BASS as a Surrogate Model for Bayesian Optimization

Adrian TJ

September 2022

Contents

Chapter 1

Stochastic Processes

Before starting, we would like to note that sections of this chapter where constructed using different references as primary guides. The section on Markov chains primarily uses the references [Ži10] and [HR05].

1.1 General Definitions

Stochastic processes play a very important role in basically every part of this work. First, the basis of traditional Bayesian Optimization is Gaussian processes, which are a specific class of stochastic process in which all the points or vectors have a conjoined multi-normal distribution. This is not the only place where stochastic processes play a role, as BASS is a Bayesian method, and modern Bayesian methods generally rely on simulation strategies such as Markov Chain Monte Carlo (MCMC) to numerically approximate complex posterior distributions. We use a particular class of MCMC algorithms called reverse jump MCMC when dealing with our alternate surrogate function, but that is detailed much later when we talk about the Bayesian Adaptive Regression Spline method.

All in all, stochastic processes are generally important to the content of this work so we include some important results and definitions, starting with the definition of a stochastic process itself, taken from [Ži10, §3.0].

Definition 1. *Let $\Theta \subseteq \mathbb{R}^+$ A stochastic process $\{X_\theta, \theta \in \Theta\}$ is a collection of random variables, indexed by a parameter θ, such that θ belongs to some index set Θ. When $\Theta = \mathbb{N}$ (or $\Theta = \mathbb{N}_0$), then it is called a discrete-time process, and if $\Theta = \mathbb{R}^+$, then it is called a continuous-time process.*

To distinguish between a discrete and continuous time process, we use the index n for discrete-time processes and t for continuous-time processes. In

all applications of this work the indexes refer to time, as the function spaces $\mathcal{X}$ on which the functions are evaluated are time dependent. This definition is very broad, as for example, if we take the case where $\Theta = \{1\}$, then the stochastic process $\{X_\theta, \theta \in \Theta\}$ is really just X_1, a random variable. As with many other areas of mathematics though, the introduction of infinities in the values the indexes can take introduces a large amount of complexity, and things like vector distributions do no extend nicely to these new cases.

Another more formal definition of a stochastic process defines it as a function of the index and an element of the sample space, as follows:

Definition 2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space*[1] *where Ω is a sample space, $\mathcal{F}$ is a σ-algebra and $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ a probability measure. Taking Θ an arbitrary index space, a stochastic process is a function*

$$X : \Omega \times \Theta \mapsto \mathbb{R}$$

such that for all $\theta \in \Theta$,

$$X_\theta : \omega \mapsto X(\omega, \theta) \equiv X_\theta : \Omega \mapsto \mathbb{R}$$

is a measurable function, which is equivalent to saying that X_θ is a random variable.

When ω is known, $\theta \mapsto X(\omega, \theta) : \Theta \mapsto \mathbb{R}$ is a sample function, also known as a realization of the stochastic process. These realizations are the trajectories or paths that are famous in diagrams such as the one shown in Figure 1.3. An example application of this type of structure are time series, where the fundamental properties of stochastic processes are used to generate a function whose realizations move through time.

As was mentioned earlier, the two main types of stochastic process that can be defined are ones where the realizations of the process occur in a

[1] Billingsley's classic book [Bil12] has all the theoretical foundation of probability spaces and provides a thorough run down of these concepts.

discrete fashion, and the other is when they occur in a continuous fashion. It so happens that the two main structures relating to stochastic processes that we need to go into detail on are Markov chains, a discrete process, and Gaussian processes, which are continuous. We go in depth in the following sections into each of these types of stochastic processes, starting with Markov chains since Gaussian processes lead nicely onto the next chapter.

1.2 Markov Chains

Markov chains are ubiquitous in their applications, since they represent a relatively simple but very powerful concept; the mathematical structure on which we can define phase transitions of position changes that are governed by probabilistic rules. The fundamental property which will be explored later is that independently of the past, the only important factor in determining the future state of the process is the current state of it. We begin with a formal definition and take a constructive approach in building this model.

Definition 3. *Let* $\{X_n, n \in \mathbb{N}_0\}$ *be a stochastic process defined on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *such that*

$$X_n : \omega \mapsto X(\omega, n) \equiv X_n : \Omega \mapsto \mathcal{S},$$

where $\mathcal{S}$ *is a finite state space of cardinality* k*, such that* $\mathcal{S} = \{s_1, s_2, \ldots, s_k\}$*. We call* $\{X_n, n \in \mathbb{N}_0\}$ *a Markov chain if*

$$\mathbb{P}(X_n = s_n | X_{n-1} = s_{n-1}, X_{n-2} = s_{n-2}, \ldots, X_0 = s_0) = \mathbb{P}(X_n = s_n | X_{n-1} = s_{n-1})$$

for all $n \in \mathbb{N}_0$ *and all* $s_0, s_1, \ldots, s_n \in \mathcal{S}$ *whenever the two conditional probabilities are well defined, this is, when* $\mathbb{P}(X_n = s_n, X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) > 0$.

Intuitively, this means that the next state in the chain is only dependent on the current state of the chain, and that the history of how the chain got to the current state is irrelevant to the next step. This behaviour is known as the Markov property, and one very valuable insight that emerges from this behaviour is being able to completely determine all properties of the chain given only the initial state of it and the knowledge of how the chain behaves at all of its possible states.

The initial state or initial probability is self-explanatory, it tells us the expected behaviour of the first element in the chain, X_0. Formally, we call $\mu^{(0)} = \{\mu_s^{(0)} : s \in \mathcal{S}\}$ with $\mu_s^{(0)} = \mathbb{P}(X_0 = s)$ the initial probability distribution of the process.

The other important component to fully describe a Markov chain is the transition probabilities from one state s_i to another s_j for all $s_i, s_j \in \mathcal{S}$. We identify this probability as

$$p_{i,j} = \mathbb{P}(X_n = s_j | X_{n-1} = s_i).$$

If we take all combinations of the states the chain can be in and map them to all other states, we get what is known as the transition matrix P, where each entry in the matrix denotes the probability of moving from one specific state to another. This matrix, because it is fundamentally a matrix made up of probabilities, has two interesting properties for the purposes of this work:

- $p_{i,j} \geq 0$ for all $i, j \in \{1, 2, \ldots, n\}$
- $\sum_{i=1}^{n} p_{i,j} = 1$ for all $j \in \{1, 2, \ldots, n\}$.

We note that these two properties are the definition of a stochastic matrix, and therefore, by construction, Markov chain transition matrices are stochastic matrices. We have only described that we can fully determine the Markov chain from its initial state and its transition probabilities, but we have not shown this concept concretely. The following theorem ties these loose ends together and formally describes this property.

Theorem 1. *Let $\{X_n, n \in \mathbb{N}_0\} : \Omega \mapsto \mathcal{S}$ be a stochastic process defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathcal{S}$ a finite state space. Then, $\{X_n, n \in \mathbb{N}_0\}$ is a Markov chain if and only if there exists a stochastic matrix P such that*

$$\mathbb{P}(X_n = s_n | X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) = p_{n-1,n}$$

for all $n \in \mathbb{N}_0$ and all $s_0, s_1, \ldots, s_n \in \mathcal{S}$ such that $\mathbb{P}(X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) > 0$.

This theorem is of great importance in the area of study of Markov chains, and we refer the interested reader to [Chu60, §1.2] for a detailed proof. this theorem establishes a bridge between the theory of Markov processes in general and the properties of linear algebra and matrix theory. One such interesting property is the use of eigenvectors and eigenvalues of the transition matrix used to define a stationary distribution, which is what we are ultimately building towards this section.

We have seen that the transition matrix P governs the behaviour and the changes that happen from one state of the Markov chain to the next. A natural follow up question arises from trying to understand the behaviour of the chain an arbitrary $k > 1$ steps in the future instead of a single step,

such as if we wanted to calculate $\mathbb{P}(X_{i+k} = s_{i+k}|X_i = s_i)$. We first note that this multi-step transition probability can be simplified, assuming that the Markov chain is homogenous, to the following:

$$\mathbb{P}(X_{i+k} = s_{i+k}|X_i = s_i) = \mathbb{P}(X_k = s_k|X_0 = s_0)$$

which we denote as $p_{i,k}^{(k)}$. Because we can define the individual probabilities of transitioning from one step to another after k steps, it stands to reason that there would be a transition matrix to describe this behaviour. The theorem that establishes the way these relationships work is the Chapman-Kolmogorov theorem and equation, which we present now. The formulation of this theorem presented in this work is taken from [Ži10, §8.3].

Theorem 2 (Chapman-Kolmogorov Theorem)**.** *Let $n, m \in \mathbb{N}_0$ and $s_i, s_j \in \mathcal{S}$. Then, $p_{i,j}^{(n+m)}$, which is the probability that starting from state i at step n we end up in state j after m steps is given by:*

$$p_{i,j}^{(n+m)} = \sum_{k \in \mathcal{S}} p_{i,k}^{(n)}\, p_{k,j}^{(m)}.$$

Proof. We will prove the case when the Markov process is homogenous. We begin by noting that $p_{i,j}^{(n+m)} = \mathbb{P}(X_{n+m} = s_j|X_n = s_i) = \mathbb{P}(X_m = s_j|X_0 = s_i) = p_{i,j}^{(m)}$. Applying the law of total probability and using the Markov property,

$$\begin{aligned} p_{i,j}^{(n+m)} &= \\ &= \mathbb{P}(X_{n+m} = s_j|X_n = s_i) \\ &= \mathbb{P}(X_m = s_j|X_0 = s_i) \\ &= \sum_{k \in \mathcal{S}} \frac{\mathbb{P}(X_m = s_j, X_n = s_k, X_0 = s_i)}{\mathbb{P}(X_0 = s_i)} \\ &= \sum_{k \in \mathcal{S}} \mathbb{P}(X_m = s_j|X_n = s_k)\mathbb{P}(X_n = s_k|X_0 = s_i) \\ &= \sum_{k \in \mathcal{S}} p_{i,k}^{(n)}\, p_{k,j}^{(m-n)}. \end{aligned}$$

□

We now introduce the following lemma that takes advantage of this property to provide an alternative way of computing and understanding the transition matrix P.

Lemma 1. *Let P^k be the k-th order matrix power of the transition matrix P for a homogeneous Markov chain $\{X_n, n \in \mathbb{N}_0\}$. Then,*

$$p_{i,j}^{(k)} = (P^k)_{i,j} \text{ for } i, j \in \mathcal{S}.$$

There is a way to prove this lemma by induction directly, but we opt instead to use the Chapman-Kolmogorov Equation n-1 times to achieve the final result, as follows.

Proof. First, we begin by noting that $p_{i,j}^{(k)} = \mathbb{P}(X_{i+j} = s_{i+k} | X_i = s_i) = \mathbb{P}(X_j = s_k | X_0 = s_i)$. Therefore, NEED TO FINISH THIS PROOF. □

This lemma becomes useful in proving the following theorem.

Theorem 3. *Let $n, m \in \mathbb{N}_0$ and $i, j \in \mathcal{S}$. Then, $p_{i,j}^{(n+m)}$, which is the probability that starting from state i in step n we end up in state j after m steps is given by:*

$$p_{i,j}^{(n+m)} = \sum_{k \in \mathcal{S}} p_{i,k}^{(m)} \cdot p_{k,j}^{(n)}.$$

Proof. Supposing a homogeneous Markov chain, then we begin by noting that

$$p_{i,j}^{(n+m)} = \mathbb{P}(X_{n+m} = j | X_n = i) = \mathbb{P}(X_m = j | X_0 = i) = p_{i,j}^{(m)}.$$

Using the law of total probability and the Markov property, NEED TO FINISH THIS PROOF. □

From here, we now present two definitions relating to the properties of matrices which will become important when further exploring the types of Markov chains we are interested in.

Definition 4. *A matrix $A \in \mathbb{R}^{n \times m}$ is called non-negative if $a_{i,j} \geq 0$ for all $1 \leq i \leq n$ and $1 \leq j \leq m$.*

Before continuing, we note that with stochastic matrices being just glorified assortments of probabilities that are neatly ordered, they are always non-negative.

Definition 5. *Let A be a non-negative matrix. If there exists $N_0 \geq 1$ such that all elements of A^{n_0} are positive, then we call A a quasi-positive matrix.*

These two definitions are necessary building blocks to get to the property we are truly interested in, ergodicity.

Definition 6. *A Markov chain* $\{X_n : n \in \mathbb{N}_0\}$ *is called ergodic if the limit* $\pi_j \coloneqq \lim_{n\to\infty} p_{i,j}^{(n)}$

- *Exists for all* $j \in \{1, 2, \ldots, k\}$.
- *Is positive and does not depend on* i.
- $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ *is a probability distribution on* $\mathcal{S}$.

Ergodicity is an incredibly powerful property that is not native to stochastic processes per say, but rather it is inherited from measure theory. In a sense, for a system to be ergodic it needs to be governed by constant or unchanging probability laws, irregardless of a change in state, or in the context of time-centric processes, changes in time. As a simple example, we consider throwing a fair coin in the air 100 times. We would (on average) get the same result if 100 people throw one coin compared to one person throwing the coin 100 times. This would be an ergodic system. On the other hand, if we change the coin to one that is imbalanced or unfair after 50 throws are completed, then the laws governing the change themselves would not remain constantas time passes or we change states, and this would be an example of a non-ergodic system.

The scope of this work does not cover this theory in as much detail as the author would like, but the interested reader is directed to sections 24 and 36 of [Bil12] for a much more complete rundown of Ergodicity looked at through the lenses of measure theory and stochastic processes. The result ergodicity provides that is the foundation for MCMC applications is the following theorem.

Theorem 4. *Let* $\{X_n : n \in \mathbb{N}_0\}$ *be an ergodic Markov chain, such that*

$$\pi_j = \lim_{n\to\infty} p_{i,j}^{(n)}, \text{ for all } j \in \mathcal{S}.$$

Then, the vector π *is a unique solution of*

$$\pi^T = \pi^T P$$

further, π *is a probability distribution on* $\mathcal{S}$.

Proof. Let j be an arbitrary member of $\mathcal{S}$. Then,

$$\begin{aligned}
\pi_j &= \lim_{n\to\infty} p_{i,j}^{(n)} \\
&= \lim_{n\to\infty} p_{i,j}^{(n+1)} \\
&= \lim_{n\to\infty} p_{i,j}^{(n)} \cdot p_{i,j} \\
&= \pi_j \cdot p_{i,j}.
\end{aligned}$$

Since this hods true for all $j \in \mathcal{S}$, then in particular it holds true for the vectorized version of this equation, this being

$$\begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \pi_1, & \pi_2, & \dots \end{bmatrix} \begin{bmatrix} p_{i,1} \\ p_{i,2} \\ \vdots \end{bmatrix}$$

Which is equivalent to saying that $\pi^T = \pi^T P$. □

If we construct a Markov chain in a way that it is ergodic an control the distribution π so that it is something that we want, we can guarantee that after a sufficiently large amount of iterations n we will be sampling from the stationary distribution of the process, which is invariant through time because of the property presented in the above theorem. This is the crux of understanding the Markov Chain Monte Carlo technique which is the central object of study of this section. Delving into the construction of the Metropolis-Hastings algorithm and the fundamental building blocks needed to construct these chains that eventually lead to interesting stationary distributions would make this work exceedingly long, but MISSING REFERENCE and MISSING REFERENCE are good resources for a detailed rundown of these topics. We will return much later in this work when discussing the BASS method to MCMC and why it is so essential to this process. We will not look at pure MCMC though but an interesting modification to the base algorithm known as the Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm.

For now though we move on to the other interesting sub-case of stochastic processes relevant to this work, the continuous-time Gaussian Process.

1.3 Gaussian Processes

A Gaussian process can be thought of as a generalization of the Gaussian (normal) distribution when viewed from the lens of stochastic processes. Stochastic processes, as we saw before, are indexed by a parameter, which itself belongs to an index set. The Gaussian process extends the definition of the Gaussian distribution from being valued on scalars or vectors to a more general index set, to a set $T \subseteq \mathbb{R}^+$. Note that this set can contain an infinite number of points, which is the main difference between the Gaussian distribution and the Gaussian process. This extension is not trivial though, and to specify the structure of these particular types of stochastic processes, we aid ourselves with finite dimensional distributions. We now provide a formal definition of these types of distributions.

Definition 7. *Let $\{t_i\}_{i\leq k}$ be a sequence of points*[2] *such that $\{t_i\}_{i\leq k} \subseteq T$. The k-dimensional random vector $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ has a distribution $\mu_{t_1,t_2,\ldots,t_k}$ over $\mathbb{R}^k$. These measures $\mu_{t_1,t_2,\ldots,t_k}$ are the finite dimensional distributions of the process.*

With this auxiliary distribution out of the way, we can now proceed to define the Gaussian process.

Definition 8. *Let $\{X_t : t \in \mathbb{R}^+\}$ be a continuous-time stochastic process. We call $\{X_t : t \in \mathbb{R}^+\}$ a Gaussian process if for every set $\{t_i\}_{i\leq k}$ and every finite k, the finite-dimensional distribution of the k-dimensional random vector $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ has a Gaussian multivariate distribution when $k > 1$ and a Gaussian univariate distribution when $k = 1$.*

An important extension of this is that we can consider the Gaussian process a function f and evaluate it at certain points x, where $x \in \mathcal{X}^{\Theta}$, the input space. This is, for each x, there is a random variable $f(x)$ that represents the possible output of the function f at this location, or a realization of a certain Gaussian distribution at location x. This can also be understood as taking a sample from the distribution $f(x)$. As with any version of a Gaussian distribution, $f(x)$ can be completely be defined by it's mean and variance. Unlike the finite-dimensional cases though, by definition we can have infinitely many values in the index space T, and therefore the reasonable approach to defining the mean and variance of functions of elements in the index space. We denote the mean function of the process as:

$$m(x) = \mathbb{E}[f(x)]$$

and the covariance function as:

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))].$$

The main reason why the covariance function is denoted by k is that this name alludes to the fact that k is a kernel of the terms in the input space. We now present a formal definition of this type of function, followed by some intuition on its interpretation.

Definition 9. *Let $\mathcal{X}$ be an input space like the ones we have previously defined (in the case of Gaussian processes, $\mathcal{X} = \{X_t : t \in T \subseteq \mathbb{R}^+\}$), and let V be a vector space equipped with an inner product. Also, let φ be a mapping such that $\varphi : \mathcal{X} \times \mathcal{X} \to V$. We call the function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a kernel or kernel function when $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_V$.*

[2]Sometimes referred to as time points, considering this index not as an abstract set but a representation of the progress of time as indexed by a subset of the positive reals.

What this kernel function allows us to do is basically borrow an inner product from a different space, in this definition V, and apply it to the transformed elements of our input or feature space. This allows us to cheat in some regard and compute dot products or metrics in the input space without knowing or needing to know the properties of that space. Tying back the kernel to Gaussian processes, we note that while the mean only controls the average value across realizations of the process f, the covariance or kernel function actually controls the underlying behavior of the realizations or Gaussian processes. As an example of this, we borrow the following figure from [Gha11].

Basically, all major properties such as smoothness, differentiability, sparcity and range (to name only a few) are controlled by the choice of the kernel function.

While the kernel function can be basically anything that fulfills the above definition, there are some particularly useful or popular kernel functions when dealing with Gaussian processes. Chief among them is the squared exponential kernel, also known as the radial basis function. This kernel is defined as

$$k(x, x') = \sigma_f^2 \exp\left(\frac{-\ell(x - x')}{2}\right),$$

where σ_f^2 and ℓ are hyperparameters.

Chapter 2

Gaussian Process Regression

Bibliography

[BBBK] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. page 9.

[Bil12] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 2012.

[BYC+13] James Bergstra, Dan Yamins, David D Cox, et al. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference*, volume 13, page 20. Citeseer, 2013.

[Chu60] Kai Lai Chung. *Markov chains with stationary transition probabilities*. Springer, Berlin, 1960. OCLC: 682058891.

[DMS98] D. G. T. Denison, B. K. Mallick, and A. F. M. Smith. Bayesian mars. *Statistics and Computing*, 8(4):337–346, 1998.

[DMTK20] Erik Daxberger, Anastasia Makarova, Matteo Turchetta, and Andreas Krause. Mixed-variable bayesian optimization. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, page 2633–2639, Jul 2020. arXiv:1907.01329 [cs, stat].

[Fra18] Peter I. Frazier. A tutorial on bayesian optimization. (arXiv:1807.02811), Jul 2018. arXiv:1807.02811 [cs, math, stat].

[Fri91] Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), Mar 1991.

[FS20] Devin Francom and Bruno Sansó. Bass : An r package for fitting and performing sensitivity analysis of bayesian adaptive spline surfaces. *Journal of Statistical Software*, 94(8), 2020.

[Gha11] Zoubin Ghahramani. A tutorial on gaussian processes (or why i don't use svms). In *Proc. MLSS Workshop Talk Gaussian Processes*, 2011.

[GMHL20] Eduardo C. Garrido-Merchán and Daniel Hernández-Lobato. Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes. *Neurocomputing*, 380:20–35, Mar 2020. arXiv:1805.03463 [cs, stat].

[Gre95] Peter J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[HR05] Johannes M Hohendorff and J Rosenthal. An introduction to markov chain monte carlo. *University of Toronto, Department of Statistics, supervised reading report (http://www. probability. ca/jeff/grad. html)*, 2005.

[KFB+17] Aaron Klein, Stefan Falkner, Simon Bartels, Philipp Hennig, and Frank Hutter. Fast bayesian optimization of machine learning hyperparameters on large datasets. (arXiv:1605.07079), Mar 2017. arXiv:1605.07079 [cs, stat].

[LGN+19] Phuc Luong, Sunil Gupta, Dang Nguyen, Santu Rana, and Svetha Venkatesh. *Bayesian Optimization with Discrete Variables*, volume 11919 of *Lecture Notes in Computer Science*, page 473–484. Springer International Publishing, Cham, 2019.

[Ros96] Sheldon M. Ross. *Stochastic processes*. Wiley series in probability and statistics. Wiley, New York, 2nd ed edition, 1996.

[SB] S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets. Published: Retrieved September 10, 2022, from http://www.sfu.ca/ ssurjano.

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. (arXiv:1206.2944), Aug 2012. arXiv:1206.2944 [cs, stat].

[Ži10] Gordan Žitković. Introduction to stochastic processes-lecture notes. *Department of Mathematics, The University of Texas at Austin*, 2010.