Prediction-Calibrated Refinement of Analytic Bounds with an Application to Boolean Formula Complexity

J. R. Landers

Abstract

Many difficult optimization problems have rigorous but loose analytic bounds and exact solutions only for small instances. We combine these two sources of information: a model predicts where the answer lies inside the analytic range, held-out exact cases calibrate its error, and the resulting interval is intersected with the proved bounds. This never weakens the analytic envelope and retains the calibration guarantee. We also prove that normalizing errors by the analytic gap gives a deterministic, per-instance lower bound on how much that gap is reduced.

We apply the framework to minimum Boolean formula size over NAND, NOR, and AND/OR/NOT gates. Across all 65,536 four-input functions, nominal 95% intervals cover 96.3%, 96.6%, and 96.4% while removing 82.0%, 79.2%, and 71.5% of the classical envelope on average. Gap-normalized calibration guarantees at least 81.4%, 79.0%, and 70.8% shrinkage for every nontrivial held-out envelope. Maximum-residual intervals cover the complete finite universe while still removing roughly half its width. A frozen five-input test retains strong NAND and NOR performance but exposes an AND/OR/NOT upper- tail weakness. The framework therefore turns useful predictions into tighter bounds while keeping statistical, finite-domain, and universal claims clearly separated.

1 Introduction

Analytic bounds and learned predictions answer different questions. A proof may certify that an unknown quantity lies in a broad interval. A predictor may locate it accurately without supplying any certificate. We ask how these objects can be combined without weakening the proof or overstating the prediction.

Our approach has four steps. Theory first brackets the answer. A model then predicts where the answer lies inside that bracket. Held-out examples measure how far the predictions can miss. Finally, we widen around the prediction by those measured errors and intersect with the original theoretical bracket. Thus theory supplies validity, while learning supplies localization.

The construction is simple: bracket, predict, calibrate, and intersect. Its value is that these four steps keep their meanings. The analytic envelope remains the guaranteed outer constraint; predictor quality controls how much the interval can shrink; and the calibration protocol determines whether the refined interval has statistical, finite-domain, or universal validity.

Boolean formula complexity provides a demanding case study. For a Boolean function f, $K_{\mathcal{L}}(f)$ is the smallest number of gates in a tree formula over library $\mathcal{L}$ that computes f. Exact synthesis finds K only in small dimensions. Classical theory gives dimension-independent lower bounds and explicit constructive upper bounds, but often leaves a wide gap. We use exact small instances to learn and calibrate a location inside that gap. The prediction $\widehat{K}_{\mathcal{L}}(f)$ is not a formula and is not itself a bound; it becomes useful only through calibrated error and intersection with the analytic envelope.

Section 2 states the general refinement principle and separates its three guarantee regimes. Section 3 instantiates the envelope, predictor, and calibration for Boolean formulas. Sections 4 and 5 give a complete four-input experiment and a frozen five-input stress test. Sections 6 and 7 discuss the route from statistical refinement to certified mathematical improvement.

Contributions. First, we state a general prediction-calibrated operator for refining any valid analytic envelope and show that intersection preserves the available error guarantee while never increasing width. We distinguish statistical, exhaustive finite-domain, and universal versions of that guarantee. Second, feeding the Boolean normalization back into the general framework, we prove a gap-normalized refinement theorem: calibrated coverage is accompanied by a pointwise fractional-shrinkage certificate. Third, we instantiate the framework for exact Boolean formula size using semantic lower bounds, constructive upper bounds, and a learned center. Over every four-input function the learned interval is much narrower than both the analytic envelope and a calibrated midpoint baseline; a maximum-error version covers the entire finite universe. Finally, a frozen five-input test preserves strong NAND/NOR shrinkage and exposes a specific AON upper-tail failure.

Exact synthesis has long been used to classify small functions and benchmark logic representations [4, 5]. Formula lower bounds, including Khrapchenko's boundary method, pursue dimension-independent obstructions [1, 2]. Conformal

prediction supplies finite-sample marginal coverage under exchangeability [6, 7]. The present work connects these strands through a reusable abstraction: theory supplies an outer envelope, prediction proposes a center, and calibration supplies an explicitly qualified refinement. Normalized nonconformity scores are standard tools for locally adaptive conformal regression [8, 9], and conformal-set size has recently received direct finite-sample analysis [10]. To our knowledge, the analytic gap has not previously been used as the normalization to obtain both conformal coverage and a deterministic, pointwise certificate on the fraction of a proved bound removed.

2 Calibrated refinement of analytic envelopes

Let x be an instance, $Y(x)$ an unknown real-valued target, and suppose analysis supplies a valid envelope

$$I_0(x) = [L(x), H(x)], \qquad L(x) \le Y(x) \le H(x). \tag{1}$$

Let $\widehat{Y}(x)$ be any point predictor. Calibration supplies nonnegative one-sided allowances $E^-(x)$ and $E^+(x)$, producing the prediction interval $C(x) = [\widehat{Y}(x) - E^-(x), \widehat{Y}(x) + E^+(x)]$. We define the refined envelope by the intersection

$$I_*(x) = I_0(x) \cap C(x) = \left[\max\{L, \widehat{Y} - E^-\}, \min\{H, \widehat{Y} + E^+\}\right]. \tag{2}$$

If the displayed lower endpoint exceeds the upper endpoint, $I_*(x)$ is the empty set and its width is defined as zero.

Proposition 1 (Envelope refinement)**.** *For every x, $I_*(x) \subseteq I_0(x)$ and therefore* $\mathrm{width}(I_*(x)) \le \mathrm{width}(I_0(x))$. *Moreover, if I_0 always contains Y and C contains Y with probability at least $1-\alpha$, then I_* contains Y with probability at least $1-\alpha$. The same statement holds with "probability" replaced by validity on a fixed finite domain or validity for every instance.*

Proof. The width statement follows from intersection. Because $Y(x) \in I_0(x)$, the events $Y(x) \in I_*(x)$ and $Y(x) \in C(x)$ coincide. The probabilistic and deterministic claims follow immediately. □

The Boolean application predicts normalized position inside the analytic gap. Applying that normalization to calibration as well yields a stronger general statement. Write $W(x) = H(x) - L(x)$. For $W(x) > 0$, assume without loss of generality that $\widehat{Y}(x) \in [L(x), H(x)]$ (clipping can only improve absolute error) and define

$$p(x) = \frac{\widehat{Y}(x) - L(x)}{W(x)}, \qquad S^-(x) = \frac{(\widehat{Y}(x) - Y(x))_+}{W(x)}, \qquad S^+(x) = \frac{(Y(x) - \widehat{Y}(x))_+}{W(x)}.$$

Theorem 2 (Certified fractional tightening)**.** *Let $q^-, q^+ \ge 0$ be any bounds or calibrated quantiles for the normalized one-sided scores, and set*

$$I_q(x) = \left[\max\{L, \widehat{Y} - q^- W\}, \min\{H, \widehat{Y} + q^+ W\}\right]. \tag{3}$$

For every nondegenerate analytic envelope,

$$\frac{\mathrm{width}(I_q(x))}{W(x)} = \min\{p(x), q^-\} + \min\{1 - p(x), q^+\} \le \min\{1, q^- + q^+\}. \tag{4}$$

Consequently every instance is tightened by at least $\max\{0, 1 - q^- - q^+\}$ *of its original analytic width. If q^- and q^+ are split-conformal quantiles at tail errors α^- and α^+, then I_q also covers Y with probability at least $1 - \alpha^- - \alpha^+$. Exhaustive or universal score bounds give the corresponding deterministic coverage statement.*

Proof. The distance from $\widehat{Y}$ to the left endpoint of I_q is $\min\{pW, q^- W\}$; the distance to its right endpoint is $\min\{(1-p)W, q^+ W\}$. Their sum proves (4), and each minimum is at most its second argument. Coverage follows by applying the two one-sided score guarantees and a union bound, then Proposition 1. □

The theorem supplies a property absent from a generic conformal interval: efficiency is certified before seeing the test label. Whenever $q^- + q^+ < 1$, no nondegenerate test envelope can remain unchanged, even on a miss. The score normalization is also dimensionless, so it states the calibration tax in units of the proved uncertainty already present in each instance. It does not remove the exchangeability requirement for statistical coverage.

The proposition separates tightness from validity. A more accurate predictor usually needs smaller allowances and hence produces a narrower refinement. The source of the allowances determines the claim. Held-out quantiles give a *statistical* interval under their sampling assumption. The largest error over an exhaustively enumerated domain gives a

finite-domain certificate. A proved error bound that holds for every instance gives a *universal* refined bound. The intersection operator is identical in all three cases; only the status of E^- and E^+ changes.

For exact calibration labels, define one-sided residuals

$$R^-(x) = (\widehat{Y}(x) - Y(x))_+, \qquad R^+(x) = (Y(x) - \widehat{Y}(x))_+, \tag{5}$$

where $(z)_+ = \max\{z, 0\}$. In grouped split conformal calibration, instances with the same observable profile form one class, and the class score is its largest member residual. With m calibration classes and total error rate α, E^- and E^+ are the $\lceil (m+1)(1-\alpha/2) \rceil$-th ordered class scores, using an unbounded allowance if that rank exceeds m. If calibration and test classes are exchangeable, meaning their ordering carries no information about their errors, a union bound over the two tails gives coverage at least $1 - \alpha$ for every member of the test class [6, 7].

3 Boolean formula complexity instantiation

3.1 Exact cost and analytic envelope

For $f : \{0,1\}^n \to \{0,1\}$ and gate library $\mathcal{L}$, define

$$K_{\mathcal{L}}(f) = \min\{|e| : \mathrm{eval}_{\mathcal{L}}(e) = f\}.$$

Here e ranges over tree formulas, $|e|$ counts gates, variables are free leaves, fanout is one, and constants are not free. We study NAND, NOR, and AON $= \{\mathrm{AND}, \mathrm{OR}, \mathrm{NOT}\}$.

Write $Z = f^{-1}(0)$ and $O = f^{-1}(1)$ for the zero and one inputs. Two inputs are neighbors when they differ in one bit, and E_{01} is the set of neighboring pairs on which f changes value. The Khrapchenko quantity is

$$B(f) = \frac{|E_{01}|^2}{|Z||O|} = \bar{s}_0(f)\bar{s}_1(f), \tag{6}$$

where $\bar{s}_0$ and $\bar{s}_1$ are the average numbers of value-changing neighbors seen from Z and O; set $B = 0$ for constants. Khrapchenko's theorem lower-bounds formula leaves by B. Moving NOT gates to the inputs does not change the leaf count, so all three libraries satisfy

$$B(f) \le K_{\mathcal{L}}(f) + 1. \tag{7}$$

Let $U(f)$ be the smallest number of leaves in a deterministic decision tree for f. Recursively converting the tree to a formula gives

$$\begin{aligned} K_{\mathrm{NAND}}(f) + 1,\ K_{\mathrm{NOR}}(f) + 1 &\le 6 + 9(U(f) - 1), \\ K_{\mathrm{AON}}(f) + 1 &\le 3 + 6(U(f) - 1). \end{aligned} \tag{8}$$

For AON, $Q_{\min}(f)$ is the smallest gate count among the prime disjunctive and conjunctive normal-form constructions we enumerate, including variants with one outer NOT. It is a valid constructive upper bound. Finding the least such cover is weighted set cover [3], and every returned cover is an explicit formula witness.

Exhaustive synthesis tightens the coefficients on the four-input universe:

$$\begin{aligned} B(f) &\le K_{\mathrm{NAND/NOR}}(f) + 1 \le 6 + \tfrac{9}{4}(U(f) - 1), \\ B(f) &\le K_{\mathrm{AON}}(f) + 1 \le \min\{3 + \tfrac{13}{9}(U(f) - 1), Q_{\min}(f)\}. \end{aligned} \tag{9}$$

These coefficients are finite-domain facts, not arbitrary-dimension claims.

3.2 Learned center and calibration

Set $Y_{\mathcal{L}}(f) = K_{\mathcal{L}}(f) + 1$ because the lower bound naturally counts leaves. Let $[L_{\mathcal{L}}(f), H_{\mathcal{L}}(f)]$ be the envelope from (7) and either (8) or (9). We predict the normalized location

$$\rho_{\mathcal{L}}(f) = \frac{Y_{\mathcal{L}}(f) - L_{\mathcal{L}}(f)}{H_{\mathcal{L}}(f) - L_{\mathcal{L}}(f)}, \qquad \widehat{Y} = L + \widehat{\rho}(\Phi(f))(H - L), \qquad \widehat{K}_{\mathcal{L}} = \widehat{Y} - 1. \tag{10}$$

If $L = H$, the answer is known and we set $\rho = 1/2$. Thus $\widehat{K}_{\mathcal{L}}(f)$ is the predicted minimum gate count, not the cost of a constructed formula and not a proved bound.

The 16-entry feature vector $\Phi(f)$ contains (B, U), four summaries of explicit constructions, counts of monomials by degree in the algebraic normal form (the unique polynomial over the two-element field), and counts of negative Fourier–Walsh coefficients by degree. A gradient-boosted tree predicts ρ, clipped to $[0, 1]$. For transfer beyond four inputs, degrees are grouped into bins $0, 1, 2, 3$, and $4+$.

Functions with the same Φ form one calibration class. We apply the grouped residual construction from Section 2, using the largest error in each class. Five rotating train/calibrate/test splits provide the reported 95% coverage. Separately, setting $E^{\pm}$ to the largest cross-fitted residual on all 65,536 four-input functions gives an exhaustively verified finite-domain instance of Proposition 1. It makes no claim about dimension five.

4 Complete four-input experiment

We synthesize exact minima for every four-input function in all three bases. Among the 65,536 truth tables there are 3,952 distinct feature vectors Φ. We divide these profile classes into five folds, keeping functions with identical features together. In each rotation, three folds fit the model, one sets the error allowances, and one tests the resulting intervals. The stated 95% target assigns 2.5% error to each tail. "Coverage" is the fraction of held-out exact values inside the interval. "Shrinkage" is the fraction of the original interval width removed. Candidate counts round the endpoints inward and count the remaining integer values that $K + 1$ could take.

basis	function cov.	profile cov.	shrink	candidates	full shrink
NAND	.963	.951	.820	$18.35 \to 3.23$	.547
NOR	.966	.953	.792	$18.35 \to 3.74$	.571
AON	.964	.952	.715	$9.15 \to 2.48$	.479

Table 1: Four-input results. Coverage and shrinkage use 95% profile-grouped calibration. "Candidates" is the mean number of possible integer values for $K + 1$, before and after calibration around $\widehat{K}$. "Full shrink" uses the largest observed errors and covers the entire four-input universe.

Table 1 gives the central result. Calibration around $\widehat{K}$ reduces roughly 9–18 possible gate counts to 2–4, while held-out coverage is near the stated 95% level. Function coverage is slightly higher than profile coverage because every profile is scored by its worst member. If we instead use the largest cross-fitted error anywhere in the complete data set, all 65,536 functions are covered and the intervals still lose roughly half their original width.

We then rerun the identical held-out protocol with the normalized scores in Theorem 2. Table 2 reports the result. The guarantee column is not an average: it is the smallest fractional tightening certified for every noncollapsed test envelope across all five rotations, computed before observing its held-out target.

basis	function cov.	profile cov.	mean shrink	guaranteed shrink
NAND	.961	.950	.824	.814
NOR	.962	.951	.798	.790
AON	.962	.954	.729	.708

Table 2: Gap-normalized 95% calibration. The final column is the pointwise lower bound $1 - \max_{\text{folds}}(q^- + q^+)$ from Theorem 2; coverage remains held-out and statistical.

Thus the Boolean experiment feeds back a general lesson stronger than its mean shrinkage numbers. At nominal 95% coverage, the calibrated quantiles certify in advance that every nontrivial analytic envelope in these held-out folds loses at least 81.4%, 79.0%, and 70.8% of its width for NAND, NOR, and AON, respectively. The realized mean improvements are only slightly larger, so the useful tightening is not carried by a small subset of easy functions.

Figure 1 shows where that tightening occurs. After sorting functions by exact complexity, the broad analytic sandwich remains visible in gray while the calibrated band in gold tracks the exact black curve. The reduction persists across the complexity distribution rather than coming only from its easiest region. AON begins with the tightest analytic upper construction, which also explains its smaller fractional improvement.

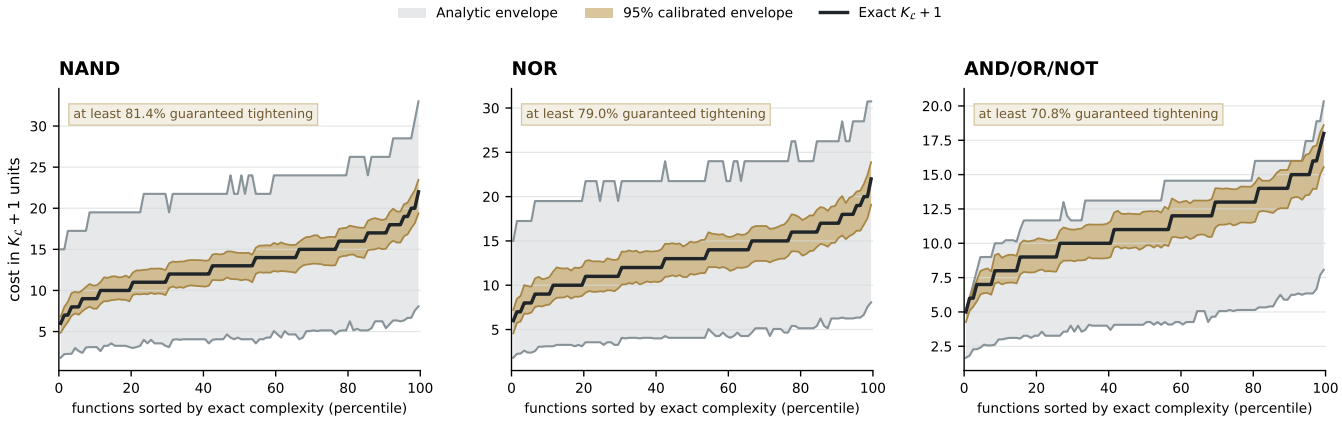

Figure 1: The prediction-calibrated complexity sandwich. Curves are medians within 100 equally populated bins after sorting by exact $K_{\mathcal{L}}+1$. Gray is the classical analytic envelope, gold is the gap-normalized 95% held-out envelope, and black is exact complexity. Annotations give the pointwise tightening certified by Theorem 2; they are not average reductions.

Does learning matter? A calibrated interval around any center can shrink a loose outer bound. We therefore repeat the complete protocol with a fixed midpoint and with feature families removed. Table 3 reports the fraction of integral candidates removed at comparable realized coverage.

center/features	NAND	NOR	AON
fixed midpoint	.612	.612	.547
theory features B, U	.643	.639	.615
construction features	.715	.717	.582
ANF + Fourier features	.737	.697	.619
all non-endpoint features	.799	.777	.650
all 16 features	**.820**	**.792**	**.717**

Table 3: Mean reduction in feasible integer candidates at 95% coverage. Each non-midpoint row trains a separate predictor using only the named feature family; "all non-endpoint" omits B and U. The 16-feature model improves materially over the fixed center of the theoretical interval.

The full model leaves 3.23, 3.74, and 2.48 candidates on average, compared with 6.96, 6.97, and 4.03 for the midpoint. Thus the improvement is not merely a consequence of calibrating any center inside a broad interval. Explicit construction features and the ANF/Fourier features provide information beyond the two theoretical endpoints (B, U).

Figure 2 summarizes the same gain in the discrete units used by the synthesis problem. Calibration reduces the mean number of feasible integer values of $K+1$ from roughly 9–18 to 2–4 across the three bases.

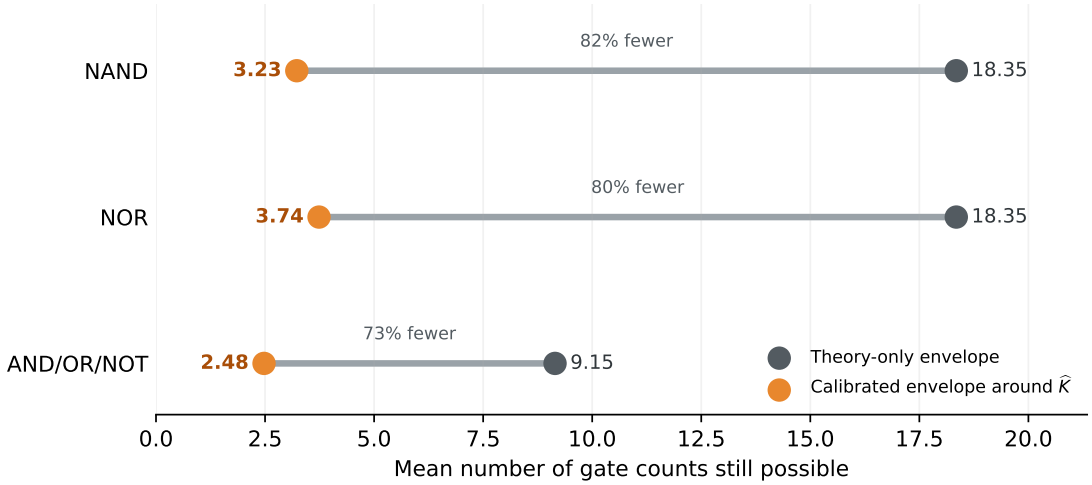

Figure 2: Mean number of integer gate counts allowed by theory alone and after 95% calibration around $\widehat{K}$. The percentage is the reduction in remaining candidates. Intersecting with the theory-only interval preserves all established lower and upper bounds.

5 Frozen five-input stress test

Exhaustively synthesizing all 2^{32} five-input truth tables is impractical with the four-input method. Instead, sparse dynamic programming enumerates every function that can first be computed using at most 11 gates. "First reached at cost 11" certifies that no formula with fewer gates computes that function. This gives exact cost for 4,434,723 NAND functions, the same number of NOR functions, and 15,631,538 AON functions. From each exact-cost layer we select at most 40 functions by a fixed seed. The resulting distribution is balanced by formula cost and is not uniform over truth tables.

We refit the predictor on four-input data, now using the general theoretical endpoints (8) rather than the four-input-only coefficients (9). The AON upper endpoint still uses $Q_{\min}$ when smaller. Initial checks showed that error size changes with both input dimension and formula cost, so a single fixed error allowance would be inappropriate. Before enumerating cost 11, we therefore freeze two models that predict lower- and upper-side error size from $\widehat{Y}$, the original interval width, (B, U), construction features, and ANF/Fourier counts. Samples through cost 9 fit these error models. Cost 10 sets one final maximum-residual correction. Cost 11 is then evaluated once, without revising the method. The frozen protocol and its SHA-256 digest are checked in with the experimental artifacts.

basis	train	cal.	test	covered	mean width	shrink
NAND	340	40	40	40/40	4.38	.947
NOR	340	40	40	40/40	5.09	.938
AON	350	40	40	37/40	4.06	.719

Table 4: Prospective cost-11 conditional-envelope results. The corresponding mean classical widths are 83.08, 81.25, and 14.57.

Table 4 asks whether the same story survives a harder, untouched test: does $\widehat{K}$ still locate exact size, and do the previously calibrated errors remain large enough? This is evidence of transfer, not a coverage theorem. Costs 10 and 11 are different distributions, and a sample of 40 functions per library has substantial uncertainty. NAND and NOR retain more than 93% shrinkage with no sampled miss. AON starts from a much tighter theoretical interval, but three exact sizes fall above the learned upper endpoint. Those failures suggest that the current features miss some recursive or factorized AON constructions.

6 Scope and next tests

Proposition 1 is an elementary intersection principle, not a new complexity lower bound. Theorem 2 adds a nontrivial efficiency certificate, but likewise does not strengthen either analytic endpoint. Their general value is organizational: they identifies exactly which component supplies outer validity, localization, and error control. The framework can be used whenever a target has a valid analytic envelope and representative exact or certified calibration labels. It is useful only to the extent that the predictor reduces residual error and the chosen calibration population matches the intended test population.

The complete four-input result is the strongest claim: every target size is exact, functions with identical features stay together, and the maximum-error interval is verified over the entire finite universe. The 95% interval is narrower but makes the exchangeability assumption stated in Section 2. In neither case does $\widehat{K}$ become a theorem by itself.

The five-input experiment tests a harder question: whether the center and its error geometry transfer across dimension. Its layer-balanced sampling is useful for controlled extrapolation but differs sharply from a random Boolean function, a random semantic profile, or a benchmark distribution. Larger tests must freeze both the sampling measure and the conditional calibration before exact synthesis. In particular, the next untouched experiment should oversample the AON upper tail and record prime-cover slack, factorability, and restriction decompositions without revising the cost-11 result.

There are two routes from a useful prediction to a proof. For an upper bound, a predictor can choose among decision trees, covers, restrictions, and factorizations; once it outputs an actual formula, its gate count is certified regardless of prediction error. A new lower bound is harder: it requires either exhaustive finite verification or a new inequality that holds for all formulas. This asymmetry makes prediction-guided construction the clearest next route to a dimension-general certified improvement.

7 Conclusion

A valid analytic envelope can be refined without confusing prediction with proof. Intersecting it with a calibrated interval around $\widehat{Y}$ never widens the envelope and inherits whatever statistical, finite-domain, or universal error guarantee supports the calibrated interval. Predictor quality determines tightness; calibration determines the strength of the claim. Normalizing residuals by the analytic gap adds a simultaneous pointwise certificate: every nontrivial interval loses at least $\max\{0, 1 - q^- - q^+\}$ of its proved outer width.

Boolean formula complexity demonstrates the distinction. $\widehat{K}$ locates exact size inside a theory-supplied range. Over all four-input functions, calibration removes roughly three quarters of the uncertainty at 95% coverage and roughly half under a 100% finite certificate. A frozen five-input experiment preserves strong NAND/NOR shrinkage and reveals a specific AON upper-tail weakness. Thus the general method provides a clean bridge from analytic bounds to useful predictions, while leaving the source and limits of validity visible.

Reproducibility. All exact targets, folds, predictions, calibration tables, protocol hashes, figures, and scripts are included in the accompanying repository. Automated tests verify the classical inequalities, finite learned intervals, generic feature implementation, exact sparse layer separation, and held-out cost boundaries.

References

[1] V. M. Khrapchenko, "Method of determining lower bounds for the complexity of Π-schemes," *Math. Notes*, 10(1):474–479, 1971.

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

[3] E. Allender, L. Hellerstein, P. McCabe, T. Pitassi, and M. Saks, "Minimizing DNF formulas and AC^0 circuits given a truth table," in *Proc. IEEE CCC*, 2006.

[4] W. Haaswijk, E. Testa, M. Soeken, and G. De Micheli, "Classifying functions with exact synthesis," in *Proc. IEEE ISMVL*, 2017.

[5] M. S. Turan and R. C. Peralta, "The multiplicative complexity of Boolean functions on four and five variables," in *LIGHTSEC*, 2015.

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

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

[8] H. Boström and U. Johansson, "Mondrian conformal regressors," in *Proc. COPA*, PMLR 128:114–133, 2020.

[9] N. Seedat, A. Jeffares, F. Imrie, and M. van der Schaar, "Improving adaptive conformal prediction using self-supervised learning," in *Proc. AISTATS*, PMLR 206:10160–10177, 2023.

[10] G. S. Dhillon, G. Deligiannidis, and T. Rainforth, "On the expected size of conformal prediction sets," in *Proc. AISTATS*, PMLR 238:1549–1557, 2024.