When Selection Breaks Calibration

Constructive Collapse in Learned Boolean-Complexity Bounds

J. R. Landers

Abstract

A learned model can identify where a mathematical upper bound is loose while still failing precisely on the cases selected for the largest improvements. We isolate this phenomenon in a frozen prospective experiment on five-input Boolean formula complexity. A semantic model predicted the residual slack in a certified AND/OR/NOT construction bound; a held-out maximum residual was subtracted before any reduction was applied. At exact cost 14, six of 120 refined endpoints fell below the truth. All six belong to one exact boundary regime: the construction endpoint is 16, the true value is 15, and therefore only one gate is removable. Among ten matched targeted functions in this regime, the six largest model scores are exactly the six misses. We derive the boundary identity behind this correspondence and show how top-score selection enriches a rare low-headroom subgroup. Its prevalence rose from 1/46 in cost-13 high-score calibration cases to 9/78 in the cost-14 targeted high-score cohort. The functions are structurally distinct, but share a discordance: restriction nearly closes the bound while global decision-tree and algebraic features continue to signal difficulty. The case provides a finite, reproducible warning that calibration must match the complete selection policy, not merely the predictor's coarse score stratum.

1 A sharper failure than ordinary prediction error

Suppose analysis supplies a certified upper bound $Y(f) \le U(f)$ for an unknown integer-valued quantity $Y(f)$. A learned model predicts the *headroom*

$$h(f) = U(f) - Y(f) \ge 0,$$

and proposes to remove part of it. This is more demanding than ordinary prediction. Overpredicting h does not merely increase squared error; it can turn a valid mathematical bound into a false statement.

Our instance is minimum Boolean tree-formula complexity. For $f : \{0,1\}^5 \to \{0,1\}$, let $K(f)$ be the minimum number of gates in a formula over $\mathrm{AON} = \{\mathrm{AND}, \mathrm{OR}, \mathrm{NOT}\}$, and set $Y(f) = K(f) + 1$. The shift counts formula leaves under the usual tree convention and aligns with classical lower and constructive upper bounds [1]. The certified endpoint $U(f)$ is the best of a general analytic construction, exact four-variable restriction constructions, and disjoint-variable factorizations.

A histogram gradient-boosted regressor [2] maps 23 semantic and construction-aware truth-table features to a headroom score $z(f)$. A calibration sample supplies a one-sided correction q_g for each score stratum g. The proposed reduction is

$$r_\tau(f) = \begin{cases} \big(z(f) - q_{g(f)}\big)_+, & \big(z(f) - q_{g(f)}\big)_+ \ge \tau, \\ 0, & \text{otherwise}, \end{cases} \qquad U_{\mathrm{sel}}(f) = U(f) - r_\tau(f), \tag{1}$$

where $\tau = 0.25$ suppresses tiny reductions. On the held-out cost-13 sample, q_g was the largest dangerous residual $z - h$ in stratum g, enlarged by a cross-layer buffer. This resembles the basic logic of split calibration [3, 4], though here the maximum rather than a nominal quantile was used.

The frozen cost-14 experiment seemed favorable in aggregate. It reduced the construction endpoint on 99 of 120 functions and removed 0.8795 gate on average. Nevertheless, coverage was 95% rather than the required 99%. The six misses have a more exact explanation than "distribution shift."

2 The boundary identity and selection enrichment

Define the amount by which the refined endpoint falls below the truth,

$$v(f) = (Y(f) - U_{\text{sel}}(f))_+ .$$

Lemma 1 (Headroom boundary)**.** *For the policy in* (1)*,*

$$v(f) = (r_\tau(f) - h(f))_+ . \tag{2}$$

Whenever $z(f) - q_{g(f)} \geq \tau$ *and* $h(f) = m$*,*

$$v(f) = \big(z(f) - (q_{g(f)} + m)\big)_+ . \tag{3}$$

Thus an m*-headroom case fails exactly when* $z(f) > q_{g(f)} + m$*.*

Proof. Substituting $U_{\text{sel}} = U - r_\tau$ and $h = U - Y$ gives $Y - U_{\text{sel}} = r_\tau - h$, proving (2). In the active branch of (1), $r_\tau = z - q_g$; substituting $h = m$ proves (3). □

The lemma shows why small headroom is a boundary layer. A score may be quite reasonable for the population yet become unsafe as soon as it crosses one sharp threshold. Top-score selection makes this boundary more dangerous. Let $H_m = \{f : h(f) = m\}$ and let $S_t = \{f : z(f) \geq t\}$ be a score-based selection event.

Proposition 2 (Selection enrichment)**.** *For any population with* $\Pr(S_t) > 0$*,*

$$\Pr(H_m \mid S_t) = \frac{\Pr(H_m)\Pr(S_t \mid H_m)}{\Pr(S_t)}. \tag{4}$$

Consequently a rare low-headroom class can be enriched arbitrarily strongly by top-score selection whenever its upper score tail is heavier than the population tail. Marginal validity within a coarse score stratum does not imply validity conditional on S_t *unless the calibration error is stable under that selection.*

The identity is Bayes' rule. Its implication is the important part: selecting instances because they promise the largest bound reductions is not an innocent use of a marginally calibrated model. It changes the mixture of headroom types inside the evaluation set.

3 An exact ten-function boundary layer

The high-stratum cost-14 correction was

$$q_{\text{high}} = 2.360706.$$

Every miss had $Y = 15$ and $U = 16$, hence $h = 1$. Lemma 1 therefore places the failure threshold at

$$z > q_{\text{high}} + 1 = 3.360706. \tag{6}$$

Exactly ten targeted functions had $U = 16$. Table 1 lists all ten. The first six scores exceed (6), and all six fail. The remaining four scores lie below the boundary, and all four are covered. The largest score produces the largest violation automatically: $3.761361 - 3.360706 = 0.400656$.

Table 1: The matched one-gate-headroom group, sorted by frozen model score. The cofactor column gives the exact AON gate costs of the unique best pair of four-input restrictions.

Truth table	Covered	Ones	z	r	v	Leaves	Cofactors
`58760a6e`	no	15	3.7614	1.4007	.4007	15	$4+7$
`b0aae8c0`	no	13	3.6240	1.2633	.2633	14	$5+6$
`f15c00a8`	no	12	3.5883	1.2276	.2276	13	$7+4$
`05e402e8`	no	11	3.5862	1.2255	.2255	13	$8+3$
`0a172737`	no	15	3.5271	1.1664	.1664	12	$5+6$
`cdd46774`	no	18	3.3803	1.0196	.0196	13	$5+6$
`7676ec28`	yes	17	3.3238	.9631	0	12	$4+7$
`f8ba1102`	yes	13	3.3210	.9603	0	13	$8+3$
`fb8aba32`	yes	18	3.2809	.9202	0	13	$6+5$
`21252111`	yes	9	2.9984	.8854	0	12	$7+4$

This perfect score ordering could still be a duplicated-orbit artifact. It is not. The ten functions occupy ten distinct NPN equivalence classes under input permutations, input negations, and output negation. Each has one unique best restriction variable. After ignoring cofactor order, their optimal four-input cost pairs span $3+8$, $4+7$, and $5+6$; covered and uncovered cases share these patterns. None is canalizing, and none has a nontrivial disjoint-variable AND or OR factorization. There is no single exceptional normal form hiding behind the six rows.

The commonality is instead a relationship between two kinds of structure. Restriction has already produced an endpoint just one gate above exact cost, but the remaining global semantics still appear difficult. In a local one-coordinate perturbation against the four covered controls, the largest upward-score contrasts are decision-tree leaves (0.2104), linear ANF support (0.1694), quadratic ANF support (0.0938), and degree-two Fourier signs (0.0566). Restriction gain has contrast only 0.0203. These values are descriptive, not causal, because the coordinates are correlated. They do, however, motivate a name for the pattern: *construction–semantics discordance*. A specialized construction has nearly closed the bound while global difficulty features continue to predict room for improvement.

4 Why calibration did not see the boundary

The cost-13 calibration sample contained only one direct analogue among its 46 targeted high-stratum functions: a one-headroom case with score 3.099887. Its dangerous residual was $3.099887 - 1 = 2.099887$. Adding the previous cross-layer buffer 0.260819 produced the frozen correction 2.360706.

At cost 14, nine of 78 targeted high-stratum cases had one gate of headroom. The relevant prevalence therefore changed from

$$\frac{1}{46} = 2.17\% \qquad \text{to} \qquad \frac{9}{78} = 11.54\%,$$

a 5.31-fold enrichment. More importantly, their scores extended to 3.761361, so the maximum dangerous residual became 2.761361. The shortfall $2.761361 - 2.360706 = 0.400656$ is exactly the worst violation.

The model did retain useful information. All 80 targeted functions were tightened, compared with 19 of 40 random references, and their mean reduction was 1.1744 rather than 0.2895. The ranking found opportunity. The failure is that calibration pooled unlike headroom regimes and then selection queried the most aggressive tail of that pool.

This distinction matters for mathematical learning systems. A safety margin calibrated on individual predictions need not survive a downstream policy that searches many candidates and reports only the largest proposed gains. The calibration unit should be the complete procedure—candidate generation, scoring, top-k selection, and refinement—or it should condition on enough construction information to make residual error stable under selection.

5 Implications

The case suggests two changes, neither of which can be validated on these same six functions. First, the model should expose the complete restriction witness profile: both cofactor costs, their imbalance, the number of optimal restriction variables, and the gap to the second-best restriction. A scalar restriction gain did not capture how completely restriction had already explained the function.

Second, the learned target should match the one-sided risk. Squared-error prediction of headroom rewards average accuracy. A direct model of the dangerous residual $z - h$, a one-sided quantile objective, or calibration that replays top-k selection would focus on the event that invalidates the bound. Any new rule derived here is post-hoc. It must be frozen before a new sample is labeled.

The six misses therefore form an interesting mathematical case not because they share one Boolean normal form, but because they instantiate an exact failure geometry. Constructive collapse places them one gate from the truth; semantic persistence assigns large scores; selection concentrates the upper tail; and the boundary identity converts that tail into six deterministic violations. This is a small finite example with a general lesson: when learning is used to sharpen mathematics, calibration must follow the query that will actually be made.

Reproducibility. The frozen predictions, model and protocol hashes, NPN audit, restriction witnesses, matched casebook, perturbation table, and support-shift calculation are stored with the companion experiment. The structural report is generated without refitting the model or changing the prospective result.

References

[1] S. Jukna, *Boolean Function Complexity: Advances and Frontiers*. Springer, 2012.

[2] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, 29(5):1189–1232, 2001.

[3] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.

[4] J. Lei, M. G'Sell, A. Rinaldo, R. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *J. Amer. Stat. Assoc.*, 113(523):1094–1111, 2018.