RECOURSE AS A CONTRACT: CERTIFYING REALIZED RECOURSE UNDER BEHAVIORAL AND MODEL DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Algorithmic recourse recommends changes intended to reverse an unfavorable decision, but a recommendation need not be implemented exactly, and the classifier may change before the individual returns. The relevant deployment question is therefore not whether the prescribed action remains valid, but whether the *realized response* will be accepted by the model in force at the realized return time. We formalize this requirement as a *recourse contract*: a high-confidence upper bound on terminal recourse failure. In practice, behavioral responses to recommendations and model evolution are typically observed in separate, unpaired data sources. Consequently, their marginal distributions may be estimable while the dependence between them, which determines terminal risk, is not. We characterize this partial-identification problem sharply, showing that the best coupling-robust set certificate attains the worst-case endpoint of the identified risk interval, and that this endpoint can become vacuous under a Strassen-type coupling condition. We then develop finite-sample contracts that remain valid for arbitrary dependence between the two channels and for action-dependent return times. We also establish a distribution-free certification floor of order $\log(1/\delta)/\min\{N, n_d\}$, showing that limited model history creates an intrinsic barrier to strong guarantees, and identify a complementary tension between recommendation effectiveness and the transferability of behavioral evidence. Experiments on controlled synthetic environments and temporally evolving LendingClub classifiers show that independence-based and model-only certificates can substantially understate realized risk, while our contracts remain valid and expose when the available data are insufficient to support a nontrivial guarantee.

1 INTRODUCTION

Algorithmic recourse tells a rejected applicant what could be changed to obtain a favorable decision (Wachter et al., 2018; Ustun et al., 2019; Karimi et al., 2022). Much work makes such prescriptions robust to model updates, multiplicity, distribution shift and causal misspecification (Upadhyay et al., 2021; Black et al., 2021; Pawelczyk et al., 2020; Nguyen et al., 2023; Karimi et al., 2020; Stępka et al., 2025; Marzari et al., 2026; Turbal et al., 2025). An institution that issues advice is, however, accountable for whether that advice still works in deployment. Recipients follow advice fully, partially or not at all, substitute other changes, or never return (Xuan et al., 2025; Upadhyay et al., 2025); and returning applicants are scored by the model in force *then* (Rawal et al., 2020; Fonseca et al., 2023). The operational question is: *among rejected applicants in context g who receive recommendation σ, how often does the returning applicant still fail?* A high-confidence upper bound on this probability is a *recourse contract*.

A contract involves two channels. The behavioral channel (recommendation $\rightarrow$ response) can be learned from randomized recommendation logs, as in encouragement designs (Angrist et al., 1996; Zhou, 2023). The institutional channel (current $\rightarrow$ future model) can be learned from the history of model updates. The two sources are *unpaired*, so the coupling between responses and future models—on which the terminal risk depends—is not identified. Independence is a tempting default, but common shocks break it: a downturn can simultaneously limit what applicants can do and tighten the institution's criteria. Robustness to the unknown coupling has a price, and that price can be total exactly under the predictive multiplicity that motivates robust recourse. Recent work on model change names imperfect implementation as an open direction (Kayastha et al., 2024).

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.

Contributions. (1) We define the contract for *realized* recourse, including partial compliance, substitution, non-return and action-dependent return times (Section 2). (2) We characterize what unpaired data can certify: the identified set is an interval whose upper endpoint equals the best set certificate (Theorem 3.1) and equals one exactly under Strassen's condition, which model multiplicity meets (Corollary 3.2); a recorded environment recovers much of the gap, while a generic dependence budget recovers it only near independence (Propositions 3.3 and 3.4). (3) We give non-asymptotic contracts that combine behavioral evidence via exact binomial tests with path-wise model-history tolerance families and remain valid for every coupling and every return-time mechanism (Theorems 4.5 and 4.6), and we show that no distribution-free procedure certifies below roughly $\log(1/\delta)/\min(n_d, N)$, a floor our procedure attains up to the split of the error budget (Proposition 4.7). (4) We give an exact decomposition of recommendation leverage (Theorem 5.1) and show that leverage and certifiability conflict: recommendations that move behavior enough to matter are those whose evidence cannot be borrowed (Corollary 5.3). (5) We run experiments designed to falsify each claim, including a common-shock sweep that breaks independence-based certificates and the certification of five recourse generators on LendingClub (Section 7).

The behavioral certificate is an instance of risk control by multiple testing (Bates et al., 2021); Theorem 3.1 and Corollary 3.2 are finite cases of Strassen's coupling theorem (Strassen, 1965); tolerance limits go back to Wilks (1941), supremum scores covering a whole trajectory are standard in functional conformal prediction (Lei et al., 2015; Diquigiovanni et al., 2022), and sampling-based guarantees already certify counterfactual robustness under model change (Marzari et al., 2024). The novelty lies in the deployment-level contract, its unpaired behavioral–model structure, the characterization of what these data can and cannot identify, and their synthesis into an end-to-end finite-sample certificate.

2 Problem formulation

An institution accepts features x when $f_{\theta_0}(x) \geq 0$. A rejected applicant X in context $G(X) = g$ receives a recommendation σ from a menu $\mathcal{S}$ shared by the context (a recommendation may also be a policy $x \mapsto$ prescription; Section 7). It induces a realized action $A(\sigma) \sim K_\sigma(\cdot \mid X)$ (the *response kernel*) and a post-action state $v_X(A)$, and the applicant returns at $T \in \{1, \ldots, \bar{t}\}$ or never ($T = \infty$). The *response record* $Z = (X, A, T)$ has law $Q_{g,\sigma}$. The institution's model follows a path $\vartheta = (\vartheta_t)_{t \leq \bar{t}} \sim \nu$, and a returning applicant is judged by ϑ_T. Terminal failure is $\ell^\star(Z, \vartheta) = \mathbf{1}\{T = \infty\} + \mathbf{1}\{T \leq \bar{t}, f_{\vartheta_T}(v_X(A)) < 0\} \in \{0, 1\}$.

Definition 2.1 (Recourse contract)**.** For a coupling $\varpi \in \Pi(Q_{g,\sigma}, \nu)$ the terminal risk is $J_{g,\sigma}(\varpi) = \mathbb{E}_\varpi[\ell^\star(Z, \vartheta)]$. Recommendation σ meets the contract at level τ in context g if $J_{g,\sigma}(\varpi) \leq \tau$ for the deployment coupling. A *certificate* is a data-dependent set of certified cells; it is valid at level δ if, with probability at least $1 - \delta$ over the calibration data, every certified cell meets the contract.

The certificates of Section 4 satisfy the stronger requirement $\sup_\varpi J_{g,\sigma}(\varpi) \leq \tau$. The contract is cell-level: individual contracts are impossible without smoothness (Theorem A.5), and for a cohort facing one model path the model budget is a cohort-wide breach probability (Proposition A.2). Non-return counts as failure, which keeps the loss observed. Because the contract concerns realized responses, it can fail even when every prescribed action is robust (Section 7).

Data. Randomized recommendation logs identify $Q_{g,\sigma}$ (Assumption 4.2); with covariate-dependent logging, inverse-propensity weighting identifies it when every recommendation has positive propensity. Historical increment paths identify ν if they are i.i.d. draws from the deployment increment law (Assumption 4.3). Nothing links a response to a future model. If A and ϑ are uniform on $\{\pm 1\}$ and failure means $A = \vartheta$, the product coupling gives risk $1/2$, the couplings $\vartheta = \pm A$ give 1 and 0, and every value in between is compatible with the data.

3 Certifying without the coupling

The difficulty is not merely estimating two distributions. Terminal recourse failure depends jointly on the realized behavioral response Z and the future model path ϑ, but the available logs identify only their marginals $Q_{g,\sigma}$ and ν. We therefore ask what terminal recourse risk is identified by these two data sources alone.

Theorem 3.1 (Sharp identified set of the terminal risk). *Fix a cell (g, σ), write $Q = Q_{g,\sigma}$ for the law of the response record Z and ν for the law of the model path ϑ, and let $F = \{(z, \vartheta) : \ell^\star(z, \vartheta) = 1\}$. For a set C of model paths let*

$$B_C = \{z : \exists\, \vartheta \in C,\ (z, \vartheta) \in F\}, \qquad A_C = \{z : (z, \vartheta) \in F\ \ \forall\, \vartheta \in C\}$$

be the responses that fail under some*, respectively* every*, path in C.*

(i) (Validity.) *For every measurable C with B_C, A_C measurable and every coupling $\varpi \in \Pi(Q, \nu)$,*

$$Q(A_C) + \nu(C) - 1 \ \le\ J(\varpi)\ \le\ Q(B_C) + \nu(C^c).$$

(ii) (Sharpness.) *If Z and ϑ take finitely many values, the set $\{J(\varpi) : \varpi \in \Pi(Q, \nu)\}$ is the interval $[J^\downarrow, J^\uparrow]$ with*

$$J^\uparrow = \min_C \big\{Q(B_C) + \nu(C^c)\big\}, \qquad J^\downarrow = \max_C \big\{Q(A_C) + \nu(C) - 1\big\},$$

where C ranges over all subsets of the support of ν.

Corollary 3.2 (When dependence is catastrophic). *In the finite setting of Theorem 3.1, $J^\uparrow = 1$ if and only if $Q(B_C) \ge \nu(C)$ for every set C of paths (Strassen's condition for a coupling supported on F; a Hall-type condition). In particular, if m equally likely models fail m disjoint, equally likely groups of responses, then the product-coupling risk is $J_\otimes = 1/m$ while $J^\uparrow = 1$.*

Part (i) is the certificate we use: choose a set C of model paths, pay for responses that fail under *some* path in C, and pay for the mass outside C. Part (ii) says nothing better is possible: the best set certificate *equals* the worst coupling, so a dependence-robust certificate is loose relative to $J^\uparrow$ only because of the restricted family of sets and finite samples. Corollary 3.2 shows the price of not knowing the coupling can be total, and predictive multiplicity (Pawelczyk et al., 2020; Black et al., 2021) is exactly its condition: robust certification is vacuous precisely where robust recourse is motivated.

Proposition 3.3 (Dependence budget). *Let $J_\otimes = (Q \otimes \nu)(F)$ and $\mathrm{kl}_+^{-1}(q, \iota) = \sup\{p \in [q, 1] : \mathrm{kl}(p \,\|\, q) \le \iota\}$, where kl is the Bernoulli relative entropy. If the deployment coupling satisfies $I_\varpi(Z; \vartheta) = \mathrm{KL}(\varpi \,\|\, Q \otimes \nu) \le \iota$, then $J(\varpi) \le \mathrm{kl}_+^{-1}(J_\otimes, \iota)$. The map $q \mapsto \mathrm{kl}_+^{-1}(q, \iota)$ is nondecreasing, so an upper confidence bound on $J_\otimes$ may be substituted. If Z and ϑ are conditionally independent given an environment E, then $I_\varpi(Z; \vartheta) \le \min\{I(Z; E), I(\vartheta; E)\} \le H(E)$.*

Proposition 3.4 (Observed common environment). *Suppose Z and ϑ are conditionally independent given an environment $E \in \{1, \dots, k\}$ that is recorded in both data sources, and that the deployment probabilities $p_e = \mathbb{P}(E = e)$ are known. Then $J(\varpi) = \sum_e p_e\, J_\otimes(Q_e, \nu_e)$. If U_e is a $(1 - \delta/k)$ upper confidence bound for $J_\otimes(Q_e, \nu_e)$ computed from the stratum-e data, then $J(\varpi) \le \sum_e p_e U_e$ with probability at least $1 - \delta$.*

The inversion kl_+^{-1} is the worst-case Bernoulli probability over a Kullback–Leibler ball (Ben-Tal et al., 2013) and is much tighter than Pinsker's bound, but a budget is informative only near independence: at $J_\otimes = 0.15$ and $\iota = H(E) = 0.50$ nats (a boom/bust environment) it gives 0.59, and ι is not identified from unpaired data. A recorded environment is far more informative; the stratified certificate of Proposition 3.4 is the best dependence-aware method in every experiment. If independence is simply assumed, $J_\otimes$ can be certified directly.

Proposition 3.5 (Product-coupling certificates). *Let $Z_1, \dots, Z_n \overset{iid}{\sim} Q$ and $\vartheta^{(1)}, \dots, \vartheta^{(n_d)} \overset{iid}{\sim} \nu$ be independent, and let $\hat{J} = (n n_d)^{-1} \sum_{i,k} \ell^\star(Z_i, \vartheta^{(k)})$.*

(a) (Bounded differences.) *With probability at least $1 - \delta$, $J_\otimes \le \hat{J} + \sqrt{\frac{1}{2}\big(\frac{1}{n} + \frac{1}{n_d}\big) \log \frac{1}{\delta}}$.*

(b) (Block empirical Bernstein.) *Let $B = \min(n, n_d)$ and, using independent uniformly random permutations drawn independently of the data, split the responses into B disjoint blocks of size $\lfloor n/B \rfloor$ and the paths into B disjoint blocks of size $\lfloor n_d/B \rfloor$ (leftover units are discarded), and let W_j be the average of $\ell^\star$ over the j-th pair of blocks. Then $W_1, \dots, W_B$ are i.i.d. in $[0, 1]$ with mean $J_\otimes$, and with probability at least $1 - \delta$,*

$$J_\otimes \le \bar{W} + \sqrt{\frac{2\hat{V} \log(2/\delta)}{B}} + \frac{7 \log(2/\delta)}{3(B - 1)},$$

where $\bar{W}$ and $\hat{V}$ are the sample mean and variance of the W_j.

$\hat{J}$ is a two-sample U-statistic of degree $(1,1)$, and the $1/n + 1/n_d$ scaling mirrors Hoeffding's variance decomposition (Hoeffding, 1963); empirical Bernstein bounds for U-statistics go back to Peel et al. (2010). Bound (a) never falls below $\sqrt{\log(1/\delta)/(2n_d)}$. Bound (b) adapts to agreement between models and was the tightest certificate whenever independence held in our experiments, but at $\delta = 0.1$ its $(B-1)^{-1}$ term makes (a) tighter for $n_d < 45$ even at zero empirical variance. Direct certification is therefore not the less efficient route; it is the less *robust* one.

4 End-to-End Certification from Unpaired Evidence

Definition 4.1 (Time-indexed tolerance family and robust loss)**.** A *tolerance family* is a collection $\mathcal{C}(r) = (C_1(r), \dots, C_{\bar{t}}(r))$ of model sets indexed by a radius $r \geq 0$, of the form $C_t(r) = \{\theta : \psi_t(\theta) \leq r\}$ for score functions ψ_t that are fixed before the calibration data are seen (a priori, or on an independent pilot split), with every $C_t(r)$, $r \geq 0$, nonempty. Its *path score* is $\Psi(\vartheta) = \max_{t \leq \bar{t}} \psi_t(\vartheta_t)$ and its *miss mass* is $\nu(\mathcal{C}(r)^c) = \nu\{\vartheta : \exists t,\ \vartheta_t \notin C_t(r)\} = \nu\{\Psi > r\}$. The *robust loss* of a response record $z = (x, a, t)$ is

$$\ell_{\mathcal{C}}(z) = \mathbf{1}\{t = \infty\} + \mathbf{1}\{t \leq \bar{t},\ \exists\, \theta \in C_t(r) : f_\theta(v_x(a)) < 0\},$$

and $R_{g,\sigma}(\mathcal{C}) = \mathbb{E}_{Q_{g,\sigma}}[\ell_{\mathcal{C}}(Z)]$ is the corresponding robust behavioral risk.

Ellipsoids $\psi_t(\theta) = \|M^{-1/2}(\theta - \theta_0)\|/c_t$ with an a priori time normalization c_t are the default; unions of such balls around pilot-estimated centres give *mixture* families. For linear scores and a ball $\{\|M^{-1/2}(\theta - \mu)\| \leq \rho\}$, the robust check is closed form, $\inf_\theta \theta^\top v = \mu^\top v - \rho\|M^{1/2}v\|$, the worst case also used by ElliCE (Turbal et al., 2025).

The family is fixed before calibration: a priori, or on a pilot split using a robust scatter estimate for M and pilot quantiles for c_t.

Assumption 4.2 (Randomized recommendation logging)**.** The behavioral sample consists of i.i.d. units $(X_i, G_i, \Sigma_i, Z_i)$ with $G_i = G(X_i)$, $\Sigma_i \mid X_i \sim \pi_0(\cdot \mid G_i)$ for a known logging policy with $\pi_0(\sigma \mid g) > 0$, and $Z_i \mid (X_i, \Sigma_i = \sigma)$ distributed as the response record that recommendation σ would induce (consistency). Non-return within the horizon is recorded as $T_i = \infty$.

Assumption 4.3 (Model evolution)**.** Model paths are written $\vartheta = \theta_0 + \Delta$. The drift sample consists of n_d i.i.d. increment paths $\Delta^{(1)}, \dots, \Delta^{(n_d)}$ from the law ν_Δ of the deployment increment path, independent of the behavioral sample. The deployment path has law ν (the shift of ν_Δ by θ_0) whatever the cell (g, σ); its dependence on the deployment responses is *unrestricted*.

Assumption 4.4 (Behavioral shift budget)**.** For every cell (g, σ) and every tolerance set used, $R^{\mathrm{dep}}_{g,\sigma}(\mathcal{C}) \leq R^{\mathrm{cal}}_{g,\sigma}(\mathcal{C}) + \varepsilon$. This holds whenever $\mathrm{TV}(Q^{\mathrm{dep}}_{g,\sigma}, Q^{\mathrm{cal}}_{g,\sigma}) \leq \varepsilon$, and $\varepsilon = 0$ when there is no shift.

Assumption 4.3 restricts only the *marginal* law of the deployment path; its dependence on deployment responses is unrestricted. A loss-specific budget in Assumption 4.4 can be much smaller than a TV budget but is not estimable before deployment; the consequences of behavioral-law misspecification are examined in Section C.7. Given hypotheses $H_{g,\sigma} : R_{g,\sigma} \geq \eta$, exact binomial p-values with Holm's procedure control the family-wise error at δ_b (Theorem A.1).

Because the return-time model is ϑ_T and T may depend on the realized response, fixed-time coverage is insufficient; we therefore require path-wise model control.

Theorem 4.5 (Path-wise PAC tolerance set)**.** *Let $\Psi_{(1)} \leq \cdots \leq \Psi_{(n_d)}$ be the ordered path scores of the drift sample, let k^* be the smallest k with $\mathbb{P}\{\mathrm{Bin}(n_d, 1-\gamma) \geq k\} \leq \delta_d$, and let $\hat{r} = \Psi_{(k^*)}$ (with $\hat{r} = \infty$ if no such k exists). Then $\mathbb{P}\big\{\nu(\mathcal{C}(\hat{r})^c) \leq \gamma\big\} \geq 1 - \delta_d$. On this event, for every random return time T, however it depends on the response or on the path, $\mathbb{P}\{T \leq \bar{t},\ \vartheta_T \notin C_T(\hat{r})\} \leq \gamma$. A finite radius exists if and only if $n_d \geq \log \delta_d / \log(1-\gamma)$.*

Supremum scores over the horizon are the standard route to simultaneous coverage of a trajectory (Lei et al., 2015; Diquigiovanni et al., 2022); pointwise calibration undercovers simultaneously (Diquigiovanni et al., 2021), which is the failure that single-horizon families exhibit in (Figure 4). We now combine the two statistically separate channels into a single deployment contract.

Algorithm 1: Recourse-as-a-contract certification (RAC)

Input: behavioral log; drift sample; tolerance family fixed in advance; menu $\mathcal{S}$; level τ; shift budget ε; grid Γ; δ_b, δ_d

foreach $\gamma \in \Gamma$ **do**
  $\hat{r}_\gamma \leftarrow$ Wilks tolerance limit of the path scores at content $1-\gamma$, level $\delta_d/|\Gamma|$;
  foreach *cell* (g, σ) **do**
    $N^f \leftarrow$ number of logged units with robust loss one under $\mathcal{C}(\hat{r}_\gamma)$ (non-return counts);
    $p_{g,\sigma,\gamma} \leftarrow \mathbb{P}\{\mathrm{Bin}(N_{g,\sigma}, \tau - \gamma - \varepsilon) \le N^f\}$;
Holm's procedure on all $p_{g,\sigma,\gamma}$ at level δ_b; certify (g, σ) if some $p_{g,\sigma,\gamma}$ is rejected;
In each context, issue the certified recommendation of highest utility, or abstain;

Theorem 4.6 (End-to-end recourse contract)**.** *Let Γ be a finite grid of model budgets. For each $\gamma \in \Gamma$, compute $\hat{r}_\gamma$ by Theorem 4.5 at level $\delta_d/|\Gamma|$, compute robust losses with $\mathcal{C}(\hat{r}_\gamma)$, and set $\eta_\gamma := \tau - \gamma - \varepsilon$. For candidates with $\eta_\gamma > 0$, test $H_{g,\sigma,\gamma} : R_{g,\sigma}(\mathcal{C}(\hat{r}_\gamma)) \ge \eta_\gamma$ with Theorem A.1 jointly over all (g, σ, γ) at level δ_b; candidates with $\eta_\gamma \le 0$ are non-certifying. Certify (g, σ) if some $H_{g,\sigma,\gamma}$ is rejected. Under Assumptions 4.2 to 4.4, with probability at least $1 - \delta_b - \delta_d$ over the two calibration samples, every certified cell satisfies $\sup_{\varpi \in \Pi(Q^{\mathrm{dep}}_{g,\sigma}, \nu)} J_{g,\sigma}(\varpi) \le \tau$. The same conclusion holds for a single fixed family $\mathcal{C}(r_0)$ if γ is replaced by the Clopper–Pearson upper bound $\hat{\gamma}$ on $\nu(\mathcal{C}(r_0)^c)$ computed from the drift sample at level δ_d (the* set certificate*).*

The target level decomposes as $\tau = \eta + \varepsilon + \gamma$, where η controls realized behavioral failure, ε behavioral shift, and γ model-path miss. Thus Theorem 4.6 combines behavioral and model-history evidence into a single recourse guarantee without requiring independence between the two deployment channels.

Budgets and timing. A certified level splits as $\tau = \eta + \varepsilon + \gamma$: a behavioral threshold tested exactly (Theorem A.1), a behavioral shift budget, and a model budget controlled by Theorem 4.5; the grid Γ lets Holm's procedure choose the split. A family calibrated at one horizon controls the model only at that time, so slower actions that return later can meet a temporary tightening; Theorem 4.5 controls the whole path, which makes action-dependent return times harmless. A natural question is whether RAC's conservatism is due to its construction or is intrinsic to the two-channel evidence.

Proposition 4.7 (Contract floor)**.** *Let N be the number of logged units in a cell and $\gamma_{\min} = 1 - (\delta_d/|\Gamma|)^{1/n_d}$. (a)* (Algorithm 1.) *Suppose the unbounded family $\mathcal{C}(\infty)$ contains, for every state, a model rejecting it (as for ellipsoids with linear scores and an intercept). If Algorithm 1 certifies the cell, then $\tau \ge \varepsilon + \gamma_{\min} + 1 - \delta_b^{1/N}$. The same holds for the set certificate with $|\Gamma| = 1$. (b)* (Any procedure.) *Consider any procedure whose probability of certifying the cell while $\sup_\varpi J_{g,\sigma}(\varpi) > \tau$ is at most δ for every response law and every model-path law. If, conditionally on N, it certifies the cell with probability at least $\omega > \delta$ under some pair (Q, ν_0), then $\tau \ge 1 - (\delta/\omega)^{1/\min(n_d, N)}$.*

Thus the short-history barrier is not merely an artifact of Algorithm 1: any distribution-free procedure with nontrivial certification power faces the same $\min(N, n_d)$ bottleneck. With $|\Gamma| = 1$ and $\delta_d = 0.05$, $\gamma_{\min} \approx 3/n_d$: a rule of three for model histories. Part (b) also covers product-coupling certificates, since the least-favourable perturbation raises the risk of every coupling. In the drift-limited regime with abundant behavioral data and high certification power, Algorithm 1 matches the same first-order $1/n_d$ floor up to the error-budget split and grid correction (a factor $\log(|\Gamma|/\delta_d)/\log(1/\delta) = 1.30$ for $|\Gamma| = 1$), while the product certificates of Proposition 3.5 have larger floors at the history lengths in Table 1. For $n_d = 10$, the value $\tau \approx 0.21$ is the high-power lower bound, not an unconditional prohibition on all procedures; escaping the model-history floor requires additional structure on the model channel.

5 Recommendation Leverage

Let $u_x : \mathcal{A} \to [0, 1]$ be any *valuation* of realized actions: the acceptance probability $u_x(a) = \mathbb{P}\{\text{return and } f_{\vartheta_T}(v_x(a)) \ge 0 \mid x, a\}$ under the product coupling, or the robust success probability $u_x(a) = 1 - \mathbb{E}[\ell_{\mathcal{C}}(Z) \mid x, a]$, which requires no coupling assumption. Write $\mathrm{Acc}_g(\sigma) =$

Table 1: Smallest certifiable contract level with abundant behavioral data (Proposition 4.7), versus the number of past model updates n_d. "Any valid": part (b) with $\delta = 0.10$ and certification probability $\omega \to 1$; RAC: part (a) with $\delta_d = 0.05$; product certificates: Proposition 3.5 with $\delta = 0.10$.

n_d	any valid	RAC, $\lvert\Gamma\rvert = 1$	RAC, $\lvert\Gamma\rvert = 6$	product, bounded diff.	product, emp. Bernstein
10	0.206	0.259	0.380	0.339	0.777
30	0.074	0.095	0.148	0.196	0.241
100	0.023	0.030	0.047	0.107	0.071

$\mathbb{E}[u_X(A(\sigma)) \mid G = g]$; when return and attrition depend on the realized action only, $\mathrm{Acc}_g(\sigma) = 1 - J_{g,\sigma}(Q_{g,\sigma} \otimes \nu)$ for the acceptance valuation and $1 - R_{g,\sigma}(\mathcal{C})$ for the robust one. For two recommendations, $D_x = K_\sigma(\cdot \mid x) - K_{\sigma'}(\cdot \mid x) = \mathrm{TV}_x\,(P_x^+ - P_x^-)$ for the normalized Jordan decomposition ($P_x^\pm$ are probability measures with disjoint supports and $\mathrm{TV}_x = D_x^+(\mathcal{A})$; products with TV_x are read as 0 when $\mathrm{TV}_x = 0$).

Theorem 5.1 (Exact two-channel decomposition)**.** *For all σ, σ' and every valuation,*

$$\mathrm{Acc}_g(\sigma) - \mathrm{Acc}_g(\sigma') = \mathbb{E}\Big[\mathrm{TV}_X\left(\mathbb{E}_{P_X^+} u_X - \mathbb{E}_{P_X^-} u_X\right) \;\Big|\; G = g\Big].$$

The first factor measures how much the recommendation moves behaviour; the second is the future-model acceptance gap between the actions that σ promotes and those it demotes. The identity is an exact rewriting by the Jordan decomposition; its value is interpretive.

Corollary 5.2 (Oscillation and transport bounds)**.** *(i) With $S_X = \mathrm{supp} P_X^+ \cup \mathrm{supp} P_X^-$,*

$$|\,\mathrm{Acc}_g(\sigma) - \mathrm{Acc}_g(\sigma')| \le \mathbb{E}[\mathrm{TV}_X\ \mathrm{osc}_{S_X}\, u_X \mid G = g] \le \mathbb{E}[\mathrm{TV}_X\ \mathrm{osc}_{\mathcal{A}}\, u_X \mid G = g].$$

(ii) If $\mathcal{A}$ is a metric space and each u_x is Lipschitz, $|\,\mathrm{Acc}_g(\sigma) - \mathrm{Acc}_g(\sigma')| \le \mathbb{E}[\mathrm{Lip}(u_X)\, W_1(K_\sigma(\cdot \mid X), K_{\sigma'}(\cdot \mid X)) \mid G = g]$.

The rightmost bound in (i) is nearly vacuous for recourse: "do nothing" is always realizable and fails, while the recommended action usually succeeds, so $\mathrm{osc}_{\mathcal{A}}\, u_x \approx 1$ and the bound collapses to the response-only bound $\mathbb{E}[\mathrm{TV}_X]$.

In our synthetic environment two recommendations with nearly disjoint kernels ($\mathbb{E}[\mathrm{TV}_X] = 0.96$) and oscillation 0.98 differ in acceptance by only 0.010, while the exact identity reproduces every true difference (Table 11): large behavioral movement need not translate into leverage. The converse obstruction concerns evidence.

Corollary 5.3 (Leverage–certifiability tension)**.** *Fix g and σ, σ', suppose return and attrition depend on the realized action but not on the recommendation, and let $\Lambda_u = |\,\mathrm{Acc}_g(\sigma) - \mathrm{Acc}_g(\sigma')|$ for a valuation u.*

(i) $\mathrm{TV}(Q_{g,\sigma}, Q_{g,\sigma'}) = \mathbb{E}[\mathrm{TV}_X \mid G = g] \ge \Lambda_u$.

(ii) Certifying σ from the logs of σ' under a TV shift budget (Assumption 4.4 with $Q^{\mathrm{cal}} = Q_{g,\sigma'}$ and $Q^{\mathrm{dep}} = Q_{g,\sigma}$) requires $\varepsilon \ge \Lambda_u$. For the robust valuation, the transferred bound $R_{g,\sigma'}(\mathcal{C}) + \varepsilon$ exceeds $R_{g,\sigma}(\mathcal{C})$ by at least $\mathbb{E}[\mathrm{TV}_X \mid G = g] + R_{g,\sigma'}(\mathcal{C}) - R_{g,\sigma}(\mathcal{C})$, hence by at least $2\Lambda_u$ whenever σ improves on σ'.

The recommendations worth issuing—those that move behavior toward actions the future model accepts—are therefore those whose evidence cannot be borrowed: each unit of leverage costs at least one unit of shift budget, and a borrowed certificate is slack by at least twice the improvement it would certify. Between the recommendations in Appendix C.8, $\mathbb{E}[\mathrm{TV}_X] \in [0.90, 0.98]$, so no TV-transferred contract below $\tau \approx 0.9$ exists, and each recommendation must be certified from its own randomized logs. A loss-specific budget can be far smaller than TV, but it is a statement about the very risk being certified and cannot be verified from the logs of σ'.

6 Related work

Robust recourse. Robust recourse studies how prescribed actions remain effective under model changes through adversarial or distributional uncertainty (Upadhyay et al., 2021; Nguyen et al.,

2023; Bui et al., 2025), predictions of future models (Kayastha et al., 2024), deep-model instability (Black et al., 2021; Hamman et al., 2023), predictive multiplicity (Pawelczyk et al., 2020), and noisy implementation (Pawelczyk et al., 2022; Nguyen et al., 2022; Xuan et al., 2025). Closest to our model channel, BetaRCE provides probabilistic guarantees under model change (Stępka et al., 2025); Marzari et al. (2024; 2026) study robustness under plausible parameter shifts; and ElliCE optimizes over a Rashomon ellipsoid (Turbal et al., 2025). RAC complements this literature by moving the certified object from the prescribed action to the *realized response at the realized return time*, combining behavioral uncertainty and model evolution learned from separate data sources. Existing recourse generators can therefore be evaluated as policies within the same contract framework.

Behavior and time. The behavioral channel connects to causal recourse (Karimi et al., 2020; 2021; Plecko et al., 2026), strategic and informational models (Tsirtsis & Gomez Rodriguez, 2020; Harris et al., 2022), and work on temporal validity and durable recourse (Fonseca et al., 2023; Segal et al., 2024; Ceccon et al., 2025). König et al. (2025) study performative validity, where recourse affects the population used to train future models. Our formulation focuses instead on deployment uncertainty in the realized response, return time, and future model path. Randomized recommendation logging is closely related to encouragement-design ideas (Angrist et al., 1996; Della Penna et al., 2016; Zhou, 2023; Della Penna, 2026).

Distribution-free calibration. Our finite-sample contracts draw on risk-controlling prediction sets and learn-then-test procedures (Bates et al., 2021; Angelopoulos et al., 2025), conformal and PAC-style prediction sets (Vovk et al., 2005; Vovk, 2012; Park et al., 2019; Tibshirani et al., 2019; Angelopoulos et al., 2024), and distribution-free trajectory coverage. RAC combines behavioral risk calibration with path-wise control of model evolution so that the relevant event is evaluated at the subject's random return time.

Dependence uncertainty and data combination. Classical results on marginal bounds, couplings, and partial identification (Makarov, 1982; Rüschendorf, 1982; Embrechts & Puccetti, 2006; Strassen, 1965; Kellerer, 1984; Ridder & Moffitt, 2007; Manski, 2003; Fan & Park, 2010) provide tools for reasoning about joint quantities when only marginal information is available. RAC brings this perspective to algorithmic recourse: behavioral responses and model paths are observed through separate data sources, and the resulting coupling uncertainty becomes part of the deployment contract.

7 Experiments

We test four predictions of the theory: (Q1) certificates that assume independent channels fail once responses and drift share an environment, while RAC does not (Theorem 3.1, Propositions 3.3–3.5); (Q2) RAC's remaining slack is mostly the identification gap; (Q3) random return times require path-wise model sets (Theorem 4.5); and (Q4) on LendingClub, what RAC certifies for standard generators and how much model history this needs (Proposition 4.7). Throughout, $\delta_b = \delta_d = 0.025$, $\Gamma = \{0.05, 0.10, \ldots, 0.50\}$, $\varepsilon = 0$, Holm's procedure runs over all tested (g, σ, γ) (on LendingClub also over τ), and evaluation data are never used for calibration. The full protocol is in Appendix C.1.

Synthetic setup. The environment has two contexts and three recommendations with partial compliance, substitution, non-return and action-dependent return times over $\bar{t} = 3$ updates. A latent environment E shifts responses and drift with intensity $s \in [0, 1]$. In the baseline parametrization the dependence gap $J^{\mathrm{dep}} - J_{\otimes}$ never exceeds 0.024, below the slack of product certificates (0.145 at $N = 800$, $n_d = 120$), so finite-sample product certificates do not fail there. We therefore add a *stress* variant with a matching-type coupling (Corollary 3.2): stressed applicants substitute toward a feature that stressed future models devalue. Its gap reaches 0.055, 0.145 and 0.20 at $s = 0.2$, 0.35 and $s \geq 0.5$. Baselines are the plug-in $\hat{J}$, the product oracle $J_{\otimes}$, the product certificates of Proposition 3.5 (bounded differences, BD; block empirical Bernstein, EB), the dependence budget of Proposition 3.3 with $\iota = H(E)$, the recorded-environment certificate of Proposition 3.4, and a BetaRCE-style model-only certificate (Stępka et al., 2025) that bounds the failure of the *prescribed* action under future models. A repetition violates if some certified cell has $J^{\mathrm{dep}} > \tau^{\star}$, with J^{dep} the Monte Carlo deployment risk.

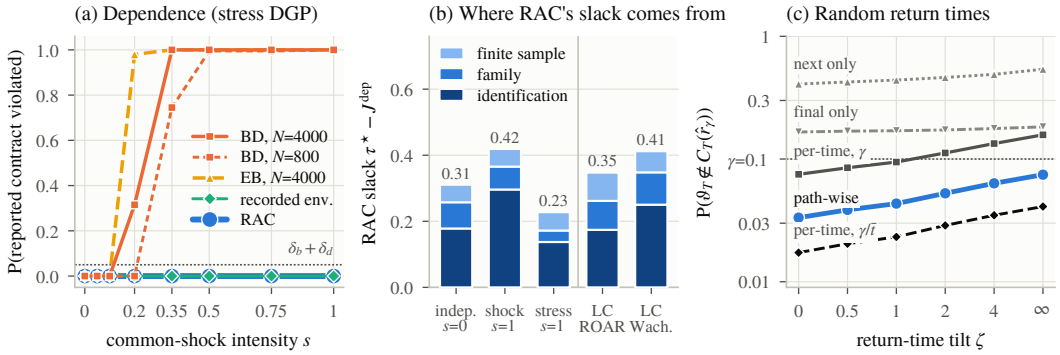

Figure 1: **(a)** How often (200 repetitions) some reported contract falls below the deployment risk in the stress environment. Product certificates fail once the dependence gap exceeds their slack, and earlier with more data; RAC and the recorded-environment certificate never do. **(b)** Exact decomposition of RAC's slack (cell means); the first two terms are population quantities. **(c)** Model-miss probability at the realized return time ($\gamma = 0.10$, $n_d = 1000$, 200 calibrations) as return times tilt toward atypical updates ($\zeta = \infty$: adversarial stopping).

Table 2: Stress environment, 500 repetitions. Viol.: frequency of some $\tau^\star < J^{\mathrm{dep}}$. Slack: median $\tau^\star - J^{\mathrm{dep}}$.

	Independent		Shock $s = 1$	
Certificate	Viol.	Slack	Viol.	Slack
Plug-in $\hat{J}$	0.79	0.01	1.00	−0.18
Product oracle $J_\otimes$	0.00	0.02	1.00	−0.18
Product BD (3.5a)	0.00	0.15	0.99	−0.04
Product EB (3.5b)	0.00	0.15	0.77	0.01
Model-only CP	1.00	−0.06	1.00	−0.44
KL budget (3.3)	0.00	0.71	0.00	abstains
Recorded env. (3.4)	0.00	0.23	0.00	0.23
RAC (Alg. 1)	0.00	0.35	0.00	0.28

Table 3: LendingClub (g_0/g_1). "–": RAC abstains. †: model-only claim below the realized risk $J_\otimes$.

Generator	$J_\otimes$	$J^\uparrow$	RAC $\tau^\star$	Model-only
AR	.840/.693	.943/.927	–/–	.773†/.764
PROBE	.835/.696	.945/.927	–/–	.767†/.754
RBR	.830/.687	.944/.929	–/–	.769†/.746
ROAR	.712/.423	.879/.597	–/.77	.569†/.133†
Wachter	.816/.568	.935/.818	–/.98	.754†/.536†

Q1: dependence breaks product certificates, not RAC. Product certificates are valid and tightest under independence (Table 2, $N = 800$, $n_d = 120$) but fail once the gap exceeds their slack (Figure 1a). At $N = 800$, BD is violated in 74% of repetitions at $s = 0.35$ and in at least 99% for $s \geq 0.5$. At $N = 4000$, $n_d = 600$ its slack shrinks to 0.067 while the gap does not, so it is violated under dependent deployment already at $s = 0.2$ (32%) and always from $s = 0.35$, by up to 0.13 on average. EB fails in 30–100% of repetitions for $s \geq 0.35$, the product oracle for every $s \geq 0.05$. None of RAC's 69,316 synthetic contracts was violated (smallest margin $\tau^\star - J^{\mathrm{dep}} = 0.17$) or fell below $J^\uparrow$. The recorded-environment certificate is the tightest one valid at every s (median slack 0.10 vs. 0.23–0.35 for RAC at $N = 4000$). The dependence budget is valid but nearly vacuous (slack 0.40–0.70; abstains in 51–100% of cells for $s \geq 0.5$). Model-only certificates, which certify the prescribed action, not the realized response, are violated in at least 96% of repetitions even at $s = 0$.

Q2: where the slack comes from. No set certificate goes below $J^\uparrow$ (Theorem 3.1(ii)), so RAC's slack splits exactly as $\tau^\star - J^{\mathrm{dep}} = (J^\uparrow - J^{\mathrm{dep}}) + (J_{\mathrm{fam}} - J^\uparrow) + (\tau^\star - J_{\mathrm{fam}})$. Here $J_{\mathrm{fam}} = \min_r\{R(\mathcal{C}(r)) + \nu(\mathcal{C}(r)^c)\}$ is the best certificate the tolerance family allows, and the last term is finite-sample error (Appendix C.4). Identification and family account for 75–87% of the slack at $N = 4000$ and for 76–84% in LendingClub; the finite-sample part is only 0.05–0.09 (Figure 1b). RAC's conservativeness is thus mostly the price of not observing the coupling, which more data cannot remove but recording E can (Proposition 3.4). The family matters too: in the stress environment a single-ball family cannot cover the bimodal drift law and abstains in over 99% of cells for $s \geq 0.35$.

Q3: random return times. We tilt the return-time law toward atypical model updates, $\Pr(T = t) \propto p_0(t)\, e^{\zeta z_t}$, where z_t is the standardized deviation of ϑ_t. Path-wise sets (Theorem 4.5) are compared with per-time sets at level γ for each t, a Bonferroni version at $\gamma/\bar{t}$, and next-update and final-horizon sets. Path-wise sets miss at most 0.075 for every ζ (Figure 1c), and their realized-time miss exceeded γ in at most 2 of 200 calibrations in any setting (target $\delta_d/|\Gamma| = 0.0025$; binomial $p \geq 0.09$). Per-time sets are valid for the untilted law, but at $n_d = 1000$ their miss grows to 0.157 and exceeds γ in 92–100% of calibrations for $\zeta \geq 2$; single-horizon sets miss 0.14–0.54. Bonferroni sets are valid but need $n_d \geq 210$ paths for a finite radius at $\gamma = 0.10$ (path-wise: 57), so at $n_d = 120$ they abstain in 4–19% of cells. End to end, no family's contracts were violated here: the robust loss leaves enough slack (median 0.36–0.58) to absorb the extra miss (Appendix C.6). Corollary 5.3 is also borne out: as response kernels separate (mean TV $\to 1$), leverage saturates at 0.052 and all transferred bounds become vacuous (Appendix C.8).

Q4: LendingClub. The setup (Appendix D) combines real applicant covariates (George, 2019), a classifier trained on 2007–2012 loans and refit on later years, and five generators on a common support of 40 applicants in two contexts. Responses are semi-synthetic ($50{,}000$ calibration and $100{,}000$ evaluation units, uniform logging), and there are $n_d = 200$ bootstrap model paths. In every cell $J^{\uparrow}$ exceeds $J_{\otimes}$ by 0.10–0.25 (Table 3). On a 0.01 contract grid, RAC certifies ROAR ($\tau^{\star} = 0.77$) and Wachter (0.98) in g_1 and abstains in the other eight cells, where $J^{\uparrow} \geq 0.879$: no coupling-robust contract below 0.879 exists there at any sample size. Model-only certificates understate the realized risk in 7 of 10 cells, by up to 0.29. Positive target margins help only partly: at margin 0.25, $J_{\otimes}$ in g_1 falls by at most 0.14 and RAC newly certifies AR and PROBE ($\tau^{\star} = 0.96$), but $J^{\uparrow} - J_{\otimes}$ stays at or above 0.17.

How much model history is enough? On random subsamples of the path bank, ROAR in g_1 is never certified at $n_d = 10$, is certified in 65% of subsamples at $n_d = 20$ and always from $n_d = 30$, and its median $\tau^{\star}$ falls from 0.98 to 0.77 at $n_d = 200$; Wachter needs $n_d \geq 150$. These paths, however, bootstrap the realized 2014/2015 refits and so measure estimation noise around known future models. A backtest using only history available at issuance (19 monthly origins in 2008–2009, refit with the same update timing) gives path scores about 14 times larger (median 33.9 vs. 2.5). Only $\gamma \geq 0.30$ is then feasible and RAC abstains in every cell, and the realized path one year later lies outside the bootstrap-calibrated set (Appendix D.6). In practice, then, model history rather than behavioral data limits what can be certified (Proposition 4.7).

8 Conclusion and Limitations

We studied realized algorithmic recourse when behavioral responses and future model evolution are observed through separate, unpaired data sources. Under independence, these data permit tight certification; when a common environment is observed, conditioning on it can recover much of the lost information. Without such structure, the sharp guarantee is determined by the worst-case coupling of the two marginals, which equals one under the Hall–Strassen condition. This highlights two institutional levers: recording variables that jointly explain behavioral and model variation (Proposition 3.4), and constraining the model channel, for example by evaluating returning applicants with the issuing model or restricting future updates to a pre-announced family.

Limitations. (i) Certification is limited by the amount of model history available (Proposition 4.7), and real update histories may be short or non-exchangeable. (ii) Parameter-space ellipsoids may be conservative in high dimensions (Appendix D.8). (iii) The LendingClub study uses genuine temporal covariate and classifier variation, but recommendation responses and return times are semi-synthetic. (iv) Exact cell-level certification requires randomized logging, although recorded propensities can support off-policy extensions. (v) The contract certifies future-model acceptance rather than improvement in the underlying outcome. (vi) Performative feedback is not modeled.

AI Use Statement

In this work, we used generative AI tools for manuscript review, language editing, assistance in checking selected mathematical arguments, and reference searches. All references suggested or located with AI assistance were independently verified against their public records. The authors

reviewed all AI-assisted content and take full responsibility for the final manuscript, including its text, mathematical claims, experiments, and other artifacts. .

ETHICS STATEMENT

Recourse advice affects access to credit and other opportunities. A certified contract can create unwarranted assurance if its assumptions fail, so our procedure abstains when certification fails, makes the behavioral shift budget explicit, and supports post-deployment monitoring. Abstention has a cost borne by the applicant, and it concentrates in thin-data cells; because contexts can include protected attributes, group-conditional contracts are directly available and should be used to audit such disparities, while Theorem A.5 rules out individual-level alternatives. Randomized recommendation logging is an experiment on consequential decisions and may interact with adverse-action notice rules that require stating the actual principal reasons for a denial; we recommend randomizing only among options believed to be acceptable, with disclosure, legal review and oversight. No human subjects were involved; the LendingClub data is public and anonymized, and the behavioral responses are simulated.

REPRODUCIBILITY STATEMENT

All assumptions and complete proofs are provided in Appendices A and B, and Algorithm 1 gives the certification procedure. Section 7 and Appendix C specify the synthetic environments, tolerance families, baselines, evaluation metrics, parameter grids, and experimental protocol. Appendix D describes the LendingClub data splits, preprocessing, recourse generators, model-update construction, and simulated behavioral-response mechanism. Together, these sections document the theoretical and experimental procedures used to obtain the reported results.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

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.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Emily Black, Zifan Wang, Matt Fredrikson, and Anupam Datta. Consistent counterfactuals for deep models. *arXiv preprint arXiv:2110.03109*, 2021.

Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. Coverage-validity-aware algorithmic recourse. *Operations Research*, 73(6):3294–3310, 2025.

T. Tony Cai and Mark G. Low. An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32(5):1805–1840, 2004.

Marina Ceccon, Alessandro Fabris, Goran Radanović, Asia J Biega, and Gian Antonio Susto. Reinforcement learning for durable algorithmic recourse. *arXiv preprint arXiv:2509.22102*, 2025.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Nicolás Della Penna. What do we care about in bandits with noncompliance? BRACE: Bandits with recommendations, abstention, and certified effects. *arXiv preprint arXiv:2603.09532*, 2026.

Nicolás Della Penna, Mark D. Reid, and David Balduzzi. Compliance-aware bandits. *arXiv preprint arXiv:1602.02852*, 2016.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. The importance of being a band: Finite-sample exact distribution-free prediction sets for functional data. *arXiv preprint arXiv:2102.06746*, 2021.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022.

Paul Embrechts and Giovanni Puccetti. Bounds for functions of dependent risks. *Finance and Stochastics*, 10(3):341–352, 2006.

Yanqin Fan and Sang Soo Park. Sharp bounds on the distribution oftreatment effects and their statisticalinference. *Econometric Theory*, 26(3):931–951, 2010.

João Fonseca, Andrew Bell, Carlo Abrate, Francesco Bonchi, and Julia Stoyanovich. Setting the right expectations: Algorithmic recourse over time. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–11, 2023.

Lester R Ford Jr and Delbert R Fulkerson. Maximal flow through a network. *Canadian journal of Mathematics*, 8:399–404, 1956.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Nathan George. All lending club loan data. Kaggle dataset, 2019. 2007–2018 Lending Club accepted and rejected loan data.

Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust counterfactual explanations for neural networks with probabilistic guarantees. In *International Conference on Machine Learning*, pp. 12351–12367. PMLR, 2023.

Keegan Harris, Valerie Chen, Joon Kim, Ameet Talwalkar, Hoda Heidari, and Steven Wu. Bayesian persuasion for algorithmic recourse. *Advances in Neural Information Processing Systems*, 35: 11131–11144, 2022.

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

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: A probabilistic approach. In *Advances in Neural Information Processing Systems*, volume 33, pp. 265–277, 2020.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 353–362, 2021.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

Kshitij Kayastha, Vasilis Gkatzelis, and Shahin Jabbari. Learning-augmented robust algorithmic recourse. *arXiv preprint arXiv:2410.01580*, 2024.

Hans G Kellerer. Duality theorems for marginal problems. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 67(4):399–432, 1984.

Gunnar König, Hidde Fokkema, Timo Freiesleben, Celestine Mendler-Dünner, and Ulrike von Luxburg. Performative validity of recourse explanations. *Advances in Neural Information Processing Systems*, 38:139334–139370, 2025.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, 3 edition, 2005.

Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):29–43, 2015.

Mark G Low. On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554, 1997.

GD Makarov. Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability & its Applications*, 26(4):803–806, 1982.

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

Luca Marzari, Francesco Leofante, Ferdinando Cicalese, and Alessandro Farinelli. Rigorous probabilistic guarantees for robust counterfactual explanations. In *ECAI 2024: 27th European Conference on Artificial Intelligence*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pp. 1059–1066, 2024.

Luca Marzari, Francesco Leofante, Ferdinando Cicalese, and Alessandro Farinelli. Probabilistically robust counterfactual explanations under model changes. *Artificial Intelligence*, pp. 104459, 2026.

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

Colin McDiarmid. On the method of bounded differences. In J. Siemons (ed.), *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989.

Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. Distributionally robust recourse action. *arXiv preprint arXiv:2302.11211*, 2023.

Tuan-Duy H Nguyen, Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. Robust bayesian recourse. In *Uncertainty in Artificial Intelligence*, pp. 1498–1508. PMLR, 2022.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. Pac confidence sets for deep neural networks via calibrated prediction. *arXiv preprint arXiv:2001.00106*, 2019.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 809–818, 2020.

Martin Pawelczyk, Teresa Datta, Johannes Van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. Probabilistically robust recourse: Navigating the trade-offs between costs and robustness in algorithmic recourse. *arXiv preprint arXiv:2203.06768*, 2022.

Thomas Peel, Sandrine Anthoine, and Liva Ralaivola. Empirical bernstein inequalities for u-statistics. *Advances in Neural Information Processing Systems*, 23, 2010.

Drago Plecko, Collin Wang, and Elias Bareinboim. Causal algorithmic recourse: Foundations and methods. *arXiv preprint arXiv:2605.11373*, 2026.

Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv preprint arXiv:2012.11788*, 2020.

Geert Ridder and Robert Moffitt. The econometrics of data combination. In *Handbook of Econometrics*, volume 6B, pp. 5469–5547. Elsevier, 2007.

Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14 (3):623–632, 1982.

Meirav Segal, Anne-Marie George, Ingrid Chieh Yu, and Christos Dimitrakakis. Better luck next time: About robust recourse in binary allocation problems. In *World Conference on Explainable Artificial Intelligence*, pp. 374–394. Springer, 2024.

Ignacy Stępka, Jerzy Stefanowski, and Mateusz Lango. Counterfactual explanations with probabilistic guarantees on their robustness to model change. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1277–1288, 2025.

Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Stratis Tsirtsis and Manuel Gomez Rodriguez. Decisions, counterfactual explanations and strategic behavior. *Advances in Neural Information Processing Systems*, 33:16749–16760, 2020.

Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

Bohdan Turbal, Iryna Voitsitska, and Lesia Semenova. Ellice: Efficient and provably robust algorithmic recourse via the rashomon sets. *Advances in Neural Information Processing Systems*, 38: 137267–137288, 2025.

Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16926–16937, 2021.

Sohini Upadhyay, Himabindu Lakkaraju, and Krzysztof Z Gajos. Counterfactual explanations may not be the best algorithmic recourse approach. In *Proceedings of the 30th International conference on intelligent user interfaces*, pp. 446–462, 2025.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 10–19, 2019.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31 (2):841–887, 2018.

Samuel S Wilks. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96, 1941.

Yueqing Xuan, Kacper Sokol, Mark Sanderson, and Jeffrey Chan. Perfect counterfactuals in imperfect worlds: modelling noisy implementation of actions in sequential algorithmic recourse. *Machine Learning*, 114(8):187, 2025.

Angela Zhou. Optimal and fair encouragement policy evaluation and learning. *Advances in Neural Information Processing Systems*, 36:13745–13772, 2023.

A Additional results and assumptions

A.1 Exact certification and cohorts

Theorem A.1 (Exact simultaneous certification). *Fix a tolerance set independently of the behavioral sample. For hypotheses $H_j : R_j \geq \eta_j$ indexed by $j = (g, \sigma, \cdot)$, let N_j be the number of logged units in cell j, N_j^f the number with robust loss one, and $p_j = \mathbb{P}\{\mathrm{Bin}(N_j, \eta_j) \leq N_j^f\}$. Holm's step-down procedure at level δ_b rejects a family of hypotheses such that, conditionally on the tolerance set, with probability at least $1 - \delta_b$ every rejected H_j is false, i.e. $R_j < \eta_j$.*

Proposition A.2 (Cohort-level contract). *Suppose the N deployment subjects of a certified cell face the same model path and that their robust-loss indicators are conditionally i.i.d. given a common environment E, with conditional mean $R(E)$ (independent subjects correspond to a trivial E). On the event of Theorem 4.6, for every $\delta' \in (0, 1)$ and $t \in [0, 1]$,*

$$\mathbb{P}\Big\{ \tfrac{1}{N} \textstyle\sum_{i \leq N} \ell^\star(Z_i, \vartheta) > t + \sqrt{\log(1/\delta')/(2N)} \Big\} \leq \gamma + \delta' + \mathbb{P}\{R(E) > t\}.$$

For independent subjects $R(E) \equiv R^{\mathrm{dep}} \leq \tau - \gamma$, and $t = \tau - \gamma$ makes the last term vanish. Under a common shock, Markov's inequality gives $\mathbb{P}\{R(E) > t\} \leq (\tau - \gamma)/t$. With probability up to γ the realized path leaves the tolerance set and the cohort failure rate is then uncontrolled: γ is a probability of a cohort-wide breach, not a per-subject rate.

A.2 Leverage and limits

Proposition A.3 (Drift smooths valuations). *In the linear model $f_\theta(v) = \theta^\top v$ with $v = (1, \tilde{v})$, $\vartheta \sim N(\theta_0, s^2 I)$ and additive actions $\tilde{v} = \tilde{x} + a$, the acceptance valuation $u_x(a) = \Phi\big(\theta_0^\top v/(s\|v\|)\big)$ satisfies $\mathrm{Lip}(u_x) \leq \|\theta_0\|/(s\sqrt{2\pi})$. Larger model uncertainty therefore reduces how much any change of recommendation can change acceptance; as $s \to \infty$, $u_x \to 1/2$ uniformly, so the reduction reflects decisions becoming arbitrary rather than recourse becoming easier.*

Proposition A.4 (Certified impossibility under arbitrary dependence). *Fix a set C of paths independently of the data. Let $L_{g,\sigma}$ be a lower confidence bound for $Q_{g,\sigma}(A_C)$ and L_ν a lower confidence bound for $\nu(C)$, jointly valid with probability at least $1 - \delta$ over all σ in context g. If $L_{g,\sigma} + L_\nu - 1 > \tau$ for every $\sigma \in \mathcal{S}$, then with probability at least $1 - \delta$ no recommendation in the menu meets the contract at level τ in context g, under any coupling.*

Theorem A.5 (No distribution-free pointwise contract). *Consider logged pairs (X_i, Y_i) with $Y_i = \ell_{\mathcal{C}}(Z_i)$ for a fixed recommendation, and suppose that near a point x_0 with $\mathbb{P}(X = x_0) = 0$ both a robustly failing and a robustly succeeding action are available, so that every continuous $r(x) = \mathbb{P}(Y = 1 \mid X = x)$ is realizable. If U is an upper confidence bound with $\mathbb{P}\{r(x_0) \leq U\} \geq 1 - \delta$ for every such law with continuous r, then $\mathbb{P}\{U \geq 1\} \geq 1 - \delta$ under* every *such law.*

Theorem A.6 (Certification complexity). *Let ϕ be any test of $H_0 : R \geq \eta$ based on n i.i.d. Bernoulli(R) losses with level δ that rejects with probability at least $1 - \zeta > \delta$ when $R = \eta - \Delta$. Then*

$$n \;\geq\; \frac{\mathrm{kl}(1 - \zeta \,\|\, \delta)}{\mathrm{kl}(\eta - \Delta \,\|\, \eta)} \;\geq\; \frac{\eta(1 - \eta)\ \mathrm{kl}(1 - \zeta \,\|\, \delta)}{\Delta^2}.$$

Conversely, the test that rejects when $N^f \leq n(\eta - \Delta/2)$ has level δ and power $1 - \zeta$ as soon as $n \geq 8\eta \log\big(1/\min(\delta, \zeta)\big)/\Delta^2$.

For completeness, the next two results quantify what smoothness buys: cell-level contracts can be refined to uniform local contracts at the minimax rate, and this rate cannot be improved; the main text does not rely on them.

Theorem A.7 (Uniform local contracts). *Let $\mathcal{X}_g \subseteq [0, 1]^d$ and suppose (a) $x \mapsto r_\sigma(x) = \mathbb{P}(\ell_{\mathcal{C}}(Z) = 1 \mid X = x, \Sigma = \sigma)$ is (L, β)-Hölder on $\mathcal{X}_g$ with known L and $\beta \in (0, 1]$; (b) for all $x \in \mathcal{X}_g$ and $h \leq h_0$, $\mathbb{P}\{X \in B_h(x) \cap \mathcal{X}_g, \Sigma = \sigma\} \geq c_0 h^d$. Let $\mathcal{N}_h \subseteq \mathcal{X}_g$ be an h-net of size $N_h \leq (1 + \sqrt{d}/h)^d$, and for $y \in \mathcal{N}_h$ let $\hat{R}_{2h}(y)$ be the empirical robust risk of the units with $\Sigma = \sigma$ in $B_{2h}(y)$. If $nc_0(2h)^d \geq 8 \log(2N_h/\delta)$ and $2h \leq h_0$, then with probability at least $1 - \delta$, simultaneously for all $x \in \mathcal{X}_g$ and the nearest net point $y(x)$,*

$$r_\sigma(x) \leq \hat{R}_{2h}(y(x)) + 3^\beta L h^\beta + \sqrt{\frac{\log(2N_h/\delta)}{nc_0(2h)^d}}.$$

With $h \asymp (\log n/n)^{1/(2\beta+d)}$ the excess is of order $(\log n/n)^{\beta/(2\beta+d)}$. The constant L cannot be learned from the data: honest adaptation over Hölder classes is impossible (Cai & Low, 2004).

Theorem A.8 (Width lower bound for local contracts)**.** *In the oracle experiment with n i.i.d. pairs (X_i, Y_i), X_i uniform on $[0,1]^d$ and $r \in \mathrm{Höl}(\beta, L)$ with values in $[1/4, 3/4]$, let U be any upper confidence bound for $r(x_0)$ that is valid at level $1-\delta$ over this class. Then there are constants $c, c' > 0$, depending only on (β, L, d, δ), such that for the constant function $r_0 \equiv 1/2$,*

$$\mathbb{P}_{r_0}\{U - r_0(x_0) \geq c\, n^{-\beta/(2\beta+d)}\} \geq c'.$$

The logged recourse experiment (binary actions with $\ell(x, a_-) = 1$, $\ell(x, a_+) = 0$ and a randomized recommendation) is a garbling of the oracle experiment, so the same lower bound applies to it.

B Proofs

B.1 Proof of Theorem 3.1

Validity. If $(z, \vartheta) \in F$ and $\vartheta \in C$ then $z \in B_C$; hence $F \subseteq (B_C \times \mathbf{\Theta}) \cup (\mathcal{Z} \times C^c)$, where $\mathbf{\Theta}$ is the path space, and for every coupling $\varpi(F) \leq Q(B_C) + \nu(C^c)$. Conversely $A_C \times C \subseteq F$, so by the Fréchet inequality $\varpi(F) \geq \varpi(A_C \times C) \geq Q(A_C) + \nu(C) - 1$.

Sharpness. Restrict Q and ν to their (finite) supports $\mathcal{Z}$ and $\mathbf{\Theta}$. The set $\Pi(Q, \nu)$ is a nonempty compact polytope and $\varpi \mapsto \varpi(F)$ is linear, so its image is a closed interval; it remains to identify the endpoints. Consider the network with a source s, a sink t, arcs $s \to z$ of capacity $Q(z)$, arcs $z \to \vartheta$ of infinite capacity for $(z, \vartheta) \in F$, and arcs $\vartheta \to t$ of capacity $\nu(\vartheta)$.

Step 1: $\max_\varpi \varpi(F)$ equals the maximum flow. Restricting a coupling to F gives a feasible flow of value $\varpi(F)$. Conversely, let f be a feasible flow of value v. If $v = 1$, f is a coupling supported on F. If $v < 1$, the residual masses $Q'(z) = Q(z) - \sum_\vartheta f(z, \vartheta) \geq 0$ and $\nu'(\vartheta) = \nu(\vartheta) - \sum_z f(z, \vartheta) \geq 0$ both have total mass $1 - v$, and $\varpi = f + Q' \otimes \nu'/(1 - v)$ is a coupling with $\varpi(F) \geq v$.

Step 2: the minimum cut equals $\min_C\{Q(B_C) + \nu(C^c)\}$. Let $S \ni s$, $t \notin S$ define a cut of finite capacity and put $C = \mathbf{\Theta} \setminus S$. Finite capacity means that no arc of F leads from $\mathcal{Z} \cap S$ into C, so $\mathcal{Z} \cap S \subseteq \mathcal{Z} \setminus B_C$ and the capacity $Q(\mathcal{Z} \setminus S) + \nu(\mathbf{\Theta} \cap S)$ is at least $Q(B_C) + \nu(C^c)$. Conversely, for every C the cut $S = \{s\} \cup (\mathcal{Z} \setminus B_C) \cup (\mathbf{\Theta} \setminus C)$ has no arc of F from $\mathcal{Z} \setminus B_C$ into C and capacity exactly $Q(B_C) + \nu(C^c)$.

By the max-flow min-cut theorem (Ford Jr & Fulkerson, 1956), $J^\uparrow = \min_C\{Q(B_C) + \nu(C^c)\}$. For the lower endpoint apply this to the complement F^c, for which the "some-path" set is $\mathcal{Z} \setminus A_C$: $J^\downarrow = 1 - \max_\varpi \varpi(F^c) = 1 - \min_C\{1 - Q(A_C) + 1 - \nu(C)\} = \max_C\{Q(A_C) + \nu(C) - 1\}$. □

Remark B.1 (Relation to the contract certificate and to general spaces). Let C be any set of paths and write $C_t = \{\vartheta_t : \vartheta \in C\}$ for its time-t section. If the robust loss is evaluated against the sections,

$$\ell_C(z) = \mathbf{1}\{t = \infty\} + \mathbf{1}\{t \leq \bar{t},\ \exists\, \theta \in C_t :\ f_\theta(v_x(a)) < 0\},$$

then B_C is exactly the event $\{\ell_C(Z) = 1\}$: for $z = (x, a, t)$ with $t \leq \bar{t}$, $(z, \vartheta) \in F$ for some $\vartheta \in C$ if and only if $f_{\vartheta_t}(v_x(a)) < 0$ for some $\vartheta \in C$, that is, if and only if some $\theta \in C_t$ rejects the state, and for $t = \infty$ both sides equal one whenever $C \neq \emptyset$. Since $\nu(C^c)$ is the miss mass, the certificate of Theorem 4.6 is the validity bound of Theorem 3.1(i) for the pre-specified set $C = \mathcal{C}(\hat{r}_\gamma) = \{\vartheta : \Psi(\vartheta) \leq \hat{r}_\gamma\}$. For a product family $\mathcal{C}(r) = \{\vartheta : \vartheta_t \in \mathcal{C}_t(r)\ \forall t\}$ the sections are the sets $\mathcal{C}_t(r)$ of Definition 4.1, because every $\theta \in \mathcal{C}_t(r)$ can be completed to a path in $\mathcal{C}(r)$; for a mixture family $\mathcal{C}(r) = \bigcup_k \prod_t \mathcal{B}_{k,t}(r)$, as in Appendix C.2, the sections are the unions $\bigcup_k \mathcal{B}_{k,t}(r)$. In either case the looseness of the certificate relative to $J^\uparrow$ comes only from the restricted family of sets and from finite samples.

B.2 Proofs of Corollary 3.2 and Propositions 3.3 to 3.5

Corollary 3.2. $J^\uparrow = 1$ iff $Q(B_C) + 1 - \nu(C) \geq 1$ for all C (the empty set attains the value 1). In the example, with groups $G_1, \ldots, G_m$ and model j failing exactly G_j, $Q(B_C) = |C|/m = \nu(C)$, so $J^\uparrow = 1$ (the coupling $\vartheta = j$ on G_j attains it), while $J_\otimes = \sum_j m^{-2} = 1/m$. The characterization is the finite, exact-coupling case of Strassen's theorem (Strassen, 1965).

Proposition 3.3. Since the marginals of ϖ are Q and ν, $I_\varpi(Z;\vartheta) = \mathrm{KL}(\varpi \,\|\, Q \otimes \nu)$. The data-processing inequality for the map $(z,\vartheta) \mapsto \mathbf{1}_F(z,\vartheta)$ gives $\mathrm{kl}(J(\varpi) \,\|\, J_\otimes) \le \iota$. If $J(\varpi) \le J_\otimes$ there is nothing to prove; otherwise $J(\varpi)$ belongs to $\{p \in [J_\otimes, 1] : \mathrm{kl}(p\|J_\otimes) \le \iota\}$. For monotonicity, let $q \le q'$; for $p \ge q'$ the map $q \mapsto \mathrm{kl}(p\|q)$ is nonincreasing on $(0,p]$, so $\mathrm{kl}(p\|q') \le \mathrm{kl}(p\|q)$ and $\{p \ge q : \mathrm{kl}(p\|q) \le \iota\} \cap [q',1] \subseteq \{p \ge q' : \mathrm{kl}(p\|q') \le \iota\}$; hence $\mathrm{kl}_+^{-1}(q,\iota) \le \max\{q', \mathrm{kl}_+^{-1}(q',\iota)\} = \mathrm{kl}_+^{-1}(q',\iota)$. Under conditional independence, $Z - E - \vartheta$ is a Markov chain, and data processing gives $I(Z;\vartheta) \le \min\{I(Z;E), I(\vartheta;E)\} \le H(E)$.

Proposition 3.4. By conditional independence, $J(\varpi) = \sum_e p_e \mathbb{E}[\ell^\star(Z,\vartheta) \mid E = e] = \sum_e p_e (Q_e \otimes \nu_e)(F)$. Conditionally on the stratum counts, the stratum samples are i.i.d. from Q_e and ν_e; a union bound over the k strata completes the proof.

Proposition 3.5. (a) $\hat{J}$ is a function of $n + n_d$ independent variables; replacing one Z_i changes it by at most $1/n$ and replacing one $\vartheta^{(k)}$ by at most $1/n_d$. McDiarmid's inequality (McDiarmid, 1989) gives $\mathbb{P}\{J_\otimes - \hat{J} \ge t\} \le \exp\{-2t^2/(1/n + 1/n_d)\}$ because $\mathbb{E}\hat{J} = J_\otimes$. (b) The permutations are independent of the data, so conditionally on them the blocks are fixed index sets; distinct block pairs use disjoint sets of mutually independent variables, so the W_j are independent; they are identically distributed by exchangeability, and each is an average of $\ell^\star(Z_i, \vartheta^{(k)})$ over pairs with $Z_i \perp \vartheta^{(k)}$, so $\mathbb{E}W_j = J_\otimes$. The bound is Theorem 4 of Maurer & Pontil (2009); empirical Bernstein inequalities for general U-statistics are due to Peel et al. (2010), and the blocking is Hoeffding's representation of U-statistics as averages of independent block means (Hoeffding, 1963). □

When $n \ge n_d$ the blocks contain one path each and, with $\bar{g}(\vartheta) = \mathbb{E}_Q \ell^\star(Z,\vartheta)$,

$$\mathrm{Var}(W_j) = \mathrm{Var}_\nu(\bar{g}(\vartheta)) + \mathbb{E}_\nu \mathrm{Var}_Q(\ell^\star \mid \vartheta)/\lfloor n/n_d \rfloor;$$

the model-channel part of the width is therefore $\sqrt{2\,\mathrm{Var}_\nu(\bar{g})\log(2/\delta)/n_d}$ plus a term of order $\log(1/\delta)/n_d$, whereas (a) never falls below $\sqrt{\log(1/\delta)/(2n_d)}$.

B.3 Proofs of Theorems A.1, 4.5 and 4.6 and Propositions A.2 and 4.7

Theorem A.1. Condition on the tolerance set and on the cell counts. Under Assumption 4.2, the N_j units of cell j are i.i.d. from the cell's conditional law, so $N_j^f \sim \mathrm{Bin}(N_j, R_j)$. If $R_j \ge \eta_j$, N_j^f stochastically dominates $N' \sim \mathrm{Bin}(N_j, \eta_j)$; since $k \mapsto F_{\eta_j}(k) = \mathbb{P}\{N' \le k\}$ is nondecreasing, $\mathbb{P}\{p_j \le \alpha\} \le \mathbb{P}\{F_{\eta_j}(N') \le \alpha\} \le \alpha$, i.e. p_j is a valid p-value (Clopper & Pearson, 1934; Lehmann & Romano, 2005). Holm's procedure controls the family-wise error rate for arbitrarily dependent p-values (Holm, 1979).

Theorem 4.5. Let G be the distribution function of Ψ under ν and $q = \inf\{r : G(r) \ge 1-\gamma\}$, so that $G(q^-) \le 1-\gamma$. The failure event $\{\nu(\Psi > \hat{r}) > \gamma\} = \{G(\hat{r}) < 1-\gamma\}$ is contained in $\{\hat{r} < q\} = \{\#\{k : \Psi_k < q\} \ge k^*\}$, and $\#\{k : \Psi_k < q\} \sim \mathrm{Bin}(n_d, G(q^-))$ is stochastically dominated by $\mathrm{Bin}(n_d, 1-\gamma)$; by the choice of k^* the failure probability is at most δ_d (Wilks, 1941). Because the deployment path is independent of the drift sample, on the complementary event $\mathbb{P}\{\Psi(\vartheta) > \hat{r}\} \le \gamma$; and for any return time, $\{T \le \bar{t},\ \vartheta_T \notin C_T(\hat{r})\} \subseteq \{\exists t : \psi_t(\vartheta_t) > \hat{r}\} = \{\Psi(\vartheta) > \hat{r}\}$. The tail $\mathbb{P}\{\mathrm{Bin}(n_d, 1-\gamma) \ge k\}$ decreases in k, so some k^* exists iff $k = n_d$ qualifies, i.e. iff $(1-\gamma)^{n_d} \le \delta_d$.

Theorem 4.6. Let $\mathcal{E}_d$ be the event that $\nu(\mathcal{C}(\hat{r}_\gamma)^c) \le \gamma$ for all $\gamma \in \Gamma$; by Theorem 4.5 and a union bound, $\mathbb{P}(\mathcal{E}_d) \ge 1 - \delta_d$. Conditionally on the drift sample, let $\mathcal{E}_b$ be the event that every rejected hypothesis is false; $\mathbb{P}(\mathcal{E}_b) \ge 1 - \delta_b$ by Theorem A.1. On $\mathcal{E}_d \cap \mathcal{E}_b$, a certified cell has some γ with $R^{\mathrm{cal}}(\mathcal{C}(\hat{r}_\gamma)) < \tau - \gamma - \varepsilon$, hence $R^{\mathrm{dep}}(\mathcal{C}(\hat{r}_\gamma)) < \tau - \gamma$ by Assumption 4.4. For any coupling, a failure requires either a robust failure (non-return, or rejection by some model in C_T) or a path leaving the family; by Theorem 3.1(i) with the product set of Remark B.1, $J(\varpi) \le R^{\mathrm{dep}}(\mathcal{C}(\hat{r}_\gamma)) + \nu(\mathcal{C}(\hat{r}_\gamma)^c) < \tau$. For the set certificate, the number of drift paths with $\Psi > r_0$ is $\mathrm{Bin}(n_d, \nu(\Psi > r_0))$, so $\hat{\gamma} \ge \nu(\mathcal{C}(r_0)^c)$ with probability at least $1 - \delta_d$; since $\hat{\gamma}$ depends only on the drift sample, the behavioral test at level $\tau - \hat{\gamma} - \varepsilon$ remains valid conditionally, and the argument is unchanged.

Proposition A.2. On $\{\Psi(\vartheta) \le \hat{r}_\gamma\}$, whose complement has probability at most γ on the event of Theorem 4.6, every subject satisfies $\ell^\star(Z_i,\vartheta) \le Y_i$ because all subjects face the same path. Conditionally on E the indicators Y_i are i.i.d. with mean $R(E)$, so Hoeffding's inequality (Hoeffding, 1963) gives $\mathbb{P}\{\bar{Y} > R(E) + \sqrt{\log(1/\delta')/(2N)} \mid E\} \le \delta'$, hence $\mathbb{P}\{\bar{Y} > t + \sqrt{\log(1/\delta')/(2N)}\} \le$

$\mathbb{P}\{R(E) > t\} + \delta'$. A union bound concludes; Markov's inequality applies because $\mathbb{E}R(E) = R^{\mathrm{dep}}(\mathcal{C}(\hat{r}_\gamma)) < \tau - \gamma$ on the event of Theorem 4.6.

Proposition 4.7. (a) By Theorem 4.5 at level $\delta_d/|\Gamma|$, a finite radius $\hat{r}_\gamma$ exists only if $(1-\gamma)^{n_d} \le \delta_d/|\Gamma|$, i.e. $\gamma \ge \gamma_{\min}$. If $\hat{r}_\gamma = \infty$, every returning unit is rejected by some model in $C_T(\infty)$, so $N^f = N$ and $p_{g,\sigma,\gamma} = 1$. For finite $\hat{r}_\gamma$ let $\eta = \tau - \gamma - \varepsilon$. Since $N^f \ge 0$, $p_{g,\sigma,\gamma} = \mathbb{P}\{\mathrm{Bin}(N,\eta) \le N^f\} \ge (1-\eta)^N$ (and $p_{g,\sigma,\gamma} = 1$ if $\eta \le 0$). Holm's procedure rejects only p-values at most δ_b, so a rejection requires $(1-\eta)^N \le \delta_b$, i.e. $\eta \ge 1 - \delta_b^{1/N}$; together with $\gamma \ge \gamma_{\min}$ this gives the claim. For the set certificate, the Clopper–Pearson bound is nondecreasing in the number of misses and equals $1 - \delta_d^{1/n_d}$ when there are none. (b) Let ϑ^{bad} be a path whose models reject every state at every time (for linear scores, $\theta = (-m, 0, \dots, 0)$ with $m > 0$), and let $\nu_p = (1-p)\nu_0 + p\,\delta_{\vartheta^{\mathrm{bad}}}$. Every unit fails on $\{\vartheta = \vartheta^{\mathrm{bad}}\}$, so $J(\varpi) \ge p$ for every coupling, and for $p > \tau$ validity gives $\mathbb{P}_{Q,\nu_p}(\text{certify}) \le \delta$. With probability $(1-p)^{n_d}$ all drift paths come from the ν_0 component, and on this event the data have exactly their law under (Q, ν_0) because the behavioral sample is independent of the drift sample. Hence $(1-p)^{n_d}\omega \le \delta$ for every $p > \tau$, and letting $p \downarrow \tau$ gives $\tau \ge 1 - (\delta/\omega)^{1/n_d}$. The same argument with $Q_q = (1-q)Q + q\,Q^{\mathrm{nr}}$, where Q^{nr} is the law of a non-returning unit with the same covariates (which fails under every coupling), applied to the N logged units of the cell, gives $\tau \ge 1 - (\delta/\omega)^{1/N}$. □

B.4 Proofs of Theorem 5.1, Corollaries 5.2 and 5.3, and Propositions A.3 and A.4

Theorem 5.1. Under randomized assignment the context population is the same for σ and σ'. For each x, $D_x(\mathcal{A}) = 0$, so $D_x^+(\mathcal{A}) = D_x^-(\mathcal{A}) = \mathrm{TV}_x$ and $\int u_x\, dD_x = \mathrm{TV}_x\big(\int u_x\, dP_x^+ - \int u_x\, dP_x^-\big)$; integrate over $X \mid G = g$.

Corollary 5.2. (i) $\int u\, dP^+ - \int u\, dP^- \le \sup_{\mathrm{supp}P^+} u - \inf_{\mathrm{supp}P^-} u$, and symmetrically. (ii) is the Kantorovich–Rubinstein duality $|\int u\, d(K - K')| \le \mathrm{Lip}(u)\, W_1(K, K')$ (Villani, 2009).

Proposition A.3. $\vartheta^\top v \sim N(\theta_0^\top v, s^2\|v\|^2)$, so $u_x(a) = \Phi(\rho(v))$ with $\rho(v) = \theta_0^\top v/(s\|v\|)$. Since $\nabla\rho(v) = \theta_0/(s\|v\|) - (\theta_0^\top v)v/(s\|v\|^3) = (I - \hat{v}\hat{v}^\top)\theta_0/(s\|v\|)$ with $\hat{v} = v/\|v\|$ is a scaled orthogonal projection of θ_0, we have $\|\nabla\rho\| \le \|\theta_0\|/(s\|v\|) \le \|\theta_0\|/s$ because the intercept gives $\|v\| \ge 1$. With $\Phi' \le (2\pi)^{-1/2}$ and $\partial v/\partial a = [0; I]$, the gradient of $a \mapsto u_x(a)$ has norm at most $\|\theta_0\|/(s\sqrt{2\pi})$, and the mean-value theorem gives the Lipschitz bound.

Proposition A.4. On the joint confidence event, Theorem 3.1(i) gives $J(\varpi) \ge Q(A_C) + \nu(C) - 1 \ge L_{g,\sigma} + L_\nu - 1 > \tau$ for every coupling and every σ. Under a behavioral shift with $\mathrm{TV} \le \varepsilon$, subtract ε from $L_{g,\sigma}$.

Corollary 5.3. (i) Under Assumption 4.2 both laws have the covariate law of context g, and the return time (with $T = \infty$ for non-return) has a common kernel given (X, A). Two joint laws that share a marginal and a conditional kernel are at total variation equal to the average total variation of their remaining conditional kernels, so $\mathrm{TV}(Q_{g,\sigma}, Q_{g,\sigma'}) = \mathbb{E}[\mathrm{TV}(K_\sigma(\cdot \mid X), K_{\sigma'}(\cdot \mid X)) \mid G = g]$. Theorem 5.1 with $|\mathbb{E}_{P^+}u - \mathbb{E}_{P^-}u| \le 1$ gives $\Lambda_u \le \mathbb{E}[\mathrm{TV}_X \mid G = g]$. (ii) A TV justification of Assumption 4.4 needs $\varepsilon \ge \mathrm{TV}(Q^{\mathrm{dep}}, Q^{\mathrm{cal}}) \ge \Lambda_u$ by (i). For the robust valuation $\mathrm{Acc}_g(\cdot) = 1 - R_{g,\cdot}(\mathcal{C})$, so $R_{g,\sigma'}(\mathcal{C}) - R_{g,\sigma}(\mathcal{C}) = \mathrm{Acc}_g(\sigma) - \mathrm{Acc}_g(\sigma')$, and the excess of the transferred bound is $\varepsilon + R_{g,\sigma'}(\mathcal{C}) - R_{g,\sigma}(\mathcal{C}) \ge \mathbb{E}[\mathrm{TV}_X \mid G = g] + \mathrm{Acc}_g(\sigma) - \mathrm{Acc}_g(\sigma')$, which is at least $2\Lambda_u$ when the difference is positive. □

B.5 Proofs of Theorems A.5 to A.8

Theorem A.5. Fix a law P in the class and $\epsilon > 0$. The function $r_\epsilon(x) = \max\{r(x), 1 - \|x - x_0\|/\epsilon\}$ is continuous, equals 1 at x_0, and coincides with r outside $B_\epsilon(x_0)$; by assumption it is realizable by changing the response kernel only on $B_\epsilon(x_0)$. The corresponding law P_ϵ of a logged unit satisfies $\mathrm{TV}(P, P_\epsilon) \le P\{X \in B_\epsilon(x_0)\}$, hence $\mathrm{TV}(P^n, P_\epsilon^n) \le nP\{X \in B_\epsilon(x_0)\}$. Validity under P_ϵ gives $P_\epsilon^n\{U \ge 1\} \ge 1 - \delta$, so $P^n\{U \ge 1\} \ge 1 - \delta - nP\{X \in B_\epsilon(x_0)\}$. Letting $\epsilon \downarrow 0$ and using $P\{X = x_0\} = 0$ proves the claim; the argument applies to any procedure that uses the full logged sample (cf. Foygel Barber et al., 2021).

Theorem A.6. Let $p_1 = \mathbb{P}_{\eta-\Delta}(\phi = 1) \geq 1-\zeta$ and $p_0 = \mathbb{P}_{\eta}(\phi = 1) \leq \delta < 1-\zeta$. By data processing and the monotonicity of $\mathrm{kl}(p\|q)$ (nondecreasing in $p \geq q$, nonincreasing in $q \leq p$), $n\ \mathrm{kl}(\eta-\Delta\|\eta) = \mathrm{KL}(P^n_{\eta-\Delta}\|P^n_{\eta}) \geq \mathrm{kl}(p_1\|p_0) \geq \mathrm{kl}(1-\zeta\|\delta)$, and $\mathrm{kl}(\eta-\Delta\|\eta) \leq \chi^2 = \Delta^2/(\eta(1-\eta))$. For the converse, the multiplicative Chernoff bounds give $\mathrm{kl}(p\|q) \geq (q-p)^2/(2q)$ for $p \leq q$ and $\mathrm{kl}(p\|q) \geq (p-q)^2/(p+q)$ for $p \geq q$. With threshold $\eta - \Delta/2$, the type-I error at any $R \geq \eta$ is at most $\exp\{-n\ \mathrm{kl}(\eta-\Delta/2\|\eta)\} \leq e^{-n\Delta^2/(8\eta)}$, and the type-II error at $R = \eta - \Delta$ is at most $\exp\{-n\ \mathrm{kl}(\eta-\Delta/2\|\eta-\Delta)\} \leq \exp\{-n(\Delta/2)^2/(2\eta-3\Delta/2)\} \leq e^{-n\Delta^2/(8\eta)}$.

Theorem A.7. Fix x and $y = y(x)$ with $\|x-y\| \leq h$. Every $x' \in B_{2h}(y)$ satisfies $\|x-x'\| \leq 3h$, hence $r_\sigma(x) \leq r_\sigma(x') + 3^\beta L h^\beta$; averaging over the conditional law of X given $X \in B_{2h}(y) \cap \mathcal{X}_g$ and $\Sigma = \sigma$ (which does not depend on σ under randomization) gives $r_\sigma(x) \leq R_{2h}(y) + 3^\beta L h^\beta$, where $R_{2h}(y)$ is the robust risk in the ball. The number N_y of units with $\Sigma = \sigma$ in the ball is $\mathrm{Bin}(n, p_y)$ with $p_y \geq c_0(2h)^d$; by Chernoff's bound $\mathbb{P}\{N_y < np_y/2\} \leq e^{-np_y/8} \leq \delta/(2N_h)$. Conditionally on N_y, Hoeffding's inequality gives $R_{2h}(y) \leq \hat{R}_{2h}(y) + \sqrt{\log(2N_h/\delta)/(2N_y)}$ except with probability $\delta/(2N_h)$, and $2N_y \geq nc_0(2h)^d$. A union bound over the N_h net points completes the proof. A regular grid of spacing $h/\sqrt{d}$ provides centres within $h/2$ of every point, and replacing each non-empty cell by one of its points in $\mathcal{X}_g$ yields an internal h-net with $N_h \leq (1+\sqrt{d}/h)^d$.

Theorem A.8. Let $\psi : \mathbb{R}^d \to [0,1]$ be $(1,\beta)$-Hölder with $\psi(0) = 1$ and support in the unit ball, let $a = \min\{L, 1/4\}$ and $r_1(x) = \frac{1}{2} + ah^\beta\psi((x-x_0)/h)$ for $h \leq 1$; then $r_1 \in \mathrm{Höl}(\beta, L)$ with values in $[1/4, 3/4]$, as is $r_0 \equiv \frac{1}{2}$. Using $\mathrm{kl}(p\|\frac{1}{2}) \leq 4(p-\frac{1}{2})^2$, $\mathrm{KL}(P^n_{r_1}\|P^n_{r_0}) \leq 4na^2h^{2\beta}\mathbb{P}\{\|X-x_0\| \leq h\} \leq 4na^2v_dh^{2\beta+d}$, where v_d is the volume of the unit ball. Choose $h = (\kappa/n)^{1/(2\beta+d)}$ with κ small enough that Pinsker's inequality gives $\mathrm{TV}(P^n_{r_1}, P^n_{r_0}) \leq (1-\delta)/2$. Validity under r_1 gives $\mathbb{P}_{r_1}\{U \geq r_1(x_0)\} \geq 1-\delta$, hence $\mathbb{P}_{r_0}\{U - r_0(x_0) \geq ah^\beta\} \geq (1-\delta)/2$, which is the claim with $c = a\kappa^{\beta/(2\beta+d)}$ and $c' = (1-\delta)/2$ (cf. Tsybakov & Zaiats, 2009; Low, 1997).

For the logged experiment, take $\mathcal{A} \supseteq \{a_-, a_+\}$ with $\ell(x, a_-) = 1$ and $\ell(x, a_+) = 0$, let recommendation σ induce $r(x)\delta_{a_-} + (1-r(x))\delta_{a_+}$, and let every other recommendation induce a fixed kernel that does not depend on r. A logged sample can then be simulated from an oracle sample by drawing Σ_i independently and setting $A_i = a_-$ if $Y_i = 1$ and $A_i = a_+$ otherwise whenever $\Sigma_i = \sigma$ (drawing A_i from the fixed kernel otherwise). Composing any bound that is valid in the logged experiment with this simulation yields a bound that is valid in the oracle experiment and has the same distribution under every r, so the lower bound transfers. □

C Experimental protocol and synthetic experiments

This appendix gives the complete protocol behind Section 7 and the results that did not fit there. Synthetic experiments isolate one assumption or phenomenon at a time (Appendices C.1–C.8). The LendingClub study runs the complete pipeline on real covariates and temporally refit classifiers; its behavioral and return-time channels are semi-synthetic (Appendix D).

C.1 Common certification protocol

RAC. For each budget γ in $\Gamma = \{0.05, 0.10, \ldots, 0.50\}$, the Wilks radius $\hat{r}_\gamma$ of Theorem 4.5 is computed from the n_d calibration path scores at content $1-\gamma$ and level $\delta_d/|\Gamma|$ ($\hat{r}_\gamma = \infty$ when no order statistic qualifies). A logged unit is a robust failure under $\mathcal{C}(\hat{r}_\gamma)$ if it does not return, or if some model in $\mathcal{C}_T(\hat{r}_\gamma)$ rejects its realized state at its realized return time T. For ball families this check is closed form; for the LendingClub polyhedra it is a linear program (Appendix D.4). With $N^f_{g,\sigma,\gamma}$ robust failures among $N_{g,\sigma}$ logged units, the hypothesis $H_{g,\sigma,\gamma} : R_{g,\sigma}(\mathcal{C}(\hat{r}_\gamma)) \geq \tau - \gamma - \varepsilon$ is tested with the exact binomial p-value $\Pr\{\mathrm{Bin}(N_{g,\sigma}, \tau-\gamma-\varepsilon) \leq N^f_{g,\sigma,\gamma}\}$. Throughout, $\delta_b = \delta_d = 0.025$ and $\varepsilon = 0$.

Holm's procedure is applied in one of two ways.

- *Synthetic studies (fixed-τ).* Holm runs over all (g, σ, γ) separately at each τ of the grid.

Table 4: Certificates compared in the experiments. In the synthetic studies, all finite-sample certificates except the CP model-only claim use $\delta = \delta_b + \delta_d = 0.05$ and Holm's procedure over cells.

Certificate	Bounds	Valid if
RAC (Algorithm 1)	$\sup_{\varpi} J(\varpi)$ over all couplings	Assumptions 4.2–4.4
Recorded env. (Prop. 3.4)	$\sum_e p_e J_\otimes(Q_e, \nu_e)$	$Z \perp \vartheta \mid E$; E recorded; p_e known
Dependence budget (Prop. 3.3)	$\mathrm{kl}_+^{-1}(J_\otimes, \iota)$	$I_\varpi(Z;\vartheta) \le \iota$ (here $\iota = H(E)$)
Product BD / EB (Prop. 3.5)	$J_\otimes$	$Z \perp \vartheta$
Plug-in $\hat{J}$	$J_\otimes$ (no correction)	$Z \perp \vartheta$, $N, n_d \to \infty$
Product oracle	population $J_\otimes$	$Z \perp \vartheta$
Model-only BD / CP	prescribed action, fixed update	full compliance, return at that update

- **LendingClub (all-τ).** Holm runs jointly over all $(g, \sigma, \gamma, \tau)$ to account for selecting the smallest certified level over the full grid.

In both cases $\tau^\star$ is the smallest certified grid value of a cell. In the synthetic studies we count a violation whenever $J^{\mathrm{dep}} > \tau^\star$ in some certified cell. This is stricter than the fixed-τ guarantee, which concerns one pre-specified τ at a time. Throughout, N denotes the number of logged units *per cell*.

Comparators. Table 4 lists every certificate, the quantity it bounds and the assumption under which it is valid. In the synthetic studies, every finite-sample comparator except the Clopper–Pearson model-only claim spends the total error budget $\delta = \delta_b + \delta_d = 0.05$, with Holm's procedure over cells at each τ; the plug-in rule and the product oracle use no error budget. On LendingClub the product bounds and the BD model-only bounds are instead Bonferroni-corrected over the 10 cells.

- *Product certificates.* These invert Proposition 3.5: a p-value is obtained by solving the bound for δ.
- *Dependence budget.* This certifies τ when the bounded-difference upper bound on $J_\otimes$ is at most $q_{\max}(\tau, \iota) = \sup\{q : \mathrm{kl}_+^{-1}(q, \iota) \le \tau\}$ (Proposition 3.3). Here $\iota = H(E)$ is the entropy of the environment: 0.500 nats in the baseline environment and $\ln 2$ in the stress environment.
- *Recorded-environment certificate.* This uses stratum-wise bounded-difference bounds combined with the known probabilities p_e and a union bound over strata (Proposition 3.4).
- *Model-only certificates.* These ignore the behavioral channel. They bound the probability that the *prescribed* action $x + \sigma$ fails under the model at a fixed update: the next one, the last one, or the worst over updates. Non-return is ignored.
  - The BD variant bounds the cell average with the p-value of Proposition 3.5(a); the worst-horizon version multiplies the largest per-horizon p-value by $\bar{t}$.
  - The Clopper–Pearson (CP) variant follows BetaRCE (Stępka et al., 2025). It averages per-applicant upper bounds at level 0.05, computed from the n_d calibration paths, and claims every τ above this average, without a multiplicity correction. For the worst-horizon version, the synthetic studies take the maximum over horizons per applicant, with level $0.05/\bar{t}$ per horizon; on LendingClub the claim is the larger of the two per-horizon averages.

Metrics. A repetition *violates* if at least one certified cell has deployment risk above its reported contract, $J^{\mathrm{dep}} > \tau^\star$. A reported frequency of zero means zero observed violations in the stated number of repetitions, not a zero population probability. We also report:

- the maximum over the τ grid of the violation frequency at fixed τ;
- the coverage, i.e. the share of contexts with a certified recommendation;
- $\tau_{.4}$ and $\tau_{.8}$, the smallest grid levels reaching coverage 0.4 and 0.8;
- the median slack $\tau^\star - J^{\mathrm{dep}}$ over certified cells, and the abstention rate.

Population quantities. Each cell gets a conditional risk matrix $R(e_b, e_m)$: the terminal failure rate when responses are drawn under environment e_b and model paths under e_m. It is computed with common random numbers from 2.5×10^5 units per cell. From it, $J_\otimes = \sum_{e_b,e_m} p_{e_b} p_{e_m} R(e_b, e_m)$. The deployment risk is $J^{\mathrm{dep}} = \sum_e p_e R(e, e)$ under the common shock and $J^{\mathrm{dep}} = J_\otimes$ under independent deployment. The sharp endpoint $J^\uparrow$ is the maximum bipartite matching between $8{,}000$ response records and $8{,}000$ model paths drawn from the two marginals, divided by $8{,}000$ and averaged over two seeds. On fresh instances this equals the optimum of the transport linear program over all couplings to machine precision. Population quantities are never available to any certificate.

Reproducibility firewall. The following use distinct seeds:

- geometry pilots;
- behavioral calibration samples;
- model-path calibration samples;
- evaluation samples;
- population truths.

Evaluation data never choose tolerance families, budgets, grids or generators. All tables and figures are regenerated from saved CSV/Parquet artifacts.

C.2 Synthetic environments and tolerance families

baseline environment. There are two contexts $g \in \{0, 1\}$, three recommendations $\sigma_1 = (0.90, 0.30, 0)$, $\sigma_2 = (0.30, 1.00, 0)$ and $\sigma_3 = (0.65, 0.65, 0)$, and a substitute response along $(0, 0, 1)$. Rejected applicants have current score $f_{\theta_0}(x) = -0.25 + x_1 + 0.8x_2 + 0.7x_3 \in [-1.25, 0)$. Effort costs are $c_0 = (1.10, 0.40, 0.45)$ and $c_1 = (0.40, 1.10, 0.45)$. Realized responses are full compliance, partial compliance, substitution or no action. Return times on $\{1, 2, 3\}$, or non-return, depend on the realized action's cost and on the environment (Appendix C.6). Models evolve as $\vartheta_t = \theta_0 + \sum_{u \le t} \Delta_u$, with Gaussian increments of standard deviations $(0.05, 0.03, 0.03, 0.03)$ for (intercept, $\beta_1, \beta_2, \beta_3$). A latent environment $E \in \{0, 1\}$ with $\Pr(E = 1) = 0.2$ moves both channels, scaled by the intensity $s \in [0, 1]$. Under $E = 1$, responses shift toward substitution and future models devalue x_3, with mean model shifts $(-0.03, 0, 0, -0.35)$, $(-0.10, 0, 0, -0.90)$ and $(-0.06, 0, 0, -0.55)$ at $t = 1, 2, 3$.

Stress environment. In the baseline environment the shock yields a dependence gap of at most 0.024 (Figure 2a). That is below the finite-sample slack of every product certificate, so their independence premise is never tested. The stress environment changes only the parameters in Table 5. Stressed applicants substitute strongly toward x_3, and stressed future models devalue x_3. Because pre-action x_3 has small spread, the devaluation mainly hurts the applicants who substituted: a matching-type coupling in the sense of Corollary 3.2.The stress environment is designed to make the dependence between the behavioral and model-drift channels large enough to be statistically consequential; its parameters were fixed before the reported experiments.

Tolerance families. Scores are standardized by the a priori scales $s_t = \sqrt{t}\,(0.05, 0.03, 0.03, 0.03)$.

- *Mixture family (main).* This is a union of $K = 2$ balls whose centre paths $m_{k,\cdot}$ come from 2-means clustering of an independent geometry pilot of 40 paths (200 in the settings with $n_d = 600$). The path score is $\Psi(\vartheta) = \min_k \max_{t \le \bar{t}} \|(\vartheta_t - m_{k,t}) \oslash s_t\|_2$, where $\oslash$ denotes coordinate-wise division; the tolerance set $\{\Psi \le r\} = \bigcup_k \prod_t \mathcal{B}_{k,t}(r)$ is a union of product sets whose time-t section is the union of balls $\bigcup_k \{\theta : \|(\theta - m_{k,t}) \oslash s_t\|_2 \le r\}$, against which the robust loss is evaluated (Remark B.1).
- *Single-ball family.* This fixes the centre at θ_0.

Write $v = (1, x + a)$ for the post-action state of a unit with features x and realized action a. A unit returning at time t fails robustly at radius r exactly when $\min_k m_{k,t}^\top v < r\, \|s_t \odot v\|_2$.

Table 5: Parameters that differ between the baseline and the stress synthetic environments. Utility shifts are the changes to the utilities of no action / partial compliance / full compliance / substitution under $E = 1$; model shifts are the changes to the mean of ϑ_t under $E = 1$, in (intercept, $\beta_1, \beta_2, \beta_3$). All shifts are multiplied by the intensity s.

Parameter	Baseline	Stress
$\Pr(E = 1)$	0.2	0.5
Spread of pre-action x_3 (multiplier)	1	0.1
Substitute action	$(0, 0, 1)$	$(0, 0, 2)$
Utility shifts	$+0.15$ / $+0.35$ / -0.90 / $+1.10$	0 / -1.67 / -5.0 / $+5.0$
Model shift, $t = 1$	$(-0.03, 0, 0, -0.35)$	$(0, 0, 0, -1.5)$
Model shift, $t = 2$	$(-0.10, 0, 0, -0.90)$	$(0, 0, 0, -1.5)$
Model shift, $t = 3$	$(-0.06, 0, 0, -0.55)$	$(0, 0, 0, -1.5)$

C.3 Dependence and validity

Design. The dependence sweep uses $s \in \{0, 0.05, 0.10, 0.20, 0.35, 0.50, 0.75, 1\}$ and 200 repetitions per point in three settings:

- the baseline environment at $N = 800$, $n_d = 120$;
- the stress environment at $N = 800$, $n_d = 120$;
- the stress environment at $N = 4000$, $n_d = 600$.

Contracts use the grid $\tau \in \{0.10, 0.11, \ldots, 0.99\}$. A second validity study covers four deployment laws, with 500 repetitions, $N = 800$, $n_d = 120$, a 40-path geometry pilot and the grid $\{0.15, 0.175, \ldots, 0.95\}$. The four laws are the baseline and stress environments, each under independent deployment and under the common shock at $s = 1$.

Results. Figure 2 and Tables 6–7 complete Figure 1(a) and Table 2.

- *baseline environment.* The dependence gap remains within the finite-sample slack of the product certificates throughout the sweep, and no violations of the finite-sample certificates are observed. The plug-in rule (at least 92% of repetitions), the product oracle (all repetitions for $s \geq 0.2$) and the model-only certificates (all repetitions) are.
- *Four deployment laws.* Table 7 reports every certificate under the four laws. In the baseline environment under the common shock, RAC reaches coverage 0.4 at $\tau = 0.725$ and coverage 0.8 at $\tau = 0.850$, and the plug-in rule's largest fixed-τ violation frequency is 0.92.

In the stress environment at $N = 4000$ the price of robustness is visible in Figure 2(c). RAC's median slack equals the identification gap $J^{\uparrow} - J^{\text{dep}}$ plus 0.10–0.15 at every s, while the median slack of product BD turns negative between $s = 0.2$ and $s = 0.35$. The recorded-environment certificate is valid at every s and, once the product certificates fail, it is the tightest valid method, with slack 0.10 at $N = 4000$ and 0.22 at $N = 800$.

In the stress environment the single-ball RAC family remains valid but abstains in 81% ($N = 800$) and 46% ($N = 4000$) of cells at $s = 0.2$. It abstains in at least 99% for $s \geq 0.35$, because one ball cannot cover the bimodal drift law. The family term in Appendix C.4 quantifies this. Model-only certificates are violated in 96–100% of repetitions in every setting, including independent deployment. They ignore partial compliance, substitution and non-return, and they evaluate a fixed update rather than the realized one.

Sharp endpoint. Three checks concern $J^{\uparrow}$.

- *Finite-support check.* Across 500 random finite coupling instances, the analytic and flow-based endpoints of Theorem 3.1 agree to 5.6×10^{-16} (upper) and 4.4×10^{-16} (lower). A multiplicity construction with ten model states has product risk 0.1 and sharp identified interval $[0, 1]$ (Corollary 3.2).

Table 6: Dependence sweep: frequency (200 repetitions) of a reported contract below J^{dep}, and median slack $\tau^\star - J^{\mathrm{dep}}$ over certified cells ("abst.": abstains in every repetition).

	Violation frequency					Median $\tau^\star - J^{\mathrm{dep}}$				
Certificate	s=0	0.2	0.35	0.5	1	s=0	0.2	0.35	0.5	1
Stress environment, $N = 800$, $n_d = 120$										
Plug-in $\hat{J}$	0.93	1.00	1.00	1.00	1.00	0.01	−0.04	−0.13	−0.19	−0.19
Product oracle	0.00	1.00	1.00	1.00	1.00	0.01	−0.04	−0.12	−0.19	−0.18
Product BD (3.5a)	0.00	0.00	0.74	0.99	1.00	0.15	0.10	0.01	−0.05	−0.05
Product EB (3.5b)	0.00	0.00	0.30	0.83	0.80	0.15	0.11	0.04	−0.01	−0.01
KL budget (3.3)	0.00	0.00	0.00	0.00	0.00	0.70	0.65	0.56	0.42	0.41
Recorded env. (3.4)	0.00	0.00	0.00	0.00	0.00	0.22	0.22	0.22	0.22	0.22
RAC, mixture	0.00	0.00	0.00	0.00	0.00	0.35	0.41	0.39	0.27	0.27
RAC, single ball	0.00	0.00	0.00	0.00	0.00	0.33	0.73	0.45	0.31	0.33
Model-only BD, next	1.00	1.00	1.00	1.00	1.00	0.00	−0.08	−0.25	−0.38	−0.39
Model-only CP, worst	1.00	1.00	1.00	1.00	1.00	−0.09	−0.16	−0.32	−0.47	−0.47
Stress environment, $N = 4000$, $n_d = 600$										
Plug-in $\hat{J}$	0.82	1.00	1.00	1.00	1.00	0.01	−0.04	−0.13	−0.19	−0.19
Product oracle	0.00	1.00	1.00	1.00	1.00	0.01	−0.04	−0.12	−0.19	−0.18
Product BD (3.5a)	0.00	0.32	1.00	1.00	1.00	0.07	0.02	−0.06	−0.13	−0.12
Product EB (3.5b)	0.00	0.98	1.00	1.00	1.00	0.05	0.00	−0.08	−0.13	−0.13
KL budget (3.3)	0.00	0.00	0.00	0.00	0.00	0.63	0.58	0.48	0.40	0.40
Recorded env. (3.4)	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.10	0.10
RAC, mixture	0.00	0.00	0.00	0.00	0.00	0.28	0.35	0.35	0.23	0.23
RAC, single ball	0.00	0.00	0.00	0.00	0.00	0.27	0.68	abst.	abst.	abst.
Model-only BD, next	1.00	1.00	1.00	1.00	1.00	−0.08	−0.16	−0.32	−0.46	−0.46
Model-only CP, worst	1.00	1.00	1.00	1.00	1.00	−0.10	−0.19	−0.35	−0.48	−0.49
baseline environment, $N = 800$, $n_d = 120$										
Plug-in $\hat{J}$	0.95	0.98	0.98	0.99	1.00	0.00	0.01	0.00	0.00	−0.01
Product oracle	0.00	1.00	1.00	1.00	1.00	0.01	0.01	0.00	−0.01	−0.01
Product BD (3.5a)	0.00	0.00	0.00	0.00	0.00	0.15	0.15	0.14	0.14	0.13
Product EB (3.5b)	0.00	0.00	0.00	0.00	0.00	0.15	0.15	0.15	0.15	0.14
KL budget (3.3)	0.00	0.00	0.00	0.00	0.00	0.62	0.63	0.62	0.61	0.59
Recorded env. (3.4)	0.00	0.00	0.00	0.00	0.00	0.21	0.21	0.21	0.21	0.21
RAC, mixture	0.00	0.00	0.00	0.00	0.00	0.38	0.41	0.43	0.45	0.48
RAC, single ball	0.00	0.00	0.00	0.00	0.00	0.36	0.50	0.62	0.60	0.58
Model-only BD, next	1.00	1.00	1.00	1.00	1.00	−0.03	−0.04	−0.06	−0.08	−0.12
Model-only CP, worst	1.00	1.00	1.00	1.00	1.00	−0.12	−0.13	−0.14	−0.15	−0.17

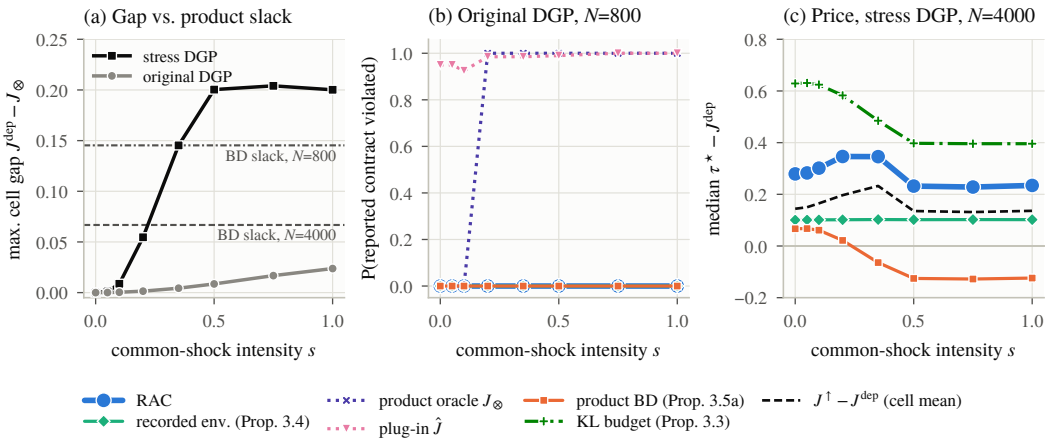

Figure 2: Dependence sweep, 200 repetitions per point. **(a)** Largest cell gap $J^{\text{dep}} - J_{\otimes}$ against the median slack of product BD at $s = 0$. **(b)** baseline environment: the gap stays below the slack, so of the certificates shown only the plug-in rule and the product oracle fail. **(c)** Stress environment at $N = 4000$: median $\tau^\star - J^{\text{dep}}$ over certified cells (negative means violated) and the cell-mean identification gap.

Table 7: Validity under four deployment laws (500 repetitions, $N = 800$, $n_d = 120$). Viol.: frequency of a reported contract below J^{dep}; $\tau_{.8}$: smallest grid level at which at least 80% of contexts have a certified recommendation ("–": never reached). The maximum fixed-τ violation frequencies of the plug-in rule are 0.62, 0.92, 0.53 and 1.00, those of product BD 0, 0, 0 and 0.99, and those of product EB 0, 0, 0 and 0.71 (columns left to right).

	baseline environment				Stress environment			
	Independent		Shock $s = 1$		Independent		Shock $s = 1$	
Certificate	Viol.	$\tau_{.8}$	Viol.	$\tau_{.8}$	Viol.	$\tau_{.8}$	Viol.	$\tau_{.8}$
Plug-in $\hat{J}$	0.84	0.275	0.99	0.350	0.79	0.225	1.00	0.425
Product oracle	0.00	0.275	1.00	0.350	0.00	0.225	1.00	0.425
Product BD (3.5a)	0.00	0.400	0.00	0.500	0.00	0.375	0.99	0.575
Product EB (3.5b)	0.00	0.425	0.00	0.500	0.00	0.375	0.77	0.625
KL budget (3.3)	0.00	0.900	0.00	0.950	0.00	0.925	0.00	–
Recorded env. (3.4)	0.00	0.475	0.00	0.575	0.00	0.450	0.00	0.850
RAC, mixture	0.00	0.650	0.00	0.850	0.00	0.600	0.00	0.900
RAC, single ball	0.00	0.625	0.00	0.950	0.00	0.575	0.00	–
Model-only BD, next	1.00	0.225	1.00	0.250	1.00	0.225	1.00	0.225
Model-only BD, worst	1.00	0.250	1.00	0.300	0.96	0.250	1.00	0.250
Model-only CP, worst	1.00	0.150	1.00	0.200	1.00	0.150	1.00	0.150

- *Matching versus transport LP.* On fresh samples from both environments, the matching estimate equals the transport linear program over all couplings to 5.6×10^{-16}.
- *Finite-sample bias.* As a plug-in estimate from finite samples, $\hat{J}^{\uparrow}$ is biased *downward*: in the stress environment (g_0, σ_1, $s = 1$) it rises from 0.767 at $n = 500$ to 0.790 at $n = 8{,}000$ and 16,000. Extrapolation in $n^{-1/2}$ and $n^{-1/3}$ gives limits 0.796 and 0.803. The estimate used in our tables ($n = 8{,}000$) therefore understates $J^{\uparrow}$ by about 0.01–0.02. This is well below RAC's smallest margin over $\hat{J}^{\uparrow}$ (0.067), so no validity statement changes. It does slightly understate the identification term and overstate the family term in Appendix C.4; their sum is unaffected.

C.4 Decomposition of the contract

For a certified cell, let $R(r)$ and $\nu(r)$ be the population robust risk and miss mass of the repetition's own tolerance family at radius r. They are estimated from 6×10^4 response records per cell and

Table 8: Decomposition of RAC's slack (means over cells and 100 repetitions; LendingClub: certified cells, held-out banks). $J^{\uparrow}$ is the plug-in endpoint (see Appendix C.3).

Setting	Regime	Ident.	Family	Radius	Nom. γ	Behavior	Total
Baseline, independent	N=800, n_d=120	0.178	0.099	0.002	0.053	0.060	0.392
	N=4000, n_d=600	0.178	0.079	0.001	0.023	0.029	0.311
Baseline, $s = 1$	N=800, n_d=120	0.296	0.074	0.003	0.049	0.053	0.475
	N=4000, n_d=600	0.296	0.069	0.002	0.025	0.026	0.419
Stress, $s = 1$	N=800, n_d=120	0.137	0.037	0.004	0.042	0.047	0.267
	N=4000, n_d=600	0.137	0.035	0.009	0.024	0.023	0.228
LendingClub, ROAR g_1	$N \approx 5000$, n_d=200	0.174	0.088	0.011	0.050	0.025	0.347
LendingClub, Wachter g_1	$N \approx 5000$, n_d=200	0.250	0.098	0.002	0.050	0.012	0.412

6×10^4 paths. Then

$$\tau^\star - J^{\mathrm{dep}} = \underbrace{(J^{\uparrow} - J^{\mathrm{dep}})}_{\text{identification}} + \underbrace{(J_{\mathrm{fam}} - J^{\uparrow})}_{\text{family}} + \underbrace{\big(\min_{\gamma\in\Gamma}[R(\hat{r}_\gamma) + \nu(\hat{r}_\gamma)] - J_{\mathrm{fam}}\big)}_{\text{model radius}}$$
$$+ \underbrace{\big(\min_{\gamma\in\Gamma}[R(\hat{r}_\gamma) + \gamma] - \min_{\gamma\in\Gamma}[R(\hat{r}_\gamma) + \nu(\hat{r}_\gamma)]\big)}_{\text{nominal budget}} + \underbrace{\big(\tau^\star - \min_{\gamma\in\Gamma}[R(\hat{r}_\gamma) + \gamma]\big)}_{\text{behavior}},$$

with $J_{\mathrm{fam}} = \min_r\{R(r) + \nu(r)\}$ over a grid of 3,000 radii. The terms have the following meanings:

- The first two are population quantities. Identification is non-negative by Theorem 3.1(ii), and family is non-negative because J_{fam} is a set certificate.
- The model-radius term is the cost of the Wilks radii and of the grid Γ.
- The nominal-budget term pays γ instead of the realized miss mass.
- The behavior term collects the exact binomial test, Holm's correction and the 0.005 contract grid.

The identity holds to 4.4×10^{-16} over all 3,600 certified cell-repetitions (100 repetitions per scenario and regime; RAC certified every cell). Identification is at least 0.105 and family at least 0.024 in every one of them. The behavior term is negative in 1 of the 3,600 cell-repetitions and the nominal-budget term in the six cells of one repetition, each by at most 0.006; this is what the error allowances δ_b and δ_d permit. Table 8 gives cell means.

Going from $N = 800$, $n_d = 120$ to $N = 4000$, $n_d = 600$ leaves identification unchanged (it is a population quantity) and roughly halves the nominal-budget and behavior terms. The family term shrinks slightly because the larger geometry pilot gives better centres, and the model-radius term is below 0.01 on average in every setting. For LendingClub the same identity is evaluated on the held-out banks for the two certified cells.

C.5 Finite-sample phase diagram

We vary the behavioral sample size $N \in \{50, 100, 250, 500, 1000, 2500, 5000\}$ and the model history $n_d \in \{5, 10, 20, 30, 50, 100, 200, 500\}$ independently, in the baseline environment (Figure 3).

- *Behavioral side.* The zero-failure floor falls from 0.1442 at $N = 50$ to 0.0016 at $N = 5000$.
- *Model side.* No budget on Γ is feasible at $n_d = 5$. The smallest feasible budget is then $.50, .30, .20, .15, .10, .05, .05$ for $n_d = 10, 20, 30, 50, 100, 200, 500$.

Neither source of data compensates for a lack of the other: the statistical floor falls only when both grow. The median level certifying 80% of cells decreases as either grows, and it is systematically lower under independent deployment than under the common shock. At the fixed levels $\tau \in \{.50, .70, .85, .95\}$ there are zero observed violations under both deployment laws, so the transition is one of certifiability, not validity.

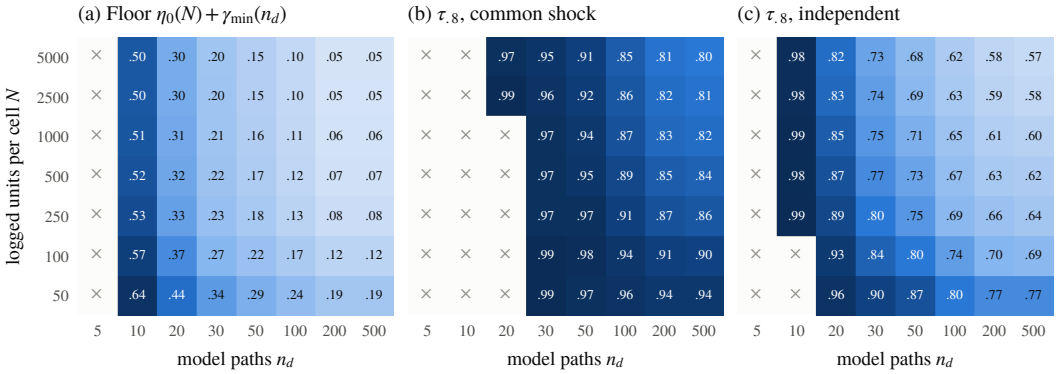

Figure 3: Finite-sample phase diagram in the baseline environment (cell values printed). **(a)** Statistical floor $\eta_0(N) + \gamma_{\min}(n_d)$; $\times$: no budget on Γ has a finite Wilks radius. **(b, c)** Median smallest τ certifying at least 80% of cells, under common-shock and independent deployment; $\times$: never reached on the contract grid.

C.6 Random return times

Return-time mechanism. In the baseline environment at $s = 1$, a unit with realized action cost c and environment E fails to return with probability $p_{\mathrm{nr}}(c, E)$. Otherwise it returns at $t \in \{1, 2, 3\}$ with probability $p_0(t \mid c, E) \propto \exp\{-(t - \mu(c, E))^2/(2 \cdot 0.55^2)\}$. We keep p_{nr} and tilt the return time toward selected updates, $\Pr(T = t) \propto p_0(t)\, e^{\xi z_t}$, where $\xi = \infty$ returns at $\arg\max_t z_t$ (an adversarial stopping time). Let $\mu_{E,t} = \mathbb{E}[\vartheta_t \mid E]$ and let v be the post-action state. We use two triggers:

- *Atypical updates:* $z_t = \|(\vartheta_t - \mu_{E,t}) \oslash s_t\|_2$. This targets the event the model set must cover.
- *Tightening updates:* $z_t = -(\vartheta_t - \mu_{E,t})^\top v / \|s_t \odot v\|_2$. This returns applicants when the update lowers their own score, which targets the contract.

Tilting moves mass from $T = 2$ to $T = 1$ and $T = 3$. Under the atypical trigger, $\Pr(T = 1, 2, 3)$ moves from $(0.244, 0.578, 0.115)$ at $\xi = 0$ to $(0.305, 0.466, 0.167)$ at $\xi = 4$ and $(0.350, 0.268, 0.319)$ at $\xi = \infty$. The tightening trigger is similar: $(0.315, 0.447, 0.175)$ at $\xi = 4$ and $(0.364, 0.251, 0.322)$ at $\xi = \infty$. Non-return stays at 0.063.

Families compared. All families use budget γ and level $\delta_d/|\Gamma|$.

- *Path-wise* (Theorem 4.5): the mixture family of Appendix C.2.
- *Per-time:* at each t, a two-ball set from 2-means clustering of the pilot slice $\{\vartheta_t\}$, with a Wilks radius on the slice scores at content $1 - \gamma$.
- *Per-time Bonferroni:* the same sets at content $1 - \gamma/\bar{t}$ and level $\delta_d/(|\Gamma|\bar{t})$.
- *Next-update* and *final-horizon* sets: a single slice set applied at every return time.

Model miss (Part A). We use 200 calibrations at each $n_d \in \{120, 1000\}$ with a 40-path pilot. Each calibration is evaluated on 10^4 units per cell. Every unit carries its own model path, and return times are drawn with common random numbers across ξ. For each calibration we record the miss rate $\Pr(\vartheta_T \notin \mathcal{C}_T)$ among returners, pooled over cells, and whether it exceeds γ (PAC failure).

- *Untilted return times.* At $\xi = 0$, $n_d = 120$ and $\gamma = 0.10$ the miss rates are 0.013 (path-wise), 0.330 (next-update) and 0.154 (final-horizon): single-horizon sets exceed γ even without tilting.
- *Path-wise sets.* Across all 48 settings (two values of n_d, two triggers, six values of ξ, and $\gamma \in \{0.10, 0.20\}$), at most 2 of 200 calibrations had a miss above γ, against a target of 0.0025 (smallest binomial $p = 0.09$). The mean miss at each return time also stays below

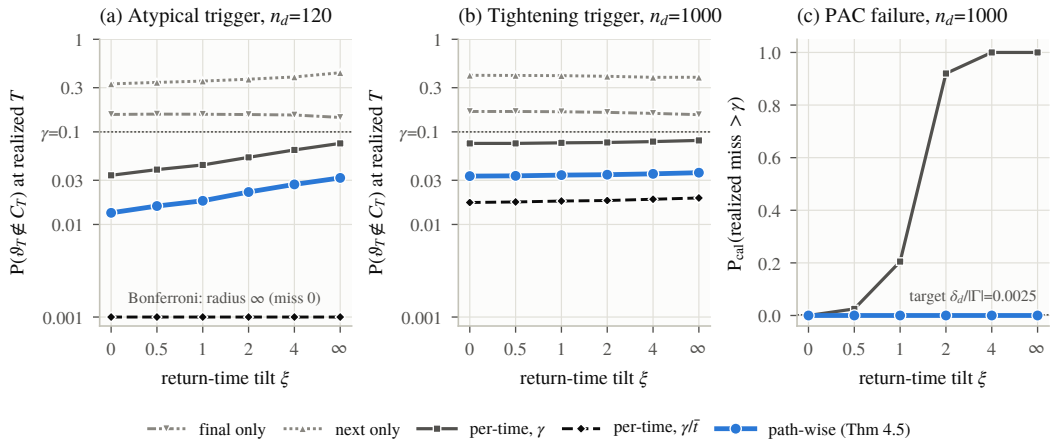

Figure 4: Random return times, $\gamma = 0.10$, 200 calibrations. **(a, b)** Mean model miss at the realized return time under the atypical trigger ($n_d = 120$) and the tightening trigger ($n_d = 1000$). At $n_d = 120$ the Bonferroni radius is infinite, so its miss is zero (drawn at 10^{-3}). **(c)** Fraction of calibrations whose realized-time miss exceeds γ (atypical trigger, $n_d = 1000$).

γ in every setting; for example it is 0.071, 0.079 and 0.075 at $T = 1, 2, 3$ for $n_d = 1000$, $\xi = \infty$ and the atypical trigger.

- *Per-time sets at* $\gamma = 0.20$. They fail in the same way as at $\gamma = 0.10$: miss up to 0.316, and PAC failure 1.0 for $\xi \geq 2$ at $n_d = 1000$.
- *Tightening trigger.* It moves the per-time miss only to 0.081, below γ (Figure 4b). The atypical trigger is the one that exposes the difference between families.

End-to-end contracts (Part B). RAC is run with the path-wise, per-time and Bonferroni families under the tightening trigger. The settings are $\xi \in \{0, 4, \infty\}$, $N = 4000$ per cell, $n_d \in \{120, 1000\}$ and 200 repetitions, with deployment risks from 2×10^5 units per cell. No family's contract was violated in any setting (Table 10).

- *Per-time contracts* are 0.01–0.02 tighter than path-wise ones. They are, however, not covered by Theorem 4.5, and in this environment their extra model miss is absorbed by the slack of the robust loss.
- *Bonferroni contracts* are valid. At $n_d = 120$ they are 0.13–0.14 looser than path-wise ones and abstain in 4–19% of cells, since a finite radius at $\gamma = 0.10$ needs $n_d \geq 210$ (path-wise: 57). At $n_d = 1000$ they are only 0.01 looser.

C.7 Behavioral shift and misspecification

These experiments probe Assumption 4.4 (Figure 5).

Cost of a valid shift budget. We interpolate between the calibration and a stressed deployment response law with $\lambda \in \{0, .1, .2, .4, .6, .8, 1\}$. The exact total-variation distance grows linearly to $\varepsilon_{\mathrm{TV}} = 0.42$. The mean terminal-risk shift reaches 0.183, with a maximum cellwise shift of 0.202. With the valid TV correction, $\tau_{.4}$ rises from 0.52 to 0.94. $\tau_{.8}$ rises from 0.59 to 0.92 at $\lambda = 0.8$, and coverage 0.8 is no longer reached at $\lambda = 1$ (maximum coverage 0.685). The uncorrected rule is tighter and happened not to be violated here, but its transport premise is false for every $\lambda > 0$.

More data does not repair a wrong behavioral law. We use targets midway between the calibration-law and deployment-law risks, which the unshifted rule certifies asymptotically while the correct law rejects them. At $\lambda = 1$ the false-certification probability of the misspecified rule rises with N: 0.253 at $N = 50$, 0.617 at 100, 0.987 at 250 and 1.000 from 500 on. It reaches 0.98 by $N = 5000$ even at $\lambda = 0.2$. The correct-deployment reference has at most 0.0133, and the TV-corrected rule has no false certification.

Table 9: Model miss at the realized return time, atypical trigger, $\gamma = 0.10$ (mean over 200 calibrations), and PAC failure frequency.

		Mean miss					PAC failure		
n_d	ξ	Path	Time	Bonf.	Next	Final	Path	Time	Bonf.
120	0	.013	.034	.000	.330	.154	.000	.000	.000
	0.5	.016	.039	.000	.342	.156	.000	.000	.000
	1	.018	.044	.000	.354	.155	.000	.025	.000
	2	.022	.053	.000	.371	.154	.000	.035	.000
	4	.027	.064	.000	.392	.152	.000	.070	.000
	∞	.032	.075	.000	.435	.143	.010	.170	.000
1000	0	.033	.075	.017	.408	.166	.000	.000	.000
	0.5	.039	.085	.020	.424	.169	.000	.025	.000
	1	.043	.094	.023	.440	.170	.000	.205	.000
	2	.053	.112	.029	.462	.173	.000	.920	.000
	4	.063	.134	.035	.489	.177	.000	1.00	.000
	∞	.075	.157	.041	.539	.183	.000	1.00	.000

Table 10: End-to-end RAC contracts, tightening trigger: median $\tau^\star - J^{\mathrm{dep}}$ and abstention. No violations in any setting.

		Median slack			Abst.
ξ	n_d	Path	Time	Bonf.	Bonf.
0	120	.454	.441	.580	.192
	1000	.430	.414	.444	.000
4	120	.427	.417	.557	.127
	1000	.403	.386	.416	.000
∞	120	.408	.394	.544	.038
	1000	.382	.362	.394	.000

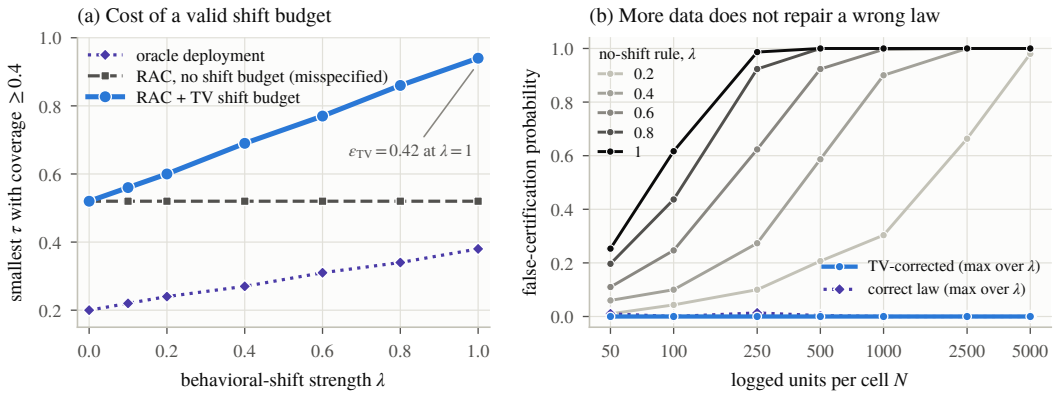

Figure 5: Behavioral shift. **(a)** Smallest τ with coverage at least 0.4 as deployment behavior moves away from the calibration law. The TV-corrected rule is valid; the no-shift rule is a misspecified diagnostic; the oracle knows the deployment law. **(b)** The false-certification probability of the no-shift rule grows with N at every shift strength λ, while the correct deployment law and the TV-corrected rule (maximum over λ) stay near zero.

C.8 Leverage and transferability

The finite-menu identity of Theorem 5.1 is first checked on the baseline environment. Some response kernels there have nearly disjoint supports: two recommendations with $\mathbb{E}[\mathrm{TV}_X] = 0.96$ and oscillation 0.98 differ in acceptance by only 0.010. To separate support mismatch from behavioral separation, a common-support variant varies the separation between recommendation-induced kernels through $\kappa \in \{0, .25, .5, 1, 2, 4, 8\}$. Table 11 and Figure 6 show the results. Mean total variation rises to essentially 1, while mean leverage saturates near 0.052 (at $\kappa = 2$: TV 0.907, leverage 0.049). No TV-transferred bound is vacuous for $\kappa \leq 0.5$, a third are vacuous at $\kappa = 1$, and all are vacuous for $\kappa \geq 2$. The decomposition of Theorem 5.1 holds to 2.0×10^{-16}. This is the leverage–certifiability tension of Corollary 5.3: evidence cannot be borrowed across recommendations long before a recommendation delivers a comparable benefit.

Table 11: Leverage and transfer on the common-support variant (all context–recommendation pairs). Λ: leverage; vacuous: fraction of TV-transferred bounds that are at least one; error: largest absolute error of the Theorem 5.1 decomposition.

κ	Mean TV	Max TV	Mean Λ	Max Λ	Vacuous	Error
0	0.000	0.000	0.0000	0.0000	0.000	0
0.25	0.152	0.176	0.0084	0.0179	0.000	2.1×10^{-17}
0.5	0.311	0.350	0.0171	0.0358	0.000	2.1×10^{-17}
1	0.599	0.645	0.0327	0.0675	0.333	5.2×10^{-17}
2	0.907	0.926	0.0492	0.1015	1.000	5.9×10^{-17}
4	0.997	0.998	0.0526	0.1091	1.000	1.3×10^{-16}
8	1.000	1.000	0.0524	0.1087	1.000	2.0×10^{-16}

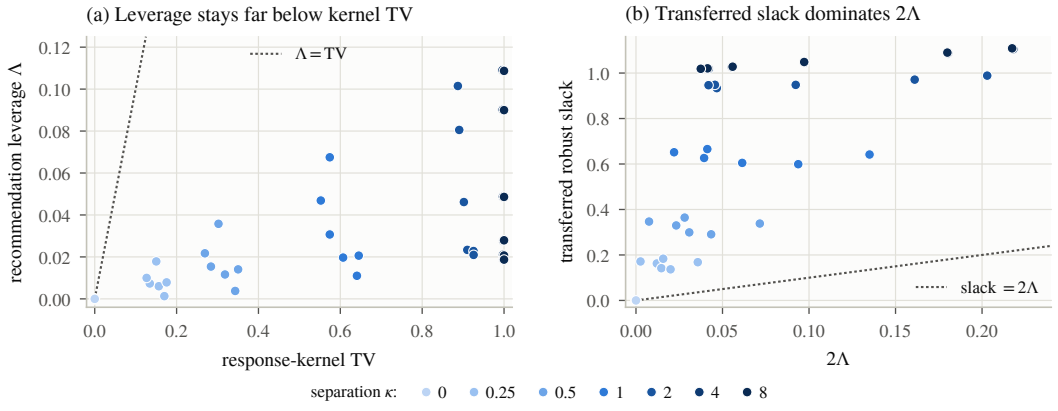

Figure 6: Leverage and transfer on the common-support variant: all context–recommendation pairs, colored by the separation κ. **(a)** Pairwise leverage lies far below the response-kernel TV (dotted: $\Lambda = \mathrm{TV}$). **(b)** The transferred robust slack dominates the 2Λ term of Corollary 5.3 (dotted: slack $= 2\Lambda$).

D LendingClub experiments

D.1 Data, models and the model-path law

The cleaned LendingClub data (George, 2019) contain 850,945 loans and 26 retained columns. They are split chronologically:

- current-model training, 2007–2012: 72,566 loans;
- behavioral reference period, 2013: 100,422;
- model-drift pool, 2014–2015: 445,596;
- external mismatch period, 2016: 232,361.

Outcomes are constructed from 36-month loans with a final status: "Fully Paid" is favorable, and "Charged Off" or "Default" is unfavorable. The actionable variables and their allowed directions are:

- loan amount (decrease);
- debt-to-income ratio (decrease);
- annual income (increase);
- revolving balance (decrease);
- revolving utilization (decrease).

Table 12: Confirmatory model-path calibration ($n_d = 200$, level $\delta_d/|\Gamma| = 0.0025$): Wilks order statistic k, radius $\hat{r}_\gamma$, exact binomial tail, and calibration and held-out miss rates.

γ	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
k	198	192	184	176	168	159	150	140	131	121
$\hat{r}_\gamma$	3.704	3.421	3.226	3.134	2.999	2.893	2.818	2.769	2.695	2.624
Binomial tail	.00234	.00139	.00214	.00196	.00145	.00161	.00154	.00213	.00163	.00182
Cal. miss	.010	.040	.080	.120	.160	.205	.250	.300	.345	.395
Held-out miss	.000	.080	.125	.150	.210	.250	.285	.330	.375	.405

Preprocessing is fit on source data only. The current logistic model uses a source-only out-of-fold threshold, giving out-of-fold AUC 0.662, balanced accuracy 0.618 and a rejection rate of 0.495.

Yearly refits with the same preprocessing accept a decreasing share of a fixed set of 2013 rejected applicants: 0.227 (2013), 0.157 (2014), 0.109 (2015) and 0.012 (2016). Their coefficient displacements from the current model have norms 1.173, 1.222, 1.253 and 1.263, and consecutive directions have cosine similarities 0.955, 0.990 and 0.983.

Model-path law and tolerance family. Each path bootstrap-samples up to $50{,}000$ loans from each of the 2014 and 2015 pools and refits the frozen logistic model. A pilot bank of 200 paths defines the prediction-space geometry. For the 800 reference states a_j, let $m_{tj} = \bar{\theta}_t^\top a_j$ and s_{tj} be the pilot mean and pilot standard deviation of $\theta_t^\top a_j$. The score and sets are

$$\psi_t(\theta) = \max_j \big[(m_{tj} - a_j^\top \theta)/s_{tj}\big]_+, \qquad \mathcal{C}_t(r) = \{\theta : a_j^\top \theta \geq m_{tj} - r\, s_{tj}\ \forall j\}.$$

The path score Ψ is the maximum of ψ_t over the two future times. The 800 states are 40 applicants, times five generators, times four response modes. After this family is frozen, fresh confirmatory calibration and evaluation banks of 200 paths each are drawn. Their path scores are close:

- mean 2.514 vs. 2.527;
- median 2.481 vs. 2.500;
- 95% quantile 3.395 vs. 3.536.

Table 12 gives the calibration. Recomputing it from the saved path scores reproduces every radius and binomial tail exactly.

D.2 Recourse generators and common support

All five generators act on the same actionable domain:

- *AR* solves the exact linear problem with normalized ℓ_1 cost.
- *Wachter* uses MAD-scaled ℓ_1 distance and a squared target penalty, with its coefficient chosen on a source-only grid.
- *ROAR* uses an exact linear robust margin under an ℓ_2 ball of radius 0.1 of the normalized parameter scale.
- *PROBE* uses the closed-form individual-robustness target under Gaussian implementation noise (variance 0.01, target 0.35).
- *RBR* uses the authors' robust loss on the same domain.

Recommendations are generated for 100 candidate applicants, 50 per context. At target margin zero, 40 applicants (8 in g_0, 32 in g_1) receive a recommendation from every generator (Table 13). Margins 0.10 and 0.25 leave no five-generator common support because RBR issues nothing there. They are studied with four generators in Appendix D.5.

Table 13: Generators at margin zero: fraction of the 50 applicants per context with a candidate / issued recommendation, and mean normalized cost on the common support.

Generator	Cand./issued		Mean cost	
	g_0	g_1	g_0	g_1
AR	.48/.48	1.00/1.00	2.047	0.510
PROBE	.42/.42	1.00/1.00	2.144	0.588
RBR	.48/.18	1.00/.64	2.173	0.766
Wachter	1.00/.48	1.00/1.00	2.913	2.234
ROAR	1.00/.48	1.00/1.00	4.278	4.183

Table 14: Held-out behavioral bank (10^5 units): response modes, return rate, validity under the current model, mean realized cost, and $\Pr(T = 1 \mid T < \infty)$.

Gen.	g	Full	Part.	Sub.	None	Ret.	Valid	Cost	$T{=}1$
AR	0	.465	.211	.132	.192	.800	.465	1.366	.462
	1	.489	.209	.142	.160	.867	.366	0.357	.565
PROBE	0	.460	.210	.141	.189	.780	.460	1.435	.458
	1	.487	.205	.147	.162	.873	.510	0.410	.555
RBR	0	.461	.211	.137	.191	.798	.461	1.453	.467
	1	.483	.211	.147	.159	.861	.585	0.533	.557
ROAR	0	.391	.231	.135	.242	.733	.507	2.594	.426
	1	.456	.210	.144	.190	.784	.804	2.805	.479
Wachter	0	.451	.213	.139	.197	.766	.522	1.926	.447
	1	.477	.209	.148	.166	.833	.804	1.535	.517

D.3 Behavioral and return-time channel

LendingClub has no logs of responses to recourse, so the behavioral channel is simulated from a mechanism frozen in advance. Each issued recommendation leads to full compliance, partial compliance, substitution or no action. Choice probabilities follow a softmax utility built from four ingredients:

- current-model score improvement;
- normalized action burden;
- a response-type bonus;
- context-dependent cost sensitivity, perturbed log-normally per individual.

The mechanism has no access to future classifiers. Given the response and context, the applicant returns after one or two updates or never ($T \in \{1, 2, \infty\}$); non-return counts as failure.

Confirmatory units draw $G \sim \mathrm{Bernoulli}(1/2)$ and an applicant uniformly from the common support in G, and then assign one of the five generators with probability 0.2. The calibration and evaluation banks hold $50{,}000$ and $100{,}000$ units. Assignment frequencies deviate from 0.2 by at most 0.0024, and context shares from 0.5 by at most 0.0024.

The banks depart materially from prescribed recourse (Table 14). Full compliance is 0.39–0.49 and return rates are 0.73–0.87. A pre-specified audit of 80 calibration-versus-evaluation statistics has no Holm rejection at family-wise level 0.05. The largest standardized discrepancy is 2.87 (Holm-adjusted $p = 0.334$).

D.4 Certification details and full results

Exact robust thresholds. Following the confirmatory protocol, a realized state a returning at time t is a robust success at radius r only if $\min_{\theta \in \mathcal{C}_t(r)} a^\top \theta \geq \epsilon = 10^{-9}$; borderline states count as failures. By linear-programming duality this holds exactly when $r \leq r^\star(a)$, where

$$r^\star(a) = \max\Big\{ \frac{m_t^\top w - \epsilon}{s_t^\top w} : A^\top w = a,\ w \geq 0 \Big\}$$

and A stacks the reference states $a_j^\top$ ($r^\star = -\infty$ when the program is infeasible). We solve this linear-fractional program after the Charnes–Cooper transformation with HiGHS, and fall back to bisection on the primal problem if the solver reports neither optimality nor infeasibility. The thresholds agree with solving the robust-failure LP directly at each budget for all 16,000 state–budget pairs and give the same 140 contract decisions. They also make contracts on the finer grid $\{0.30, 0.31, \ldots, 0.99\}$ cheap to evaluate.

Multiplicity. There are three nested hypothesis families:

Table 15: LendingClub risks on the held-out banks. Plug-in, BD and EB are product-coupling estimates and bounds (Bonferroni over cells) from the calibration bank; J_{fam} is the best certificate in the prediction-space family; "presc." is $J_{\otimes}$ for the prescribed full action; the model-only columns are claims about the prescribed action (†: below the realized $J_{\otimes}$).

Gen.	g	Plug-in	BD	EB	$J_{\otimes}$	$J^{\uparrow}$	$J^{\uparrow} - J_{\otimes}$	J_{fam}	Presc.	MO-BD next	MO-CP worst
AR	0	.828	.946	.918	.840	.943	.103	1.000	.764	1.000	.773†
	1	.701	.818	.796	.693	.927	.234	.952	.672	.855	.764
PROBE	0	.838	.956	.925	.835	.945	.110	1.000	.761	1.000	.767†
	1	.685	.802	.784	.696	.927	.231	.961	.670	.837	.754
RBR	0	.833	.951	.921	.830	.944	.114	1.000	.760	1.000	.769†
	1	.681	.798	.777	.687	.929	.243	.960	.662	.813	.746
ROAR	0	.705	.823	.800	.712	.879	.167	.894	.508	.648†	.569†
	1	.419	.536	.514	.423	.597	.174	.684	.263	.313†	.133†
Wachter	0	.813	.931	.900	.816	.935	.120	.999	.730	1.000	.754†
	1	.573	.691	.668	.568	.818	.250	.916	.478	.555†	.536†

- fixed-τ Holm: 100 hypotheses (10 cells $\times$ 10 budgets);
- all-τ Holm on the seven-point grid: 700 hypotheses;
- all-τ Holm on the 0.01 grid: 7,000 hypotheses.

Every all-τ rejection is also a fixed-τ rejection. ROAR in g_0 shows the difference: $\tau = 0.99$ survives fixed-τ Holm (adjusted $p = 6.7 \times 10^{-3}$) but not the all-τ correction.

Results. Tables 15 and 16 collect all cells.

- *Sharp endpoint.* $J^{\uparrow}$ exceeds $J_{\otimes}$ by 0.103–0.250 in every cell. Recomputed on the calibration banks, pooled banks or either half of the evaluation bank, it moves by at most 0.028.
- *Product certificates.* They bound $J_{\otimes}$ correctly here, because the simulation draws the two channels independently. Their bounds lie below $J^{\uparrow}$ in all 10 cells for EB and in 7 of 10 for BD (for example, BD gives 0.818 against $J^{\uparrow} = 0.927$ for AR in g_1), so they claim levels that no coupling-robust procedure can certify.
- *Prescribed action.* Replacing the realized response by the prescribed action lowers $J_{\otimes}$ in every cell, by up to 0.204 (ROAR in g_0).
- *Model-only certificates.* Each is valid for its own estimand, the prescribed action at a fixed update. Combined with a fixed evaluation time, the optimism of the prescribed action makes them understate the realized risk in 3 (BD, next update) and 7 (CP, worst horizon) of 10 cells.
- *RAC.* The robust failure rate at $\gamma = 0.05$ is already 0.93–1.00 in the abstaining cells, so their empirical floor exceeds 0.97. In the two certified cells the contract decomposes as in Table 8. For ROAR in g_1, $J_{\mathrm{fam}} = 0.684$ exceeds $J_{\otimes}$ by 0.174 (identification) plus 0.088 (family); for Wachter in g_1, $J_{\mathrm{fam}} = 0.916$ exceeds $J_{\otimes}$ by 0.250 plus 0.098.

D.5 Positive target margins

The confirmatory study uses margin-zero recommendations, the only margin at which all five generators share a common support. The frozen recommendations at margins 0.10 and 0.25 exist for AR, Wachter, ROAR and PROBE. For these four generators we run the same pipeline at each margin. We rebuild the reference states and the prediction-space family (centres and scales from the same 200-path pilot) and draw fresh behavioral banks of $50{,}000$ and $100{,}000$ units from the frozen mechanism. The confirmatory path banks are reused, robust thresholds are recomputed exactly with the LP above, and contracts use all-τ Holm on the 0.01 grid $\{0.05, \ldots, 0.99\}$.

For an across-margin comparison on a fixed population, we use the intersection cohort of 46 applicants who receive a recommendation from all four generators at every margin. Analyses using the

Table 16: RAC on LendingClub. Robust fail: calibration robust-failure rate at $\gamma = 0.05$; floor: $\min_\gamma[\hat{R}(\hat{r}_\gamma) + \gamma]$; $\tau^\star$ under fixed-τ and all-τ Holm on the seven-point grid $\{.50, .60, .70, .80, .90, .95, .99\}$ and all-τ on the 0.01 grid. Every witness is $\gamma^\star = 0.05$; "–": abstains.

Generator	g	Robust fail	Floor	$\tau^\star$ fixed-τ	$\tau^\star$ all-τ	$\tau^\star$ all-τ, 0.01 grid
AR	0 / 1	1.000 / .956	1.050 / 1.006	– / –	– / –	– / –
PROBE	0 / 1	1.000 / .956	1.050 / 1.006	– / –	– / –	– / –
RBR	0 / 1	1.000 / .959	1.050 / 1.009	– / –	– / –	– / –
ROAR	0 / 1	.927 / .684	.977 / .734	.99 / .80	– / .80	– / .77
Wachter	0 / 1	1.000 / .910	1.050 / .960	– / .99	– / .99	– / .98

Table 17: Positive-margin recommendations in g_1 (fixed cohort of 40 applicants).

Generator	Margin	Mean cost	$J_\otimes$	$J^\uparrow$	Product BD	RAC $\tau^\star$	Identification	Family
AR	0.00	0.599	.708	.934	.821	–	.225	.034
	0.10	1.072	.667	.901	.785	–	.235	.062
	0.25	1.977	.569	.830	.680	.96	.261	.077
PROBE	0.00	0.680	.707	.923	.817	–	.215	.045
	0.10	1.072	.661	.902	.787	–	.241	.062
	0.25	1.977	.564	.827	.683	.96	.263	.078
ROAR	0.00	4.164	.432	.610	.541	.77	.177	.077
	0.10	4.164	.433	.600	.558	.78	.167	.088
	0.25	4.164	.431	.606	.549	.75	.174	.071
Wachter	0.00	2.342	.583	.816	.701	.98	.233	.091
	0.10	2.342	.581	.813	.689	.97	.232	.101
	0.25	2.517	.543	.798	.656	.95	.255	.095

larger per-margin common supports (71, 62, and 46 applicants) yield the same qualitative pattern. Because this cohort excludes RBR and requires a recommendation at every margin, it differs from the confirmatory one, and its margin-zero risks differ slightly from Table 15.

In context g_1 (Table 17, Figure 7), raising the margin to 0.25 lowers $J_\otimes$ by up to 0.14 (AR and PROBE) at a higher recommended cost. The identification gap stays at 0.17–0.26. RAC additionally certifies AR and PROBE at $\tau^\star = 0.96$ and tightens ROAR to 0.75 and Wachter to 0.95. The six g_0 applicants are too few for a separate analysis; there RAC abstains for every generator and margin.

D.6 How much model history is enough?

Subsampling the path bank. A single nested prefix of the confirmatory bank per n_d gives only one realization of the certification outcome. We therefore draw 200 random subsamples per $n_d \in \{10, 20, 30, 50, 75, 100, 150\}$, add the full bank ($n_d = 200$), and repeat the all-$\tau$ certification on the 0.01 grid (Table 18, Figure 8a,b). Feasibility follows $\gamma_{\min}(n_d) = 1 - 0.0025^{1/n_d}$ exactly.

- *ROAR in g_1.* Never certified at $n_d = 10$, certified in 65% of subsamples at $n_d = 20$ and always from $n_d = 30$. Its median contract falls from 0.98 to 0.77.
- *Wachter in g_1.* Certified in 32% of subsamples at $n_d = 150$ and always at $n_d = 200$ (0.98).
- *Other cells.* Apart from ROAR in g_0 (0.5% of subsamples at $n_d = 150$), no other cell is ever certified.
- *Held-out check.* The fraction of subsamples in which the held-out miss exceeds some finite-radius γ is at most 0.02, within $\delta_d = 0.025$.

A temporal backtest with genuine history. The confirmatory paths bootstrap the realized 2014 and 2015 pools, so they measure estimation noise around the future models that actually occurred. They do not measure update-to-update drift known at issuance. The backtest uses only data available by the end of 2012.

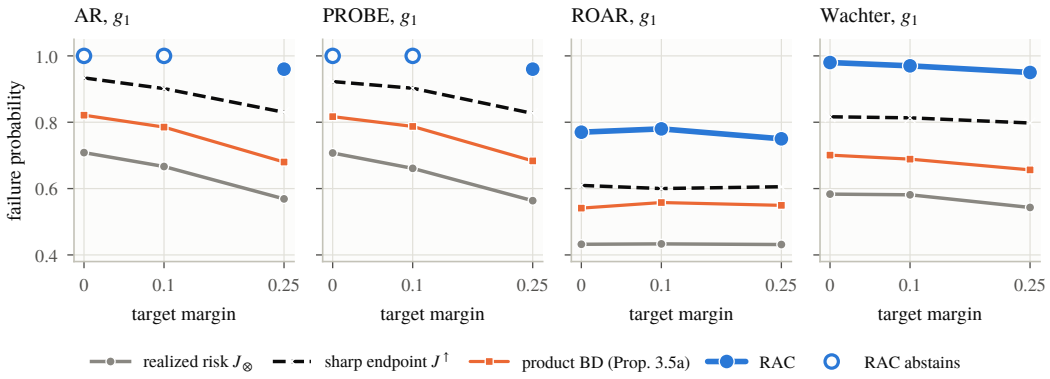

Figure 7: Positive-margin recommendations in g_1 (fixed cohort): realized risk, sharp endpoint, product BD bound and RAC contract (hollow: abstains).

Table 18: Model history in LendingClub. $\gamma_{\min}(n_d) = 1 - 0.0025^{1/n_d}$ and the number of feasible budgets on Γ. Nested prefix: the first n_d paths of the bank (all-τ Holm, seven-point grid; $\tau^\star$ with witness $\gamma^\star$). Random subsamples: 200 subsamples per $n_d < 200$ and the full bank at $n_d = 200$ (all-τ Holm, 0.01 grid; certification probability and median $\tau^\star$). "·": not run; "–": abstains.

			Nested prefix		Random subsamples			
			ROAR g_1	Wachter g_1	ROAR g_1		Wachter g_1	
n_d	$\gamma_{\min}$	Feasible γ	$\tau^\star$ ($\gamma^\star$)	$\tau^\star$ ($\gamma^\star$)	P(cert.)	$\tau^\star$	P(cert.)	$\tau^\star$
5	.698	0/10	–	–	·	·	·	·
10	.451	1/10	–	–	0.00	–	0.00	–
20	.259	5/10	.99 (.30)	–	0.65	.98	0.00	–
30	.181	7/10	.90 (.20)	–	1.00	.90	0.00	–
50	.113	8/10	.90 (.15)	–	1.00	.87	0.00	–
75	.077	9/10	·	·	1.00	.82	0.00	–
100	.058	9/10	.80 (.10)	–	1.00	.79	0.00	–
150	.039	10/10	·	·	1.00	.78	0.32	.98
200	.030	10/10	.80 (.05)	.99 (.05)	1.00	.77	1.00	.98

- *Origins.* At each of 19 monthly origins o from 2008-06 to 2009-12, the current model is refit on all loans issued up to o (1,124–6,529 loans).
- *Updates.* Its two updates are refit on the windows $(o + 12, o + 24]$ and $(o + 24, o + 36]$ months (7,113–14,101 and 10,109–43,470 loans). This is the timing of current $\to$ 2014 $\to$ 2015.
- *Increment paths.* The increments $\theta_t(o) - \theta_{\text{cur}}(o)$ are applied to the frozen current model (Assumption 4.3). The median first-increment norm is 2.115, against 1.222 for the realized 2014 update.

Three tolerance families are calibrated on these paths:

- *P:* the confirmatory prediction-space shape, with its radius from history ($n_d = 19$);
- *H1:* an a priori family, fixed before any data, with centre at the current-model score $\theta_{\text{cur}}^\top a_j$ and scale $\sqrt{t}\,\|a_j\|$ ($n_d = 19$);
- *H2:* centre and scale from alternate origins (the 1st, 3rd, …, 19th), with the Wilks radius from the remaining 9.

Table 19 and Figure 8(c) show the results.

- *Scale of genuine drift.* In family P, genuine historical path scores (median 33.9) are about 14 times the bootstrap scores (median 2.48).

Table 19: Temporal backtest. Path scores: medians over the 200 bootstrap paths and over the historical paths, and the scores of the realized 2014/15 and 2015/16 paths (*: outside the history-calibrated set). γ: smallest feasible budget, with its radius. Min. fail: smallest robust-failure rate over cells and budgets. RAC abstains in every cell under every family.

		Path score						
Family	n_d	Bootstrap	History	2014/15	2015/16	γ	Radius	Min. fail
P: confirmatory shape	19	2.481	33.895	0.469	13.966	0.30	43.465	1.000
H1: a priori	19	0.162	0.400	0.155	0.204	0.30	0.618	1.000
H2: history split	9	4.150	1.399	4.189*	4.498*	0.50	2.945	0.992

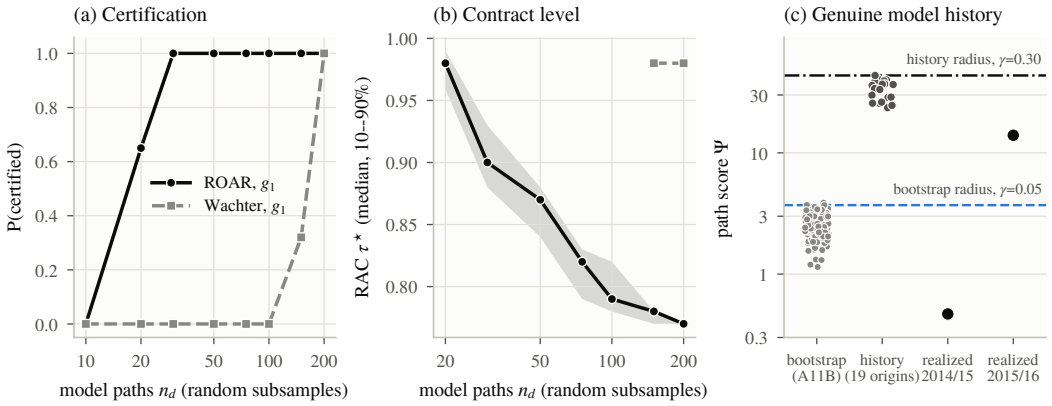

Figure 8: Model history in LendingClub. **(a, b)** Certification probability and median contract (10–90% band) over 200 random subsamples of the confirmatory path bank per n_d (the full bank at $n_d = 200$). **(c)** Path scores in the confirmatory family: bootstrap paths, 19 genuine historical paths, and the realized 2014/15 and 2015/16 paths, with the bootstrap radius at $\gamma = 0.05$ and the history radius at the smallest feasible γ.

- *Feasibility.* With 19 paths only $\gamma \geq 0.30$ is feasible (radius 43.5 in family P); with the 9 calibration paths of H2 only $\gamma = 0.50$ is.
- *Contracts.* Under every family RAC abstains in all 10 cells: even the smallest robust-failure rate over cells and feasible budgets is 0.99.
- *Realized paths.* The realized path $(\theta_{2014}, \theta_{2015})$ lies inside every set except H2. The path one year later, $(\theta_{2015}, \theta_{2016})$, scores 13.97 in family P, far outside the bootstrap radius (3.70 at $\gamma = 0.05$) but inside the historical one.
- *Realized cohort failure.* Under the actual 2014/2015 models it is 0.43–0.83, close to the bootstrap-law product risks (0.42–0.84).

Monthly origins overlap in time, so the historical paths are not i.i.d. as Assumption 4.3 requires, and the radii are, if anything, optimistic. The bootstrap-calibrated contracts of Section 7 should therefore be read as certifying against estimation uncertainty around known updates. Certifying against genuine drift requires a longer or pooled history, or a committed model channel.

D.7 External 2016 mismatch diagnostic

The 2016 classifier is excluded from all calibration and evaluation. Relative to the 2015 centre, its prediction-space RMS displacement is 0.3785. That is about 5.4 times the median displacement under the 2015 same-law bootstrap and above the entire bootstrap distribution; its parameter displacement is $\|\theta_{2016} - \theta_{2015}\|_2 = 0.3221$. Together with the backtest, this shows that a later update can fall far outside the variation represented by the bootstrap path law. The guarantees of Section 7 apply to the pre-specified path law, not to arbitrary future shifts.

D.8 Developmental tolerance families

During method development, we also evaluated two parameter-space tolerance families. We then fixed the prediction-space family before drawing the independent confirmatory calibration and evaluation samples; the developmental families are reported here for completeness and were not used for confirmatory inference.

- *RMS parameter-space score.* It calibrates well but is operationally vacuous. No hypothesis is certified, the minimum robust-failure rate over Γ is 0.884 (ROAR in g_1), and the minimum empirical budget is 1.05 in every cell.
- *Covariance-aware parameter family.* Its calibration and evaluation score distributions agree (mean 6.568 vs. 6.582, 95% quantile 8.150 vs. 8.298), yet it is equally vacuous.

E Notation

Symbol	Meaning
$X,\ G = G(X)$	features of a rejected applicant; context (cell index)
$\mathcal{S} = \{\sigma_1, \dots, \sigma_m\}$	recommendation menu
$K_\sigma(\cdot \mid x)$	response kernel: law of the realized action $A(\sigma)$
$Z = (X, A, T)$, $v_x(a)$	response record (action and return time $T \in \{1, \dots, \bar{t}\} \cup \{\infty\}$); post-action state
$Q_{g,\sigma}$	law of Z in cell (g, σ)
$\vartheta = (\vartheta_t)_{t \le \bar{t}}$, ν	future model path and its law; θ_0 is the issuing model
$\ell^\star(Z, \vartheta)$	terminal failure: no return, or rejection by ϑ_T
$\varpi \in \Pi(Q, \nu)$	coupling of responses and model paths
$J(\varpi)$, $J_\otimes$	terminal risk under ϖ; under the product coupling
$J^\uparrow,\ J^\downarrow$	sharp upper and lower endpoints of the identified set
B_C, A_C	responses failed by some, respectively every, path in C
$\mathcal{C}(r) = (C_t(r))_t$, Ψ	tolerance family and path score (Definition 4.1)
$\ell_\mathcal{C}$, $R_{g,\sigma}(\mathcal{C})$	robust loss and robust behavioral risk
$\tau = \eta + \varepsilon + \gamma$	contract level = behavioral threshold + shift budget + model budget
$\gamma_{\min}$, ω	smallest feasible model budget; certification probability (Proposition 4.7)
$\delta_b,\ \delta_d$	error levels of the behavioral and model calibrations
$n,\ n_d$	logged units; model paths (drift sample)
ι, E, s	dependence budget (mutual information); common environment; shock intensity (E3)
J_{shock}	terminal risk under the common-shock deployment coupling
coverage, $\tau_{0.4}$	share of contexts with a certified recommendation; smallest τ with coverage ≥ 0.4
u_x, $\mathrm{Acc}_g(\sigma)$	valuation of realized actions; expected valuation of a recommendation
TV_x, $P_x^\pm$	total variation between two response kernels; normalized Jordan parts
Λ_u	leverage $\lvert \mathrm{Acc}_g(\sigma) - \mathrm{Acc}_g(\sigma') \rvert$ (Corollary 5.3)
π_0	randomized recommendation logging policy (Assumption 4.2)
kl, kl_+^{-1}	Bernoulli relative entropy; its upper inverse
ζ	type-II error in Theorem A.6
β, L, h	Hölder exponent and constant; bandwidth (Theorems A.7 and A.8)