1 INTRODUCTION

Optimization is an innate human behavior. On an individual level, we all strive to better ourselves and our surroundings. On a collective level, societies struggle to allocate limited resources seeking to improve the welfare of their members, and optimization has been an engine of societal progress since the domestication of crops through selective breeding over 12 000 years ago – an effort that continues to this day.

Given its pervasiveness, it should perhaps not be surprising that optimization is also *difficult.* While searching for an optimal design, we must spend – sometimes quite significant – resources evaluating suboptimal alternatives along the way. This observation compels us to seek methods of optimization that, when necessary, can carefully allocate resources to identify optimal parameters as efficiently as possible. This is the goal of mathematical optimization.

Since the 1960s, the statistics and machine learning communities have refined a *Bayesian* approach to optimization that we will develop and explore in this book. Bayesian optimization routines rely on a statistical model of the objective function, whose beliefs guide the algorithm in making the most fruitful decisions. These models can be quite sophisticated, and maintaining them throughout optimization may entail significant cost of its own. However, the reward for this effort is unparalleled sample efficiency. For this reason, Bayesian optimization has found a niche in optimizing objectives that:

- are costly to compute, precluding exhaustive evaluation,
- lack a useful expression, causing them to function as "black boxes,"
- cannot be evaluated exactly, but only through some indirect or noisy mechanism, and/or
- offer no efficient mechanism for estimating their gradient.

Let us consider an example setting motivating the machine learning community's recent interest in Bayesian optimization. Consider a data scientist crafting a complex machine learning model – say a deep neural network – from training data. To ensure success, the scientist must carefully tune the model's hyperparameters, including the network architecture and details of the training procedure, which have massive influence on performance. Unfortunately, effective settings can only be identified via trial-and-error: by training several networks with different settings and evaluating their performance on a validation dataset.

The search for the best hyperparameters is of course an exercise in optimization. Mathematical optimization has been under continual development for centuries, and numerous off-the-shelf procedures are available. However, these procedures usually make assumptions about the objective function that may not always be valid. For example, we might assume that the objective is cheap to evaluate, that we can easily compute its gradient, or that it is convex, allowing us to reduce from global to local optimization.

https://doi.org/10.1017/9781108348973.002 Published online by Cambridge University Press

In hyperparameter tuning, all of these assumptions are invalid. Training a deep neural network can be extremely expensive in terms of both time and energy. When some hyperparameters are discrete – as many features of network architecture naturally are – the gradient does not even *exist.* Finally, the mapping from hyperparameters to performance may be highly complex and multimodal, so local refinement may not yield an acceptable result.

The Bayesian approach to optimization allows us to relax all of these assumptions when necessary, and Bayesian optimization algorithms can deliver impressive performance even when optimizing complex "black box" objectives under severely limited observation budgets. Bayesian optimization has proven successful in settings spanning science, engineering, and beyond, including of course hyperparameter tuning.[1] In light of this broad success, GELMAN and VEHTARI identified adaptive decision analysis – and Bayesian optimization in particular – as one of the eight most important statistical ideas of the past 50 years.[2]

Annotated Bibliography of Applications: Appendix D, p. 313

Covering all these applications and their nuances could easily fill a separate volume (although we do provide an overview of some important application domains in an annotated bibliography), so in this book we will settle for developing the mathematical foundation of Bayesian optimization underlying its success. In the remainder of this chapter we will lay important groundwork for this discussion. We will first establish the precise formulation of optimization we will consider and important conventions of our presentation, then outline and illustrate the key aspects of the Bayesian approach. The reader may find an outline of and reading guide for the chapters to come in the Preface.

outline and reading guide: p. x

1.1 FORMALIZATION OF OPTIMIZATION

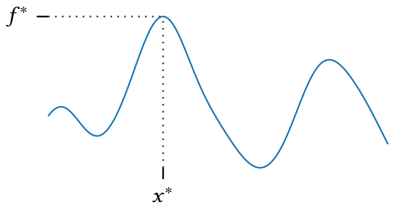
Throughout this book we will consider a simple but flexible formulation of sequential global optimization outlined below. There is nothing inherently Bayesian about this model, and countless solutions are possible.

objective function, f

domain of objective function, $\mathcal{X}$

We begin with a real-valued objective function defined on some domain $\mathcal{X}$; $f\colon \mathcal{X} \to \mathbb{R}$. We make no assumptions regarding the nature of the domain. In particular, it need not be Euclidean but might instead, for example, comprise a space of complex structured objects. The goal of optimization is to systematically search the domain for a point x^* attaining the globally maximal value f^*:[3]

$$x^* \in \operatorname*{arg\,max}_{x \in \mathcal{X}} f(x); \qquad f^* = \max_{x \in \mathcal{X}} f(x) = f(x^*). \tag{1.1}$$



An objective function with the location, x^*, and value, f^*, of the global optimum marked.

Before we proceed, we note that our focus on maximization rather than minimization is entirely arbitrary; the author simply judges maximization to be the more optimistic choice. If desired, we can freely transform one problem to the other by negating the objective function. We caution the reader that some translation may be required when comparing expressions derived here to what may appear in parallel texts focusing on minimization.

[1] R. TURNER et al. (2021). Bayesian Optimization Is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020. *Proceedings of the NeurIPS 2020 Competition and Demonstration Track.*

[2] A. GELMAN and A. VEHTARI (2021). What Are the Most Important Statistical Ideas of the Past 50 Years? *Journal of the American Statistical Association* 116(536):2087–2097.

[3] A skeptical reader may object that, without further assumptions, a global maximum may not exist at all! We will sidestep this issue for now and pick it up again in § 2.7, p. 34.

https://doi.org/10.1017/9781108348973.002 Published online by Cambridge University Press

```
input: initial dataset $\mathcal{D}$                    ▸ can be empty
repeat
    $x \leftarrow \textsc{policy}(\mathcal{D})$          ▸ select the next observation location
    $y \leftarrow \textsc{observe}(x)$                  ▸ observe at the chosen location
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x, y)\}$  ▸ update dataset
until termination condition reached     ▸ e.g., budget exhausted
return $\mathcal{D}$
```

Algorithm 1.1: Sequential optimization.

In a significant departure from classical mathematical optimization, we do not require that the objective function have a known functional form or even be computable directly. Rather, we only require access to a mechanism revealing *some* information about the objective function at identified points on demand. By amassing sufficient information from this mechanism, we may hope to infer the solution to (1.1). Avoiding the need for an explicit expression for f allows us to consider so-called "black box" optimization, where a system is optimized through indirect measurements of its quality. This is one of the greatest strengths of Bayesian optimization.[4]

Optimization policy

Directly solving for the location of global optima is infeasible except in exceptional circumstances. The tools of traditional calculus are virtually powerless in this setting; for example, enumerating and classifying every stationary point in the domain would be tedious at best and perhaps even impossible. Mathematical optimization instead takes an indirect approach: we design a sequence of experiments to probe the objective function for information that, we hope, will reveal the solution to (1.1).

The iterative procedure in Algorithm 1.1 formalizes this process. We begin with an initial (possibly empty) dataset $\mathcal{D}$ that we grow incrementally through a sequence of observations of our design. In each iteration, an *optimization policy* inspects the available data and selects a point $x \in \mathcal{X}$ where we make our next observation.[5] This action in turn reveals a corresponding value y provided by the system under study. We append the newly observed information to our dataset and finally decide whether to continue with another observation or terminate and return the current data. When we inevitably do choose to terminate, the returned data can be used by an external consumer as desired, for example to inform a subsequent decision.

terminal recommendations: § 5.1, p. 90

We place no restrictions on how an optimization policy is implemented beyond mapping an arbitrary dataset to some point in the domain for evaluation. A policy may be deterministic or stochastic, as demonstrated respectively by the prototypical examples of grid search and random search. In fact, these popular policies are *nonadaptive* and completely ignore the observed data. However, when observations only come at significant cost, we will naturally prefer policies that adapt their behavior in light of evolving information. The primary challenge in opti-

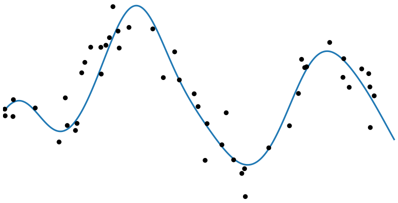
[4] Of course, we do not *require* but merely *allow* that the objective function act as a black box. Access to a closed-form expression does not preclude a Bayesian approach!

[5] Here "policy" has the same meaning as in other decision-making contexts: it maps our state (indexed by our data, $\mathcal{D}$) to an action (the location of our next observation, x).

https://doi.org/10.1017/9781108348973.002 Published online by Cambridge University Press

mization is designing policies that can *rapidly* optimize a broad class of objective functions, and intelligent policy design will be our focus for the majority of this book.

Observation model



Inexact observations of an objective function corrupted by additive noise.

For optimization to be feasible, the observations we obtain must provide information about the objective function that can guide our search and in aggregate determine the solution to (1.1). A near-universal assumption in mathematical optimization is that observations yield *exact* evaluations of the objective function at our chosen locations. However, this assumption is unduly restrictive: many settings feature inexact measurements due to noisy sensors, imperfect simulation, or statistical approximation. A typical example featuring additive observation noise is shown in the margin. Although the objective function is not observed directly, the noisy measurements nonetheless constrain the plausible options due to strong dependence on the objective.

measured value, y

observation location, x

objective function value, $\phi = f(x)$

We thus relax the assumption of exact observation and instead assume that observations are realized by a stochastic mechanism depending on the objective function. Namely, we assume that the value y resulting from an observation at some point x is distributed according to an observation model depending on the underlying objective function value $\phi = f(x)$:

$$p(y \mid x, \phi). \tag{1.2}$$

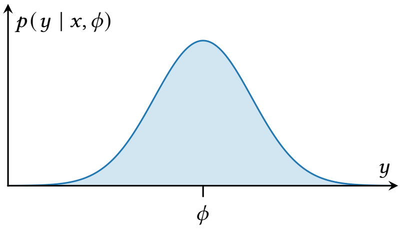
Through judicious design of the observation model, we may consider a wide range of observation mechanisms.

conditional independence of observations given objective values

As with the optimization policy, we do not make any assumptions about the nature of the observation model, save one. Unless otherwise mentioned, we assume that a set of *multiple* measurements $\mathbf{y}$ are conditionally independent given the corresponding observation locations $\mathbf{x}$ and objective function values $\boldsymbol{\phi} = f(\mathbf{x})$:

$$p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\phi}) = \prod_i p(y_i \mid x_i, \phi_i). \tag{1.3}$$

This is not strictly necessary but is overwhelmingly common in practice and will simplify our presentation considerably.



Additive Gaussian noise: the distribution of the value y observed at x is Gaussian, centered on the objective function value ϕ.

One particular observation model will enjoy most of our attention in this book: *additive Gaussian noise*. Here we model the value y observed at x as

$$y = \phi + \varepsilon,$$

where ε represents measurement error. Errors are assumed to be Gaussian distributed with mean zero, implying a Gaussian observation model:

$$p(y \mid x, \phi, \sigma_n) = \mathcal{N}(y; \phi, \sigma_n^2). \tag{1.4}$$

observation noise scale, σ_n

heteroskedastic noise: § 2.2, p. 25

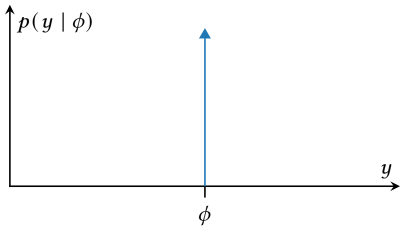
Here the observation noise scale σ_n may optionally depend on x, allowing us to model both homoskedastic or heteroskedastic errors.

https://doi.org/10.1017/9781108348973.002 Published online by Cambridge University Press

If we take the noise scale to be identically zero, we recover the special case of exact observation, where we simply have $y = \phi$ and the observation model collapses to a Dirac delta distribution:

$$p(y \mid \phi) = \delta(y - \phi).$$

Although not universally applicable, many settings do feature exact observations such as optimizing the output of a deterministic computer simulation. We will sometimes consider the exact case separately as some results simplify considerably in the absence of measurement error.



Exact observations: every value measured equals the corresponding function value, yielding a Dirac delta observation model.

We will focus on additive Gaussian noise as it is a reasonably faithful model for many systems and offers considerable mathematical convenience. This observation model will be most prevalent in our discussion on Gaussian processes in the next three chapters and on the explicit computation of Bayesian optimization policies with this model class in Chapter 8. However, the general methodology we will build in the remainder of this book is not contingent on this choice, and we will occasionally address alternative observation mechanisms.

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

optimization with non-Gaussian observations: § 11.11, p. 282

Termination

The final decision we make in each iteration of optimization is whether to terminate immediately or continue with another observation. As with the optimization policy, we do not assume any particular mechanism by which this decision is made. Termination may be deterministic – such as stopping after reaching a certain optimization goal or exhausting a preallocated observation budget – or stochastic, and may optionally depend on the observed data. In many cases, the time of termination may in fact not be under the control of the optimization routine at all but instead decided by an external agent. However, we will also consider scenarios where the optimization procedure can dynamically choose when to return based upon inspection of the available data.

optimal termination: § 5.4, p. 103

practical termination: § 9.3, p. 210

1.2 THE BAYESIAN APPROACH

Bayesian optimization does not refer to one particular algorithm but rather to a philosophical approach to optimization grounded in Bayesian inference from which an extensive family of algorithms have been derived. Although these algorithms display significant diversity in their details, they are bound by common themes in their design.

Optimization is fundamentally a sequence of decisions: in each iteration, we must choose where to make our next observation and then whether to terminate depending on the outcome. As the outcomes of these decisions are governed by the system under study and outside our control, the success of optimization rests entirely on effective decision making.

Increasing the difficulty of these decisions is that they must be made under *uncertainty,* as it is impossible to know the outcome of an observation before making it. The optimization policy must therefore design each

https://doi.org/10.1017/9781108348973.002 Published online by Cambridge University Press

observation with some measure of faith that the outcome will ultimately prove beneficial and justify the cost of obtaining it. The sequential nature of optimization further compounds the weight of this uncertainty, as the outcome of each observation not only has an immediate impact, but also forms the basis on which all future decisions are made. Developing an effective policy requires somehow addressing this uncertainty.

The Bayesian approach systematically relies on probability and Bayesian inference to reason about the uncertain quantities arising during optimization. This critically includes the objective function itself, which is treated as a random variable to be inferred in light of our prior expectations and any available data. In Bayesian optimization, this belief then takes an active role in decision making by guiding the optimization policy, which may evaluate the merit of a proposed observation location according to our belief about the value we might observe. We introduce the key ideas of this process with examples below, starting with a refresher on Bayesian inference.

Bayesian inference

To frame the following discussion, we offer a quick overview of Bayesian inference as a reminder to the reader. This introduction is far from complete, but there are numerous excellent references available.[6]

Bayesian inference is a framework for inferring uncertain features of a system of interest from observations grounded in the laws of probability. To illustrate the basic ideas, we may begin by identifying some unknown feature of a given system that we wish to reason about. In the context of optimization, this might represent, for example, the value of the objective function at a given location, or the location x^* or value f^* of the global optimum (1.1). We will take the first of these as a running example: inferring about the value of an objective function at some arbitrary point x, $\phi = f(x)$. We will shortly extend this example to inference about the *entire* objective function.

In the Bayesian approach to inference, *all* unknown quantities are treated as random variables. This is a powerful convention as it allows us to represent beliefs about these quantities with probability distributions reflecting their plausible values. Inference then takes the form of an inductive process where these beliefs are iteratively refined in light of observed data by appealing to probabilistic identities.

prior distribution, $p(\phi \mid x)$

As with any induction, we must start somewhere. Here we begin with a so-called *prior distribution* (or simply *prior*) $p(\phi \mid x)$, which encodes what we consider to be plausible values for ϕ before observing any data.[7] The prior distribution allows us to inject our knowledge about and experience with the system of interest into the inferential process, saving us from having to begin "from scratch" or entertain patently absurd possibilities. The left panel of Figure 1.1 illustrates a prior distribution for our example, indicating support over a range of values.

Once a prior has been established, the next stage of inference is to refine our initial beliefs in light of observed data. Suppose in our

[6] The literature is vast. The following references are excellent, but no list can be complete:

D. J. C. MACKAY (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

A. O'HAGAN and J. FORSTER (2004). *Kendall's Advanced Theory of Statistics*. Vol. 2B: Bayesian Inference. Arnold.

J. O. BERGER (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer–Verlag.

[7] Here we assume the location of interest x is known, hence our conditioning the prior on its value.

https://doi.org/10.1017/9781108348973.002 Published online by Cambridge University Press

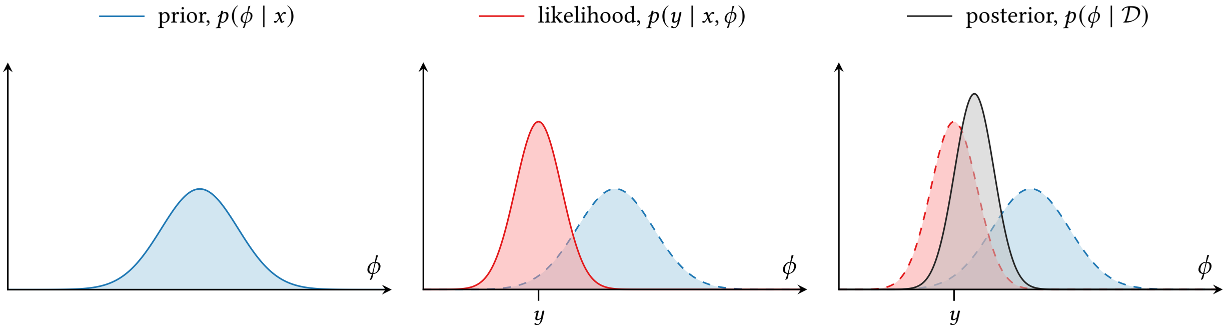


Figure 1.1: Bayesian inference for an unknown function value $\phi = f(x)$. Left: a prior distribution over ϕ; middle: the likelihood of the marked observation y according to an additive Gaussian noise observation model (1.4) (prior shown for reference); right: the posterior distribution in light of the observation and the prior (prior and likelihood shown for reference).

example we make an observation of the objective function at x, revealing a measurement y. In our model of optimization, the distribution of this measurement is assumed to be determined by the value of interest ϕ through the observation model $p(y \mid x, \phi)$ (1.2). In the context of Bayesian inference, a distribution explaining the observed values (here y) in terms of the values of interest (here ϕ) is known as a *likelihood function* or simply a *likelihood.* The middle panel of Figure 1.1 shows the likelihood – as a function of ϕ – for a given measurement y, here assumed to be generated by additive Gaussian noise (1.4).

likelihood function (observation model), $p(y \mid x, \phi)$

Finally, given the observed value y, we may derive the updated *posterior distribution* (or simply *posterior*) of ϕ by appealing to Bayes' theorem:

posterior distribution, $p(\phi \mid x, y)$

$$p(\phi \mid x, y) = \frac{p(\phi \mid x)\, p(y \mid x, \phi)}{p(y \mid x)}. \tag{1.5}$$

The posterior is proportional to the prior weighted by the likelihood of the observed value. The denominator is a constant with respect to ϕ that ensures normalization:

$$p(y \mid x) = \int p(y \mid x, \phi)\, p(\phi \mid x)\, \mathrm{d}\phi. \tag{1.6}$$

The right panel of Figure 1.1 shows the posterior resulting from the measurement in the middle panel. The posterior represents a compromise between our experience (encoded in the prior) and the information contained in the data (encoded in the likelihood).

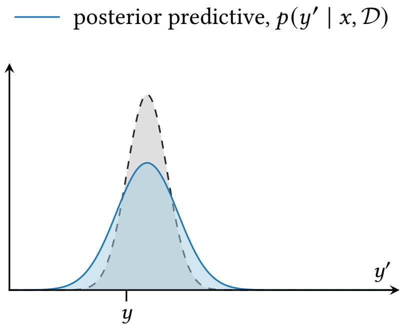
Throughout this book we will use the catchall notation $\mathcal{D}$ to represent all the information influencing a posterior belief; here the relevant information is $\mathcal{D} = (x, y)$, and the posterior distribution is then $p(\phi \mid \mathcal{D})$.

data informing posterior belief, $\mathcal{D}$

As mentioned previously, Bayesian inference is an inductive process whereby we can continue to refine our beliefs through additional observation. At this point, the induction is trivial: to incorporate a new

https://doi.org/10.1017/9781108348973.002 Published online by Cambridge University Press

observation, what was our posterior serves as the prior in the context of the new information, and multiplying by the likelihood and renormalizing yields a new posterior. We may continue in this manner as desired.



Posterior predictive distribution for a repeated measurement at x for our running example. The location of our first measurement y and the posterior distribution of ϕ are shown for reference. There is more uncertainty in y' than ϕ due to the effect of observation noise.

The posterior distribution is not usually the end result of Bayesian inference but rather a springboard enabling follow-on tasks such as prediction or decision making, both of which are integral to Bayesian optimization. To address the former, suppose that after deriving the posterior (1.5), we wish to predict the result of an independent, *repeated* noisy observation at x, y'. Treating the outcome as a random variable, we may derive its distribution by integrating our posterior belief about ϕ against the observation model (1.2):[8]

$$p(y' \mid x, \mathcal{D}) = \int p(y' \mid x, \phi)\, p(\phi \mid x, \mathcal{D})\, \mathrm{d}\phi; \qquad (1.7)$$

this is known as the *posterior predictive distribution* for y'. By integrating over all possible values of ϕ weighted by their plausibility, the posterior predictive distribution naturally accounts for uncertainty in the unknown objective function value; see the figure in the margin.

The Bayesian approach to decision making also relies on a posterior belief about unknown features affecting the outcomes of our decisions, as we will discuss shortly.

Bayesian inference of the objective function

stochastic process

At the heart of any Bayesian optimization routine is a probabilistic belief over the objective function. This takes the form of a *stochastic process*, a probability distribution over an infinite collection of random variables – here the objective function value at every point. The reasoning behind this inference is, in essence, the same as our single-point example above.

objective function prior, $p(f)$

We begin by encoding any assumptions we may have about the objective function, such as smoothness or other features, in a *prior process* $p(f)$. Conveniently, we can specify a stochastic process via the distribution of the function values $\boldsymbol{\phi}$ corresponding to an arbitrary *finite* set of locations $\mathbf{x}$:

$$p(\boldsymbol{\phi} \mid \mathbf{x}). \qquad (1.8)$$

Chapter 2: Gaussian Processes, p. 15

Chapter 3: Modeling with Gaussian Processes, p. 45

Chapter 4: Model Assessment, Selection, and Averaging, p. 67

The family of *Gaussian processes* – where these finite-dimensional distributions are multivariate Gaussian – is especially convenient and widely used in Bayesian optimization. We will explore this model class in depth in the following three chapters; here we provide a motivating illustration.

Figure 1.2 shows a Gaussian process prior on a one-dimensional objective function, constructed to reflect a minimal set of assumptions we will elaborate on later in the book:

differentiability: § 2.6, p. 30

characteristic length scales: § 3.4, p. 56

stationarity: § 3.2, p. 50

- that the objective function is smooth (that is, infinitely differentiable),
- that correlations among function values have a characteristic scale, and
- that the function's expected behavior does not depend on location (that is, the prior process is *stationary*).

[8] This expression takes the same form as (1.6), which is simply the (prior) predictive distribution evaluated at the actual observed value.

https://doi.org/10.1017/9781108348973.002 Published online by Cambridge University Press

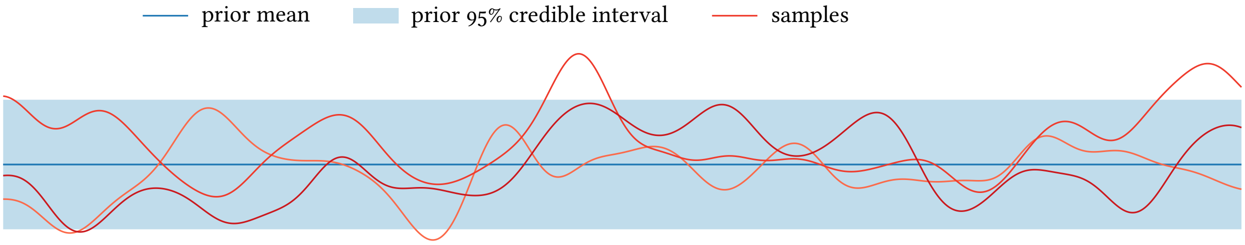


Figure 1.2: An example prior process for an objective defined on an interval. We illustrate the marginal belief at every location with its mean and a 95% credible interval and also show three example functions sampled from the prior process.

We summarize the marginal belief of the model, for each point in the domain showing the prior mean and a 95% credible interval for the corresponding function value. We also show three functions sampled from the prior process, each exhibiting the assumed behavior. We encourage the reader to become comfortable with this plotting convention, as we will use it throughout this book. In particular we eschew axis labels, as they are always the same: the horizontal axis represents the domain $\mathcal{X}$ and the vertical axis the function value. Further, we do not mark units on axes to stress relative rather than absolute behavior, as scale is arbitrary in this illustration.

plotting conventions

We can encode a vast array of information into the prior process and can model significantly more complex structure than in this simple example. We will explore the world of possibilities in Chapter 3, including interaction at different scales, nonstationarity, low intrinsic dimensionality, and more.

nonstationarity, warping: § 3.4, p. 56

low intrinsic dimensionality: § 3.5, p. 61

With the prior process in hand, suppose we now make a set of observations at some locations $\mathbf{x}$, revealing corresponding values $\mathbf{y}$; we aggregate this information into a dataset $\mathcal{D} = (\mathbf{x}, \mathbf{y})$. Bayesian inference accounts for these observations by forming the *posterior process* $p(f \mid \mathcal{D})$.

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

objective function posterior, $p(f \mid \mathcal{D})$

The derivation of the posterior process can be understood as a two-stage process. First we consider the impact of the data on the corresponding function values $\boldsymbol{\phi}$ alone (1.5):

$$p(\boldsymbol{\phi} \mid \mathcal{D}) \propto p(\boldsymbol{\phi} \mid \mathbf{x})\, p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\phi}). \tag{1.9}$$

The quantities on the right-hand side are known: the first term is given by the prior process (1.8), and the second by the observation model (1.3), which serves the role of a likelihood. We now extend the posterior on $\boldsymbol{\phi}$ to all of f:[9]

$$p(f \mid \mathcal{D}) = \int p(f \mid \mathbf{x}, \boldsymbol{\phi})\, p(\boldsymbol{\phi} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\phi}. \tag{1.10}$$

The posterior encapsulates our belief regarding the objective in light of the data, incorporating both the assumptions of the prior process and the information contained in the observations.

We illustrate an example posterior in Figure 1.3, where we have conditioned our prior from Figure 1.2 on three exact observations. As the

[9] The given expression sweeps some details under the rug. A careful derivation of the posterior process proceeds by finding the posterior of an arbitrary *finite*-dimensional vector $\boldsymbol{\phi}_* = f(\mathbf{x}_*)$:

$$p(\boldsymbol{\phi}_* \mid \mathbf{x}_*, \mathcal{D}) = \int p(\boldsymbol{\phi}_* \mid \mathbf{x}_*, \mathbf{x}, \boldsymbol{\phi})\, p(\boldsymbol{\phi} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\phi},$$

which specifies the process. The distributions on the right-hand side are known: the posterior on $\boldsymbol{\phi}$ is in (1.9), and the posterior on $\boldsymbol{\phi}_*$ given the *exact* function values $\boldsymbol{\phi}$ can be found by computing their joint prior (1.8) and conditioning.

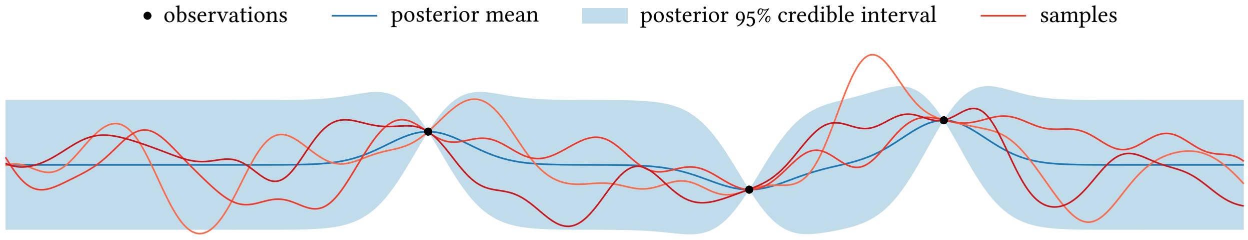


Figure 1.3: The posterior process for our example scenario in Figure 2.1 conditioned on three exact observations.

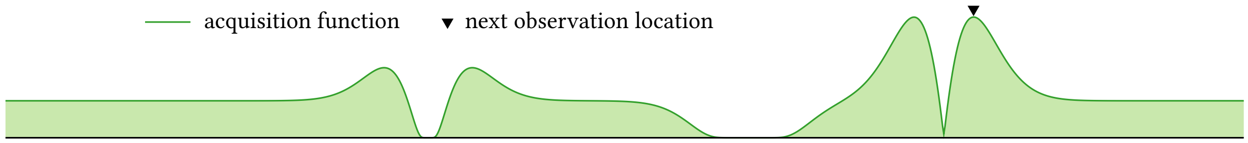


Figure 1.4: A prototypical acquisition function corresponding to our example posterior from Figure 1.3.

observations are assumed to be exact, the objective function posterior collapses onto the observed values. The posterior mean interpolates through the data, and the posterior credible intervals reflect increased certainty regarding the function near the observed locations. Further, the posterior continues to reflect the structural assumptions encoded in the prior, demonstrated by comparing the behavior of the samples drawn from the posterior process to those drawn from the prior.

Uncertainty-aware optimization policies

Bayesian inference provides an elegant means of reasoning about an uncertain objective function, but the success of optimization is measured not by the fidelity of our beliefs but by the outcomes of our actions. These actions are determined by the optimization policy, which examines available data to design each successive observation location. Each of these decisions is fraught with uncertainty, as we must commit to each observation before knowing its result, which will form the context of all following decisions. Bayesian inference enables us to express this uncertainty, but effective decision making additionally requires us to establish preferences over outcomes and act to maximize those preferences.

To proceed we need to establish a framework for decision making under uncertainty, an expansive subject with a world of possibilities. A natural and common choice is *Bayesian decision theory*, the subject of Chapters 5–6. We will discuss this and other approaches to policy construction at length in Chapter 7 and derive popular optimization policies from first principles.

Chapter 5: Decision Theory for Optimization, p. 87

Chapter 6: Utility Functions for Optimization, p. 109

Chapter 7: Common Bayesian Optimization Policies, p. 123

Ignoring details in policy design, a thread running through all Bayesian optimization policies is a uniform handling of uncertainty in the objective function and the outcomes of observations via Bayesian inference. Instrumental in connecting our beliefs about the objective function to decision making is the posterior predictive distribution (1.7), representing our belief about the outcomes of proposed observations. Bayesian optimization policies are designed with reference to this distribution, which guides the policy in discriminating between potential actions.

acquisition functions: § 5, p. 88

In practice, Bayesian optimization policies are defined indirectly by optimizing a so-called *acquisition function* assigning a score to potential observation locations commensurate with their perceived ability to benefit the optimization process. Acquisition functions tend to be cheap to evaluate with analytically tractable gradients, allowing the use of off-the-shelf optimizers to efficiently design each observation. Numerous acquisition functions have been proposed for Bayesian optimization, each derived from different considerations. However, all notable acquisition functions address the classic tension between *exploitation* – sampling where the objective function is expected to be high – and *exploration* – sampling where we are uncertain about the objective function to inform future decisions. These opposing concerns must be carefully balanced for effective global optimization.

example and discussion

An example acquisition function is shown in Figure 1.4, corresponding to the posterior from Figure 1.3. Consideration of the exploitation–exploration tradeoff is apparent: this example acquisition function attains relatively large values both near local maxima of the posterior mean and in regions with significant marginal uncertainty. Local maxima of the acquisition function represent optimal compromises between these concerns. Note that the acquisition function vanishes at the location of the current observations: the objective function values at these locations are already known, so observing there would be pointless. Maximizing the acquisition function determines the policy; here the policy chooses to search around the local optimum on the right-hand side.

Figure 1.5 demonstrates an entire session of Bayesian optimization, beginning from the belief and initial decision from Figure 1.4 and progressing iteratively following Algorithm 1.1. The true (unknown) objective function is also shown for reference; its maximum is near the center of the domain. The running marks below each posterior show the locations of each measurement made, progressing in sequence from top to bottom, and we show the objective function posterior at four waypoints.

Dynamic consideration of the exploitation–exploration tradeoff is evident in the algorithm's behavior. The first two observations map out the neighborhood of the initially best-seen point, exhibiting exploitation. Once sufficiently explored, the policy continues exploitation around the second best-seen point, discovering and refining the global optimum in iterations 7–8. Finally, the policy switches to exploration in iterations 13–19, systematically covering the domain to ensure nothing has been missed. At termination, there is clear bias in the collected data toward higher objective values, and all remaining uncertainty is in regions where the credible intervals indicate the optimum is unlikely to reside.

The "magic" of Bayesian optimization is that the intuitive behavior of this optimization policy is not the result of ad hoc design, but rather

https://doi.org/10.1017/9781108348973.002 Published online by Cambridge University Press

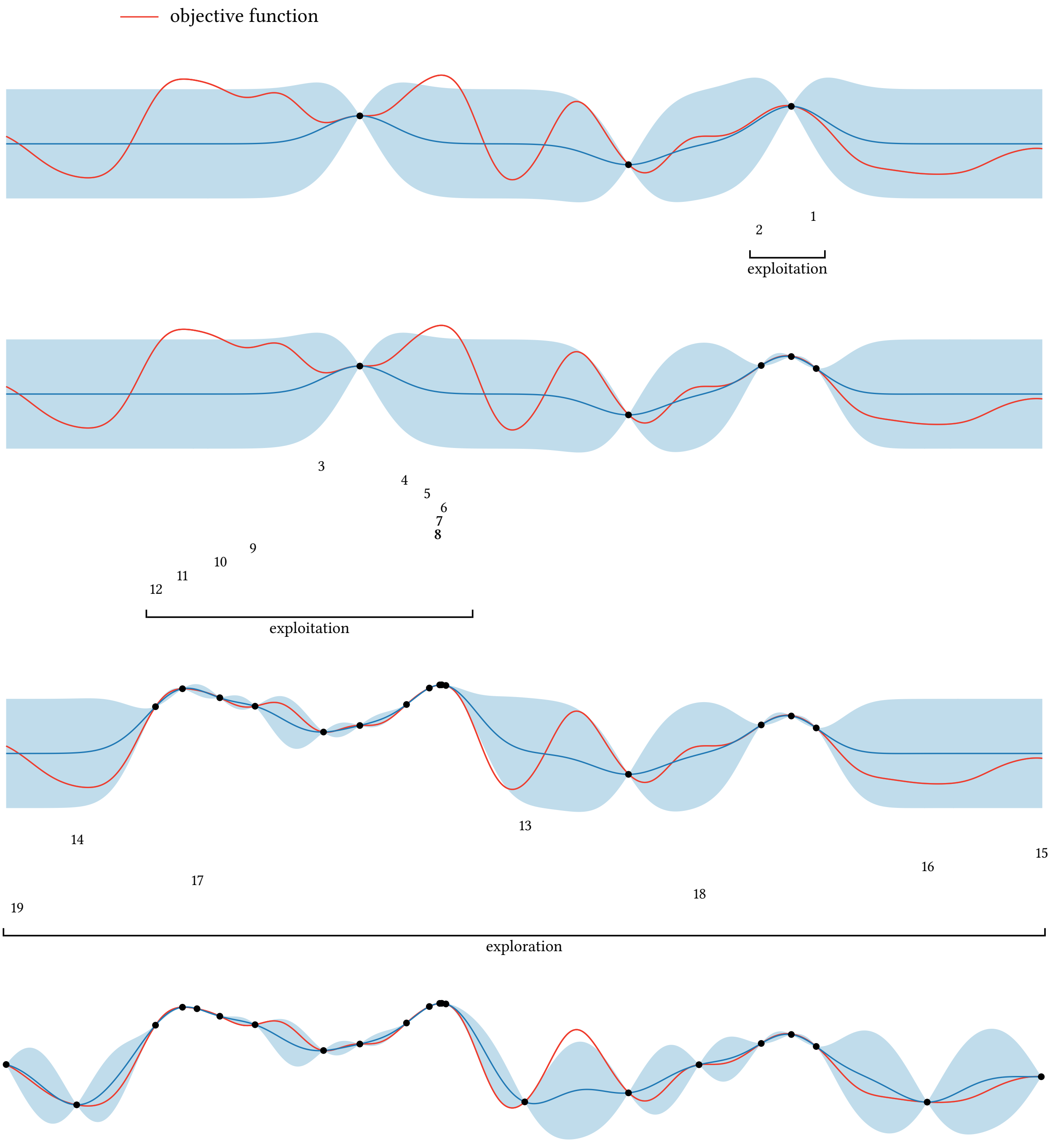


Figure 1.5: The posterior after the indicated number of steps of an example Bayesian optimization policy, starting from the posterior in Figure 1.4. The marks show the points chosen by the policy, progressing from top to bottom. Observations sufficiently close to the optimum are marked in bold; the optimum was located on iteration 7.

https://doi.org/10.1017/9781108348973.002 Published online by Cambridge University Press

emerges *automatically* through the machinery of Gaussian processes and Bayesian decision theory that we will develop over the coming chapters. In this framework, building an optimization policy boils down to:

- choosing a model of the objective function,
- deciding what sort of data we seek to obtain, and
- systematically transforming these beliefs and preferences into an optimization policy.

Over the following chapters, we will develop tools for achieving each of these goals: Gaussian processes (Chapters 2–4) for expressing what we believe about the objective function, utility functions (Chapter 6) for expressing what we value in data, and Bayesian decision theory (Chapter 5) for building optimization policies aware of the uncertainty encoded in the model and guided by the preferences encoded in the utility function. In Chapter 7 we will combine these fundamental components to realize complete Bayesian optimization policies, at which point we will be equipped to replicate this example from first principles.

https://doi.org/10.1017/9781108348973.002 Published online by Cambridge University Press

https://doi.org/10.1017/9781108348973.002 Published online by Cambridge University Press