Journal of Machine Learning Research () Submitted ;

Probabilistic Programming with Gaussian Process Memoization

Editor:

Abstract

This paper describes the *Gaussian process memoizer*, a probabilistic programming technique that uses Gaussian processes to provide a statistical alternative to memorization. Memoizing a target procedure results in a self-caching wrapper that remembers previously computed values. Gaussian process memoization additionally produces a statistical emulator based on a Gaussian process whose predictions automatically improve whenever a new value of the target procedure becomes available. This paper also introduces an efficient implementation, named `gpmem`, that can use kernels given by a broad class of probabilistic programs. The flexibility of `gpmem` is illustrated via three applications: (i) GP regression with hierarchical hyper-parameter learning, (ii) Bayesian structure learning via compositional kernels generated by a probabilistic grammar, and (iii) a bandit formulation of Bayesian optimization with automatic inference and action selection. All applications share a single 50-line Python library and require fewer than 20 lines of probabilistic code each.

1. Introduction

Probabilistic programming could be revolutionary for machine intelligence due to universal inference engines and the rapid prototyping for novel models (Ghahramani, 2015). This propels the design and testing of new models as well as the incorporation of complex prior knowledge which currently is a difficult and time consuming task. Probabilistic programming languages aim to provide a formal language to specify probabilistic models in the style of computer programming and can represent any computable probability distribution as a program. In this work, we will introduce new features of Venture, a recently developed probabilistic programming language. We consider Venture the most compelling of the probabilistic programming languages because it is the first probabilistic programming language suitable for general purpose use (Mansinghka et al., 2014). Venture comes with scalable performance on hard problems and with a general purpose inference engine. The inference engine deploys Markov Chain Monte Carlo (MCMC) methods (for an introduction, see Andrieu, De Freitas, Doucet, and Jordan (2003)). MCMC lends itself to models with complex structures such as probabilistic programs or hierarchical Bayesian non-parametric models since they can provide a vehicle to express otherwise intractable integrals necessary for a fully Bayesian representation. MCMC is scalable, often distributable and also compositional. That is, one can arbitrarily chain MCMC kernels to infer over several hierarchically connected or nested models as they will emerge in probabilistic programming.

One very powerful model yet unseen in probabilistic programming languages are Gaussian Processes (GPs). GPs are gaining increasing attention for representing unknown func-

tions by posterior probability distributions in various fields such as machine learning, signal processing, computer vision and bio-medical data analysis. Making GPs available in probabilistic programming is crucial to allow a language to solve a wide range of problems. Hard problems include but are not limited to hierarchical prior construction (Neal, 1997), Bayesian Optimization Snoek et al. (2012) and systems for inductive learning of symbolic expressions such as the one introduced in the Automated Statistician project Duvenaud et al. (2013); Lloyd et al. (2014). Learning such symbolic expressions is a hard problem that requires careful design of approximation techniques since standard inference method do not apply.

In the following, we will present `gpmem` as a novel probabilistic programming technique that solves such hard problems. `gpmem` introduces a statistical alternative to standard memoization. Our contribution is threefold:

- we introduce an efficient implementation of `gpmem` in form of a self-caching wrapper that remembers previously computed values;
- we illustrate the statistical emulator that `gpmem` produces and how it improves with every data-point that becomes available; and
- we show how one can solve hard problems of state-of-the-art machine learning related to GP using `gpmem` in a Bayesian fashion and with only a few lines of Venture code.

We evaluate the contribution on problems posed by the GP community using real world and synthetic data by assessing quality in terms of posterior distributions of symbolic outcome and in terms of the residuals produced by our probabilistic programs. The paperis structured as follows, we will first provide some background on memoization. We will explain programming in Venture and provide a brief introduction to GPs. We introduce `gpmem` and its use in probabilistic programming and Bayesian modeling. Finally, we will show how we can apply `gpmem` on problems of causally structured hierarchical priors for hyper-parameter inference, structure discovery for Gaussian Processes and Bayesian Optimization including experiments with real world and synthetic data.

1.1 Memoization

Memoization is the practice of storing previously computed values of a function so that future calls with the same inputs can be evaluated by lookup rather than recomputation. Research on the Church language (Goodman et al., 2008) pointed out that although memoization does not change the semantics of a determinstic program, it does change that of a stochastic program. The authors provide an intuitive example: let f be a function that flips a coin and return "head" or "tails". The probability that two calls of f are equivalent is 0.5. However, if the function call is memoized, it is 1.

In fact, there is an infinite range of possible caching policies (specifications of when to use a stored value and when to recompute), each potentially having a different semantics. Any particular caching policy can be understood by random world semantics (Poole, 1993; Sato, 1995) over the stochastic program: each possible world corresponds to a mapping from function input sequence to function output sequence (McAllester et al., 2008). In Venture, these possible worlds are first-class objects, known as *traces* (Mansinghka et al., 2014).

1.2 Venture

Venture is a compositional language for custom inference strategies that comes with a Scheme-like and a JavaScript-like front-end syntaxes. Its implementation is based on on three concepts: (i) *stochastic procedures* that specify and encapsulate random variables, analogously to conditional probability tables in a Bayesian network; (ii) *execution traces* that represent (partial) execution histories and track the conditional dependencies of the random variables occurring therein; and (iii) *scaffolds* that partition execution histories and factor global inference problems into sub-problems. These building blocks provide a powerful and concise way to represent probability distributions, including distributions with a dynamically determined and unbounded set of random variables. In this paperwe will use only the four basic Venture directives: ASSUME, OBSERVE, SAMPLE and INFER.

- ASSUME induces a hypothesis space for (probabilistic) models including random variables by binding the result of a supplied expression to a supplied symbol.

- Whereas in Scheme an expression is evaluated within an environment, in Venture an expression is evaluated within a (partial) trace of the model program. Thus, the value of an expression within a model program is a random variable, whose randomness comes from the distribution on possible execution traces of the program. The SAMPLE directive samples the value of the supplied expression within the current model program.

- OBSERVE constrains the supplied expression to have the supplied value. In other words, all samples taken after an OBSERVE are conditioned on the observed data.

- INFER uses the supplied inference program to mutate the execution trace. For a correct inference program, this will result approximate sampling from the true posterior on execution traces, conditioned on the model and constraints introduced by ASSUME and OBSERVE. The posterior on any random variable can then be approximately sampled by calling SAMPLE to extract values from the trace.

INFER is commonly done using the Metropolis–Hastings algorithm (MH) (Metropolis et al., 1953). Many of the most popular MCMC algorithms can be interpreted as special cases of MH (Andrieu et al., 2003). We can outline the MH algorithm as follows. The following two-step process is repeated as long as desired (say, for T iterations): First we sample x^* from a proposal distribution q:

$$x^* \sim q(x^* \mid x^t); \tag{1}$$

then we accept this proposal ($x^{t+1} \leftarrow x^*$) with probability

$$\alpha = \min\left\{1, \frac{p(x^*)q(x^t \mid x^*)}{p(x^t)q(x^* \mid x^t)}\right\}; \tag{2}$$

if the proposal is not accepted then we take $x^{t+1} \leftarrow x^t$.

Venture includes a built-in generic MH inference program which performs the above steps on any specified set of random variables in the model program. In that inference program, partial execution traces play the role of x above.

2. Gaussian Processes Memoization

We now introduce GP related theory and notations. We work exclusively with two-variable regression problems. Let the data be pairs of real-valued scalars $\{(x_i, y_i)\}_{i=1}^n$ (complete data will be denoted by column vectors $\mathbf{x}$, $\mathbf{y}$). In regression, one tries to learn a functional relationship $y_i = f(x_i)$, where the function f is to be learned. GPs present a non-parametric way to express prior knowledge on the space of possible functions f. Formally, a GP is an infinite-dimensional extension of the multivariate Gaussian distribution. For any finite set of inputs $\mathbf{x}$, the marginal prior on $f(\mathbf{x})$ is the multivariate Gaussian

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x})),$$

where $m(\mathbf{x}) = \mathbb{E}_f[f(\mathbf{x})]$ is the mean function and $k(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}_f(f(\mathbf{x}), f(\mathbf{x}'))$ is the covariance function, a.k.a. kernel.[1] Together m and k characterize the distribution of f; we write

$$f \sim \mathcal{GP}(m, k).$$

In all examples below, our prior mean function m is identically zero; this is the most common choice. The marginal likelihood can be expressed as:

$$p(f(\mathbf{x}) = \mathbf{y} \mid \mathbf{x}) = \int p(f(\mathbf{x}) = \mathbf{y} \mid f, \mathbf{x})\, p(f|\mathbf{x})\, df \tag{3}$$

where here $p(f|\mathbf{x}) = p(f) \sim \mathcal{GP}(m, k)$ since we assume no dependence of f on $\mathbf{x}$. We can sample a vector of unseen data $\mathbf{y}^* = f(\mathbf{x}^*)$ from the predictive posterior with

$$\mathbf{y}^* \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4}$$

a multivariate normal with mean vector

$$\boldsymbol{\mu} = k(\mathbf{x}^*, \mathbf{x})\, k(\mathbf{x}, \mathbf{x})^{-1}\, \mathbf{y} \tag{5}$$

and covariance matrix

$$\boldsymbol{\Sigma} = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{x}) k(\mathbf{x}, \mathbf{x})^{-1} k(\mathbf{x}, \mathbf{x}^*). \tag{6}$$

Often one assumes the values $\mathbf{y}$ are noisily measured, that is, one only sees the values of $\mathbf{y}_{\text{noisy}} = \mathbf{y} + \mathbf{w}$ where $\mathbf{w}$ is Gaussian white noise with variance σ^2_{noise}. In that case, the log-likelihood is

$$\log p(\mathbf{y}_{\text{noisy}} \mid \mathbf{x}) = -\frac{1}{2}\mathbf{y}^\top (\boldsymbol{\Sigma} + \sigma^2_{\text{noise}}\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log\left|\boldsymbol{\Sigma} + \sigma^2_{\text{noise}}\mathbf{I}\right| - \frac{n}{2}\log 2\pi \tag{7}$$

where n is the number of data points. Both log-likelihood and predictive posterior can be computed efficiently in a Venture SP with an algorithm that resorts to Cholesky factorization(Rasmussen and Williams, 2006, chap. 2) resulting in a computational complexity of $\mathcal{O}(n^3)$ in the number of data points.

The covariance function governs high-level properties of the observed data such as linearity, periodicity and smoothness. The most widely used form of covariance function is the squared exponential:

$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right), \tag{8}$$

where σ and ℓ are hyperparameters: σ is a scaling factor and ℓ is the typical length-scale.

1. Note that $m(\mathbf{x}) = (m(x_i))_{i=1}^n$ and $k(\mathbf{x}, \mathbf{x}') = \left(k(x_i, x'_{i'})\right)_{1 \le i' \le n'}^{1 \le i \le n}$, where n' is the number of entries in $\mathbf{x}'$.

Adjusting hyperparameters results in a new covariance function with the same qualitative human-interpretation; more drastically different covariance functions are achieved by changing the structure of the covariance function. Note that covariance function structures are compositional: adding or multiplying two valid covariance functions results in another valid covariance function.

It has been suggested that adding covariance structures k_1, k_2 together,

$$k_3(x, x') = k_1(x, x') + k_2(x, x'), \tag{9}$$

corresponds to combining global structures, while multiplying covariance functions,

$$k_4(x, x') = k_1(x, x') \times k_2(x, x'), \tag{10}$$

corresponds to combining local structures (Duvenaud et al., 2013). Note that both k_3 and k_4 are valid covariance function structures.

2.1 Venture GPs

The Venture procedure `make-gp` takes as input a mean function and a covariance function, and outputs a procedure for sampling from a Gaussian process. In effect, each call to this procedure samples from (4) conditioned on the return values of all previous samples. `make-gp` allows us to perform GP inference in Venture with only a few lines of code. We can concisely express a wide variety of GPs: simple smoothing with fixed hyper-parameters, or a prior on hyper-parameters, or a custom covariance function. Inference on hyper-parameters can be performed using Venture's built-in MH operator or a custom inference strategy.

Venture code to create and sample from a GP with a smoothing kernel and hyperparameters is shown in Listing 2.1.

```
// HYPER-PARAMETERS
assume sf = tag(quote(hyper), 0, gamma(1,1)))
assume l = tag(quote(hyper), 1, gamma(1,1)))

// COVARIANCE FUNCTION
assume se = make_squaredexp(sf, l)

// MAKE GAUSSIAN PROCESS
assume gp = make_gp( 0, se)

// INCORPORATE OBSERVATIONS
observe gp(array x[1],...x[n])= array(y[1],...,y[n])

// INFER HYPER-PARAMETERS
infer mh(quote(hyper), one, 1)))
```

The first two lines declare the hyper-parameters. We tag both of them to belong to the "scope" `quote(hyper)`. Scopes may be further subdivided into blocks, on which block proposals can be made. In the program above we assign two blocks, 0 and 1. These tags are supplied to the inference program (in this case, MH) to specify on which random variables

inference should be done. In this paper, we use MH inference throughout and do not use block proposals; MH inference is done on one variable at a time.

The ASSUME directives describe the GP model: `sf` and `l` (corresponding to σ and ℓ) are drawn from independent $\Gamma(1, 3)$ distributions. The squared exponential covariance function can be defined outside the Venture code in a conventional programming language (e.g. Python) and imported as a foreign SP. In that way, the user can define custom covariance functions using his or her language and libraries of choice, without having to port existing code into Venture's modelling language. In the above, the factory function `make-se`, which produces a squared exponential function with the supplied hyperparameters, is imported from Python (we have omitted the Python code). In the next line `make-se` is used to produce a covariance function `SE`, whose (random) hyperparameters are `l` and `sf`. Finally, we declare `GP` to be a Gaussian process with mean zero and covariance function `SE`.

2.2 Gaussian process memoization: `gpmem`

`gpmem` is an extension of previous approaches to stochastic memoization. Previous approaches such as the one introduced for Dirichlet Processes in the Church language (Goodman et al., 2008) deal with the discrete-valued domain. We introduce memoization for the continuous valued domain. This is important for problems of regression and optimization where target functions should not be artificially discretized. We present a probabilistic programming technique that uses a GP to provide a statistical variant of memoization.

We implement this by memoizing a target procedure in a wrapper that remembers previously computed values. The technique gives rise to a number of use-cases of GPs in a fully Bayesian setting which are otherwise hard to implement.

2.3 A Bayesian interpretation

We illustrate and compare Bayesian and frequentist view points on GP with a simple example (Fig. 1). We show how in a simple model, two outliers can bias a maximum a posteriori inference. The data where generated with:

$$y = 2x + 15 \tag{11}$$

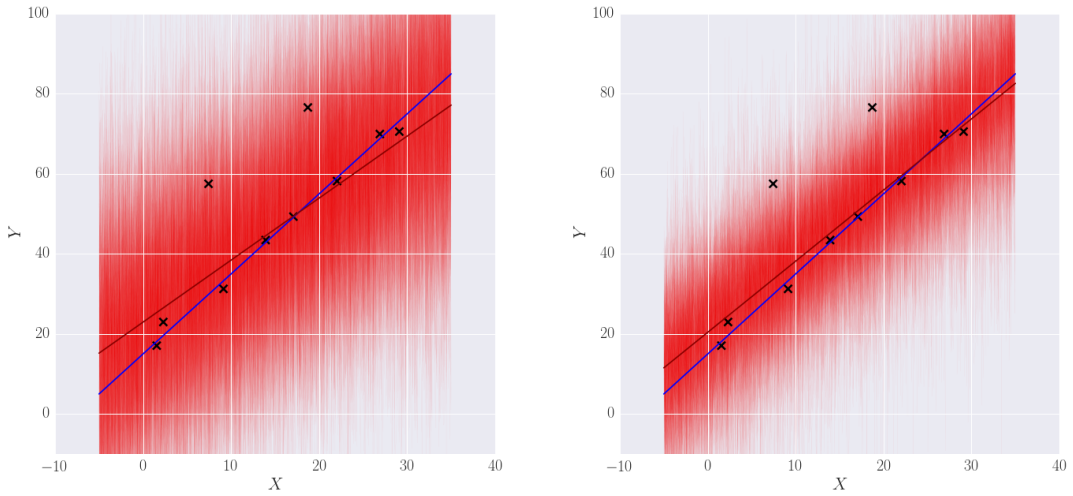
and outliers are generated with a parallel line:

$$\hat{y} = 2x + 40. \tag{12}$$

We add some small amount of white noise. We generate eight data points with (11) and two with (12). Since we suspect the underlying data generating mechanism to be linear, we fit a linear kernel with a constant covariance as intercept and some white noise.

$$\mathbf{K} = \mathrm{LIN} + \mathrm{C} + \mathrm{WN} \tag{13}$$

where we the upper case matrix notation denotes the covariance matrix of the complete training data. Omitting the scaling parameter for the linear kernel, there are two hyper-parameters to learn, that is the noise variance and the hyper-parameter for the constant function. Maximum a posteriori inference fits the single one best line and accounts for the outliers with a large noise scaling parameter. MH does better. It assigns a small amount of probability mass to a different scaling parameter and a larger constant. The resulting prediction (indicating with the predictive mean in figure 1) is closer to the true underlying function.



(a) MAP

(b) MH

Figure 1: (a) depicts MAP inference on the data, (b) depicts MH for hyperparameter inference. The blue line is the actual data generating function. Red are samples drawn from the posterior. The dark red line is the posterior predictive mean. We see that the MH shifts the posterior closer to the ground truth than MAP.

2.3.1 Data modelling as a special case of `gpmem`

From the standpoint of computation, a data set of the form $\{(x_i, y_i)\}$ can be thought of as a function $y = f_{\text{restr}}(x)$, where f_{restr} is restricted to only allow evaluation at a specific set of inputs x. Modelling the data set with a GP then amounts to trying to learn a smooth function f_{emu} ("emu" stands for "emulator") which extends f to its full domain. Indeed, if f_{restr} is a foreign procedure made available as a black-box to Venture, whose secret underlying source code is:

```
def f_restr(x):
  if x in D:
    return D[x]
  else:
    raise Exception('Illegal input')
```

Then the `OBSERVE` code in Listing 2.1 can be rewritten using `gpmem` as follows (where here the data set `D` has keys `x[1]`,...,`x[n]`):

```
[ASSUME (list f_compute f_emu) (gpmem f_restr)]
for i=1 to n:
  [PREDICT (f_compute x[i])]
  [INFER (MH {hyper-parameters} one 100)]
[SAMPLE (f_emu (array 1 2 3))]
```

This rewriting has at least two benefits: (i) readability (in some cases), and (ii) amenability to active learning. As to (i), the statistical code of creating a Gaussian process is replaced with a memoization-like idiom, which will be more familiar to programmers. As to (ii), when

using `gpmem`, it is quite easy to decide incrementally which data point to sample next: for example, the loop from `x[1]` to `x[n]` could be replaced by a loop in which the next index `i` is chosen by a supplied decision rule. In this way, we could use `gpmem` to perform online learning using only a subset of the available data.

3. Example Applications

3.0.2 Hyperparameter estimation with structured hyper-prior

The probability of the hyper-parameters of a GP with assumptions as above and given covariance function structure $\mathbf{K}$ can be described as:

$$P(\boldsymbol{\theta} \mid \mathbf{D}, \mathbf{K}) = \frac{P(\mathbf{D} \mid \boldsymbol{\theta}, \mathbf{K}) P(\boldsymbol{\theta} \mid \mathbf{K})}{P(\mathbf{D} \mid \mathbf{K})}. \tag{14}$$

Let the $\mathbf{K}$ be the sum of a smoothing and a white noise (WN) kernel. For this case, Neal suggested the problem of outliers in data as a use-case for a hierarchical Bayesian treatment of Gaussian processes (1997)[2]. The work suggests a hierarchical system of hyper-parameterization. Here, we draw hyper-parameters from a Γ distributions:

$$\ell^{(t)} \sim \Gamma(\alpha_1, \beta_1),\ \sigma^{(t)} \sim \Gamma(\alpha_2, \beta_2) \tag{15}$$

and in turn sample the α and β from Γ distributions as well:

$$\alpha_1^{(t)} \sim \Gamma(\alpha_\alpha^1, \beta_\alpha^1),\ \alpha_2^{(t)} \sim \Gamma(\alpha_\alpha^2, \beta_\alpha^2), \cdots \tag{16}$$

Assuming the covariance structure is an additive comprised of a smoothing and a white noise kernel, one can represent this kind of model using `gpmem` with only a few lines of code:

2. In (Neal, 1997) the sum of an SE plus a constant kernel is used. We stick to the WN kernel for illustrative purposes.

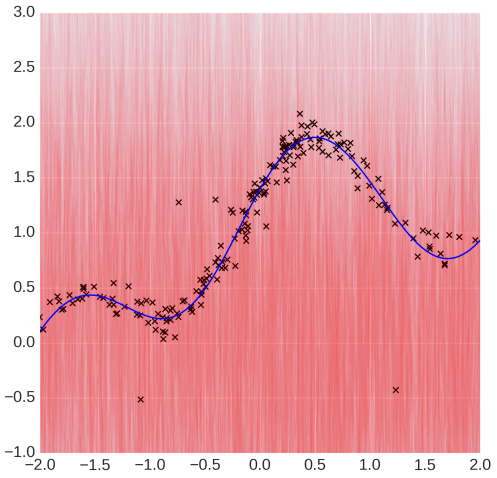
```
/// SETTING UP THE MODEL
assume alpha_sf = tag(quote(hyperhyper), 0, gamma(7, 1))
assume beta_sf = tag(quote(hyperhyper), 1, gamma(7, 1))
assume alpha_l = tag(quote(hyperhyper), 2, gamma(7, 1))
assume beta_l = tag(quote(hyperhyper), 3, gamma(7, 1))

// Parameters of the covariance function
assume sf = tag(quote(hyper), 0, gamma(alpha_sf, beta_sf)))
assume l = tag(quote(hyper), 1, gamma(alpha_l, beta_l)))
assume sigma = tag(quote(hyper), 2, uniform_continuous(0, 2))

// The covariance function
assume se = make_squaredexp(sf, l)
assume wn = make_whitenoise(sigma)
assume covariance = add_funcs(se, wn)

/// PERFORMING INFERENCE
// Create a prober and emulator using gpmem
assume f_restr = get_neal_blackbox()
assume (f_compute, f_emu) = gpmem(f_restr, covariance)

// Probe all data points
predict mapv(f_compute, get_neal_data_xs())

// Infer hypers and hyperhypers
infer repeat(100, do(
    mh(quote(hyperhyper), one, 2),
    mh(quote(hyper), one, 1)))
```

Neal provides a custom inference algorithm setting and evaluates it using the following synthetic data problem. Let f be the underlying function that generates the data:

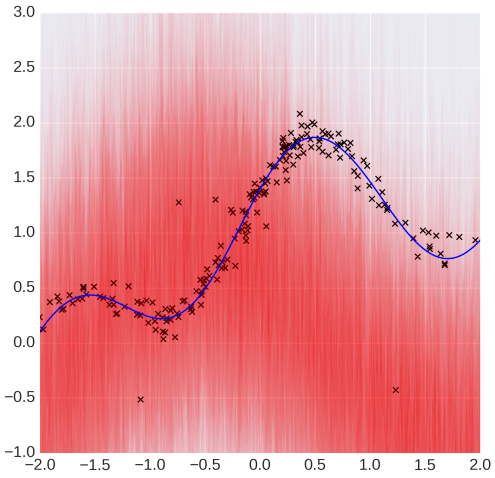
$$f(x) = 0.3 + 0.4x + 0.5\sin(2.7x) + \frac{1.1}{(1+x^2)} + \eta \quad with \;\; \eta \sim \mathcal{N}(0, \sigma) \tag{17}$$

We synthetically generate outliers by setting $\sigma = 0.1$ in 95% of the cases and to $\sigma = 1$ in the remaining cases. `gpmem` can capture the true underlying function within only 100 MH steps on the hyper-parameters to get a good approximation for their posterior (see Fig. 2). Note that Neal devices an additional noise model and performs a large number of Hybrid-Monte Carlo and Gibbs steps. We illustrate the hyper-parameter by showing the shift of the distribution on the noise parameter σ (Fig. 3). We see that `gpmem` learns the posterior distribution well, the posterior even exhibits a bimodal histogram when sampling σ 100 times reflecting the two modes of data generation, that is normal noise and outliers[3].

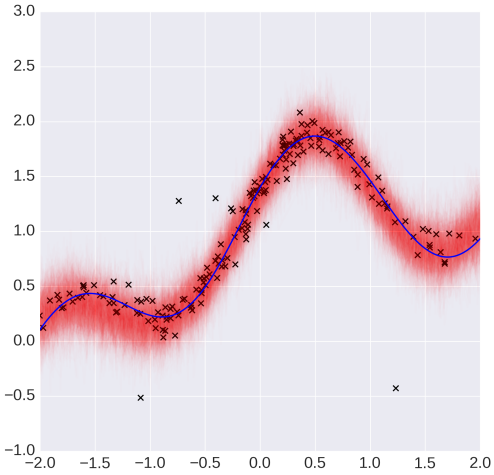
3. For this pedagogical example we have increased the probability for outliers in the data generation slightly from 0.05 to 0.2



(a) Prior Inference

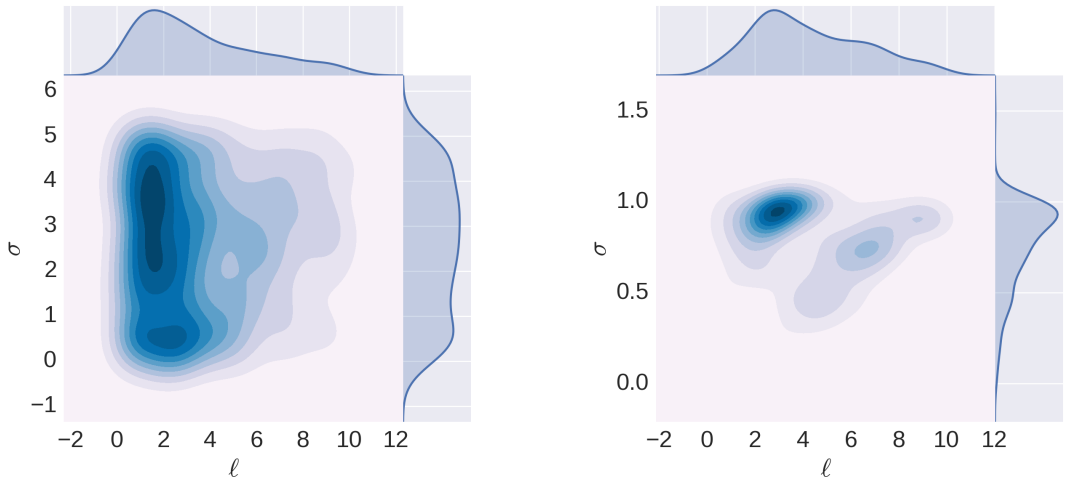


(b) Observed



(c) Inferred

Figure 2: (a)-(c) shows `gpmem` on Neal's example. We see that prior renders functions all over the place (a). After `gpmem` observes a some data-points an arbitrary smooth trend with a high level of noise is sampled. After running inference on the hierarchical system of hyper-parameters we see that the posterior reflects the actual curve well. Outliers are treated as such and do not confound the GP.



(a) Before (b) After

Figure 3: Hyper-parameter inference on the parameter of the noise kernel. We show a 100 samples drawn from the distribution on σ. One can clearly recognise the shift from the uniform prior $\mathcal{U}(0, 5)$ to a double peak distribution around the two modes - normal and outlier.

3.0.3 Broader applicability of `gpmem`

More generally, `gpmem` is relevant not just when a data set is available, but also whenever we have at hand a function f_{restr} which is expensive or impractical to evaluate many times. `gpmem` allows us to model f_{restr} with a GP-based emulator f_{emu}, and also to use f_{emu} during the learning process to choose, in an online manner, an effective set of probe points $\{x_i\}$ on which to use our few evaluations of f_{restr}. This idea is illustrated in detail in Section 3.2. Before doing this, we will illustrate another benefit of having a probabilistic programming apparatus for GP modelling: the linguistically unified treatment of inference over structure and inference over parameters. This unification makes interleaved joint inference over structure and parameters very natural, and allows us to give a short, elegant description of what it means to "learn the covariance function," both in prose and in code. Furthermore, the example in Section 3.1 below recovers the performance of current state-of-the-art GP-based models.

3.1 Discovering symbolic models for time series

The space of possible kernel composition is infinite. Combining inference over this space with the problem of finding a good parameterization that could potentially explain the observed data best poses a hard problem. The natural language interpretation of the meaning of a kernel and its composition renders this a problem of symbolic computation. Duvenaud and colleagues note that a sum of kernels can be interpreted as logical OR operations and kernel multiplication as logical AND (2013). This is due to the kernel rendering two points similar if k_1 OR k_2 outputs a high value in the case of a sum. Respectively, multiplication of two kernels results in high values only if k_1 AND k_2 have high values (see Fig. 4 exemplifies how

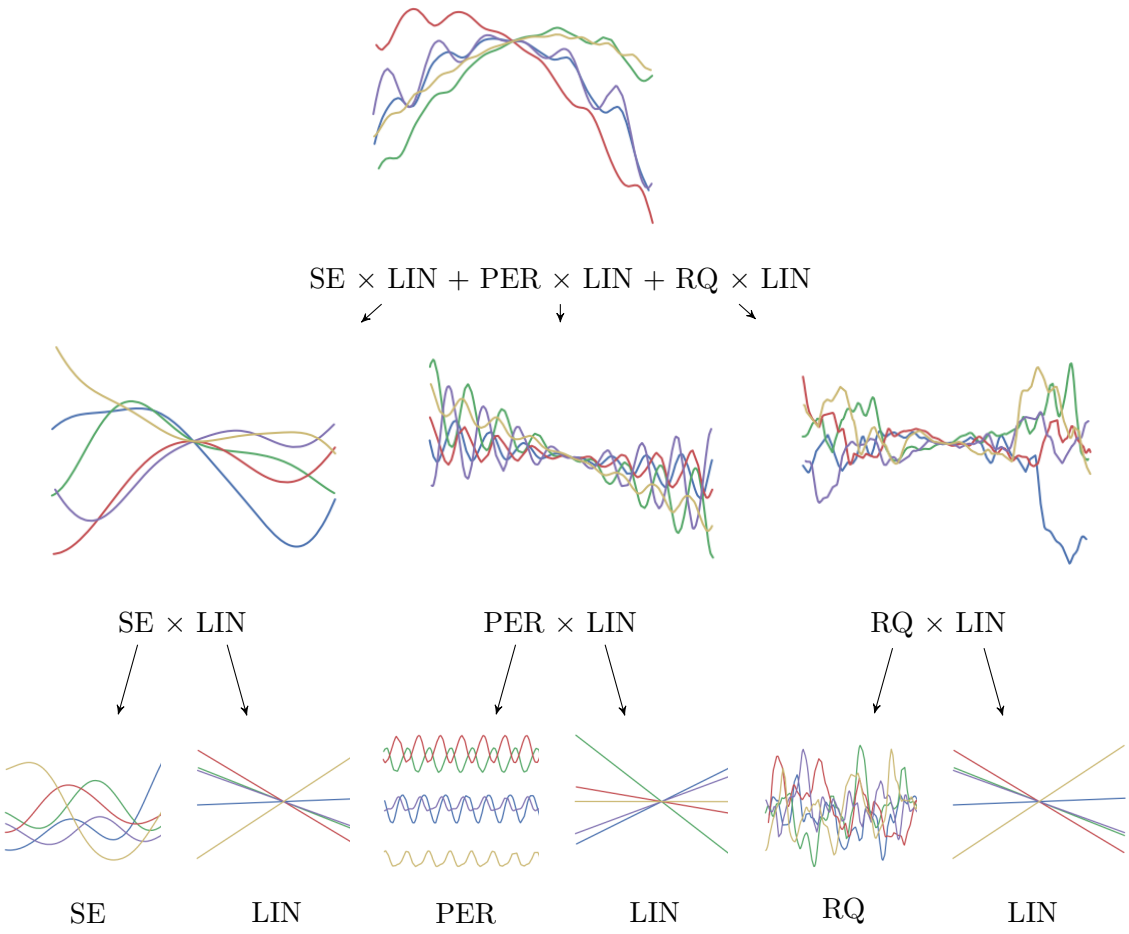


Figure 4: Composition of covariance function parsed using the example SE × LIN + PER × LIN + RQ × LIN

to interpret global vs. local aspects and its symbolic analog respectively). In the following, we will refer to covariance functions that are not composite as base covariance functions.

Knowledge about the composite nature of covariance functions is not new, however, until recently, the choice and the composition of covariance functions were done ad-hoc. The Automated Statistician Project came up with an approximate search over the possible space of kernel structures (Duvenaud et al., 2013; Lloyd et al., 2014). However, a fully Bayesian treatment of this was not done before. To the best of our knowledge, we are the first one to present a fully Bayesian solution to this. We deploy a probabilistic context free grammar for our prior on structures(see Fig. 5)

Our probabilistic programming based MCMC framework approximates the following intractable integrals of the expectation for the prediction:

$$\mathbb{E}[y^* \mid x^*, \mathbf{D}, \mathbf{K}] = \iint f(x^*, \boldsymbol{\theta}, \mathbf{K})\, P(\boldsymbol{\theta} \mid \mathbf{D}, \mathbf{K})\, P(\mathbf{K}|\boldsymbol{\Omega}, s, n)\, \mathbf{d}\boldsymbol{\theta}\mathbf{d}\mathbf{K}. \tag{18}$$

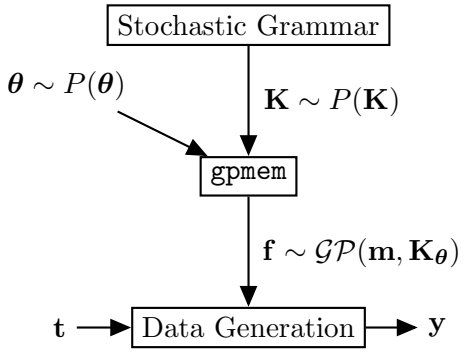


Figure 5: Graphical description of Bayesian GP structure learning.

This is done by sampling from the posterior probability distribution of the hyper-parameters and the possible kernel:

$$y^* \approx \frac{1}{T} \sum_{t=1}^{T} f(x^* | \boldsymbol{\theta}^{(t)}, \mathbf{K}^{(t)}). \tag{19}$$

In order to provide the sampling of the kernel, we introduce a stochastic process to the SP that simulates the grammar for algebraic expressions of covariance function algebra:

$$\mathbf{K}^{(t)} \sim P(\mathbf{K} \mid \boldsymbol{\Omega}, s, n) \tag{20}$$

Here, we start with a set of possible kernels and draw a random subset. For this subset of size n, we sample a set of possible operators that operate on the base kernels.

The marginal probability of a kernel structure which allows us to sample is characterized by the probability of a uniformly chosen subset of the set of n possible covariance functions times the probability of sampling a global or a local structure which is given by a binomial distribution:

$$P(\mathbf{K} \mid \boldsymbol{\Omega}, s, n) = P(\boldsymbol{\Omega} \mid s, n) \times P(s \mid n) \times P(n), \tag{21}$$

with

$$P(\boldsymbol{\Omega} \mid s, n) = \binom{n}{r} p_{+\times}^{k} (1 - p_{+\times})^{n-k} \tag{22}$$

and

$$P(s \mid n) = \frac{n!}{\mid s \mid !} \tag{23}$$

where $P(n)$ is a prior on the number of base kernels used which can sample from a discrete uniform distribution. This will strongly prefer simple covariance structures with few base kernels since individual base kernels are more likely to be sampled in this case due to (23). Alternatively, we can approximate a uniform prior over structures by weighting $P(n)$ towards higher numbers. It is possible to also assign a prior for the probability to sample global or local structures, however, we have assigned complete uncertainty to this with the probability of a flip $p = 0.5$.

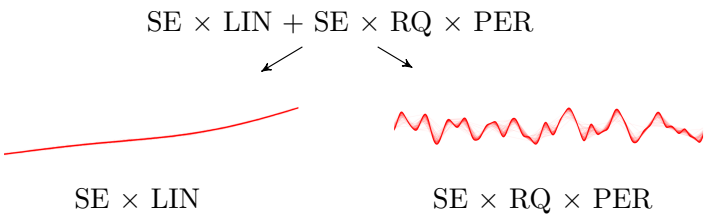
Many equivalent covariance structures can be sampled due to covariance function algebra and equivalent representations with different parameterization (Lloyd et al., 2014). Certain covariance functions can differ in terms of the hyper-parameterization but can be absorbed into a single covariance function with a different parameterization. To inspect the posterior

of these equivalent structures we convert each kernel expression into a sum of products and subsequently simplify. Rules for this simplification can be found in appendix B.

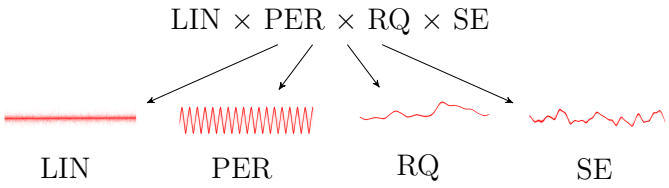
For reproducing results from the Automated Statistician Project in a Bayesian fashion we first define a prior on the hypothesis space. Note that, as in the implementation of the Automated Statistician, we upper-bound the complexity of the space of covariance functions we want to explore. We also put vague priors on hyper-parameters.

```
// GRAMMAR FOR KERNEL STRUCTURE
assume base_kernels = list(se, wn, lin, per, rq) // as above

// prior on the number of kernels
assume p_number_k = tag(quote(pnk), 0, uniform_structure(n))
assume s = tag(quote(choice), 0, subset(base_kernels, p_number_k))

assume cov_compo = proc(l) {
  // kernel composition
  if (size(l) <= 1)
  then { first(l) }
  else { if (flip())
                then { add_funcs(first(l), cov_compo(rest(l))) }
                else { mult_funcs(first(l), cov_compo(rest(l))) }
       }
  }

assume K = tag(quote(composit), 0, cov_compo(s))

assume (f_compute f_emu) = gpmem(f_restr, K)
predict mapv(f_compute(get_data_xs)) // probe all data points

// PERFORMING INFERENCE
infer  repeat(2000, do(
                        mh(quote(pnk), one, 1),
                        mh(quote(choice), one, 1),
                        mh(quote(composit), one, 1),
                        for kernel ∈ K
                          mh(quote(hyper_kernel), one, 10)))
```

We defined the space of covariance structures in a way allowing us to reproduce results for covariance function structure learning as in the Automated Statistician. This lead to coherent results, for example for the airline data set describing monthly totals of international airline passengers (Box et al., 1997, according to Duvenaud et al., 2013. The sample is identical with the highest scoring result reported in previous work using a search-and-score method (Duvenaud et al., 2013) for the CO_2 data set (see Rasmussen and Williams, 2006 for a description) and the predictive capability is comparable. However, the components factor in a different way due to different parameterization of the individual base kernels. We see that that the most probable alternatives for a structural description both recover the data dynamics (Fig. 6 for the airline data set).



(a)



(b)

Figure 6: (a) most frequent sample drawn from the posterior on structure. We have found two global components. First, A smooth trend (LINxSE) with a non-linear increasing slope. Second, a periodic component with increasing variation and noise. (b) second most frequent sample drawn from the posterior on structure. We found one global component. It is comprised of local changes that are periodic and with changing variation.

We further investigated the quality of our stochastic processes by running a leave one out cross-validation computing residuals to gain confidence on the posterior. Confident about our results, we can now query the data for certain structures being present. We illustrate this using the Mauna Loa data used in previous work on automated kernel discovery (Duvenaud et al., 2013). We assume a relatively simple hypothesis space consisting of only four kernels, a linear, a smoothing, a periodic and a white noise kernel. In this experimemt, we resort to the white noise kernel instead RQ (similar to (Lloyd et al., 2014)). We can now run the algorithm, compute a posterior of structures. We can also query this posterior distribution for the marginal of certain simple structures to occur. We demonstrate this in Fig. 7

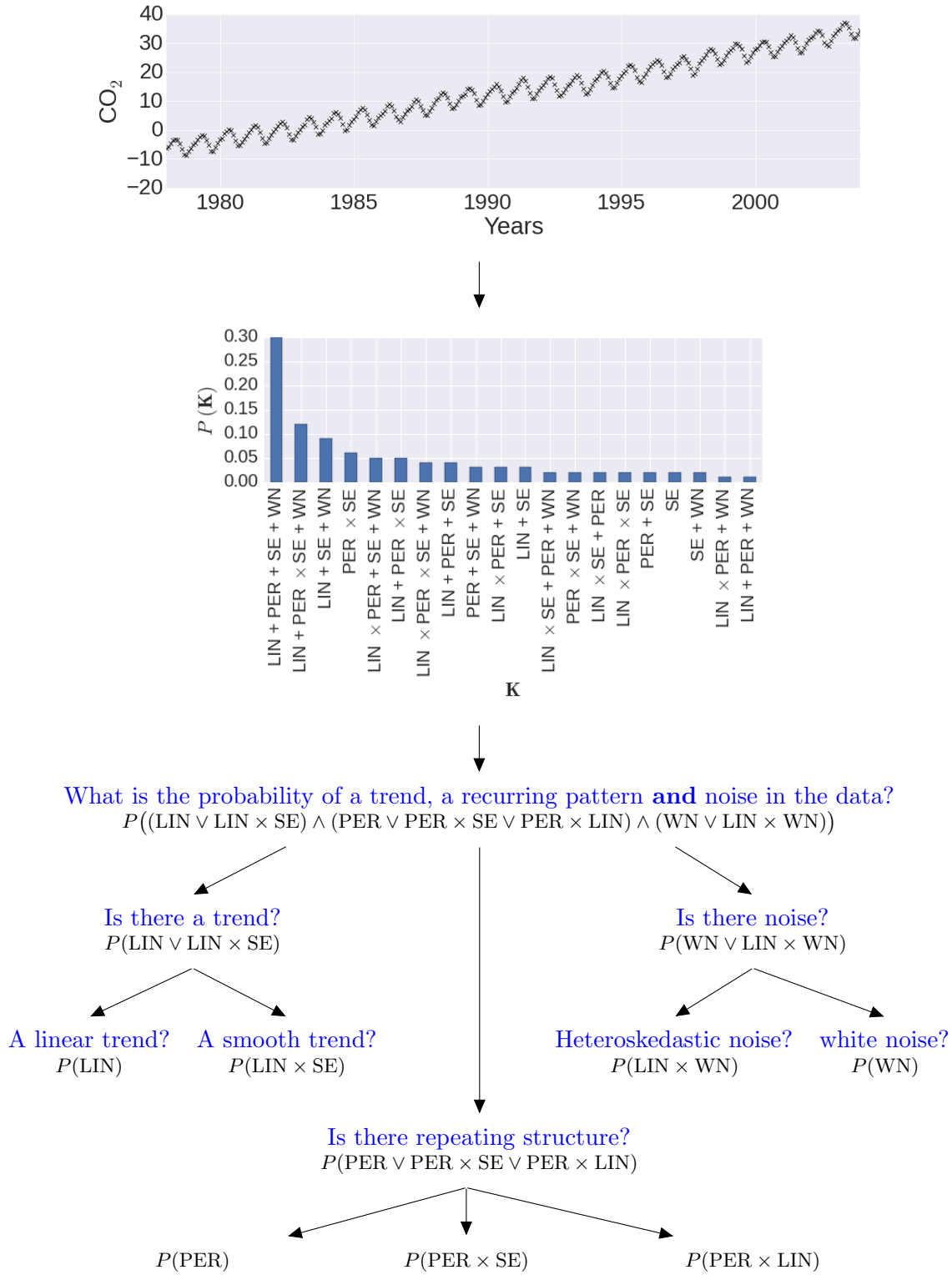


Figure 7: We can query the data, asking, similar to symbolic reasoning, if some statements are probable to be true. In this example, we are interested in three qualitative aspects of data, which we formulate in terms of the following: 1. Is it true that there is a trend? Is it true that there are recurring patterns and is it true that that there is noise in the data? Given the answers to these questions, we can then ask more specific questions.

3.2 Bayesian optimization via Thompson sampling

Bayesian optimization casts the problem of finding the global maximum of an unknown function as a hierarchical decision problem (Ghahramani, 2015). Evaluating the actual function may be very expensive, either in computation time or in some other resource. For one example, when searching for the best configuration for the learning algorithm of a large convolutional neural network, a large amount of computational work is required to evaluate a candidate configuration, and the space of possible configurations is high-dimensional. Another common example, alluded to in Section 3.0.3, is data acquisition: for machine learning problems in which a large body of data is available, it is often desirable to choose the right queries to produce a data set on which learning will be most effective. In continuous settings, many Bayesian optimization methods employ GPs (e.g. Snoek et al., 2012).

We have implemented a version of Thompson sampling using GPs in Venture. Thompson sampling Thompson (1933) is a widely-used Bayesian framework for solving exploration-exploitation problems. Our implementation has two notable features: (i) the ability to search over a broader space of contexts than the parametric families that are typically used, and (ii) the parsimony of the resulting probabilistic program.

3.2.1 Thompson sampling framework

We now lay out the setup of Thompson sampling for Markov decision processes (MDPs). An agent is to take a sequence of actions $a_1, a_2, \ldots$ from a (possibly infinite) set of possible actions $\mathcal{A}$. After each action, a reward $r \in \mathbb{R}$ is received, according to an unknown conditional distribution $P_{\text{true}}(r|a)$. The agent's goal is to maximize the total reward received for all actions. In Thompson sampling, the Bayesian agent accomplishes this by placing a prior distribution $P(\theta)$ on the possible "contexts" $\theta \in \Theta$. Here a context is a believed model of the conditional distributions $\{P(r|a)\}_{a\in\mathcal{A}}$, or at least, a believed statistic of these conditional distributions which is sufficient for deciding an action a. One example of such a sufficient statistic is the conditional mean $V(a|\theta) = \mathbb{E}[r|a,\theta]$, which can be thought of as a value function. Thompson sampling thus has the following steps, repeated as long as desired:

1. Sample a context $\theta \sim P(\theta)$.
2. Choose an action $a \in \mathcal{A}$ which (approximately) maximizes $V(a|\theta) = \mathbb{E}[r|a,\theta]$.
3. Let r_{true} be the reward received for action a. Update the believed distribution on θ, i.e., $P(\theta) \leftarrow P_{\text{new}}(\theta)$ where $P_{\text{new}}(\theta) = P(\theta \mid a \mapsto r_{\text{true}})$.

Note that when $\mathbb{E}[r|a,\theta]$ (under the sampled value of θ for some points a) is far from the true value $\mathbb{E}_{P_{\text{true}}}[r|a]$, the chosen action a may be far from optimal, but the information gained by probing action a will improve the belief θ. This amounts to "exploration." When $\mathbb{E}[r|a,\theta]$ is close to the true value except at points a for which $\mathbb{E}[r|a,\theta]$ is low, exploration will be less likely to occur, but the chosen actions a will tend to receive high rewards. This amounts to "exploitation." Roughly speaking, exploration will happen until the context θ is reasonably sure that the unexplored actions are probably not optimal, at which time the sampler will exploit by choosing actions in regions it knows to have high value.

Typically, when Thompson sampling is implemented, the search over contexts $\theta \in \Theta$ is limited by the choice of representation. In traditional programming environments, θ often consists of a few numerical parameters for a family of distributions of a fixed functional form. With work, a mixture of a few functional forms is possible; but without probabilistic programming machinery, implementing a rich context space Θ would be an unworkably large technical burden. In a probabilstic programing language, however, the representation of heterogeneously structured or infinite-dimensional context spaces is quite natural. Any computable model of the conditional distributions $\{P(r|a)\}_{a\in\mathcal{A}}$ can be represented as a stochastic procedure $(\lambda(a)\ldots)$. Thus, for computational Thompson sampling, the most general context space $\widehat{\Theta}$ is the space of program texts. Any other context space Θ has a natural embedding as a subset of $\widehat{\Theta}$.

3.2.2 Thompson sampling in Venture

Because Venture supports sampling and inference on (stochastic-)procedure-valued random variables (and the generative models which produce those procedures), Venture can capture arbitrary context spaces as described above. To demonstrate, we have implemented Thompson sampling in Venture in which the contexts θ are Gaussian processes over the action space $\mathcal{A} = \mathbb{R}$. That is, $\theta = (\mu, K)$, where the mean μ is a computable function $\mathcal{A} \to \mathbb{R}$ and the covariance K is a computable (symmetric, positive-semidefinite) function $\mathcal{A} \times \mathcal{A} \to \mathbb{R}$. This represents a Gaussian process $\{R_a\}_{a\in\mathcal{A}}$, where R_a represents the reward for action a. Computationally, we represent a context not as a pair of infinite lookup tables for μ and K, but as a finite data structure $\theta = (K_{\text{prior}}, \sigma, \ell, \mathbf{a}_{\text{past}}, \mathbf{r}_{\text{past}})$, where

- $K_{\text{prior}} = K_{\text{prior},\sigma,\ell}$ is a procedure, with parameters σ, ℓ, to be used as the prior covariance function: $K_{\text{prior}}(a, a') = \sigma^2 \exp\left(-\frac{(a-a')^2}{2\ell^2}\right)$
- σ and ℓ are (hyper)parameters for K_{prior}
- $\mathbf{a}_{\text{past}} = (a_i)_{i=1}^n$ are the previously probed actions
- $\mathbf{r}_{\text{past}} = (r_i)_{i=1}^n$ are the corresponding rewards

To simplify the treatment, we take prior mean $\mu_{\text{prior}} \equiv 0$. The mean and covariance for θ are then gotten by the usual conditioning formula:

$$\begin{aligned}
\mu(\mathbf{a}) &= \mu\left(\mathbf{a} \mid \mathbf{a}_{\text{past}}, \mathbf{r}_{\text{past}}\right) \\
&= K_{\text{prior}}(\mathbf{a}, \mathbf{a}_{\text{past}})\, K_{\text{prior}}(\mathbf{a}_{\text{past}}, \mathbf{a}_{\text{past}})^{-1}\, \mathbf{r}_{\text{past}} \\
K(\mathbf{a}, \mathbf{a}) &= K\left(\mathbf{a}, \mathbf{a} \mid \mathbf{a}_{\text{past}}, \mathbf{r}_{\text{past}}\right) \\
&= K_{\text{prior}}(\mathbf{a}, \mathbf{a}) - K_{\text{prior}}(\mathbf{a}, \mathbf{a}_{\text{past}})\, K_{\text{prior}}(\mathbf{a}_{\text{past}}, \mathbf{a}_{\text{past}})^{-1}\, K_{\text{prior}}(\mathbf{a}_{\text{past}}, \mathbf{a}).
\end{aligned}$$

Note that even in this simple example, the context space Θ is not a finite-dimensional parametric family, since the vectors $\mathbf{a}_{\text{past}}$ and $\mathbf{r}_{\text{past}}$ grow as more samples are taken. Θ is, however, quite easily representable as a computational procedure together with parameters and past samples, as we do in the representation $\theta = (K_{\text{prior}}, \sigma, \ell, \mathbf{a}_{\text{past}}, \mathbf{r}_{\text{past}})$.

3.2.3 Implementation with `gpmem`

As a demonstration, we use Thompson sampling to optimize an unknown function $V(x)$ (the value function) using `gpmem`. We assume V is made available to Venture as a black-box. The code for optimizing V is given in Listing 1. For step 3 of Thompson sampling, the Bayesian update, we not only condition on the new data (the chosen action a and the received reward r), but also perform inference on the hyperparameters σ, ℓ using a Metropolis–Hastings sampler. These two inference steps take 1 line of code: 0 lines to condition on the new data (as this is done automatically by `gpmem`), and 1 line to call Venture's built-in `MH` operator. The results are shown in Figure 8. We can see from the figure that, roughly speaking, each successive probe point a is chosen either because the current model V_{emu} thinks it will have a high reward, or because the value of $V_{\text{emu}}(a)$ has high uncertainty. In the latter case, probing at a decreases this uncertainty and, due to the smoothing kernel, also decreases the uncertainty at points near a. We thus see that our Thompson sampler simultaneously learns the value function and optimizes it.

Listing 1: Code for Bayesian optimization using `gpmem`. In the loop, `V_compute` is called to probe the value of `V` at a new argument. The new argument, `(mc_argmax V_emu_pointwise mc_sampler)`, is a Monte Carlo estimate of the maximum pointwise sample of `V_emu` (itself a stochastic quantity), with the Monte Carlo samples being drawn in this case uniformly between -20 and 20. After each new call to `V_compute`, the Metropolis–Hastings algorithm is used to perform inference on the hyperparameters of the covariance function in the GP model in light of the new conditioning data. Once enough calls to `V_compute` have been made (in our case we stopped at 15 calls), we can inspect the full list of probed (a, r) pairs with `extract_stats`. The answer to our maximization problem is simply the pair having the highest r; but our algorithm also learns more potentially useful information.

```
assume sf = tag(quote(hyper), 0, uniform_continuous(0, 10))
assume l = tag(quote(hyper), 1, uniform_continuous(0, 10))
assume se = make_squaredexp(sf, l)
assume blackbox_f = get_bayesopt_blackbox()
assume (f_compute, f_emu) = gpmem(blackbox_f, se)

define get_uniform_candidate = proc(prev_xs) {
  uniform_continuous(-20, 20)
}

define mc_argmax = proc(func, prev_xs) {
  // Monte Carlo estimator for the argmax of func.
  run(do(
    candidate_xs <- mapv(proc(i) {get_uniform_candidate(prev_xs)},
                         linspace(0, 19, 20)),
    candidate_ys <- mapv(func, candidate_xs),
    lookup(candidate_xs, argmax_of_array(candidate_ys))))
}

define emulator_point_sample = proc(x) {
  run(sample(lookup(
    f_emu(array(x)),
    0)))
}

infer repeat(15, do(pass,
     // Phase 2: Call f_compute on the next probe point
     predict f_compute(
                 mc_argmax(emulator_point_sample, '_)),
     // Phase 1: Hyperparameter inference
     mh(quote(hyper), one, 50)))
```

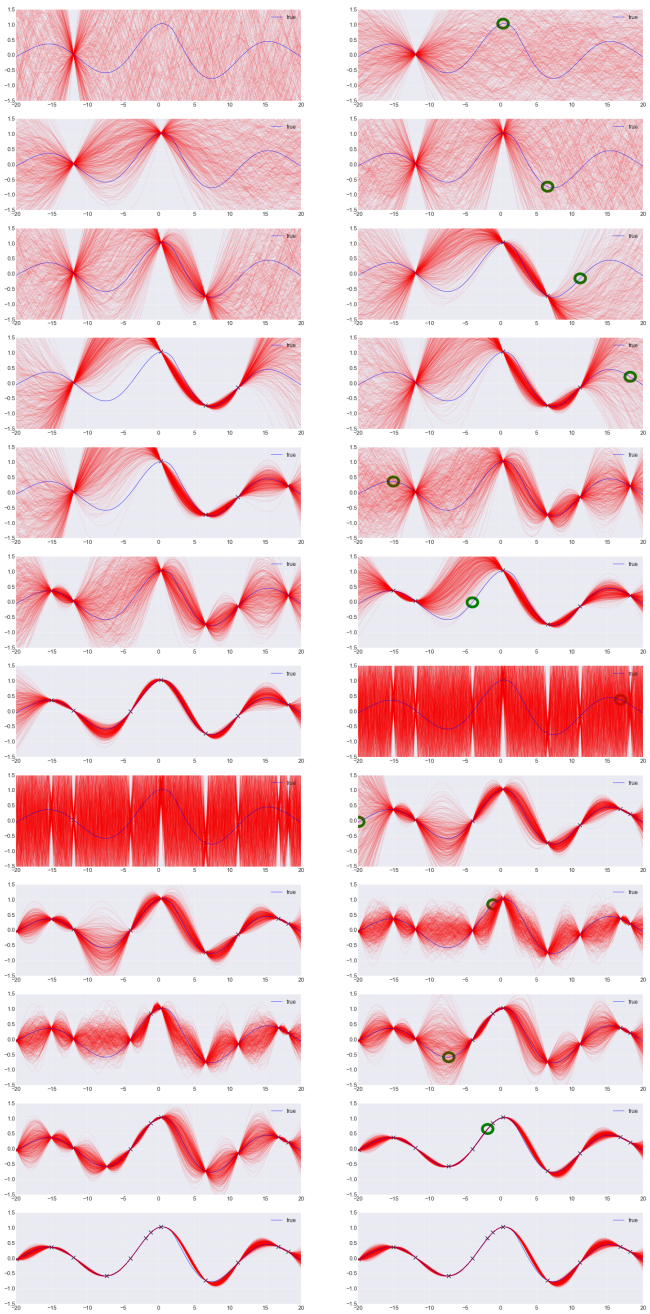


Figure 8: Dynamics of Thompson sampling in Venture. The blue curve is the true function V, and the red region is a blending of 100 samples of the curve generated (jointly) by a GP-based emulator V_{emu}. The left and right columns show the state of V_{emu} before and after hyperparameter inference is run on the new data, respectively. (We can see, for example, that after the seventh probe point, the Metropolis–Hastings sampler chose a "crazy" set of hyperparameters, which was corrected at the next inference step.) In the right column, the next chosen probe point is circled in green. Each successive probe point a is the (stochastic) maximum of V_{emu}, sampled pointwise and conditioned on the values of the previously probed points. Note that probes tend to happen at points either where the value of V_{emu} is high, or where V_{emu} has high uncertainty.

4. Discussion

We have shown Venture GPs. We have introduced novel stochastic processes for a probabilistic programming language. We showed how flexible non-parametric models can be treated in Venture in only a few lines of code. We evaluated our contribution on a range of hard problems for state-of-the-art Bayesian non-parametrics. Venture GPs showed competitive performance in all of them.

4.1 Contributions

Appendix

A Covariance Functions

SE and WN are defined in the text above, for completeness we will introduce the covariance:

$$k_{LIN}(x, x') = \theta(xx') \tag{24}$$

$$k_{PER}(x, x') = \theta \exp\left(\frac{2\sin^2(\pi(x - x')/p}{\ell^2}\right) \tag{25}$$

$$k_{RQ}(x, x') = \theta\left(1 + \frac{(x - x')^2}{2\alpha\ell^2}\right)^{-\alpha} \tag{26}$$

B Covariance Simplification

`SE`	$\times$	`SE`
`{SE,PER,C,WN}`	$\times$	`WN`
`LIN`	+	`LIN`
`{SE,PER,C,WN,LIN}`	$\times$	`C`
	$\rightarrow$	`{SE,PER,C,WN,LIN}`

Rule 1 is derived as follows:

$$\begin{aligned}
\sigma_c^2 \exp(-\frac{(x - x')^2}{2\ell_c^2}) &= \sigma_a^2 \exp(-\frac{(x - x')^2}{2\ell_a^2}) \times \sigma_b^2 \exp(-\frac{(x - x')^2}{2\ell_b^2}) \\
&= \sigma_c^2 \exp(-\frac{(x - x')^2}{2\ell_a^2}) \times \exp(-\frac{(x - x')^2}{2\ell_b^2}) \\
&= \sigma_c^2 \exp\left(-\frac{(x - x')^2}{2\ell_a^2} - \frac{(x - x')^2}{2\ell_b^2}\right) \\
&= \sigma_c^2 \exp\left(-\frac{(x - x')^2}{2\ell_c^2}\right)
\end{aligned} \tag{27}$$

Rule 3 is derived as follows:

$$\theta_c(x \times x') = \theta_a(x \times x') + \theta_b(x \times x') \tag{28}$$

For stationary kernels that only depend on the lag vector between x and x' it holds that multiplying such a kernel with a WN kernel we get another WN kernel. Take for example the SE kernel:

$$\sigma_a^2 \exp\left(-\frac{(x - x')^2}{2\ell_c^2}\right) \times \sigma_b \delta_{x,x'} = \sigma_a \sigma_b \delta_{x,x'} \tag{29}$$

Multiplying any kernel with a constant obviously changes only the scale parameter of a kernel.

References

C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time series analysis: forecasting and control*. 1997.

D. Duvenaud, J. R. Lloyd, R. Grosse, J. Tenenbaum, and Z. Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1166–1174, 2013.

Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553): 452–459, 2015.

V. K. Goodman, N. D .and Mansinghka, D. Roy, K. Bonawitz, and J.B. Tenenbaum. Church: A language for generative models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, UAI 2008*, pages 220–229, 2008.

J. R. Lloyd, D. Duvenaud, R. Grosse, J. Tenenbaum, and Z. Ghahramani. Automatic construction and natural-language description of nonparametric regression models. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

V. K. Mansinghka, D. Selsam, and Y. Perov. Venture: a higher-order probabilistic programming platform with programmable inference. *arXiv preprint arXiv:1404.0099*, 2014.

D. McAllester, B. Milch, and N. D. Goodman. Random-world semantics and syntactic independence for expressive languages. Technical report, 2008.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21 (6):1087–1092, 1953.

R. M. Neal. Monte carlo implementation of gaussian process models for bayesian regression and classification. *arXiv preprint physics/9701026*, 1997.

D. Poole. Probabilistic horn abduction and bayesian networks. *Artificial Intelligence*, 64 (1):81–129, 1993.

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

T. Sato. A statistical learning method for logic programs with distribution semantics. In *In Proceedings of the 12th International Conference on Logic Programming*. Citeseer, 1995.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.