CONSERVGP: CERTIFIED DIRECTIONAL COMPRESSION OF GAUSSIAN PROCESS UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Compressing a Gaussian process (GP) changes its predictive uncertainty, and the direction of the change matters: a planner needs the surrogate at least as uncertain as the teacher, a monitor no more uncertain. We present CONSERVGP, which fits a fixed-size sparse surrogate under a directional variance contract and certifies the contract over the continuous input domain, with every arithmetic step rounded outward on both models. Plain interval arithmetic loses the lower bound of the GP variance on most subintervals. The certifier instead encloses the mean coefficients through a verified linear solve and evaluates the variance in a form that keeps the cancellation between its two correlated terms. The certificate proves a signed inequality with a chosen margin, in either direction, between the teacher and the surrogate. Previous guarantees instead bound the range of one model or the size of an approximation error, or give a one-directional conservative posterior at a margin the construction fixes. In experiments, certified surrogates meet their contracts on every one-dimensional channel from synthetic data and two drone platforms, on the full domain volume of five problems with two to four inputs under certification-aware compression, and for six-output drone teachers under a matrix ordering on the joint covariance. On drone state channels the surrogate stores 54 times fewer parameters than the teacher and predicts about 34 times faster. In the selection step of safe Bayesian optimization the certified surrogate proposes only candidates inside the teacher's safe set, which unconstrained compression leaves in a fifth of the seeds. As a certified anomaly monitor it matches the teacher's discrimination at a matched false-alarm rate. For any input dimension, contracting the variance (the monitor contract) costs strictly more KL divergence than inflating it by the same margin (the envelope contract), with a gap that grows cubically at small margins and faster beyond them.

1 INTRODUCTION

Deploying a Gaussian process (GP) trained on a robot's sensor data requires compressing the full posterior into a fixed-size surrogate, because the time and memory of one GP prediction grow with the number of training observations, and a teacher of a thousand or more observations exceeds what an on-board computer can spend per query. A neural network is often compressed, for example by quantization, and then certified (Gholami et al., 2022; Ugare et al., 2024). For GPs the second step has no counterpart that proves the surrogate's variance stays on a chosen side of the teacher's, by a chosen margin. Compression changes the predictive variance, and the data the surrogate is fitted to set the direction of the change. On twenty synthetic teachers compressed to $K{=}25$ pseudo-observations (Section 2), unconstrained sparse GP regression (Snelson & Ghahramani, 2005) keeps the variance ratio σ_q/σ_p between 0.989 and 1.024, but the compressed variance is wider on about half the domain and narrower on the rest (Section 5.3). Even small deviations matter, because a contract is a signed inequality. In a safe Bayesian optimization step, an unconstrained surrogate proposed candidates outside the teacher's safe set in 4 of 20 seeds (Section 5.5). When the surrogate is more confident than the teacher (Figure 1, center), a planner that keeps to inputs whose lower predictive bound clears a safety threshold accepts inputs the teacher would have rejected. When the surrogate is less confident, a monitor that flags readings outside a predictive interval misses anomalies the teacher would have flagged.

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.

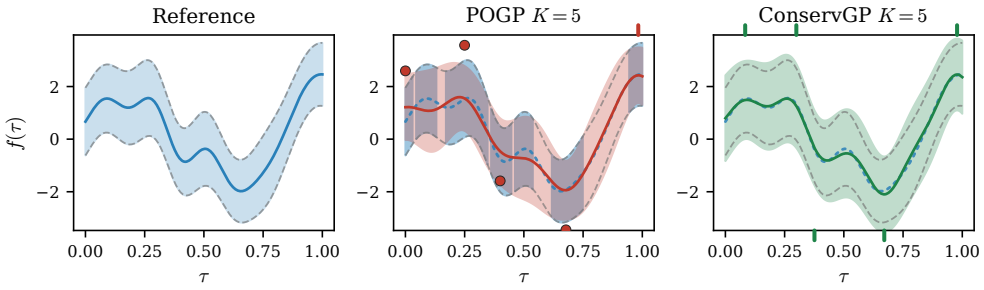

Figure 1: Predictive intervals ($\mu \pm 2\sigma$, K=5). Left: teacher Combined GP (blue). Center: unconstrained POGP, with blue shading where the compressed interval violates enclosure. Right: CONSERVGP (δ=0.15) maintains certified enclosure everywhere. Dotted blue line: teacher mean overlaid for comparison. Grey dashed lines: teacher interval bounds.

Previous GP guarantees bound either one model's prediction range, over a local region or over a tiling of a whole domain (Patane et al., 2022; Skovbekk et al., 2025) or the magnitude of the approximation error (Huggins et al., 2019; Burt et al., 2019). Several constructions produce one-directional conservative posteriors (Wenger et al., 2022; 2024; Travis & Ray, 2023; Ciosek et al., 2020) at zero margin. Capone et al. (2022) order two posteriors at a factor set by the lengthscale ratio. None of these combines a chosen margin, either direction and a whole-domain certificate (Section 6). CONSERVGP has two stages. The first fits the surrogate by minimizing the KL divergence subject to sufficient conditions that separate mean and variance. The second certifies the contract over the continuous input domain. A check on a grid gives no guarantee between its points, and a single violation anywhere invalidates the safety claim. Standard interval arithmetic (Rump, 2010) fails for GP variances because $\sigma^2 = s - \|\mathbf{v}\|^2$ is a difference of correlated quantities: it overestimates $\|\mathbf{v}\|^2$ and collapses σ^2 to zero on most subintervals. We avoid this by tiling the domain with boxes and enclosing both models' posterior means and variances on each box in outward-rounded ball arithmetic (Johansson, 2017): the mean coefficients through a verified linear solve (Rump, 2010), and the variance in a form that keeps the cancellation between its two correlated terms (Section 4.2). For more than one input the certifier halves boxes adaptively and tracks volumes as exact rational numbers. Our contributions are:

1. **Directional contracts and constrained compression.** We formalize three directional uncertainty contracts for GP compression (envelope, monitor, faithful), signed variance inequalities with a chosen margin, and fit a fixed-size surrogate by minimizing KL divergence under sufficient conditions (Lemma 1) that separate mean and variance (Sections 2 and 4.1).
2. **The cost of each direction.** We prove KL lower bounds for each direction: contraction is strictly more expensive than inflation, with the gap cubic in the margin (Theorems 1–2). We show that the optimizer closely matches these bounds in practice (Section 5.3).
3. **A whole-domain certificate.** The certifier encloses both models on a tiling of the input domain with every step rounded outward and returns a verdict that holds at every point, in either direction (Section 4.2).
4. **Certified compression at no cost in size or speed.** On synthetic teachers and drone trajectories from two platforms, with one to four inputs and one or six outputs, the certified surrogates meet their contracts at the size and speed of unconstrained ones (Sections 5.1–5.4). In a safe Bayesian optimization step the certified surrogate proposes no candidate outside the teacher's safe set, and as an anomaly detector the certified monitor keeps the teacher's detection rate at a matched false-alarm rate (Section 5.5).

2 PROBLEM SETUP AND DIRECTIONAL CONTRACTS

Teacher GP. Consider a teacher GP posterior p on $\tau \in [0, 1]$ with mean $\mu_p(\tau)$ and variance $\sigma_p^2(\tau)$. In our motivating application (autonomous robot monitoring), the teacher is a *Combined GP*: a distributional model of one coordinate of normal behavior within an *operating regime* (a class of

maneuvers sharing similar dynamics, e.g., "straight-leg cruise" or "hover-to-land transition"). The formulation applies to any GP posterior with pointwise mean and variance (Appendix L).

Compressed surrogate parameterization. For the compressed GP q we adopt a pseudo-observation parameterization (abbreviated POGP below): q is defined by K synthetic pseudo-observations $\{(\tau_k, y_k, \sigma^2_{\text{noise},k})\}_{k=1}^K$ and hyperparameters (the method applies to any stationary kernel, and the experiments use the squared-exponential with lengthscale ℓ, outputscale s, mean constant, and predictive noise). Unlike the teacher (which grows with the number of observed trajectories), this surrogate has fixed size and $O(K^2)$ prediction cost. The surrogate is the posterior of a GP prior conditioned on the K pseudo-observations with per-point noise. All parameters (τ_k, y_k, $\sigma^2_{\text{noise},k}$, kernel hyperparameters) are jointly optimized to minimize KL divergence to the teacher subject to directional constraints (Section 4.1). The resulting model has the standard GP predictive form (Rasmussen & Williams, 2006) with no dependency on the original training data:

$$\mu_q(\tau) = \mathbf{k}_*^\top \mathbf{A}^{-1}\mathbf{y}, \quad \sigma_q^2(\tau) = k(\tau,\tau) - \mathbf{k}_*^\top \mathbf{A}^{-1}\mathbf{k}_*, \tag{1}$$

where $\mathbf{A} = \mathbf{K} + \text{diag}(\boldsymbol{\sigma}^2_{\text{noise}})$. Our parameterization extends the pseudo-input model of Snelson & Ghahramani (2005) with heteroscedastic per-point noise (Kersting et al., 2007), giving $3K + 4$ parameters ($O(K)$ storage) with per-point variance control, and the certifier accepts any surrogate with closed-form mean and variance under the enclosure kernel (Appendix F).

Directional contracts. We formalize three pointwise conservatism contracts for all $\tau \in [0,1]$, where $c > 0$ controls coverage and $\delta > 0$ is the contract margin:

$$\begin{aligned} \text{Envelope:} \quad & \mu_q - c\sigma_q \le \mu_p - c\sigma_p \ \text{ and } \ \mu_q + c\sigma_q \ge \mu_p + c\sigma_p, \\ \text{Monitor:} \quad & \mu_q - c\sigma_q \ge \mu_p - c\sigma_p \ \text{ and } \ \mu_q + c\sigma_q \le \mu_p + c\sigma_p, \\ \text{Faithful:} \quad & (1-\delta)\,\sigma_p \le \sigma_q \le (1+\delta)\,\sigma_p. \end{aligned} \tag{2}$$

These differ only in inequality direction: envelope requires enclosure (safe for planning), monitor requires containment (sensitive for detection), faithful requires proximity (drop-in replacement).

Lemma 1 (Sufficient conditions for envelope conservatism)**.** *If $\sigma_q(\tau) \ge (1+\delta)\,\sigma_p(\tau)$ and $|\mu_q(\tau) - \mu_p(\tau)| \le c\,\delta\,\sigma_p(\tau)$ for all $\tau \in [0,1]$, then q satisfies the envelope contract* (2) *w.r.t. p.*

Proof. Upper enclosure: $\mu_q + c\sigma_q \ge \mu_p - c\delta\sigma_p + c(1+\delta)\sigma_p = \mu_p + c\sigma_p$. Lower enclosure: $\mu_q - c\sigma_q \le \mu_p + c\delta\sigma_p - c(1+\delta)\sigma_p = \mu_p - c\sigma_p$. □

3 The Cost of the Two Contract Directions

The theorems below bound the KL cost of each contract direction from below for a given variance gap. Burt et al. (2019) bound it from above for a given approximation. The bounds give the optimizer a benchmark and quantify the asymmetry between the two directions.

Theorem 1 (KL cost of variance inflation)**.** *If $\sigma_q(\tau_i) \ge (1+\delta)\,\sigma_p(\tau_i)$ for all evaluation points $\{\tau_i\}_{i=1}^M$, then*

$$\frac{1}{M}\sum_{i=1}^{M} \text{KL}(p_i \| q_i) \ \ge \ \Phi(\delta) \ := \ \ln(1+\delta) + \frac{1}{2(1+\delta)^2} - \frac{1}{2}\,. \tag{3}$$

Equality holds when $\mu_q = \mu_p$ and $\sigma_q = (1+\delta)\sigma_p$ everywhere.

Theorem 2 (KL cost of variance contraction)**.** *Under $\sigma_q(\tau_i) \le (1-\delta)\,\sigma_p(\tau_i)$ for $\delta \in (0,1)$:*

$$\frac{1}{M}\sum_{i=1}^{M} \text{KL}(p_i \| q_i) \ \ge \ \Psi(\delta) \ := \ \ln(1-\delta) + \frac{1}{2(1-\delta)^2} - \frac{1}{2}\,. \tag{4}$$

Both follow from monotonicity of $f(r) = \ln r + 1/(2r^2) - 1/2$, and both are attained with equality for every δ (proofs and Corollary 1 in Appendix B). The two bounds are asymmetric, with $\Psi(\delta) > \Phi(\delta)$ for all $\delta \in (0,1)$. Expanding near $\delta = 0$: $\Psi(\delta) - \Phi(\delta) = \frac{10}{3}\delta^3 + O(\delta^4)$, so the gap is cubic for small margins but grows rapidly (e.g., $\Phi = 0.018$, $\Psi = 0.030$ at $\delta = 0.15$ and $\Phi = 0.058$, $\Psi = 0.164$ at $\delta = 0.30$, a ratio of 1.7 to 2.8). As $\delta \to 1^-$, $\Psi \to \infty$ while Φ stays finite. Tightening intervals for detection costs more than widening them for safety. Section 5.3 compares the optimizer with these bounds at $\delta = 0.15$, and Table 7 across δ.

Algorithm 1 CONSERVGP compression

Require: Teacher (μ_p, σ_p) on grid $\{\tau_i\}_{i=1}^M$; margin δ; coverage c; iterations T
1: Initialize support points by interpolating teacher
2: **for** $t = 1, \ldots, T$ **do**
3: Evaluate KL + penalty (Eq. 5); gradient step (Adam + cosine annealing)
4: **if** constraints satisfied on grid **then**
5: Update best feasible state if KL improved
6: **end if**
7: **end for**
8: Warm-start SLSQP from best feasible; refine with explicit inequality constraints
9: **Return** feasible model with lowest KL

4 METHOD

CONSERVGP has two stages: constrained compression (Section 4.1) and continuous-domain verification (Section 4.2).

4.1 STAGE 1: CONSTRAINED COMPRESSION

We minimize $\mathrm{KL}(p\|q)$ on a grid $\{\tau_i\}_{i=1}^M$ subject to the sufficient conditions of Lemma 1, using a quadratic penalty with ramping weight $\lambda(t)$:

$$\min_\theta \ \frac{1}{M}\sum_{i=1}^{M} \mathrm{KL}(p_i\|q_i) + \lambda(t)\Big[\frac{1}{M}\sum_i [\max(0,\ \sigma^2_{\text{target},i} - \sigma^2_{q,i})]^2 + \frac{1}{M}\sum_i [\max(0,\ |\mu_{q,i} - \mu_{p,i}| - \epsilon_i)]^2\Big], \tag{5}$$

where $\sigma^2_{\text{target},i} = ((1+\delta)\,\sigma_{p,i})^2$ and $\epsilon_i = c\,\delta\,\sigma_{p,i}$. Equation 5 shows the *envelope* formulation. For the *monitor* contract, the variance penalty reverses direction: $[\max(0,\ \sigma^2_{q,i} - ((1-\delta)\,\sigma_{p,i})^2)]^2$ enforces an upper bound instead of a lower bound, while the mean penalty is unchanged. For the *faithful* contract, both an upper and a lower variance penalty apply simultaneously, bounding σ^2_q into the band $[((1-\delta)\,\sigma_p)^2,\ ((1+\delta)\,\sigma_p)^2]$. Algorithm 1 takes the constraint direction as a parameter, and the certifier of Section 4.2 checks both sides of the chosen inequality on every box.

Parameters $\theta = \{\tau_k, y_k, \log \sigma^2_{\text{noise},k}, \log \ell, \log s\}$ are optimized with Adam (1500 steps) then warm-started into SLSQP. Starting SLSQP from the initial parameters was unreliable in our runs because that point is deeply infeasible. The penalty warm-start places the iterate near the feasible region, and SLSQP then refines to strict feasibility.

For problems with two or more input dimensions, the optimizer targets $\sigma^2_q \geq (1+\delta+\gamma)^2\,\sigma^2_p$ instead of $(1+\delta)^2\,\sigma^2_p$, where $\gamma > 0$ gives the certifier headroom beyond the contract boundary. Appendix G reports the effect of γ on KL and coverage on Meuse 2D. The same SQP refinement applies to constrained SVGP (Appendix F).

4.2 STAGE 2: CONTINUOUS-DOMAIN CERTIFICATE

A grid check at M points can miss violations between the points. We therefore tile the input domain with boxes whose union provably covers it, enclose both models' posterior mean and variance over each box in outward-rounded ball arithmetic (Arb (Johansson, 2017), through python-flint), and read the verdict from a ball endpoint.

Interval enclosure of GP variance. The posterior variance $\sigma^2(\tau) = s - \|\mathbf{L}^{-1}\mathbf{k}_*(\tau)\|^2$ is a difference of correlated quantities, where each component of $\mathbf{v} = \mathbf{L}^{-1}\mathbf{k}_*$ depends on the same input τ. Standard interval arithmetic treats them as independent (the *dependency problem* of interval arithmetic (Rump, 2010)), overestimates $\|\mathbf{v}\|^2$, and collapses the variance lower bound to zero on all but a thin boundary layer of subintervals.

Mean enclosure. The mean coefficients $\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y}$ are enclosed as Arb balls by the verified linear solve of Rump (2010). A floating-point solution $\boldsymbol{\alpha}_0 = R\,\mathbf{y}$ is computed with a preconditioner $R \approx \mathbf{K}^{-1}$. The residual $\mathbf{r}_\mu = \mathbf{K}\boldsymbol{\alpha}_0 - \mathbf{y}$ and $\rho = \|\mathbf{I} - R\mathbf{K}\|_\infty$ are evaluated in ball arithmetic.

When $\rho < 1$, each coefficient is carried as a ball of radius $\|R\,\mathbf{r}_\mu\|_\infty/(1-\rho)$ around the centre $\boldsymbol{\alpha}_0$. For $d > 1$ the working precision is doubled until $\rho < 1$. Plain interval Gaussian elimination on the 200-point Gram matrix of one synthetic teacher fails at 128 bits and succeeds from 256 bits, with coefficient widths that shrink as the precision grows (Appendix O). The mean over a box is then the intersection of two enclosures, one from monotone bounds on each kernel entry and one from a first-order affine form (Stolfi & de Figueiredo, 2003).

Variance enclosure. The variance quadratic form $\mathbf{k}_*^\top \mathbf{K}^{-1}\mathbf{k}_*$ must not be evaluated directly on interval kernel balls, because the preconditioner $R \approx \mathbf{K}^{-1}$ has large entries on ill-conditioned Gram matrices and ball arithmetic loses the cancellation between their large positive and negative products. On the same teacher, a box of width 0.008 turns a pointwise margin of $+0.29$ into -3.8×10^3, and the affine form gives -0.63. We use the residual-corrected preconditioner identity: $\mathbf{k}^\top\mathbf{K}^{-1}\mathbf{k} = \mathbf{k}^\top R\,\mathbf{k} + \mathbf{k}^\top\mathbf{K}^{-1}\mathbf{r}_k$, where $\mathbf{r}_k = (\mathbf{I} - \mathbf{K}R)\mathbf{k}$ is the small residual. The first term applies R to the box-center kernel vector, where cancellation is preserved. The residual term is bounded in ball arithmetic (Appendix O).

Verdict. On each box the certifier composes the band margin $g = c(\sigma_q - \sigma_p) - |\mu_q - \mu_p|$ in ball arithmetic. The contract holds on the box when the lower endpoint of the margin ball is non-negative. This verdict is sound: every enclosure rounds outward on both models, so a certified box carries the contract at every point of the box for the teacher and the surrogate as fitted. Subdivision drives the enclosure widths down to a numerical limit set by the float preconditioner and the coefficient balls. In exact arithmetic, when the uniform band margin exceeds that limit, a finite subdivision certifies every box (Proposition 1, Appendix B). Rounding adds a further allowance that the proposition does not bound. The same construction certifies the envelope and monitor contracts by exchanging which model's standard deviation enters from above and which from below.

Two to four inputs. For $d > 1$ the certifier halves every axis adaptively, splitting boxes where the enclosure is too loose to decide the contract and re-verifying the sub-boxes with tighter bounds. Certified and unresolved volumes are tracked as exact rational numbers (Python `fractions.Fraction`). A reported 100% is therefore an exact quotient with no floating-point error (Appendix P).

Multi-output teachers. For a D-output teacher with joint covariance $\boldsymbol{\Sigma}_p$, the contract requires $\boldsymbol{\Sigma}_p \preceq \boldsymbol{\Sigma}_q \preceq (1+\delta)^2\,\boldsymbol{\Sigma}_p$ in the Loewner order, plus a per-channel mean band at $c = 2$. The certifier decides each positive-semidefinite test from a certified lower bound on the smallest eigenvalue of the difference matrix, computed in ball arithmetic (Appendix Q).

5 Experiments

Unless noted otherwise, synthetic experiments use Combined GP teachers ($N = 5$, RBF kernel $\ell = 0.15$, $s = 1.0$, inverse-gamma prior), $K = 25$ support points, $M = 200$ grid points. The SQP warm start runs 1500 Adam steps and then up to 200 SLSQP iterations, about 100 seconds per seed on one CPU core. The unconstrained and post-hoc baselines use 1500 Adam steps with five restarts (lowest KL kept), followed by an L-BFGS-B polish of the same objective (up to 2000 iterations). Each platform records six state coordinates (x, y, z, v_x, v_y, v_z). Flights are segmented into maneuver phases and grouped into operating regimes, and one Combined GP is built per regime and coordinate, which gives 24 Terra channels from four regimes and 18 KIOS channels from three.

5.1 Certificate Coverage

All 62 one-input channels certify (Table 1): 38 channels (20 synthetic, 18 KIOS from DJI Matrice 300 RTK flights (Grigoriou et al., 2024)) at $\delta = 0.15$, and 24 Terra channels from DJI Matrice 100 flights at $\delta = 0.20$. Margins are strictly positive on every subinterval. The total ball-arithmetic time is 11.5 hours (Table 20). Figure 2 shows the Terra altitude channel: the ConservGP envelope encloses the teacher, while unconstrained POGP deflates variance.

Five problems with two to four inputs, including the Meuse river dataset (Pebesma & Bivand, 2005) and the Ishigami function (Ishigami & Homma, 1990), certify the whole domain volume with no unresolved box (Table 2). The certificate takes three minutes at d=2 and thirteen hours at d=4, at a KL of 0.2 to 2.1. We also evaluated the floating-point posterior at random points in random boxes, and no sample fell outside its ball. The monitor direction certifies on the same 20 synthetic teachers:

Table 1: One-input suite. Both the teacher Gaussian process and its compressed surrogate are enclosed in ball arithmetic with outward rounding, and every channel is certified on the whole input interval. Margins are the smallest distance to the contract boundary over all subintervals of a channel, minimised again over the channels of a phase; both are strictly positive everywhere. Times are the rigorous-arithmetic wall-clock per channel.

			Smallest margin			Wall-clock			Teacher
Phase	Channels	Certified	upper	lower	KL range	median (s)	max (s)	total (h)	support/traj.
Synthetic	20	20	0.171	0.181	0.0180–0.0185	21	22	0.12	200
KIOS	18	18	0.257	0.057	0.0179–0.0205	163	192	0.68	317–361
Terra	24	24	0.324	0.571	0.0297–0.0330	1068	8361	10.69	280–1592
All	62	62	0.171	0.057	0.0179–0.0330	164	8361	11.49	200–1592

Table 2: Two, three and four inputs: the certified share of the input box is an exact rational number. Both Gaussian posteriors are enclosed in ball arithmetic with outward rounding over each sub-box. The sub-boxes are halved along every axis until each one resolves, and the certified and unresolved volumes are added as fractions that partition the domain exactly. The minimum margin is the smallest distance to each side of the contract among the sub-boxes that passed. Wall-clock is end-to-end on the number of worker processes given in the table.

				KL	Volume (exact)		Min. margin	Max		Wall-clock
d	Problem	K	γ	posted / meas.	certified	unresolved	(upper, lower)	depth	Sub-boxes	(worker processes)
2	Meuse	25	0.5	0.23 / 0.225	$1/1$	$0/1$	$(9.4\times10^{-6}, 2.8\times10^{-6})$	7	190 416	174 s (8w)
3	Synthetic	64	0.5	0.92 / 0.922	$1/1$	$0/1$	$(8.4\times10^{-6}, 8.0\times10^{-7})$	5	2 059 368	983 s (16w)
3	Ishigami	125	1.5	1.49 / 1.484	$1/1$	$0/1$	$(8.1\times10^{-5}, 5.3\times10^{-6})$	6	18 480 048	10.6 h (14w)
3	Meuse-3D	64	3.0	1.29 / 1.300	$1/1$	$0/1$	$(2.5\times10^{-6}, 1.1\times10^{-7})$	6	10 693 984	2.1 h (20w)
4	Synthetic	216	4.0	2.06 / 2.058	$1/1$	$0/1$	$(2.2\times10^{-6}, 3.6\times10^{-7})$	4	11 091 824	13.4 h (8w)

20 of 20 at $\delta = 0.15$, $K = 25$ (KL 0.030, matching $\Psi(0.15)$) and 20 of 20 at $\delta = 0.10$, $K = 10$ (Appendix E.5, per channel in Appendix E).

5.2 Joint Multi-Output Certificates

Ten six-output drone teachers certify the joint matrix contract at $\delta = 0.15$ over the whole horizon (Table 3, wall-clock times in Appendix J): seven Terra flights and three KIOS regimes. The joint surrogate preserves cross-channel correlation at 4 to 109 times lower joint KL than six independent scalar surrogates at the same tolerance. On flight 6, whose channels are nearly independent ($\rho_{\max} = 1.078$ against 1 for independent channels, Table 3), the best diagonal surrogate on a finite grid (Adam, 500 iterations) achieves $\delta = 0.175$, and the search found none that meets the joint contract at $\delta = 0.15$.

5.3 Certification Ablation

The full method (penalty warm-up followed by SQP, minimizing KL under the conditions of Lemma 1) certifies every seed at KL $= 0.018$ over 20 seeds, matching $\Phi(0.15)$ (Table 4). Two ablations remove the constraint and the SQP stage in turn. Unconstrained compression (minimizing KL, no directional constraint, 20 seeds) reaches KL < 0.0001 but certifies 0% on both grid and dense, with the smallest ratio σ_q/σ_p over seeds 0.989 (mean over seeds 0.993, Table 4). Low KL divergence therefore does not give a directional guarantee: the unconstrained optimum is wider than the teacher on about half the grid points and narrower on the rest. The penalty-only optimizer (no SQP) certifies 95%.

Under the same SQP refinement, constrained full-covariance SVGP also certifies every seed. The two families then differ only in single precision, where the certified variance moves by 0.11 for SVGP and by 4.2×10^{-7} for the pseudo-observation surrogate (Appendix F).

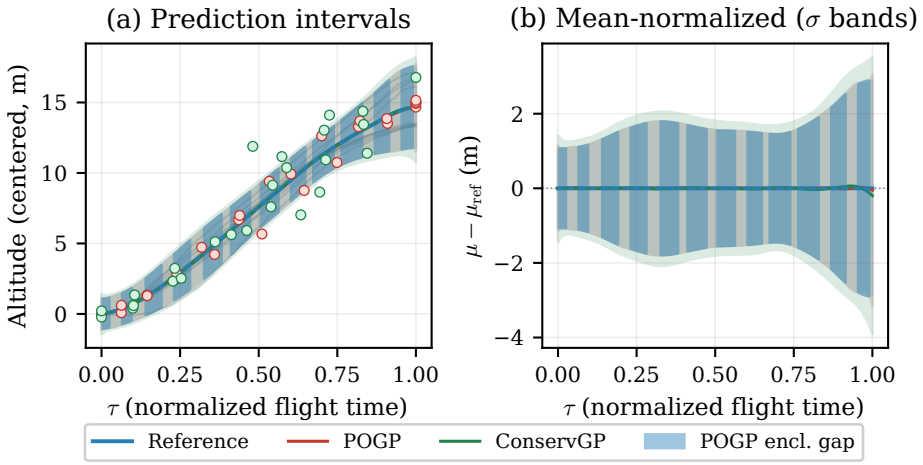

Figure 2: Terra altitude channel (K=25). **(a)** Raw intervals: the Combined GP aggregates trajectory GPs into a distributional model. **(b)** Mean-normalized: ConservGP (green) encloses teacher (blue), and POGP (red) deflates variance ($\min \sigma_q/\sigma_p = 0.959$).

Table 3: Ten six-output drone teachers, each certified against a correlation-preserving surrogate on the whole flight horizon at tolerance $\delta = 0.15$. The contract is two-sided. The surrogate covariance must sit above and below the teacher's in the matrix ordering, and its mean must stay inside a band around the teacher's. Terra rows are seven flights of one platform; KIOS rows are three regimes of a separate replayed dataset. The ratio column shows how much joint KL the correlation-preserving surrogate saves over a diagonal one at the same tolerance. $\rho_{\max}$ is the largest over the horizon of the square root of the top eigenvalue of the teacher's normalised correlation matrix, equal to 1 for independent channels.

	Flight /		Teacher	K_{cov}	Min.	Cert.	Joint KL		KL	Floats	
Dataset	regime	$\rho_{\max}$	l.scale	/ K_μ	δ	share	kept	diag.	ratio	surr. / diag.	Verdict
Terra	flight 0	1.805	0.051	30 / 30	0.030	1.000	0.0285	3.115	109.3×	868 / 552	certified
Terra	flight 1	1.418	0.020	120 / 150	0.031	1.000	0.0269	2.048	76.1×	3 719 / 2 172	certified
Terra	flight 2	1.809	0.217	30 / 30	4×10^{-6}	1.000	0.0280	1.989	71.1×	868 / 552	certified
Terra	flight 3	1.362	0.043	60 / 150	0.002	1.000	0.0279	0.314	11.2×	2 399 / 1 092	certified
Terra	flight 4	1.479	0.451	30 / 30	0.002	1.000	0.0280	1.371	49.0×	868 / 552	certified
Terra	flight 5	1.412	0.020	120 / 200	0.008	1.000	0.0280	1.516	54.1×	4 069 / 2 172	certified
Terra	flight 6	1.078	0.025	120 / 150	0.016	1.000	0.0284	0.117	4.1×	3 719 / 2 172	certified
KIOS	regime 0	1.301	0.138	30 / 30	0.086	1.000	0.0287	0.199	6.9×	868 / 552	certified
KIOS	regime 1	1.308	0.138	30 / 30	0.089	1.000	0.0289	0.207	7.2×	868 / 552	certified
KIOS	regime 2	1.132	0.183	30 / 30	0.036	1.000	0.0282	0.126	4.5×	868 / 552	certified

10 of 10 certified on the whole horizon at $\delta = 0.15$; joint KL 4.1 to 109.3 times lower than a diagonal surrogate.

5.4 Compression Benefit

The surrogate stores 54 times fewer parameters than the teacher per drone state channel and answers a query about 34 times faster, averaged over six channels and 300 repeats on one CPU with 8 threads (Table 5). The counts and timings match an unconstrained fit at the same K. The contract therefore leaves size and query time unchanged.

5.5 Downstream: Safe Exploration and Anomaly Monitoring

Safe exploration with envelope certificates. The envelope certificate verifies that the surrogate's c-band contains the teacher's, $[\mu_q \pm c\sigma_q] \supseteq [\mu_p \pm c\sigma_p]$, for all τ. SafeOpt (Sui et al., 2015) constructs a safe set from the lower predictive bounds $\mu - \beta_t^{1/2}\sigma$. Its safety guarantee assumes that the confidence bounds contain the true function. A surrogate whose variance is below the teacher's narrows those bounds and loses the teacher's coverage there.

Table 4: Certification comparison ($\delta = 0.15$, $K = 25$, $c = 2$). All rows 20 seeds (Faithful 5). Grid and Dense check the row's contract (band enclosure for Envelope and Monitor) at M and $10M$ points ($3M$ for Penalty). Ball-arithmetic certificates for the SQP rows are in Tables 1 and 11.

Contract	Method	Grid	Dense	KL	mean min σ_q/σ_p
Envelope	Unconstrained	0%	0%	< 0.0001	0.993
	Penalty	95%	95%	0.047 ± 0.078	1.146
	SQP	**100%**	**100%**	$\mathbf{0.018 \pm 0.0002}$	1.150
Monitor	SQP	**100%**	**100%**	0.030 ± 0.002	—
Faithful	SQP	**100%**	—	<0.001	—

Table 5: What the surrogate costs to store and to run, on the six real drone state channels. The teacher is the ensemble of per-flight exact Gaussian processes that the surrogate replaces. It keeps every training point and grows with the data; the surrogate is fixed at $K = 25$ support points. Prediction time is the median wall-clock to return mean and variance on a 200-point grid over 300 repeats, single CPU with 8 threads, using the same prediction routine on both sides.

	Stored floats			Retained data points			Prediction (ms / 200 pts)		
Channel	teacher	surrogate	ratio	teacher	surrogate	ratio	teacher	surrogate	speedup
x	4 228	79	53.5×	1 400	25	56.0×	2.603	0.075	34.9×
y	4 228	79	53.5×	1 400	25	56.0×	2.648	0.075	35.4×
z	4 228	79	53.5×	1 400	25	56.0×	2.621	0.074	35.4×
vx	4 228	79	53.5×	1 400	25	56.0×	2.501	0.077	32.5×
vy	4 228	79	53.5×	1 400	25	56.0×	2.443	0.074	33.0×
vz	4 228	79	53.5×	1 400	25	56.0×	2.392	0.078	30.9×
Mean	4228	79	53.5×	1400	25	56.0×	2.535	0.075	33.7×

We rank grid points by GP-UCB ($\beta^{1/2}$=2, 20 seeds) on a fixed posterior and select the 15 highest-ranked inside each model's safe set, comparing the teacher GP, the certified envelope (δ=0.15, K=5), and an unconstrained surrogate (K=5). A *violation* is a selected point outside the teacher's safe set.

In this selection step the certified envelope proposes no candidate outside the teacher's safe set over 20 seeds (Table 6). The unconstrained surrogate proposes 7 such candidates across 4 seeds (Figure 7, Appendix N). Its variance is below the teacher's on 48% of the grid, and the 7 candidates lie one to three grid cells past the teacher's boundary. The certificate's price is a smaller safe set, 24 grid points on average against the teacher's 35, and fewer than 15 candidates in 6 of 20 seeds (the teacher in 1). In a 25-step loop at K=10 that recompresses the surrogate after each observation (certified at 492 of 500 steps), the final safe set is on average within one grid point of the exact GP's (Appendix N).

Anomaly monitoring with certified contractive surrogates. The monitor contract certifies that the surrogate's band lies inside the teacher's over the whole continuous domain (Section 5.1). Every exceedance the teacher would flag is therefore also flagged by the surrogate. We tune each of the ten certified monitors, one per seed, to a 5% false-alarm rate on its own scores. Each then alarms on exactly the anomalous test trajectories on which the teacher alarms, 84% of all anomalous ones, and so does the unconstrained surrogate once tuned (Table 8, Appendix E.3). Only the certified monitor is guaranteed to keep every teacher alarm when the teacher's threshold is used unchanged, and at that threshold its tighter band, which raises every score, also adds false alarms.

6 Related Work

Approximate GPs and their uncertainty. Inducing-point methods (Titsias, 2009; Quiñonero-Candela & Rasmussen, 2005; Snelson & Ghahramani, 2005) produce the surrogates we certify. Their predictive variance can fall on either side of the teacher's (Section 5 and Wenger et al., 2024). Several methods produce conservative posteriors (Wenger et al., 2022; 2024; Travis & Ray, 2023) at zero margin. Capone et al. (2022) order two posteriors under two lengthscale vectors at a factor set by

Table 6: Safe exploration: candidates proposed outside the teacher's safe set, over 20 seeds, and the mean size of each model's safe set on the 150-point grid.

Method	Seeds w/violations	Total violations	Safe set (of 150)	Cont. certified
Teacher GP	0 / 20	0	35.45	N/A
ConservGP (envelope)	**0 / 20**	0	24.10	**100%**
Unconstrained	4 / 20	7	36.05	N/A

their ratio, and Ciosek et al. (2020) give an analogous result for neural uncertainty. Approximation-error bounds are two-sided or aggregate: Burt et al. (2019; 2020) from the KL, Huggins et al. (2019) pointwise, Calandriello et al. (2019) as a multiplicative ratio with high probability. $\mathrm{KL}(p\|q)$ penalizes a surrogate narrower than the teacher more than one wider by the same factor (Minka, 2005; Margossian et al., 2025). Theorems 1–2 give the size of that gap under a fixed variance ratio. Our contracts point in either direction and carry a margin the user sets, for a fixed-size surrogate that discards the training data, and the certificate of Section 4.2 proves them.

Verification of a GP over a region. Patane et al. (2022) bound one GP's mean and variance over a box by branch and bound, noting that a compact set can be covered by a finite union of boxes. Skovbekk et al. (2025) extend such bounds to a whole-domain tiling. Lederer et al. (2019) bound one model's mean error over a compact domain with high probability. Our certifier encloses two models on each box and reads the signed inequality from the composed margin.

Multi-output GPs and downstream uses. Bonilla et al. (2008) model multi-task GPs with a shared input kernel and free-form task covariance. In conservative data fusion, a covariance estimate is consistent when it is at least the true error covariance in the Loewner order (Noack et al., 2017). The matrix contract (Section 4.2) asks for that ordering with a margin, between two models. Safe Bayesian optimization (Sui et al., 2015) and safe model-based control (Berkenkamp et al., 2017) both rely on the pessimistic side of a confidence interval, which a deflated σ invalidates. GP-based monitoring flags exceedances of $\mu \pm c\sigma$. Tonini et al. (2026) measure coverage collapse from 90% to 5–15% at 20 to 40 inducing points on a polynomial system, and Jagdt et al. (2026) obtain probabilistic safety for a finite-feature GP without a teacher, whereas our certificate is deterministic. Conformal methods (Vovk et al., 2005) bound the gap between data and a model, while our certificate bounds the gap between two models.

7 Conclusion

We presented a constrained compression that fits a fixed-size sparse surrogate under a directional variance contract, and a certificate that proves the contract over the continuous input domain. The certificate holds on every one-input channel of the suite, on all of the volume of five problems with two to four inputs, and on ten six-output teachers. The certified surrogate is as small and as fast as an unconstrained one and keeps the teacher's safe-set or detection guarantee at no cost per query. Because the certifier subdivides the input box, its cost grows with the input dimension, from minutes at two inputs to more than half a day at four (Table 2). The largest certified dimension is four, and correlation length sets this limit because short lengths force small boxes. A 3D weather field reached 49.9997% in about 20 h, and a denser run did not finish in 34 days. Each $d > 1$ problem here is a single instance, and the tasks are small. On the Meuse two-input problem, the certifier headroom γ raises the KL from 0.041 at $\gamma{=}0$ to 0.226 at $\gamma{=}0.5$, while the verifier finds no contract violation at any setting (Appendix G). On the same problem, the exact certificate holds at $\gamma{=}0$ (KL 0.039) over the whole domain, at about five times the boxes of the $\gamma{=}0.5$ run (1.02 million against 190 thousand).

The envelope contract preserves every predictive interval of the teacher at the certificate's multiplier c, $[\mu_p \pm c\sigma_p] \subseteq [\mu_q \pm c\sigma_q]$ for all τ. SafeOpt safe sets and GP-UCB bounds that use $\beta_t^{1/2} \geq c$ therefore carry over (Appendix N).

AI USE STATEMENT

Language-model assistants (Claude, GPT), directed by the authors, were used throughout this work. They polished the writing and drafted sections from the authors' outlines and results. They retrieved related work, which the authors then checked against the cited texts. Starting from the authors' problem statement and safety framing of the two contract directions, they proposed formulations of the contracts and the certifier, wrote and ran the experiment code to the authors' specification and under author guidance, and gave first readings of the results. They proposed theorem statements and proofs, which a jury of language models checked before they reached the authors as suggestions. No data were generated by a language model. The authors have read and verified the whole paper and take responsibility for its content.

REPRODUCIBILITY STATEMENT

The anonymized archive at `https://anonymous.4open.science/r/conservgp-paper-B82F/` carries the code and the data for every result reported here. It gives the environment (a `requirements.txt` and an import check, with `python-flint`, the Python wrapper of Arb, for the outward-rounded ball arithmetic), the datasets, and a table in `REPRODUCE.md` that maps each table and figure of this paper to the script, the flags and the runtime that produce it. The seven Terra flights are in the archive. The KIOS flights are public under CC-BY 4.0 and download automatically from Zenodo (DOI 10.5281/zenodo.13682870). The Meuse dataset ships with the R `sp` package. The Ishigami function, $f(x) = \sin x_1 + a \sin^2 x_2 + b\, x_3^4 \sin x_1$ on $[-\pi, \pi]^3$ with a=7 and b=0.1 (Ishigami & Homma, 1990), is analytic and needs no data. How the drone logs become the certified channels is described in Section 5 and Appendix L, and the proofs of all theorems and propositions are in Appendix B. Every certificate in this paper was computed on CPU only, on workstations with 8 to 20 worker processes.

REFERENCES

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems*, volume 30, pp. 908–918, 2017.

Edwin V. Bonilla, Kian Ming A. Chai, and Christopher K. I. Williams. Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems*, volume 20, pp. 153–160, 2008.

David Burt, Carl Edward Rasmussen, and Mark van der Wilk. Convergence of sparse variational inference in Gaussian processes regression. *Journal of Machine Learning Research*, 21(131):1–63, 2020.

David R. Burt, Carl Edward Rasmussen, and Mark van der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *International Conference on Machine Learning*, volume 97, pp. 862–871. PMLR, 2019.

Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference on Learning Theory*, volume 99, pp. 533–557. PMLR, 2019.

Alexandre Capone, Armin Lederer, and Sandra Hirche. Gaussian process uniform error bounds with unknown hyperparameters for safety-critical applications. In *International Conference on Machine Learning*, volume 162, pp. 2609–2624. PMLR, 2022.

Kamil Ciosek, Vincent Fortuin, Ryota Tomioka, Katja Hofmann, and Richard Turner. Conservative uncertainty estimation by fitting prior networks. In *International Conference on Learning Representations*, 2020.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient Neural Network inference. In *Low-Power Computer Vision*, pp. 291–326. Chapman and Hall/CRC, 2022. doi: 10.1201/9781003162810-13.

Yiannis Grigoriou, Nicolas Souli, Panagiotis Kardaras, and Panayiotis Kolios. Drone onboard multi-modal sensor dataset for complex outdoor scenarios, 2024. CC-BY 4.0.

James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *UAI*, pp. 282–290, 2013.

Jonathan H. Huggins, Trevor Campbell, Mikolaj Kasprzak, and Tamara Broderick. Scalable Gaussian process inference with finite-data mean and variance guarantees. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pp. 796–805. PMLR, 2019.

Tsutomu Ishigami and Toshimitsu Homma. An importance quantification technique in uncertainty analysis for computer models. In *Proceedings of the First International Symposium on Uncertainty Modeling and Analysis*, pp. 398–403. IEEE, 1990. doi: 10.1109/ISUMA.1990.151285.

Julius Jagdt, Johanna Menn, Sebastian Trimpe, Melanie N. Zeilinger, and Anna Scampicchio. Scalable Gaussian process regression via deterministic trigonometric features: Uniform bounds for safe model predictive control. arXiv preprint arXiv:2608.16415, 2026.

Fredrik Johansson. Arb: Efficient arbitrary-precision midpoint-radius interval arithmetic. *IEEE Transactions on Computers*, 66(8):1281–1292, 2017. doi: 10.1109/TC.2017.2690633.

Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most likely heteroscedastic Gaussian process regression. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 393–400, 2007. doi: 10.1145/1273496.1273546.

Armin Lederer, Jonas Umlauft, and Sandra Hirche. Uniform error bounds for Gaussian process regression with application to safe control. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Charles C. Margossian, Loucas Pillaud-Vivien, and Lawrence K. Saul. Variational inference for uncertainty quantification: an analysis of trade-offs. *Journal of Machine Learning Research*, 26 (202):1–41, 2025.

Thomas Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, 2005.

Arnold Neumaier. Complete search in continuous global optimization and constraint satisfaction. *Acta Numerica*, 13:271–369, 2004. doi: 10.1017/S0962492904000194.

Benjamin Noack, Joris Sijs, Marc Reinhardt, and Uwe D. Hanebeck. Decentralized data fusion with inverse covariance intersection. *Automatica*, 79:35–41, 2017. doi: 10.1016/j.automatica.2017.01.019.

Andrea Patane, Arno Blaas, Luca Laurenti, Luca Cardelli, Stephen Roberts, and Marta Kwiatkowska. Adversarial robustness guarantees for Gaussian processes. *Journal of Machine Learning Research*, 23(146):1–55, 2022.

Edzer J. Pebesma and Roger S. Bivand. Classes and methods for spatial data in R. *R News*, 5(2): 9–13, 2005.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(65):1939–1959, 2005.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. doi: 10.7551/mitpress/3206.001.0001.

Siegfried M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449, 2010. doi: 10.1017/S096249291000005X.

John Skovbekk, Luca Laurenti, Eric Frew, and Morteza Lahijanian. Formal verification of unknown dynamical systems via Gaussian process regression. *IEEE Transactions on Automatic Control*, 70 (8):4960–4975, 2025. doi: 10.1109/TAC.2025.3532812.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, volume 18, pp. 1257–1264, 2005.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012. doi: 10.1109/TIT.2011.2182033.

Jorge Stolfi and Luiz Henrique de Figueiredo. An introduction to affine arithmetic. *TEMA (Tendências em Matemática Aplicada e Computacional)*, 4(3):297–312, 2003. doi: 10.5540/tema.2003.04.03.0297.

Yanan Sui, Alkis Gotovos, Joel W. Burdick, and Andreas Krause. Safe exploration for optimization with Gaussian processes. In *International Conference on Machine Learning (ICML)*, volume 37, pp. 997–1005. PMLR, 2015.

Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *AISTATS*, volume 5, pp. 567–574, 2009.

Stefano Tonini, Soroush Rastegarpour, Hamid Reza Feyzmahdavian, Nicola Bastianello, and Karl Henrik Johansson. Safe exploration for nonlinear processes using online Gaussian process learning. arXiv preprint arXiv:2605.09772, 2026.

Luke Travis and Kolyan Ray. Pointwise uncertainty quantification for sparse variational Gaussian process regression with a Brownian motion prior. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7419–7442, 2023.

Shubham Ugare, Tarun Suresh, Debangshu Banerjee, Gagandeep Singh, and Sasa Misailovic. Incremental randomized smoothing certification. In *International Conference on Learning Representations (ICLR)*, 2024.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Jonathan Wenger, Geoff Pleiss, Marvin Pfoertner, Philipp Hennig, and John P. Cunningham. Posterior and computational uncertainty in Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10876–10890, 2022.

Jonathan Wenger, Kaiwen Wu, Philipp Hennig, Jacob R. Gardner, Geoff Pleiss, and John P. Cunningham. Computation-aware Gaussian processes: Model selection and linear-time inference. In *Advances in Neural Information Processing Systems*, volume 37, pp. 31316–31349, 2024.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. Deep kernel learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51, pp. 370–378, 2016.

A Reproducibility and Baselines

Data and code release. The KIOS dataset is publicly available (CC-BY 4.0 (Grigoriou et al., 2024)). The Terra dataset (7 DJI Matrice 100 flights) is included in the repository. Code will be publicly released upon acceptance.

Baselines. None of the baselines below enforces a directional contract. Post-hoc noise inflation raises the mean smallest σ_q/σ_p after inflation to 1.02. The variance-domination sufficient condition holds on no seed, while the band-enclosure check passes on 3 of 10. On the same polished fits, inflating the noise by 1.15^2 moves the mean smallest ratio from 0.993 to 1.023, about 3 per cent against the 15 the condition needs. SVGP with post-hoc global variance scaling achieves grid certification at $10\times$ higher KL (0.182 vs. 0.018). The gap arises because the SVGP maximizes the ELBO, and the global factor that corrects the collapsed variance inflates the divergence. Constrained full-covariance SVGP also certifies every seed at $\mathrm{KL} = 0.018$. Mean-field SVGP (diagonal $\mathbf{S}$), which has $O(K)$ independent parameters, reaches only 30% certification under the penalty stage. Each noise term $\sigma^2_{\mathrm{noise},k}$ of the pseudo-observation surrogate acts mainly on its own support point, so the inflation can differ from point to point although the noise is diagonal (Appendix F). Full comparisons in Table 15.

B Proofs

Proof of Theorem 1.

Proof. Let $r_i = \sigma_q(\tau_i)/\sigma_p(\tau_i) \geq 1+\delta$. Then $\mathrm{KL}(p_i\|q_i) = \ln r_i + (\sigma_p^2 + (\mu_p - \mu_q)^2)/(2\sigma_q^2) - 1/2 \geq \ln r_i + 1/(2r_i^2) - 1/2 =: f(r_i)$. Since $f'(r) = (r^2-1)/r^3 > 0$ for $r > 1$, f is increasing on $[1,\infty)$ with $f(1) = 0$, so $f(r_i) \geq f(1+\delta) = \Phi(\delta)$. □

Proof of Theorem 2.

Proof. Same $f(r)$ with $r_i \leq 1 - \delta < 1$. Since $f'(r) < 0$ for $r < 1$, $f(r_i) \geq f(1-\delta) = \Psi(\delta)$. □

Corollary 1 (Both bounds are attained)**.** *For every $\delta \in (0,1)$, $\mu_q = \mu_p$ with $\sigma_q = (1+\delta)\sigma_p$ at every evaluation point attains equality in Theorem 1, and $\mu_q = \mu_p$ with $\sigma_q = (1-\delta)\sigma_p$ attains equality in Theorem 2.*

Proof. With $\mu_q = \mu_p$ and $\sigma_q = r\sigma_p$ at every point, $\mathrm{KL}(p_i\|q_i) = f(r)$ exactly, and $f(1+\delta) = \Phi(\delta)$, $f(1-\delta) = \Psi(\delta)$. □

Derivation of Remark 4.

Proof. $\mathrm{KL}(p\|q) = \mathrm{KL}(p\|\prod_i p_i) + \sum_i \mathrm{KL}(p_i\|q_i)$. The first term is $\mathcal{I}(p)$ by the Gaussian KL formula. $\mathcal{I}(p) \geq 0$ by Hadamard's inequality. □

The enclosure machinery follows Neumaier (2004) (interval subdivision) and Patane et al. (2022) (GP mean and variance bounds over a box).

Proposition 1 (Subdivision against the numerical floor)**.** *Let the teacher and the surrogate be RBF posteriors on a compact box, each with positive definite Gram matrix $\mathbf{K}$, symmetric float preconditioner R with $\rho = \|\mathbf{I} - R\mathbf{K}\|_\infty < 1$, and coefficient balls of radius e_α (Appendix O), with every enclosure evaluated in exact arithmetic. For a point τ, let $\ell(\tau) \geq 0$ be the amount by which the certifier's lower bound on the band margin $g = c(\sigma_q - \sigma_p) - |\mu_q - \mu_p|$, evaluated on the degenerate box $\{\tau\}$, falls short of $g(\tau)$, and let $\varepsilon_{\mathrm{num}} = \sup_\tau \ell(\tau)$. Then: (1) on every box B of diameter h the certified lower bound is at least $\inf_B g - \varepsilon_{\mathrm{num}} - C\sqrt{h}$, with C independent of B; (2) for a single GP with kernel vector $\mathbf{k}(\tau)$, the mean enclosure at τ has width at most $2e_\alpha\|\mathbf{k}\|_1$ and the variance enclosure at most $2e\,(e + \sqrt{\mathbf{k}^\top R\,\mathbf{k}})$ with $e = \|(\mathbf{I} - \mathbf{K}R)\mathbf{k}\|_2/\sqrt{\nu}$, where $0 < \nu \leq \lambda_{\min}(\mathbf{K})$ is the certified lower bound on the smallest eigenvalue that the certifier divides by, so $\varepsilon_{\mathrm{num}} = 0$ for exact solves; (3) if $\inf_\tau g(\tau) > \varepsilon_{\mathrm{num}}$, uniform subdivision certifies every box at some finite depth.*

Proof. Fix a box B of diameter h with centre τ_0 and write $\mathbf{k}(\tau) = \mathbf{k}_0 + \mathbf{J}(\tau - \tau_0) + \mathbf{r}$ with $\|\mathbf{r}\| = O(h^2)$; the RBF Hessian is bounded on the compact domain, so every constant below is uniform over boxes. The mean enclosure $\boldsymbol{\alpha}^\top \mathbf{k}$ has width $2e_\alpha \|\mathbf{k}_0\|_1 + O(h)$. For the variance put $A = \mathbf{k}^\top R\, \mathbf{k}$, $Q = \mathbf{k}^\top \mathbf{K}^{-1} \mathbf{k}$ and $\mathbf{s} = (\mathbf{I} - \mathbf{K}R)\mathbf{k}$. Then $Q - A = \mathbf{k}^\top \mathbf{K}^{-1} \mathbf{s}$, and Cauchy–Schwarz in the $\mathbf{K}^{-1}$ inner product gives $|Q - A| \le \|\mathbf{s}\|_2 \sqrt{Q} / \sqrt{\lambda_{\min}(\mathbf{K})} \le e\sqrt{Q}$, the second step because $\nu \le \lambda_{\min}(\mathbf{K})$. Since $\rho < 1$, the spectrum of $R\mathbf{K}$ lies in $(0, 2)$, so $R \succ 0$ and $A \ge 0$. With affine bounds $A_- \le A \le A_+$ and $e \le \bar{e}$ over B, the quadratic inequality gives $\sqrt{Q} \le U := (\bar{e} + \sqrt{\bar{e}^2 + 4A_+})/2$ and $Q \ge A_- - U\bar{e}$, an enclosure of width $A_+ - A_- + 2U\bar{e}$. Here $A_+ - A_- = O(h)$, $\bar{e} = e(\tau_0) + O(h)$ and $U \le \bar{e} + \sqrt{A_+}$, which gives the variance width in (2) plus $O(h)$. The teacher's mean and variance are a finite average of trajectory posteriors plus the between-trajectory sum of squares (Appendix L); finite sums and squares of enclosures of width $w + O(h)$ have width $w' + O(h)$. Because $|\sqrt{a} - \sqrt{b}| \le \sqrt{|a - b|}$ for $a, b \ge 0$, an $O(h)$ change in a variance bound moves the standard deviation by $O(\sqrt{h})$, so each endpoint computed on B lies within $O(\sqrt{h})$ of the endpoint computed on $\{\tau_0\}$; with $g(\tau_0) \ge \inf_B g$ this proves (1). For (3), (1) gives a positive lower bound on every box once $C\sqrt{h} < \inf_\tau g(\tau) - \varepsilon_{\text{num}}$. □

The implemented certifier evaluates the same enclosures with outward rounding at the working precision and converts some intermediate quantities and the reported endpoints to binary64, which adds to every enclosure a width of the order of the binary64 unit roundoff that does not decrease with the working precision. The certifier's shortfall at a point therefore exceeds $\ell(\tau)$ by this rounding width, which the bounds in (2) do not cover; they bound the algebraic part of the floor.

C Extended Theory

Remark 1 (Exact c-conservatism vs. variance domination). Lemma 1 is sufficient but not necessary. At $(d, r) = (0, 1)$ (perfect mean match, no inflation), KL $= 0$ under exact c-conservatism. ConservGP enforces Lemma 1 rather than the exact condition for two reasons. First, the decomposed conditions admit separate interval-arithmetic bounding. Second, the tight lower bound $\Phi(\delta)$ enables optimizer benchmarking. The cost is extra KL relative to exact-c-conservative pairs with smaller mean error.

Proposition 2 (Teacher perturbation tolerance)**.** *If q satisfies the envelope contract w.r.t. p with gap $g_{\min} = \min_\tau [c(\sigma_q - \sigma_p) - |\mu_q - \mu_p|] > 0$, and p' satisfies $|\mu_{p'} - \mu_p| \le \varepsilon_\mu \sigma_p$ and $|\sigma_{p'} - \sigma_p| \le \varepsilon_\sigma \sigma_p$, then q is conservative w.r.t. p' whenever $(c\varepsilon_\sigma + \varepsilon_\mu)\sigma_{\max} < g_{\min}$.*

Proof. $g'(\tau) \ge g(\tau) - (c\varepsilon_\sigma + \varepsilon_\mu)\sigma_p(\tau) \ge g_{\min} - (c\varepsilon_\sigma + \varepsilon_\mu)\sigma_{\max}$. □

D Joint Multi-Channel Certificates

Our per-channel certificates address independent coordinate-wise bounds. In multi-output GP deployments (e.g., the 6D drone state), a downstream monitor may aggregate channels into a joint test statistic. Two complementary results address this:

Bonferroni composition. If each of D channels has individual false-alarm rate $\le \alpha/D$, the union bound gives a joint rate $\le \alpha$. This is conservative but distribution-free.

Proposition 3 (Expected chi-squared shift under per-channel envelope)**.** *Let the deployed monitor compute the chi-squared statistic $T = \sum_{i=1}^{D} z_i^2$ where $z_i = (x_i - \mu_i)/\sigma_i$ is the per-channel Mahalanobis distance. If the compressed model satisfies the envelope contract on every channel with mean tolerance ε (i.e., $\sigma_{q,i} \ge (1 + \delta)\sigma_{p,i}$ and $|\mu_{q,i} - \mu_{p,i}| \le \varepsilon\, \sigma_{p,i}$, with $\varepsilon \le c\delta$ per Lemma 1), then the expected shift of T under the teacher distribution satisfies*

$$\mathbb{E}[T_{\text{comp}} - T_{\text{ref}}] \;\le\; \frac{D}{(1 + \delta)^2} \big(\varepsilon^2 - \delta(2 + \delta)\big). \tag{6}$$

In particular, whenever $\varepsilon < \sqrt{\delta(2 + \delta)}$, the expected statistic under the compressed model is smaller in expectation*. This does not imply stochastic ordering in general, but indicates that the compressed model's chi-squared scores are on average lower (more permissive) than the teacher's.*

Proof. Writing $r_i = \sigma_{p,i}/\sigma_{q,i} \leq (1+\delta)^{-1}$ and $u_i = (\mu_{q,i} - \mu_{p,i})/\sigma_{p,i}$, each compressed Mahalanobis component is $z^2_{q,i} = r_i^2(z_{p,i} - u_i)^2$. Taking expectations: $\mathbb{E}[z^2_{q,i}] = r_i^2(1+u_i^2)$. Summing and using $r_i \leq (1+\delta)^{-1}$, $|u_i| \leq \varepsilon$ gives (6). □

As an illustration, for $D=6$, $\delta=0.15$, $\varepsilon=0.04$: the bound gives -1.46, indicating that per-channel envelope conservatism makes the joint chi-squared test more permissive in expectation.

E Additional Experiments

E.1 Contract Visualization and Pareto Frontier

Figure 3 shows all three contracts at δ=0.15 on a single synthetic teacher. Figure 4 plots the KL cost of each direction against δ.

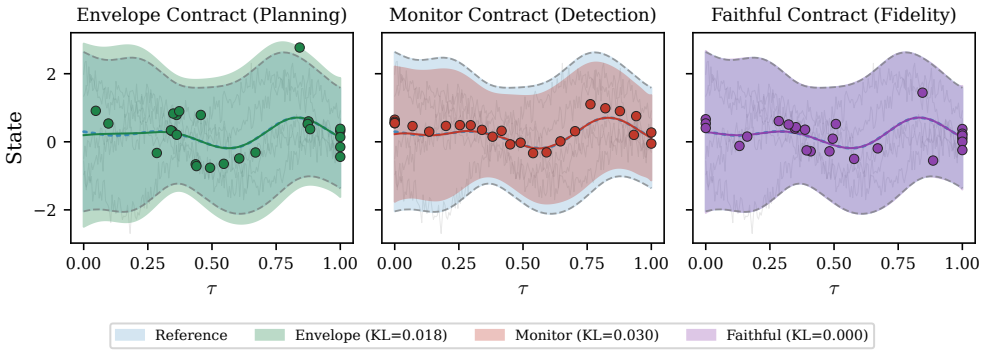

Figure 3: Three contracts at $\delta = 0.15$: envelope (wider), monitor (tighter), faithful (tracks teacher).

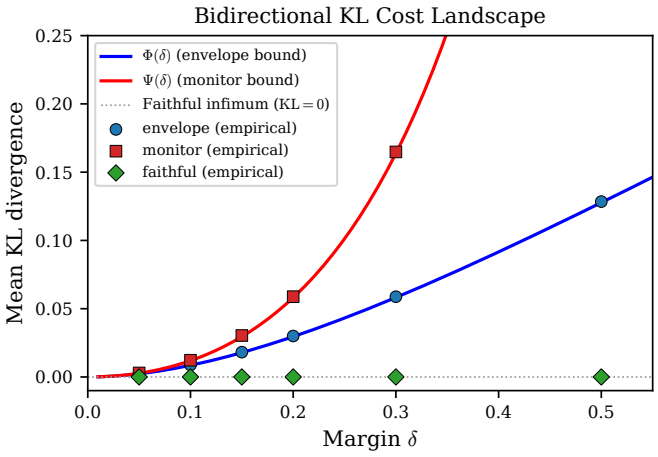

Figure 4: Bidirectional KL cost: $\Phi(\delta)$ (blue, envelope) and $\Psi(\delta)$ (red, monitor) with empirical medians (5 seeds, hybrid SQP optimizer, feasibility-filtered). Contraction is consistently more expensive.

E.2 Margin Sweep and Model Complexity

SQP achieves 100% continuous-domain certification for $\delta \leq 0.30$. At $\delta = 0.50$ (a stress test rather than a deployment target), SQP reaches only 70%.

Table 7: Margin δ vs. certification (10 seeds each).

	Penalty only			Penalty + SQP		
δ	Grid	Cont.	KL	Grid	Cont.	KL
0.10	60%	60%	0.017	**100%**	**100%**	0.009
0.15	90%	90%	0.031	**100%**	**100%**	0.018
0.20	100%	100%	0.044	**100%**	**100%**	0.030
0.30	100%	100%	0.083	**100%**	**100%**	0.059
0.50	40%	40%	0.186	80%	70%	–

Model complexity. Varying $K \in \{10, 25, 50\}$ and $N \in \{3, 5, 10\}$ (5 seeds per configuration, $\delta = 0.15$): certification is 100% across all 9 combinations, and KL stays within $\pm 7\%$ of $\Phi(0.15) = 0.018$. Theorem 1 gives the lower bound; the empirical insensitivity to K and N is consistent with it but not a consequence of it.

Detection cost. At $\delta = 0.15$ with SQP (the certified regime), the envelope contract reduces true detection rate by 6.3 pp (72.6% vs. 78.9% for the unconstrained surrogate) with 0% false alarms for both methods. Wider intervals are more permissive. The monitor contract certifies band containment over the whole domain. The surrogate therefore raises every alarm the teacher would raise.

E.3 Monitoring Detection Results

Table 8: Detection rate of the anomaly monitor at three false-alarm rates (mean $\pm$ 1 s.d. over 10 seeds, 200 normal and 200 anomalous test trajectories per seed). Each monitor's threshold is tuned on its own scores: every observed score is a candidate threshold, a trajectory raises an alarm when its score is strictly above the threshold, and each entry is the largest detection rate whose empirical false-alarm rate is at most the column's. The last column states whether the surrogate's band is certified to lie inside the teacher's over the whole horizon.

	Detection rate at a false-alarm rate of				Band inside
Monitor	0.05	0.10	0.225	Seeds	teacher's
Teacher GP	0.836 ± 0.030	0.841 ± 0.031	0.845 ± 0.031	10	—
Certified monitor (ConservGP)	0.836 ± 0.030	0.841 ± 0.031	0.845 ± 0.031	10	certified, 10/10
Unconstrained surrogate	0.836 ± 0.029	0.841 ± 0.031	0.846 ± 0.032	10	not certified

At its own 5% threshold (Table 8) the certified monitor alarms on the same 1671 of the 2000 anomalous test trajectories as the teacher, and the unconstrained surrogate on those 1671 and one more. With the teacher's 5% threshold applied unchanged, the certified monitor keeps every one of the teacher's alarms, because its tighter band raises the score of every trajectory, and the same rise adds false alarms on 1788 of the 2000 normal trajectories. The unconstrained surrogate keeps all of the teacher's alarms on this benchmark and adds 11 false alarms, but no property of the surrogate prevents a loss.

At a fixed operating point of a second detector, the certified monitor raises false alarms at rate 0.225 against the teacher's 0.059 (Table 9).

E.4 Per-Channel Certification Results

All 62 channels (38 at $\delta = 0.15$, 24 Terra at $\delta = 0.20$) certify with strictly positive margins on every subinterval. Channels with large spatial spans use up to 5000 subintervals, and all certify under the ball-arithmetic pipeline.

E.5 Monitor-Direction Certificates

At both configurations the contract holds on all 20 synthetic teachers, with KL matching the theoretical lower bound $\Psi(\delta)$.

Table 9: Anomaly detection at unequal false-alarm rates. Two operating points of the same windowed exceedance detector (a Beta posterior on the rate of readings outside the P=0.95 interval, updated with the last 10 readings) on the certified monitor (δ=0.10, K=10), the teacher GP, and an unconstrained surrogate (10 seeds, 100 normal + 100 anomalous trajectories each). The monitor's higher false-alarm rate is a consequence of its tighter bands rather than a detection gain. The unconstrained surrogate was not submitted to the certifier.

Method	TDR@1σ	TDR@1.5σ	TDR@2σ	FAR	Cont.
Teacher GP	0.260	0.747	0.984	0.059	N/A
ConservGP (monitor)	0.582	0.948	1.000	0.225	100%
POGP (unconstr., K=10)	0.260	0.749	0.984	0.058	—

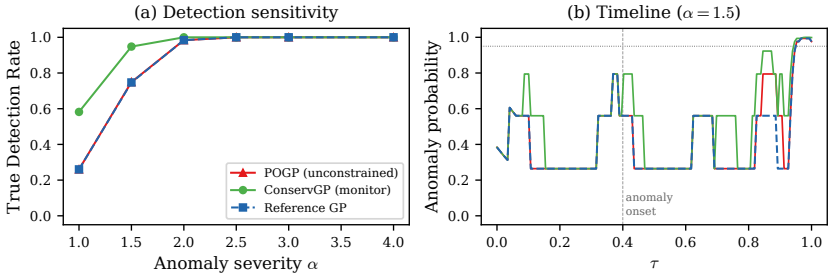

Figure 5: K=10 for all surrogates (matched capacity). Left: true detection rate vs. anomaly severity α. Right: anomaly probability timeline for a 1.5σ offset anomaly (onset at $\tau = 0.4$).

E.6 Detection Sensitivity

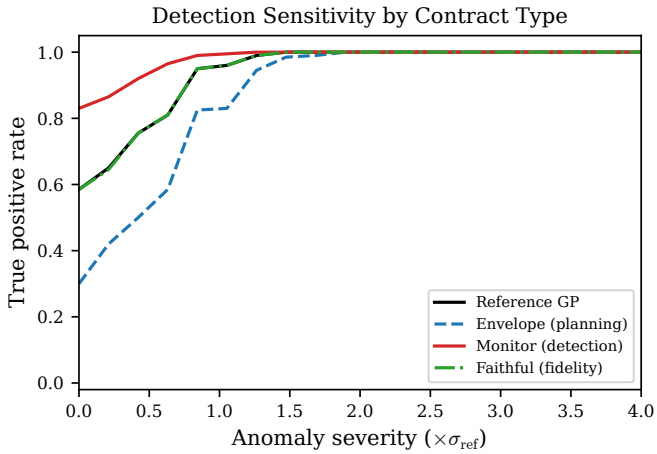

Figure 6: Detection rate vs. anomaly severity. The monitor contract certifies band containment, guaranteeing that the surrogate raises every alarm the teacher would raise.

Pooled AUROC is 0.879 for both the envelope surrogate and the teacher: wider bands reduce both false positives and mild-anomaly true positives, leaving discriminability unchanged.

Severity sweep. Table 12 injects segment anomalies at varying magnitudes (10 seeds, 50+50 per severity, $\delta = 0.15$, SQP). The conservative model's TDR converges to the teacher's for anomalies $\geq 2\sigma$.

Correlated anomalies. Table 13 evaluates three correlated anomaly types: mean-shifted GP, shorter-lengthscale GP, and drift+GP path (10 seeds, 50+50 per kind, FAR calibrated to 5%, $\delta = 0.10$).

Table 10: Every channel of the one-input suite, in full. Subintervals is the number of pieces the input interval was split into; the certificate holds on each piece as a whole, not at its endpoints. Upper and lower are the smallest distances to the two sides of the envelope contract over all subintervals.

	Traj.	Support				Smallest margin		Time	
Channel	(N)	/traj.	K	Subint.	KL	upper	lower	(s)	Certified
synthetic_seed0	5	200	25	1000	0.0185	0.210	0.214	21	yes
synthetic_seed1	5	200	25	1000	0.0183	0.192	0.193	21	yes
synthetic_seed2	5	200	25	1000	0.0184	0.202	0.196	21	yes
synthetic_seed3	5	200	25	1000	0.0180	0.171	0.186	21	yes
synthetic_seed4	5	200	25	1000	0.0181	0.191	0.198	21	yes
synthetic_seed5	5	200	25	1000	0.0182	0.200	0.199	21	yes
synthetic_seed6	5	200	25	1000	0.0182	0.218	0.209	21	yes
synthetic_seed7	5	200	25	1000	0.0185	0.220	0.216	21	yes
synthetic_seed8	5	200	25	1000	0.0181	0.186	0.194	21	yes
synthetic_seed9	5	200	25	1000	0.0181	0.209	0.181	21	yes
synthetic_seed10	5	200	25	1000	0.0184	0.207	0.199	22	yes
synthetic_seed11	5	200	25	1000	0.0181	0.204	0.201	21	yes
synthetic_seed12	5	200	25	1000	0.0182	0.216	0.214	21	yes
synthetic_seed13	5	200	25	1000	0.0183	0.234	0.237	21	yes
synthetic_seed14	5	200	25	1000	0.0182	0.198	0.199	21	yes
synthetic_seed15	5	200	25	1000	0.0181	0.256	0.255	22	yes
synthetic_seed16	5	200	25	1000	0.0182	0.196	0.204	21	yes
synthetic_seed17	5	200	25	1000	0.0181	0.197	0.203	21	yes
synthetic_seed18	5	200	25	1000	0.0184	0.181	0.192	21	yes
synthetic_seed19	5	200	25	1000	0.0184	0.208	0.204	21	yes
kios_c2_x	6	317	25	2000	0.0189	1.285	1.281	161	yes
kios_c2_y	6	317	25	2000	0.0193	1.259	1.273	164	yes
kios_c2_z	6	317	25	2000	0.0181	1.341	1.341	155	yes
kios_c2_vx	6	317	25	2000	0.0186	0.395	0.388	174	yes
kios_c2_vy	6	317	25	2000	0.0186	0.375	0.367	163	yes
kios_c2_vz	6	317	25	2000	0.0179	0.423	0.419	163	yes
kios_c1_x	6	319	25	2000	0.0190	1.256	1.231	174	yes
kios_c1_y	6	319	25	2000	0.0193	1.273	1.268	166	yes
kios_c1_z	6	319	25	2000	0.0180	1.341	1.335	168	yes
kios_c1_vx	6	319	25	2000	0.0183	0.409	0.397	168	yes
kios_c1_vy	6	319	25	2000	0.0185	0.381	0.351	154	yes
kios_c1_vz	6	319	25	2000	0.0180	0.424	0.424	162	yes
kios_c0_x	3	361	25	2000	0.0198	1.465	1.449	55	yes
kios_c0_y	3	361	25	2000	0.0195	1.401	1.391	58	yes
kios_c0_z	3	361	25	2000	0.0201	1.579	1.627	50	yes
kios_c0_vx	3	361	25	2000	0.0205	0.386	0.304	57	yes
kios_c0_vy	3	361	25	2000	0.0195	0.334	0.334	55	yes
kios_c0_vz	3	361	25	5000	0.0195	0.257	0.057	192	yes
terra_c0_x	8	280	25	2000	0.0302	2.186	2.143	484	yes
terra_c0_y	8	280	25	2000	0.0308	2.198	2.113	491	yes
terra_c0_z	8	280	25	2000	0.0298	2.178	2.139	493	yes
terra_c0_vx	8	280	25	2000	0.0300	0.709	0.700	474	yes
terra_c0_vy	8	280	25	2000	0.0302	0.588	0.657	495	yes
terra_c0_vz	8	280	25	2000	0.0298	0.710	0.684	481	yes
terra_c1_x	7	1592	25	5000	0.0330	2.205	2.392	8361	yes
terra_c1_y	7	1592	25	2000	0.0316	2.208	2.315	3331	yes
terra_c1_z	7	1592	25	2000	0.0298	2.270	2.207	3355	yes
terra_c1_vx	7	1592	25	2000	0.0307	0.324	0.571	3333	yes
terra_c1_vy	7	1592	25	2000	0.0297	0.695	0.696	2950	yes
terra_c1_vz	7	1592	25	2000	0.0297	0.714	0.713	2780	yes
terra_c2_x	4	1421	25	2000	0.0324	2.491	2.406	1744	yes
terra_c2_y	4	1421	25	2000	0.0314	2.462	2.389	1722	yes
terra_c2_z	4	1421	25	2000	0.0297	2.586	2.642	1739	yes
terra_c2_vx	4	1421	25	2000	0.0298	0.707	0.711	1654	yes
terra_c2_vy	4	1421	25	2000	0.0299	0.809	0.813	1651	yes
terra_c2_vz	4	1421	25	2000	0.0298	0.826	0.826	1641	yes
terra_c3_x	3	586	25	2000	0.0316	2.531	2.683	221	yes
terra_c3_y	3	586	25	2000	0.0319	2.660	2.546	220	yes
terra_c3_z	3	586	25	2000	0.0298	2.745	2.809	214	yes
terra_c3_vx	3	586	25	2000	0.0304	0.828	0.836	221	yes
terra_c3_vy	3	586	25	2000	0.0304	0.788	0.758	222	yes
terra_c3_vz	3	586	25	2000	0.0298	0.861	0.870	220	yes

Table 11: Monitor-direction certificates on the 20 synthetic teachers, with both models in outward-rounded ball arithmetic. The contract requires the surrogate's band to lie inside the teacher's. At the Table 1 configuration (δ=0.15, K=25) all 20 seeds certify with KL matching the theoretical lower bound $\Psi(0.15) = 0.030$. At the anomaly monitor's own configuration (δ=0.10, K=10) all 20 again certify. Margins are the smallest distance to the contract boundary, minimised over all subintervals and seeds.

		Smallest margin			Wall-clock	
Configuration	Certified	upper	lower	KL range	med. (s)	total (h)
Table 1 monitor (δ=0.15, K=25)	20/20	0.1579	0.1850	0.0298–0.0306	186	1.05
Anomaly monitor (δ=0.10, K=10)	20/20	0.0785	0.0900	0.0121–0.0134	191	1.05

Table 12: Detection rate vs. anomaly severity (mean over 10 seeds, 2σ-band detector). FPR is per-trajectory: the fraction of normal trajectories with at least one exceedance, hence high for pointwise bands (cf. the windowed exceedance detector in Table 9).

Severity	TDR ref	TDR comp	FPR ref	FPR comp
0.5σ	0.96	0.62	0.92	0.54
1.0σ	0.98	0.81	0.92	0.57
1.5σ	0.99	0.93	0.92	0.56
2.0σ	1.00	0.97	0.91	0.50
3.0σ	1.00	0.99	0.90	0.57
4.0σ	1.00	0.99	0.92	0.55

CONSERVGP matches teacher TDR within 2 pp on both GP-path models. On drift anomalies, all monitors struggle.

Table 13: Correlated anomaly detection (10 seeds, 50+50 per kind, FAR calibrated to 5%, $\delta = 0.10$, $c = 2$).

Anomaly model	Monitor	TDR (%)	FAR (%)
Mean-shift GP	Teacher	83.6	6.0
Mean-shift GP	Unconstrained	80.8	6.0
Mean-shift GP	ConservGP	83.6	6.0
Shorter lengthscale GP	Teacher	87.0	6.0
Shorter lengthscale GP	Unconstrained	86.4	6.0
Shorter lengthscale GP	ConservGP	85.2	6.0
Drift + GP path	Teacher	28.6	6.0
Drift + GP path	Unconstrained	38.6	6.0
Drift + GP path	ConservGP	35.6	6.0

E.7 SENSITIVITY TO MODEL COMPLEXITY

F POST-HOC AND SVGP BASELINES

Post-hoc inflation. Unconstrained POGP compression followed by post-hoc noise inflation (multiplying σ^2_{noise} by $(1+\delta)^2$) raises the mean smallest σ_q/σ_p after inflation to 1.02 at $\delta = 0.15$ (10 seeds) and 1.04 at $\delta = 0.30$. At both deltas the band-enclosure check passes on 3 of 10 seeds, but the variance-domination sufficient condition holds on none. The unconstrained fit (Section 5) reaches KL below 0.0001 with smallest $\sigma_q/\sigma_p = 0.989$. On the same polished fits, inflating the noise by 1.15^2 moves the mean smallest ratio from 0.993 to 1.023, about 3 per cent against the 15 the condition needs.

Table 14: K and N ablation ($\delta = 0.15$, 5 seeds).

K	N	Grid	Dense	Mean KL	Min σ_q/σ_p
10	3	100%	100%	0.019	1.150
10	5	100%	100%	0.018	1.150
10	10	100%	100%	0.019	1.150
25	3	100%	100%	0.018	1.150
25	5	100%	100%	0.018	1.150
25	10	100%	100%	0.018	1.150
50	3	100%	100%	0.018	1.150
50	5	100%	100%	0.018	1.150
50	10	100%	100%	0.018	1.150

SVGP baseline (post-hoc). SVGP (Titsias, 2009; Hensman et al., 2013) ($K = 25$, ELBO-trained, 20 seeds) collapses variance drastically ($\min \sigma_q/\sigma_p < 0.001$). Post-hoc global inflation (scaling σ_q^2 uniformly by the factor α needed so that $\sigma_q \geq (1+\delta)\sigma_p$ at every grid point, with mean $\alpha \approx 2{,}240{,}000$ across seeds) brings grid certification to 100% but with $10\times$ higher KL (0.182 vs. 0.018). The gap arises because the SVGP maximizes the ELBO, and the global factor that corrects the collapsed variance inflates the divergence. When the same SVGP architecture minimizes the pointwise KL instead, it reaches $\sigma_q/\sigma_p \approx 1.000$ before scaling and KL $= 0.018$ after a global factor of 1.32, matching CONSERVGP.

Constrained SVGP (full covariance). A natural question is whether SVGP can be directly optimized with contract penalties instead of post-hoc inflation. We apply the same quadratic penalty method (Eq. 5) to SVGP's variational parameters ($\mathbf{m} \in \mathbb{R}^K$, Cholesky factor $\mathbf{L}_S \in \mathbb{R}^{K\times K}$, plus kernel hyperparameters, giving 352 parameters in total against POGP's 79). Constrained full-covariance SVGP also certifies every seed at KL $= 0.018$ and mean $\min \sigma_q/\sigma_p = 1.149$, matching the pseudo-observation surrogate (Table 15).

Mean-field (diagonal) SVGP. Restricting to a diagonal $\mathbf{S} = \mathrm{diag}(s_1^2, \ldots, s_K^2)$ removes the coupling but does not resolve the difficulty. Diagonal SVGP has $2K + 2 = 52$ parameters ($O(K)$, comparable to POGP's 79) and each s_k^2 adjusts the variance contribution of inducing point k. However, diagonal SVGP achieves only 30% grid certification (3/10 seeds), with mean KL $= 0.102$, compared to POGP's 95%. Each s_k^2 is independent, but the mapping from s_k^2 to the predictive variance $\sigma_q^2(\tau)$ passes through $\mathbf{K}_{mm}^{-1}$, coupling the effect of every s_k^2 to every test point. In contrast, POGP's noise variance $\sigma^2_{\text{noise},k}$ enters the predictive variance through $\mathbf{A}^{-1} = (\mathbf{K} + \mathrm{diag}(\boldsymbol{\sigma}^2_{\text{noise}}))^{-1}$, where each $\sigma^2_{\text{noise},k}$ acts primarily on its own pseudo-observation. Under SQP refinement, SVGP also reaches 100% (Table 16), so the gap is one of optimization difficulty under the penalty method.

Table 15: Baseline comparison (envelope contract, $\delta = 0.15$, $K = 25$). Post-hoc inflation (row 1) passes the sampled band-enclosure check on 3 of 10 seeds but meets the variance-domination sufficient condition on none. Both full-covariance SVGP (row 3) and the pseudo-observation surrogate reach KL $= 0.018$ (the latter under SQP, Table 4). Diagonal SVGP (row 4) certifies 30%. †10 seeds, all others 20 seeds.

Method	KL_{avg}	Grid	Dense	mean min σ_q/σ_p	Params
Post-hoc inflate†	0.005	30%	30%	1.023	79
SVGP + α·var	0.182	100%	100%	1.150	352
Constrained SVGP (full $\mathbf{S}$)	0.018	100%	100%	1.149	352
Constrained SVGP (diag $\mathbf{S}$)†	0.102	30%	30%	1.155	52
POGP/CONSERVGP (penalty)	0.047	95%	95%	1.146	79

Table 16: The same constrained fit, the same optimiser and the same certification routine, applied to a sparse variational Gaussian process and to our surrogate, over 20 seeds. Under the penalty stage alone our parameterisation reaches the certified region on 19 of 20 seeds and the variational one on 3 of 20. Given the same sequential quadratic programming refinement, the variational model also certifies on every seed, so the gap is one of optimisation difficulty, not of what the two families can represent. What does not close is numerical robustness: the variational certificate rests on inverting a dense inducing matrix whose condition number is about 10^{17}, so its certified predictive variance moves by about 0.11 between single and double precision, five to six orders of magnitude more than ours. Since the refinement drives the margin exactly onto the contract boundary (smallest ratio 1.150), a shift of that size is enough to break the variational certificate in a single-precision pipeline.

				Certified		Smallest	Single-prec.	Fit	
Model	Optimiser	Params	Mean KL	grid	dense	σ ratio	shift	(s)	Seeds
Sparse variational GP (full cov.)	penalty only	352	0.0243	15%	15%	0.875	1.6×10^{-1}	1.7	20
Sparse variational GP (full cov.)	penalty + SQP	352	0.0179	100%	100%	1.150	1.1×10^{-1}	21.0	20
ConservGP	penalty only	79	0.0469	95%	95%	1.146	8.6×10^{-7}	1.1	20
ConservGP	penalty + SQP	79	0.0183	100%	100%	1.150	4.2×10^{-7}	19.0	20

Table 17: Compression target sweep on the Meuse two-input problem. The target γ is the extra variance the compressor is asked to carry above what the contract strictly needs. Across the whole range the verifier finds no contract violation, while the cost in KL rises by a factor of five.

γ	KL	Grid envelope (%)	Largest violation (upper, lower)	Verified sub-boxes	Compress (s)	Verify (s)
0.0	0.0414	100.00	(0.0, 0.0)	173 528	12	124
0.1	0.0755	100.00	(0.0, 0.0)	56 568	13	71
0.2	0.1066	100.00	(0.0, 0.0)	34 664	12	63
0.3	0.1462	100.00	(0.0, 0.0)	25 336	12	54
0.4	0.1861	100.00	(0.0, 0.0)	19 072	21	54
0.5	0.2257	100.00	(0.0, 0.0)	14 936	20	73

G Compression Target Sweep

Turning the compression target γ up from 0.0 to 0.5 on Meuse 2D costs a factor of five in KL. The grid envelope passes at 100% for every γ, and the verifier finds no violation at any setting.

H KL vs. Tolerance Sweep

Table 18 sweeps the compression target γ against the contract tolerance δ for three and four inputs, reporting the surrogate's KL and the smallest containment margin.

I Kernel Comparison

Smoother kernels are harder to certify under tight margins: RBF lags Matérn variants in certification rates despite similar KL. Matérn-3/2 achieves the lowest KL and highest certification, suggesting more flexible kernels better match the synthetic teacher under tight δ.

J Computational Cost

Compression with the penalty method takes about 1.4 s (1000 Adam steps, $K = 25$, $M = 200$). SQP refinement (up to 200 SLSQP iterations over the $3K+2$ optimized coordinates) raises the compression time to about 98 s. Two-sided interval verification ($N_{\text{sub}} = 2000$) takes about 16 s, dominated by evaluating the teacher bounds. The full pipeline (SQP compress plus verify) averages about 115 s per seed on a single CPU core.

Each six-output joint certificate of Table 3 took between 6 minutes and 7.8 hours of wall-clock time.

Table 18: KL and sampled containment margin as the contract tolerance is loosened, for three and four inputs. Each cell is the teacher-to-surrogate KL of the compressed model and, in brackets, the smallest sampled margin from 200,000 Sobol points and local searches (not a continuous-domain certificate). The pre-fit rows reuse heavily inflated models and only exist at one tolerance; their KL is an upper bound.

			Contract tolerance δ				
d	γ	Source	0.10	0.15	0.20	0.25	0.30
3	0.25	fresh	1.038 (+0.048)	0.860 (+0.030)	0.735 (+0.025)	0.752 (+0.027)	0.687 (+0.016)
3	0.50	fresh	infeas.	1.018 (+0.058)	0.910 (+0.053)	1.307 (+0.082)	0.792 (+0.026)
3	1.00	fresh	infeas.	1.258 (+0.100)	1.135 (+0.083)	1.013 (+0.053)	1.014 (+0.055)
3	1.50	fresh	1.644 (+0.180)	1.413 (+0.123)	1.314 (+0.107)	1.230 (+0.084)	1.168 (+0.077)
3	2.00	fresh	infeas.	1.624 (+0.164)	1.523 (+0.144)	1.390 (+0.102)	1.315 (+0.076)
4	0.50	fresh	0.930 (+0.124)	0.888 (+0.099)	0.857 (+0.123)	0.827 (+0.075)	0.783 (+0.073)
4	1.00	fresh	1.217 (+0.192)	1.152 (+0.191)	1.071 (+0.174)	1.040 (+0.159)	1.000 (+0.146)
4	2.00	fresh	1.556 (+0.371)	1.535 (+0.304)	1.444 (+0.341)	1.423 (+0.275)	1.369 (+0.292)
4	3.00	fresh	1.885 (+0.514)	1.787 (+0.434)	1.693 (+0.468)	1.697 (+0.401)	1.634 (+0.429)
4	3.00	pre-fit	—	1.834 (+0.527)	—	—	—
4	4.00	pre-fit	—	2.058 (+0.800)	—	—	—
4	5.00	pre-fit	—	2.234 (+0.852)	—	—	—
4	6.00	pre-fit	—	2.450 (+1.007)	—	—	—
4	8.00	pre-fit	—	2.655 (+1.335)	—	—	—
4	10.00	pre-fit	—	2.860 (+1.762)	—	—	—
4	12.00	pre-fit	—	3.019 (+2.362)	—	—	—

Table 19: Kernel comparison ($\delta = 0.10$, $K = 25$, 10 seeds, 1000 iters).

Kernel	KL	Grid	Cont.	Compress time
RBF	0.017	60%	60%	$\sim$1.3s
Matérn-5/2	0.018	90%	90%	$\sim$2.2s
Matérn-3/2	0.015	100%	100%	$\sim$2.3s

K Deep-Kernel Teachers and 2D Grid Certification

The compression stage of ConservGP takes as input only the teacher predictions $(\mu_p(\mathbf{x}), \sigma_p^2(\mathbf{x}))$ without requiring the teacher's kernel structure. The continuous-domain certifier requires kernel enclosures for the surrogate (Appendix M). The teacher can use any kernel, including Deep Kernel Learning (DKL) (Wilson et al., 2016).

We demonstrate on the Meuse 2D benchmark. A Sparse DKL ($K = 25$) teacher achieves RMSE 0.44 with 100% test coverage, and ConservGP compresses it into a $K = 25$ stationary-kernel GP while certifying the envelope contract at 97% of grid points (KL 0.052). The same pipeline applied to a stationary ARD-RBF teacher achieves 97% grid certification (KL 0.030), confirming that the compression does not depend on the teacher's kernel.

The same envelope-constrained compression, verified by standard interval arithmetic (not the adaptive exact-rational certifier of Table 2), certifies 83.1% of the Meuse benchmark area, with the contract violated at about 3% of grid points.

L Combined GP Reference Construction

Multiple observed executions of the same maneuver type are each modeled as a trajectory GP over normalized progress τ, with mean $\mu_n(\tau)$ and variance $\sigma_n^2(\tau)$ for the n-th execution, then aggregated

Table 20: What certification costs once, and what evaluation costs every time. Certifying is an offline job whose cost grows with the number of inputs, because the domain has to be split into sub-boxes and every sub-box has to be resolved; in one input it is minutes per channel, and in four inputs it is most of a day. Evaluating the certified surrogate afterwards is microsecond work that grows about linearly in the number of support points, and it does not depend on how the certificate was obtained. Two certifiers appear: one bounds both Gaussian posteriors in ball arithmetic with outward rounding, so no step relies on IEEE double precision, and one uses interval arithmetic with a Lipschitz bound in IEEE double precision, which is cheaper. The single-query figures are dominated by Python and tensor-library call overhead rather than by the linear algebra, so quote the batched figure as the throughput number.

(a) One-time certification cost

d	Problem or suite	Certifier	Resolution	Median	Total	Workers
1	Synthetic 1D suite (20 ch.)	ball arith., both sides	1000 subint./ch.	21 s/ch.	426 s	–
1	KIOS 1D suite (18 ch.)	ball arith., both sides	2000 subint./ch.	163 s/ch.	2437 s	–
1	Terra 1D suite (24 ch.)	ball arith., both sides	2000 subint./ch.	1068 s/ch.	10.7 h	–
2	Meuse	ball arith., both sides	190 416 sub-boxes	–	174 s	8
3	Synthetic	ball arith., both sides	2 059 368 sub-boxes	–	983 s	16
3	Ishigami	ball arith., both sides	18 480 048 sub-boxes	–	10.6 h	14
3	Meuse-3D	ball arith., both sides	10 693 984 sub-boxes	–	2.1 h	20
4	Synthetic	ball arith., both sides	11 091 824 sub-boxes	–	13.4 h	8
1	Real drone z	interval arith. + Lipschitz	500 subint.	–	6 s	8 thr.
1	Real drone z	interval arith. + Lipschitz	1 000 subint.	–	12 s	8 thr.
1	Real drone z	interval arith. + Lipschitz	2 000 subint.	–	24 s	8 thr.
1	Real drone z	interval arith. + Lipschitz	4 000 subint.	–	49 s	8 thr.

(b) Online evaluation of the certified surrogate

Support points K	Stored floats	Single query (μs)	Batch of 1000 (ms)	Per query (μs)
5	19	46.62	0.067	0.067
10	34	49.85	0.074	0.074
25	79	59.73	0.101	0.101
50	154	74.41	0.179	0.179
100	304	132.47	0.342	0.342

Table 21: Grid certification with stationary vs. DKL teachers (Meuse 2D, $K = 25$, $\delta = 0.15$, $c = 2$). No continuous-domain certificate was produced for these runs.

Reference	Contract	$\mathrm{KL}_{\mathrm{avg}}$	Grid	Test RMSE	Test 95%
Stationary GP	envelope	0.030	97%	0.454	66%
Stationary GP	monitor	0.048	100%	0.456	49%
DKL	envelope	0.052	97%	0.477	100%
DKL	monitor	0.131	96%	0.479	97%

across all N trajectories:

$$\mu_p(\tau) = \tfrac{1}{N}\textstyle\sum_{n=1}^{N} \mu_n(\tau), \tag{7}$$

$$\sigma_p^2(\tau) = \underbrace{\tfrac{1}{N}\textstyle\sum_{n=1}^{N} \sigma_n^2(\tau)}_{\text{within-trajectory}} + \underbrace{\frac{\beta_0 + \frac{1}{2}\sum_n (\mu_n(\tau) - \mu_p(\tau))^2}{\alpha_0 + N/2 - 1}}_{\text{between-trajectory}}. \tag{8}$$

The within-trajectory term captures measurement noise and per-execution uncertainty. The between-trajectory term (an Inverse-Gamma posterior with α_0=3, β_0=2) captures variability *across* executions of the same maneuver. The Combined GP thus encodes both "how uncertain is a single flight?" and "how much do flights of this type differ from each other?"

The Combined GP is the natural teacher for trajectory monitoring, but Sections 3–4.2 apply to arbitrary teacher GPs.

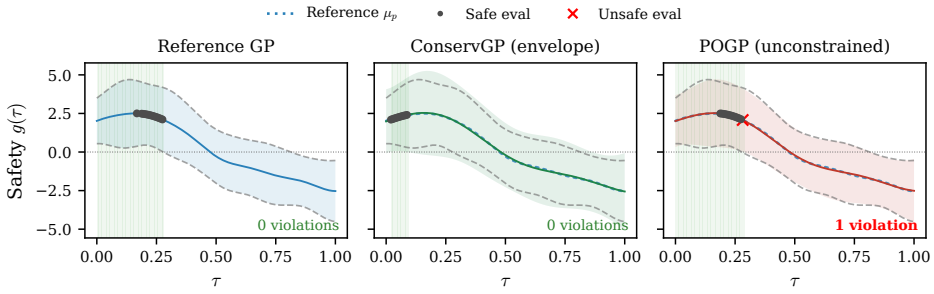

Figure 7: Safe-set regions and top 15 GP-UCB candidates (K=5, fixed posterior, seed 7). Left: teacher GP. Center: certified envelope surrogate (safe set $\subseteq$ teacher, 0 violations). Right: unconstrained POGP (safe set exceeds teacher, red ✗ marks a candidate outside the teacher's safe set).

M Framework Generality

Remark 2 (Framework generality). The certification pipeline (directional contracts, KL theory, and continuous-domain verification) applies to any surrogate whose posterior mean and variance are closed-form under the kernel the enclosures are written for (SVGP, FITC, PITC, or spectral methods with an RBF kernel). Constrained full-covariance SVGP certifies every seed at KL $= 0.018$ and ratio 1.149 (Table 15), matching the pseudo-observation surrogate. The surviving difference is numerical: SVGP's certified predictive variance moves by 0.071 (polished fits; 0.11 under SQP, Table 16) between single and double precision, against 8.1×10^{-7} for the pseudo-observation surrogate.

N Downstream Guarantee Preservation

The envelope contract preserves safe sets in SafeOpt (Sui et al., 2015) and upper confidence bounds in GP-UCB (Srinivas et al., 2012), because a surrogate whose c-band contains the teacher's at every input also keeps its safe set inside the teacher's. Section 5.5 demonstrates this on 20 fixed-posterior comparisons: in every seed the certified surrogate's candidates lie inside the teacher's safe set (Figure 7). For the monitor contract, when the means match, $\sigma_q \leq \sigma_p$ gives $z_q \geq z_p$ pointwise. Under mean error the certified band remains the sound guarantee because raw z ordering may not hold. The monitor contract enforces band containment, so the surrogate can add alarms but never miss one the teacher would raise.

Remark 3 (Downstream predictive-interval preservation). Algorithms such as GP-UCB (Srinivas et al., 2012) and SafeOpt (Sui et al., 2015) rely on their confidence intervals, the predictive intervals $[\mu \pm \beta_t^{1/2}\sigma]$ of the GP posterior, holding uniformly. Under Lemma 1's separate mean and variance premises, every teacher interval with multiplier $\beta_t^{1/2} \geq c$ is contained in the corresponding surrogate interval. The reported continuous certificate directly verifies the band at $c = 2$. Section 5.5 demonstrates this empirically. Over 20 seeds the certified surrogate proposes no candidate outside the teacher's safe set, whereas unconstrained compression does in 20% of the seeds.

A sequential loop. We also ran the safe-exploration loop. Each of 20 seeds has a hidden safety function on the 150-point grid ($2.5\cos(\pi\tau)$ plus a draw from the teacher's prior), and three arms start from the same 8 safe observations and take 25 steps, each proposing the unevaluated point of largest variance whose lower bound $\mu - 2\sigma$ is at least zero: the exact GP, the certified surrogate (K=10, δ=0.15) recompressed and recertified after every observation, and the unconstrained surrogate (K=10) recompressed after every observation. Each arm evaluated one point with a negative safety value in its 500 proposals, all three in the same seed. The certified surrogate's band contained the exact GP's at all 500 steps, on the grid and on a 15 001-point check. The continuous certificate proved this at 492 steps, and at the other 8 the loop used the surrogate without a certificate. The certified surrogate never proposed outside the exact GP's safe set. The unconstrained surrogate's band was narrower than the exact GP's somewhere at every step (ratio σ_q/σ_p down to 0.94), and it did not propose outside the exact GP's safe set either. The exact GP fitted to the data the certified loop collected ends with a safe set of 69.05 grid points on average (at most 7 fewer than the exact GP's

own loop in any seed), against 69.75 for the exact GP's own loop and 69.65 for the unconstrained loop, of 74.85 truly safe.

Remark 4 (Joint KL for factorized compression). If the compressed distribution factorizes as $q = \mathcal{N}(\mu_q, \mathrm{diag}(\sigma_q^2))$ while the teacher has full covariance Σ_p, then

$$\mathrm{KL}(p\|q) = \mathcal{I}(p) + \sum_{i=1}^{D} \mathrm{KL}(p_i\|q_i), \tag{9}$$

where $\mathcal{I}(p) = \frac{1}{2}(\sum_i \ln \sigma_p(i)^2 - \ln|\Sigma_p|) \geq 0$ is the total correlation of p, which is nonnegative by Hadamard's inequality, $|\Sigma_p| \leq \prod_i \sigma_p(i)^2$.

This adds a *constant* $\mathcal{I}(p)$ reflecting the unavoidable cost of dropping correlations, independent of the contract chosen.

O Ball-Arithmetic Enclosure Construction

This section details the enclosure construction summarised in Section 4.2. The certifier receives a trained GP (teacher or surrogate) with kernel matrix $\mathbf{K}$, observations $\mathbf{y}$, and a box $B \subset [0,1]^d$. Every matrix and vector operation below is evaluated in Arb ball arithmetic (Johansson, 2017) at working precision p bits.

Mean coefficient enclosure (Neumann bound). The mean coefficients $\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y}$ are not computed by interval Gaussian elimination, whose success depends on the working precision. On the traced 200-point Gram matrix (condition number 1.4×10^3), Gaussian elimination with partial pivoting in ball arithmetic fails at 128 bits with a singular pivot, and from 256 bits it solves with coefficient enclosure widths between 5.0×10^{-32} and 4.2×10^{-10}, shrinking to at most 3.6×10^{-87} at 512 bits and 2.7×10^{-241} at 1024 bits. Instead, a floating-point solution $\boldsymbol{\alpha}_0 = R\,\mathbf{y}$ is computed with a preconditioner $R \approx \mathbf{K}^{-1}$, and the residual $\mathbf{r}_\mu = \mathbf{K}\boldsymbol{\alpha}_0 - \mathbf{y}$ and the matrix $\rho = \|\mathbf{I} - R\mathbf{K}\|_\infty$ are evaluated in ball arithmetic. When $\rho < 1$, each coefficient is enclosed as a ball of radius $\|R\,\mathbf{r}_\mu\|_\infty/(1-\rho)$ around the floating-point centre $\boldsymbol{\alpha}_0$. For $d > 1$ the working precision is doubled until $\rho < 1$.

Mean evaluation over a box. Given the enclosed coefficient vector $\boldsymbol{\alpha}$, the mean $\mu(\tau) = \boldsymbol{\alpha}^\top \mathbf{k}_*(\tau)$ is bounded over B by two methods whose intersection is kept. First, because the squared-exponential kernel is monotone in squared distance, each kernel evaluation $k(\tau, \tau_j)$ has tight interval bounds on B and the resulting mean bound has width $2e_\alpha\|\mathbf{k}\|_1 + O(h)$ in exact arithmetic (Proposition 1). Second, a first-order affine form centred at the box midpoint provides an independent enclosure whose radius is $O(h)$, with the second-order Taylor remainder contributing $O(h^2)$.

Variance enclosure (residual-corrected preconditioner). The posterior variance involves the quadratic form $\mathbf{k}_*^\top \mathbf{K}^{-1}\mathbf{k}_*$. Direct interval evaluation of the quadratic form over a box loses the cancellation between large positive and negative products when the Gram matrix is ill-conditioned, even with the residual correction below. On the traced one-input channel at 160 bits ($\delta = 0.15$, $c = 2$), the lower bound on the band margin is $+0.29$ at the point $\tau = 0.5$ and -3.8×10^3 under direct interval evaluation on the box $[0.496, 0.504]$ of width 0.008, because the enclosed teacher standard deviation grows from 1.0 to 1.9×10^3. The first-order affine form of the kernel entries described below gives -0.63 on the same box (teacher standard deviation at most 1.3). The certifier decomposes the quadratic form as $\mathbf{k}^\top R\,\mathbf{k} + \mathbf{k}^\top \mathbf{K}^{-1}\mathbf{r}_k$ where $\mathbf{r}_k = (\mathbf{I} - \mathbf{K}R)\mathbf{k}$ is the residual and R is the float preconditioner. The first term applies a fixed matrix to the box-centre kernel vector and preserves cancellation. The residual term $\mathbf{k}^\top \mathbf{K}^{-1}\mathbf{r}_k$ is bounded in ball arithmetic. It has a fixed part set by $\|\mathbf{I} - \mathbf{K}R\|$ and an $O(h)$ box-width remainder (Proposition 1). Each kernel entry over the box is represented in first-order affine form, so that the linear dependence on the input is tracked symbolically and the overestimation from the dependency problem is confined to the second-order remainder.

Band margin and verdict. On each box the certifier computes the margin $g = c(\sigma_q - \sigma_p) - |\mu_q - \mu_p|$ in ball arithmetic, with $\sigma = \sqrt{\sigma^2}$ evaluated as an Arb square root of the variance enclosure. The

contract holds on the box when the lower endpoint of the margin ball satisfies $\underline{g} \geq 0$. For more than one input, boxes for which $\underline{g} < 0$ are split up to the depth cap and otherwise remain unresolved; for one input the partition is fixed and a box with $\underline{g} < 0$ fails the channel at that subdivision count.

Noise lower bound. The kernel diagonal includes an additive noise term σ^2_{noise}, which enters the Gram matrix as $\mathbf{K} + \sigma^2_{\text{noise}}\mathbf{I}$. The certifier subtracts the Gram rounding perturbation from the noise lower bound so that the enclosed Gram matrix is guaranteed positive-definite at the working precision.

P Exact Rational Volume Accounting

For $d > 1$ the certifier subdivides the unit hypercube adaptively, halving boxes along every axis when the enclosure is too loose. The certified and unresolved volumes are tracked as exact rational numbers (Python `fractions.Fraction`). Each initial box has volume $(1/N)^d$ for N initial divisions per axis. Each split halves one coordinate, so every sub-box volume is a power of $\frac{1}{2}$ times the parent's rational volume, and the sum of all sub-box volumes equals the domain volume exactly.

Q Matrix-Valued Certifier for Multi-Output Teachers

For a D-output teacher with joint cross-channel covariance $\mathbf{\Sigma}_p(\tau) \in \mathbb{R}^{D\times D}$ and a surrogate with covariance $\mathbf{\Sigma}_q(\tau)$, the contract requires

$$\mathbf{\Sigma}_p(\tau) \preceq \mathbf{\Sigma}_q(\tau) \preceq (1+\delta)^2\, \mathbf{\Sigma}_p(\tau), \tag{10}$$

in the Loewner (positive-semidefinite) order at every input τ, together with per-channel band containment at $c = 2$: $\mu_{q,i}(\tau) + c\,\sigma_{q,i}(\tau) \geq \mu_{p,i}(\tau) + c\,\sigma_{p,i}(\tau)$ and $\mu_{q,i}(\tau) - c\,\sigma_{q,i}(\tau) \leq \mu_{p,i}(\tau) - c\,\sigma_{p,i}(\tau)$ for each channel i.

The certifier evaluates the two positive-semidefinite conditions by computing a certified lower bound on the smallest eigenvalue of each difference matrix $\mathbf{\Sigma}_q - \mathbf{\Sigma}_p$ and $(1+\delta)^2\,\mathbf{\Sigma}_p - \mathbf{\Sigma}_q$ in Arb ball arithmetic, using Gershgorin bounds or a validated midpoint eigensolve with a radius bound. The contract holds on a sub-box when the certified lower bound on $\lambda_{\min}$ is non-negative there.

Two design choices govern the surrogate structure for multi-output teachers. First, the mean must be decoupled from the covariance and given its own dense support when the teacher's lengthscale is short (0.02 to 0.04 in normalised coordinates), because the covariance support points optimised for the variance contract are too sparse to track rapid mean oscillations. Second, the covariance support count must grow with the teacher's roughness: $K = 30$ suffices for smooth teachers, while teachers with short lengthscale and high signal variance require $K = 60$ to 120.

On a six-output drone teacher whose channels are nearly independent ($\rho_{\max} = 1.078$, the largest over the horizon of the square root of the largest eigenvalue of the teacher's normalised correlation matrix at that input, equal to 1 when the channels are independent), no diagonal (channel-independent) surrogate found by a numerical search (Adam, 500 iterations, on a finite input grid) satisfies the joint contract (10) at the paper's $\delta = 0.15$. The best observed diagonal tolerance is $\delta = 0.175$.