	FM	0.47517 ± 0.00223	0.76541 ± 0.00064
	Embedding MLP	0.48440 ± 0.00509	0.76059 ± 0.00145

Table 11: Paired common-candidate diagnostics averaged over three report seeds. Only A/B is the mean fraction certified only at the indicated checkpoint.

Dataset	A	B	LCB Spearman	Agreement	Only A	Only B
Avazu	Main	FM	0.875 ± 0.066	0.765 ± 0.062	0.000	0.235
	Main	MLP	0.607 ± 0.286	0.349 ± 0.292	0.000	0.651
	FM	MLP	0.640 ± 0.340	0.584 ± 0.234	0.000	0.416
Criteo	Main	FM	0.868 ± 0.052	0.984 ± 0.027	0.008	0.008
	Main	MLP	0.862 ± 0.095	0.517 ± 0.443	0.000	0.483
	FM	MLP	0.763 ± 0.171	0.517 ± 0.443	0.000	0.483

Main-effect and FM checkpoints have similar candidate decisions on Criteo. The MLP certifies more candidates and has higher base log-loss, linking its larger repair gains to the residual structure of these fitted checkpoints.

G Reproducibility and compute

Controlled diagnostics use CPU workers. The strict-null, hierarchical-recovery, common-candidate, and low-order experiments use Python 3.10.19, PyTorch 2.9.1, and CUDA 12.8 on an NVIDIA RTX 5070 Ti laptop GPU with 15.2 GB of system memory. Most audit operations use CPU cell aggregation and sparse linear algebra; peak allocated PyTorch tensor memory remains below 0.1 GB.

The strict-null, hierarchical-recovery, and common-candidate suites take approximately 713.7 seconds in total:

strict-null stress test	2.0 s,
hierarchical recovery	90.7 s,
common-candidate audits	621.0 s.

The low-order experiments take 10764.6 seconds, including 3855 unregularized pair-update residualizations.

The hierarchical experiment uses seeds $3, 4, 5$; the common-candidate audits use $17, 271, 2027$; and the joint-repair audits use $101, 503, 4099$. Every real-covariate experiment uses

$$|D_0|/|D_1|/|D_2|/|D_3|/|D_4| = 60\text{K}/30\text{K}/40\text{K}/30\text{K}/40\text{K}.$$

The supplementary package includes preprocessing and candidate-generation scripts, sampling and split seeds, locked manifests, configuration files, run scripts, and compact seed-level logs.

H Low-order repair controls

H.1 Experimental setting

The joint-repair experiments use separate splits of the fixed Avazu and Criteo samples with report seeds

$$\{101, 503, 4099\}.$$

For report seed s, the split seed is $41000 + s$, and

$$|D_0|/|D_1|/|D_2|/|D_3|/|D_4| = 60\text{K}/30\text{K}/40\text{K}/30\text{K}/40\text{K}.$$

We audit main-logistic, FM, and embedding-MLP checkpoints. Preprocessing retains at most 1000 categorical levels, bins numerical Criteo fields into at most 32 quantile bins, and replaces Avazu `hour` with day and hour-of-day. The resulting representations contain 23 Avazu fields and 39 Criteo fields.

Checkpoint training uses AdamW (Loshchilov & Hutter, 2019), batch size 4096, weight decay 10^{-6}, at most 30 epochs, and early-stopping patience 5 on an internal 80/20 split of D_0. The learning rate is 10^{-2} for main logistic and 10^{-3} for FM and MLP. FM and MLP use embedding dimension 16; the MLP has hidden widths $(128, 64)$ and dropout 0.1. After epoch selection, each model is retrained on all of D_0.

Each checkpoint emits at most 40 pairs using the locked D_1 generator. Their canonicalized union is audited at every checkpoint. For seeds 101, 503, and 4099, the union sizes are $(46, 43, 43)$ on Avazu and $(41, 43, 41)$ on Criteo.

H.2 Repair families

For checkpoint logit $f_i = f(X_i)$, define

$$p_i = \mathrm{sigmoid}(f_i), \qquad s_i = Y_i - p_i, \qquad w_i = \max\{p_i(1-p_i), 10^{-5}\}, \qquad r_i = \frac{s_i}{w_i}.$$

Every update is fitted on D_2.

Intercept repair (I). The intercept update is

$$z_I(x) = \widehat{c}, \qquad \widehat{c} = \frac{\sum_{i \in D_2} s_i}{\sum_{i \in D_2} w_i}.$$

Main-effect repair (A). Let Q_0 contain an explicit intercept and full one-hot bases for all encoded fields. Fit

$$z_A(x) = Q_0(x)^\top \widehat{\alpha},$$

where

$$\widehat{\alpha} = \arg\min_{\alpha} \left\{ \sum_{i \in D_2} w_i \big(r_i - Q_0(X_i)^\top \alpha\big)^2 + \lambda_A \left\|\alpha_{-0}\right\|_2^2 \right\},$$

α_{-0} excludes the unpenalized intercept, and

$$\lambda_A \in \{0.1, 1, 10, 100, 1000\}.$$

Permissive pair repair (P). For source pair (A, B), define

$$z_P(x) = \widehat{\beta}_{A(x), B(x)},$$

where

$$\widehat{\beta}_{ab} = \frac{\sum_{i \in D_2 : (A_i, B_i) = (a,b)} s_i}{\sum_{i \in D_2 : (A_i, B_i) = (a,b)} w_i + \lambda_P}, \qquad \lambda_P \in \{0.001, 0.01, 0.1, 1, 10\}.$$

Cells absent from D_2 receive zero update.

Post-fit residualized pair repair (R). For each fitted permissive pair, solve

$$\widehat{a}_P = \arg\min_{a} \sum_{i \in D_2} w_i \left(z_P(X_i) - Q_0(X_i)^\top a\right)^2$$

without ridge regularization, and define

$$z_R(x) = z_P(x) - Q_0(x)^\top \widehat{a}_P.$$

The projection coefficients fitted on D_2 are reused unchanged on D_3 and D_4. This residualizes the fitted update rather than the full pair basis, so conclusions apply to this specific pipeline.

Shuffled control (S). Remove the intercept Newton direction,

$$s_i^{\mathrm{ctrl}} = s_i - w_i \frac{\sum_{k\in D_2} s_k}{\sum_{k\in D_2} w_k},$$

then permute these scores over D_2 and fit pair tables using the same ridge grid.

H.3 Bounded updates and joint selection

For an additive function

$$q(x) = a_0 + \sum_{t=1}^{F} a_t(C_t),$$

define

$$a_t^- = \min_c a_t(c), \qquad a_t^+ = \max_c a_t(c).$$

Then

$$M(q) = \max\left\{ \left| a_0 + \sum_t a_t^- \right|, \left| a_0 + \sum_t a_t^+ \right| \right\}$$

bounds $\sup_x |q(x)|$ over the encoded product domain.

For a pair table, let

$$M_P = \max_{(a,b)\in\mathcal{O}} |\widehat{\beta}_{ab}|,$$

where $\mathcal{O}$ is the retained D_2 support and coefficients outside $\mathcal{O}$ are zero. For $z_R = z_P - q$, define

$$b^- = \min\left\{ -a_A^+ - a_B^+,\ \min_{(a,b)\in\mathcal{O}} \left[\widehat{\beta}_{ab} - a_A(a) - a_B(b) \right] \right\},$$

$$b^+ = \max\left\{ -a_A^- - a_B^-,\ \max_{(a,b)\in\mathcal{O}} \left[\widehat{\beta}_{ab} - a_A(a) - a_B(b) \right] \right\}.$$

A full-domain residualized bound is

$$M_R = \max\left\{ \left| b^- - a_0 - \sum_{t\notin\{A,B\}} a_t^+ \right|, \left| b^+ - a_0 - \sum_{t\notin\{A,B\}} a_t^- \right| \right\}.$$

This bound includes pair cells absent from D_2. The implementation applies the numerical guard

$$M \leftarrow (1 + 10^{-10})M + 10^{-12}.$$

The nominal step grid is

$$\gamma \in \{0.125, 0.25, 0.5, 1\}.$$

For any $M \geq \sup_x |z(x)|$, global scalar shrinkage uses

$$\eta = \min\left\{ \gamma, \frac{2}{M} \right\}, \qquad \widetilde{z} = \eta z,$$

with $2/M = +\infty$ when $M = 0$. A matched permissive and residualized pair uses

$$\eta_{P,R} = \min\left\{ \gamma, \frac{2}{\max(M_P, M_R)} \right\},$$

giving exact logistic gain bounds

$$B_P = \eta_{P,R} M_P, \qquad B_R = \eta_{P,R} M_R.$$

All updates, bounds, tuple identities, and prior weights are fixed before validation. For K common source pairs, the joint family contains

$$H = 3\,[4 + 5\cdot 4 + 3K\cdot 5\cdot 4] = 72 + 180K$$

tuples across the three checkpoints. The terms correspond to intercept, main-effect, and the three pair-based families P, R, and S. With uniform weights and $\delta = 0.1$, the two-sided penalty uses

$$L = \log \frac{4H}{\delta}.$$

For each tuple,

$$\mathrm{LCB}_j = \widehat{\Delta}_j - \mathrm{pen}_j, \qquad \mathrm{UCB}_j = \widehat{\Delta}_j + \mathrm{pen}_j\,.$$

On the simultaneous event,

$$\Delta_j - \Delta_k \geq \mathrm{LCB}_j - \mathrm{UCB}_k.$$

We report a null-inclusive selection for each repair family. The global policy compares I, A, P, R, and null; S remains a control. Every selection is locked before D_4 is evaluated.

H.4 Repair selection and held-out gains

Table 12: Joint-repair certification counts over three seeds. I: intercept; A: main effect; P: permissive pair; R: residualized pair; S: shuffled control. Global selects among I, A, P, R, and null.

Dataset	Checkpoint	I	A	P	R	S	Global
Avazu	Main logistic	0/3	0/3	1/3	0/3	0/3	1/3
	FM	0/3	0/3	0/3	0/3	0/3	0/3
	MLP	2/3	2/3	3/3	0/3	0/3	3/3
Criteo	Main logistic	0/3	0/3	0/3	0/3	0/3	0/3
	FM	0/3	1/3	0/3	0/3	0/3	1/3
	MLP	2/3	3/3	3/3	0/3	0/3	3/3
Total		4/18	6/18	7/18	0/18	0/18	8/18

Table 12 reports certification counts, and Table 13 reports the corresponding null-inclusive empirical D_4 gains.

Table 13: Selected empirical D_4 log-loss reductions in units of 10^{-3}, reported as mean $\pm$ sample standard deviation over seeds 101, 503, and 4099. Null contributes zero. R and S are omitted because they select null in every setting.

Dataset	Checkpoint	I	A	P	Global
Avazu	Main logistic	0	0	0.099 ± 0.171	0.099 ± 0.171
	FM	0	0	0	0
	MLP	0.144 ± 0.158	0.522 ± 0.544	0.755 ± 0.642	0.793 ± 0.611
Criteo	Main logistic	0	0	0	0
	FM	0	0.196 ± 0.339	0	0.196 ± 0.339
	MLP	0.651 ± 0.641	5.022 ± 2.562	0.894 ± 0.201	3.992 ± 3.667

For the Criteo MLP with seed 503, the simultaneous intervals certify

$$\begin{aligned}\mathrm{LCB}_A - \mathrm{UCB}_P &= 0.0026810846 - (0.0009610230 + 0.0008115808)\\ &= 0.0009084808.\end{aligned}$$

Thus, on the simultaneous event, the conditional population gain of the selected main-effect update exceeds that of the selected pair update by at least

$$9.08 \times 10^{-4}.$$

Different report seeds repartition the same fixed 200K-row sample, so the reported standard deviations measure split-to-split variation.

H.5 Matched residualization comparisons

For every certified permissive pair, we evaluate its residualized partner at the same source pair, ridge parameter, and actual scalar step. Table 14 reports all seven comparisons.

Table 14: Matched fitted-update comparisons. Gains and LCBs are in units of 10^{-3}. Gain ratio is the residualized partner's empirical D_4 gain divided by the permissive gain. Partners are determined by the D_3-selected permissive tuples and reported regardless of certification.

Dataset	Checkpoint/seed	P gain	R partner gain	R partner LCB	Gain ratio (%)
Avazu	Main/4099	0.296993	0.002463	−0.439318	0.829
	MLP/101	0.366381	0.025128	−1.007061	6.858
	MLP/503	1.496660	0.020995	−4.393079	1.403
	MLP/4099	0.402403	0.063691	−1.064063	15.828
Criteo	MLP/101	1.120223	−0.215332	−4.422784	−19.222
	MLP/503	0.829794	−0.211243	−4.517025	−25.457
	MLP/4099	0.733093	−0.321212	−4.310347	−43.816

For the Avazu MLP, the mean matched residualized gain is

$$(3.660 \pm 2.355) \times 10^{-5},$$

and the mean per-seed gain ratio is

$$8.03\% \pm 7.28\%.$$

For the Criteo MLP, the mean matched residualized gain is

$$(-2.493 \pm 0.623) \times 10^{-4}.$$

None of the seven residualized partners receives a positive certificate. The all-main-effect projection therefore removes most of the selected permissive pair utility in these runs.

H.6 Numerical diagnostics

Residualization uses column-scaled LSQR (Paige & Saunders, 1982) with

$$\texttt{atol} = \texttt{btol} = 10^{-10}, \qquad \texttt{iter_lim} = 20000.$$

Across all settings, we compute 3855 unregularized pair-update projections. The maximum leakage ratio is

$$\frac{\left\|Q_0^\top W z_R\right\|_2}{\left\|Q_0^\top W z_P\right\|_2 + 10^{-12}} = 2.93 \times 10^{-8}.$$

Every projection satisfies the pre-specified 10^{-6} relative leakage threshold. None reaches the iteration limit, and the largest iteration count across main-effect fitting and residualization is 3629.

Global shrinkage applies to 79.2–81.5% of pair tuples on Avazu and 90.0–90.7% on Criteo. The highest-LCB residualized tuple in each audit uses

$$\lambda_P = 10, \qquad \gamma = 0.125,$$

with scale fraction one. Supplementary logs include tuple manifests, selected hyperparameters, validation statistics, update bounds, and seed-level results.

I Scope and limitations

Fitting model. The exposure–cost analysis gives sufficient conditions for fixed-design Gaussian linear learners. Population discoverability also depends on the covariate distribution, cell coverage, and the generalization of the fitted update. The independent validation layer assesses population gain for the resulting candidate functions.

Checkpoints and data. The CTR experiments use finite-budget checkpoints and 200K-row samples. Report seeds repartition the same samples, so their standard deviations measure split-to-split variation. Base risks differ across checkpoints; longer training or joint retraining can change the residual directions available for repair.

Reference spaces. The interpretation of a repair depends on its reference space and fitting procedure. Permissive pair tables include lower-order directions. The real-label residualized updates remove a fitted empirical all-main-effect component after pair fitting. Their certification outcomes characterize that pipeline. The planted construction similarly removes an empirical constituent-field additive component under the observed covariate distribution.

Validation design. The certificates require validation observations that are i.i.d. from the target distribution and independent of the fixed candidate pipeline. Adaptive changes informed by validation results require new independent validation. Temporal dependence, repeated-user clustering, and deployment shift require a corresponding evaluation design.