Introduction to Gaussian Process Regression

xxx

October 28, 2025

1 Introduction

Gaussian Process Regression (GPR) is a Bayesian nonparametric framework for learning an unknown mapping $f : \mathbb{R}^d \to \mathbb{R}$ from noisy observations while quantifying predictive uncertainty. The central modeling choice is to place a Gaussian process prior on the entire function so that the mean function and covariance kernel jointly determine all finite-dimensional laws and encode structural assumptions such as smoothness, periodicity, and invariances. Given supervised data and a simple noise model, our objective throughout is to obtain a posterior distribution over functions that characterizes f together with its uncertainty, rather than a single point estimate, thereby enabling principled decision making under limited data.

GPR underpins a range of applications in science and engineering where calibrated uncertainty is essential. In geostatistics, it specializes to kriging for spatial interpolation of environmental and subsurface fields [2]. In computer experiments, it serves as a surrogate model that supports prediction, design, and Bayesian calibration when simulations are expensive [4]. In PDE-constrained inference and inverse problems, kernels and linear operators allow data to be fused with governing equations to learn latent states or parameters [10]. In global black-box optimization, GPR models the objective while acquisition rules leverage uncertainty to balance exploration and exploitation [8]. For general foundations and methodology, we refer to Rasmussen and Williams [7].

Next we briefly state the mathematical setup of the GPR. We have $\{(x_i, y_i)\}_{i=1}^N$ with inputs $x_i \in \mathbb{R}^d$ and noisy responses modeled by

$$y_i = f(x_i) + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_{\mathrm{n}}^2)$ are independent measurement errors. The unknown mapping $f : \mathbb{R}^d \to \mathbb{R}$ is assigned a Gaussian process prior $f \sim \mathcal{GP}(m, k)$, where the mean function m and the covariance kernel k jointly determine all finite-dimensional laws and encode structural assumptions such as smoothness or periodicity. Under this Gaussian noise model, conjugacy implies that the posterior over functions is again a Gaussian process, which we denote by $f \mid \mathcal{D} \sim \mathcal{GP}(m_{\mathrm{post}}, k_{\mathrm{post}})$; our goal here is not to derive closed forms for m_{post} and k_{post} but to emphasize that this posterior process furnishes a distribution on functions that provides point predictions through its mean and principled uncertainty through its covariance, with hyperparameters either learned from data or endowed with hyperpriors to reflect domain knowledge.

2 Gaussian processes: basic theory

We now summarize the basic mathematical theory for Gaussian processes, also known as Gaussian random fields when the index set is a spatial domain. Throughout, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an index set $\mathcal{X} \subseteq \mathbb{R}^d$. A stochastic process $f = \{f(x) : x \in \mathcal{X}\}$ is called Gaussian if for every finite $X = (x_1, \ldots, x_n) \subset \mathcal{X}$ the random vector $(f(x_1), \ldots, f(x_n))$ is multivariate normal. Such a process is *uniquely* characterized by:

- **Mean function:** $m(x) = \mathbb{E}[f(x)]$;
- **Covariance function:** $k(x, x') = \mathrm{Cov}(f(x), f(x'))$.

For any finite $X = (x_1, \ldots, x_n)$ one then has

$$(f(x_1), \ldots, f(x_n)) \sim \mathcal{N}(m(X), K(X, X)),$$

where the mean vector $m(X)$ and the covariance matrix are

$$m(X) = \begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix} \in \mathbb{R}^n, \quad K(X, X) = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix} \in \mathbb{R}^{n \times n}$$

Since multivariate Gaussian laws are completely determined by first and second moments, specifying m and k specifies the finite-dimensional distributions of f [7, Ch. 2].

The preceding construction is valid whenever k is symmetric and positive semidefinite, meaning that for any finite $X = (x_1, \ldots, x_n)$ one has

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geqslant 0$$

for all $c_1, \ldots, c_n \in \mathbb{R}$. Under this consistency requirement, the Kolmogorov extension theorem guarantees the existence of a process whose finite-dimensional laws are exactly the prescribed Gaussian marginals [3, Thm. 6.16]. Regarding sample-path regularity, the Kolmogorov continuity theorem yields a version with continuous (Hölder) trajectories under suitable increment bounds; for Matérn kernels, this translates into almost-sure Hölder smoothness controlled by the kernel smoothness parameter [9].

It is useful to record several closure properties. If L is any fixed linear operator acting on functions of x (for example, differentiation or integration in the weak sense), then Lf is again a Gaussian process with mean Lm and covariance $(x, x') \mapsto L_x L_{x'} k(x, x')$, where subscripts indicate the argument of action; this follows from the preservation of Gaussianity under linear transformations. In particular, if k is sufficiently smooth, pointwise derivatives $x \mapsto \partial^\alpha f(x)$ exist in mean square and form a jointly Gaussian family, which is the basis for encoding linear constraints or constructing likelihoods from residuals in PDE-informed models [7, 10]. Conditioning a jointly Gaussian family remains Gaussian as well, and therefore posterior processes induced by Gaussian likelihoods are also Gaussian; this observation underlies the GPR formulation in the previous section.

On a compact set $D \subset \mathbb{R}^d$ with a continuous, symmetric, positive semidefinite kernel k, the associated integral operator

$$(T\varphi)(x) = \int_D k(x, x')\varphi(x')\, dx'$$

is compact and self-adjoint. By Mercer's theorem, there exist orthonormal eigenfunctions $\{\phi_j\}_{j\geqslant 1}$ in $L^2(D)$ with nonnegative eigenvalues $\{\lambda_j\}_{j\geqslant 1}$ such that

$$k(x, x') = \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\phi_j(x')$$

with convergence in $L^2(D \times D)$ and, under continuity, uniformly on $D \times D$ [6]. If f is zero-mean with covariance k, then the Karhunen–Loève expansion reads

$$f(x) = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \xi_j\, \phi_j(x),$$

where $\{\xi_j\}_{j\geqslant 1}$ are independent standard normal variables; the series converges in $L^2(\Omega \times D)$ and, under mild assumptions, uniformly in x [5, Ch. 2]. This representation clarifies that sample paths live in the closure of the span of the eigenfunctions and that truncations yield optimal mean-square low-rank approximations.

We distinguish the global index domain $\mathcal{X}$, the compact subdomain $D \subseteq \mathcal{X}$ used to invoke Mercer's theorem and the Karhunen–Loève expansion, and the finite sampling set $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ used to form finite-dimensional marginals.

When the kernel is stationary, meaning $k(x, x') = k(x - x')$, a spectral representation is available. Bochner's theorem states that a continuous function $k : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it is the Fourier transform of a finite nonnegative measure μ on $\mathbb{R}^d$, that is, $k(h) = \int_{\mathbb{R}^d} e^{i\omega\cdot h}\, \mu(d\omega)$. In this case, one may realize the process as a random Fourier integral

$$f(x) = \int_{\mathbb{R}^d} e^{i\omega\cdot x} \sqrt{S(\omega)}\, W(d\omega),$$

where S is the spectral density when μ is absolutely continuous and W is a complex Gaussian random measure with orthogonal increments [9, Chs. 2–3]. This representation explains the frequency content induced by a stationary kernel and underlies random Fourier feature approximations.

Finally, we note the connection with reproducing kernel Hilbert spaces (RKHS) and the Cameron–Martin space. Every positive semidefinite kernel k defines an RKHS $\mathcal{H}_k$ with reproducing property $\langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} = f(x)$. For a zero-mean Gaussian process with covariance k, the Cameron–Martin space is isometrically isomorphic to $\mathcal{H}_k$, and it describes the directions that yield absolutely continuous shifts of the law [1, Chs. 1–3]. This viewpoint clarifies how linear functionals act on f and provides a bridge between probabilistic modeling and deterministic approximation theory [7, Ch. 6]. We adopt these conventions throughout the remainder of this document.

3 Gaussian processes: examples and realizations

We collect several canonical Gaussian processes by specifying their covariance kernels and, for each, give representative realizations either as a Karhunen–Loève (KL) series on a compact domain or as a random Fourier integral on $\mathbb{R}^d$ when the kernel is stationary. Throughout, amplitudes and noise levels use the notational convention σ_{f} and σ_{n}, lengths are denoted by ℓ, and $r = \|x - x'\|$. The acronym ARD (Automatic Relevance Determination) means that each input coordinate is given its own length scale; algebraically one replaces the isotropic metric $\ell^2 I$ by a diagonal metric $\Lambda = \mathrm{diag}(\ell_1^2, \dots, \ell_d^2)$ so that sensitivity along each coordinate can differ.

3.1 RBF (squared exponential): periodic domains and full space

The isotropic RBF (Gaussian, squared exponential) kernel prescribes exponentially decaying correlations with respect to squared distance,

$$k_{\mathrm{RBF}}(x, x') = \sigma_{\mathrm{f}}^2 \exp\Big(- \frac{\|x - x'\|^2}{2\,\ell^2} \Big),$$

and its ARD variant replaces $\ell^2 I$ by the diagonal matrix Λ, which yields $k_{\mathrm{RBF}}(x, x') = \sigma_{\mathrm{f}}^2 \exp\big(- \frac{1}{2}(x - x')^\top \Lambda^{-1} (x - x') \big)$. On a periodic box $D = [0, L]^d$ with Fourier modes $\omega_k = \frac{2\pi}{L} k$ for $k \in \mathbb{Z}^d$, one obtains a discrete spectral (Fourier–KL) representation that expands the process in complex exponentials whose coefficients are Gaussian,

$$f(x) = \sum_{k \in \mathbb{Z}^d} \sqrt{S_{\mathrm{RBF}}(\omega_k)}\, \xi_k\, e^{i\,\omega_k \cdot x},$$

where $\{\xi_k\}$ are independent complex Gaussian variables constrained by $\xi_{-k} = \overline{\xi_k}$ so that f is real valued. The corresponding spectral density under the stated Fourier convention is

$$S_{\mathrm{RBF}}(\omega) = \sigma_{\mathrm{f}}^2\, (2\pi)^{d/2}\, \ell^d\, \exp\Big(- \frac{\ell^2}{2} \|\omega\|^2 \Big).$$

On the full space $\mathbb{R}^d$, stationarity allows the random Fourier integral realization

$$f(x) = \int_{\mathbb{R}^d} e^{i\,\omega \cdot x} \sqrt{S_{\mathrm{RBF}}(\omega)}\, W(d\omega),$$

where W is a complex Gaussian random measure with orthogonal increments, meaning that the increments over disjoint frequency sets are independent. On a general compact $D \subset \mathbb{R}^d$ without periodic structure, one uses the KL expansion $f(x) = \sum_{j \geqslant 1} \sqrt{\lambda_j}\, \xi_j\, \phi_j(x)$, where (λ_j, ϕ_j) are the Mercer eigenpairs of the integral operator defined by k.

Figure 1 displays zero-mean Gaussian-process prior samples on $[-4, 4]$ using the RBF (Gaussian/squared exponential) kernel with hyperparameters $\ell = 1.0$ and $\sigma_{\mathrm{f}} = 1.0$. For numerical stability we add a small jitter $\sigma_{\mathrm{n}} = 10^{-6}$ to the covariance diagonal. The grid has $N = 300$ points and we draw $m = 6$ independent trajectories from $f \sim \mathcal{GP}(0, k_{\mathrm{RBF}})$, where $k_{\mathrm{RBF}}(x, x') = \sigma_{\mathrm{f}}^2 \exp(-\|x - x'\|^2/(2\ell^2))$.

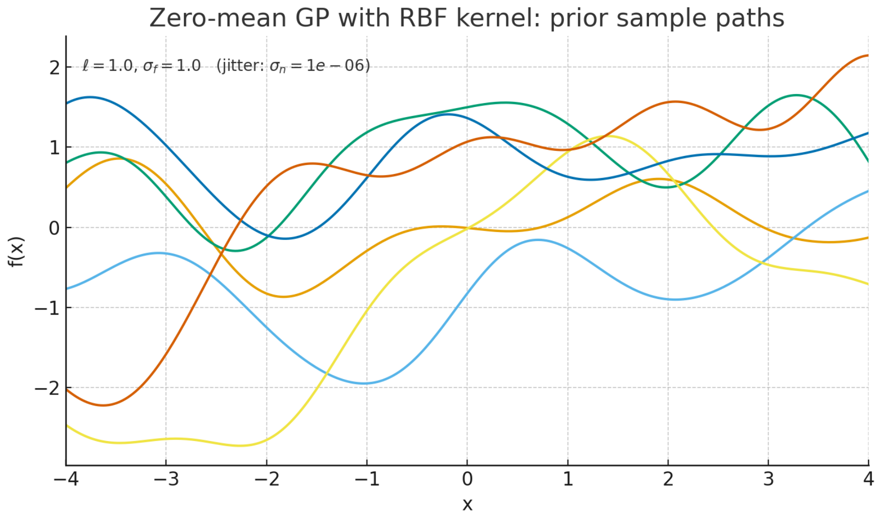

Figure 1: Zero-mean GP prior samples with the RBF kernel on $[-4, 4]$; $m = 6$ independent draws on a grid of $N = 300$ points.

3.2 Matérn family: smoothness controlled by ν

The Matérn family introduces a smoothness parameter $\nu > 0$ that directly controls mean-square differentiability. In isotropic form,

$$k_{\text{Matérn}}(r) = \sigma_{\mathrm{f}}^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\frac{\sqrt{2\nu}\, r}{\ell}\Big)^{\nu} K_{\nu}\Big(\frac{\sqrt{2\nu}\, r}{\ell}\Big),$$

where K_ν is the modified Bessel function; sample paths are mean-square $\lfloor \nu \rfloor$-times differentiable, and as $\nu \to \infty$ the kernel approaches the RBF limit. For half-integers $\nu = p + \frac{1}{2}$, one can realize the process in one dimension via a finite-order Markov state-space model after augmenting the state with derivatives. The spectral density on $\mathbb{R}^d$ has a rational form,

$$S_{\text{Matérn}}(\omega) = \sigma_{\mathrm{f}}^2\, C_{d,\nu}\, \ell^{-2\nu} \Big(\frac{2\nu}{\ell^2} + \|\omega\|^2\Big)^{-(\nu + d/2)},$$

with $C_{d,\nu} = \frac{(2\pi)^{d/2}(2\nu)^{\nu}\Gamma(\nu + \frac{d}{2})}{\Gamma(\nu)}$. Hence on a periodic box the process admits the discrete spectral series

$$f(x) = \sum_{k \in \mathbb{Z}^d} \sqrt{S_{\text{Matérn}}(\omega_k)}\, \xi_k\, e^{i\,\omega_k \cdot x},$$

where the same real-valuedness constraint on coefficients applies, and on $\mathbb{R}^d$ the realization is the random Fourier integral

$$f(x) = \int_{\mathbb{R}^d} e^{i\,\omega\cdot x}\,\sqrt{S_{\text{Matérn}}(\omega)}\,W(d\omega).$$

On a general compact D, one again reverts to the KL expansion with Mercer eigenpairs; an ARD version is obtained by replacing r with the anisotropic distance induced by a diagonal metric Λ.

Figure 2 displays zero-mean Gaussian-process prior samples on $[-4, 4]$ using the Matérn kernel with hyperparameters $\ell = 1.2$, $\sigma_{\mathrm{f}} = 1.0$, and $\nu = 1.5$; for numerical stability a small jitter $\sigma_{\mathrm{n}} = 10^{-6}$ is added to the covariance diagonal. The grid has $N = 300$ points and we draw $m = 6$ independent trajectories from $f \sim \mathcal{GP}(0, k_{\text{Matérn}})$.

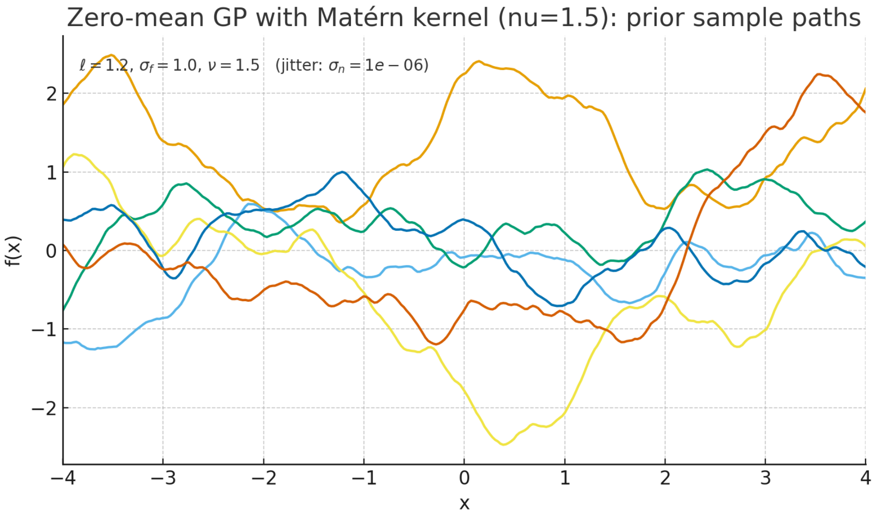

Figure 2: Zero-mean GP prior samples with the Matérn kernel on $[-4, 4]$; $m = 6$ independent draws on a grid of $N = 300$ points.

3.3 Other Gaussian processes: kernels and realizations

Beyond RBF and Matérn, several kernels are widely used and fit the same realization patterns. The rational quadratic kernel

$$k_{\mathrm{RQ}}(x, x') = \sigma_{\mathrm{f}}^2 \Big(1 + \frac{\|x - x'\|^2}{2\alpha\,\ell^2}\Big)^{-\alpha}$$

can be interpreted as a continuous mixture of RBF kernels over length scales, producing a heavy-tailed spectrum that captures multi-scale correlations. An exactly periodic one-dimensional kernel with period $p > 0$ is

$$k_{\mathrm{per}}(x, x') = \sigma_{\mathrm{f}}^2 \exp\Big(-\frac{2}{\ell^2}\sin^2\Big(\pi\frac{x - x'}{p}\Big)\Big),$$

and multiplying this kernel by an RBF or Matérn kernel yields smooth periodic priors with tunable local regularity. Trend and measurement noise are modeled, respectively, by linear kernels and by the white-noise kernel $k_{\mathrm{wn}}(x,x') = \sigma_{\mathrm{n}}^2\,\mathbf{1}\{x = x'\}$. More expressive stationary structure is available through spectral mixture kernels whose spectral density is a finite Gaussian mixture; these lead to discrete spectral series on periodic domains and random Fourier integrals on $\mathbb{R}^d$. In all cases, ARD variants replace $\ell^2 I$ by Λ to decouple relevance by coordinate, while realizations on compact sets follow from KL expansions and realizations on the full space follow from the spectral integral with the appropriate $S(\omega)$.

4 Characterization of posterior distribution

We derive the posterior law in Gaussian process (GP) regression directly from Bayes' formula and show that it is again a GP. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with inputs $x_i \in \mathbb{R}^d$ and observation model

$$y_i = f(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_{\mathrm{n}}^2) \text{ independent of } f.$$

The prior on the latent function is $f \sim \mathcal{GP}(m,k)$. Denote $X = (x_1, \ldots, x_N)$, $y = (y_1, \ldots, y_N)^\top$, $m(X) = (m(x_1), \ldots, m(x_N))^\top$, and $K_{XX} = k(X,X)$. For any test inputs $X_* = (x_1^*, \ldots, x_M^*)$ write $K_{X*} = k(X, X_*)$, $K_{*X} = K_{X*}^\top$, and $K_{**} = k(X_*, X_*)$. The noise covariance is $\Sigma_{\mathrm{n}} = \sigma_{\mathrm{n}}^2 I_N$.

Step 1: Bayes formula at the training points. The likelihood and prior in finite dimensions are

$$p(y \mid f(X)) = \mathcal{N}\big(y; f(X), \Sigma_{\mathrm{n}}\big), \qquad p\big(f(X)\big) = \mathcal{N}\big(f(X); m(X), K_{XX}\big).$$

Bayes' rule gives

$$\begin{aligned} p\big(f(X) \mid y\big) \propto\ & \exp\Big(-\tfrac{1}{2}\big(y - f(X)\big)^\top \Sigma_{\mathrm{n}}^{-1}\big(y - f(X)\big)\Big) \\ & \exp\Big(-\tfrac{1}{2}\big(f(X) - m(X)\big)^\top K_{XX}^{-1}\big(f(X) - m(X)\big)\Big). \end{aligned}$$

Collect the quadratic and linear terms in $f(X)$; with $K = K_{XX}$ and $m_X = m(X)$,

$$-\log p\big(f(X) \mid y\big) = \tfrac{1}{2} f^\top \Big(K^{-1} + \Sigma_{\mathrm{n}}^{-1}\Big) f - \Big(\Sigma_{\mathrm{n}}^{-1} y + K^{-1} m_X\Big)^\top f + \text{const},$$

where $f = f(X)$. Completing the square shows that the posterior is Gaussian with precision $K^{-1} + \Sigma_{\mathrm{n}}^{-1}$, hence

$$\Sigma_{\mathrm{post}} = \Big(K^{-1} + \Sigma_{\mathrm{n}}^{-1}\Big)^{-1}, \qquad m_{\mathrm{post}}(X) = \Sigma_{\mathrm{post}}\Big(\Sigma_{\mathrm{n}}^{-1} y + K^{-1} m_X\Big).$$

Using the matrix inversion lemma and its linear-form identity yields the standard GP expressions

$$\Sigma_{\mathrm{post}} = K - K\big(K + \Sigma_{\mathrm{n}}\big)^{-1} K, \qquad m_{\mathrm{post}}(X) = m_X + K\big(K + \Sigma_{\mathrm{n}}\big)^{-1}\big(y - m_X\big).$$

These formulas show that the kernel hyperparameters (for example σ_{f} and length scales in k) and the noise variance σ_{n}^2 enter only through K and Σ_{n}.

Step 2: Conditioning to obtain predictions at new inputs. Under the prior,

$$\begin{pmatrix} y \\ f(X_*) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(X) \\ m(X_*) \end{pmatrix}, \begin{pmatrix} K_{XX} + \Sigma_{\mathrm{n}} & K_{X*} \\ K_{*X} & K_{**} \end{pmatrix} \right).$$

Since conditioning a jointly Gaussian vector is Gaussian, the posterior predictive law is

$$f(X_*) \mid \mathcal{D} \sim \mathcal{N}\Big(m_{\mathrm{post}}(X_*),\, K_{\mathrm{post}}(X_*, X_*)\Big),$$

with

$$\begin{aligned} m_{\mathrm{post}}(X_*) &= m(X_*) + K_{*X}\big(K_{XX} + \Sigma_{\mathrm{n}}\big)^{-1}\big(y - m(X)\big), \\ K_{\mathrm{post}}(X_*, X_*) &= K_{**} - K_{*X}\big(K_{XX} + \Sigma_{\mathrm{n}}\big)^{-1} K_{X*}. \end{aligned}$$

Conclusion: posterior is again a GP. Passing from finite sets to the process level, the posterior is a Gaussian process

$$f \mid \mathcal{D} \sim \mathcal{GP}\big(m_{\mathrm{post}}, k_{\mathrm{post}}\big),$$

where for any $x, x' \in \mathbb{R}^d$

$$\boxed{m_{\mathrm{post}}(x) = m(x) + k(x, X)\big(K_{XX} + \Sigma_{\mathrm{n}}\big)^{-1}\big(y - m(X)\big)}$$

$$\boxed{k_{\mathrm{post}}(x, x') = k(x, x') - k(x, X)\big(K_{XX} + \Sigma_{\mathrm{n}}\big)^{-1} k(X, x')}$$

These are the formulas for the posterior mean m_{post} and covariance $K_{\mathrm{post}}(\cdot, \cdot)$, obtained by an explicit Bayes derivation via completing the square and Gaussian conditioning.

To sample from the posterior distribution it suffices to work on a finite set of inputs $X_* = (x_j^*)_{j=1}^M$, since a Gaussian process is fully determined by its finite-dimensional Gaussian marginals. Concretely, with m_{post} and k_{post} as derived above, the posterior vector of function values satisfies $f(X_*) \mid \mathcal{D} \sim \mathcal{N}(m_*, K_*)$, where $m_* = \big(m_{\mathrm{post}}(x_1^*), \ldots, m_{\mathrm{post}}(x_M^*)\big)^\top$ and $K_* = \big[k_{\mathrm{post}}(x_i^*, x_j^*)\big]_{i,j=1}^M$. In practice one adds a tiny numerical jitter $\varepsilon_{\mathrm{num}}^2$ to the diagonal for stability, computes a Cholesky factorization $K_* + \varepsilon_{\mathrm{num}}^2 I_M = LL^\top$, draws $z \sim \mathcal{N}(0, I_M)$, and sets $f(X_*) = m_* + Lz$; repeating this step yields independent posterior sample paths on X_*. If noisy predictions are needed, form $y_* = f(X_*) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma_{\mathrm{n}}^2 I_M)$ (or a general noise covariance). On a compact domain one may also approximate continuous trajectories by truncating the Karhunen–Loève expansion of k_{post}, while for stationary kernels on $\mathbb{R}^d$ random Fourier integral constructions offer spectral samplers; nevertheless, the finite-set Cholesky scheme above is the default, robust method for posterior sampling in computation and visualization.

Algorithm 1: Gaussian Process Regression: Posterior Characterization

Input : Training inputs $X = (x_i)_{i=1}^N$, training outputs $y = (y_1, \dots, y_N)^\top$; mean function $m : \mathbb{R}^d \to \mathbb{R}$; kernel $k(\cdot, \cdot; \theta)$ with hyperparameters θ; noise variance σ_n^2; optional test inputs $X_* = (x_j^*)_{j=1}^M$.

Output: Posterior GP $f \mid \mathcal{D} \sim \mathcal{GP}(m_\mathrm{post}, k_\mathrm{post})$;

1 **Build blocks.** Set

$$m_X = m(X), \qquad K_{XX} = k(X, X), \qquad \Sigma_\mathrm{n} = \sigma_\mathrm{n}^2 I_N.$$

If X_* is provided, also set

$$m_* = m(X_*), \qquad K_{X*} = k(X, X_*), \qquad K_{*X} = K_{X*}^\top, \qquad K_{**} = k(X_*, X_*).$$

2 **Solve the data-fit system.** Form the centered targets $r = y - m_X$ and solve

$$\alpha = \left(K_{XX} + \Sigma_\mathrm{n}\right)^{-1} r.$$

Numerical note: use a Cholesky factorization of $K_{XX} + \Sigma_\mathrm{n}$.

3 **Posterior GP (process level).** For any $x, x' \in \mathbb{R}^d$,

$$\boxed{m_\mathrm{post}(x) = m(x) + k(x, X)\, \alpha}$$

$$\boxed{k_\mathrm{post}(x, x') = k(x, x') - k(x, X)\left(K_{XX} + \Sigma_\mathrm{n}\right)^{-1} k(X, x')}$$

which defines $f \mid \mathcal{D} \sim \mathcal{GP}(m_\mathrm{post}, k_\mathrm{post})$.

5 numerical experiments

We present a PDE-informed Gaussian process (GP) regression experiment to visualize the posterior distribution of a function on a one-dimensional domain. The latent field u on $[0, 1]$ is given a zero-mean GP prior with the RBF (Gaussian/squared exponential) kernel $k(x, x') = \sigma_\mathrm{f}^2 \exp(-\|x - x'\|^2/(2\ell^2))$. To encode physics, we introduce *linear* observations arising from the differential operator $L = \frac{d^2}{dx^2} + c$; specifically, we use the homogeneous Helmholtz/Poisson-type relation $Lu = 0$ with Dirichlet boundary conditions. For a concrete target we set $c = \pi^2$, so that the reference solution $u_\mathrm{true}(x) = \sin(\pi x)$ satisfies $Lu_\mathrm{true} = 0$ and $u_\mathrm{true}(0) = u_\mathrm{true}(1) = 0$. The dataset combines three kinds of linear information: (i) boundary values $u(0)$ and $u(1)$, (ii) a few noisy interior pointwise values $u(x_i)$, and (iii) noisy PDE residuals $(Lu)(x_j) \approx 0$ at collocation points. Measurement noise is modeled as independent Gaussian with variances collected in a diagonal matrix, adding a nugget σ_n^2 to ensure numerical stability. Because GPs are closed under linear maps, the likelihood for all constraints is jointly Gaussian, and the posterior $u \mid \mathcal{D}$ is again a GP with mean and covariance obtained by the standard conditioning formulas derived earlier; in practice we compute the posterior mean and the pointwise variances on a dense grid via a Cholesky factorization of the combined covariance of all observations.

The resulting figure reports the posterior mean curve together with a 95% pointwise credible band (mean ± 2 standard deviations), as well as the underlying reference solution and the locations of the measurements. As expected, the credible band tightens near boundary and interior value observations and remains narrow where the PDE residual constraints are dense, demonstrating how the linear-physics information effectively reduces posterior uncertainty while preserving Bayesian coherence.

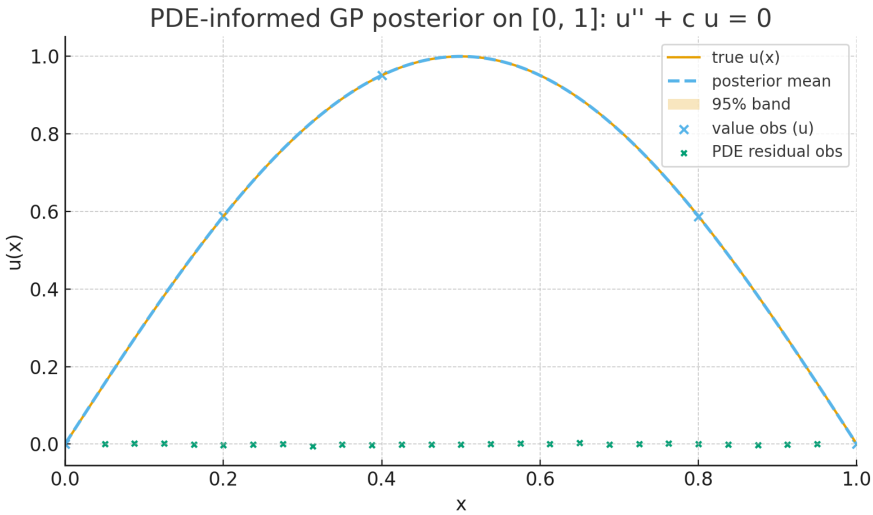

Figure 3: PDE-informed GP posterior on $[0, 1]$ for $Lu = 0$ with $L = \frac{d^2}{dx^2} + \pi^2$. The plot shows the posterior mean (dashed) and its 95% credible band, together with the reference solution and the observation locations (boundary values, interior pointwise values, and PDE residual collocation points).

6 concluding remarks

In summary, we presented Gaussian Process Regression (GPR) as a Bayesian nonparametric framework in which a prior over functions is specified by a mean and a covariance kernel, finite–dimensional laws are characterized explicitly, and sample–path representations arise via Karhunen–Loève expansions on compact domains and spectral representations for stationary kernels. Building on these foundations, we derived the posterior distribution directly from Bayes' formula and Gaussian conditioning, thereby obtaining closed-form expressions for the posterior mean and covariance, and we demonstrated in a PDE–informed experiment how linear physics constraints and noisy data can be fused to deliver calibrated uncertainty quantification. Looking ahead, the same principles extend to learning hyperparameters by marginal likelihood or hierarchical priors, to designing structured kernels (ARD, periodic, spectral mixtures) that encode inductive bias, to handling non-Gaussian likelihoods through approximate inference, and to scalable approximations (inducing points, low-rank factorizations, iterative solvers) that preserve posterior calibration. These directions, together with multi-output and operator-valued extensions, provide a practical pathway from the theory summarized here to large-scale scientific and engineering applications.

References

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2004.

[2] N. Cressie. *Statistics for Spatial Data*. Wiley, 1993.

[3] O. Kallenberg. *Foundations of Modern Probability*. Springer, 2nd edition, 2002.

[4] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, 63(3):425–464, 2001.

[5] M. Loève. *Probability Theory II*. Springer, 4th edition, 1977.

[6] J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209: 415–446, 1909.

[7] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[8] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, 2012.

[9] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

[10] A. M. Stuart. Inverse problems: A bayesian perspective. *Acta Numerica*, 19:451–559, 2010.