Certified Low-Rank Gaussian Processes via the Noise-Floor Rule

When kernel compression is safe for posterior linear solves

A certified alternative to silent low-rank GP solves

Dr. Tamás Nagy

ORCID: 0009-0004-8079-4679

Independent Researcher

tnagyphd@gmail.com

https://thalens.org

Working Paper • 2026-08-19

Build-std v2.0 | 2026-08-19 10:25 | 92500127

Overview

Gaussian-process regression is one of the cleanest tools in applied machine learning: it gives predictions and uncertainty from a single kernel model. Its computational bottleneck is equally clean. With n observations, the standard solve and log-determinant require an $n \times n$ kernel matrix and cubic linear algebra. This cost is why large Gaussian processes are routinely replaced by low-rank approximations like Nyström or inducing-point methods.

The practical problem is not just speed. The practical problem is trust. A low-rank approximation may be fast and still yield the wrong posterior solve. Usual kernel-level error measures do not, by themselves, certify that solve or the solve-induced component of mean-prediction error.

This paper provides a simple rule for certifying a low-rank Gaussian-process linear solve. The truncation error must be controlled below the observation-noise floor. If K is the kernel matrix, K_r is its low-rank replacement, and the observation variance satisfies $\sigma^2 > 0$, choosing the rank so that

$$\|K - K_r\|_F \leq \varepsilon\sigma^2$$

implies a direct relative perturbation bound on the posterior solve. The rule is deliberately conservative, but it is explicitly checkable.

The method is implemented as a certifying latent compiler. It discovers a low-rank factor, certifies the residual, accepts the replacement only if the rank is cost-beneficial, and otherwise falls back to exact inference. On a smooth RBF kernel with $n = 4000$, the compiler selects rank 239, yields an 855× speedup for the recorded fresh factorization-and-log-determinant invocation, and produces a posterior mean relative error of 2.1×10^{-5}, using the exact test–training cross-covariance vector, under a certified 5% solve bound. A rough near-full-rank kernel is rejected.

The contribution is not a new kernel trick. It is a safety rule for deciding when a kernel approximation is allowed to replace the exact posterior linear solve.

Abstract

Low-rank approximations scale Gaussian-process regression, but they are typically judged by kernel accuracy rather than posterior-solve safety; this paper introduces a noise-floor rule for certified low-rank GP linear solves. Given strictly positive observation-noise variance σ^2, a low-rank kernel replacement K_r is accepted only when its residual satisfies $\|K - K_r\|_F \le \varepsilon\sigma^2$ and the resulting rank provides a cost advantage. For a nonzero response vector, a resolvent perturbation argument then guarantees

$$\frac{\|\alpha - \hat{\alpha}\|_2}{\|\alpha\|_2} \le \frac{\|K - K_r\|_2}{\sigma^2} \le \frac{\|K - K_r\|_F}{\sigma^2} \le \varepsilon,$$

where $\alpha = (K + \sigma^2 I)^{-1} y$ and $\hat{\alpha} = (K_r + \sigma^2 I)^{-1} y$. An accompanying absolute bound covers $y = 0$. The accepted replacement is evaluated via Woodbury solves and the matrix determinant lemma. On a smooth RBF kernel with $n = 4000$, the certified method selects rank 239, delivers an 855× speedup for the recorded fresh factorization-and-log-determinant invocation, and produces a posterior mean relative error of 2.1×10^{-5}, using the exact test–training cross-covariance vector, under a certified 5% posterior-solve bound. A rough near-full-rank control is correctly rejected; timings are hardware-dependent, discovery is much slower than one exact solve in this implementation, and cached repeated-right-hand-side performance was not benchmarked.

1. Introduction

Gaussian processes are attractive because they keep two pieces of information together: a prediction and a measure of uncertainty. This is why they remain essential in Bayesian optimization, spatial statistics, active learning, and small-data modeling. Their computational obstacle is also well known. Exact Gaussian-process regression requires solving

$$(K + \sigma^2 I)\alpha = y$$

and evaluating $\log\det(K + \sigma^2 I)$, where K is the kernel matrix. The standard Cholesky route is robust but cubic in n.

The common response is to replace K by a cheaper approximation. The Nyström method is the canonical example: choose landmarks, build a low-rank surrogate, and solve a smaller problem (Williams and Seeger, 2001). Random features, inducing points, and variational sparse Gaussian processes follow the same practical instinct: keep enough structure to be useful while making the algebra cheaper. This instinct is correct. It is also incomplete.

Several lines of work have studied the quality of sparse GP approximations. Titsias's variational formulation gives a lower bound on the marginal likelihood and an implicit measure of approximation quality (Titsias, 2009). Burt, Rasmussen, and van der Wilk established convergence rates for sparse variational GP regression, showing that under regularity conditions the posterior approximation improves as the number of inducing points grows (Burt et al., 2019). These results are asymptotic: they characterize when a sparse GP will eventually become accurate. Daskalakis, Dellaportas, and Panos give finite-sample guarantees for particular random-feature and Mercer-truncation approximations, including predictive mean and covariance errors (Daskalakis et al., 2022). The present rule is narrower: it is an implementation-level acceptance gate that turns an already computed Frobenius residual and the stated observation-noise variance into a direct bound for the posterior linear solve. It is not claimed to replace those approximation guarantees.

Greengard, Rachh, and Barnett derive kernel-approximation and posterior-mean error bounds for a fast Fourier GP representation (Greengard et al., 2025). Their result is method-specific and

substantially more sophisticated. The contribution here is instead a method-agnostic, conservative runtime gate for a supplied positive-semidefinite low-rank factor. The resolvent inequality behind that gate is standard; the technical contribution is its explicit use as an accept-or-reject rule tied to the factor actually used by the Woodbury solve.

The missing question is not whether K_r is close to K. The missing question is whether inference with $K_r + \sigma^2 I$ is close to inference with $K + \sigma^2 I$. These are not the same question. When σ^2 is small, the inverse can amplify directions that look harmless at the kernel level. A one-percent kernel residual can produce a useless posterior if the discarded subspace lies below the numerical eye but above the noise floor.

This paper isolates that failure mode and turns it into a rule:

> A low-rank Gaussian-process replacement is safe only when the certified kernel residual lies below the observation-noise floor **under this paper's conservative posterior-solve criterion**.

Formally, if K_r satisfies

$$\|K - K_r\|_F \leq \varepsilon\sigma^2,$$

then the posterior solve satisfies a relative perturbation bound of order ε. The bound is conservative, but it is explicit, data-dependent, and easy to audit.

The method proposed here is a certified low-rank GP pipeline:

1. discover a low-rank factorization $K_r = WW^\top$;
2. certify the residual $\|K - K_r\|_F$;
3. accept the replacement only if the residual is below the noise floor and the rank is cost-beneficial;
4. solve and evaluate the log determinant using Woodbury and the determinant lemma;
5. reject near-full-rank kernels and fall back to exact inference.

The paper is intentionally modest. It does not claim that every Gaussian process is low-rank. It does not claim to replace inducing-point methods in all regimes. It claims that low-rank GP inference should be gated by a posterior-relevant certificate, and it gives one such gate.

Throughout this paper, "certified" means that the error bound is explicit, computable from the data, and verifiable at runtime—not that it has been machine-checked in a formal proof assistant. The certificate is a numerical inequality that can be audited by any implementation.

2. Background

2.1 Gaussian-process regression

We first fix the exact inference target that any low-rank replacement must preserve.

Let $X = (x_i)_{i=1}^n$ be training inputs and let $y \in \mathbb{R}^n$ be observations. For a positive semidefinite kernel k, write

$$K_{ij} = k(x_i, x_j).$$

With independent Gaussian observation noise of variance $\sigma^2 > 0$, the posterior mean at a test input x_* is

$$m(x_*) = k_*^\top \alpha, \qquad \alpha = (K + \sigma^2 I)^{-1} y,$$

where $k_* = (k(x_*, x_i))_{i=1}^n$. The log marginal likelihood is

$$-\frac{1}{2}y^\top (K + \sigma^2 I)^{-1} y - \frac{1}{2}\log\det(K + \sigma^2 I) - \frac{n}{2}\log(2\pi).$$

Throughout this paper, k_* is evaluated exactly and only the training linear solve is replaced. Approximating k_* would introduce a separate cross-covariance error term that is not certified here. This is the classical GPML setup (Rasmussen and Williams, 2006).

2.2 Low-rank kernel approximations

Suppose K is approximated by a rank-r positive semidefinite matrix

$$K_r = WW^\top, \qquad W \in \mathbb{R}^{n\times r}.$$

Then

$$K_r + \sigma^2 I = \sigma^2 I + WW^\top,$$

and the inverse can be applied by the Woodbury identity:

$$(\sigma^2 I + WW^\top)^{-1} = \sigma^{-2} I - \sigma^{-4} W \left(I + \sigma^{-2} W^\top W\right)^{-1} W^\top.$$

The log determinant follows from the matrix determinant lemma:

$$\log\det(\sigma^2 I + WW^\top) = n\log\sigma^2 + \log\det(I + \sigma^{-2} W^\top W).$$

The solve and log determinant therefore cost $O(nr^2 + r^3)$ rather than $O(n^3)$, after the factor W has been built.

The algebra is standard; see, for example, Golub and Van Loan for the matrix identities and numerical linear algebra background (Golub and Van Loan, 2013). The question in this paper is not the algebra. The question is when the replacement is safe.

2.3 Certified discovery

The implementation used in the experiment discovers the low-rank subspace by a randomized range finder and then computes an exact residual certificate for the produced factor. Randomized low-rank discovery is classical (Halko et al., 2011). Here it is used as a search mechanism, not as a certificate by itself.

The certificate is the relative Frobenius residual

$$\eta_r = \frac{\|K - K_r\|_F}{\|K\|_F}.$$

For symmetric positive semidefinite kernels, the best rank-r approximation is given by the leading spectral factors. The certificate supplies the absolute residual

$$e_r = \|K - K_r\|_F = \eta_r \|K\|_F.$$

The GP gate then compares e_r to σ^2.

3. The Noise-Floor Rule

The core issue is that Gaussian-process inference uses an inverse. A kernel approximation error is filtered through a resolvent. That resolvent can turn a small-looking kernel error into a large posterior error.

Assume throughout this section that K and K_r are positive semidefinite and that $\sigma^2 > 0$. Let

$$A = K + \sigma^2 I, \qquad \widehat{A} = K_r + \sigma^2 I, \qquad E = K - K_r.$$

Let

$$\alpha = A^{-1} y, \qquad \hat{\alpha} = \widehat{A}^{-1} y.$$

Since K and K_r are positive semidefinite, both A and $\widehat{A}$ have smallest eigenvalue at least σ^2. Hence

$$\|A^{-1}\|_2 \le \sigma^{-2}, \qquad \|\widehat{A}^{-1}\|_2 \le \sigma^{-2}.$$

For every response vector, the useful identity is

$$\hat{\alpha} - \alpha = \widehat{A}^{-1}(A - \widehat{A})\alpha = \widehat{A}^{-1} E \alpha.$$

It first gives the absolute bound

$$\|\hat{\alpha} - \alpha\|_2 \le \frac{\|E\|_2}{\sigma^4} \|y\|_2,$$

which also covers $y = 0$. If $y \neq 0$, invertibility of A implies $\alpha \neq 0$, so division by $\|\alpha\|_2$ is valid and

$$\frac{\|\hat{\alpha} - \alpha\|_2}{\|\alpha\|_2} \le \|\widehat{A}^{-1}\|_2 \|E\|_2 \le \frac{\|E\|_2}{\sigma^2} \le \frac{\|E\|_F}{\sigma^2}.$$

This gives the rule.

Noise-floor rule. Let $K, K_r \succeq 0$, $\sigma^2 > 0$, and $\varepsilon > 0$. Accept a low-rank GP replacement only if

$$\|K - K_r\|_F \le \varepsilon \sigma^2.$$

Then the absolute bound above holds for every y, and for $y \neq 0$ the relative error in the posterior solve is bounded by

$$\frac{\|\hat{\alpha} - \alpha\|_2}{\|\alpha\|_2} \le \varepsilon.$$

For a prediction that retains the exact k_*, the same bound immediately controls the solve-induced pointwise posterior-mean error:

$$|k_*^\top (\hat{\alpha} - \alpha)| \le \|k_*\|_2 \, \|\hat{\alpha} - \alpha\|_2.$$

The rule is stronger than a generic kernel approximation criterion. It says that the truncation scale must be calibrated to the observation noise. A fixed one-percent residual is not intrinsically safe. It may be safe when σ^2 is large and unsafe when σ^2 is small.

4. Algorithm

The certified low-rank GP algorithm is:

1. Build the kernel matrix K.
2. Run low-rank discovery with increasing rank until
$$\|K - K_r\|_F \le \varepsilon\sigma^2.$$
3. If the required rank is not cost-beneficial, reject the low-rank replacement and use exact GP inference.
4. If accepted, form $W = U_r \operatorname{diag}(\sqrt{s_r})$, where $K_r = U_r \operatorname{diag}(s_r) U_r^\top$.
5. Compute
$$\hat{\alpha} = \sigma^{-2} y - \sigma^{-4} W (I + \sigma^{-2} W^\top W)^{-1} W^\top y.$$
6. Compute
$$\log\det(\sigma^2 I + WW^\top) = n \log\sigma^2 + \log\det(I + \sigma^{-2} W^\top W).$$
7. Return the posterior quantities together with the certificate
$$\|K - K_r\|_F / \sigma^2 \le \varepsilon.$$

The algorithm has two separate costs. Discovery is a build cost. Subsequent factorization, solves, and log-determinants are query-side costs. This distinction matters, but the benchmark below times a fresh exact Cholesky factorization and a fresh low-rank solve/log-determinant invocation. It does not measure a cached-factor repeated-right-hand-side workload. Reuse across right-hand sides, online prediction, hyperparameter-local sweeps, or batched active-learning updates is a plausible deployment regime that requires a separate cached-factor benchmark.

5. Numerical Experiment

5.1 Setup

We test the method on a synthetic Gaussian-process regression problem. Training inputs are sampled in $\mathbb{R}^4$, and the kernel is the squared-exponential RBF kernel

$$k(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\ell^2}\right).$$

The response is generated from a smooth latent function plus Gaussian noise. The experiment compares:

1. exact GP inference by Cholesky factorization;
2. certified low-rank GP inference using the noise-floor rule;
3. a Nyström baseline at the same rank, without certificate;
4. a rough short-lengthscale control kernel.

The code path is the same for all accepted low-rank replacements: discover the factor, certify the residual, solve with Woodbury, and compute the log determinant by the determinant lemma.

5.2 Smooth RBF kernel

The smooth-kernel benchmark uses 4,000 observations in four input dimensions. Its RBF length-scale is 2.5, the observation-noise variance is 0.05, and the target relative solve tolerance is 5%. Under this protocol, the method selects rank 239.

Quantity	Value
n	4000
selected rank	239
certified solver-factor $\|K - K_r\|_F / \|K\|_F$ bound	1.0202×10^{-6}
certified $\|\alpha - \hat{\alpha}\| / \|\alpha\|$ bound	4.9895×10^{-2}
actual relative solve error	2.12×10^{-3}
posterior mean relative error	2.08×10^{-5}
exact GP solve/logdet time (recorded host)	2.75 s
certified low-rank solve/logdet time (recorded host)	3.21 ms
fresh factorization + solve/logdet invocation speedup (recorded host)	855×
one-shot speedup including discovery (recorded host)	0.0044×
Nyström posterior mean relative error, same rank	1.97×10^{-2}

The first number to read is not the speedup. The first number is the bound: the certified relative solve error is below 5%, and the observed error is about 0.21%. With the exact test–training cross-covariance vector retained, the posterior mean is much more accurate than the solve bound, with relative error 2.08×10^{-5}.

The comparison with Nyström is also informative. At the same rank, the random landmark Nyström approximation has posterior mean relative error 1.97×10^{-2}, roughly three orders of magnitude larger than the certified spectral replacement in this run. More importantly, it has no certificate of posterior accuracy.

A note on variance: the reported certified run fixes the randomized-discovery seed at zero and then transfers the discovery residual to the exact PSD factor used by the Woodbury solve. The table is one source-bound execution, not a claim of seed-invariant timing or rank. The Nyström baseline also depends on random landmark selection; its comparison uses one fixed seed. The key point is not that the certified method beats this particular Nyström run, but that the certified method comes with a runtime-checkable bound while the baseline does not.

The one-shot timing is intentionally reported. On the recorded host, discovery takes about 631 seconds, far longer than one exact factorization-and-solve invocation. Therefore the current prototype is not a one-off replacement for Cholesky. Its certified factor can in principle be reused across right-hand sides, test points, or online posterior queries, but the recorded 855× ratio does not measure that cached regime. A matched cached-factor benchmark is needed before making a repeated-solve speed claim. The discovery cost remains a major engineering limitation of the present prototype.

5.3 Rough kernel control

The same pipeline is run on a short-lengthscale RBF kernel with $\ell = 0.25$ and $\sigma^2 = 0.01$. The required rank is 4000, i.e. full rank, and the cost model reports no useful advantage. The method rejects the replacement and falls back to exact GP inference.

This is the behavior we want. A compiler that always returns a low-rank GP is not a safe compiler. A safe compiler must sometimes say no.

6. Discussion

The noise-floor rule changes the interpretation of low-rank GP approximation. The question is no longer "is the kernel approximately low-rank?" The question is "is the discarded kernel energy below the scale at which the posterior can resolve it?"

This distinction explains an easy failure. If one truncates a smooth kernel at a fixed relative kernel error while the observation noise is small, the posterior solve can still be badly wrong: a kernel-relative certificate does not control the amplification scale of the regularized inverse.

The noise-floor rule fixes the target. The residual must be small compared with σ^2, not merely small compared with $\|K\|_F$. This is natural: Gaussian observation noise defines the scale used by the present conservative certificate. Directions below that scale can be certified for compression by this rule. Directions above it are not certified by this rule; a tighter spectral-norm or task-local analysis may still accept them.

The method is deliberately conservative. The bound uses $\|E\|_F$ to control $\|E\|_2$, and it controls the whole posterior solve rather than a specific prediction task. Tighter bounds are possible. For example, one could use spectral-norm certificates, task-local bounds involving the test-point kernel vector, or posterior covariance bounds. The advantage of the present rule is that it is simple and hard to misuse.

7. Limitations and Non-Claims

This paper makes five explicit non-claims.

First, it does not claim that every GP kernel is low-rank. Rough kernels, short lengthscales, weak noise, and high-dimensional data can force the required rank to approach n. The method rejects those cases.

Second, it does not claim that low-rank discovery is free. In the recorded prototype benchmark, discovery cost about 230 times one exact factorization. Any practical deployment must therefore reuse the certified factor enough to amortize that cost; the present experiment does not quantify the required cached-workload break-even point.

Third, it does not claim that Nyström or inducing-point methods are obsolete. Those methods can be faster to build and may perform well empirically. The claim here is different: a solve-relevant certificate should decide whether a low-rank replacement may replace the exact posterior linear solve.

Fourth, it does not claim a sharp posterior covariance or uncertainty-calibration bound. The paper certifies the posterior linear solve and, when the exact test–training cross-covariance vector is retained, the solve-induced component of derived mean-prediction error. Cross-covariance approximation and posterior variances require separate certificates.

Fifth, "certified" here means numerically verifiable with an explicit bound—not formally verified in a theorem prover. The perturbation argument in Section 3 is a standard resolvent bound from numerical linear algebra; it is correct but not machine-checked.

8. Conclusion

Low-rank Gaussian-process linear solves should be governed by the noise floor. A kernel approximation is not solve-safe merely because it is visually accurate, fast, or standard. Under the stated positive-noise and positive-semidefinite assumptions, it is solve-safe when the discarded operator satisfies the declared noise-floor criterion.

The noise-floor rule provides a practical gate:

$$\|K - K_r\|_F \leq \varepsilon\sigma^2.$$

Together with Woodbury algebra and determinant-lemma log-determinants, this turns low-rank GP approximation into a certified factorized solver. The method accepts smooth compressible kernels, rejects near-full-rank controls, and reports a posterior-relevant error certificate.

The next step is engineering rather than ideology: wrap the method as a drop-in GP backend, add sharper spectral-norm and posterior-variance certificates, and benchmark it against inducing-point and structured-kernel methods on real Bayesian-optimization workloads.

Code Availability

A local reproducibility archive containing the certified low-rank GP implementation, the benchmark script, and the immutable result receipt has been prepared for deposit with this manuscript. The archive will be released on Zenodo together with the paper; only after that release is verified will the same bytes be mirrored in the public Thalens repository.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used Cursor and its integrated AI models (Claude and GPT families) in order to assist with manuscript drafting, structural editing, and language polishing. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Williams, C. K. I. and Seeger, M. (2001). Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems* 13, 682–688. MIT Press.

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review* 53(2):217–288.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix Computations*, 4th ed. Johns Hopkins University Press.

Titsias, Michalis (2009). Variational Learning of Inducing Variables in Sparse Gaussian Processes. *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, 5, 567-574. https://proceedings.mlr.press/v5/titsias09a.html

Burt, David; Rasmussen, Carl Edward; van der Wilk, Mark (2019). Rates of Convergence for Sparse Variational Gaussian Process Regression. *Proceedings of the 36th International Conference on Machine Learning*, 97, 862-871. https://proceedings.mlr.press/v97/burt19a.html

Daskalakis, Constantinos; Dellaportas, Petros; Panos, Aristeidis (2022). How Good Are Low-Rank Approximations in Gaussian Process Regression?. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 6463-6470. DOI: 10.1609/aaai.v36i6.20598

Greengard, Philip R.; Rachh, Manas; Barnett, Alex H. (2025). Equispaced Fourier Representations for Efficient Gaussian Process Regression from a Billion Data Points. *SIAM/ASA Journal on Uncertainty Quantification*, 13, 63-89. DOI: 10.1137/23M1565310