2

GAUSSIAN PROCESSES

The central object in optimization is an objective function $f\colon \mathcal{X} \to \mathbb{R}$, and the primary challenge in algorithm design is inherent *uncertainty* about this function: most importantly, where is the function maximized and what is its maximal value? Prior to optimization, we may very well have no idea. Optimization affords us the opportunity to acquire information about the objective – through observations of our own design – to shed light on these questions. However, this process is itself fraught with uncertainty, as we cannot know the outcomes and implications of these observations at the time of their design. Notably, we face this uncertainty even when we have a closed-form expression for the objective function, a favorable position as many objectives act as "black boxes."

Reflecting on this situation, DIACONIS posed an intriguing question:[1] "what does it mean to 'know' a function?" The answer is unclear when an analytic expression, which might at first glance seem to encapsulate the essence of the function, is insufficient to determine features of interest. However, DIACONIS argued that although we may not know *everything* about a function, we often have *some* prior knowledge that can facilitate a numerical procedure such as optimization. For example, we may expect an objective function to be smooth (or rough), or to assume values in a given range, or to feature a relatively simple underlying trend, or to depend on some hidden low-dimensional representation we hope to uncover.[2] All of this knowledge could be instrumental in accelerating optimization if it could be systematically captured and exploited.

Having identifiable information about an objective function prior to optimization motivates the Bayesian approach we will explore throughout this book. We will address uncertainty in the objective function through the unifying framework of Bayesian inference, treating f – as well as ancillary quantities such as x^* and f^* (1.1) – as random variables to be inferred from observations revealed during optimization.

To pursue this approach, we must first determine how to build useful prior distributions for objective functions and how to compute a posterior belief given observations. If the system under investigation is well understood, we may be able to identify an appropriate parametric form $f(x;\boldsymbol{\theta})$ and infer the parameters $\boldsymbol{\theta}$ directly. This approach is likely the best course of action when possible;[3] however, many objective functions have no obvious parametric form, and most models used in Bayesian optimization are thus nonparametric to avoid undue assumptions.[4]

In this chapter we will introduce *Gaussian processes* (GPs), a convenient class of nonparametric regression models widely used in Bayesian optimization. We will begin by defining Gaussian processes and deriving some basic properties, then demonstrate how to perform inference from observations. In the case of exact observation and additive Gaussian noise, we can perform this inference *exactly,* resulting in an updated posterior Gaussian process. We will continue by considering some theoretical properties of Gaussian processes relevant to optimization and inference with non-Gaussian observation models.

1 P. DIACONIS (1988). Bayesian Numerical Analysis. In: *Statistical Decision Theory and Related Topics IV.*

2 We will explore all of these possibilities in the next chapter, p. 45.

Bayesian inference of the objective function: § 1.2, p. 8

3 V. DALIBARD et al. (2017). BOAT: Building Auto-Tuners with Structured Bayesian Optimization. *WWW 2017.*

4 The term"nonparametric" is something of a misnomer. A nonparametric objective function model has parameters but their dimension is infinite – we effectively parameterize the objective by its value at every point.

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

The literature on Gaussian processes is vast, and we do not intend this chapter to serve as a standalone introduction but rather as companion to the existing literature. Although our discussion will be comprehensive, our focus on optimization will sometimes bias its scope. For a broad overview, the interested reader may consult RASMUSSEN and WILLIAMS's classic monograph.[5]

5 C. E. RASMUSSEN and C. K. I. WILLIAMS (2006). *Gaussian Processes for Machine Learning*. MIT Press.

2.1 DEFINITION AND BASIC PROPERTIES

multivariate normal distribution: § A.2, p. 296

A *Gaussian process* is an extension of the familiar multivariate normal distribution suitable for modeling functions on infinite domains. Gaussian processes inherit the convenient mathematical properties of the multivariate normal distribution without sacrificing computational tractability. Further, by modifying the structure of a GP, we can model functions with a rich variety of behavior; we will explore this capability in the next chapter. This combination of mathematical elegance and flexibility in modeling has established Gaussian processes as the workhorse of Bayesian approaches to numerical tasks, including optimization.[6,7]

Chapter 3: Modeling with Gaussian Processes, p. 45

6 P. HENNIG et al. (2015). Probabilistic Numerics and Uncertainty in Computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 471(2179):20150142.

7 P. HENNIG et al. (2022). *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press.

Definition

Consider an objective function $f\colon \mathcal{X} \to \mathbb{R}$ of interest over an arbitrary infinite domain $\mathcal{X}$.[8] We will take a nonparametric approach and reason about the function as an infinite collection of random variables, one corresponding to the function value at every point in the domain. Mutual dependence between these random variables will then determine the statistical properties of the function's shape.

8 If $\mathcal{X}$ is finite, there is no distinction between a Gaussian process and a multivariate normal distribution, so only the infinite case is interesting for this discussion.

It is perhaps not immediately clear how we can specify a useful distribution over infinitely many random variables, a construction known as a *stochastic process*. However, a result known as the *Kolmogorov extension theorem* allows us to construct a stochastic process by defining only the distribution of arbitrary *finite* sets of function values, subject to natural consistency constraints.[9] For a Gaussian process, these finite-dimensional distributions are all multivariate Gaussian, hence its name.

9 B. ØKSENDAL (2013). *Stochastic Differential Equations: An Introduction with Applications*. Springer–Verlag. [§ 2.1]

In this light, we build a Gaussian process by replacing the parameters in the finite-dimensional case – a mean vector and a positive semidefinite covariance matrix – by analogous *functions* over the domain. We specify a Gaussian process on f:[10]

$$p(f) = \mathcal{GP}(f; \mu, K)$$

10 Writing the process as if it were a function-valued probability density function is an abuse of notation, but a useful and harmless one.

mean function, μ
covariance function (kernel), K
value of objective at x, ϕ

by a *mean function* $\mu\colon \mathcal{X} \to \mathbb{R}$ and a positive semidefinite *covariance function* (or *kernel*) $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The mean function determines the expected function value $\phi = f(x)$ at any location x:

$$\mu(x) = \mathbb{E}[\phi \mid x],$$

thus serving as a location parameter representing the function's central tendency. The covariance function determines how deviations from the

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

mean are structured, encoding expected properties of the function's behavior. Defining $\phi' = f(x')$, we have:

value of objective at x', ϕ'

$$K(x, x') = \operatorname{cov}[\phi, \phi' \mid x, x']. \tag{2.1}$$

The mean and covariance functions of the process allow us to compute any finite-dimensional marginal distribution on demand. Let $\mathbf{x} \subset \mathcal{X}$ be finite and let $\boldsymbol{\phi} = f(\mathbf{x})$ be the corresponding function values, a vector-valued random variable. For the Gaussian process (2.1), the distribution of $\boldsymbol{\phi}$ is multivariate normal with parameters determined by the mean and covariance functions:

values of objective at $\mathbf{x}$, $\boldsymbol{\phi} = f(\mathbf{x})$

$$p(\boldsymbol{\phi} \mid \mathbf{x}) = \mathcal{N}(\boldsymbol{\phi}; \boldsymbol{\mu}, \Sigma), \tag{2.2}$$

where

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{\phi} \mid \mathbf{x}] = \mu(\mathbf{x}); \qquad \Sigma = \operatorname{cov}[\boldsymbol{\phi} \mid \mathbf{x}] = K(\mathbf{x}, \mathbf{x}). \tag{2.3}$$

Here $K(\mathbf{x}, \mathbf{x})$ is the matrix formed by evaluating the covariance function for each pair of points: $\Sigma_{ij} = K(x_i, x_j)$, also called the *Gram matrix* of $\mathbf{x}$.

Gram matrix of $\mathbf{x}$, $\Sigma = K(\mathbf{x}, \mathbf{x})$

In many ways, Gaussian processes behave like "really big" Gaussian distributions, and one can intuit many of their properties from this heuristic alone. For example, the Gaussian marginal property in (2.2–2.3) corresponds precisely with the analogous formula in the finite-dimensional case (A.13). Further, this property automatically ensures global consistency in the following sense.[11] If $\mathbf{x}$ is an arbitrary set of points and $\mathbf{x}' \supset \mathbf{x}$ is a superset, then we arrive at the same belief about $\boldsymbol{\phi}$ whether we compute it directly from (2.2–2.3) or indirectly by first computing $p(\boldsymbol{\phi}' \mid \mathbf{x}')$ then marginalizing $\mathbf{x}' \setminus \mathbf{x}$ (A.13).

11 In fact, this is precisely the consistency required by the Kolmogorov extension theorem mentioned on the facing page.

marginalizing multivariate normal distributions, § A.2, p. 299

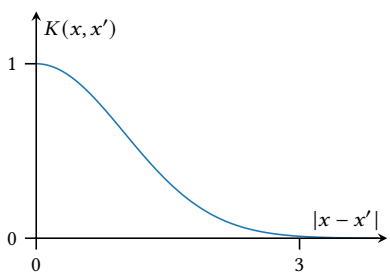
Example and basic properties

Let us construct and explore an explicit Gaussian process for a function on the interval $\mathcal{X} = [0, 30]$. For the mean function we take the zero function $\mu \equiv 0$, indicating a constant central tendency. For the covariance function, we take the prototypical *squared exponential* covariance:

squared exponential covariance: § 3.3, p. 51

$$K(x, x') = \exp\left(-\tfrac{1}{2}|x - x'|^2\right). \tag{2.4}$$

Let us pause to consider the implications of this choice. First, note that $\operatorname{var}[\phi \mid x] = K(x, x) = 1$ at every point $x \in \mathcal{X}$, and thus the covariance function (2.4) also measures the *correlation* between the function values ϕ and ϕ'. This correlation decreases with the distance between x and x', falling from unity to zero as these points become increasingly separated; see the illustration in the margin. We can loosely interpret this as a statistical consequence of continuity: function values at nearby locations are highly correlated, whereas function values at distant locations are effectively independent. This assumption also implies that observing the function at some point x provides nontrivial information about the function at sufficiently nearby locations (roughly when $|x - x'| < 3$). We will explore this implication further shortly.



The squared exponential covariance (2.4) as a function of the distance between inputs.

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

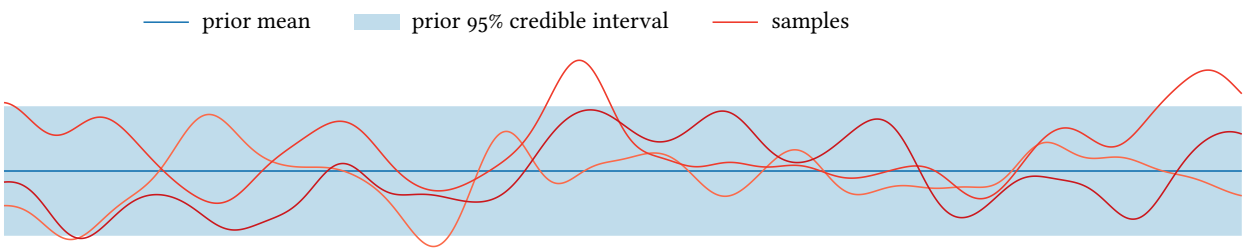


Figure 2.1: Our example Gaussian process on the domain $\mathcal{X} = [0, 30]$. We illustrate the marginal belief at every location with its mean and a 95% credible interval and also show three example functions sampled from the process.

predictive credible intervals

For a Gaussian process, the marginal distribution of any *single* function value is univariate normal (2.2):

$$p(\phi \mid x) = \mathcal{N}(\phi; \mu, \sigma^2); \qquad \mu = \mu(x); \qquad \sigma^2 = K(x, x), \tag{2.5}$$

where we have abused notation slightly by overloading the symbol μ. This allows us to derive pointwise credible intervals; for example, the familiar $\mu \pm 1.96\sigma$ is a 95% credible interval for ϕ. Examining our example GP, the marginal distribution of every function value is in fact *standard* normal. We provide a rough visual summary of the process via its mean function and pointwise 95% predictive credible intervals in Figure 2.1. There is nothing terribly exciting we can glean from these marginal distributions alone, and no interesting structure in the process is yet apparent.

Sampling

We may gain more insight by inspecting samples drawn from our example process reflecting the *joint* distribution of function values. Although it is impossible to represent an arbitrary function on $\mathcal{X}$ in finite memory, we can approximate the sampling process by taking a dense grid $\mathbf{x} \subset \mathcal{X}$ and sampling the corresponding function values from their joint multivariate normal distribution (2.2). Plotting the sampled vectors against the chosen grid reveals curves approximating draws from the Gaussian process. Figure 2.1 illustrates this procedure for our example using a grid of 1000 equally spaced points. Each sample is smooth and has several local optima distributed throughout the domain – for some applications, this might be a reasonable model for an objective function on $\mathcal{X}$.

sampling from a multivariate normal distribution: § A.2, p. 299

2.2 INFERENCE WITH EXACT AND NOISY OBSERVATIONS

We now turn to our attention to *inference*: given a Gaussian process prior on an objective function, how can we condition this initial belief on observations obtained during optimization?

example and discussion

Let us look at an example to build intuition before diving into the details. Figure 2.2 shows the effect of conditioning our example GP from

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

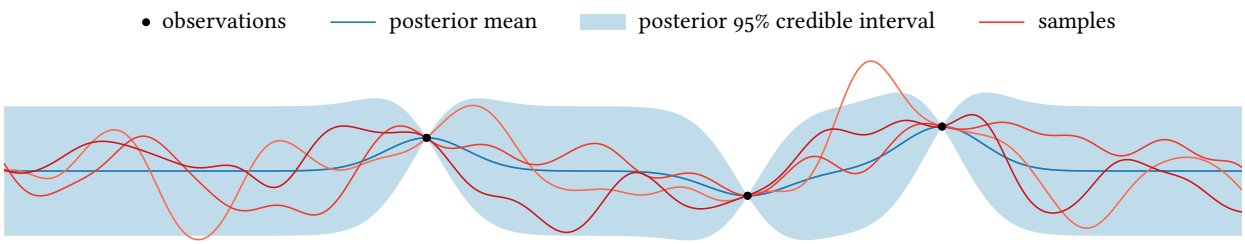


Figure 2.2: The posterior for our example scenario in Figure 2.1 conditioned on three exact observations.

the previous section on three exact measurements of the function. The updated belief reflects both our prior assumptions and the information contained in the data, the hallmark of Bayesian inference. To elaborate, the posterior mean smoothly interpolates through the observed values, agreeing with both the measured values and the smoothness encoded in the prior covariance function. The posterior credible intervals are reduced in the neighborhood of the measured locations – where the prior covariance function encodes nontrivial dependence on at least one observed value – and vanish where the function value has been exactly determined. On the other hand, our marginal belief remains effectively unchanged from the prior in regions sufficiently isolated from the data, where the prior covariance function encodes effectively no correlation.

Conveniently, inference is straightforward for the pervasive observation models of exact measurement and additive Gaussian noise, where the self-conjugacy of the normal distribution yields a *Gaussian process* posterior with updated parameters we can compute in closed form. The reasoning underlying inference for both observation models is identical and is subsumed by a flexible general argument we will present first.

Inference from arbitrary jointly Gaussian observations

We may exactly condition a Gaussian process $p(f) = \mathcal{GP}(f; \mu, K)$ on the observation of *any* vector $\mathbf{y}$ sharing a joint Gaussian distribution with f:

vector of observed values, $\mathbf{y}$

$$p(f, \mathbf{y}) = \mathcal{GP}\left(\begin{bmatrix} f \\ \mathbf{y} \end{bmatrix}; \begin{bmatrix} \mu \\ \mathbf{m} \end{bmatrix}, \begin{bmatrix} K & \kappa^\top \\ \kappa & \mathbf{C} \end{bmatrix}\right). \tag{2.6}$$

This notation, analogous to (A.12), extends the Gaussian process on f to include the entries of $\mathbf{y}$; that is, we assume the distribution of any finite subset of function and/or observed values is multivariate normal. We specify the joint distribution via the marginal distribution of $\mathbf{y}$:[12]

12 We assume $\mathbf{C}$ is positive definite; if it were only positive *semi*definite, there would be wasteful linear dependence among observations.

observation mean and covariance, $\mathbf{m}$, $\mathbf{C}$

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mathbf{m}, \mathbf{C}) \tag{2.7}$$

and the cross-covariance function between $\mathbf{y}$ and f:

cross-covariance between observations and function values, κ

$$\kappa(x) = \operatorname{cov}[\mathbf{y}, \phi \mid x]. \tag{2.8}$$

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

Although it may seem absurd that we could identify and observe a vector satisfying such strong restrictions on its distribution, we can already deduce several examples from first principles, including:

inference from exact observations: § 2.2, p. 22

- any vector of function values (2.2),

affine transformations: § A.2, p. 298

- any affine transformation of function values (A.10), and

derivatives and expectations: § 2.6, p. 30

- limits of such quantities, such as partial derivatives or expectations.

Further, we may condition on any of the above even if corrupted by independent additive Gaussian noise, as we will shortly demonstrate.

conditioning a multivariate normal distribution: § A.2, p. 299

We may condition the joint distribution (2.6) on $\mathbf{y}$ analogously to the finite-dimensional case (A.14), resulting in a Gaussian process posterior on f. Writing $\mathcal{D} = \mathbf{y}$ for the observed data, we have:

$$p(f \mid \mathcal{D}) = \mathcal{GP}(f; \mu_{\mathcal{D}}, K_{\mathcal{D}}), \tag{2.9}$$

posterior mean and covariance, $\mu_{\mathcal{D}}$, $K_{\mathcal{D}}$

where

$$\begin{aligned} \mu_{\mathcal{D}}(x) &= \mu(x) + \kappa(x)^\top \mathbf{C}^{-1}(\mathbf{y} - \mathbf{m}); \\ K_{\mathcal{D}}(x, x') &= K(x, x') - \kappa(x)^\top \mathbf{C}^{-1} \kappa(x'). \end{aligned} \tag{2.10}$$

This can be verified by computing the joint distribution of an arbitrary finite set of function values and $\mathbf{y}$ and conditioning on the latter (A.14).[13]

13 This is a useful exercise! The result will be a stochastic process with multivariate normal finite-dimensional distributions, a Gaussian process by definition (2.5).

The above result provides a simple procedure for GP posterior inference from any vector of observations satisfying (2.6):

1. compute the marginal distribution of $\mathbf{y}$ (2.7),
2. derive the cross-covariance function κ (2.8), and
3. find the posterior distribution of f via (2.9–2.10).

We will realize this procedure for several special cases below. However, we will first demonstrate how we may seamlessly handle measurements corrupted by additive Gaussian noise and build intuition for the posterior distribution by dissecting its moments in terms of the statistics of the observations and the correlation structure of the prior.

Corruption by additive Gaussian noise

We pause to make one observation of immense practical importance: any vector satisfying (2.6) would continue to suffice even if corrupted by independent additive Gaussian noise, and thus we can use the above result to condition a Gaussian process on *noisy* observations as well.

noisy observation of $\mathbf{y}$, $\mathbf{z}$
vector of random errors, $\boldsymbol{\varepsilon}$
noise covariance matrix, $\mathbf{N}$

Suppose that rather than observing $\mathbf{y}$ exactly, our measurement mechanism only allowed observing $\mathbf{z} = \mathbf{y} + \boldsymbol{\varepsilon}$ instead, where $\boldsymbol{\varepsilon}$ is a vector of random errors independent of $\mathbf{y}$. If the errors are normally distributed with mean zero and known (arbitrary) covariance $\mathbf{N}$:

$$p(\boldsymbol{\varepsilon} \mid \mathbf{N}) = \mathcal{N}(\boldsymbol{\varepsilon}; \mathbf{0}, \mathbf{N}), \tag{2.11}$$

sums of normal vectors: § A.2, p. 300

then we have

$$p(\mathbf{z} \mid \mathbf{N}) = \mathcal{N}(\mathbf{z}; \mathbf{m}, \mathbf{C} + \mathbf{N}); \qquad \operatorname{cov}[\mathbf{z}, \phi \mid x] = \operatorname{cov}[\mathbf{y}, \phi \mid x] = \kappa(x).$$

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

Thus we can condition on an observation of the corrupted vector $\mathbf{z}$ by simply replacing $\mathbf{C}$ with $\mathbf{C} + \mathbf{N}$ in the prior (2.6) and posterior (2.10).[14] Note that the posterior converges to that from a direct observation of $\mathbf{y}$ if we take the noise covariance $\mathbf{N} \to \mathbf{0}$ in the positive semidefinite cone, a reassuring result.

14 Assuming zero-mean errors is not strictly necessary but is overwhelmingly common in practice. A nonzero mean $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{n}$ is possible by further replacing $(\mathbf{y} - \boldsymbol{\mu})$ with $(\mathbf{y} - [\boldsymbol{\mu} + \mathbf{n}])$ in (2.10), where $\boldsymbol{\mu} + \mathbf{n} = \mathbb{E}[\mathbf{y}]$.

Interpretation of posterior moments

The moments of the posterior Gaussian process (2.10) contain update terms adjusting the prior moments in light of the data. These updates have intuitive interpretations in terms of the nature of the prior process and the observed values, which we may unravel with some care.

We can gain some initial insight by considering the case where we observe a *single* value with y distribution $\mathcal{N}(y; m, s^2)$ and breaking down its impact on our belief. Consider an arbitrary function value ϕ with prior distribution $\mathcal{N}(\phi; \mu, \sigma^2)$ (2.5) and define

z-score of measurement y, z

$$z = \frac{y - m}{s}$$

to be the z-score of the observed value y and

correlation between measurement y and function value ϕ, ρ

$$\rho = \operatorname{corr}[y, \phi \mid x] = \frac{\kappa(x)}{\sigma s}$$

to be the correlation between y and ϕ. Then the posterior mean and standard deviation of ϕ are, respectively:

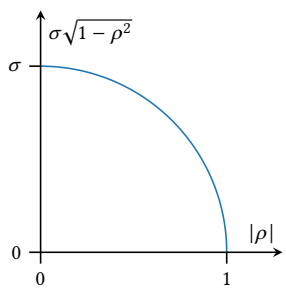
posterior moments of ϕ from a scalar observation

$$\mu + \sigma\rho z; \qquad \sigma\sqrt{1 - \rho^2}. \tag{2.12}$$

The z-score of the posterior mean, with respect to the prior distribution of ϕ, is ρz. An independent measurement with $\rho = 0$ thus leaves the prior mean unchanged, whereas a perfectly dependent measurement with $|\rho| = 1$ shifts the mean up or down by z standard deviations (depending on the sign of the correlation) to match the magnitude of the measurement's z-score. Measurements with partial dependence result in outcomes between these extremes. Further, *surprising* measurements – that is, those with large $|z|$ – yield larger shifts in the mean, whereas an entirely expected measurement with $y = m$ leaves the mean unchanged.

interpretation of moments

Turning to the posterior standard deviation, the measurement reduces our uncertainty in ϕ by a factor depending on the correlation ρ, but *not* on the value observed. An independent measurement again leaves the prior intact, whereas a perfectly dependent measurement collapses the posterior standard deviation to zero as the value of ϕ would be completely determined. The relative reduction in posterior uncertainty as a function of the absolute correlation is illustrated in the margin.



The posterior standard deviation of ϕ as a function of the strength of relationship with y, $|\rho|$.

In the case of vector-valued observations, we can interpret similar structure in the posterior, although dependence between entries of $\mathbf{y}$ must also be accounted for. We may factor the observation covariance matrix as

$$\mathbf{C} = \mathbf{SPS}, \tag{2.13}$$

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

where $\mathbf{S}$ is diagonal with $S_{ii} = \sqrt{C_{ii}} = \text{std}[y_i]$ and $\mathbf{P} = \text{corr}[\mathbf{y}]$ is the observation correlation matrix. We may then rewrite the posterior mean of ϕ as

$$\mu + \sigma \boldsymbol{\rho}^\top \mathbf{P}^{-1} \mathbf{z},$$

where $\mathbf{z}$ and $\boldsymbol{\rho}$ represent the vectors of measurement z-scores and the cross-correlation between ϕ and $\mathbf{y}$, respectively:

$$z_i = \frac{y_i - m_i}{s_i}; \qquad \rho_i = \frac{[\kappa(x)]_i}{\sigma s_i}.$$

The posterior mean is now in the same form as the scalar case (2.12), with the introduction of the observation correlation matrix moderating the z-scores to account for dependence between the observed values.[15]

The posterior standard deviation of ϕ in the vector-valued case is

$$\sigma \sqrt{1 - \boldsymbol{\rho}^\top \mathbf{P}^{-1} \boldsymbol{\rho}},$$

again analogous to (2.12). Noting that the inverse correlation matrix $\mathbf{P}^{-1}$ is positive definite,[16] the posterior covariance again reflects a global reduction in the marginal uncertainty of every function value. In fact, the *joint* distribution of any set of function values has reduced uncertainty in the posterior in terms of the differential entropy (A.16), as[17]

$$|K(\mathbf{x}, \mathbf{x}) - \kappa(\mathbf{x})^\top \mathbf{C}^{-1} \kappa(\mathbf{x})| \leq |K(\mathbf{x}, \mathbf{x})|.$$

The reduction of uncertainty again depends on the strength of dependence between function values and the observed data, with independence ($\boldsymbol{\rho} = \mathbf{0}$) resulting in no change. The reduction also depends on the precision of the measurements: all other things held equal, observations with greater precision in terms of the Löwner order[18] on the precision matrix $\mathbf{C}^{-1}$ provide a globally better informed posterior. In particular, as $(\mathbf{C} + \mathbf{N})^{-1} \prec \mathbf{C}^{-1}$ for any noise covariance $\mathbf{N}$, noisy measurements (2.11) categorically provide *less* information about the function than direct observations, as one should hope.

15 It can be instructive to contrast the behavior of the posterior when conditioning on two highly correlated values versus two independent ones. In the former case, the posterior does not change much as a result of the second measurement, as dependence reduces the effective number of measurements.

16 $\mathbf{P}$ is congruent to $\mathbf{C}$ (2.13) and is thus positive definite from Sylvester's law of inertia.

17 For positive semidefinite $\mathbf{A}$, $\mathbf{B}$, $|\mathbf{A}| \leq |\mathbf{A} + \mathbf{B}|$.

18 The Löwner order is the partial order induced by the convex cone of positive-semidefinite matrices. For symmetric $\mathbf{A}$, $\mathbf{B}$, we define $\mathbf{A} \prec \mathbf{B}$ if and only if $\mathbf{B} - \mathbf{A}$ is positive definite:

K. Löwner (1934). Über monotone Matrixfunktionen. *Mathematische Zeitschrift* 38:177–216.

Inference with exact function evaluations

We will now explicitly demonstrate the general process of Gaussian process inference for important special cases, beginning with the simplest possible observation mechanism: exact observation.

observed data, $\mathcal{D} = (\mathbf{x}, \boldsymbol{\phi})$

Suppose we have observed f at some set of locations $\mathbf{x}$, revealing the corresponding function values $\boldsymbol{\phi} = f(\mathbf{x})$, and let $\mathcal{D} = (\mathbf{x}, \boldsymbol{\phi})$ denote this dataset. The observed vector shares a joint Gaussian distribution with any other set of function values by the GP assumption on f (2.2), so we may follow the above procedure to derive the posterior. The marginal distribution of $\boldsymbol{\phi}$ is Gaussian (2.3):

$$p(\boldsymbol{\phi} \mid \mathbf{x}) = \mathcal{N}(\boldsymbol{\phi}; \boldsymbol{\mu}, \Sigma),$$

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

and the cross-covariance between an arbitrary function value and $\boldsymbol{\phi}$ is by definition given by the covariance function:

$$\kappa(x) = \operatorname{cov}[\boldsymbol{\phi}, \phi \mid \mathbf{x}, x] = K(\mathbf{x}, x).$$

Appealing to (2.9–2.10) we have:

$$p(f \mid \mathcal{D}) = \mathcal{GP}(f; \mu_{\mathcal{D}}, K_{\mathcal{D}}),$$

where

$$\begin{aligned} \mu_{\mathcal{D}}(x) &= \mu(x) + K(x, \mathbf{x})\Sigma^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu}); \\ K_{\mathcal{D}}(x, x') &= K(x, x') - K(x, \mathbf{x})\Sigma^{-1}K(\mathbf{x}, x'). \end{aligned} \tag{2.14}$$

example and discussion

Our previous Figure 2.2 illustrates the posterior resulting from conditioning our GP prior in Figure 2.1 on three exact measurements, with high-level analysis of its behavior in the accompanying text.

Inference with function evaluations corrupted by additive Gaussian noise

With the notable exception of optimizing the output of a deterministic computer program or simulation, observations of an objective function are typically corrupted by noise due to measurement limitations or statistical approximation; we must be able to handle such noisy observations to maximize utility. Fortunately, in the important case of additive Gaussian noise, we may perform exact inference following the general procedure described above. In fact, the derivation below follows directly from our previous discussion on arbitrary additive Gaussian noise, but the case of additive Gaussian noise in function evaluations is important enough to merit its own discussion.

arbitrary additive Gaussian noise: § 2.2, p. 20

Suppose we make observations of f at locations $\mathbf{x}$, revealing corrupted values $\mathbf{y} = \boldsymbol{\phi} + \boldsymbol{\varepsilon}$. Suppose the measurement errors $\boldsymbol{\varepsilon}$ are independent of $\boldsymbol{\phi}$ and normally distributed with mean zero and covariance $\mathbf{N}$, which may optionally depend on $\mathbf{x}$:

$$p(\boldsymbol{\varepsilon} \mid \mathbf{x}, \mathbf{N}) = \mathcal{N}(\boldsymbol{\varepsilon}; \mathbf{0}, \mathbf{N}). \tag{2.15}$$

19 Allowing nondiagonal $\mathbf{N}$ departs from our typical convention of assuming conditional independence between observations (1.3), but doing so does not complicate inference, so there is no harm in this generality.

As before we aggregate the observations into a dataset $\mathcal{D} = (\mathbf{x}, \mathbf{y})$.

The observation noise covariance can in principle be arbitrary;[19] however, the most common models in practice are independent homoskedastic noise with scale σ_n:

special case: independent homoskedastic noise

$$\mathbf{N} = \sigma_n^2 \mathbf{I}, \tag{2.16}$$

and independent heteroskedastic noise with scale depending on location according to a function $\sigma_n : \mathcal{X} \to \mathbb{R}_{\geq 0}$:

special case: independent heteroskedastic noise

$$\mathbf{N} = \operatorname{diag} \sigma_n^2(\mathbf{x}). \tag{2.17}$$

For a given observation location x, we will simply write σ_n for the associated noise scale, leaving any dependence on x implicit.

observation noise scale, σ_n

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

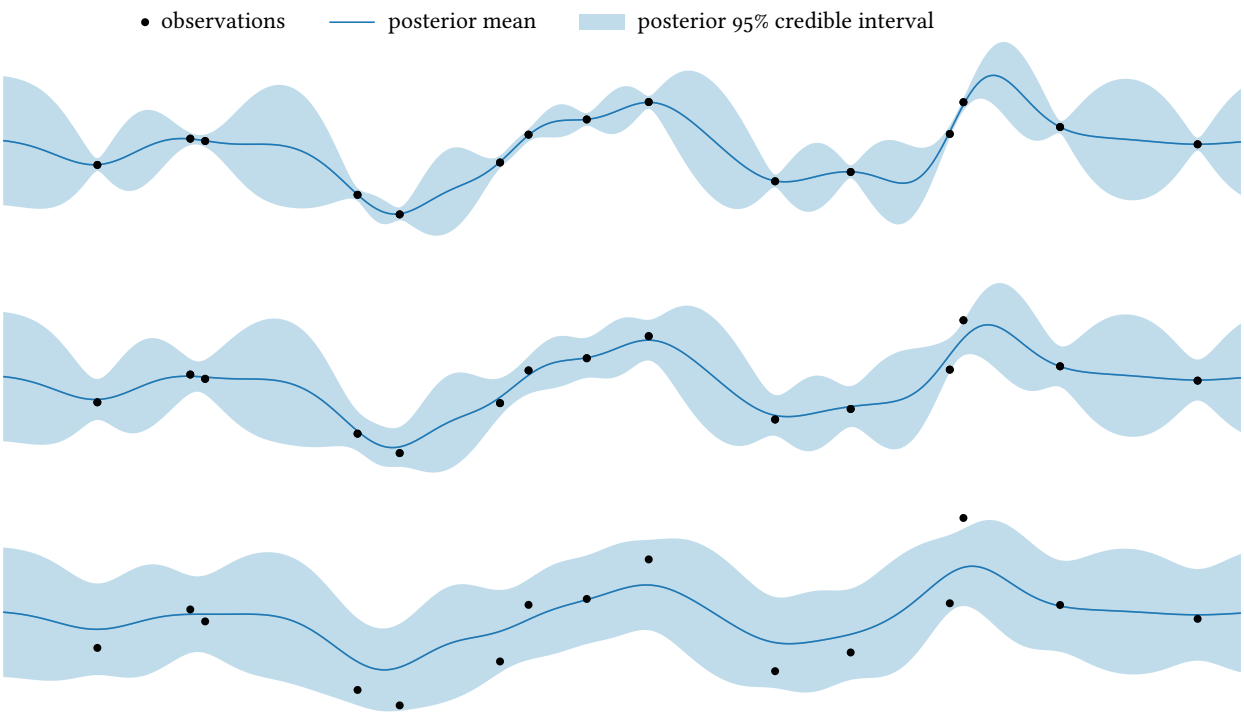


Figure 2.3: Posteriors for our example GP from Figure 2.1 conditioned on 15 noisy observations with independent homoskedastic noise (2.16). The signal-to-noise ratio is 10 for the top example, 3 for the middle example, and 1 for the bottom example.

The prior distribution of the observations is now multivariate normal (2.3, A.15):

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{N}) = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}, \Sigma + \mathbf{N}). \tag{2.18}$$

Due to independence of the noise, the cross-covariance remains the same as in the exact observation case:

$$\kappa(x) = \operatorname{cov}[\mathbf{y}, \phi \mid \mathbf{x}, x] = K(\mathbf{x}, x).$$

Conditioning on the observed value now yields a GP posterior with

$$\begin{aligned} \mu_{\mathcal{D}}(x) &= \mu(x) + K(x, \mathbf{x})(\Sigma + \mathbf{N})^{-1}(\mathbf{y} - \boldsymbol{\mu}); \\ K_{\mathcal{D}}(x, x') &= K(x, x') - K(x, \mathbf{x})(\Sigma + \mathbf{N})^{-1} K(\mathbf{x}, x'). \end{aligned} \tag{2.19}$$

homoskedastic example and discussion

Figure 2.3 shows a sequence of posterior distributions resulting from conditioning our example GP on data corrupted by increasing levels of homoskedastic noise (2.16). As the noise level increases, the observations have diminishing influence on our belief, with some extreme values eventually being partially explained away as outliers. As measurements are assumed to be inexact, the posterior mean is not compelled to interpolate perfectly through the observations, as in the exact case (Figure

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

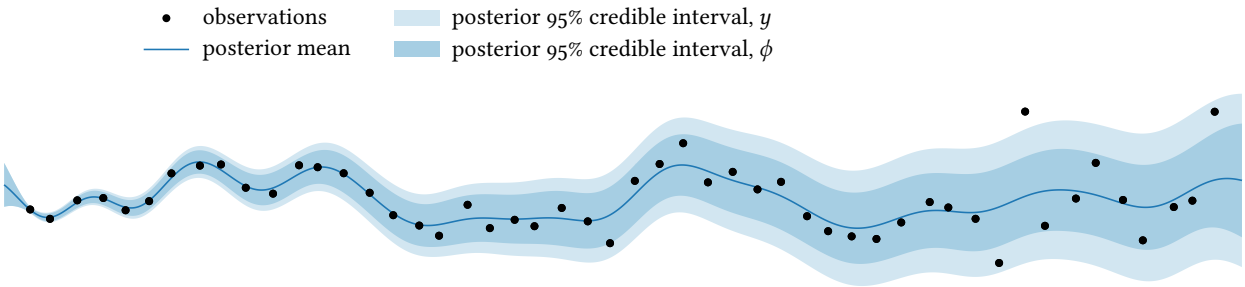


Figure 2.4: The posterior distribution for our example GP from Figure 2.1 conditioned on 50 observations with heteroskedastic observation noise (2.17). We show predictive credible intervals for both the latent objective function and noisy observations; the standard deviation of the observation noise increases linearly from left-to-right.

2.2). Further, with increasing levels of noise, our posterior belief reflects significant residual uncertainty in the function, even in regions with multiple nearby observations.

heteroskedastic example and discussion

We illustrate an example of Gaussian process inference with heteroskedastic noise (2.17) in Figure 2.4, where the signal-to-noise ratio decreases smoothly from left-to-right over the domain. Although the observations provide relatively even coverage, our posterior uncertainty is minimal on the left-hand side of the domain – where the measurements provide maximal information – and increases as our observations become more noisy and less informative.

posterior predictive distribution for noisy observations

We will often require the posterior predictive distribution for a noisy measurement y that would result from observing at a given location x. The posterior distribution on f (2.19) provides the posterior predictive distribution for the latent function value $\phi = f(x)$ (2.5):

$$p(\phi \mid x, \mathcal{D}) = \mathcal{N}(\phi; \mu, \sigma^2); \qquad \mu = \mu_{\mathcal{D}}(x); \qquad \sigma^2 = K_{\mathcal{D}}(x, x),$$

but does not account for the effect of observation noise. In the case of independent additive Gaussian noise (2.16–2.17), deriving the posterior predictive distribution is trivial; we have (A.15):

$$p(y \mid x, \mathcal{D}, \sigma_n) = \mathcal{N}(y; \mu, \sigma^2 + \sigma_n^2). \tag{2.20}$$

This predictive distribution is illustrated in Figure 2.4; the credible intervals for noisy measurements reflect inflation of the credible intervals for the underlying function value commensurate with the scale of the noise.

predictive distribution with correlated noise

If the noise contains nondiagonal correlation structure, we must account for dependence between training and test errors in the predictive distribution. The easiest way to proceed is to recognize that the noisy observation process $y = \phi + \varepsilon$, as a function of x, is itself a Gaussian process with mean function μ and covariance function

covariance function for noisy measurements, C

$$C(x, x') = \operatorname{cov}[y, y' \mid x, x'] = K(x, x') + N(x, x'),$$

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

covariance function for observation noise, N

where N is the noise covariance: $N(x, x') = \text{cov}[\varepsilon, \varepsilon' \mid x, x']$. The posterior of the observation process is then a GP with

$$\begin{aligned} \mathbb{E}[y \mid x, \mathcal{D}] &= \mu(x) + C(x, \mathbf{x})(\Sigma + \mathbf{N})^{-1}(\mathbf{y} - \boldsymbol{\mu}); \\ \text{cov}[y, y' \mid x, x', \mathcal{D}] &= C(x, x') - C(x, \mathbf{x})(\Sigma + \mathbf{N})^{-1}C(\mathbf{x}, x'), \end{aligned} \tag{2.21}$$

from which we can derive predictive distributions via (2.2).

2.3 OVERVIEW OF REMAINDER OF CHAPTER

In the remainder of this chapter we will cover some additional, somewhat niche and/or technical aspects of Gaussian processes that see occasional use in Bayesian optimization. Modulo mathematical nuances irrelevant in practical settings, an *intuitive* (but not entirely accurate!) summary follows:[20]

20 In particular the claims regarding continuity and differentiability are slightly more complicated than stated below.

- a *joint Gaussian process* (discussed below) allows us to model *multiple* related functions simultaneously, which is critical for some scenarios such as multifidelity and multiobjective optimization;

multifidelity optimization: § 11.5, p. 263
multiobjective optimization: § 11.7, p. 269

- GP sample paths are continuous if the mean function is continuous and the covariance function is continuous along the "diagonal" $x = x'$;

sample path continuity: § 2.5, p. 28

- GP sample paths are differentiable if the mean function is differentiable and the covariance function is differentiable along the "diagonal" $x = x'$;

sample path differentiability: § 2.6, p. 30

- a function with a sufficiently smooth GP distribution shares a joint GP distribution with its gradient; among other things, this allows us to condition on (potentially noisy) derivative observations via exact inference;

derivative observations: § 2.6, p. 32

- GP sample paths attain a maximum when sample paths are continuous and the domain is compact;

existence of global maxima: § 2.7, p. 34

- GP sample paths attain a *unique* maximum under the additional condition that no two unique function values are perfectly correlated; and

uniqueness of global maxima: § 2.7, p. 34

- several methods are available for approximating the posterior process of a GP conditioned on information incompatible with exact inference.

inference with non-Gaussian observations and constraints: § 2.8, p. 35

If satisfied with the above summary, the reader may safely skip this material for now and move on with the next chapter. For those who wish to see the gritty details, dive in below!

2.4 JOINT GAUSSIAN PROCESSES

In some settings, we may wish to reason *jointly* about two-or-more related functions, such as an objective function and its gradient or an expensive objective function and a cheaper surrogate. To this end we can extend Gaussian processes to yield a joint distribution over the values assumed by multiple functions. The key to the construction is to "paste together" a collection of functions into a single function on a larger domain, then construct a standard GP on this combined function.

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

Definition

To elaborate, consider a set of functions $\{f_i \colon \mathcal{X}_i \to \mathbb{R}\}$ we wish to model.[21] We define the *disjoint union* of these functions $\sqcup f$ – defined on the disjoint union[22] of their domains $\mathcal{X} = \bigsqcup \mathcal{X}_i$ – by insisting its restriction to each domain be compatible with the corresponding function:

disjoint union of $\{f_i\}$, $\sqcup f$
disjoint union of $\{\mathcal{X}_i\}$, $\mathcal{X}$

$$\sqcup f \colon \mathcal{X} \to \mathbb{R}; \qquad \sqcup f|_{\mathcal{X}_i} \equiv f_i.$$

21 The domains need not be equal, but they often are in practice.

We now can define a GP on $\sqcup f$ by choosing mean and covariance functions on $\mathcal{X}$ as desired:

22 A disjoint union represents a point $x \in \mathcal{X}_i$ by the pair (x, i), thereby combining the domains while retaining their identities.

$$p(\sqcup f) = \mathcal{GP}(\sqcup f; \mu, K). \tag{2.22}$$

We will call this construction a *joint Gaussian process* on $\{f_i\}$.

joint Gaussian process

It is often convenient to decompose the moments of a joint GP into their restrictions on relevant subspaces. For example, consider a joint GP (2.22) on $f \colon \mathcal{F} \to \mathbb{R}$ and $g \colon \mathcal{G} \to \mathbb{R}$. After defining

$$\begin{aligned} &\mu_f \equiv \mu|_{\mathcal{F}}; \qquad \mu_g \equiv \mu|_{\mathcal{G}}; \\ &K_f \equiv K|_{\mathcal{F}\times\mathcal{F}}; \quad K_g \equiv K|_{\mathcal{G}\times\mathcal{G}}; \quad K_{fg} \equiv K|_{\mathcal{F}\times\mathcal{G}}; \quad K_{gf} \equiv K|_{\mathcal{G}\times\mathcal{F}}, \end{aligned}$$

we can see that f and g in fact have marginal GP distributions:[23]

23 In fact, *any* restriction of a GP-distributed function has a GP (or multivariate normal) distribution.

$$p(f) = \mathcal{GP}(f; \mu_f, K_f); \qquad p(g) = \mathcal{GP}(g; \mu_g, K_g), \tag{2.23}$$

that are coupled by the *cross-covariance functions* K_{fg} and K_{gf}. Given vectors $\mathbf{x} \subset \mathcal{F}$ and $\mathbf{x}' \subset \mathcal{G}$, these compute the covariance between the corresponding function values $\boldsymbol{\phi} = f(\mathbf{x})$ and $\boldsymbol{\gamma} = g(\mathbf{x}')$:

$$\begin{aligned} K_{fg}(\mathbf{x}, \mathbf{x}') &= \operatorname{cov}[\boldsymbol{\phi}, \boldsymbol{\gamma} \mid \mathbf{x}, \mathbf{x}']; \\ K_{gf}(\mathbf{x}, \mathbf{x}') &= \operatorname{cov}[\boldsymbol{\gamma}, \boldsymbol{\phi} \mid \mathbf{x}, \mathbf{x}'] = K_{fg}(\mathbf{x}, \mathbf{x}')^\top \end{aligned} \tag{2.24}$$

When convenient we will notate a joint GP in terms of these decomposed functions, here writing:[24]

24 We also used this notation in (2.6), where the "domain" of the vector $\mathbf{y}$ can be taken to be some finite index set of appropriate size.

$$p(f, g) = \mathcal{GP}\left(\begin{bmatrix} f \\ g \end{bmatrix}; \begin{bmatrix} \mu_f \\ \mu_g \end{bmatrix}, \begin{bmatrix} K_f & K_{fg} \\ K_{gf} & K_g \end{bmatrix}\right). \tag{2.25}$$

With this notation, the marginal GP property (2.23) is perfectly analogous to the marginal property of the multivariate Gaussian distribution (A.13).

We can also use this construction to define a GP on a *vector*-valued function $\mathbf{f} \colon \mathcal{X} \to \mathbb{R}^d$ by defining a joint Gaussian process on its d coordinate functions $\{f_i\} \colon \mathcal{X} \to \mathbb{R}$. In this case we typically write the resulting model using the standard notation $\mathcal{GP}(\mathbf{f}; \mu, K)$, where the mean and covariance functions are now understood to map to $\mathbb{R}^d$ and $\mathbb{R}^{d\times d}$.

extension to vector-valued functions

Example

We can demonstrate the behavior of a joint Gaussian process by extending our running example GP on $f \colon [0, 30] \to \mathbb{R}$. Recall the prior on f has

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

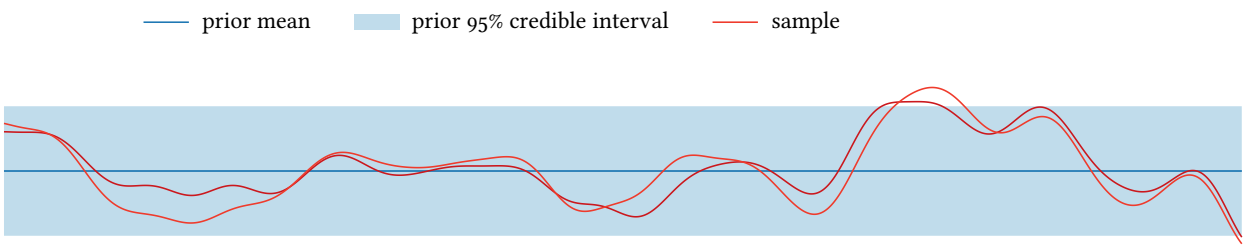


Figure 2.5: A joint Gaussian process over two functions on the shared domain $\mathcal{X} = [0, 30]$. The marginal belief over both functions is the same as our example GP from Figure 2.1, but the cross-covariance (2.26) between the functions strongly couples their behavior. We also show a sample from the joint distribution illustrating the strong correlation induced by the joint prior.

zero mean function $\mu \equiv 0$ and squared exponential covariance function (2.4). We augment our original function with a companion function g, defined on the same domain, that has exactly the same marginal GP distribution. However, we couple the distribution of f and g by defining a nontrivial cross-covariance function K_{fg} (2.24):

$$K_{fg}(x, x') = 0.9K(x, x'), \tag{2.26}$$

where K is the marginal covariance function of f and g. A consequence of this choice is that for any given point $x \in \mathcal{X}$, the correlation of the corresponding function values $\phi = f(x)$ and $\gamma = g(x)$ is quite strong:

$$\operatorname{corr}[\phi, \gamma \mid x] = 0.9. \tag{2.27}$$

We illustrate the resulting joint GP in Figure 2.5. The marginal credible intervals for f (and now g) have not changed from our original example in Figure 2.1. However, drawing a sample of the functions from their joint distribution reveals the strong coupling encoded in the prior (2.26–2.27).

Inference for joint Gaussian processes

The construction in (2.22) allows us to reason about a joint Gaussian process as if it were a single GP. This allows us to condition a joint GP on observations of jointly Gaussian distributed values following the procedure outlined previously. In Figure 2.6, we condition the joint GP prior from Figure 2.5 on ten observations: five exact observations of f on the left-hand side of the domain and five exact observations of g on the right-hand side. Due to the strong correlation between the two functions, an observation of either function strongly informs our belief about the other, even in regions where there are no direct observations.

inference from jointly Gaussian distributed observations: § 2.2, p. 18

2.5 CONTINUITY

In this and the following sections we will establish some important properties of Gaussian processes determined by the properties of their

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

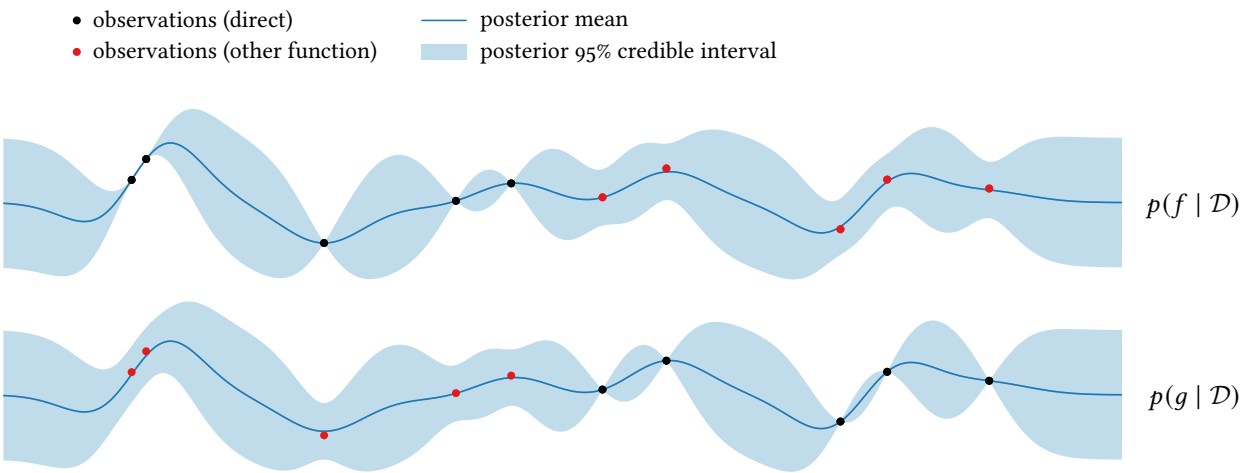


Figure 2.6: The joint posterior for our example joint GP prior in Figure 2.5 conditioned on five exact observations of each function.

moments. As a GP is completely specified by its mean and covariance functions, it should not be surprising that the nature of these functions has far-reaching implications regarding properties of the function being modeled. A good familiarity with these implications can help guide model design in practice – the focus of the next two chapters.

To begin, a fundamental question regarding Gaussian processes is whether sample paths are almost surely continuous, and if so how many times differentiable they may be. This is obviously an important consideration for modeling and is also critical to ensure that global optimization is a well-posed problem, as we will discuss later in this chapter. Fortunately, continuity of Gaussian processes is a well-understood property that can be guaranteed almost surely under simple conditions on the mean and covariance functions.

existence of global maxima: § 2.7, p. 34

Suppose $f\colon \mathcal{X} \to \mathbb{R}$ has distribution $\mathcal{GP}(f; \mu, K)$. Recall that f is continuous at x if $f(x) - f(x') = \phi - \phi' \to 0$ when $x' \to x$. Continuity is thus a limiting property of differences in function values. But under the Gaussian process assumption, this difference is Gaussian distributed (2.5, A.9)! We have

$$p(\phi - \phi' \mid x, x') = \mathcal{N}(\phi - \phi'; m, s^2),$$

where

$$m = \mu(x) - \mu(x'); \qquad s^2 = K(x,x) - 2K(x,x') + K(x',x').$$

Now if μ is continuous at x and K is continuous at $x = x'$, then both $m \to 0$ and $s^2 \to 0$ as $x \to x'$, and thus $\phi - \phi'$ converges in probability to 0. This intuitive condition of continuous moments is known as *continuity in mean square* at x; if μ and K are both continuous over the entire domain (the latter along the "diagonal" $x = x'$), then we say the entire process is continuous in mean square.

continuity in mean square

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

sample path continuity

It turns out that continuity in mean square is not quite sufficient to guarantee that f is simultaneously continuous at every $x \in \mathcal{X}$ with probability one, a property known as *sample path continuity*. However, very slightly stronger conditions on the moments of a GP are sufficient to guarantee sample path continuity.[25] The following result is adequate for most settings arising in practice and may be proven as a corollary to the slightly weaker (and slightly more complicated) conditions assumed in ADLER and TAYLOR's theorem 1.4.1.

25 R. J. ADLER and J. E. TAYLOR (2007). *Random Fields and Geometry*. Springer–Verlag. [§§ 1.3–1.4]

Theorem. *Suppose $\mathcal{X} \subset \mathbb{R}^d$ is compact and $f\colon \mathcal{X} \to \mathbb{R}$ has Gaussian process distribution $\mathcal{GP}(f; \mu, K)$, where μ is continuous and K is Hölder continuous.*[26] *Then f is almost surely continuous on $\mathcal{X}$.*

26 Hölder continuity is a generalization of Lipschitz continuity. Effectively, the covariance function must, in some sense, be "predictably" continuous.

The condition that $\mathcal{X} \subset \mathbb{R}^d$ be compact is equivalent to the domain being closed and bounded, by the Heine–Borel theorem.[27] Applying this result to our example GP in Figure 2.1, we conclude that samples from the process are continuous with probability one as the domain $\mathcal{X} = [0, 30]$ is compact and the squared exponential covariance function (2.4) is Hölder continuous. Indeed, the generated samples are very smooth.[28]

27 W. RUDIN (1976). *Principles of Mathematical Analysis*. McGraw–Hill. [theorem 2.41]

28 Following the discussion in the next section, they in fact are *infinitely* differentiable.

Sample path continuity can also be guaranteed on non-Euclidean domains under similar smoothness conditions.[25]

2.6 DIFFERENTIABILITY

We can approach the question of differentiability by again reasoning about the limiting behavior of linear transformations of function values. Suppose $f\colon \mathcal{X} \to \mathbb{R}$ with $\mathcal{X} \subset \mathbb{R}^d$ has distribution $\mathcal{GP}(f; \mu, K)$, and consider the ith partial derivative of f at $\mathbf{x}$, if it exists:

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h},$$

where $\mathbf{e}_i$ is the ith standard basis vector. For $h > 0$, the value in the limit is Gaussian distributed as a linear transformation of Gaussian-distributed random variables (A.9). Assuming the corresponding partial derivative of the mean exists at $\mathbf{x}$ and the corresponding partial derivative with respect to each input of the covariance function exists at $\mathbf{x} = \mathbf{x}'$, then as $h \to 0$ the partial derivative converges in distribution to a Gaussian:

sequences of normal RVs: § A.2, p. 300

$$p\left(\frac{\partial f}{\partial x_i}(\mathbf{x}) \mid \mathbf{x}\right) = \mathcal{N}\left(\frac{\partial f}{\partial x_i}(\mathbf{x}); \frac{\partial \mu}{\partial x_i}(\mathbf{x}), \frac{\partial^2 K}{\partial x_i \, \partial x_i'}(\mathbf{x}, \mathbf{x})\right).$$

differentiability in mean square

If this property holds for each coordinate $1 \leq i \leq d$, then f is said to be *differentiable in mean square* at $\mathbf{x}$.

joint GP between function and gradient

If f is differentiable in mean square everywhere in the domain, the process itself is called differentiable in mean square, and we have the remarkable result that the function and its gradient have a *joint* Gaussian process distribution:

$$p(f, \nabla f) = \mathcal{GP}\left(\begin{bmatrix} f \\ \nabla f \end{bmatrix}; \begin{bmatrix} \mu \\ \nabla \mu \end{bmatrix}, \begin{bmatrix} K & K\nabla^\top \\ \nabla K & \nabla K \nabla^\top \end{bmatrix}\right). \tag{2.28}$$

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

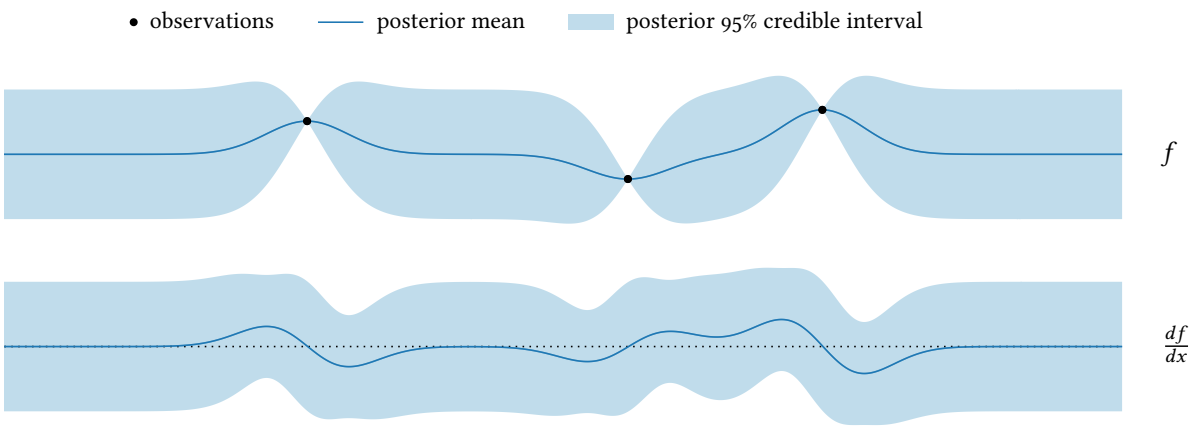


Figure 2.7: The joint posterior of the function and its derivative for our example Gaussian process from Figure 2.2. The dashed line in the lower plot corresponds to a derivative of zero.

Here by writing the gradient operator ∇ on the left-hand side of K we mean the result of taking the gradient with respect to its *first* input, and by writing $\nabla^\top$ on the right-hand side of K we mean taking the gradient with respect to its *second* input and transposing the result. Thus $\nabla K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$ maps pairs of points to column vectors:

covariance between $\nabla f(\mathbf{x})$ and $f(\mathbf{x}')$, ∇K

$$\left[\nabla K(\mathbf{x}, \mathbf{x}')\right]_i = \operatorname{cov}\left[\frac{\partial f}{\partial x_i}(\mathbf{x}), f(\mathbf{x}') \;\middle|\; \mathbf{x}, \mathbf{x}'\right] = \frac{\partial K}{\partial x_i}(\mathbf{x}, \mathbf{x}'),$$

and $K\nabla^\top \colon \mathcal{X} \times \mathcal{X} \to (\mathbb{R}^d)^*$ maps pairs of points to row vectors:

transpose of covariance between $f(\mathbf{x})$ and $\nabla f(\mathbf{x}')$, $K\nabla^\top$

$$K\nabla^\top(\mathbf{x}, \mathbf{x}') = \left[\nabla K(\mathbf{x}', \mathbf{x})\right]^\top.$$

Finally, the function $\nabla K \nabla^\top \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d \times d}$ represents the result of applying both operations, mapping a pair of points to the covariance matrix between the entries of the corresponding gradients:

covariance between $\nabla f(\mathbf{x})$ and $\nabla f(\mathbf{x}')$, $\nabla K \nabla^\top$

$$\left[\nabla K \nabla^\top(\mathbf{x}, \mathbf{x}')\right]_{ij} = \operatorname{cov}\left[\frac{\partial f}{\partial x_i}(\mathbf{x}), \frac{\partial f}{\partial x'_j}(\mathbf{x}') \;\middle|\; \mathbf{x}, \mathbf{x}'\right] = \frac{\partial^2 K}{\partial x_i \, \partial x'_j}(\mathbf{x}, \mathbf{x}').$$

As the gradient of f has a Gaussian process marginal distribution (2.28), we can reduce the question of *continuous* differentiability to sample path continuity of the gradient process following the discussion above.

continuous differentiability

Figure 2.7 shows the posterior distribution for the derivative of our example Gaussian process alongside the posterior for the function itself. We can observe a clear correspondence between the two distributions; for example, the posterior mean of the derivative vanishes at critical points of the posterior mean of the function. Notably, we have a great deal of residual uncertainty about the derivative, even at the observed locations. That is because the relatively high spacing between the existing observations limits our ability to accurately estimate the derivative

example and discussion

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

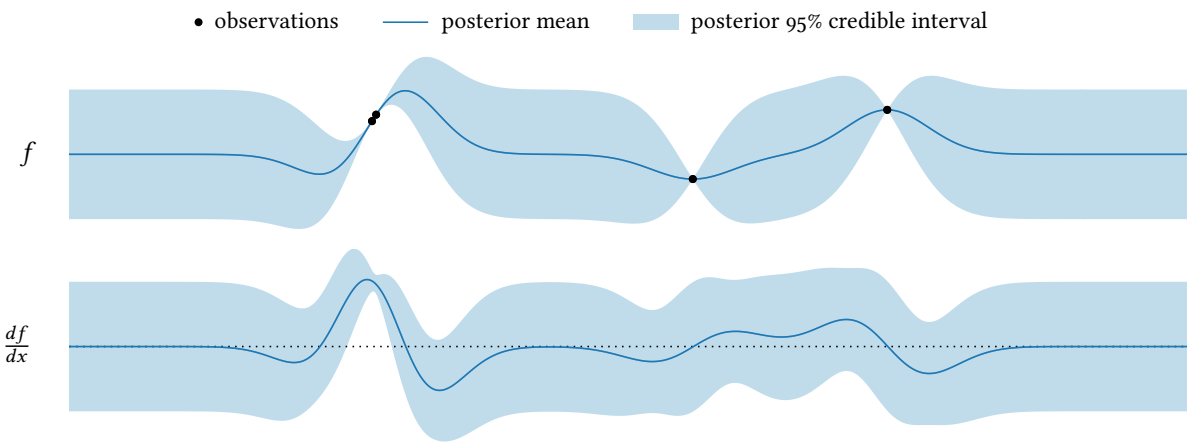


Figure 2.8: The joint posterior of the derivative of our example Gaussian process after adding a new observation nearby another suggesting a large positive slope. The dashed line in the lower plot corresponds to a derivative of zero.

anywhere. Adding an observation immediately next to a previous one significantly reduces the uncertainty in the derivative in that region by effectively providing a finite-difference approximation; see Figure 2.8.

Conditioning on derivative observations

However, we can be more direct in specifying derivatives than finite differencing. We can instead condition the joint GP (2.28) *directly* on a derivative observation, as described previously. Figure 2.9 shows the joint posterior after conditioning on an exact observation of the derivative at the left-most observation location, where the uncertainty in the derivative now vanishes entirely. This capability allows the seamless incorporation of derivative information into an objective function model. Notably, we can even condition a Gaussian process on *noisy* derivative observations as well, as we might obtain in stochastic gradient descent.

inference from jointly Gaussian distributed observations: § 2.2, p. 18

We can reason about derivatives past the first recursively. For example, if μ and K are *twice* differentiable,[29] then the (e.g., half-vectorized[30]) Hessian of f will also have a joint GP distribution with f and its gradient. Defining $\mathbf{h}$ to be the operator mapping a function to its half-vectorized Hessian:

$$\mathbf{h}f = \operatorname{vech} \nabla\nabla^\top f,$$

for a Gaussian process with suitably differentiable moments, we have

$$p(\mathbf{h}f) = \mathcal{GP}\big(\mathbf{h}f; \mathbf{h}\mu, \mathbf{h}K\mathbf{h}^\top\big), \tag{2.29}$$

where we have used the same notational convention for the transpose. Further, f, ∇f, and $\mathbf{h}f$ will have a joint Gaussian process distribution given by augmenting (2.28) with the marginal in (2.29) and the cross-covariance functions

$$\operatorname{cov}[\mathbf{h}f, f] = \mathbf{h}K; \qquad \operatorname{cov}[\mathbf{h}f, \nabla f] = \mathbf{h}K\nabla.^\top$$

29 For K we again only need to consider the "diagonal" $\mathbf{x} = \mathbf{x}'$.

30 Recall the Hessian is symmetric (assuming the second partial derivatives are continuous) and thus redundant. The *half-vectorization* operator vech $\mathbf{A}$ maps the upper triangular part of a square, symmetric matrix $\mathbf{A}$ to a vector.

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

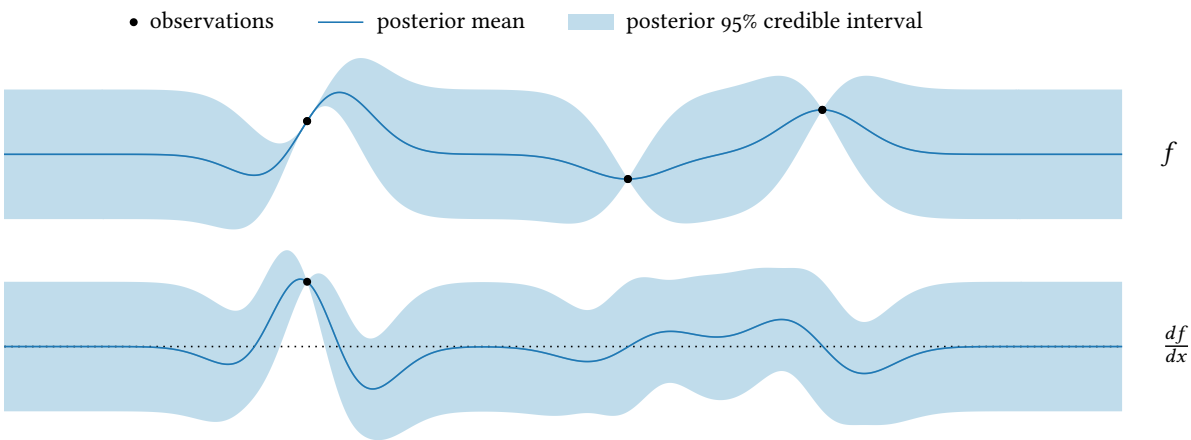


Figure 2.9: The joint posterior of the derivative of our example Gaussian process after adding an exact observation of the derivative at the indicated location. The dashed line in the lower plot corresponds to a derivative of zero.

We can continue further in this vein if needed; however, we rarely reason about derivatives of third-or-higher order in Bayesian optimization.[31]

31 This is true in classical optimization as well!

Other linear transformations

The joint GP distribution between a suitably smooth GP-distributed function and its gradient (2.28) is simply an infinite-dimensional analog of the general result that Gaussian random variables are jointly Gaussian distributed with arbitrary linear transformations (A.10), after noting that differentiation is a linear operator. We can extend this result to reason about other linear transformations of GP-distributed functions. DIACONIS's original motivation for studying Bayesian numerical methods was *quadrature*, the numerical estimation of intractable integrals.[32] It turns out that Gaussian processes are a rather convenient model for this task: if $p(f) = \mathcal{GP}(f; \mu, K)$ and we want to reason about the expectation

$$Z = \int f(x)\, p(x)\, \mathrm{d}x,$$

then (under mild conditions) we again have a joint Gaussian process distribution over f and Z.[33] This enables both inference about Z and conditioning on noisy observations of integrals, such as a Monte Carlo estimate of an expectation. The former is the basis for *Bayesian quadrature*, an analog of Bayesian optimization bringing Bayesian experimental design to bear on numerical integration.[32,34,35]

32 P. DIACONIS (1988). Bayesian Numerical Analysis. In: *Statistical Decision Theory and Related Topics IV*.

33 This can be shown, for example, by considering the limiting distribution of Riemann sums.

34 A. O'HAGAN (1991). Bayes–Hermite Quadrature. *Journal of Statistical Planning and Inference* 29(3):245–260.

35 C. E. RASMUSSEN and Z. GHAHRAMANI (2002). Bayesian Monte Carlo. *NeurIPS 2002*.

2.7 EXISTENCE AND UNIQUENESS OF GLOBAL MAXIMA

The primary use of GPs in Bayesian optimization is to inform optimization decisions, which will be our focus for the majority of this book. Before continuing down this path, we pause to consider whether global

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

optimization of a GP-distributed function is a well-posed problem, in particular, whether the model guarantees the existence of a global maximum at all.

Consider a function $f\colon \mathcal{X} \to \mathbb{R}$ with distribution $\mathcal{GP}(f; \mu, K)$, and consider the location and value of its global optimum, if one exists:

$$x^* = \arg\max_{x \in \mathcal{X}} f(x); \qquad f^* = \max_{x \in \mathcal{X}} f(x) = f(x^*).$$

mutual information and entropy search: § 7.6, p. 135

As f is unknown, these quantities are random variables. Many Bayesian optimization algorithms operate by reasoning about the distributions of (and uncertainties in) these quantities induced by our belief on f.

There are two technical issues we must address. The first is whether we can be certain that a globally optimal value f^* exists when the objective function is random. If existence is not guaranteed, then its distribution is meaningless. The second issue is one of uniqueness: assuming the objective does attain a maximal value, can we be certain the optimum is unique? In general x^* is a *set*-valued random variable, and thus its distribution might have support over arbitrary subsets of the domain, rendering it complicated to reason about. However, if we could ensure the uniqueness of x^*, its distribution would have support on $\mathcal{X}$ rather than its power set, allowing more straightforward inference.

Both the existence of f^* and uniqueness of x^* are tacitly assumed throughout the Bayesian optimization literature when building algorithms based on distributions of these quantities, but these properties are not guaranteed for arbitrary Gaussian processes. However, we can ensure these properties hold almost surely under mild conditions.

Existence of global maxima

To begin, guaranteeing the existence of an optimal value is straightforward if we suppose the domain $\mathcal{X}$ is compact, a pervasive assumption in optimization. This is no coincidence! In this case, if f is continuous then it achieves a global optimum by the extreme value theorem.[36] Thus sample path continuity of f and compactness of $\mathcal{X}$ is sufficient to ensure that f^* exists almost surely. Both conditions can be readily established: sample path continuity by following our previous discussion, and compactness of the domain by standard arguments (for example, ensuring that $\mathcal{X} \subset \mathbb{R}^d$ be closed and bounded).

36 W. RUDIN (1976). *Principles of Mathematical Analysis*. McGraw–Hill. [theorem 4.16]

sample path continuity: § 2.5, p. 28

Uniqueness of global maxima

We now turn to the question of uniqueness of x^*, which obviously only becomes a meaningful question after presupposing that f^* exists. Again, this condition is easy to ensure almost surely under simple conditions on the covariance function of a Gaussian process.

KIM and POLLARD considered this issue and provided straightforward conditions under which the uniqueness of x^* is guaranteed for a centered Gaussian process.[37,38] Namely, no two unique points in the domain can

37 A centered Gaussian process has identically zero mean function $\mu \equiv 0$.

38 J. KIM and D. POLLARD (1990). Cube Root Asymptotics. *The Annals of Statistics* 18(1):191–219. [lemma 2.6]

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

have perfectly correlated function values, a natural condition that can be easily verified.

Theorem (KIM and POLLARD, 1990). *Let $\mathcal{X}$ be a compact metric space.*[39] *Suppose $f : \mathcal{X} \to \mathbb{R}$ has distribution $\mathcal{GP}(f; \mu \equiv 0, K)$, and that f is sample path continuous. If for all $x, x' \in \mathcal{X}$ with $x \neq x'$ we have*

$$\operatorname{var}[\phi - \phi' \mid x, x'] = K(x, x) - 2K(x, x') + K(x', x') \neq 0,$$

then f almost surely has a unique maximum on $\mathcal{X}$.

39 Although unlikely to matter in practice, KIM and POLLARD allow $\mathcal{X}$ to be σ-compact and show that the supremum (rather than the maximum) is unique under the same conditions.

ARCONES provided slightly weaker conditions for uniqueness of the supremum, avoiding the requirement of sample path continuity.[40]

40 M. A. ARCONES (1992). On the arg max of a Gaussian Process. *Statistics & Probability Letters* 15(5):373–374.

Counterexamples

Although the above conditions for ensuring existence of f^* and uniqueness of x^* are fairly mild, it is easy to construct counterexamples.

Consider a function on the closed unit interval, which we note is compact: $f : [0, 1] \to \mathbb{R}$. We endow f with a "white noise"[41] Gaussian process with

$$\mu(x) \equiv 0; \qquad K(x, x') = [x = x'].$$

41 It turns out this naïve model of white noise has horrible mathematical properties, but it is sufficient for this counterexample.

Now f almost surely does not have a maximum. Roughly, because the value of f at every point in the domain is independent of every other, there will almost always be a point with value exceeding any putative maximum.[42] However, the conditions of sample path continuity were violated as the covariance is discontinuous at $x = x'$.

We may also construct a Gaussian process that almost surely achieves a maximum that is not unique. Consider a random function f defined on the (compact) interval $[0, 4\pi]$ defined by the parametric model

$$f(x) = \alpha \cos x + \beta \sin x,$$

where α and β are independent standard normal random variables. Then f has a Gaussian process distribution with

$$\mu(x) \equiv 0; \qquad K(x, x') = \cos(x - x'). \tag{2.30}$$

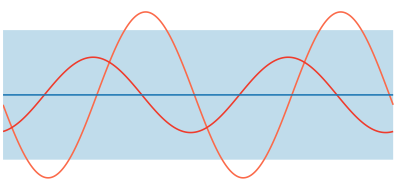
Here μ is continuous and K is Hölder continuous, and thus f is sample path continuous and almost surely achieves a global maximum. However, f is also periodic with period 2π with probability one and will thus almost surely achieve its maximum *twice*. Note that the covariance function does not satisfy the conditions outlined in the above theorem, as any input locations separated by 2π have perfectly correlated function values.

42 Let $Q = \mathbb{Q} \cap [0, 1] = \{q_i\}$ be the rationals in the domain and let f^* be a putative maximum. Defining $\phi_i = f(q_i)$, we must have $\phi_i \leq f^*$ for every i; call this event A.

Define the event A_k by f^* exceeding the first k elements of Q. From independence,

$$\Pr(A_k) = \prod_{i=1}^{k} \Pr(\phi_i \leq f^*) = \Phi(f^*)^k,$$

so $\Pr(A_k) \to 0$ as $k \to \infty$. But $\{A_k\} \nearrow A$, so $\Pr(A) = 0$, and f^* is almost surely not the maximum.



Our counterexample GP without a unique maximum. Every sample achieves its maximum twice.

2.8 INFERENCE WITH NON-GAUSSIAN OBSERVATIONS AND CONSTRAINTS

Gaussian process inference is tractable when the observed values are jointly Gaussian distributed with the function of interest (2.6). However, this may not always hold for all relevant information we may receive.

inference from jointly Gaussian distributed observations: § 2.2, p. 18

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

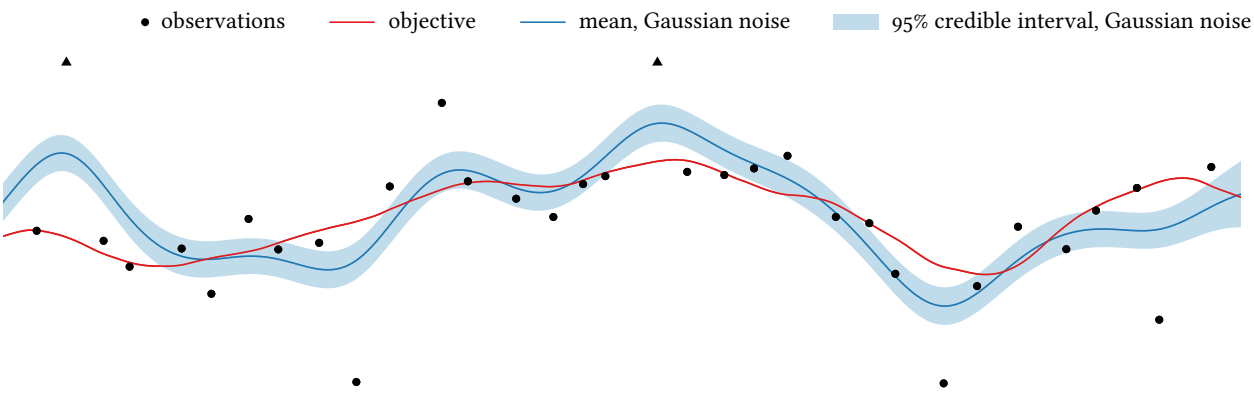
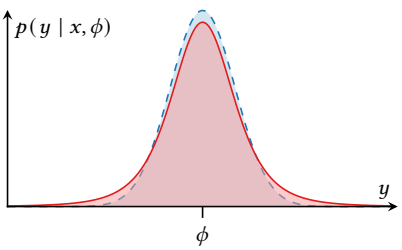


Figure 2.10: Regression with observations corrupted with heavy-tailed noise. The triangular marks indicate observations lying beyond the plotted range. Shown is the posterior distribution of an objective function (along with ground truth) modeling the errors as Gaussian. The posterior is heavily affected by the outliers.



A Student-t error model (solid) with a Gaussian error model (dashed) for reference. The heavier tails of the Student-t model can better explain large outliers.

One obvious limitation is an incompatibility with naturally non-Gaussian observations. A scenario particularly relevant to optimization is heavy-tailed noise. Consider the data shown in Figure 2.10, where some observations represent extreme outliers. These errors are poorly modeled as Gaussian, and attempting to infer the underlying objective function with the additive Gaussian noise model leads to overfitting and poor predictive performance. A Student-t error model with $\nu \approx 4$ degrees of freedom provides a robust alternative:[43]

$$p(y \mid x, \phi) = \mathcal{T}(y; \phi, \sigma_n^2, \nu). \tag{2.31}$$

The heavier tails of this model can better explain large outliers; unfortunately, the non-Gaussian nature of this model also renders exact inference impossible. We will demonstrate how to overcome this impasse.

Constraints on an objective function, such as bounds on given function values, can also provide valuable information during optimization, but many natural constraints cannot be reduced to observations that can be handled in closed form. Several Bayesian optimization policies impose hypothetical constraints on the objective function when designing each observation, requiring inference from intractable constraints even when the observations themselves pose no difficulties.

To see how constraints might arise in optimization, consider a Gaussian process belief on a one-dimensional objective f, and suppose we wish to condition on f on having a *local* maximum at a given location x. Assuming the function is twice differentiable, we can invoke the second-derivative test to encode this information in two constraints:

$$f'(x) = 0; \qquad f''(x) < 0. \tag{2.32}$$

We can condition a GP on the first of these conditions by following our previous discussion. However, *no* GP is compatible with the second

43 K. L. LANGE et al. (1989). Robust Statistical Modeling Using the t Distribution. *Journal of the American Statistical Association* 84(408):881–896.

differentiability, derivative observations: § 2.6, p. 30

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

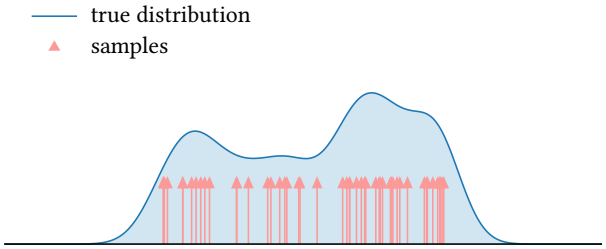


Figure 2.11: The probability density function of an example distribution along with 50 samples drawn independently from the distribution. In Monte Carlo approaches, the distribution is effectively approximated by a mixture of Dirac delta distributions at the sample locations.

condition as $f''(x)$ would necessarily have a Gaussian distribution with unbounded support (2.29). We need some other means to proceed.

Non-Gaussian observations: general case

We can address both non-Gaussian observations and constraints with the following general case, which is flexible enough to handle a large range of information. As in our discussion on exact inference, suppose there is some vector $\mathbf{y}$ sharing a joint Gaussian process distribution with a function of interest f (2.6):

$$p(f, \mathbf{y}) = \mathcal{GP}\left(\begin{bmatrix} f \\ \mathbf{y} \end{bmatrix}; \begin{bmatrix} \mu \\ \mathbf{m} \end{bmatrix}, \begin{bmatrix} K & \kappa^\top \\ \kappa & \mathbf{C} \end{bmatrix}\right).$$

Suppose we receive some information about $\mathbf{y}$ in the form of information $\mathcal{D}$ inducing a non-Gaussian posterior on $\mathbf{y}$. Here, it is convenient to adopt the language of factor graphs[44] and write the resulting posterior as proportional to the prior weighted by a function $t(\mathbf{y})$ encoding the available information, which may factorize:

$$p(\mathbf{y} \mid \mathcal{D}) \propto p(\mathbf{y})\, t(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mathbf{m}, \mathbf{C}) \prod_i t_i(\mathbf{y}). \tag{2.33}$$

The functions $\{t_i\}$ are called *factors* or *local functions* that may comprise a likelihood augmented by any desired (hard or soft) constraints. The term "local functions" arises because each factor often depends only on a low-dimensional subspace of $\mathbf{y}$, often a single entry.[45]

The posterior on $\mathbf{y}$ (2.33) in turn induces a posterior on f:

$$p(f \mid \mathcal{D}) = \int p(f \mid \mathbf{y})\, p(\mathbf{y} \mid \mathcal{D})\, \mathrm{d}\mathbf{y}. \tag{2.34}$$

At first glance, we may hope to resolve this posterior easily as $p(f \mid \mathbf{y})$ is a Gaussian process (2.9–2.10). Unfortunately, the non-Gaussian posterior on $\mathbf{y}$ usually renders the posterior on f intractable.

Monte Carlo sampling

A Monte Carlo approach to approximating the f posterior (2.34) begins by drawing samples from the $\mathbf{y}$ posterior (2.33):

$$\{\mathbf{y}_i\}_{i=1}^s \sim p(\mathbf{y} \mid \mathcal{D}).$$

44 F. R. KSCHISCHANG et al. (2001). Factor Graphs and the Sum–Product Algorithm. *IEEE Transactions on Information Theory* 47(2):498–519.

factors, local functions, $\{t_i\}$

45 For example, when observations are conditionally independent given the corresponding function values, the likelihood factorizes into a product of one-dimensional factors (1.3):

$$p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\phi}) = \prod_i p(y_i \mid x_i, \phi_i).$$

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

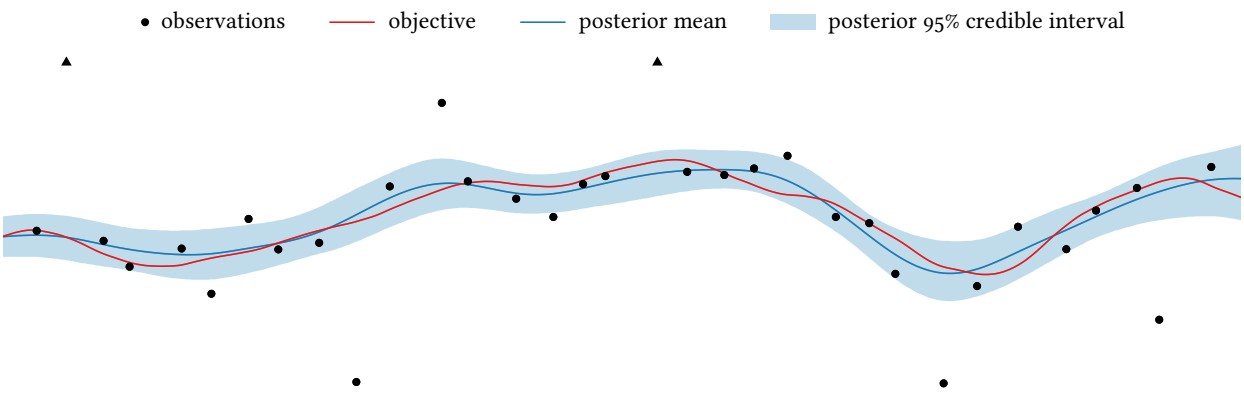


Figure 2.12: Regression with observations corrupted with heavy-tailed noise. The triangular marks indicate observations lying beyond the plotted range. Shown is the posterior distribution of an objective function (along with ground truth) modeling the errors as Student-t distributed with $\nu = 4$ degrees of freedom. The posterior was approximated from 100 000 Monte Carlo samples. Comparing with the additive Gaussian noise model from Figure 2.10, this model effectively ignores the outliers and the fit is excellent.

We may generate these by appealing to one of numerous Markov chain Monte Carlo (MCMC) routines.[46] One natural choice would be *elliptical slice sampling*,[47] which is specifically tailored for latent Gaussian models of this form. Samples from a one-dimensional toy example distribution are shown in Figure 2.11.

Given posterior samples of $\mathbf{y}$, we may then approximate (2.34) via the standard Monte Carlo estimator

$$p(f \mid \mathcal{D}) \approx \frac{1}{s}\sum_{i=1}^{s} p(f \mid \mathbf{y}_i) = \frac{1}{s}\sum_{i=1}^{s} \mathcal{GP}(f; \mu_{\mathcal{D}_i}, K_{\mathcal{D}}). \tag{2.35}$$

This is a mixture of Gaussian processes, each of the form in (2.9–2.10). The posterior mean functions depend on the corresponding $\mathbf{y}$ samples, whereas the posterior covariance functions are identical as there is no dependence on the observed values. In this approximation, the marginal belief about any function value is then a mixture of univariate Gaussians:

$$p(\phi \mid x, \mathcal{D}) \approx \frac{1}{s}\sum_{i=1}^{s} \mathcal{N}(\phi; \mu_i, \sigma^2); \quad \mu_i = \mu_{\mathcal{D}_i}(x); \quad \sigma^2 = K_{\mathcal{D}}(x, x). \tag{2.36}$$

Although slightly more complex than the Gaussian marginals of a Gaussian process, this is often convenient enough for most needs.

example: Student-t observation model

A Monte Carlo approximation to the posterior for the heavy-tailed dataset from Figure 2.10 is shown in Figure 2.12. The observations were modeled as corrupted by Student-t errors with $\nu = 4$ degrees of freedom. The posterior was approximated using a truly excessive number of samples (100 000, with a burn-in of 10 000) from the $\mathbf{y}$ posterior drawn using elliptical slice sampling.[47] The outliers in the data are ignored and the predictive performance is excellent.

46 *Handbook of Markov Chain Monte Carlo* (2011). Chapman & Hall.

47 I. MURRAY et al. (2010). Elliptical Slice Sampling. *AISTATS 2010*.

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

Gaussian approximate inference

An alternative to sampling is *approximate inference,* where we make a parametric approximation to the $\mathbf{y}$ posterior that yields a tractable posterior on f. In particular, if the posterior (2.33) were actually *normal,* it would induce a Gaussian process posterior on f. This insight is the basis for most approximation schemes.

In this vein, we proceed by first – somehow – approximating the true posterior over $\mathbf{y}$ with a multivariate Gaussian distribution:

$$p(\mathbf{y} \mid \mathcal{D}) \approx q(\mathbf{y} \mid \mathcal{D}) = \mathcal{N}(\mathbf{y}; \tilde{\mathbf{m}}, \tilde{\mathbf{C}}). \tag{2.37}$$

We are free to design this approximation as we see fit. There are several general-purpose approaches available, distinguished by how they approach maximizing the fidelity of fitting the true posterior (2.33). These include the Laplace approximation, Gaussian expectation propagation, and variational Bayesian inference. The first two of these methods are covered in Appendix B, and NICKISCH and RASMUSSEN provide an extensive survey of these and other approaches in the context of Gaussian process binary classification.[48]

Laplace approximation: § B.1, p. 301
Gaussian expectation propagation: § B.2 p. 302

48 H. NICKISCH and C. E. RASMUSSEN (2008). Approximations for Binary Gaussian Process Classification. *Journal of Machine Learning Research* 9(Oct):2035–2078.

Regardless of the details of the approximation scheme, the high-level result is the same – the normal approximation (2.37) in turn induces an approximate Gaussian process posterior on f. To demonstrate this, we consider the posterior on f that would arise from a direct observation of $\mathbf{y}$ (2.9–2.10) and integrate against the approximate posterior (2.37):

$$p(f \mid \mathcal{D}) \approx \int p(f \mid \mathbf{y})\, q(\mathbf{y} \mid \mathcal{D})\, \mathrm{d}\mathbf{y} = \mathcal{GP}(f; \mu_{\mathcal{D}}, K_{\mathcal{D}}), \tag{2.38}$$

where

$$\begin{aligned} \mu_{\mathcal{D}}(x) &= \mu(x) + \kappa(x)^{\top}\mathbf{C}^{-1}(\tilde{\mathbf{m}} - \mathbf{m}); \\ K_{\mathcal{D}}(x, x') &= K(x, x') - \kappa(x)^{\top}\mathbf{C}^{-1}(\mathbf{C} - \tilde{\mathbf{C}})\mathbf{C}^{-1}\kappa(x'). \end{aligned} \tag{2.39}$$

For most approximation schemes, the posterior covariance on f simplifies to a nicer, more familiar form. Most approximations to the $\mathbf{y}$ posterior (2.37) yield an approximate posterior covariance of the form

$$\tilde{\mathbf{C}} = \mathbf{C} - \mathbf{C}(\mathbf{C} + \mathbf{N})^{-1}\mathbf{C}, \tag{2.40}$$

where $\mathbf{N}$ is positive definite. Although this might appear mysterious, it is actually a natural form: it is the posterior covariance that would result from observing $\mathbf{y}$ corrupted by additive Gaussian noise with covariance $\mathbf{N}$ (2.19), except we are now free to design the noise covariance to maximize the fit. For approximations of this form (2.40), the approximate posterior covariance function on f simplifies to the more familiar

$$K_{\mathcal{D}}(x, x') = K(x, x') - \kappa(x)^{\top}(\mathbf{C} + \mathbf{N})^{-1}\kappa(x'). \tag{2.41}$$

example: conditioning on a local optimum

To demonstrate the power of approximate inference, we return to our motivating scenario of conditioning a one-dimensional process on having a local maximum at an identified point x, which we can achieve by

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

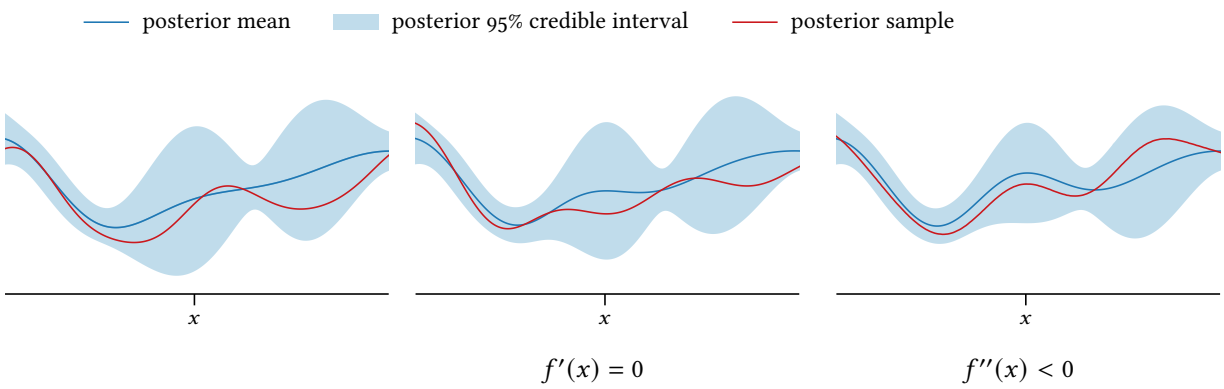


Figure 2.13: Approximately conditioning a Gaussian process to have a local maximum at the marked point x. We show each stage of the conditioning process with a sample drawn from the corresponding posterior. We begin with the unconstrained process (left), which we condition on the first derivative being zero at x using exact inference (middle). Finally we use Gaussian expectation propagation to approximately condition on the second derivative being negative at x.

derivative observations: § 2.6, p. 32

conditioning the first derivative to be zero and constraining the second derivative to be negative at x (2.32). We illustrate an approximation to the resulting posterior step-by-step in Figure 2.13, beginning with the example Gaussian process in the left-most panel. We first condition the process on the first derivative observation $f'(x) = 0$ using *exact* inference; the result is shown in the middle panel. Both the updated posterior mean and the sample reflect this information; however, the sample displays a local *minimum* at x, as the second-derivative constraint has not yet been addressed.

To incorporate the second-derivative constraint, we begin with this updated GP and consider the second derivative $h = f''(x)$, which is Gaussian distributed prior to the constraint (2.29):

$$p(h) = \mathcal{N}(h; m, s^2).$$

The negativity constraint induces a posterior on h incorporating the factor $[h < 0]$ (2.33); see Figure 2.14:

$$p(h \mid \mathcal{D}) \propto p(h)\,[h < 0].$$

The result is a truncated normal posterior on h. We may use Gaussian expectation propagation, which is especially convenient for handling bound constraints of this form, to produce a Gaussian approximation:

$$p(h \mid \mathcal{D}) \approx q(h \mid \mathcal{D}) = \mathcal{N}(h; \tilde{m}, \tilde{s}^2).$$

Incorporating the updated belief on h into the Gaussian process (2.39) yields the approximate posterior in the right-most panel of Figure 2.13. Although there is still some residual probability that the second derivative is positive at x in the approximate posterior (approximately 8%; see Figure 2.14), the belief reflects the desired information reasonably faithfully.

49 P. MCCULLAGH and J. A. NELDER (1989). *Generalized Linear Models*. Chapman & Hall.

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

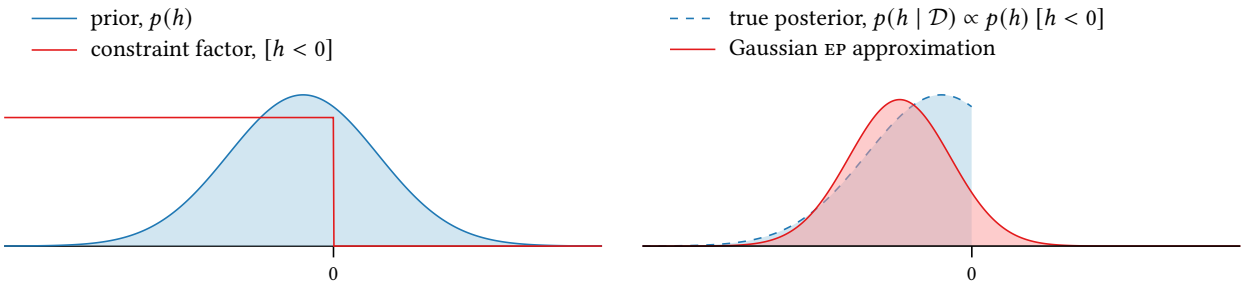


Figure 2.14: A demonstration of Gaussian expectation propagation. On the left we have a Gaussian belief on the second derivative, $p(h)$. We wish to constrain this value to be negative, introducing a step-function factor encoding the constraint, $[h < 0]$. The resulting distribution is non-Gaussian (right), but we can approximate it with a Gaussian, which induces an updated GP posterior on the function approximately incorporating the constraint.

Going beyond this example, we may use the approach outlined above to realize a general framework for Bayesian nonlinear regression by combining a GP prior on a latent function with an observation model appropriate for the task at hand, then approximating the posterior as desired. The convenience and modeling flexibility offered by Gaussian processes can easily justify any extra effort required for approximating the posterior. This can be seen as a nonlinear extension of the well-known family of *generalized linear models*.[49]

This approach is quite popular and has been realized countless times. Notable examples include binary classification using a logistic or probit observation model,[50] modeling point processes as a nonhomogeneous Poisson process with unknown intensity,[51,52] and robust regression with heavy-tailed additive noise such as Laplace[53] or Student-t[54,55] distributed errors. With regard to the latter and our previous heavy-tailed noise example, a Laplace approximation to the posterior for the data in Figures 2.10–2.12 with the Student-t observation model produces an approximate posterior in excellent agreement with the Monte Carlo approximation in Figure 2.12; see Figure 2.15. The cost of approximate inference in this case was dramatically (several orders of magnitude) cheaper than Monte Carlo sampling.

2.9 SUMMARY OF MAJOR IDEAS

Gaussian processes have been studied – in one form or another – for over 100 years.[56] Although we have covered a lot of ground in this chapter, we have only scratched the surface of an expansive body of literature. A good entry point to that literature is RASMUSSEN and WILLIAMS's monograph, which focuses on machine learning applications of Gaussian processes but also covers their theoretical underpinnings and properties in depth.[57] A good companion to this work is the book of ADLER and TAYLOR, which takes a deep dive into the properties and geometry of sample paths, including statistical properties of their maxima.[58]

50 H. NICKISCH and C. E. RASMUSSEN (2008). Approximations for Binary Gaussian Process Classification. *Journal of Machine Learning Research* 9(Oct):2035–2078.

51 J. MØLLER et al. (1998). Log Gaussian Cox Processes. *Scandinavian Journal of Statistics* 25(3): 451–482.

52 R. P. ADAMS et al. (2009). Tractable Nonparametric Bayesian Inference in Poisson Processes with Gaussian Process Intensities. *ICML 2009*.

53 M. KUSS (2006). Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning. Ph.D. thesis. Technische Universität Darmstadt.[§ 5.4]

54 R. M. NEAL (1997). *Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification*. Technical report (9702). Department of Statistics, University of Toronto.

55 P. JYLÄNKI et al. (2011). Robust Gaussian Process Regression with a Student-t Likelihood. *Journal of Machine Learning Research* 12(99): 3227–3257.

56 DIACONIS identified an early application of GPs by POINCARÉ for nonlinear regression:

P. DIACONIS (1988). Bayesian Numerical Analysis. In: *Statistical Decision Theory and Related Topics IV*.

H. POINCARÉ (1912). *Calcul des probabilités*. Gauthier–Villars.

57 C. E. RASMUSSEN and C. K. I. WILLIAMS (2006). *Gaussian Processes for Machine Learning*. MIT Press.

58 R. J. ADLER and J. E. TAYLOR (2007). *Random Fields and Geometry*. Springer–Verlag.

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

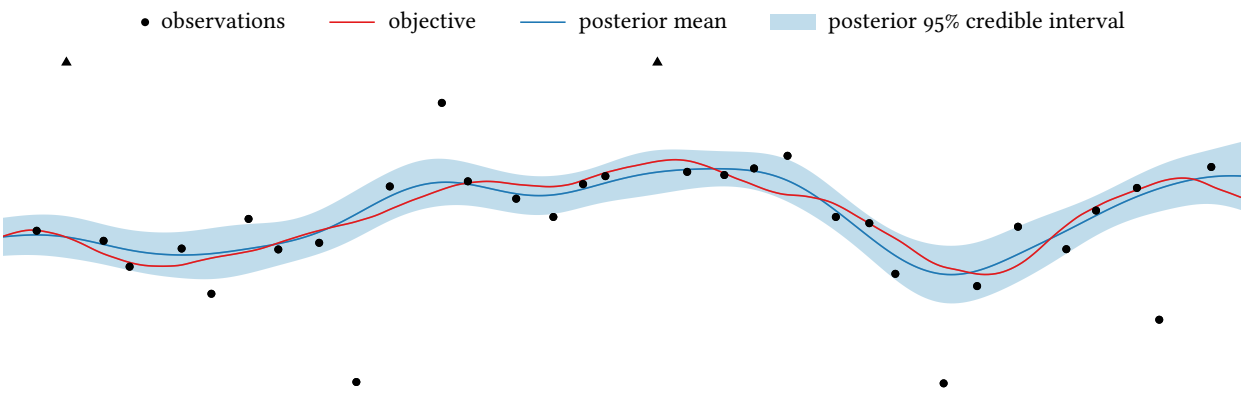


Figure 2.15: A Laplace approximation to the posterior from Figure 2.12.

Fortunately, the basic definitions and properties covered in § 2.1 and exact inference procedure covered in § 2.2 already provide a sufficient foundation for the majority of practical applications of Bayesian optimization. This material also provides sufficient background knowledge for the majority of the remainder of the book. However, we wish to underscore the major results from this chapter at a high level.

- Gaussian processes extend the multivariate normal distribution to model functions on infinite domains. As in the finite-dimensional case, Gaussian processes are specified by their first two moments – a mean function and a positive-definite covariance function – which endow any finite set of function values with a multivariate normal distribution (2.2–2.3).
- Conditioning a Gaussian process on function observations that are either exact or corrupted by additive Gaussian noise yields a Gaussian process posterior with updated moments reflecting the assumptions in the prior and the information in the observations (2.9–2.10).

inference from arbitrary joint Gaussian observations: § 2.2, p. 22

- In fact, we may condition a Gaussian process on the observation of *any* observations sharing a joint Gaussian distribution with the function of interest.

interpretation of posterior moments: § 2.2, p. 21

- In the case of exact inference, the posterior moments of a Gaussian process can be rewritten in terms of correlations among function values and z-scores of the observed values in a manner that may be more intuitive than the standard formulas.

joint Gaussian processes: § 2.4, p. 26

Extensions and Related Settings: Chapter 11, p. 245

- We may extend Gaussian processes to jointly model multiple correlated functions via careful bookkeeping, a construction known as a *joint Gaussian process*. Joint GPs are widely used in optimization settings involving multiple objectives and/or cheaper surrogates for an expensive objective.

continuity: § 2.5, p. 28

differentiability: § 2.6, p. 30

- Continuity and differentiability of Gaussian process sample paths can be guaranteed under mild assumptions on the mean and covariance functions. When these functions are sufficiently differentiable, a GP-distributed function shares a joint GP distribution with its gradient (2.28).

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press

This joint distribution allows us to condition a Gaussian process on (potentially noisy) derivative observations.

derivative observations: § 2.6, p. 32

- The existence and uniqueness of global maxima for Gaussian process sample paths can be guaranteed under mild assumptions on the mean and covariance functions. Establishing these properties ensures that the location x^* and value f^* of the global maximum are well-founded random variables, which will be critical for some optimization methods introduced later in the book.[59]

existence and uniqueness of global maxima: § 2.7, p. 33

59 In particular, policies grounded in information theory under the umbrella of "entropy search." See § 7.6, p. 135 for more.

- Inference from non-Gaussian observations and constraints is possible via Monte Carlo sampling or Gaussian approximate inference.

inference with non-Gaussian observations and constraints: § 2.8, p. 35

Looking forward, the focus of this chapter has been on theoretical rather than practical properties of Gaussian processes. A huge outstanding question is how to actually *design* a Gaussian process to model a given system. This will be our focus for the next two chapters. In the next chapter, we will explore model *construction,* and in the following chapter we will consider model *assessment* in light of available data.

Finally, we have not yet discussed any computational issues inherent to Gaussian process inference, including, most importantly, how the cost of computing the posterior grows with respect to the number of observations. We will discuss implementation details and scaling in a dedicated chapter later in the book.

implementation and scaling of Gaussian process inference: § 9.1, p. 201

https://doi.org/10.1017/9781108348973.003 Published online by Cambridge University Press