SLEWMATCH: JOINT-ENVELOPE DIAGNOSTICS FOR DECISION-RELEVANT PREDICTION ERROR

Anonymous authors
Paper under double-blind review

ABSTRACT

A response model can choose correctly at a few preference weights while misrepresenting the decision family between them. We study finite affine costs with a known penalty slope and frozen predicted responses. Classical lower envelopes provide the computational skeleton; their true–predicted joint cells expose which relative errors can affect a choice. The resulting offline diagnostic returns a worst-loss witness and separates common offsets, ineligible comparisons, margin slack and retained regret. Same-window experiments compare sparse, dense, random and structure-guided audits. In a release-response case, three probes miss 423 of 491 intrinsic-shift bound violations, whereas 1,025 uniform probes miss four; conventional hull vertices recover every exact maximum. A separately fitted synthetic capacity task has over thirteen active actions on average and confirms that finding all threshold violations is weaker than recovering worst loss. Standard marginal calibration remains a report, not a policy change. The contribution is a reusable, interpretable error audit with explicit classical antecedents, paired computational comparisons and two synthetic structural tests, rather than a new envelope algorithm or learned physical law.

1 INTRODUCTION

A learned response vector represents consequences for a downstream choice. Reconstructing every response value and preserving the choices are different requirements: adding one constant to all candidate values changes prediction error without changing a minimizer. Relative errors can instead move a switching boundary or introduce an action that should never be optimal. Evaluation at a few preference weights can then endorse a predictor whose worst decision is wrong between those weights.

We study this distinction for finite candidates with cost $P_j + \lambda v_j$: a predicted response P_j is traded against a known penalty slope v_j, using one nonnegative weight λ. This is a parametric optimization problem with established lower-envelope geometry. Minimization diagrams already encode active intervals, intersections and degenerate ties (**?**); predict-then-optimize theory already studies discontinuous decision loss and margins (**?**Elmachtoub & Grigas, 2022). We do not propose a new envelope construction or claim that decision loss was previously interchangeable with prediction error.

Our contribution is an offline diagnostic interface between *two* such diagrams: the true and predicted consequences. Their jointly active cells select the comparison directions that can affect the declared decision family. The diagnostic returns the worst loss with its weight and action-pair witness, and accounts for the absolute-error bound through common offsets, excluded comparisons, margin slack and retained regret. It distinguishes an accurate response model, a correct active-action set, and correctly located action switches. These are observable diagnostic outputs given reference responses, rather than new layers inside a learned representation.

We test what this information adds beyond reasonable probing and classical structure. On a release-response model, a 1,025-point uniform scan misses four of 491 intrinsic-shift bound violations, rather than the 423 missed by three probes. A conventional hull-vertex check matches the exact result: increasing density is not the only alternative to sparse evaluation, and exactness is not unique to our implementation. A separate synthetic capacity-shortage prediction task has over thirteen true active actions on average, uneven costs, and different learned active sets. It tests the same diagnostic with

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.

neural and polynomial predictors, not another variant of the release ODE. Figure 1 connects the physical case to the underlying information question.

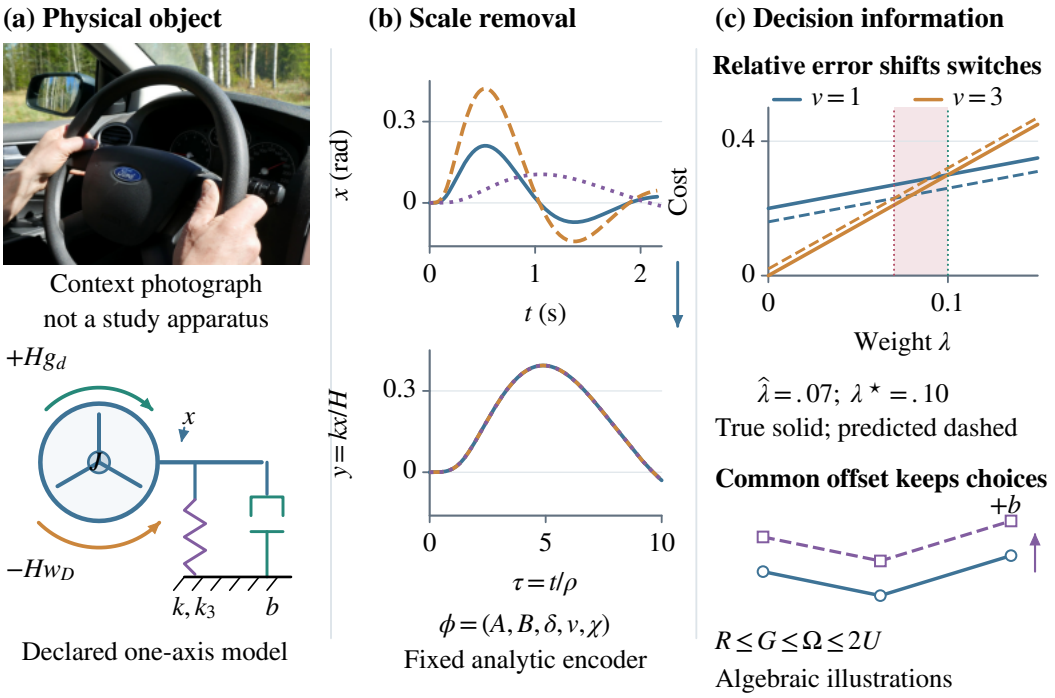


Figure 1: Physical context, scale removal, and decision information. The licensed photograph is background, not study apparatus or model input; the rotational system is schematic. Middle: computed DOP853 traces under amplitude/time factors $(1, 1), (2, 1), (.5, 2)$ (blue solid, orange dashed, purple dotted) coincide after analytic normalization. Right: the labeled two-action algebraic example has true/predicted switches $.10/.07$; solid/dashed costs identify truth/prediction and color identifies slope. A common offset preserves choices. Neither illustration is an independent test case. Photo credit: Santeri Viinamäki, CC BY-SA 4.0; crop and source in Appendix I.3.

2 Joint-envelope diagnostics for a finite decision family

2.1 What must be represented for the decision?

For context c, let $P \in \mathbb{R}^K$ be true responses, $v \in \mathbb{R}^K$ known slopes, and $p_m \in \mathbb{R}^K$ the emitted predictions of a fixed family of M models. The candidate arrays may depend on c, but true and predicted costs use the *same* slopes. Define

$$C_j(\lambda) = P_j + \lambda v_j, \quad \widehat{C}_{m,j}(\lambda) = p_{m,j} + \lambda v_j, \quad \hat{\jmath}_m \in \arg\min_j \widehat{C}_{m,j}, \quad R_{\lambda,m} = C_{\hat{\jmath}_m} - \min_i C_i. \tag{1}$$

Deployment breaks ties by smallest index; the audit includes *all* predicted minimizers. For an interval $\Lambda \subseteq [0, \infty)$, write

$$G_m^\Lambda = \sup_{\lambda \in \Lambda} \max_{j \in \widehat{\mathcal{S}}_m(\lambda)} [C_j(\lambda) - \min_i C_i(\lambda)], \quad G^\Lambda = \max_m G_m^\Lambda, \quad \widehat{\mathcal{S}}_m(\lambda) = \arg\min_j \widehat{C}_{m,j}(\lambda). \tag{2}$$

Unadorned G_m, G use the full nonnegative axis. For $e_m = p_m - P$, let $U_m = \|e_m\|_\infty$ and $\Omega_m = \max e_m - \min e_m$, with family maxima U, Ω. Half the span equals $\min_b \|e_m - b\mathbf{1}\|_\infty$, so comparison forgets common offsets. Predicted optimality gives

$$0 \le R_{\lambda,m} \le e_{m,j^\star} - e_{m,\hat{\jmath}_m} \le \Omega_m \le 2U_m, \qquad R_{\lambda,m} \le G \le \Omega \le 2U. \tag{3}$$

G is directed and objective-dependent: it vanishes exactly when every predicted minimizer is truly optimal at every weight. It is not a symmetric response distance or a learned feature.

2.2 CLASSICAL CELLS, JOINT COMPARISONS, AND AN INSPECTABLE WITNESS

Each predicted candidate has a closed minimizing interval

$$I_{m,j} = \{\lambda \geq 0 : (v_j - v_i)\lambda \leq p_{m,i} - p_{m,j} \text{ for all } i\}. \tag{4}$$

It may be empty, a singleton, or unbounded. On it, candidate-j regret is convex piecewise affine: $\max_i[(P_j - P_i) + \lambda(v_j - v_i)]$. Its maximum on a clipped cell occurs at an endpoint; the final unbounded cell uses the minimum-slope limit. Algorithm 1 gives a transparent $O(MK^2)$ reference. Slope-sorted hull construction gives the same cells; a hull-vertex audit is therefore an equally valid exact route, not a weaker approximation.

Algorithm 1 Reference endpoint audit; returns loss and an error witness

Require: True P, fixed predictions $p_1, \ldots, p_M$, slopes v, $\Lambda = [L, H]$
1: $g \leftarrow -\infty$; initialize witness w
2: **for** each fit m and candidate j **do**
3: $[l, u] \leftarrow [L, H]$; mark cell nonempty
4: **for** each i **do**
5: $d \leftarrow v_j - v_i$; $r \leftarrow p_{m,i} - p_{m,j}$
6: If $d > 0$: $u \leftarrow \min(u, r/d)$; if $d < 0$: $l \leftarrow \max(l, r/d)$
7: If $d = 0$ and $r < 0$: mark cell empty
8: **end for**
9: **if** nonempty and $l \leq u$ **then**
10: For each finite $\lambda \in \{l, u\}$, maximize $(P_j - P_i) + \lambda(v_j - v_i)$ over i
11: Update g, w with the loss and (m, j, i, λ); retain singleton cells
12: If $u = \infty$, also test $P_j - \min_{i:v_i=v_j} P_i$ and its limiting witness
13: **end if**
14: **end for**
15: **return** g, w

The diagnostic interpretation requires both diagrams. Let J_i be true-optimal intervals and $\mathcal{Q}_m = \{(i, j) : J_i \cap I_{m,j} \neq \varnothing\}$. Only a pair in this intersection relation can be true-optimal and predicted-optimal at the same weight. At such a witness,

$$R_j(\lambda) = (e_{m,i} - e_{m,j}) - \underbrace{(\widehat{C}_{m,i} - \widehat{C}_{m,j})}_{\text{nonnegative predicted margin}}. \tag{5}$$

The audit-only $B_m = \max_{(i,j)\in\mathcal{Q}_m}(e_{m,i} - e_{m,j})$ satisfies $0 \leq G_m \leq B_m \leq \Omega_m$. With the min–max midpoint $b_m = (\max e_m + \min e_m)/2$, an exact accounting is

$$2U_m = \underbrace{2|b_m|}_{\text{common offset}} + \underbrace{(\Omega_m - B_m)}_{\text{ineligible comparisons}} + \underbrace{(B_m - G_m)}_{\text{margin slack}} + G_m. \tag{6}$$

All terms are nonnegative. $B_m - G_m$ is aggregate slack, not necessarily the margin of one fixed maximizing pair. This identity does not decompose a ratio of separately calibrated quantiles. The extra output beyond a minimization diagram is the joint comparison relation, its error accounting, and the witnessed equation that explains a failure; these can be attached to a classical envelope implementation.

2.3 TWO WAYS A FINITE PROBE CAN MISS AN ERROR

For the same adjacent pair in true and predicted positive-length cells,

$$\lambda^\star_{ij} = \frac{P_i - P_j}{v_j - v_i}, \qquad \widehat{\lambda}_{ij} - \lambda^\star_{ij} = \frac{e_{m,i} - e_{m,j}}{v_j - v_i}. \tag{7}$$

Relative errors move this boundary; a common offset does not. A different failure inserts an action that should never be active. Take $v = (0, 1, 2)$, $p = (a + b, a, 0)$ and $P = (a + b, a + \eta, 0)$, with $0 < a < b$ and $\eta > (b - a)/2$. The predicted middle action occupies (a, b), but the true envelope uses only endpoints. Its worst regret is η at the cell boundaries under the all-minimizer convention

(or the corresponding one-sided supremum). Narrowing $b - a$ need not reduce this loss, even with slopes fixed in $[0, 2]$. This elementary counterexample is not a new statistical impossibility theorem. It identifies a discontinuity that invalidates a global Lipschitz interpolation argument across predicted switches; Figure 3 exposes the relevant intersections.

2.4 A MARGINAL REPORT IS SEPARATE FROM THE SELECTOR

For a score $Z \in \{U, \Omega, G\}$ fixed before exchangeable calibration contexts, standard split-conformal ranking gives (Lei et al., 2018)

$$q_Z = Z_{(\lceil (n+1)(1-\alpha) \rceil)}, \qquad q_Z = +\infty \text{ if the rank exceeds } n. \tag{8}$$

Conditional on fixed model-selection choices, with marginal probability at least $1 - \alpha$, every included fit satisfies $R_{\lambda,m} \le q_G$ for all $\lambda \ge 0$. Separately, q_Ω and $2q_U$ give analogous bounds; using the same rows and rank implies $q_G \le q_\Omega \le 2q_U$. This is a standard scalar calibration step. It does not enter the argmin, observe online truth, or give conditional validity for each context. Separate 95% events are not joint 95% validity, and changed models, candidates or slopes require their own score and calibration.

3 TWO RESPONSE-LEARNING INSTANCES

Physical release and a fixed analytic encoder. From zero initial state, the one-axis model is

$$\begin{aligned} J\ddot{x} + b\dot{x} + kx + k_3x^3 = H(g_d - w_D), \qquad & g_d(t) = e^{-\max(t-d,0)/\rho}, \\ w_D = 1 - 3s^2 + 2s^3, \qquad & s = \mathrm{clip}(t/D, 0, 1). \end{aligned} \tag{9}$$

Here x is wheel perturbation; residual effort Hg_d is prescribed. Declared metadata determine

$$\tau = t/\rho, \quad y = kx/H, \qquad \phi(c, D) = \left(\frac{J}{k\rho^2}, \frac{b}{k\rho}, \frac{d}{\rho}, \frac{D}{\rho}, \frac{k_3H^2}{k^3} \right) = (A, B, \delta, v, \chi). \tag{10}$$

Two 64-unit tanh layers learn the sampled normalized peak $P = \max_{0 \le \tau \le 20} |y(\tau)|$ from $(\log A, \log B, \delta, \log v, \chi)$, with training-only scaling and nonnegative clipping. The encoder is analytic; the regressor learns peaks, not groups, trajectories or intentions. Limits S, L yield 21 durations in $[1.5H/S, L]$ and slopes $v_j = D_j/\rho$; an empty set requests external arbitration. For $a, s > 0$,

$$(H, J, b, k, d, \rho, D, k_3, S, L) \mapsto (aH, s^2J, sb, k, sd, s\rho, sD, k_3/a^2, aS/s, sL) \tag{11}$$

preserves groups, normalized candidates, predictions and reference responses; $x_{a,s}(t) = ax(t/s)$. It transports each report event, including failures, by pointwise equality rather than an invariant sampling distribution. Hidden lag or changed normalized dynamics need not preserve it.

Capacity choice with a different prediction target. Independently, a synthetic demand context $z \in [-1, 1]^4$ defines a two-component lognormal mixture. Seventeen capacities $a_j = 2.8(j/16)^{1.3}$ have known, uneven increasing costs $v_j(z)$. The target is the *analytic conditional expectation* $P_j(z) = \mathbb{E}[(D - a_j)_+ \mid z]$, not one realized demand or a Monte Carlo estimate. A three-fit MLP and a validation-selected cubic polynomial ridge predict this target from (z, a_j); both use the same selector $p_j + \lambda v_j$. The task has no release ODE or physical-scale transport claim. Appendix D specifies every mixture and cost parameter.

4 EVIDENCE ROLES AND DIAGNOSTIC COMPARISONS

Data roles. Release predictors and their 256 calibration, 512 ID and 512 intrinsic-shift contexts are fixed. Earlier model selection preceded those outcomes; the current diagnostic comparisons reuse already-inspected records and are retrospective. Existing fair augmentation, retrained omissions, nested label budgets, hidden lag, and finer-grid audits remain available. Capacity uses newly generated independent train/validation/calibration/test contexts of sizes 1,024/256/256/512, with 17 analytic labels each. Its generator, scores and budgets were recorded before fitting; selected model hashes were frozen before generating calibration and test outcomes. This is a local protocol freeze, not external preregistration. Each task calibrates its own frozen families. Appendix E.1 records the chronology in one place.

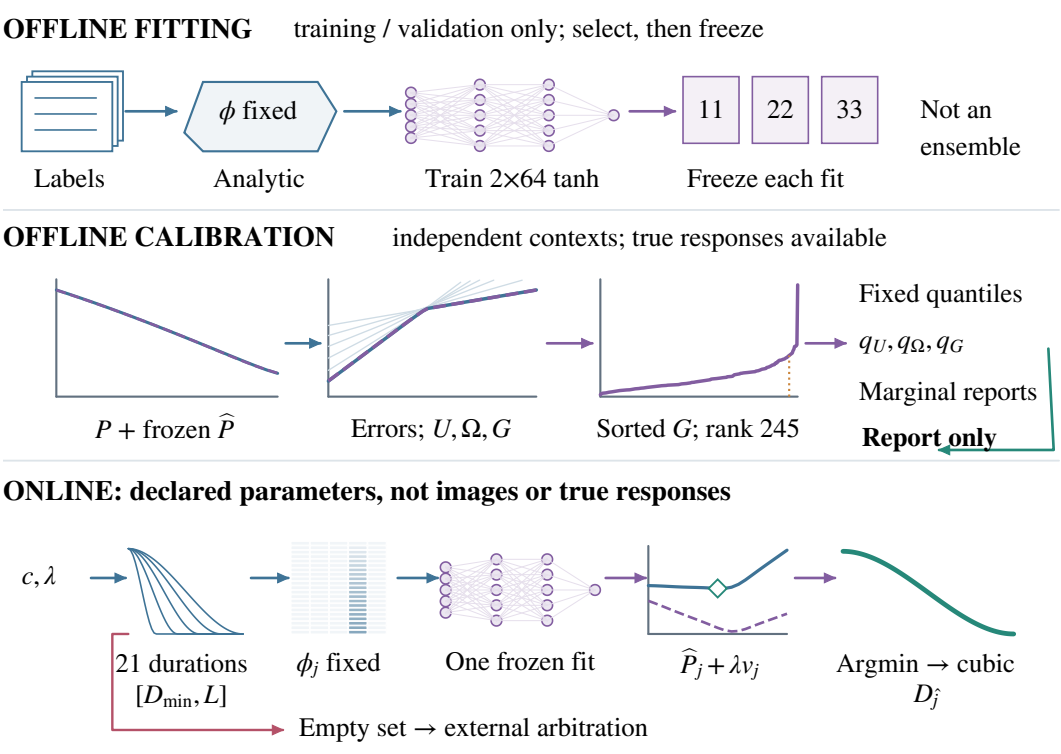


Figure 2: Learned and observed objects in the release instance. Analytic encoding is fixed; the depicted MLP has two 64-unit hidden layers. Three fits are frozen separately, not deployed as an ensemble. Offline panels use true/predicted candidates and sorted G scores. Online panels show only candidate durations, fixed features, predicted peaks (dashed) and known-cost calculations (solid) versus candidate v_j, followed by the selected cubic command. Quantiles end at the report, not the argmin. The image is not an input; there is no online true-error test or safety gate.

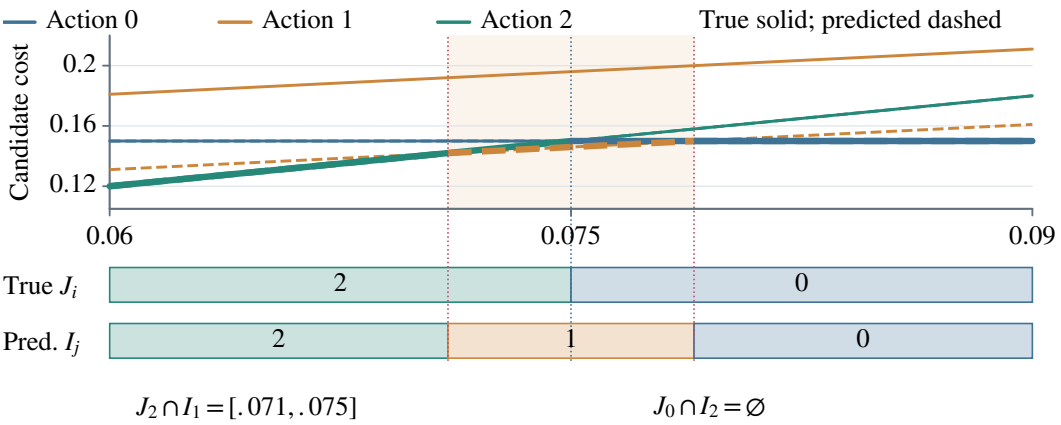
Same window, same minimizer convention. Ten audits share identical true/predicted arrays and all-minimizer semantics. Release uses $\Lambda = [0, .5]$; capacity uses $[0, 1.5]$. The full nonnegative-domain audit is reported separately. Fixed designs are the original three weights, the original 84-point mixed-log scan, and uniform grids of 257/1,025 points; random designs use one fixed nested realization at these larger budgets. Capacity scales the original probes by three. A prediction-only hull supplies either cell midpoints or all cell vertices. The two exact implementations use pairwise feasibility (Algorithm 1) or independent slope-sorted hulls followed by closed true/predicted cell intersections. The latter follows classical envelope geometry, not a new construction or a call to CGAL.

Metrics and cost. An exact violation is $G^\Lambda > q_G$; a miss is a probe value at most that *same* width. We also report absolute loss underestimation, witness value/location, and work. Relative gaps are missing when exact loss is zero. Time includes interval construction, cost evaluations and witness output, but excludes labels, loading and prediction. Two warm-ups and five repetitions use the first fixed 32 contexts per cohort/family; medians summarize contexts after their repetition medians. This is local offline Python/NumPy timing, not control latency. No scan budget, seed or threshold was adjusted after observing test performance.

5 WHAT THE JOINT GEOMETRY CHANGES IN AN EVALUATION

5.1 ERROR MAGNITUDE, ACTIVE IDENTITY AND SWITCH LOCATION ANSWER DIFFERENT QUESTIONS

The decomposition in Equation 6 separates kinds of information rather than labeling all discarded error benign. On release ID data, offsets account for 27.5% of mean per-fit $2U_m$, ineligible comparisons 26.3%, margin slack 5.4%, and retained loss 40.8%. Under intrinsic shift, the offset share grows to 59.5%, yet mean G_m rises from .002274 to .017532. A large removed offset is not evidence of a reliable selector. These component means are not an additive explanation of the 58.7% reduction between separately ranked calibration quantiles.



At $\lambda = .071$: selected 1, reference 2; relative error minus margin:

$$R = (e_2 - e_1) - (\widehat{C}_2 - \widehat{C}_1) = .050 - .000 = .050$$

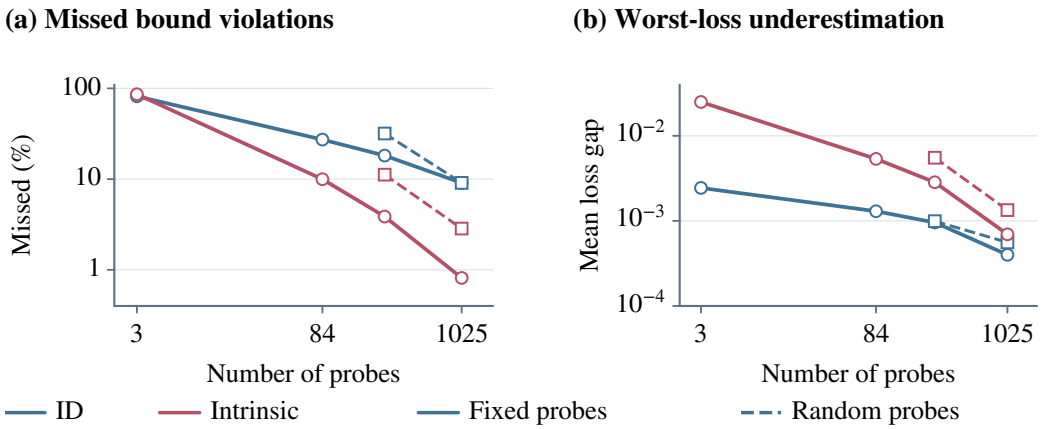
Figure 3: Joint intervals and a false active action. This *algebraic illustration* uses $a = .071$, $b = .079$, $\eta = .05$, not test records. Thick lines show lower envelopes. True action 1 is never optimal; prediction inserts it on $[a, b]$. The two interval rows distinguish an eligible comparison $J_2 \cap I_1$ from an empty one. At the left boundary, the all-minimizer witness has relative error .05, zero predicted margin and regret .05. Dashed/solid endpoint-action costs coincide. The same bounded slopes permit an arbitrarily narrow high-loss cell.

Shared active sets also do not fix the switching weights. The release ID and intrinsic cohorts have identical true/predicted active sets in 60.0% and 88.3% of fit–context pairs, but matched switches move by median .00281 and .00485. Under intrinsic shift, 444/512 contexts pass three *bound checks*, while only 21/512 pass the full-family bound. The separate operating point $\lambda = .1$ has zero regret throughout. Confusing these two statements would hide the error mechanism.

5.2 HOW MUCH PROBING IS ENOUGH, AND WHAT DOES STRUCTURE ADD?

For the full-group release family, bounded and full-domain maxima coincide on every tested context. Within the common window, the three probes miss 423/491 intrinsic violations; the 84-point scan misses 49, and uniform 257/1,025-point scans miss 19/4 (Figure 4, Table 1). Thus reasonable scanning removes most of the dramatic three-probe failure; the remaining issue is a quantified approximation gap, not a claim that dense probes are useless. On ID, 1,025 uniform probes miss 2/22 violations and underestimate G by mean .000400.

Structural vertices recover the exact maximum in every comparison, while hull-cell midpoints still miss 375/491 intrinsic violations. Merely knowing which actions become active is insufficient; boundary behavior matters. The independent hull overlay and pairwise reference agree to 1.12×10^{-16} over 8,192 context–family evaluations across both tasks. On intrinsic release, median times are .313



Window $[0, .5]$; structural vertices and both exact routes have zero misses.

Figure 4: Same-window probing on 512 release contexts per regime. The solid sequence uses the fixed 3/84 designs then uniform 257/1,025 probes; dashed curves use the single prespecified nested random design. Miss fractions divide by 22 ID or 491 intrinsic exact violations, not 512. Underestimation averages all contexts. Logarithmic axes do not show the exact methods' zero miss/gap; these are explicit in Table 1. Connecting finite budgets is descriptive, not a convergence bound.

Table 1: Same-window release audit, $\lambda \in [0, .5]$, three fixed fits. Missed counts divide by exact violations of the unchanged q_G, not by all 512 contexts. ΔG is mean underestimation ($\times 10^3$) on intrinsic shift. Time is median offline audit milliseconds, including construction and witness output, excluding prediction and labels.

Audit	ID missed	Shift missed	ΔG	Shift ms
3 probes	18/22	423/491	25.012	0.222
84 probes	6/22	49/491	5.359	0.419
Uniform 257	4/22	19/491	2.846	0.822
Uniform 1,025	2/22	4/491	0.696	2.622
Random 257	7/22	55/491	5.532	0.817
Random 1,025	2/22	14/491	1.340	2.606
Cell midpoints	20/22	375/491	24.630	0.389
Hull vertices	0/22	0/491	0.000	0.313
Pairwise exact	0/22	0/491	0.000	1.241
Hull overlay	0/22	0/491	0.000	0.537

ms for hull vertices, .537 ms for the joint overlay, 1.241 ms for pairwise exactness and 2.622 ms for 1,025 uniform probes. Classical structure is a strong exact comparator and is faster here. The joint overlay additionally supplies eligible comparisons and the error-minus-margin explanation; we make no algorithmic speed novelty claim. Full method, witness and domain results are in Appendix C.

5.3 A SECOND PREDICTIVE TASK CHANGES THE ACTIVE STRUCTURE, NOT THE DIAGNOSTIC CONTRACT

The capacity reference averages 13.22 active actions instead of the two-endpoint intrinsic release envelope. On independent test contexts, the MLP inserts at least one false active action in 11.3% of fit–context pairs and omits a true action in 36.5%; cubic ridge has corresponding rates 67.8% and 79.7%. Full active sets agree in 54.9% and 5.1%. This is genuine response fitting followed by diagnostics, not hand-set prediction errors. Figure 5 shows that relative-error screening remains relevant when many internal actions and uneven slopes enter the comparison.

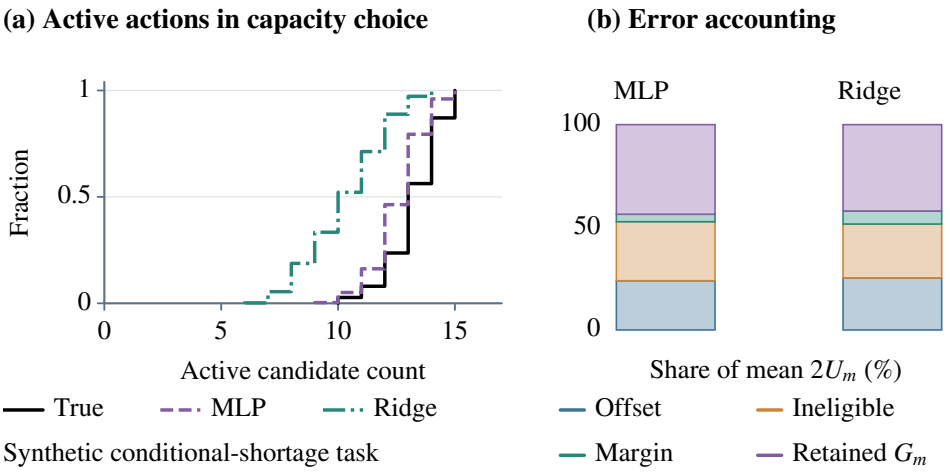


Figure 5: Independent synthetic capacity prediction. Left: distributions of positive-length active-candidate counts; the MLP curve averages three fixed fits, not extra independent contexts. Right: per-fit decomposition means divided by mean $2U_m$, separately for each predictor. Ineligible comparisons account for 28.7% for the MLP and 26.2% for ridge. The graph describes learned errors against the same analytic expectation, not an intervention that improves either model.

Table 2: Independent synthetic capacity prediction. Error and width are multiplied by 10^3. Each model family has its own 256-context calibration. Coverage is full-domain $G \leq q_G$ on 512 tests. Pred. counts all active candidates; Extra is the percentage of context–fit pairs with an extra predicted active action; Same compares full active sets. True sets average 13.22 candidates.

Model	RMSE	q_G	Covered	Pred.	Extra (%)	Same (%)
MLP (3 fits)	2.11	11.05	503/512	12.77	11.3	54.9
Cubic ridge	26.51	92.23	489/512	12.36	67.8	5.1

At their separately calibrated widths, nine MLP and 23 ridge test contexts violate the full-family report. In the common $[0, 1.5]$ window, the rescaled 84-point scan misses 9/9 and 18/23 violations; uniform 1,025 misses one for each. Random 1,025 finds all ridge violations but still underestimates its worst loss by mean .000590. Finding every threshold exceedance on a finite test set is therefore weaker than recovering the loss. Vertex and overlay audits recover both models' bounded maxima. Seven ridge contexts have larger loss outside this window; they are explicitly separate domain effects, not counted as bounded-scan errors.

5.4 Calibration, training budgets and physical scope remain distinct

Release calibration still gives $2q_U = .017137$, $q_\Omega = .012303$, $q_G = .007075$, with ID all-weight regret coverage 511/512, 510/512 and 490/512. The corresponding full-family loss has P95 .006484 and P99 .009242. Empirical-pool repartitioning gives a median envelope-width reduction of 58.1%, with dependent 5–95% range 54.7–62.3%; it is not independent replication. At the primary weight, the observable span-margin report certifies regret at most .001 for only 1.4% of fit–context pairs. Narrower global reports are not uniformly informative locally.

Nor does the geometry rely on maintaining a neural performance advantage. Doubling the bounded optimization budget gives release ID regret .000117 for analytic inputs and .000130 for normalized-target augmentation; the paired interval for their difference, $[-.000052, .000032]$, includes zero. The wider RBF search remains at its boundary. With hidden actuator lag .2, separate recalibration widens q_G to .060228 and covers 499/512 full-family losses, but leaves mean primary regret .003937 and every action unchanged. The test-only 201-candidate reference raises all-weight P95 from .006484 to .008027 without creating a new fine-grid certificate. These limitations and the original unfavorable intrinsic outcomes remain in Appendices F–H.

6 POSITIONING, LIMITS AND CONCLUSION

Classical envelopes supply active cells, vertex ties and construction/merge routines (**?**); SPO generalization studies discontinuous optimization loss and margins (**?**). Ranking-based decision-focused learning changes the fitting criterion (Mandi et al., 2022). Our joint-cell diagnostic instead consumes frozen predictions and reference responses to return eligible error directions, a worst-loss witness and its exact accounting. The experiments test which probe budgets miss these witnesses and why active-set agreement alone is insufficient. The contribution is this interpretable audit and its measured consequences, not a new minimization diagram or the nonnegative telescoping identity alone.

Known unit-equivariant prediction motivates the physical encoder (Villar et al., 2023); group discovery is a different problem (Bakarji et al., 2022). Yeh et al. (**?**) learn decision-relevant uncertainty sets that change robust decisions, unlike our post-fit report. SymmPI uses distributional group symmetry (Dobriban & Yu, 2025); the physical transport here is deterministic equality conditional on exchangeable base contexts. Appendix B separates these objects and assumptions.

The scope is finite candidates, one affine weight, known common slopes and offline reference responses. Both empirical domains are synthetic; capacity extends the decision structure, not the physical-scale theorem or real-world validation. Small numbers of ID violations limit miss-rate precision, one random grid is not an average over all grids, and local floating-point agreement is not an exact-arithmetic certificate for arbitrarily near-parallel lines. Within this scope, the diagnostic identifies what a response representation must preserve for an entire decision family and what sparse success fails to establish.

ETHICS STATEMENT

Scientific data are synthetic physical responses, analytic conditional-shortage labels, and offline software records; there are no newly collected participants, vehicles, driving traces, or personal profiles. Parameters and residual effort are declared inputs, not identified human intentions. The licensed photograph in Figure 1 supplies physical context only; it is neither a study apparatus nor a model input. Its source, crop and reuse license are recorded in Appendix I.3. A marginal simulator-regret report does not establish actuator tracking, driver consent, vehicle stability, or a safe maneuver. Hardware use requires independent identification, qualified force limits, arbitration and appropriate human-research review. The artifact is not vehicle firmware.

REPRODUCIBILITY STATEMENT

Appendices A–I specify proofs, data roles, complete comparisons and executable sources. The release data, checkpoints and earlier analyses are preserved. New work comprises same-window diagnostic comparisons, an independent hull reference, and a synthetic capacity task with three MLP fits and one validation-selected polynomial ridge. Capacity models are frozen before independent calibration/test generation; release reanalyses are retrospective. Python executes all newly reported numerical studies and figures. A single `SlewMatch_ICLR2027_V6_Figures` file renders Figures 3–18 from embedded numeric arrays using native MATLAB plotting. MATLAB/Octave is unavailable in the authoring environment, so numeric read-back is not runtime or pixel-equivalence verification. Editable figure sources and photo attribution accompany both packages.

AI USE STATEMENT

An OpenAI assistant substantially assisted research framing, primary-literature verification, theoretical development and proof writing, implementation, experiment design, synthetic-data generation, model training, numerical execution, statistical summaries, interpretation, manuscript drafting, plotting and artifact checks across revisions. Values derive from retained records or actual executable computations, not plausible-looking generated measurements. The release cohorts followed an earlier local score/model freeze; subsequent diagnostic and budget comparisons reuse examined records. The capacity generator and model budgets were recorded before fitting, with hashes frozen before its independent calibration/test generation. These are local execution records, not external preregistration or independent-laboratory replication. The contextual photograph is attributed third-party material, not AI-generated experimental evidence. No language-model controller, human adaptation model or physical-vehicle experiment is evaluated. Authors remain responsible for scientific claims, proofs, provenance, source rights, code and submission disclosures.

REFERENCES

Joseph Bakarji, Jared Callaham, Steven L. Brunton, and J. Nathan Kutz. Dimensionally consistent learning with Buckingham Pi. *Nature Computational Science*, 2:834–844, 2022. doi: 10.1038/s43588-022-00355-5.

Edgar Dobriban and Mengxin Yu. SymmPI: predictive inference for data with group symmetries. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(5):1353–1381, 2025. doi: 10.1093/jrsssb/qkaf022.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Alexandre Girard. Dimensionless policies based on the Buckingham π theorem: Is this a good way to generalize numerical results? *Mathematics*, 12(5):709, 2024. doi: 10.3390/math12050709.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Jayanta Mandi, Víctor Bucarey, Maxime Mulamba Ke Tchomba, and Tias Guns. Decision-focused learning: Through the lens of learning to rank. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14935–14947, 2022. URL https://proceedings.mlr.press/v162/mandi22a.html.

Yash P. Patel, Sahana Rayan, and Ambuj Tewari. Conformal contextual robust optimization. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2485–2493, 2024. URL https://proceedings.mlr.press/v238/patel24a.html.

William Therrien, Olivier Lecompte, and Alexandre Girard. Using the Buckingham π theorem for multi-system transfer learning: A case-study with 3 vehicles sharing a database. *Electronics*, 13 (11):2041, 2024. doi: 10.3390/electronics13112041.

Soledad Villar, Weichi Yao, David W. Hogg, Ben Blum-Smith, and Bianca Dumitrascu. Dimensionless machine learning: Imposing exact units equivariance. *Journal of Machine Learning Research*, 24(109):1–32, 2023. URL https://jmlr.org/papers/v24/22-0680.html.

Wenbin Zhou and Shixiang Zhu. Calibrating decision robustness via inverse conformal risk control. *arXiv preprint arXiv:2510.07750v3*, 2026. URL https://arxiv.org/abs/2510.07750.

A PROOFS AND THE EXACT INFORMATION BOUNDARY

A.1 EXISTENCE, ANALYTIC NORMALIZATION, AND PHYSICAL SCALING

The forcing $q(t) = H(g_d(t) - w_D(t))$ is bounded by H in absolute value. The physical state vector field is continuous in time and locally Lipschitz in state. To exclude finite-time escape, use mechanical energy

$$\mathcal{E}(x, \dot{x}) = \frac{1}{2}J\dot{x}^2 + \frac{1}{2}kx^2 + \frac{1}{4}k_3x^4, \qquad \dot{\mathcal{E}} = q\dot{x} - b\dot{x}^2 \le \frac{q^2}{2b} - \frac{b}{2}\dot{x}^2 \le \frac{H^2}{2b}. \tag{12}$$

Positive J, k make this energy coercive in both state coordinates. On every finite time interval, the energy and state remain bounded, extending the unique solution. This is a mathematical property of the declared reduced-order forced system, not a stability or passivity certificate for a coupled driver–vehicle loop.

Substituting $t = \rho\tau$ and $x = (H/k)y$ gives $Ay'' + By' + y + \chi y^3 = e^{-\max(\tau-\delta,0)} - w_v(\tau)$ with the groups in Equation 10. Under Equation 11, every group is identical: for example, $(s^2J)/(k(s\rho)^2) = A$ and $(k_3/a^2)(aH)^2/k^3 = \chi$. Initial conditions, normalized forcing and candidate bounds also coincide. Uniqueness then gives the same normalized trajectory and P; directly, $x_{a,s}(t) = ax(t/s)$. The transformed duration interval is

$$[1.5(aH)/(aS/s), sL] = s[D_{\min}, L]. \tag{13}$$

The corresponding ordered normalized candidate arrays are therefore the same. A deterministic function of these arrays, including a fixed nonnegative clipping rule, has identical outputs and selected indices. This proves the stated scale property without a learned approximation assumption. It does not establish that every physically different system with the same five numbers is on this particular two-parameter orbit, nor that changing H while keeping nonzero k_3 fixed is covered.

For our numerical target, normalized RK4 uses the same $\Delta\tau = .01$ grid and horizon 20 before and after a matched transformation, so the target computation is also invariant in exact arithmetic. Powers-of-two transformations are checked at the retained floating precision. An arbitrary real-valued transformation can introduce ordinary rounding differences; the theorem concerns the specified mathematical maps, not bitwise guarantees for every floating-point input.

A.2 CANDIDATE ERROR AS DISTANCE MODULO OFFSETS

For a single error vector $e \in \mathbb{R}^K$, write $e_{\max} = \max_j e_j$ and $e_{\min} = \min_j e_j$. For every constant b,

$$\|e - b\mathbf{1}\|_\infty \ge \max(e_{\max} - b, b - e_{\min}) \ge \frac{e_{\max} - e_{\min}}{2}. \tag{14}$$

Choosing $b = (e_{\max} + e_{\min})/2$ attains equality. This proves the quotient interpretation of half the error span. It is not an assertion that the MLP itself learns a quotient representation. The encoder is fixed, the response is learned, and the error is measured with a decision-relevant functional after training.

Let $j^\star$ minimize true cost and $\hat{j}$ minimize predicted cost. Since the known delay term is identical in both costs,

$$\begin{aligned} R_\lambda &= C_{\lambda,\hat{j}} - C_{\lambda,j^\star} \\ &= -e_{\hat{j}} + (\widehat{C}_{\lambda,\hat{j}} - \widehat{C}_{\lambda,j^\star}) + e_{j^\star} \le e_{j^\star} - e_{\hat{j}} \le e_{\max} - e_{\min}. \end{aligned} \tag{15}$$

Nonnegativity follows from true optimality. Taking maxima over fixed fits gives Equation 3, and $e_{\max} - e_{\min} \le 2\max_j |e_j|$ gives the last link. A common emitted offset may have arbitrary magnitude while leaving both span and decisions unchanged. This example demonstrates the distinction between required information, not an empirical property assumed for all fitted errors.

For the margin bound, if $j^\star = \hat{j}$ then regret is zero. Otherwise $\widehat{C}_{\lambda,j^\star} - \widehat{C}_{\lambda,\hat{j}} \ge \widehat{\gamma}_\lambda$, so $R_\lambda \le \Omega - \widehat{\gamma}_\lambda$. If this upper expression is negative, a different true minimizer is impossible on the span event, leaving the zero-regret case. Thus $R_\lambda \le [q - \widehat{\gamma}_\lambda]_+$ whenever $\Omega \le q$. The candidate set

$$\mathcal{A}_q = \{j : \widehat{C}_{\lambda,j} - \min_i \widehat{C}_{\lambda,i} \le q\} \tag{16}$$

contains a true minimizing action on the span event: inserting the true optimum into the same inequality bounds its predicted excess cost by the error span. Counting this set measures report informativeness. The implementation does not use it to choose, reject, or rerank an action.

A.3 Computing the objective-family regret envelope

For each predicted line $p_j + \lambda v_j$, all pairwise optimality constraints have the form $(v_j - v_i)\lambda \le p_i - p_j$. Positive slope differences impose upper limits; negative differences impose lower limits; equal slopes either impose no restriction or make the candidate impossible. Intersecting with $\lambda \ge 0$ produces $I_j = [\ell_j, u_j]$, an empty interval, or an unbounded final interval. The code retains isolated ties with $\ell_j = u_j$, rather than silently dropping them.

For a fixed action j, its true regret at a weight is

$$f_j(\lambda) = P_j + \lambda v_j - \min_i (P_i + \lambda v_i) = \max_i [(P_j - P_i) + \lambda(v_j - v_i)]. \tag{17}$$

This maximum of affine functions is convex. On a finite closed interval its maximum occurs at an endpoint. An action optimal for arbitrarily large λ must have a minimum candidate slope; otherwise some smaller-slope action eventually dominates it. On that final interval, the limiting regret is $P_j - \min_{i:v_i=v_j} P_i$. With strictly increasing duration candidates, the minimum-slope index is unique and this limit is zero. Equal-slope repeated candidates are handled explicitly.

The algorithm constructs all MK intervals with K pairwise comparisons each, then evaluates at most two endpoints per interval against K true lines. This is $O(MK^2)$ time and requires no dense weight scan or differentiation through a solver. The supremum over any predicted minimizer is slightly stronger than the smallest-index tie rule at an isolated tie, as declared in Equation 2. Evaluation uses pairwise cost differences instead of subtracting two very large common costs. Unit tests compare this algorithm with dense random weight checks, exact two-action switch examples, repeated slopes, and the deterministic score hierarchy.

The score G is a directed decision functional, not a norm or symmetric distance between response functions. It depends on the fixed objective family, candidates and delay slopes. A candidate-error span controls any pairwise comparison and therefore provides a more general sufficient statement; G trades that breadth for an exact description of this declared decision family. Changing the family or adding candidates requires recomputing the score and checking its assumptions. We do not transfer an old G guarantee merely because some mean predictions remain similar.

A.4 Conformal rank, simultaneous fits and deterministic transport

Conditional on fixed training data, validation choices, score definitions and included fits, the score of each fresh context is the same deterministic measurable functional of that context and its offline reference responses. Under exchangeability, the rank of the new score among the $n+1$ scores is uniform absent ties; ties make the non-strict coverage event conservative. Thus for $k = \lceil (n+1)(1-\alpha) \rceil \le n$,

$$\Pr[Z_{n+1} \le Z_{(k)}] \ge k/(n+1) \ge 1 - \alpha. \tag{18}$$

The $+\infty$ convention covers $k > n$. With $n = 256$, the ranks at nominal 90%, 95%, and 99% are 232, 245, and 255. Applying this standard argument separately to U, Ω and G, then using the deterministic reductions, establishes the marginal statement in Section 2.4. Including the maximum over three fits supplies a joint statement about that pre-fixed family; it is not a posterior over models or a guarantee over arbitrary future training runs.

Pointwise inequalities $G_i \le \Omega_i \le 2U_i$ imply the same order for their identically ranked order statistics. This establishes width non-increase without assuming the scores have similar distributions. It does not imply that the three nominal events are jointly valid at the same nominal level. A joint response-and-decision statement could, for example, allocate $\alpha_U + \alpha_G \le \alpha_{\text{total}}$; that different construction is not the reported experiment.

Under matched physical scaling, the reference vector, postprocessed prediction vector and normalized candidate slopes are identical. Every score and normalized selected cost is therefore identical, and each coverage indicator is the same random event at every scale of a context. There is no union bound over those copies and no requirement that the physical-scale sampling distribution be invariant. This

deterministic transfer applies to failure indicators too. Re-estimated metadata, hidden actuator state, a different forcing law, or an altered candidate problem need not preserve it. For a proposed approximate extension one must specify errors in target, predictor and candidates separately; especially a changed delay term multiplied by unbounded λ cannot be hidden inside an unspecified uniform additive error.

A.5 Which comparisons matter: an exact nonnegative accounting

Fix a fit and omit its subscript. Let J_i and I_j be the true and predicted minimizing intervals, and define $\mathcal{Q} = \{(i, j) : J_i \cap I_j \neq \varnothing\}$. Both envelopes have a minimizer at every nonnegative weight, so $\mathcal{Q}$ is nonempty. At any overlap weight, true optimality and predicted optimality imply

$$0 \leq C_j - C_i = (e_i - e_j) - (\widehat{C}_i - \widehat{C}_j) \leq e_i - e_j. \tag{19}$$

Consequently $B = \max_{(i,j)\in\mathcal{Q}}(e_i - e_j) \geq 0$. Every regret witness is such a pair, which proves $G \leq B \leq \Omega$. This definition is deliberately stricter than "both actions are active somewhere": their active intervals must overlap at the same weight. It is also deliberately broader than the actions chosen at any finite list of probe weights. Singleton overlap intervals are retained.

For $a = \min e$, $b = \max e$, $2\max(|a|, |b|) = (b - a) + |a + b|$. Since the min–max midpoint is $(a + b)/2$, this is $2U = \Omega + 2|b_m|$. Insert B and G between Ω and zero to obtain Equation 6. Each component is nonnegative. $\Omega - B$ measures comparison directions excluded by the paired optimality geometry. $B - G$ is the remaining slack after the admissible comparisons meet their predicted margins and weight restrictions; the maximizing pair for B need not be the regret witness for G. We therefore call it margin slack, not the predicted gap of one universal worst pair.

At an actual witness (i, j, λ) the stronger equality in Equation 5 holds, and the implementation records both its relative error and its predicted margin. At an infinite endpoint, relevant candidates have equal minimum slope, so the same calculation reduces to intercept differences. When all candidate errors equal a constant, $\Omega = B = G = 0$; when $U = 0$, all components are zero and a component ratio is undefined. Code marks such ratios missing rather than adding an arbitrary denominator. Our data have no zero-U records among the audited fits, but tests cover the limiting case.

The accounting applies to each fit–context pair before aggregation. We average those components with equal context/fit weights and divide by mean $2U_m$ for Figure 7. This is not the mean of per-record ratios, not a decomposition of $2\max_m U_m$, and not an additive explanation of $1 - q_G/(2q_U)$. The last ratio compares distinct order statistics; different contexts or fits may determine their quantiles. No new calibration score or population-level causal attribution is inferred from the diagnostic identity.

A.6 Zero decision loss, matched switches, and a finite-probe counterexample

Let $\mathcal{S}(\lambda) = \arg\min_i C_i(\lambda)$. Equation 2 gives $G = 0$ if and only if $\widehat{\mathcal{S}}(\lambda) \subseteq \mathcal{S}(\lambda)$ at every $\lambda \geq 0$. The forward direction holds because each nonnegative audited regret must then be zero. The reverse direction holds because every selectable candidate is truly optimal. Equality of complete response vectors is unnecessary; the relation is directed, and a candidate that is never selected can have different values as long as it does not enter the predicted envelope. Allowing it to enter changes the relation and cannot be ignored merely because it was inactive under the original predictor.

For a common adjacent switching pair with unequal slopes, equality of the two true affine costs gives $(v_j - v_i)\lambda^\star = P_i - P_j$; equality of the predicted costs gives the same equation with P replaced by p. Subtraction proves Equation 7. We match ordered pairs of adjacent *positive-length* intervals. Isolated ties remain in the loss audit but are not treated as duration-bearing switches. Different active sets and unmatched transitions have their own counts rather than imputed zero displacement.

For any finite collection of nonnegative probe weights, choose $0 < a < b$ in an open interval containing no probe. For $s \geq 0, d > 0$, let $v = (s, s + d)$, $P = (db, 0)$ and $p = (da, 0)$. Both selectors choose the second action below a and the first above b. Between them, prediction has switched but truth has not; its regret is $d(b - \lambda)$, with worst value $d(b - a)$ at the predicted tie (or the corresponding right limit). Thus every probe has zero regret while the all-weight envelope is positive. Scaling d makes it arbitrarily large absent a slope restriction. If the weight range is prescribed and contains a gap between probes, the same construction lies entirely within that range. This is an elementary counterexample, not a new statistical impossibility theorem.

Figure 1 uses the explicitly marked alternative numerical example $P = (.20, 0)$, $p = (.16, .02)$, $v = (1, 3)$. Its true/predicted switches are .10 and .07 and its all-minimizer envelope is .06. The three probes .03, .10, .30 have zero true regret; at .10 the true candidates tie. These algebraic values are not substituted for any simulated context.

A.7 Numerical conventions and witnesses

The exported interval implementation uses float64 comparisons on the supplied finite values, distinguishes exactly equal slopes from unequal slopes, and does not bucket close slopes or truncate large weights. A zero-width feasible interval is retained; an interval with upper endpoint below its lower endpoint is empty. Finite-endpoint losses use pairwise differences $(P_j - P_i) + \lambda(v_j - v_i)$ rather than subtracting large common costs. The real-arithmetic result is exact; ill-conditioned nearly parallel inputs can still amplify ordinary floating-point error. Tolerances in verification assertions do not authorize altering a minimizing interval or discarding a sample. Tests cover equal/repeated slopes, all-tied lines, isolated ties, the two-action counterexample, permutation/common-offset invariance, and random comparisons against a dense diagnostic scan. Agreement on those tests is not an interval-arithmetic proof for every float input.

A.8 Different finite reference sets and the unbounded endpoint

The same candidate intervals can audit the selector against another finite reference with true responses P_i^{ref} and slopes v_i^{ref}. Evaluate $(P_j - P_i^{\mathrm{ref}}) + \lambda(v_j - v_i^{\mathrm{ref}})$ at the selector interval endpoints. If a final selected action has slope greater than the reference's minimum slope, its limiting regret is $+\infty$, which is returned explicitly. If the slopes coincide, the limit is the difference from the smallest reference intercept at that slope. This prevents extending an all-weight finite bound across a changed delay term without accounting for the discrepancy. In the 21/201 audit, endpoint slopes agree and the coarse grid is a subset of the fine one.

A.9 A narrow false active cell with fixed bounded slopes

Let $0 < a < b$ and $\eta > (b - a)/2$, with $v = (0, 1, 2)$, $p = (a + b, a, 0)$ and $P = (a + b, a + \eta, 0)$. Predicted costs intersect successively at a and b, so candidate 1 is optimal exactly on $[a, b]$, candidate 2 before it, and candidate 0 after it. The two true endpoint costs meet at $h = (a + b)/2$. The middle true line lies strictly above their lower envelope: its minimum excess occurs at h and is $\eta - (b - a)/2 > 0$. Thus the true middle action has no minimizing cell.

On the false predicted cell its regret is

$$R_1(\lambda) = \begin{cases} \eta + a - \lambda, & a \le \lambda \le h, \\ \eta + \lambda - b, & h \le \lambda \le b. \end{cases} \tag{20}$$

It equals η at both endpoints and is positive in between. Outside $[a, b]$, the predicted and true selected endpoints agree. The all-minimizer supremum is η; under smallest-index deployment the same supremum is attained at a or approached from the appropriate cell side. Taking $b - a$ arbitrarily small leaves slopes in $[0, 2]$ and $\|p - P\|_\infty = \eta$ unchanged. Every fixed finite probe set has a gap in which such a cell can be placed. This is a deterministic elementary construction, not a claimed new statistical lower bound. It explains why bounded line slopes alone do not yield a global Lipschitz constant for the loss after a discontinuous argmin.

Figure 3 uses $a = .071, b = .079, \eta = .05$ and candidate IDs 0,1,2. The witness at .071 is $i = 2, j = 1$, with $e_2 - e_1 = .05$ and predicted margin zero. The eligible intersection is $J_2 \cap I_1 = [.071, .075]$; $J_0 \cap I_2$ is empty. Original probes .03, .10, .30 have zero regret. These algebraic values never replace a simulated response or a capacity prediction.

A.10 Independent classical envelope route and numerical semantics

Our independent reference sorts lines by decreasing slope and maintains their lower hull with increasing intersection locations. Exactly equal slopes retain only the smallest intercept, but every exactly coincident line identity remains. A strict stack-pop inequality preserves candidates active only at a singleton; close unequal slopes are not bucketed. This is a standard envelope construction,

consistent with the minimization-diagram cells, degenerate overlaps and line–hull duality described by **?**. Our implementation is not CGAL and does not claim CGAL's exact predicates.

The vertex route builds only the predicted diagram and evaluates true regrets for the known active identities at every clipped cell endpoint. This yields the same mathematical maximum as Algorithm 1. The independent joint route builds both diagrams, intersects closed cells and maximizes their affine true-cost differences on each intersection. It also computes B from the corresponding eligible error directions. The implemented overlay uses a straightforward pair loop, with conservative $O(K^2)$ work rather than claiming the fastest diagram merge. Hull construction, tie handling and endpoint sufficiency are inherited computation; the joint error accounting and diagnostic outputs are the added interface.

The study compares predicted cell identities and endpoints on 85,226 active cells. The pairwise and independent joint-route losses differ by at most 1.12×10^{-16} over 8,192 context–family evaluations, which share 1,536 unique test contexts across the two tasks. Unit tests also compare integer-coefficient cases against an independently enumerated rational arrangement, and cover repeated slopes, coincident lines, singleton ties, $\lambda = 0$, unbounded cells, negative slopes, common offsets, permutations and near-parallel inputs. Endpoints use long-double intersection arithmetic in the hull route and float64 in the original pairwise route. Assertions use tolerance, but neither algorithm silently widens minimizing sets or removes contexts. These checks are not an interval-arithmetic guarantee for every ill-conditioned input.

B Position relative to the closest methods

The interpretation of the inherited and added objects is explicit. Our contribution is not automatic dimensional-group discovery, exact units equivariance alone, a new MLP architecture, a new conformal quantile, or the general observation that decision loss differs from prediction loss. Table 3 separates the relevant prior objects. The empirical claims concern post-fit diagnostics on a synthetic physical task and a separate synthetic conditional-shortage prediction task; they do not constitute a benchmark against the richer systems studied by those works.

Computational geometry and SPO theory. A minimization diagram already records active intervals and tied vertices; divide-and-conquer construction, linear diagram merging and line–convex-hull duality are classical (**?**). Our sorted-hull reference makes that antecedent executable. **?** study generalization for noncontinuous SPO loss and derive margin-based analyses; Mandi et al. (2022) use ranking to train decision-focused predictors. Those are different objectives from our fixed-vector audit. We attach the relation $J_i \cap I_{m,j} \neq \varnothing$, the directed error $e_{m,i} - e_{m,j}$ and the actual predicted margin to the known cells. This explains which comparison is responsible for a loss and what a probe misses, without claiming that a new geometric data structure or fitting principle is required.

The useful distinction is between two reductions. Physical encoding removes declared scale nuisance while retaining normalized dynamics and action bounds. Candidate comparison then removes common response offsets that cannot affect the known-cost selector. Finally, the envelope restricts attention to actions reachable under the fixed one-parameter objective family. The latter two are calibrated error objects, not assertions that the response network learns an interpretable internal world model. Other architectures, discovered invariants or decision-focused losses may improve the learned approximation; this study neither evaluates nor rules out those directions.

? is published in TMLR (December 2025; revised arXiv v2 in February 2026), not merely an unreviewed title-matched preprint. Its end-to-end method learns decision-relevant uncertainty sets, including partially input-convex neural representations, and differentiates through calibration and robust optimization. SlewMatch instead leaves the response-MSE fit and nominal argmin fixed. Direct scalar loss calibration alone is not claimed as new; the reusable additional object is the exact finite-affine interval audit together with the jointly active comparison criterion and error witness. Its scope is narrower than general conditional robust optimization. We neither implement that richer optimizer as a baseline nor claim that its guarantees are subsumed by our result.

Table 3: What is inherited and where the present question differs. These are methodological distinctions, not measured external baseline scores.

Work	Main object / assumptions	Relation to this study
Villar et al. (2023)	Analytic dimensionless inputs enforce exact units equivariance.	Our physical encoder follows this known principle; novelty is not the scale-factor removal.
Bakarji et al. (2022)	Data-driven groups, including a learned BuckiNet projection, with dimensional consistency constraints.	We supply the governing groups rather than discovering them; no empirical group-discovery comparison is claimed.
Girard (2024); Therrien et al. (2024)	Dimensionally similar policies and shared vehicle-model data.	Exact policy transfer and label efficiency have precedent. Our case has no real-vehicle measurements.
Dobriban & Yu (2025)	Predictive inference under distributional group invariance and equivariant processing.	We combine base-context exchangeability with pointwise physical response/predictor equality, not an invariant physical-scale distribution.
Elmachtoub & Grigas (2022); Mandi et al. (2022)	Decision-focused and ranking-based fitting losses.	Our predictors fit response MSE; span and envelope scores calibrate a frozen selector rather than train a decision-loss model.
Patel et al. (2024)	Conformal uncertainty regions enter robust optimization.	Our scalar reports do not enter the argmin; response and decision events are separated.
Zhou & Zhu (2026)	A recent preprint calibrates robustness through an inverse risk-control problem.	This is related positioning, not an obligatory benchmark or an independently validated baseline here.
?	End-to-end learning of convex uncertainty sets using downstream robust-optimization loss, with conformal calibration.	Their uncertainty model affects decisions. Here fixed nominal decisions are audited through an exact objective-family loss and its witness; no uncertainty set is fitted or optimized.

C Same-domain audit design and complete diagnostic comparisons

C.1 Fixed probes, adaptive cells and all-minimizer meaning

The recorded plan fixes ten comparators before their execution. Release scans use $[0, .5]$; capacity scans use $[0, 1.5]$. The three release probes are $.03, .1, .3$. The 84-point design is the union of 81 logarithmically spaced values from $.001$ to $.5$ and the three original probes; it does not include zero. Uniform designs have 257 or 1,025 endpoint-inclusive points. RNG 9601 supplies one sequence of 1,023 uniform draws; the first 255 plus both endpoints define Random257, while all draws plus endpoints define Random1025. Capacity multiplies the original 3/84 coordinates by three and generates the other designs on its own window. We do not average over random grids, choose a favorable seed, or claim pointwise monotonicity between nonnested designs.

Cell midpoints use the prediction-only hull, one midpoint per nonempty bounded cell, the window endpoints and singleton identities. Hull vertices instead evaluate every known active identity at all cell endpoints. The distinction matters when an endpoint tie or one-sided loss is larger than the interior value. Fixed probes compute all predicted minimizers at a point with long-double cost arithmetic, without an arbitrary near-minimum tolerance. The two exact implementations use closed-cell all-minimizer semantics. Small float differences are checked separately; the deployed smallest-index tie rule is not silently substituted into one comparator.

Table 4: All original release predictors under the same $[0, .5]$ audit. V denotes exact violations; all subsequent counts are violations missed by that audit, using the family-specific unchanged q_G. Every family retains 512 contexts. Both exact implementations and the vertex audit have no misses for every row.

Regime	Method	V	3	84	U1,025	R1,025
ID	Raw	18	17	12	2	3
ID	Gain	20	17	12	4	7
ID	Raw + aug.	18	16	11	2	5
ID	Gain + aug.	26	25	20	6	11
ID	SLEWMATCH	22	18	6	2	2
ID	RBF	23	21	7	1	2
ID	8-NN	19	16	6	2	3
Shift	Raw	301	269	66	5	14
Shift	Gain	316	279	71	10	28
Shift	Raw + aug.	266	233	43	6	7
Shift	Gain + aug.	363	323	95	8	22
Shift	SLEWMATCH	491	423	49	4	14
Shift	RBF	456	326	44	7	7
Shift	8-NN	203	120	8	0	1

Table 5: Capacity audits share $[0, 1.5]$ and all-minimizer semantics. The 3/84 designs scale the release probes by three; the remaining fixed/random designs use this same window. Columns list missed/exact events, P95 underestimation ($\times 10^3$), and local milliseconds. Only nine MLP violations make its conditional miss proportion imprecise. Zero misses do not imply zero worst-loss underestimation.

Audit	MLP miss	Ridge miss	MLP P95	Ridge P95	MLP ms	Ridge ms
3 probes	9/9	23/23	7.741	77.983	0.220	0.155
84 probes	9/9	18/23	5.002	24.933	0.377	0.205
Uniform 257	5/9	3/23	2.064	2.713	0.695	0.313
Uniform 1,025	1/9	1/23	0.661	0.701	2.222	0.812
Random 257	5/9	9/23	2.866	7.383	0.703	0.323
Random 1,025	3/9	0/23	1.339	1.873	2.209	0.819
Cell midpoints	9/9	23/23	6.362	56.906	0.378	0.131
Hull vertices	0/9	0/23	0.000	0.000	0.545	0.171
Pairwise exact	0/9	0/23	0.000	0.000	1.301	0.420
Hull overlay	0/9	0/23	0.000	0.000	0.659	0.217

C.2 MISSED EVENTS, UNDERESTIMATED LOSS AND ERROR LOCATION

For each family/context and window, $\Delta G = G^{\Lambda} - G_{\mathrm{audit}} \geq 0$ in exact arithmetic. A miss requires both $G^{\Lambda} > q_G$ and $G_{\mathrm{audit}} \leq q_G$, using the same family's *full-domain* calibrated width as a fixed diagnostic threshold. This threshold is valid for the smaller window by domination, but the scan itself does not become a new calibrated score. We preserve zero-loss contexts and leave relative gaps undefined at zero rather than adding an arbitrary epsilon. Loss attainment is checked within 10^{-10}; weight distance uses the canonical exact witness and does not establish unique-maximizer recovery when several witnesses are equivalent.

The release full-group family has equal bounded/full-domain maxima in both cohorts. Full-domain loss exceeds the bounded maximum for one ID and one intrinsic 8-NN context; these are retained domain effects. For capacity ridge, seven contexts have a larger full-domain maximum outside $[0, 1.5]$; capacity MLP has none. These cases are not included as failures of a bounded scan to find a weight it was never asked to inspect. All 8,192 family–context rows and 81,920 audit rows retain their domain, fit, candidate pair, weights, loss gaps and event flags.

Table 6: Witness localization and work. Distances use one canonical exact maximizer and are descriptive when maximizers are not unique. Attain means loss within 10^{-10} of the exact maximum. Vectors counts candidate-vector cost evaluations, summed across fixed fits; hull construction is additional work. Hull-overlay scalar pair evaluations are separately recorded, not counted as vectors.

Task	Audit	Median $\lvert\Delta\lambda\rvert$	P95	Attain (%)	Vectors
ID	84	0.04344	0.14052	0.0	252.0
ID	U1,025	0.00024	0.00280	1.8	3075.0
ID	Vertices	0.00000	0.00000	100.0	24.0
Shift	84	0.00174	0.00497	0.0	252.0
Shift	U1,025	0.00024	0.00046	0.0	3075.0
Shift	Vertices	0.00000	0.00000	100.0	16.8
Capacity MLP	84	0.39929	0.95911	0.0	252.0
Capacity MLP	U1,025	0.00074	0.26499	1.4	3075.0
Capacity MLP	Vertices	0.00000	0.00000	100.0	76.0
Capacity ridge	84	0.03259	0.81380	0.0	84.0
Capacity ridge	U1,025	0.00069	0.00142	2.1	1025.0
Capacity ridge	Vertices	0.00000	0.00000	100.0	22.4

The 1,025-point random capacity-ridge scan finds all 23 observed violations yet has nonzero mean loss underestimation (.000590). Conversely, capacity MLP has only nine violations, so one additional miss changes its conditional miss fraction by more than eleven percentage points. We report raw denominators and do not interpret this small conditional sample as a precise population ranking. Fixed-grid success at a threshold, recovery of the worst value and recovery of a canonical witness are distinct outcomes.

C.3 What the timing and work counts include

The microbenchmark fixes the first 32 contexts per cohort/family, one numerical thread, two warm-ups and five repetitions. We first take the median repetition time for each context and then its cohort median. The measured function includes input checks, the prescribed probe construction or hull/interval construction, candidate-cost evaluation and witness output. It excludes file loading, model inference, analytic-label computation and every ODE integration. Grid arithmetic uses long double for tie comparisons; geometry uses the declared float/long-double conventions. These are local implementation times, not complexity lower bounds or an optimized-library benchmark.

A vector evaluation tests costs against the full true candidate vector. On intrinsic release, hull vertices average 16.83 such evaluations across all three fits, versus 3,075 for 1,025 probes. Hull-overlay work is different: it averages 22.83 scalar eligible-pair endpoint evaluations, with true/predicted hull construction and cell-pair intersection overhead in addition. The CSV has separate vector and scalar counters; they must not be equated. The measured median joint-overlay time (.537 ms) exceeds vertex-only time (.313 ms) because it produces additional comparison information. Timing is rerunnable but not deterministic across hosts.

D Independent synthetic conditional-shortage prediction

D.1 Reference quantity and known capacity costs

Context $z \sim \mathrm{Uniform}([-1,1]^4)$ defines a mixture of lognormal demands with weight $\omega = .15 + .7\,\mathrm{sigmoid}(1.5z_1 - .8z_2)$ and parameters

$$\mu_1 = \log(.75) + .25z_1 + .12\sin(\pi z_2), \qquad \mu_2 = \mu_1 + .55 + .1z_4, \tag{21}$$

$$\sigma_1 = .2 + .05(z_3 + 1), \qquad \sigma_2 = .3 + .05(z_2 + 1). \tag{22}$$

For a lognormal component, the expected shortfall at positive capacity a is

$$S(a;\mu,\sigma) = e^{\mu+\sigma^2/2}\Phi\left(\frac{\mu+\sigma^2-\log a}{\sigma}\right) - a\Phi\left(\frac{\mu-\log a}{\sigma}\right). \tag{23}$$

At $a = 0$ it is simply $e^{\mu+\sigma^2/2}$. The reference P_j is $\omega S(a_j; \mu_1, \sigma_1) + (1-\omega)S(a_j; \mu_2, \sigma_2)$, with $a_j = 2.8(j/16)^{1.3}$ for $j = 0, \ldots, 16$. No random realized demands or finite-reference empirical means are substituted for this conditional expectation. Numerical quadrature tests independently check the analytic expression.

Costs are increasing and uneven. Starting at $v_0 = 0$, for $r = 1, \ldots, 16$,

$$v_r - v_{r-1} = (a_r - a_{r-1})[1 + .65\sin^2(2.3r + z_3)](.85 + .15z_4). \tag{24}$$

These positive increments are known to the optimizer. Each predictor receives $(z_1, z_2, z_3, z_4, a_j)$ and learns expected shortfall; v_j is not a learned label. The mixture and uneven cost schedule were chosen to test many interior actions and altered active sets, not to model a real company's demand. This is a second synthetic structure, not a real operations dataset or evidence that the release symmetry applies to capacity.

D.2 Fitting, freezing, calibration and testing

Independent RNG streams 9611/9612/9613/9614 create 1,024 training, 256 validation, 256 calibration and 512 test contexts. This is 2,048 distinct contexts and 34,816 analytic candidate labels, not 34,816 independent decisions. Context identifiers retain their stream and index. The protocol is recorded before fitting; model hashes are frozen before the calibration and test files exist. The release arrays and quantiles are never overwritten or used as capacity labels.

The three MLPs use seeds 11/22/33, two 64-unit tanh layers, AdamW with learning rate .001 and weight decay .0001, batch size 512, and 200 epochs (6,800 updates). Training-only feature/target standardization and nonnegative output clipping define the deployed predictions used by every score. Minimum validation MSE selects epochs 197, 197 and 199, with deployed validation MSE approximately $4.076, 4.931, 3.834$ times 10^{-6}. The fixed budget does not prove convergence. No additional capacity MLP was selected using test data.

Cubic polynomial ridge uses all monomials of total degree at most three in the same five inputs, excluding the constant term and using a fitted intercept. Polynomial feature scaling uses training rows only. Validation chooses among penalties $.0001, .01, 1$; $.0001$ is selected, with validation MSE .00074328. One fitted ridge is evaluated, not three independent ridge seeds. Models are supplied as numeric, pickle-free NPZ arrays with an independent NumPy inference routine. The scoring and selector interfaces accept either predictor.

Response RMSE is the square root of mean squared error with equal context, candidate and fixed-fit weights; G instead maximizes over fits and all weights within each context before summarizing contexts. Both families calibrate U, Ω, G using their own 256 held-out contexts and rank 245. The MLP score first maximizes across its three fits; ridge has one fit. Test coverage is 503/512 for the MLP envelope and 489/512 for ridge. The higher MLP count is an empirical frequency, not a target achieved by tuning calibration. Different families' reports are separate marginal statements, not simultaneous 95% validity across models or all quantities.

Table 7: Separately calibrated nominal-95% capacity score events; all counts are out of 512 independent test contexts. Calibration uses rank 245 of 256. Counts test $U \le q_U$, $\Omega \le q_\Omega$ and $G \le q_G$, not interchangeable response or regret events. Different families are not jointly calibrated.

Method	q_U	q_Ω	q_G	U	Ω	G
MLP	0.009513	0.013761	0.011046	500	502	503
Cubic ridge	0.099615	0.141257	0.092232	499	496	489

D.3 What transfers as an empirical diagnostic

The true capacity diagram averages 13.22 active actions (positive-length counts coincide for truth); learned clipping can additionally create singleton ties at zero weight, so full and positive-length active counts remain separate fields. For the MLP, same-active-set frequency is 54.9%, compared with 5.1% for ridge. The decomposition assigns 24.0%/28.7%/3.6%/43.6% of mean $2U_m$ to offset/ineligible/margin/retained components for MLP, and 25.4%/26.2%/6.3%/42.1% for ridge. This

identity-based accounting does not show that removing an error term through training would cause that exact performance gain.

The same-window probe comparison applies without requiring the MLP to dominate a leaderboard: both fitted response types can lose or add active actions, both can pass finite checks with a larger envelope loss, and both admit a classical exact vertex audit with a diagnostic joint overlay. The experiment supports portability of the offline interface across these two synthetic structures, not universal predictive validity, a real demand forecast, or deployment cost savings over the available analytic reference.

E Data roles, model selection, and execution chronology

E.1 Chronology and independent data roles

Version names below identify source snapshots rather than separate independent replications. Original release development (V3) preceded a locally frozen score/model evaluation (V4). The error-geometry, repartition and longer-budget work (V5) reused its examined outcomes. Current same-domain release audits also reuse them. Only the capacity task adds new labels and new independent contexts in this revision; its protocol and selected-model freeze precede its held-out data. No analysis is represented as external preregistration.

The immutable historical data comprise 16,000 training responses (RNG 4101), 2,000 response-validation examples (4102), 2,000 independent response-test examples (4103), 256 calibration contexts (4201), 512 ID decision contexts (4202), and 512 intrinsic-shift contexts (4203). The four original representations have seeds 11, 22 and 33, totaling twelve fitted networks. V3 already analyzed these contexts, so reusing them to examine new scores is exploratory. The original numerical and figure records remain separately available.

The local V4 protocol file fixed U, Ω, G, all new model grids, the label-budget schedule, seeds, weights, diagnostic strata and lag levels before the new confirmation outcomes were generated. After training/validation selection, a second freeze recorded hashes of the 45 scientific neural checkpoints and three RBF models. New calibration, ID testing and intrinsic testing then used seeds 8401, 8402 and 8403. No new outcome updated a model, selected a score, tuned a kernel or removed a context. This ordering protects the newly generated evaluation from those choices, but does not turn the retrospective research motivation into external preregistration or independent-laboratory replication.

Table 8: New stored response integrations. Only the first three rows introduce independent contexts (1,280 total); later rows reuse their context identifiers with different dynamics or candidates. The sum is 178,176 response integrations, not 178,176 independent tasks.

Role	Contexts	Responses	Purpose
New calibration	256	5,376	Score quantiles
New ID test	512	10,752	Confirmation
New intrinsic test	512	10,752	Declared shift
Paired lag tests	512	43,008	Four nonzero lags
Paired lag calibration	256	5,376	Fixed $\ell = .2$
Fine-grid ID audit	512	102,912	201 candidates

The 1,280 new independent context identifiers are disjoint from all old identifiers. Each independent decision context has 21 correlated candidate responses. Four matched physical scalings produce exact derived records rather than additional ODE labels. Lag tests and the 201-candidate audit share the indicated ID contexts. The 178,176 stored response count excludes additional independent-solver checks, step-halving checks, the Figure 1 illustration, and timing-only reevaluations; those are separately recorded below. It must not be combined with derived decision counts to suggest a larger sample of independent tasks.

E.2 Exact parameterization and generator ranges

Stored physical inputs are $[H, J, b, k, d, \rho, D, k_3]$; normalized inputs are $[A, B, \delta, v, \chi]$. In the training sampler, $v \sim U(.65, 5.6)$ independently of the four dynamical groups. The scale variables are $H \sim U(.6, 1.9)$ N m, $k \sim U(2, 3)$ N m/rad and $\rho \sim U(.065, .13)$ s. Reconstruction is $J = Ak\rho^2$, $b = Bk\rho$, $d = \delta\rho$, $D = v\rho$ and $k_3 = \chi k^3/H^2$. The decision sampler instead selects $v_{\min}, v_{\max}$ as in the main text and derives $S = 1.5H/(\rho v_{\min})$, $L = \rho v_{\max}$. It therefore constructs feasible tasks; it does not estimate the frequency of feasible physical interventions.

The additional generator stream at seed plus 100 samples decision bounds and, in the intrinsic case, the changed δ and χ. Input IDs are $100000 \times \text{seed} + i$. All ranges are analyst-selected; uniform and log-uniform implementations use their ordinary half-open numerical intervals. This is not a population model of drivers. A separately retained historical 300-context stress samples H, S, L independently and reports 85 empty intervals; that original result is not recounted as new confirmation evidence.

E.3 What the neural objectives and augmentation actually control

For the physical Raw target, $x_{\text{peak}} = (H/k)P$. Squared physical error therefore equals $(H/k)^2(\widehat{P} - P)^2$ before the method's constant target standardization. Normalizing the target changes these sample weights. Raw and Gain both receive eight physical features: logs of positive H, J, b, k, raw d, logs of ρ, D, and $\log(1 + k_3)$. They differ in target units. Their feature standardization and target mean/variance use training records only. Validation reverses target standardization, applies the deployed nonnegative clipping, converts physical predictions to normalized units as needed, and computes normalized MSE.

Raw+augmentation and the new Gain+augmentation use the same declared physical transformations: independent log-uniform amplitude and time factors on $[.5, 2]$, with H, J, b, d, ρ, D, k_3 transformed exactly as in Equation 11. A fixed training-only augmented view (RNG 7299) sets feature scaling; subsequent views use RNG $7300+$ initialization seed. Raw+augmentation transforms the physical target by the amplitude factor. Gain+augmentation leaves the normalized target unchanged. Both reuse the same base ODE response labels. This comparison isolates the previously missing normalized-target augmentation combination, but does not claim an exhaustive search over augmentation distributions or optimizers.

Every new neural fit uses AdamW, learning rate .001, weight decay .0001, nominal batch size 512, 7,680 updates, and validation checks every 32 updates. Epochs are 240, 960 and 3,840 for 16k, 4k and 1k labels; final partial batches produce 3,840,000 training-example presentations in each run. A single permutation (RNG 8410) defines nested 1k and 4k subsets, shared across methods and seeds. The 16k models receive the whole dataset. Validation set size is always 2,000; the additional 256-context candidate calibration is also held fixed across budgets. Consequently the label-efficiency claim concerns *regression* labels only, not total training-plus-validation-plus-calibration expenditure.

Group-removal controls are newly fitted four-feature MLPs obtained by removing one of $\log A$, $\log B$, δ, or χ; the candidate $\log v$ and complete admissible set remain. They are not simulated by zeroing an input of an already fitted full model. Parameter counts are 4,609 for five-feature SLEWMATCH, 4,801 for eight-feature dimensional models, and 4,545 for four-feature omissions. Equal hidden widths do not make parameter counts identical. All model-selection criteria are fixed before fresh calibration and testing.

A local RBF interpolator receives the same five invariant features standardized by its training inputs. At each label budget, a validation-only grid crosses thin-plate-spline/cubic kernels, 32/64/128 neighbors, and smoothing $0, 10^{-5}, 10^{-3}$ (18 configurations, 54 in total). Cubic, 128 neighbors and zero smoothing is selected at all three budgets, with validation MSE .00018228, .00006091 and .00001809. The one fitted kernel per budget is not presented as three independent seeds. The original inverse-distance 8-NN remains for continuity. Neither these choices nor the many checkpoints near the end of the neural training curves establish full optimization of every method; the claim is restricted to the stated resource/search budgets.

Table 9: All 33 newly fitted scientific neural models. Columns 11/22/33 identify initialization seeds. Updates are selected by legacy validation MSE only; every run executes 7,680 updates. MSE is multiplied by 10^6. The separate timing-only fit is excluded.

Family	Labels	Update 11	22	33	MSE 11	22	33
Gain + aug.	16000	7488	7552	7520	10.15	11.84	10.69
Without A	16000	4832	5920	7072	467.56	466.29	467.45
Without B	16000	4928	3808	3328	9273.65	9264.35	9274.57
Without δ	16000	1120	4288	2464	20494.40	20466.40	20441.55
Without χ	16000	7392	7680	7168	65.50	64.43	65.71
Gain	1000	7456	7520	7680	25.94	20.78	25.56
Gain + aug.	1000	7584	7680	7488	13.54	12.13	12.41
SLEWMATCH	1000	7584	7424	7584	5.35	4.29	4.08
Gain	4000	7456	7488	7520	7.86	7.35	7.82
Gain + aug.	4000	7680	7680	7136	12.36	8.34	9.12
SLEWMATCH	4000	7040	7584	7552	3.21	2.64	2.81

F COMPLETE RELEASE RESULTS AND RETROSPECTIVE DIAGNOSTICS

F.1 THE PRIMARY INDEPENDENT-COHORT RESULTS ARE PRESERVED

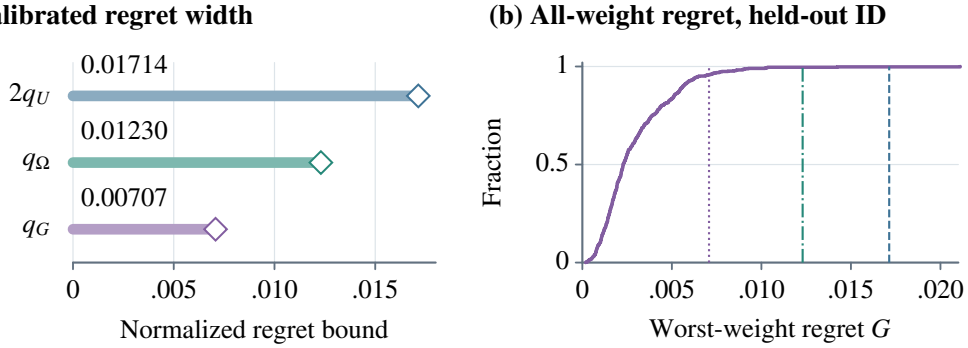
The primary comparisons below remain unchanged. Neural means first average the three fixed fits within each context. The seven learned/neighbor families are calibrated separately; the 8-NN and RBF each have one fitted rule. Tables 10 and 11 retain primary means, tails, baseline rules and calibration denominators. Subsequent diagnostics do not replace them with a more favorable model or calibration set. Quantile order is deterministic from the pointwise score order, whereas finite-test counts and width reductions are empirical.

Table 10: Retained V4 primary comparison ($\lambda = .1$), 512 independent ID contexts. RMSE and regret are multiplied by 10^3. P95 is the 95th percentile of context-mean regret; agreement averages fixed fits. Matched-scale copies do not enlarge the sample.

Method	ID RMSE	ID mean	ID P95	J2 mean	Agree. (%)
Raw	3.34	0.266	1.723	11.588	84.2
Gain	2.14	0.217	1.088	6.131	87.0
Raw + aug.	5.30	0.600	3.454	0.904	79.8
Gain + aug.	2.72	0.225	1.468	0.380	85.5
SLEWMATCH	1.43	0.115	0.522	0.115	89.3
RBF	3.88	0.511	4.044	0.511	83.4
8-NN	23.67	11.415	38.673	11.415	28.9
Fastest	–	40.939	175.753	40.939	46.9
Fixed fraction	–	40.939	175.753	40.939	46.9
Impulse	–	36.634	119.909	36.634	31.8
21-grid oracle	–	0.000	0.000	0.000	100.0

Table 11: SlewMatch nominal-95% calibration and regret counts out of 512. ID score counts test $U \le q_U$, $\Omega \le q_\Omega$, or $G \le q_G$. Regret columns test the displayed bound, jointly over fits. "3 weights means .03, .1, .3; "all uses the exact affine-envelope calculation.

Report	Width	ID score	ID all	Shift 3 weights	Shift all
$2q_U$	0.017137	493	511	464	132
q_Ω	0.012303	492	510	457	63
q_G	0.007075	490	490	444	21



Same 256 calibration contexts; ranks fixed before new test data.

Figure 6: Primary release calibration widths and all 512 held-out ID family maxima. The same 256 calibration contexts and rank 245 produce each report; quantiles and predictions are unchanged. The full-family tail, not a single-weight mean, is the object controlled by the envelope report.

Table 12: Exact per-fit decomposition, then equal-weight means within each retained cohort. Columns are percentages of mean $2U_m$: offset $2|b_m|$, ineligible comparisons $\Omega_m - B_m$, margin slack $B_m - G_m$ and retained G_m. Same set compares the full true/predicted active-candidate sets. These percentages do not decompose a calibration-quantile ratio.

Regime	Method	Offset	Ineligible	Margin	G_m	Same set
ID	SLEWMATCH	27.5	26.3	5.4	40.8	60.0
ID	Gain + aug.	32.5	22.0	6.9	38.6	56.6
ID	RBF	31.6	22.7	10.3	35.4	54.5
Intrinsic	SLEWMATCH	59.5	5.1	0.5	34.9	88.3
Intrinsic	Gain + aug.	63.7	1.3	0.0	35.0	98.3
Intrinsic	RBF	70.5	5.8	2.9	20.8	28.7

F.2 BOUNDED OPTIMIZATION CHECKS AND VALIDATION-ONLY SELECTION

A reviewer-motivated V5 plan fixes six from-scratch trajectories before execution: the full-group MLP and normalized-target Gain+augmentation, each with initialization seeds 11, 22 and 33 and 16,000 regression labels. Each uses the original architecture, preprocessing and AdamW settings, with validation after every 32 updates. A 15,360-update trajectory supplies nested best-validation snapshots at 7,680 and 15,360 updates. There are six additional optimization runs and twelve stored budget snapshots, not twelve independent runs. This design avoids resuming a best checkpoint without its optimizer state, or giving only one representation extra updates. The unchanged original validation set selects each snapshot; the previously examined V4 ID, shift and scaled cohorts are used only after the new model hashes are frozen. This is not independent confirmation.

The V5 RBF grid retains cubic kernels and crosses 128, 256, 512 neighbors with smoothing $0, 10^{-5}, 10^{-3}$ (nine configurations). The 512-neighbor zero-smoothing rule minimizes legacy valida-

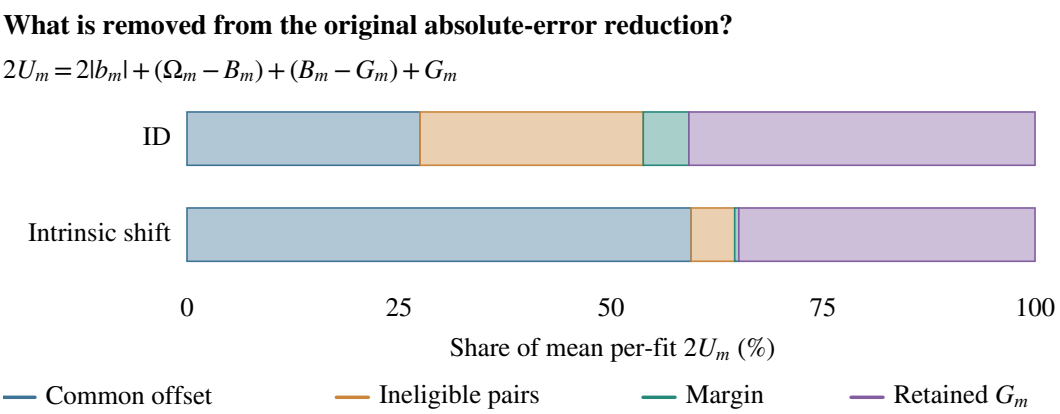


Figure 7: Per-fit information accounting on the two release cohorts. Component means are divided by mean $2U_m$ with equal fit/context weights. A large discarded offset under shift coexists with larger retained regret; these shares do not decompose a ratio of calibration quantiles.

tion MSE, improving it from 1.8087×10^{-5} at 128 neighbors to 1.5380×10^{-5}. Search still ends at its upper neighbor boundary. Full validation histories, selected-update indices, configuration times and model hashes are retained. No recalibration is performed for these changed models, so none inherits the original model family's scalar certificate.

Table 13: Validation-selected budget checks on the same examined V4 contexts. Each neural pair is selected from one newly executed trajectory, not two independent fits. RMSE, regrets and envelope means are multiplied by 10^3; R is at $\lambda = .1$, and G maximizes over all weights and included fits. No new calibration guarantee is attached to these changed models.

Method	Updates	ID RMSE	ID R	ID P95	J2 R	ID G
SLEWMATCH	7,680	1.433	0.115	0.522	0.115	2.945
SLEWMATCH	15,360	1.080	0.117	0.519	0.117	2.378
Gain + aug.	7,680	2.720	0.225	1.468	0.380	5.095
Gain + aug.	15,360	1.791	0.130	0.845	0.234	3.710
RBF (512)	–	3.727	0.469	3.389	0.469	3.828

Table 14: Expanded validation-only cubic RBF search. MSE is multiplied by 10^6. All nine configurations reuse the same 16,000 response labels and 2,000 validation labels. The selected 512-neighbor, zero-smoothing rule is still at the neighborhood boundary; no exhaustive optimization claim follows. Time is the recorded configuration evaluation time in seconds.

Neighbors	Smoothing	Validation MSE	Time (s)
128	0	18.087	0.57
128	1e-05	18.087	0.48
128	0.001	18.130	0.57
256	0	15.689	2.00
256	1e-05	15.690	1.78
256	0.001	15.726	1.86
512	0	15.380	9.15
512	1e-05	15.380	8.88
512	0.001	15.413	8.90

At doubled updates, the paired ID primary-regret difference (full-group minus normalized augmentation) is $-.0000130$, with pointwise 95% context-bootstrap interval $[-.0000522, .0000316]$. At

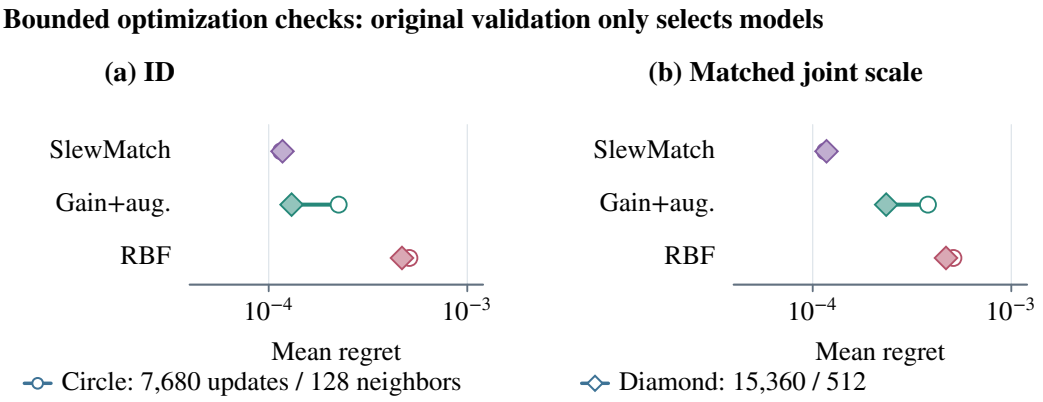


Figure 8: Validation-only budget extension on examined release contexts. Circles are 7,680-update neural snapshots or the 128-neighbor RBF; diamonds are 15,360-update snapshots or the selected 512-neighbor RBF. The neural pairs come from the same six trajectories. No updated certificate is claimed for changed models.

joint scale two it is $-.0001171$, with interval $[-.0002749, -.0000132]$. The enlarged budget thus makes the ID ordering uncertain, while the inspected scale gap remains. On intrinsic shift at .1, both extended neural families give zero regret, preserving the endpoint-insensitivity warning. Their all-weight means remain nonzero: .020255 for the extended full-group family and .020822 for extended augmentation. We do not infer a general OOD ordering from one endpoint or from these bounded runs.

F.3 Retained label budgets, missing groups and omitted lag

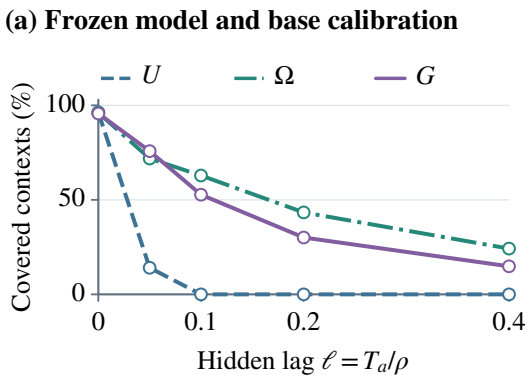
V4's nested 1k/4k/16k regression-label comparison uses 7,680 updates and the same 2,000 validation responses plus 5,376 candidate calibration responses. Thus lower regression-label error is not a proportional reduction in total labeling or computing expense. Each missing-group condition removes one input and is retrained, while retaining candidate duration and complete action bounds. Historical metadata-zeroing is a different intervention. The hidden-lag experiment likewise changes the true response system, not a predictive input alone; its recalibration at the fixed lag .2 only widens the report.

Table 15: Nested regression-label budgets on fresh ID contexts. Neural runs use identical update counts, not epochs; all budgets additionally use the same 2,000 validation labels and 5,376 candidate calibration responses. Errors/regrets are multiplied by 10^3.

Labels	Method	RMSE	Mean regret	CI low	CI high
1000	Gain	4.100	0.302	0.196	0.463
1000	Gain + aug.	2.946	0.227	0.165	0.300
1000	SlewMatch	1.795	0.121	0.082	0.165
1000	RBF	11.670	2.569	2.062	3.110
4000	Gain	2.363	0.218	0.137	0.328
4000	Gain + aug.	2.639	0.229	0.165	0.300
4000	SlewMatch	1.490	0.106	0.069	0.146
4000	RBF	6.716	1.200	0.914	1.522
16000	Gain	2.140	0.217	0.121	0.375
16000	Gain + aug.	2.720	0.225	0.163	0.298
16000	SlewMatch	1.433	0.115	0.078	0.158
16000	RBF	3.885	0.511	0.369	0.659



Figure 9: Previously executed nested-label experiment, same fixed update counts and common held-out ID cohort. The panels preserve response RMSE and primary regret; regret need not improve monotonically with more labels. RBF selection is validation-only at each budget. Shared validation/calibration labels remain overhead.



(b) Recalibrate at $\ell = .2$

Same predictions and actions

	Before	After
q_G	0.0071	0.0602
G coverage	0.3008	0.9746
Mean $R_{.1}$	0.0039	0.0039

Wider report, not a better policy.

Paired 512 contexts; nominal command limits do not certify actuator tracking.

Figure 10: Retained paired hidden-actuator-lag experiment. Five-group predictors omit $\ell = T_a/\rho$. Base quantiles fail at different rates; separate calibration at .2 restores high envelope coverage with a much wider report. Predictions, actions and regret are identical before/after recalibration. Nominal command duration does not certify physical actuator completion.

F.4 What the mean error decomposition measures

The retrospective audit computes 21,760 per-fit context records across seven families and the retained calibration, ID and intrinsic cohorts. It verifies the envelope against the original stored V4 scores and records the exact worst weight, predicted/true action pair, relative error and predicted margin. It also records both full active sets, their positive-length subsets and the jointly eligible comparison pairs. All contexts/fits remain in denominators. Table 12 summarizes selected comparisons, while the full CSV includes every family and cohort. A large common bias can inflate absolute error yet leave the selector unchanged; a large offset share under shift therefore must not be read as successful modeling. The remaining all-weight error still increases materially. Conversely, eligible-action agreement counts do not fix switch locations: even the same pair of endpoints can transition at the wrong weight. The table reports mean-component shares with explicit denominators rather than imputing a causal explanation to the 58.7% quantile reduction.

Retrained missing-group controls: same labels and update budget

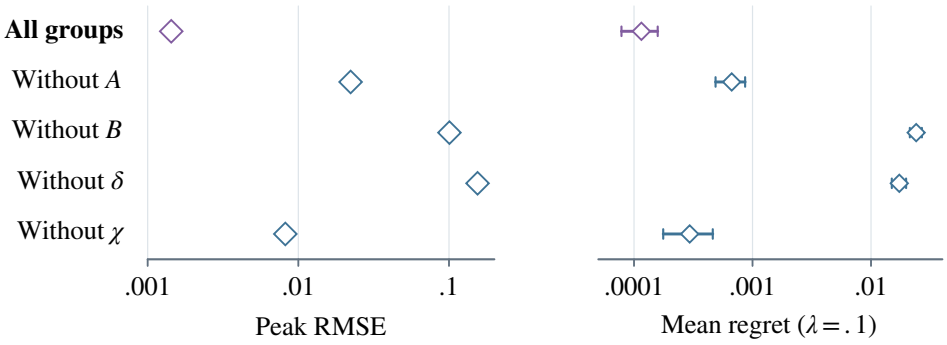


Figure 11: Retained actual retrained omissions, each with three fits on the same 16,000 labels and update budget. The candidate-duration input and full admissible set remain. Regret intervals resample contexts with fits fixed; deleting B or δ is particularly damaging in this sampled family, not universally most important in all mechanics.

Validation traces: longer optimization and a wider kernel neighborhood

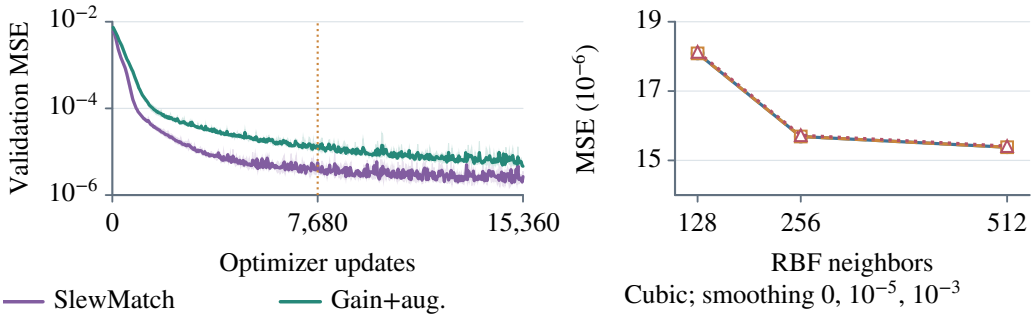


Figure 12: Bounded optimization checks. Left: complete validation traces of the six new runs, with min–max fit ranges (not confidence intervals); the dotted line marks 7,680 updates. Right: the nine cubic-RBF validation configurations. Continued improvement and a selected boundary neighborhood limit any convergence claim.

F.5 Calibrated width stability within an examined pool

Using the fixed full-group family, we pool the 256 V4 calibration and 512 ID score rows (768 examined contexts). RNG 9501 produces 200 permutations. In each, the first 64, 128 or 256 rows form nested calibration subsets; the last 256 rows are a disjoint test subset shared across those sizes. Every score uses the same subset and its own standard rank: 62, 123 or 245 at nominal 95%. Prefixes and repeated partitions are dependent. We retain membership indices and pool identifiers, rather than presenting resampled copies as extra observations.

The reduction $1 - q_G/(2q_U)$ is computed within each permutation before summarizing, not formed from the ratio of separately summarized quantiles. No primary model, score definition or reported V4 quantile is changed. The median reduction remains about 58% over this empirical pool; smaller calibration samples yield a wider range. The repeated-partition percentiles measure finite-pool sensitivity, not confidence limits for a new population or an externally repeated experiment. A fresh independent calibration-replication cohort was not generated.

Table 16: Finite empirical-pool calibration stability, 200 seeded repartitions. Reduction is $1 - q_G/(2q_U)$. P5/P95 summarize correlated repartition outcomes, not confidence limits; coverage concerns the corresponding disjoint 256-context holdout. The original primary calibration is unchanged.

n	Median q_G	Red. P5 (%)	Median (%)	P95 (%)	Cover. (%)
64	0.007075	50.1	58.8	67.8	95.5
128	0.006977	52.8	58.7	64.5	95.3
256	0.006977	54.7	58.1	62.3	95.3

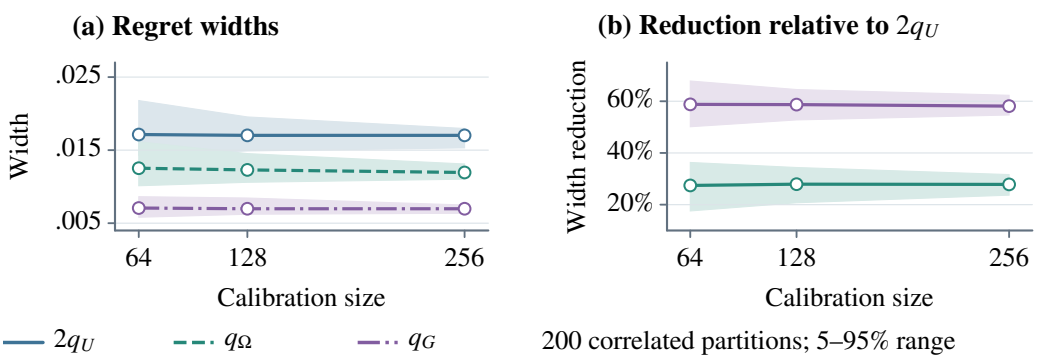


Figure 13: Retrospective calibration-membership sensitivity. Each of 200 fixed-pool permutations uses nested calibration sizes and a separate 256-context holdout. Curves show medians and bands the 5th–95th percentiles of correlated repartition results, not confidence intervals. The primary calibration set and all reported original test counts remain unchanged.

F.6 All-weight regret to the finer finite reference

The existing V4 ID fine-grid file contains 201 candidates per context, with the 21-point grid at every tenth index. The shared-target maximum discrepancy is 4.44×10^{-16}. For each unchanged coarse selector, the V5 audit keeps its 21 predicted-cost intervals but evaluates endpoint losses against all 201 true reference lines. It repeats this for the seven retained families, with all three fits where present. No ODE label or fine-grid calibration score is newly generated.

For each weight, $R_{21\to201}(\lambda) = R_{21}(\lambda) + \Delta_{\text{grid}}(\lambda)$. Taking suprema gives $G_{21} \le G_{21\to201} \le G_{21} + H_{\text{grid}}$ where $H_{\text{grid}} = \sup_{\lambda\ge0} \Delta_{\text{grid}}(\lambda)$, but equality of the upper bound is not guaranteed: the two summands may peak at different weights. The audit records all witnesses and checks this inequality. The common grid-envelope mean is .000801, P95 .004019 and maximum .008038. Neural G summaries first maximize over fits at each context; they are not means of the per-fit envelope statistics in Figure 7.

F.7 Shared switches, changed structures and narrow failure intervals

A switch is the boundary between two adjacent positive-length minimizing intervals. We match ordered pairs in the true and predicted active sequences; missing predicted counterparts retain unavailable displacement fields, not zero displacement. The exact regret algorithm separately retains singleton ties. Across all methods and cohorts the switch table has 54,944 rows. For full-group ID, 3,355 of 5,358 true transitions are matched; for intrinsic shift, 1,357 of 1,536 are matched. The median absolute shifts of those matched pairs are .002805 and .004848. The algebraic displacement identity is checked only where its shared-pair premise holds.

Figure 16 shows matched switches in a fixed $[0, .12]^2$ viewing window and labels the number outside; all rows remain in the exported data. The full active sets agree in 60.0% of ID fit–context pairs and

Table 17: Exact all-weight comparison to the finer finite reference, all 512 retained ID contexts. Values are multiplied by 10^3; neural envelopes maximize over the three fits before context summaries. The same test-only records supply both reference evaluations; these rows are not new nominal-95% calibrated guarantees.

Method	Mean G_{21}	G_{201}	P95 G_{21}	G_{201}	P99 G_{201}	Max
Raw	5.980	6.255	13.076	13.771	20.808	32.666
Gain	3.880	4.172	8.885	9.416	13.094	34.440
Raw + aug.	9.797	10.118	20.349	20.877	27.984	45.442
Gain + aug.	5.095	5.413	11.367	12.064	17.745	32.274
SLEWMATCH	2.945	3.288	6.484	8.027	10.527	21.104
RBF	4.448	4.776	14.338	16.094	23.194	27.395
8-NN	37.869	38.104	80.053	80.144	106.684	162.068

All nonnegative weights, with a finer finite reference—not a new certificate

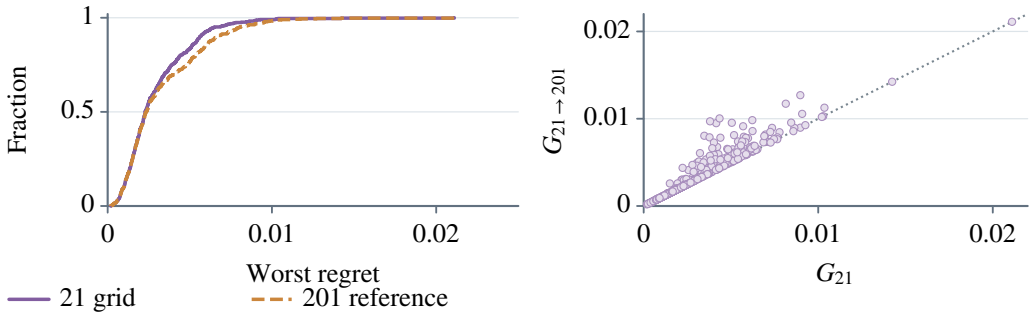


Figure 14: All-weight analysis of existing coarse and fine response records. Left: complete full-group family-envelope distributions against the 21 and 201 grids. Right: paired context envelopes, identity reference shown. All 512 contexts remain. This is a test-reference diagnostic, not a transported 95% fine-grid certificate or a continuous global optimum.

Mean regret $\times 10^3$: no omitted endpoint or unfavorable method

	ID			Intrinsic shift		
Raw	0.53	0.27	0.00	1.39	0.00	0.00
Gain	0.38	0.22	0.00	1.73	0.00	0.00
Raw + aug.	1.06	0.60	0.00	1.59	0.12	0.00
Gain + aug.	0.54	0.23	0.00	1.92	0.10	0.00
SlewMatch	**0.31**	**0.12**	**0.00**	**1.79**	**0.00**	**0.00**
RBF	1.22	0.51	0.00	3.53	8.20	0.65
8-NN	9.44	11.42	1.42	21.03	38.04	2.93
λ	.03	.10	.30	.03	.10	.30

Figure 15: Primary mean regrets at all three original weights, multiplied by 10^3. Pale cells and dark numerals preserve zero values and adverse method/regime comparisons. Both cohorts keep all 512 contexts. The complete three-regime table additionally records matched joint scaling.

Table 18: All fresh mean regrets ($\times 10^3$), including endpoint-zero and unfavorable outcomes. Each regime retains every one of its 512 contexts. These are not the already-examined V3 test contexts.

Regime	Method	$\lambda = .03$	$\lambda = .10$	$\lambda = .30$
ID	Raw	0.527	0.266	0.000
ID	Gain	0.381	0.217	0.000
ID	Raw + aug.	1.064	0.600	0.000
ID	Gain + aug.	0.537	0.225	0.000
ID	SLEWMATCH	0.311	0.115	0.000
ID	RBF	1.221	0.511	0.000
ID	8-NN	9.441	11.415	1.417
Joint $\times 2$	Raw	16.368	11.588	0.000
Joint $\times 2$	Gain	4.620	6.131	0.000
Joint $\times 2$	Raw + aug.	1.284	0.904	0.000
Joint $\times 2$	Gain + aug.	0.908	0.380	0.000
Joint $\times 2$	SLEWMATCH	0.311	0.115	0.000
Joint $\times 2$	RBF	1.221	0.511	0.000
Joint $\times 2$	8-NN	9.441	11.415	1.417
Intrinsic	Raw	1.393	0.000	0.000
Intrinsic	Gain	1.730	0.000	0.000
Intrinsic	Raw + aug.	1.591	0.123	0.000
Intrinsic	Gain + aug.	1.916	0.099	0.000
Intrinsic	SLEWMATCH	1.788	0.000	0.000
Intrinsic	RBF	3.528	8.202	0.646
Intrinsic	8-NN	21.028	38.036	2.930

88.3% of intrinsic pairs. This high intrinsic set agreement is compatible with displaced endpoint transitions and poor full-family coverage.

For the 512 intrinsic contexts, we additionally compute the union of intervals within the declared diagnostic window $[0, .5]$ on which at least one fixed fit has regret greater than the unchanged q_G. Pairwise true/predicted intervals and affine inequality intersections give their lengths without a weight grid. Median length is .00773 (including zero-length cases), and maximum is .03039. Neither length nor its ratio to the chosen window is a population probability over preferences. The median worst-weight family witness is .04847. The same sparse test covers 444 contexts while the full-family test covers 21: 423 contexts pass the sparse test but fail the full test. The original scan/endpoint outputs remain available and are not discarded as "easy samples."

F.8 REPORT USEFULNESS, FULL MEAN COMPARISONS AND MATCHED-SCALE EVENTS

An observable span-margin bound can be evaluated without online truth. In contrast, the G score itself requires offline labels and does not provide a context-conditional online error test. On all 1,536 ID fit–context pairs, the span-margin report certifies regret at most .001 in 1.4% at .1 and 99.5% at the easier .3 endpoint. The exact fractions and plausible-candidate counts remain unchanged in Table 19; a tighter global tail report is not automatically an informative operating-point certificate.

The original V4 tables retain endpoint-zero, unfavorable intrinsic and non-neural outcomes. Regret tails at a fixed weight and all-weight family-envelope tails are different summaries; the maximum of a mean curve is not the mean of contextwise maxima. Additional nominal levels and Wilson intervals remain in the raw certificate tables; finite test intervals do not establish conditional conformal validity for the realized calibration set.

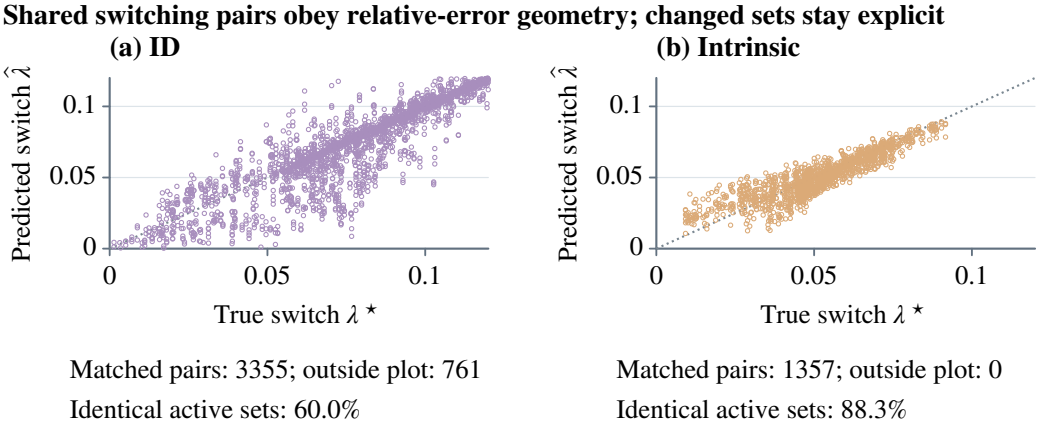


Figure 16: Exact matched-transition geometry on the existing V4 cohorts. Points match the same adjacent active pair in true and predicted envelopes; unmatched transitions are counted in the text rather than assigned zero shift. The fixed viewing window omits explicitly counted out-of-range points only from the display, never from the data. Identity means the correct boundary.

Table 19: Observable margin reports on all 1,536 ID context–fit pairs. "Near means an upper report of at most .001; "plausible counts actions with predicted excess cost at most the width. These do not observe the true online coverage event or change any action.

λ	Width	Exact (%)	Near (%)	Median plausible	Excluded (%)
0.03	$2q_U$	10.0	12.5	3	85.1
0.03	q_Ω	25.8	30.2	2	88.1
0.1	$2q_U$	0.0	0.0	5	66.5
0.1	q_Ω	1.0	1.4	4	74.5
0.3	$2q_U$	94.9	96.1	1	95.0
0.3	q_Ω	99.2	99.5	1	95.2

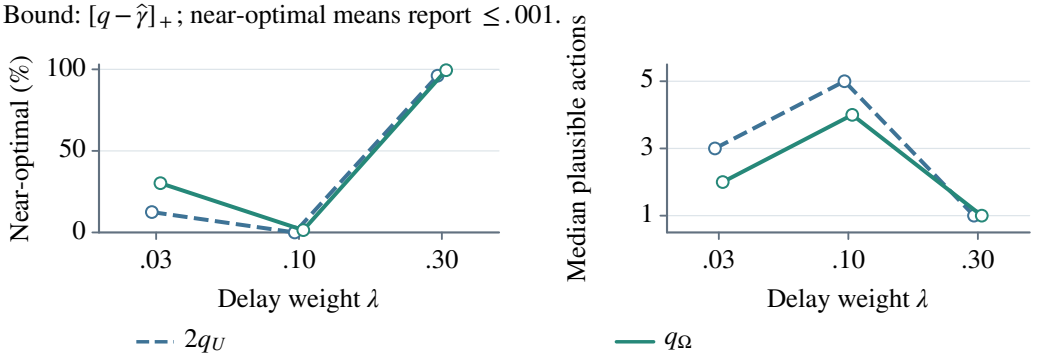


Figure 17: Retained operating-point margin reports, all ID fit–context pairs. A narrower comparison width excludes more candidate actions, but near-optimality certification is rare at the primary weight. At .3 the high fraction must be read together with its degenerate fastest-action oracle. Reports neither observe true online error nor change the selected action.

Table 20: Fresh calibration thresholds and score-coverage counts out of 512 at nominal 95%. Each family is calibrated separately. The first three counts concern ID and the last three intrinsic shift; the G score is itself the all-weight regret envelope.

Method	q_U	q_Ω	q_G	ID U	ID Ω	ID G	Shift U	Shift Ω	Shift G
Raw	0.01939	0.02078	0.01458	501	486	494	210	276	211
Gain	0.01369	0.01697	0.00921	504	504	492	168	292	196
Raw + aug.	0.02284	0.03386	0.02186	491	490	494	193	308	246
Gain + aug.	0.01513	0.02086	0.01143	493	495	486	159	276	149
SLEWMATCH	0.00857	0.01230	0.00707	493	492	490	18	58	21
RBF	0.02299	0.03351	0.01482	487	497	489	36	86	56
8-NN	0.07680	0.11731	0.08649	480	485	493	54	322	309

Table 21: True oracle structure for all 512 contexts. Near ties have a best–runner-up true cost gap at most .001 and may overlap any location stratum. Endpoints are never filtered out.

Regime	λ	Fastest	Slowest	Interior	Near tie	Median gap
ID	0.03	0	371	141	38	0.00767
ID	0.1	240	163	109	84	0.00359
ID	0.3	512	0	0	0	0.03087
Intrinsic	0.03	56	456	0	4	0.00575
Intrinsic	0.1	512	0	0	0	0.00984
Intrinsic	0.3	512	0	0	0	0.03902

Joint absolute-response coverage: exact structure versus learned approximation

	ID	**A2**	**T2**	**J2**	**J0.5**
Raw	501	372	174	158	118
Gain	504	464	151	135	210
Raw + aug.	491	486	418	444	246
Gain + aug.	493	479	456	440	408
SlewMatch	493	493	493	493	493
RBF	487	487	487	487	487
8-NN	480	480	480	480	480

Counts out of 512. Matched transforms are paired copies, not new evidence.

Figure 18: Retained absolute-response coverage counts, 512 paired contexts per scale and a separate radius per method. Analytic-input models preserve their coverage events exactly under the declared transformations; these are paired copies rather than independent confirmations. Different calibrated widths prevent treating coverage alone as an equal-width comparison.

G RETAINED NUMERICAL AND MODEL-MISMATCH BOUNDARIES

G.1 RETAINED NUMERICAL VALIDATION AND V4 INDEPENDENT CHECKS

The historical normalized solver is float64 RK4 with $\Delta\tau = .01$ and horizon 20. Its output includes sampled peak absolute displacement, peak rate, signed minimum and terminal displacement. The previously executed checks remain available: 64 independent physical-coordinate DOP853 comparisons have largest normalized-peak difference 9.59×10^{-7}; step halving changes the corresponding

sampled peaks by at most 2.91×10^{-6}; extending the historical 10,752 ID candidates to horizon 40 does not change their observed peaks. Those are retained numerical results, not first performed in V4.

The new lag variant adds $\ell z' = w_v - z$, $z(0) = 1$, and replaces the forcing by $g_\delta - z$. For positive ℓ, z is bounded between zero and one, so the existence argument extends to the three-state system. Its normalized step is .01; $\ell \geq .05$ in nonzero-lag tests. The independent physical implementation uses $T_a \dot{z} = w_D - z$ and the unnormalized mechanical equation, with DOP853 relative tolerance 10^{-11}, absolute tolerance 10^{-13}, and maximum step $.04\rho$. The first eight new ID contexts at each of $.05, .2, .4$ give 24 comparisons; RK4 and half-step RK4 add 48 targeted audit executions. The largest sampled-peak difference is 2.15×10^{-8} and largest step-halving change 3.81×10^{-7}.

Two further physical DOP853 integrations verify extended lag scaling when T_a scales with ρ, with normalized trace difference below 2.72×10^{-12}. Three DOP853 integrations generate Figure 1 from the first historical independent-response input; their collapsed traces differ by at most 1.83×10^{-11}. These three curves are explicitly an illustration. Thus there are 29 new DOP853 responses in these targeted checks/illustration, distinct from the stored study workload. A finite numerical agreement sample never establishes a uniform physical-peak or continuous-time error bound.

G.2 Keep candidate-grid error separate from learned error

The fresh 201-candidate audit has 102,912 integrations; the 21 grid is its every-tenth subset. The maximum shared-candidate response difference is 4.44×10^{-16}. For any selected coarse-grid action, total regret to the fine-grid oracle decomposes exactly into its coarse-grid learned regret plus the coarse-oracle–fine-oracle gap. This gap is nonnegative up to floating precision, depends on the weight/context, and is the same additive quantity for every coarse-grid policy on that context.

G.3 What changes near action switches

The fresh ID oracle uses an average of 4.49 positive-length optimal-action intervals; 226/512 contexts have at least one interior action on the full lower envelope. Every fresh intrinsic oracle instead uses only the two endpoint intervals. That does not make every intrinsic objective easy: a learned shift in the transition between endpoints can incur substantial regret between three sampled weights. This is why the exact objective-family score is informative even when interior optima disappear. The score targets errors in the declared affine decision family, not a claim that one particular hard-task distribution has been optimized.

G.4 Hidden lag and fixed, separately calibrated reports

Table 22: Paired hidden actuator lag, 512 ID contexts and frozen five-group models. Counts refer to U, Ω, G score events. Only the final row recalibrates at the prespecified $\ell = .2$, using 256 independent calibration contexts; it does not alter predictions or actions. Regret and RMSE are multiplied by 10^3.

ℓ	Calibration	q_G	U	Ω	G	Regret	RMSE
0.00	Base	0.00707	493	492	490	0.115	1.43
0.05	Base	0.00707	72	368	388	0.227	10.23
0.10	Base	0.00707	0	322	270	0.677	20.45
0.20	Base	0.00707	0	222	154	3.937	41.24
0.20	Matched	0.06023	491	500	499	3.937	41.24
0.40	Base	0.00707	0	124	76	18.975	82.80

For the base quantiles, increasing ℓ changes actual candidate outcomes while the declared context and all predicted candidates stay fixed. The response-score event therefore collapses earlier than some decision-score events. The independently calibrated .2 case was part of the recorded design, not chosen after searching for a successful recalibration. Its new labels come from the calibration contexts and are evaluated on distinct ID contexts; paired lag settings within either group are not treated as independent samples.

The original command duration and slew formulas remain valid for the commanded cubic withdrawal. They do not imply that the lagged actuator has completed withdrawal at that duration; its state can persist. A wider marginal simulator report cannot repair this mismatch or make the nominal deadline a physical tracking guarantee. Other unmodeled friction, relaxation laws and coupled dynamics are not evaluated, and the six-group lag observation is not presented as a learned identification result.

H HISTORICAL CONTROLS AND THE EVIDENCE THAT REMAINS UNFAVORABLE

The original V3 calibration and test records have already influenced the research questions. Their analyses are retained for traceability rather than pooled with new confirmation contexts. In particular, the historical SLEWMATCH radius is .0081738122, the old bound .0163476244, and primary ID regret .0001534566. The original response counts are 489/512 in distribution and 17/512 in intrinsic shift. These values differ from the new cohort's values because the contexts changed; no retraining of the original twelve networks explains that change.

The original full weight and coverage tables remain in the packaged historical source records. They are not pooled with the primary cohort; the adverse intrinsic comparisons, wrong-metadata outcomes and their distinct calibration radius remain explicitly identified here.

The original declared-delay experiment changes only the predictor's d by factors $1 + \beta$, $\beta \in \{-.5, -.25, 0, .25, .5\}$. True responses, candidate durations and known delay penalties stay fixed. The nonlinear omission similarly sets only predictor metadata k_3 to zero, rather than retraining or simulating a linear plant. The archived metadata-intervention arrays preserve these negative tests; Figure 11 separately answers the retrained missing-information question.

The earlier linear threshold and constructed-pulse datasets remain in the code archive with their original outputs. Their low-threshold comparison exposes an amplitude confound rather than demonstrated personalization; none of those rows enters either learned response task. They are not repeated as core evidence for the present diagnostic method.

I RUNTIME, CERTIFICATES, AND EXECUTABLE REPRODUCTION

I.1 THE DEPLOYED SELECTOR AND THE REPORT-ONLY WRAPPER

The original `ReleasePolicy` uses one stored NumPy checkpoint, constructs the nominal candidate set, predicts normalized peaks, and returns the smallest-index argmin of predicted peak plus known delay. The retained `CertifiedReleasePolicy` first calls that identical selector, verifies that its checkpoint belongs to the calibrated frozen family by SHA-256, and appends the three marginal reports plus the observable span-margin report. It explicitly marks true responses and per-context coverage as unavailable online. An infeasible result has no action and no appended action certificate. Unit tests verify identical selected indices/durations with and without the wrapper and preserve the original infeasibility and event-state semantics.

The optional historical event wrapper exposes assisting, candidate, yielding and manual states, freezes accepted releases, keeps explicit re-engagement, and records urgent cancellation as bypassing the normal ramp. Its physical acknowledgement remains false because there is no actuator. This software account is not a new evaluation of a full human–machine interface. No q threshold is used to authorize, cancel or replace an action.

I.2 LOCAL COMPUTATION AND THE LIMITS OF AMORTIZATION

The retained V4 measurements use an AMD EPYC 9V74 host, one numerical/ML thread, ten warm-up calls and 100 repetitions per method/count. At 21 candidates, median times are .0477 ms for the retained SLEWMATCH fit, .0345 ms for Gain+augmentation, 5.387 ms for the validated local RBF, and 2.029 ms for numerical enumeration. Larger candidate counts are timing workloads, not additional confirmed action sets covered by the 21-candidate certificate. RBF's local linear solves make it slower than enumeration at these counts, illustrating why predictive accuracy and query cost must be evaluated separately.

The extra timing-only fit is stored under a separate results directory, not among the 45 scientific neural checkpoints. It repeats old training/validation data solely to measure current-host cost; its outputs do not change the already frozen evaluation. Numerical timing also reuses declared inputs. These reevaluations are recorded outside the independent context and stored-response workload counts. Total deployment economics additionally depend on calibration simulations, model search, data acquisition, hardware and future query volume; the partial break-even table cannot establish a universal benefit or the necessity of a learned controller.

I.3 Contextual photograph and illustration provenance

Figure 1 uses *Driving car 20170523*, photographed by Santeri Viinamäki on 23 May 2017 and distributed through Wikimedia Commons under Creative Commons Attribution–ShareAlike 4.0 (**?**). The asset is cropped and resized for contextual explanation; it is not a photograph of our apparatus, participant or experiment, and it is not an input to the predictor. The adjacent rotational diagram is newly drawn. The photographic crop and composite Figure 1 are supplied under CC BY-SA 4.0; the license is `https://creativecommons.org/licenses/by-sa/4.0/`, and exact source/crop records accompany the asset. This attribution does not identify a depicted person as a study participant or claim endorsement.

The middle curves reuse the three independently integrated V4 DOP853 illustration trajectories, not new labels or measurements. Their declared factors $(a, s) = (1, 1), (2, 1), (.5, 2)$ are encoded by blue solid, orange dashed and purple dotted traces. The normalized trajectories coincide up to the retained numerical discrepancy. The right-hand switch example is the algebraic construction specified in Appendix A; solid costs are true, dashed costs predicted. Its colored wrong-choice interval is not a confidence band. Figure 2 uses calibration context zero and ID context zero for its response/cost exemplars; the online panel contains only emitted peaks and known-cost calculations. Layer-node counts are symbolic, with the real two-hidden-layer width 64 stated on the figure. All figure text is black at the final plotting size; the photograph remains a photographic raster embedded within an otherwise editable vector composition.

I.4 Files, commands and execution boundaries

The model package preserves all earlier scientific data and adds `results/v6`, `data/v6/capacity`, `checkpoints/v6/capacity`, and the independent audit-comparator module. Current release scans read existing reference/prediction arrays; new analytic labels and fitted models belong only to capacity. Source manifests and the chronology distinguish computation from re-rendering. Full response generation and historical timing files are not relabeled as new execution.

```
python reproduce_v6.py check
python reproduce_v6.py figures
python reproduce_v6.py analyze
```

Check validates a pristine manifest and executes tests. Figures regenerates current numeric publication plots/tables and the single MATLAB entry. Analyze repeats the fixed audit comparisons; capacity evaluation can be recomputed from saved selected models without retraining. The optional training route writes only to a working copy with an explicit overwrite flag. No route submits a paper or publishes data.

The single `SlewMatch_ICLR2027_V6_Figures(outputDir,figureNumbers)` renders Figures 3–18 in paper order using embedded numeric arrays and native MATLAB plotting. Figure 3 is explicitly algebraic; the remaining plots use executed response or analysis records. Editing arrays updates the plotted curves, labels, distributions and intervals, but does not retrain or recompute statistical estimates. Contextual Figures 1–2 retain editable Python/SVG/PDF sources and the licensed photo. Numeric read-back is checked independently; MATLAB/Octave was not available, and runtime or cross-platform pixel equivalence is not claimed. The separate LaTeX archive can re-render all publication vectors from frozen scene descriptions without importing any model. The supplied ICLR 2027 style is unchanged, and standalone font files are excluded.