CALIBRATING VERIFIED TASK SETS WITH LIMITED MEASUREMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

A learned controller can meet a task specification on several disconnected regions of physical parameter space. With few calibration measurements, a generic confidence interval may certify none of those regions. We study how a verified task set can instead determine a fixed-task acceptance rule. For a scalar Gaussian estimate, we reduce the worst unsafe-parameter risk to finitely many boundary evaluations when every acceptance component is separated from its boundary by at least one standard error. The result supports component-specific radii; maximizing a same-component power objective then becomes a convex optimization problem. Exact-rational inner-set verification connects the statistical calculation to repeated feedback with a shared uncertain restitution coefficient. On 63 tasks from 15 learned controllers, four-measurement acceptance increases from 41.0% with a gap-based test to 44.0% with a common boundary-calibrated radius and 44.7% with component allocation. Eight tasks change from zero to positive acceptance. Tighter-tolerance experiments locate this benefit in the narrow-component regime, and an independent checkpoint study compares calibration-aware selection with imitation and task-loss selection. Matched composite-null, continuum-search, and conformal controls identify which gains require Gaussian information and which are computational. The contribution is a calibrated post-training validation procedure under explicit model and measurement contracts, rather than a guarantee of unrestricted physical transfer.

1 INTRODUCTION

A low validation loss is not the same as evidence that a learned controller will meet a particular tolerance. A controller may work for some values of restitution and fail for others; feedback can split those values into several disconnected intervals. When only a few measurements are available, the relevant question is not whether every physical parameter has been accurately identified. It is whether the measurements rule out the parameter values that violate the task.

This paper studies that question after training. A verifier first computes physical parameters on which a frozen policy satisfies a specified task. Calibration then determines which measurement outcomes justify a claim of success. The two stages have different jobs: verification describes task behavior, while statistical calibration controls the chance of making a false claim. Keeping them separate also distinguishes three possible improvements: a better controller, a less conservative verifier, and a more powerful statistical rule.

The useful regime is *limited measurements relative to the width of a safe component*. For a safe interval three standard errors wide, the familiar two one-sided tests (TOST) radius of approximately 1.64 standard errors leaves an empty acceptance interval. Accounting for both finite endpoints gives a radius of approximately 1.31 and positive acceptance. This connected-interval calculation is classical equivalence testing, not a new contribution (Schuirmann, 1987; Berger & Hsu, 1996; Romano, 2005). It explains why finite widths matter before introducing the harder disconnected case. When several components are possible, their acceptance probabilities add at unsafe parameters. Calibrating each component in isolation can therefore violate the risk budget.

Our contribution is a tractable calculation for that coupling, together with a reliable interface to verified control tasks. We show that a finite set of boundary constraints controls a continuum of unsafe scalar parameters. We extend the calculation to different radii in different components and

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.

identify a convex allocation problem. This separates *where to allow acceptance* from the conventional choice of one common confidence radius. An independent directed-arithmetic check validates the returned rule, so optimizer convergence is not the basis for its risk guarantee.

The contact model provides a demanding application of the interface. The same unknown restitution coefficient multiplies velocities that already depend on earlier impacts and feedback. We preserve this repeated dependence using polynomial representations and exact-rational range validation. An inner approximation of the task-safe set is enough: unresolved intervals are excluded rather than silently merged. This is a restricted application of established set-based verification, not a new general reachability algorithm (Althoff et al., 2021; Casagrande et al., 2022).

The experiments ask two learning-relevant questions. First, which learned controller tasks remain certifiable with little calibration, and how does tightening the task tolerance change that answer? Second, does choosing a training checkpoint by imitation loss select the same controller as choosing it by anticipated certification performance on independent development tasks? Both comparisons include strong controls. Benefits beyond a gap-based statistical test are concentrated rather than universal, and a task-loss checkpoint selector captures much of the gain over imitation loss.

Scope in one place. The central exact test uses a scalar, unbiased Gaussian location estimate and a task fixed independently of that estimate. Its validity is conditional on a sound model-based task set. We derive such an estimate from a simple rebound-speed regression, give a conservative unknown-variance extension, and compare a conformal alternative with a separate residual-reference resource. None establishes that an arbitrary nonlinear system-identification output is Gaussian or that a simulator captures the true dynamics. Accordingly, the title and claims concern verified task-set calibration, not general sim-to-real transfer.

2 THE CERTIFICATION PROBLEM AND ITS VALIDITY CONTRACTS

A task c consists of a frozen policy, initial state, horizon H, goal g, and tolerance ϵ. For a parameter θ that remains fixed through the rollout, define

$$J_c(\theta) = \|W(x_H^{\pi}(\theta) - g)\|_2^2, \qquad S_{\text{true}} = \{\theta : J_c(\theta) \le \epsilon^2\}. \tag{1}$$

The verifier supplies a finite union of closed intervals

$$S = \bigcup_{i=1}^{K} [\ell_i, u_i] \subseteq S_{\text{true}}, \qquad u_i < \ell_{i+1}. \tag{2}$$

Training data and independent development data may determine c and S. We condition on those data before observing $Y \sim N(\theta, s^2)$, where s is the standard error. A measurement acceptance region A is valid at level δ when

$$\sup_{\theta \notin S} \mathbb{P}_\theta\{Y \in A\} \le \delta. \tag{3}$$

Because every truly failing parameter lies outside S, this also bounds false certification for the modeled task. Excluding additional safe parameters can reduce power, but cannot invalidate the implication. Abstention is not a prediction of task failure.

A shared confidence interval offers a different contract. If $U(Y)$ covers θ with probability at least $1 - \delta$, accepting when $U(Y) \subseteq S$ controls false claims even after choosing a task or a sound inner approximation from the same data. Coverage alone proves this statement; we treat it as a baseline. A fixed-task rule satisfying (3) need not protect such selection. Thus comparisons against a shared interval measure the benefit of using a narrower inference objective, not dominance at identical validity requirements.

For comparisons, let $p_A(\theta) = \mathbb{P}_\theta(Y \in A)$. Unless otherwise stated, power is its average under the uniform distribution on the same verified set S for all methods. This summarizes acceptance under a chosen parameter distribution; it is not a physical success rate. We also evaluate distributions concentrated near component boundaries and interiors. The checkpoint experiment instead uses a fixed prior over the full physical domain, so a controller is not rewarded merely for having a small but easy-to-certify safe set.

Table 1: Different validity contracts are not interchangeable. Every model-based claim additionally requires (2).

Rule	Measurements/resources	Protected claim
Gaussian shared interval	Known Gaussian standard error	Task selection after observing Y
Gap, boundary, allocated	Same Gaussian model	Fixed-task uniform unsafe risk
Reference conformal interval	199 additional labeled estimator residuals	Shared-set coverage, marginal over reference data; uniform in θ under translation-invariant errors

3 FINITE-BOUNDARY CALIBRATION AND COMPONENT ALLOCATION

For radii $z_i \geq 1$, define

$$A_{\boldsymbol{z}} = \bigcup_i [\ell_i + z_i s, u_i - z_i s], \tag{4}$$

where reversed or zero-length intervals contribute zero Gaussian probability. Components retain their identity; a measurement interval cannot bridge an unsafe gap.

Theorem 1 (Boundary reduction, including different radii)**.** *For* (2)*, known* $s > 0$*, and* $\min_i z_i \geq 1$*, the exact unsafe risk is*

$$\sup_{\theta \notin S} p_{A_{\boldsymbol{z}}}(\theta) = \max_{b \in \{\ell_1, u_1, \ldots, \ell_K, u_K\}} p_{A_{\boldsymbol{z}}}(b), \tag{5}$$

with boundary values interpreted as limits from unsafe parameters. Each value is a sum of Gaussian interval probabilities. Computing all constraints takes $O(K^2)$ *cumulative distribution function (CDF) evaluations.*

Why it works. For an unsafe mean, every accepted observation lies at least s away. The Gaussian density, viewed as a function of its mean, has nonnegative second derivative at those distances. Integrating over accepted observations makes risk convex on each bounded unsafe gap. Its maximum is at an endpoint. On the two exterior rays risk is monotone toward the safe set. This proof does not require the radii to be equal. Appendix A gives the complete argument.

For a common radius z, monotonicity permits a scalar search for the smallest feasible $z \geq 1$. The gap comparator instead solves

$$\Phi(-z) + \Phi(-z - g_{\min}/s) = \delta, \tag{6}$$

where $g_{\min}$ is the minimum unsafe gap and is infinite for one component. It bounds all acceptance to the left and right by half-lines. The boundary calculation retains finite component widths, subtracting probability that those half-lines incorrectly include. Consequently, the two rules nearly agree when relevant components are wide relative to s; the additional gain appears when widths are only a few standard errors.

3.1 ALLOCATING ACCEPTANCE ACROSS HETEROGENEOUS COMPONENTS

A common radius may be constrained by one part of the safe set while unnecessarily shrinking another. Write $m_i = (\ell_i + u_i)/2$ and let h_i be the accepted halfwidth. For components with $u_i - \ell_i \geq 2s$, allow

$$0 \leq h_i \leq (u_i - \ell_i)/2 - s, \qquad A_i(h_i) = [m_i - h_i, m_i + h_i]. \tag{7}$$

Narrower components are inactive under the one-standard-error restriction. Fix nonnegative weights ω_i summing to one and define the same-component power

$$Q(\boldsymbol{h}) = \sum_i \frac{\omega_i}{u_i - \ell_i} \int_{\ell_i}^{u_i} \mathbb{P}_\theta\{Y \in A_i(h_i)\}\, d\theta. \tag{8}$$

This is a lower bound on total acceptance under the corresponding mixture of uniform component distributions: it does not count a safe true mean whose noisy observation is accepted in a different component.

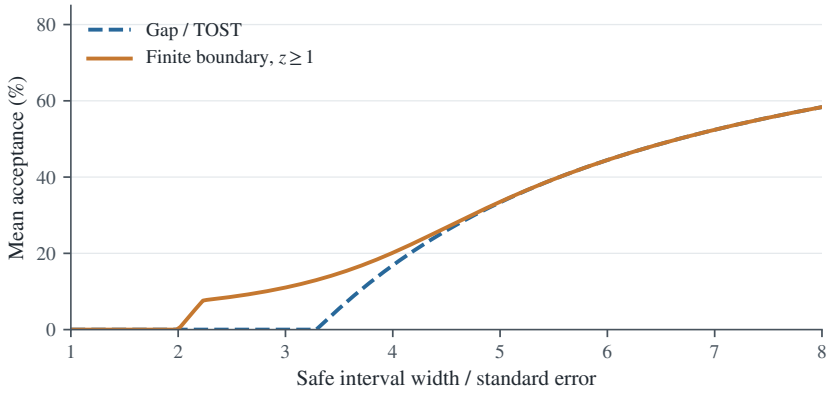


Figure 1: Why finite width matters. A gap/TOST radius can leave no accepted observations when the safe component is only a few standard errors wide. The restricted finite-boundary rule recovers a nonempty region; below width two its radius floor also loses power. The connected-interval mechanism is classical (Romano, 2005); disconnected sets require accounting for acceptance from every component.

Algorithm 1 Prepare a checked rule before measuring

Require: Sound inner set S, scale s, risk δ, and a fixed design prior

1: Compute a feasible common radius using boundary constraints.
2: Optimize component halfwidths; retain the common fallback.
3: Increase proposed radii until a directed risk upper bound is $\leq \delta$.
4: Freeze the accepted intervals and record the exact risk bound.
5: On a later measurement, certify only if its rounding enclosure lies in one accepted interval.

Theorem 2 (Convex component allocation)**.** *On* (7)*, each boundary risk $p_{A_{\boldsymbol{h}}}(b)$ is convex in $\boldsymbol{h}$, while Q is concave. Therefore maximizing Q subject to all boundary risks being at most δ is a convex optimization problem. Every feasible solution satisfies* (3)*.*

For a boundary b, each accepted interval lies on one side of b, at distance at least s. Convexity of the density there gives nonnegative second derivative in h_i. For the objective, convolving one uniform component with Gaussian noise yields a symmetric density decreasing away from its midpoint. Its centered interval mass is concave in halfwidth. These curvature properties give a convex optimization problem.

Our implementation uses analytic gradients and a common-radius feasible initialization. The main allocated rule optimizes full uniform-set power; a second run optimizes the concave objective (8). Full power need not be concave, so global optimality is claimed only for the ideal convex formulation, not for that numerical full-power search. If a proposed vector has lower evaluated full power than the common rule, the implementation keeps the common rule. Both objectives give nearly identical powers in the present tasks, but that observation is not a general equivalence theorem. In every case, a separate outward-rounded probability check, rather than an optimizer status flag, establishes feasibility.

3.2 What the radius restriction buys, and what it loses

The condition $z_i \geq 1$ is sufficient for an inexpensive exact maximum; it is not necessary for a valid test. We implement a second method that bounds risk over every unsafe interval using directed cumulative distribution function (CDF) evaluations and

$$\sup_{\theta \in [a,b]} p_A(\theta) \leq \max\{\overline{p}_A(a), \overline{p}_A(b)\} + \frac{(b-a)^2}{8s^2}. \tag{9}$$

The bound follows from $|p''_A| \le s^{-2}$ and linear interpolation. Rational subdivision makes it arbitrarily tight without a one-standard-error restriction. This provides both a strong control and a fallback for very narrow components. Where the boundary theorem applies, its main advantage over this route is avoiding a continuum subdivision, not granting a kind of validity that no other method can provide.

4 From a verified model to a measurement claim

4.1 Repeated impacts and sound inner sets

Our application is a sampled controller driving a point object between walls 0 and 1. Let v denote the previous velocity, $u = \mathrm{clip}(\pi(x, v, r), -.32, .32)$ the learned action, and $w = .35v + .65u$. With $f = x + w$, one step is

$$(x', v') = \begin{cases} (-ef, -ew), & f \le 0, \\ (1 - e(f-1), -ew), & f \ge 1, \\ (f, w), & 0 < f < 1. \end{cases} \tag{10}$$

The restitution $e \in [.1, .95]$ is shared by every impact. The speed bound permits at most one impact per sampled step. Rectified linear unit (ReLU) decisions, clipping, and contacts create different parameter branches; multiplication by e raises polynomial degree along a branch. This model has genuine feedback and repeated uncertain impacts, but one uncertain scalar.

Proposition 3 (Verified inner-set interface)**.** *Represent each state polynomial on a closed parameter cell by exact rational Bernstein coefficients. Preserve every compatible controller and contact continuation, splitting or conservatively covering unresolved guards. Accept a cell only if every terminal enclosure satisfies the task tolerance. For any finite resource budget, the union of accepted cells satisfies* (2)*.*

The proof is an induction on the retained state cover. Bernstein coefficients bound polynomial ranges, and rational subdivision preserves the cover (Farouki, 2012). A resource limit causes abstention, never an unproved positive result. Crucially, equality trajectories are included: post-impact velocity can affect later feedback even when position is continuous at the event. Accepted endpoints are verified too. Tiny unresolved gaps are not merged. Appendix B supplies the identities and trusted-computation contract.

The radius search uses binary arithmetic only to propose a rule. We then evaluate endpoint risks with directed Multiple Precision Floating-Point Reliable (MPFR) bounds and increase the radii until the rational upper bound is at most $1/20$ (Fousse et al., 2007). Exported safe endpoints round inward. Final containment can use rational arithmetic and an externally justified enclosure for a rounded observation. These checks establish a claim for the recorded rational dynamics; unmodeled discrepancies in the dynamics or measurements require separate bounds.

4.2 An explicit calibration experiment, and unknown variance

A scalar Gaussian estimator can arise from a specified measurement equation. Suppose approach speeds a_i are controlled and rebound readings have independent and identically distributed (iid) Gaussian errors:

$$V_i = b - ea_i + \xi_i, \qquad \xi_i \overset{\text{iid}}{\sim} N(0, \tau^2). \tag{11}$$

Regressing V_i on an intercept and a_i gives $Y = -\widehat{\text{slope}} \sim N(e, \tau^2 / \sum_i (a_i - \bar{a})^2)$. The intercept absorbs a constant sensor offset. Errors in the commanded approach speeds, velocity-dependent restitution, or heteroscedastic noise are not covered by this derivation.

If τ is unknown, let $\widehat{s}$ be the usual regression standard error and $\nu = n - 2$. The residual scale is independent of Y. For $\beta < \delta$, set

$$s_U = \widehat{s}\sqrt{\nu / \chi^2_{\beta,\nu}}. \tag{12}$$

It exceeds the true standard error with probability $1 - \beta$. Conditional on s_U, calibrate at scale s_U and level $\delta - \beta$, retaining radii at least one. At an unsafe mean all accepted observations are at least s_U away; their Gaussian density increases with scale up to s_U. Therefore the resulting rule has risk at most δ. This conservative extension does not replace the stronger Student confidence interval in every setting; we compare them explicitly.

Table 2: Mean acceptance (%) on 63 tasks. The last column is the common-minus-gap improvement in percentage points, with a paired 95% controller-cluster bootstrap interval. Shared Gaussian coverage also protects task selection; the other columns concern a fixed task.

n	Shared	Gap	Common	Allocated	Common–gap (95% interval)
4	38.3	41.0	44.0	44.7	2.99 [2.33, 3.70]
16	59.2	61.7	63.3	64.0	1.57 [1.13, 2.00]
64	75.7	77.7	78.5	78.8	0.77 [0.49, 1.12]
256	86.7	88.0	88.3	88.5	0.38 [0.19, 0.62]

Adaptive verification is a statistical operation. After seeing an abstention, raising the verification budget and recalibrating against a newly refined set at the same nominal level can invalidate fixed-task risk control. A valid protocol predeclares sets/budget levels and risks δ_j with $\sum_j \delta_j \leq \delta$; a union bound then protects any stopping choice using the same Y. Alternatively, use independent pilot data to choose the verifier and fresh data for calibration. A shared confidence set needs no such adjustment for refining a sound verifier; sequentially collecting new measurements instead requires a time-uniform confidence sequence or another valid sequential protocol (Howard et al., 2021). The software separates set/radius preparation from the measurement query.

5 Experiments

All experiments use synthetic data. We distinguish different tasks from repeated statistical draws. The main cohort contains 63 tasks from 15 trained controllers, not 63 independent networks. The focused cohort fixes 30 controller/context/horizon combinations and evaluates three tolerances. Thirty of its 90 task definitions repeat a subset of the main cohort. A separate checkpoint experiment uses eight independent training seeds.

5.1 Small-data power and component allocation

Table 2 uses the same verified set and known standard error $.1/\sqrt{n}$ for every method. Power is integrated analytically, not estimated from millions of nominally independent control trials. Paired intervals resample the fifteen controller blocks, retaining all their tasks, with 10,000 bootstrap replicates. They describe variation in this synthetic controller cohort; they are not distribution-free population confidence bounds for all robots.

At four measurements, common-radius calibration raises acceptance from 38.3% for the shared interval to 44.0%, a 14.9% relative increase, but only the increase over the 41.0% gap rule isolates finite-width calibration. Eight tasks have empty gap-rule acceptance regions and positive common-rule acceptance. Allocated radii raise the mean further to 44.7%. Their largest individual gain over a common radius is 5.00 percentage points. At 64 measurements the additional common-minus-gap mean is 0.77 points, its median is effectively zero, and twelve tasks exceed one point. These are different operating regimes, not conflicting conclusions.

The focused experiment tests that interpretation without selecting tasks by their resulting geometry. For each controller, the first context at horizons 12 and 24 is fixed, and tolerance is reduced from 0.10 to 0.05 and $1/40$. Table 3 includes every resulting task. At tolerance .05 and four measurements, mean acceptance rises from 15.1% with the gap rule to 18.3% with the common rule and 18.9% with allocation. Six of ten gap-empty cases become nonempty. At tolerance $1/40$, even the improved rule averages only 4.0% acceptance: recovering a nonempty region is useful, but does not make an undermeasured task easy.

The geometry predicts the gain. A gap bound treats neighboring accepted sets as infinite rays. At an endpoint it omits finite-width tail subtractions; those terms matter when component width is only a few standard errors. On the main-cohort tasks at $n = 64$, all twelve gains above one point occur in the stratum with minimum gap below one standard error. Wide, well-separated components require almost no correction beyond the gap rule. The common-radius floor is inactive in all 252 main-cohort task/budget cases; it is active in some deliberately tighter tasks. The main table omits the radius-floor count because it is zero in every row.

Table 3: Four-measurement focused cohort: 30 paired tasks per tolerance. "Rescued" counts gap-empty cases with positive common-rule acceptance. All three tolerance settings are retained.

Tolerance $\times 10^2$	Shared	Gap	Common	Allocated	Rescued
2.50	0.4	1.8	4.0	4.0	9/25
5.00	8.7	15.1	18.3	18.9	6/10
10.00	40.0	43.7	45.8	46.3	2/2

Sensitivity to the power estimand. Moving the same prior mass to the outer 10% on each side of each component lowers four-measurement acceptance to 8.2%, 10.2%, and 10.6% for gap, common, and allocated rules. Concentrating it in the middle 20% instead gives 62.2%, 65.3%, and 66.2%. Absolute acceptance is therefore prior-sensitive. The common rule contains the gap acceptance region, so its ordering is prior-independent; a vector optimized for one prior need not dominate pointwise.

5.2 Closest statistical controls and unrestricted continuum search

We instantiate the scalar normal pointwise-rejection approximation of Park & Wang (2026). Its adjusted significance $\alpha' = 2\delta$ gives radius $\Phi^{-1}(1-\delta)$ when the unsafe region consists of scalar intervals. This approximation has maximum computed unsafe risk between 9.3% and 9.8% on the main cohort, depending on budget. At $n = 64$ it exceeds 5% in 30 of 63 tasks. This is a result about the specified approximation on disconnected narrow-gap sets, not a claim that every possible calibration of their general framework fails. Replacing its adjustment with a valid continuum size calculation recovers the common-radius family studied here. The novelty is the exact finite maximum and its allocation formulation, not the principle of pointwise rejection.

For thirty fixed control task/budget cases, the directed continuum search (9) agrees with boundary-calibrated power to the numerical search tolerance. It also finds positive-power radii below one for very narrow single intervals where the restricted boundary method abstains. Thus boundary reduction primarily simplifies calibration where its condition holds; it is not uniformly more powerful than rigorous continuum optimization. Appendix A gives the continuum bound and stopping tolerances.

The conformal control uses 199 independent labeled estimator residuals in addition to the target measurements. Under Gaussian errors it gives 38.1% acceptance at four measurements, close to the 38.3% shared Gaussian interval. With standardized Laplace estimator errors, the Gaussian task rule reaches 6.20% risk at an evaluated unsafe boundary, whereas the residual-based rule retains its distribution-free rank guarantee under the stated exchangeability model. This comes with a different information resource and a marginal-over-reference-data contract. It is not a free replacement for a known-noise experiment with only four measurements. The raw rebound-regression experiment separately verifies the unbiased Gaussian derivation and shows the substantial small-sample cost of the variance-upper-bound extension.

5.3 Does calibration change checkpoint selection?

The learning experiment compares four checkpoints from each of eight independent training runs. Every selector receives the same checkpoints and three independent development tasks. It chooses by held-out imitation loss, average simulated task cost, or expected verified correct-certification mass. The latter integrates acceptance over the verified safe set under a *fixed prior on the entire physical domain*; it cannot improve its score simply by shrinking the set of safe true parameters. Four independent contexts per seed form the held-out evaluation. Neither controller selection nor radius preparation observes the eventual calibration measurement.

Imitation loss chooses the last checkpoint in every seed. Certification-based selection chooses a different checkpoint in seven seeds and increases mean held-out certificate utility from 8.4% to 12.4%. The strong task-cost selector reaches 11.8%, and differs from certificate selection in only two seeds. This study answers a post-training model-selection question, not a claim that the statistical theorem requires neural networks or that a new training algorithm has been discovered. It shows why the evaluation objective matters while keeping the stronger inexpensive selector visible.

Table 4: Held-out checkpoint selection at $n = 4$, tolerance .05. Values are means $\pm$ standard errors over eight training seeds. Certificate utility is a lower bound on the joint probability of task success and acceptance under a uniform full-domain parameter prior. Simulation success is a separate grid-evaluated diagnostic.

Selector	Certificate utility (%)	Success (%)	Cost $\times 10^3$
Imitation loss	8.4 ± 1.9	32.5 ± 3.6	7.30 ± 1.03
Task cost	11.8 ± 2.9	36.8 ± 4.6	6.65 ± 0.96
Certificate utility	12.4 ± 2.8	38.2 ± 4.2	6.35 ± 0.86

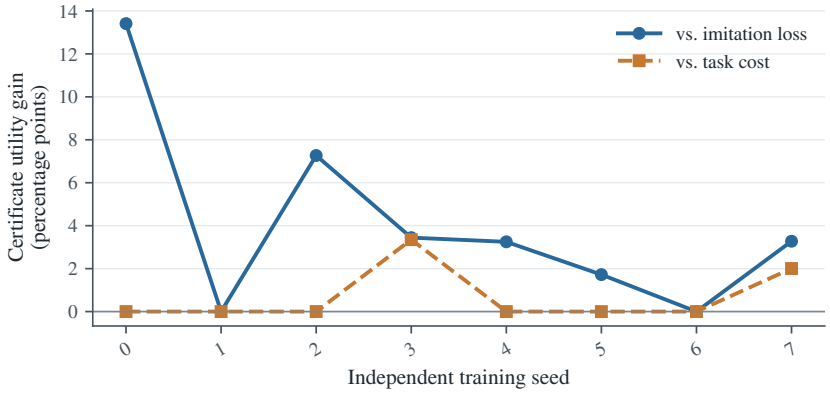


Figure 2: Paired held-out checkpoint results. The certification criterion improves over imitation-loss selection in several seeds, but differs from the stronger task-cost selector in only two. Each point averages four held-out contexts of one independently trained network; contexts are not treated as separate training replicates.

5.4 Verification, computation, and dimensional limits

The exact-rational soundness checks include close unsafe roots and isolated unsafe equality events. The branch-cover ablation changes the verifier, not the calibration rule. At the four budgets its acceptance recovery is 15.7, 13.6, 10.4, and 6.9 percentage points under the same Gaussian radius. It illustrates sensitivity to small unverified holes; this gain is attributable to verification rather than statistical calibration.

A fully cached maximum-cost check stores stationary points, values, breakpoint trajectories, and range maxima. This baseline accounts for reuse of prepared computations; comparing query cost alone does not establish an overall speed advantage. An independent dependency-preserving mixed-integer linear program (MILP) matches the exact verification decisions on all 30 compared cases. In the benign feedback-selection control, plug-in selection followed by exact verification and robust candidate selection both certify 65.5% of cases at 16 probes. These controls limit claims of superiority for the verifier and selector.

Beyond one parameter, a scalar robust profile $S_e = \{e : \forall \nu \in B,\ J(e, \nu) \leq \epsilon^2\}$ can use the same calibration if the nuisance box B is guaranteed. A full Gaussian vector alternative covers the whole unsafe set by a finite net and pays an explicit total-variation radius penalty. Neither route inherits a finite scalar-boundary theorem. Constructing an unsafe-set cover and bounding multivariate probabilities can be expensive; the scalar result does not establish efficient multivariate verification. Appendix D gives the bound and a two-dimensional counterexample to naive boundary maximization.

6 Related work and contribution boundaries

Task-directed control and identification. Domain randomization and simulator identification address model uncertainty in policy development (Tobin et al., 2017; Peng et al., 2018; Chebotar et al.,

2019; Ramos et al., 2019; Muratore et al., 2022). Value-aware model learning, value equivalence, and task-optimal exploration already target downstream decisions rather than uniform prediction accuracy (massoud Farahmand et al., 2017; Voelcker et al., 2022; Grimm et al., 2020; Wagenmaker et al., 2021). We study a different interface: converting a *verified fixed-task parameter set* into a statistically calibrated claim, and using that criterion to evaluate trained checkpoints. Robust Markov decision process (MDP) and learning-based control results provide broader uncertainty-aware decision frameworks (Iyengar, 2005; Nilim & El Ghaoui, 2005; Hewing et al., 2020).

Verification and contact. Contact-task verification and conformance identification are direct antecedents (Tang & Althoff, 2023; 2024); hybrid reachability, mixed logical dynamics, and neural-network verification already preserve parameter dependencies (Bemporad & Morari, 1999; Bemporad et al., 2002; Bak & Duggirala, 2017; Tran et al., 2020; Everett et al., 2021; Tjeng et al., 2019). Polynomial parameter synthesis and nonlinear flowpipe methods likewise precede our inner-set construction (Casagrande et al., 2022; Chen et al., 2013). Contact sensitivity and impact-data design explain why event structure can matter (Kong et al., 2024; Payne et al., 2022; Tucker et al., 2023), while known limitations of contact models preclude treating mathematical verification as validation of arbitrary physical dynamics (Fazeli et al., 2020; Le Lidec et al., 2024; Stewart & Trinkle, 1996; Anitescu & Potra, 1997). We contribute a calibrated use of these verified sets, not a generally superior verifier.

Statistical inference. The fixed connected-interval case belongs to equivalence testing (Schuirmann, 1987; Berger & Hsu, 1996; Romano, 2005). Composite-null pointwise rejection has a broader approximate formulation (Park & Wang, 2026). Our exact union-boundary result and convex component allocation concern a narrower Gaussian class. Confidence-set coverage, testing lower bounds, and sequential validity are established foundations (Neyman & Pearson, 1933; Kaufmann et al., 2016; Howard et al., 2021). Conformal prediction relaxes distributional assumptions using exchangeability and a calibration resource (Angelopoulos & Bates, 2023); our comparator states the extra translation-invariance assumption needed to turn marginal residual coverage into uniform parameter-level control. This distinction is essential rather than a claim of dominance over distribution-free inference.

7 Conclusion

Verified task sets can support more useful claims than generic confidence intervals when safe components are narrow relative to measurement noise. The finite-boundary calculation makes that fixed-task calibration tractable, and component allocation uses heterogeneous geometry more effectively. The benefit is concentrated in narrow components: it can make acceptance possible, but adds little beyond strong gap rules for wide components. Sound inner sets, checked probabilities, independent preparation, and explicit measurement assumptions are necessary parts of the procedure. The results support a model-conditional, scalar post-training validation tool and motivate better multivariate and measurement-robust methods; they do not establish broad physical-transfer guarantees.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper and polish writing. Additionally, we used generative AI tools for editing code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Matthias Althoff, Goran Frehse, and Antoine Girard. Set propagation techniques for reachability analysis. *Annual Review of Control, Robotics, and Autonomous Systems*, 4(1):369–395, 2021. doi: 10.1146/annurev-control-071420-081941. URL https://doi.org/10.1146/annurev-control-071420-081941.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://arxiv.org/abs/2107.07511.

Mihai Anitescu and Florian A. Potra. Formulating dynamic multi-rigid-body contact problems with friction as solvable linear complementarity problems. *Nonlinear Dynamics*, 14(3):231–247, 1997. doi: 10.1023/A:1008292328909. URL https://link.springer.com/article/10.1023/A:1008292328909.

Stanley Bak and Parasara Sridhar Duggirala. HyLAA: A tool for computing simulation-equivalent reachability for linear systems. In *International Conference on Hybrid Systems: Computation and Control*, pp. 173–178, 2017. doi: 10.1145/3049797.3049808. URL https://doi.org/10.1145/3049797.3049808.

Alberto Bemporad and Manfred Morari. Control of systems integrating logic, dynamics, and constraints. *Automatica*, 35(3):407–427, 1999. doi: 10.1016/S0005-1098(98)00178-2. URL https://doi.org/10.1016/S0005-1098(98)00178-2.

Alberto Bemporad, Manfred Morari, Vivek Dua, and Efstratios N. Pistikopoulos. The explicit linear quadratic regulator for constrained systems. *Automatica*, 38(1):3–20, 2002. doi: 10.1016/S0005-1098(01)00174-1. URL https://doi.org/10.1016/S0005-1098(01)00174-1.

Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996. doi: 10.1214/ss/1032280304. URL https://doi.org/10.1214/ss/1032280304.

Alberto Casagrande, Thao Dang, Luca Dorigo, Tommaso Dreossi, Carla Piazza, and Eleonora Pippia. Parameter synthesis of polynomial dynamical systems. *Information and Computation*, 289:104941, 2022. doi: 10.1016/j.ic.2022.104941. URL https://arts.units.it/handle/11368/3025969.

Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *IEEE International Conference on Robotics and Automation*, pp. 8973–8979, 2019. doi: 10.1109/ICRA.2019.8793789. URL https://arxiv.org/abs/1810.05687.

Xin Chen, Erika Ábrahám, and Sriram Sankaranarayanan. Flow*: An analyzer for non-linear hybrid systems. In *Computer Aided Verification*, volume 8044 of *Lecture Notes in Computer Science*, pp. 258–263, 2013. doi: 10.1007/978-3-642-39799-8_18. URL https://link.springer.com/chapter/10.1007/978-3-642-39799-8_18.

Michael Everett, Golnaz Habibi, and Jonathan P. How. Efficient reachability analysis of closed-loop systems with neural network controllers. In *IEEE International Conference on Robotics and Automation*, pp. 4384–4390, 2021. doi: 10.1109/icra48506.2021.9561348. URL https://arxiv.org/abs/2101.01815.

Rida T. Farouki. The Bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, 29(6):379–419, 2012. doi: 10.1016/j.cagd.2012.03.001. URL https://doi.org/10.1016/j.cagd.2012.03.001.

Nima Fazeli, Samuel Zapolsky, Evan Drumwright, and Alberto Rodriguez. Fundamental limitations in performance and interpretability of common planar rigid-body contact models. In *Robotics Research*, volume 10 of *Springer Proceedings in Advanced Robotics*, pp. 555–571. Springer, 2020. doi: 10.1007/978-3-030-28619-4_41. URL https://arxiv.org/abs/1710.04979.

Laurent Fousse, Guillaume Hanrot, Vincent Lefèvre, Patrick Pélissier, and Paul Zimmermann. MPFR: A multiple-precision binary floating-point library with correct rounding. *ACM Transactions on Mathematical Software*, 33(2):13:1–13:15, 2007. doi: 10.1145/1236463.1236468. URL https://doi.org/10.1145/1236463.1236468.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3bb585ea00014b0e3ebe4c6dd165a358-Abstract.html.

Lukas Hewing, Kim P. Wabersich, Marcel Menner, and Melanie N. Zeilinger. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):269–296, 2020. doi: 10.1146/annurev-control-090419-075625. URL https://doi.org/10.1146/annurev-control-090419-075625.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991. URL https://arxiv.org/abs/1810.08240.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005. doi: 10.1287/moor.1040.0129. URL https://pubsonline.informs.org/doi/10.1287/moor.1040.0129.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://jmlr.org/papers/v17/kaufman16a.html.

Nathan J. Kong, J. Joe Payne, James Zhu, and Aaron M. Johnson. Saltation matrices: The essential tool for linearizing hybrid dynamical systems. *Proceedings of the IEEE*, 112(6):585–608, 2024. doi: 10.1109/jproc.2024.3440211. URL https://arxiv.org/abs/2306.06862.

Quentin Le Lidec, Wilson Jallet, Louis Montaut, Ivan Laptev, Cordelia Schmid, and Justin Carpentier. Contact models in robotics: A comparative analysis. *IEEE Transactions on Robotics*, 40:3716–3733, 2024. doi: 10.1109/tro.2024.3434208. URL https://arxiv.org/abs/2304.06372.

Amir massoud Farahmand, André Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1486–1494, 2017. URL https://proceedings.mlr.press/v54/farahmand17a.html.

Fabio Muratore, Fabio Ramos, Greg Turk, Wenhao Yu, Michael Gienger, and Jan Peters. Robot learning from randomized simulations: A review. *Frontiers in Robotics and AI*, 9:799893, 2022. doi: 10.3389/frobt.2022.799893. URL https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2022.799893/full.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231(694-706):289–337, 1933. doi: 10.1098/rsta.1933.0009. URL https://doi.org/10.1098/rsta.1933.0009.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216. URL https://pubsonline.informs.org/doi/10.1287/opre.1050.0216.

Joonha Park and Ming Wang. A novel finite-sample testing procedure for composite null hypotheses via pointwise rejection. *arXiv preprint arXiv:2601.02529*, 2026. URL https://arxiv.org/abs/2601.02529.

J. Joe Payne, Nathan J. Kong, and Aaron M. Johnson. The uncertainty aware salted Kalman filter: State estimation for hybrid systems with uncertain guards. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022. doi: 10.1109/iros47612.2022.9981218. URL https://arxiv.org/abs/2202.12729.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *IEEE International Conference on Robotics and Automation*, pp. 3803–3810, 2018. doi: 10.1109/ICRA.2018.8460528. URL https://arxiv.org/abs/1710.06537.

Fabio Ramos, Rafael Carvalhaes Possas, and Dieter Fox. BayesSim: Adaptive domain randomization via probabilistic inference for robotics simulators. In *Robotics: Science and Systems*, 2019. doi: 10.15607/rss.2019.xv.029. URL https://www.roboticsproceedings.org/rss15/p29.html.

Joseph P. Romano. Optimal testing of equivalence hypotheses. *The Annals of Statistics*, 33(3): 1036–1047, 2005. doi: 10.1214/009053605000000048. URL https://arxiv.org/abs/math/0507415.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419. URL https://pubmed.ncbi.nlm.nih.gov/3450848/.

David E. Stewart and Jeffrey C. Trinkle. An implicit time-stepping scheme for rigid body dynamics with inelastic collisions and Coulomb friction. *International Journal for Numerical Methods in Engineering*, 39(15):2673–2691, 1996. doi: 10.1002/(sici)1097-0207(19960815)39:15<2673::aid-nme972>3.0.co;2-i. URL https://doi.org/10.1002/(SICI)1097-0207(19960815)39:15%3C2673::AID-NME972%3E3.0.CO;2-I.

Chencheng Tang and Matthias Althoff. Formal verification of robotic contact tasks via reachability analysis. *IFAC-PapersOnLine*, 56(2):7912–7919, 2023. doi: 10.1016/j.ifacol.2023.10.028. URL https://arxiv.org/abs/2307.13977.

Chencheng Tang and Matthias Althoff. Efficiently obtaining reachset conformance for the formal analysis of robotic contact tasks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2305–2311, 2024. doi: 10.1109/iros58592.2024.10802263. URL https://arxiv.org/abs/2410.10391.

Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.07356.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133. URL https://arxiv.org/abs/1703.06907.

Hoang-Dung Tran, Xiaodong Yang, Diego Manzanas Lopez, Patrick Musau, Luan Viet Nguyen, Weiming Xiang, Stanley Bak, and Taylor T. Johnson. NNV: The neural network verification tool for deep neural networks and learning-enabled cyber-physical systems. In *Computer Aided Verification*, volume 12224 of *Lecture Notes in Computer Science*, pp. 3–17, 2020. doi: 10.1007/978-3-030-53288-8_1. URL https://arxiv.org/abs/2004.05519.

Maegan Tucker, Noel Csomay-Shanklin, and Aaron D. Ames. Robust bipedal locomotion: Leveraging saltation matrices for gait optimization. In *IEEE International Conference on Robotics and Automation*, pp. 12218–12225, 2023. doi: 10.1109/icra48891.2023.10161309. URL https://arxiv.org/abs/2209.10452.

Claas Voelcker, Victor Liao, Animesh Garg, and Amir massoud Farahmand. Value gradient weighted model-based reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2204.01464.

Andrew Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10641–10652, 2021. URL https://proceedings.mlr.press/v139/wagenmaker21a.html.

A Complete statistical proofs

A.1 The inner-set implication and shared coverage

Write $S \subseteq S_{\text{true}}$. For any truly unsafe θ, membership in S is impossible. Thus

$$\sup_{\theta \notin S_{\text{true}}} \mathbb{P}_\theta(Y \in A) \leq \sup_{\theta \notin S} \mathbb{P}_\theta(Y \in A).$$

This implication does not require S to recover every true safe component. It does require the inclusion premise for the actual model being claimed. For a shared confidence set $U(D)$, on the event $\theta \in U(D)$ every accepted containment $U(D) \subseteq S_c(D) \subseteq S_{c,\text{true}}$ is correct, even if the task index and sound inner approximation depend on D. Hence the probability of any false containment-based claim is bounded by the coverage failure probability. By contrast, a task-specific acceptance region is not itself a confidence set and does not inherit that simultaneous implication.

A.2 Proof of Theorem 1

Let $f_s(y-\theta) = s^{-1}\phi((y-\theta)/s)$ and let $A = A_{\boldsymbol{z}}$. Empty components can be omitted. On any bounded unsafe gap (u_i, ℓ_{i+1}), every $y \in A$ satisfies $|y-\theta| \geq s$: accepted observations on the left lie at most at $u_i - s$, while those on the right lie at least at $\ell_{i+1} + s$. The Gaussian derivative is

$$\partial_\theta^2 f_s(y-\theta) = \left(\frac{(y-\theta)^2}{s^4} - \frac{1}{s^2}\right) f_s(y-\theta) \geq 0.$$

Finite interval unions are bounded, so differentiation under the integral is justified by a continuous integrable dominating derivative on every compact mean interval. Therefore $p_A(\theta) = \int_A f_s(y-\theta)dy$ is convex on each unsafe gap. Continuity extends its supremum to the closed gap, where a convex function takes its maximum at an endpoint. Those endpoints may themselves be safe; the displayed values are the limits approached through unsafe means. For $\theta < \ell_1$, all accepted observations lie to its right, so $\partial_\theta f_s(y-\theta) = (y-\theta)f_s/s^2 \geq 0$. The analogous derivative is nonpositive for $\theta > u_K$. The two exterior suprema occur at the outer endpoints. Taking the maximum over all unsafe components proves the result. Each endpoint requires at most K Gaussian interval probabilities and there are $2K$ endpoints.

This proof applies to different z_i without modification. It fails below one standard error because the second derivative can be negative near an unsafe mean. In that case risk can peak inside an unsafe gap, rather than at either boundary.

A.3 The minimum-gap comparator

Fix a common radius $r = zs$ and an unsafe mean in a bounded gap $[a, b]$ of width g. Accepted observations are contained in $(-\infty, a-r] \cup [b+r, \infty)$. At $\theta = a + d$, their probability is at most

$$F(d) = \Phi(-z - d/s) + \Phi(-z - (g-d)/s), \quad 0 \leq d \leq g.$$

For $z \geq 0$, each term is convex in d on this interval, so F is maximized at $d = 0$ or g. Exterior unsafe rays require only one tail. The largest resulting bound occurs at the smallest gap, giving (6). With a single component the second term is zero. For the risk levels used here, $\Phi^{-1}(1-\delta) > 1$, so the gap radius is also feasible for Theorem 1. Consequently the ideal smallest common boundary radius cannot exceed it. Rounding and directed inflation can alter the last few floating-point bits, not the mathematical ordering.

A.4 Finite widths and the classical connected-interval example

For one safe interval of width w, a nonempty eroded interval has endpoint risk

$$R_w(z) = \Phi(-z) - \Phi(z - w/s).$$

The second term is exactly what an infinite-half-line approximation drops. When $w/s = 3$ and $\delta = .05$, solving $R_w(z) = \delta$ gives $z \approx 1.31$. The gap radius $\Phi^{-1}(0.95) \approx 1.64$ makes $w - 2zs < 0$ and therefore yields no acceptance. The finite-width rule has mean acceptance 0.11 under the uniform

safe prior and 0.15 at the midpoint. This is a known equivalence-testing phenomenon (Romano, 2005); its role here is explanatory, not a novelty claim.

For two neighboring safe components of widths w_L, w_R and unsafe gap g, the risk at the right component's left endpoint, counting those two components, is

$$\Phi(-z) + \Phi(-z - g/s) - \Phi(z - w_R/s) - \Phi(z - (g + w_L)/s).$$

The minimum-gap bound omits the subtracted terms. Finite widths only a few standard errors wide make that omission substantial. More distant components add finite interval probabilities, not independent significance budgets. This explains why the gap alone does not predict the total improvement.

A.5 Proof of Theorem 2

Consider an active component i of width $w_i \geq 2s$. Fix an unsafe boundary b and write

$$r_{b,i}(h) = \int_{m_i - h}^{m_i + h} f_s(y - b)dy.$$

Because b is not inside this component, the whole integration interval is at distance at least s from b for every $0 \leq h \leq w_i/2 - s$. Direct differentiation gives

$$r'_{b,i}(h) = f_s(m_i + h - b) + f_s(m_i - h - b), \qquad r''_{b,i}(h) = f'_s(m_i + h - b) - f'_s(m_i - h - b).$$

The density is convex on the relevant one-sided domain, so its first derivative is nondecreasing there. Thus $r''_{b,i} \geq 0$. Summing these univariate convex functions makes every $r_b(\boldsymbol{h})$ convex. Their sublevel constraints form a convex feasible region; $\boldsymbol{h} = 0$ is feasible.

For the objective, let

$$q_i(y) = \frac{1}{w_i} \int_{\ell_i}^{u_i} f_s(y - \theta)d\theta = \frac{\Phi((y - \ell_i)/s) - \Phi((y - u_i)/s)}{w_i}.$$

It is symmetric about m_i, and for $t \geq 0$

$$q'_i(m_i + t) = \frac{\phi((t + w_i/2)/s) - \phi((t - w_i/2)/s)}{w_i s} \leq 0.$$

Therefore $\frac{d^2}{dh^2} \int_{m_i - h}^{m_i + h} q_i(y)dy = 2q'_i(m_i + h) \leq 0$. Nonnegative weighted sums preserve concavity. This proves the convex optimization characterization. Components narrower than $2s$ contribute no positive-probability acceptance under the stated restriction and are fixed inactive.

For a mean lying in component i, acceptance in A_i implies acceptance in the full union. Integrating establishes the lower-bound interpretation of Q. Additional acceptance in other components can introduce different objective curvature; no concavity claim is made for the full mixture power. The full-power optimization uses the same convex constraints, but its global optimality is not established. Independent probability validation makes any returned feasible vector valid regardless of optimization error. The fallback comparison is a numerical power safeguard, not a proof of global optimality or pointwise dominance.

A.6 A rigorous continuum route without the radius floor

For any measurable A, differentiating the Gaussian density gives

$$|p''_A(\theta)| \leq s^{-2}\mathbb{E}|Z^2 - 1| = 4\phi(1)s^{-2} < s^{-2}, \quad Z \sim N(0, 1).$$

For a twice differentiable function on $[a, b]$ with second derivative bounded in magnitude by M, the linear-interpolation remainder is at most $M(b - a)^2/8$. Replacing endpoint probabilities by outward upper bounds proves (9). Exterior rays are monotone as before because $A \subseteq S$ when radii are nonnegative. Bisection of every bounded unsafe gap preserves a complete cover, and the maximum interval upper bound remains valid even if the computational budget is exhausted. At decreasing mesh width the curvature slack vanishes. Directed CDF endpoints control arithmetic error independently of the mesh.

Our radius search maintains a feasible upper candidate. A candidate is used only if the risk upper enclosure is at most the budget. This can be slightly conservative; the search tolerance is not an optimality certificate. The control runs use radius bracket tolerance 2×10^{-5} and risk-enclosure tolerance 2×10^{-7}. For widths below two standard errors the unrestricted method can recover power that the restricted boundary class excludes. Where the floor is inactive, the methods approach the same common-radius solution; the finite-boundary result avoids the additional unsafe-space subdivision.

B Root-free verification and the numerical contract

B.1 Rational representation and range enclosure

On a closed parameter cell $[a, b]$, map $e = a + (b - a)t$ with $t \in [0, 1]$. Represent a polynomial as

$$p(t) = \sum_{k=0}^{m} c_k B_k^m(t), \qquad B_k^m(t) = \binom{m}{k} t^k (1-t)^{m-k}.$$

The basis functions are nonnegative and sum to one, so the exact coefficient range encloses every value on the closed cell. Degree elevation and addition are exact rational identities. The coefficients of multiplication by the same uncertain e are

$$d_0 = ac_0, \quad d_{m+1} = bc_m, \quad d_k = \frac{m+1-k}{m+1} ac_k + \frac{k}{m+1} bc_{k-1}, \quad 1 \le k \le m.$$

Repeated impacts therefore retain the dependence on one persistent restitution parameter rather than introducing a fresh interval-valued coefficient at each step. The de Casteljau half-cell transformation uses rational averages and exactly restricts every retained polynomial to each subcell. These are established Bernstein facts (Farouki, 2012).

B.2 Proof of Proposition 3

Initially the constant state polynomial contains the specified initial condition on the whole restitution interval. Assume every true state at a given step is represented by at least one retained continuation for its parameter. For an affine controller preactivation, coefficient bounds that are wholly positive or nonpositive justify the corresponding ReLU branch. If neither sign is established, both continuations are retained or the parameter cell is subdivided. ReLU equality causes no output disagreement, so a closed-cell sign decision is sound.

Action saturation is treated in the same way. The implementation uses the two saturated commands and the unsaturated polynomial wherever their guards might be feasible. A contact-free continuation is retained whenever $0 < f < 1$ is possible; left and right impacts are retained when $f \le 0$ and $f \ge 1$ are possible. The model explicitly assigns equality to the impact branch. The polynomial maps in (10) are exact on their respective feasible branches. Retaining algebraic extensions beyond a feasible branch only enlarges the cover. Consequently, at least one retained next state contains the true continuation for each parameter. This proves the invariant by induction.

At the terminal step, if every retained position enclosure lies within $[g - \epsilon, g + \epsilon]$, then every true terminal position does too. A range wholly outside the target can be marked unsafe; all other unresolved cells remain excluded from the positive set. A cell rejected because of a conservative enclosure need not be truly unsafe, which only reduces certification power. Splitting, a finite branch limit, or a depth limit may change the size of the returned inner set but cannot turn an unproved cell into a safe cell. Accepted cells share verified closed endpoints, so adjacent accepted cells can be merged only when their exact rational endpoints coincide. No tolerance-based gap filling occurs.

B.3 Trusted components and independent checks

Input floating-point controller coefficients are interpreted as their exact binary-rational values. The command schedule, initial state, goal, and tolerance receive the same interpretation. The GNU Multiple Precision Arithmetic Library (GMP) handles rational arithmetic; the verifier performs no floating-point sign test and does not rely on a proposed list of roots. The source implementation, compiler, GMP operations, and correspondence between this rational model and the intended

mathematical specification remain trusted components. This is not a formal proof of the complete implementation or a guarantee that deployed numerical arithmetic matches the rational model.

The default scalar validator uses depth 32 and a splitting threshold of 100,000 visited nodes. Unresolved branches at the boundary of that budget may be covered by exact polynomial continuations, with a limit of 256 simultaneous branches, or by a conservative range enclosure. Failure to establish safety returns an unresolved leaf. Pending sibling cells are still processed to preserve the domain cover; the node threshold is a stopping criterion for further splitting, not permission to discard work.

The analysis reuses the rational validations for the 63 main-cohort tasks. The additional tight-tolerance and checkpoint tasks receive independent validation with the same verifier. The tests include rational replay, interior/end-point comparisons with separate event solvers, close unsafe gaps, and isolated unsafe equality events. Such checks support implementation correctness but do not replace the inclusion proof or the stated trusted-computation contract.

B.4 From the inner set to a machine-checked risk

Each exported lower endpoint is rounded upward and each upper endpoint downward, preserving containment in the rational set. The numerical optimization proposes radii. To validate them, the implementation evaluates each Gaussian CDF with downward and upward MPFR rounding, forms lower/upper differences, sums them as exact rationals, and takes the maximum boundary upper bound. A radius vector is increased when necessary until that bound is at most the rational risk budget. The risk check retains the exact upper fraction, not just its rounded decimal display.

For a rounded observation, an external enclosure $[Y_-, Y_+]$ is sufficient if it contains the mathematical Gaussian statistic. Certification requires the entire enclosure to lie in one accepted component. The user must justify that measurement-arithmetic enclosure; assuming it is exact would not address sensor or computation bias. Optimizing a radius in ordinary arithmetic is acceptable here because feasibility is checked separately. Likewise, an unresolved numerical root is not acceptable evidence for a safe interval, which is why the positive set comes from rational validation rather than the numerical atlas.

B.5 Small holes and the branch-cover ablation

A falsely omitted unsafe point can cause a uniform-risk failure even when it has zero measure. Conversely, conservatively excluding a safe interior point can remove a substantial measurement region after erosion: an interval of radius $r = zs$ crosses that point for all means in a band of width $2r$. If no other boundaries interfere, the acceptance loss at the point is $2\Phi(z) - 1$. Thus set-length error alone does not control power error. The controlled branch-cover ablation demonstrates this effect, recovering 15.7/13.6/10.4/6.9 percentage points across the four budgets while holding the calibration radius fixed. The recovered acceptance measures the effect of verification conservatism rather than the benefit of boundary calibration.

C Measurement assumptions, conformal control, and adaptive protocols

C.1 Rebound regression and a variance-upper-bound test

Let a_i be deterministic controlled approach speeds with $S_{aa} = \sum_i (a_i - \bar{a})^2 > 0$. In model (11), ordinary least squares gives

$$Y = e - \frac{\sum_i (a_i - \bar{a})\xi_i}{S_{aa}}, \qquad s^2 = \tau^2 / S_{aa}.$$

The centered linear Gaussian contrast is independent of the orthogonal residual vector. With $\nu = n-2$, $\nu\hat{s}^2/s^2$ is χ^2_ν and is independent of Y. This proves the distribution and independence used in (12); it also shows exactly why a nonlinear ratio of two noisy velocities need not satisfy the same model.

On the event $s \le s_U$, condition on the residual scale. For every unsafe mean and every accepted observation, $|y-\theta| \ge s_U$. For $0 < t \le s_U$,

$$\partial_t f_t(y-\theta) = \left(\frac{(y-\theta)^2}{t^3} - \frac{1}{t}\right) f_t(y-\theta) \ge 0.$$

Therefore the conditional risk at the actual scale is no larger than that evaluated at s_U. Calibrating the latter to $\delta - \beta$ and paying at most β for $s_U < s$ yields total risk at most δ. Rounding s_U upward to a predeclared grid preserves the inequality. Grid overflow returns abstention. The experiment uses $\beta = 1/200$, test level $9/200$, and a grid $10^{-5}(41/40)^j$, $j = 0, \dots, 599$ in physical units, independent of Y and the true variance.

The regression control has $n \in \{4, 16, 64, 256\}$ alternating approach speeds .2 and .4, an intercept .03, and Gaussian sensor standard deviation .01. Hence $s = .1/\sqrt{n}$, matching the scale in the task-set experiments. For each budget, 20,000 independent repetitions are run at $e = 0.40 - 10^{-8}, 19/40, 0.55 + 10^{-8}$, using fixed task-safe interval $[.4, .55]$. We compare known-scale boundary calibration, the variance-upper-bound rule, and the exact Student shared interval. Monte Carlo uncertainty is summarized by standard errors. The upper-scale method is particularly conservative at four and sixteen measurements; it is a validity extension, not an empirical improvement over Student inference. The extension is proved for exact regression statistics and quantiles. Its Monte Carlo implementation uses standard floating-point chi-square and Student quantiles; unlike the known-scale boundary check, it is not an end-to-end directed-arithmetic certificate for those quantiles or the raw regression arithmetic.

C.2 Reference-residual conformal interval

Assume labeled reference estimators have residuals $E_1, \dots, E_m$, independent of the target error $E = Y - \theta$. Their common continuous distribution need not be Gaussian. To obtain uniformity for each fixed physical θ, we assume this error law is translation-invariant rather than merely exchangeable across a population of different parameters. Let $R_{(k)}$ be the kth order statistic of $|E_1|, \dots, |E_m|$ with $k = \lceil (1-\delta)(m+1) \rceil \le m$. Rank symmetry gives

$$\mathbb{P}\{|Y-\theta| \le R_{(k)}\} \ge k/(m+1) \ge 1 - \delta.$$

The interval $[Y - R_{(k)}, Y + R_{(k)}]$ is therefore a shared confidence set under this location-error assumption. It can be passed to any sound task verifier, including after task selection. Its guarantee averages over both the reference data and the target noise; conditioning on one realized reference bank changes the claim. Without translation-invariance, standard conformal coverage is generally marginal over the test population, not uniform at every unsafe physical parameter (Angelopoulos & Bates, 2023).

The control uses $m = 199$, $k = 190$. Under a continuous residual law the CDF value at the selected absolute residual has distribution $\mathrm{Beta}(190, 10)$. We integrate over this reference-bank randomness using 32-node Gauss–Jacobi quadrature. Power within each component uses 32-node Gauss–Legendre quadrature. Gaussian, unit-variance Student-t_3, and unit-variance Laplace estimator errors are evaluated on all 63 sets at $n = 4, 64$. Here n indexes the scale $.1/\sqrt{n}$ of the estimator-error experiment; non-Gaussian cases are not claimed to be exact sample means of n iid raw errors. The separate regression experiment studies actual raw Gaussian readings.

The Gaussian-task rule is not robust to every alternative error law. In particular, a Laplace error example reaches more than 5% risk at a supplied unsafe boundary. The conformal rule's proof does not rely on those boundary calculations. These values are maximum evaluated boundary risks, not numerical proofs of a continuum supremum for a new distribution. Both the additional 199 labeled residuals and the conditional-versus-marginal distinction are essential to interpreting the comparison.

C.3 Valid adaptive verification protocols

Suppose a finite or countable ladder of sound inner sets S_j is prepared without the final calibration estimate. The ladder may correspond to increasing split budgets. Prepare rule A_j with uniform unsafe risk at most δ_j, where $\sum_j \delta_j \le \delta$. For a truly failing parameter, each S_j excludes it. No

independence between the events is necessary:

$$\mathbb{P}_\theta\{\exists j : Y \in A_j\} \le \sum_j \mathbb{P}_\theta(Y \in A_j) \le \delta.$$

Hence stopping at the first certificate or choosing a level after earlier abstentions is valid. Equal spending over a fixed finite ladder is implemented with exact rational budgets. This argument does not permit arbitrarily constructing a new set around the same observed Y without a predeclared family or another uniform guarantee.

An independent pilot statistic can choose the task, verification budget, and set; conditioning on that pilot preserves the fixed-task argument for fresh calibration data. A third route is a shared confidence set: its coverage event protects any adaptive use of a sound verifier. If new observations are collected until a claim passes, a single fixed-sample interval is not sufficient in general. A confidence sequence, risk spending over sample sizes, or a justified sequential test is needed (Howard et al., 2021).

D Dimension, robust profiles, and a non-boundary alternative

D.1 A robust scalar profile with nuisance parameters

For physical parameters (e, ν) with nuisance vector known to lie in B, define

$$S_e = \{e : \forall \nu \in B,\ J_c(e, \nu) \le \epsilon^2\}.$$

Any verified inner approximation of S_e can use the scalar measurement rule for e. A false claim is impossible when e lies in that approximation and $\nu \in B$. If B is itself a statistically justified set with failure probability β, that failure must be accounted for as well; assuming the nuisance bounds are known does not identify them for free. This route trades power for robustness to nuisance parameters without asserting an exact multivariate boundary reduction.

D.2 A finite cover of a genuinely multivariate unsafe set

Let $Y \sim N_d(\theta, \Sigma)$ with known positive-definite covariance and fixed measurable acceptance region A. For two means θ, η, equal-covariance Gaussian total variation gives

$$|\mathbb{P}_\theta(Y \in A) - \mathbb{P}_\eta(Y \in A)| \le 2\Phi\left(\frac{\left\|\Sigma^{-1/2}(\theta - \eta)\right\|}{2}\right) - 1.$$

To verify the formula, whiten and rotate so the two means differ only in the first coordinate. The likelihood-ratio separating hyperplane is halfway between them, and integrating the two one-dimensional normal tails gives the expression. The bound is at most $\left\|\Sigma^{-1/2}(\theta - \eta)\right\| / \sqrt{2\pi}$.

Suppose every unsafe parameter in a known compact model domain lies within whitened distance ρ of some net point η_j. Then

$$\sup_{\theta \text{ unsafe}} \mathbb{P}_\theta(Y \in A) \le \max_j \mathbb{P}_{\eta_j}(Y \in A) + 2\Phi(\rho/2) - 1.$$

The cover must contain *all* unsafe parameters, not only their boundary. Certified Gaussian probability bounds at the net points yield a sound finite calculation. Constructing the cover and evaluating multidimensional probabilities can be expensive, with dimension-dependent covering costs. This is a path to a conservative multidimensional method, not a dimension-free extension of Theorem 1.

D.3 Why checking only the boundary fails in two dimensions

In whitened two-dimensional space, consider a safe annulus $.2 \le \|\theta\| \le 2.4$ and acceptance annulus $1.2 \le \|Y\| \le 1.4$. Every accepted point lies at least one unit from the unsafe region. Nevertheless, the probability at the unsafe center exceeds its value on either safe-set boundary. Rotational symmetry reduces the comparison to a noncentral-χ^2_2 CDF at radii 0,.2,2.4. There is also an analytic check. At mean radius t, polar integration gives

$$p(t) = \int_{1.2}^{1.4} r \exp(-(r^2 + t^2)/2) I_0(rt)\, dr.$$

The power series of the modified Bessel functions imply $I_1(x) < (x/2)I_0(x)$ for $x > 0$: the corresponding coefficient in $I_1(x)/(x/2)$ has the additional factor $1/(k+1)$. Since every integration radius satisfies $r^2 < 2$, the derivative integrand is proportional to $rI_1(rt) - tI_0(rt) < 0$. Thus $p(t)$ is strictly decreasing for $t > 0$, and the unsafe center strictly exceeds both boundary values. The recorded probabilities are 11.14%, 11.07%, and 3.21%, respectively. The scalar proof fails because the Gaussian Hessian has negative tangential directions; a boundary net alone cannot repair this failure.