
Probabilistic Programming with Gaussian Process Memoization

Anonymous Author(s)

Affiliation

Address

email

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 levitates 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 et al. (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 functions 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

054 to hierarchical prior construction (Neal, 1997), Bayesian Optimization Snoek et al. (2012) and sys-
055 tems for inductive learning of symbolic expressions such as the one introduced in the Automated
056 Statistician project Duvenaud et al. (2013); Lloyd et al. (2014). Learning such symbolic expressions
057 is a hard problem that requires careful design of approximation techniques since standard inference
058 method do not apply.

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

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

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

078 2 Background

079 2.1 Memoization

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

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

094 2.2 Venture

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

- 107 • 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 :

$$x^* \sim q(x^* | x^t); \quad (1)$$

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

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

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.

2.3 Gaussian Processes

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 $k(\mathbf{x}, \mathbf{x}') = \text{Cov}_f (f(\mathbf{x}), f(\mathbf{x}'))$ 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:

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

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

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

a multivariate normal with mean vector

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

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

162 and covariance matrix

$$163 \quad \Sigma = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{x})k(\mathbf{x}, \mathbf{x})^{-1}k(\mathbf{x}, \mathbf{x}^*). \quad (6)$$

165 Often one assumes the values \mathbf{y} are noisily measured, that is, one only sees the values of $\mathbf{y}_{\text{noisy}} =$
 166 $\mathbf{y} + \mathbf{w}$ where \mathbf{w} is Gaussian white noise with variance σ_{noise}^2 . In that case, the log-likelihood is
 167

$$168 \quad \log p(\mathbf{y}_{\text{noisy}} | \mathbf{x}) = -\frac{1}{2}\mathbf{y}^\top (\Sigma + \sigma_{\text{noise}}^2 \mathbf{I})^{-1}\mathbf{y} - \frac{1}{2} \log |\Sigma + \sigma_{\text{noise}}^2 \mathbf{I}| - \frac{n}{2} \log 2\pi \quad (7)$$

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

174 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:
 175

$$177 \quad k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right), \quad (8)$$

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

180 Adjusting hyperparameters results in a new covariance function with the same qualitative human-
 181 interpretation; more drastically different covariance functions are achieved by changing the structure
 182 of the covariance function. Note that covariance function structures are compositional: adding or
 183 multiplying two valid covariance functions results in another valid covariance function. We suggest
 184 (**TODO** cite or point to later in paper) that adding covariance structures k_1, k_2 together,

$$185 \quad k_3(x, x') = k_1(x, x') + k_2(x, x'), \quad (9)$$

186 corresponds to combining global structures, while multiplying covariance functions,
 187

$$188 \quad k_4(x, x') = k_1(x, x') \times k_2(x, x'), \quad (10)$$

189 corresponds to combining local structures. Note that both k_3 and k_4 are valid covariance function
 190 structures.

192 3 Venture GPs

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

201 Venture code to create and sample from a GP with a
 202 smoothing kernel and hyperparameters is shown in Listing 3.

```
// HYPER-PARAMETERS
assume log_sf = tag('hyper, log(gamma(1, 1)))
assume log_l = tag('hyper, log(gamma(1, 1)))

// COVARIANCE FUNCTION
assume se = make_squaredexp(log_sf, log_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('hyper, one, 1))
```

The figure consists of two side-by-side scatter plots. Both plots have a horizontal axis labeled X ranging from -10 to 40 and a vertical axis labeled Y ranging from 0 to 100. Each plot contains several data points represented by black 'x' marks and a solid blue line representing a linear regression model. In the left plot, the background is a uniform light gray. In the right plot, the background is a vertical gradient from light gray at the bottom to red at the top. The data points and the regression line are identical in both plots.

Figure 1: Left, MAP estimate, right Metropolis Hastings

Figure 1: Left, MAP estimate, right Metropolis Hastings

The first two lines declare the hyper-parameters. We tag both of them to belong to the “scope” ‘hyper. 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. Scopes may be further subdivided into blocks, on which block proposals can be made. In this paper we 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`.

I don't know what these two paragraphs mean –Ben

In the case where hyper-parameters are unknown they can be found deterministically by optimizing the marginal likelihood using a gradient based optimizer. Non-deterministic, Bayesian representations of this case are also known (Neal, 1997).

We have already implemented this in listing 3. We draw the hyper-parameters from a Γ -prior for a Bayesian treatment of hyper-parameters. This is simple using the build in stochastic procedure that simulates drawing samples from a gamma distribution. The program gives rise to a Bayesian representation of GPs, which we will explore in the following.

3.1 Gaussian process memoization: `gpmem`

TODO write

270 **3.2 A Bayesian interpretation**
 271

272 **3.2.1 Data modelling as a special case of `gpmem`**
 273

274 From the standpoint of computation, a data set of the form $\{(x_i, y_i)\}$ can be thought of as a function
 275 $y = f_{\text{restr}}(x)$, where f_{restr} is restricted to only allow evaluation at a specific set of inputs x . Modelling
 276 the data set with a GP then amounts to trying to learn a smooth function f_{emu} (“emu” stands for
 277 “emulator”) which extends f to its full domain. Indeed, if f_{restr} is a foreign procedure made available
 278 as a black-box to Venture, whose secret underlying source code is:

```
279 def f_restr(x):
280     if x in D:
281         return D[x]
282     else:
283         raise Exception('Illegal input')
```

284 Then the `OBSERVE` code in Listing 3 can be rewritten using `gpmem` as follows (where here the data
 285 set `D` has keys $x[1], \dots, x[n]$):

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

291 This rewriting has at least two benefits: (i) readability (in some cases), and (ii) amenability to active
 292 learning. As to (i), the statistical code of creating a Gaussian process is replaced with a memoization-
 293 like idiom, which will be more familiar to programmers. As to (ii), when using `gpmem`, it is quite
 294 easy to decide incrementally which data point to sample next: for example, the loop from $x[1]$ to
 295 $x[n]$ could be replaced by a loop in which the next index i is chosen by a supplied decision rule.
 296 In this way, we could use `gpmem` to perform online learning using only a subset of the available
 297 data.

298
 299 **3.2.2 The efficacy of learning hyperparameters**
 300

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

$$305 P(\boldsymbol{\theta} | \mathbf{D}, \mathbf{K}) = \frac{P(\mathbf{D} | \boldsymbol{\theta}, \mathbf{K})P(\boldsymbol{\theta} | \mathbf{K})}{P(\mathbf{D} | \mathbf{K})}. \quad (11)$$

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

$$314 \ell^{(t)} \sim \Gamma(\alpha_1, \beta_1), \sigma^{(t)} \sim \Gamma(\alpha_2, \beta_2) \quad (12)$$

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

$$319 \alpha_1^{(t)} \sim \Gamma(\alpha_\alpha^1, \beta_\alpha^1), \alpha_2^{(t)} \sim \Gamma(\alpha_\alpha^2, \beta_\alpha^2), \dots \quad (13)$$

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

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

```

326 //////////////////////////////////////////////////////////////////
327 // SETTING UP THE MODEL
328 assume alpha_sf = tag('hyperhyper, gamma(7, 1))
329 assume beta_sf = tag('hyperhyper, gamma(7, 1))
330 assume alpha_l = tag('hyperhyper, gamma(7, 1))
331 assume beta_l = tag('hyperhyper, gamma(7, 1))

332 // Parameters of the covariance function
333 assume log_sf = tag('hyper, log(gamma(alpha_sf, beta_sf)))
334 assume log_l = tag('hyper, log(gamma(alpha_l, beta_l)))
335 assume log_sigma = tag('hyper, log(uniform_continuous(0, 2)))

336 // The covariance function
337 assume se = make_squaredexp(log_sf, log_l)
338 assume wn = make_whitenoise(log_sigma)
339 assume composite_covariance = add_funcs(se, wn)

340 //////////////////////////////////////////////////////////////////
341 // PERFORMING INFERENCE
342 // Create a prober and emulator using gpmem
343 assume f_restr = get_neal_blackbox()
344 assume (f_compute, f_emu) = gpmem(f_restr, composite_covariance)

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

347 // Infer hypers and hyperhypers
348 infer repeat(100, do(
349   mh('hyperhyper, one, 2),
350   mh('hyper, one, 1)))
351
352

```

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

$$355 \quad 356 \quad f(x) = 0.3 + 0.4x + 0.5 \sin(2.7x) + \frac{1.1}{(1+x^2)} + \eta \quad \text{with } \eta \sim \mathcal{N}(0, \sigma) \quad (14)$$

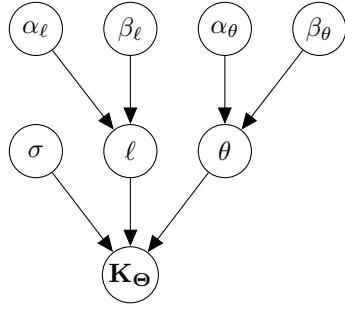
357 We synthetically generate outliers by setting $\sigma = 0.1$ in 95% of the cases and to $\sigma = 1$ in the
 358 remaining cases. `gpmem` can capture the true underlying function within only 100 MH steps on the
 359 hyper-parameters to get a good approximation for their posterior (see Fig. 2). Note that Neal devices
 360 an additional noise model and performs large number of Hybrid-Monte Carlo and Gibbs steps. We
 361 illustrate the hyper-parameter by showing the shift of the distribution on the noise parameter σ (Fig.
 362 3). We see that `gpmem` learns the posterior distribution well, the posterior even exhibits a bimodal
 363 histogram when sampling σ 100 times reflecting the two modes of data generation, that is normal
 364 noise and outliers³.

365 3.2.3 Broader applicability of `gpmem`

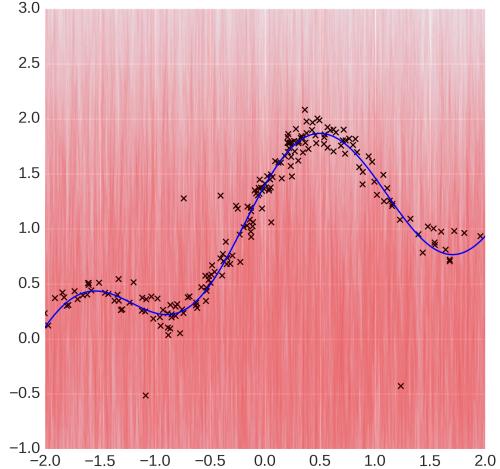
366 More generally, `gpmem` is relevant not just when a data set is available, but also whenever we
 367 have at hand a function f_{restr} which is expensive or impractical to evaluate many times. `gpmem`
 368 allows us to model f_{restr} with a GP-based emulator f_{emu} , and also to use f_{emu} during the learning
 369 process to choose, in an online manner, an effective set of probe points $\{x_i\}$ on which to use our few
 370 evaluations of f_{restr} . This idea is illustrated in detail in Section 4. Before doing this, we will illustrate
 371 another benefit of having a probabilistic programming apparatus for GP modelling: the linguistically
 372 unified treatment of inference over structure and inference over parameters. This unification makes
 373 interleaved joint inference over structure and parameters very natural, and allows us to give a short,
 374 elegant description of what it means to “learn the covariance function,” both in prose and in code.

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

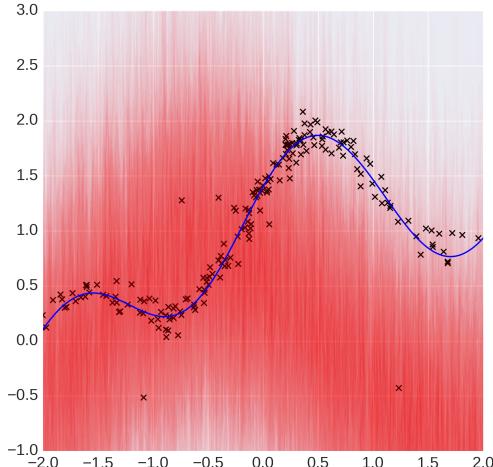
378
 379
 380
 381
 382
 383
 384
 385
 386
 387
 388
 389
 390
 391
 392
 393
 394
 395
 396
 397
 398
 399
 400



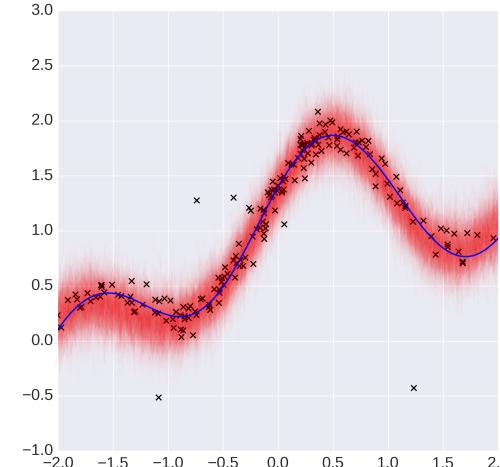
(a) Hierarchical Prior



(b) Prior Inference



(c) Observed



(d) Inferred

422 Figure 2: (a) depicts the hierarchical structure of the hyper-parameter as constructed in the work by
 423 Neal as a Bayesian Network. (b)-(d) shows gpmem on Neal's example. We see that prior renders
 424 functions all over the place (a). After gpmem observes a some data-points an arbitrary smooth trend
 425 with a high level of noise is sampled. After running inference on the hierarchical system of hyper-
 426 parameters we see that the posterior reflects the actual curve well. Outliers are treated as such and
 427 do not confound the GP.
 428
 429
 430
 431

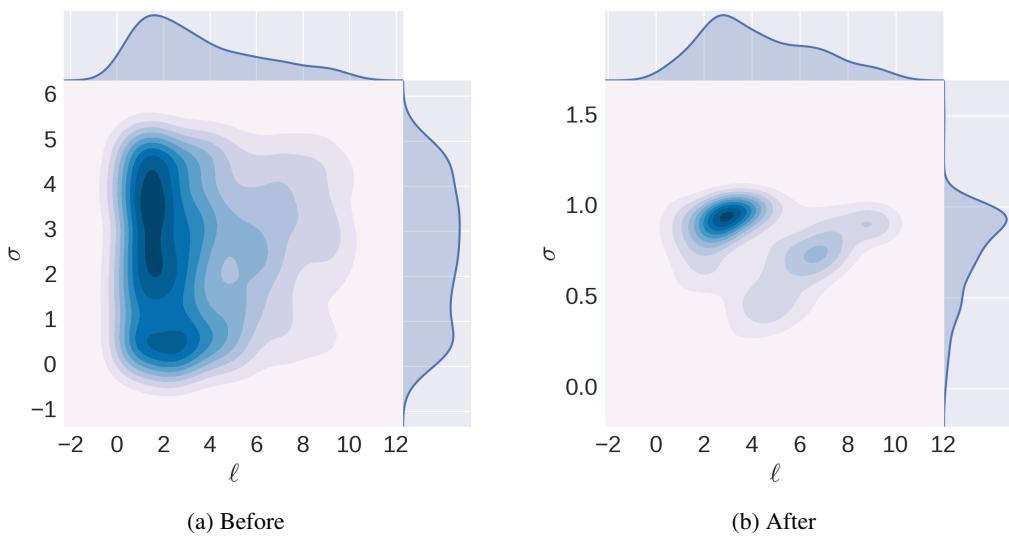


Figure 3: Hyper-parameter inference on the parameter of the noise kernel. We show a 100 samples drawn from the distribution on σ . One can clearly recognise the shift from the uniform prior $\mathcal{U}(0, 5)$ to a double peak distribution around the two modes - normal and outlier.

Furthermore, the example in Section 3.3 below recovers the performance of current state-of-the-art GP-based models.

3.3 Structure Learning

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 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. The case where the covariance structure is not given is even more interesting. Our probabilistic programming based MCMC framework approximates the following intractable integrals of the expectation for the prediction:

$$\mathbb{E}[y^* \mid x^*, \mathbf{D}, \mathbf{K}] = \iint f(x^*, \boldsymbol{\theta}, \mathbf{K}) P(\boldsymbol{\theta} \mid \mathbf{D}, \mathbf{K}) P(\mathbf{K} \mid \boldsymbol{\Omega}, s, n) \, d\boldsymbol{\theta} d\mathbf{K}. \quad (15)$$

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

$$y^* \approx \frac{1}{T} \sum_{t=1}^T f(x^* | \theta^{(t)}, \mathbf{K}^{(t)}). \quad (16)$$

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:

$$\mathbf{K}^{(t)} \sim P(\mathbf{K} \mid \Omega, s, p) \quad (17)$$

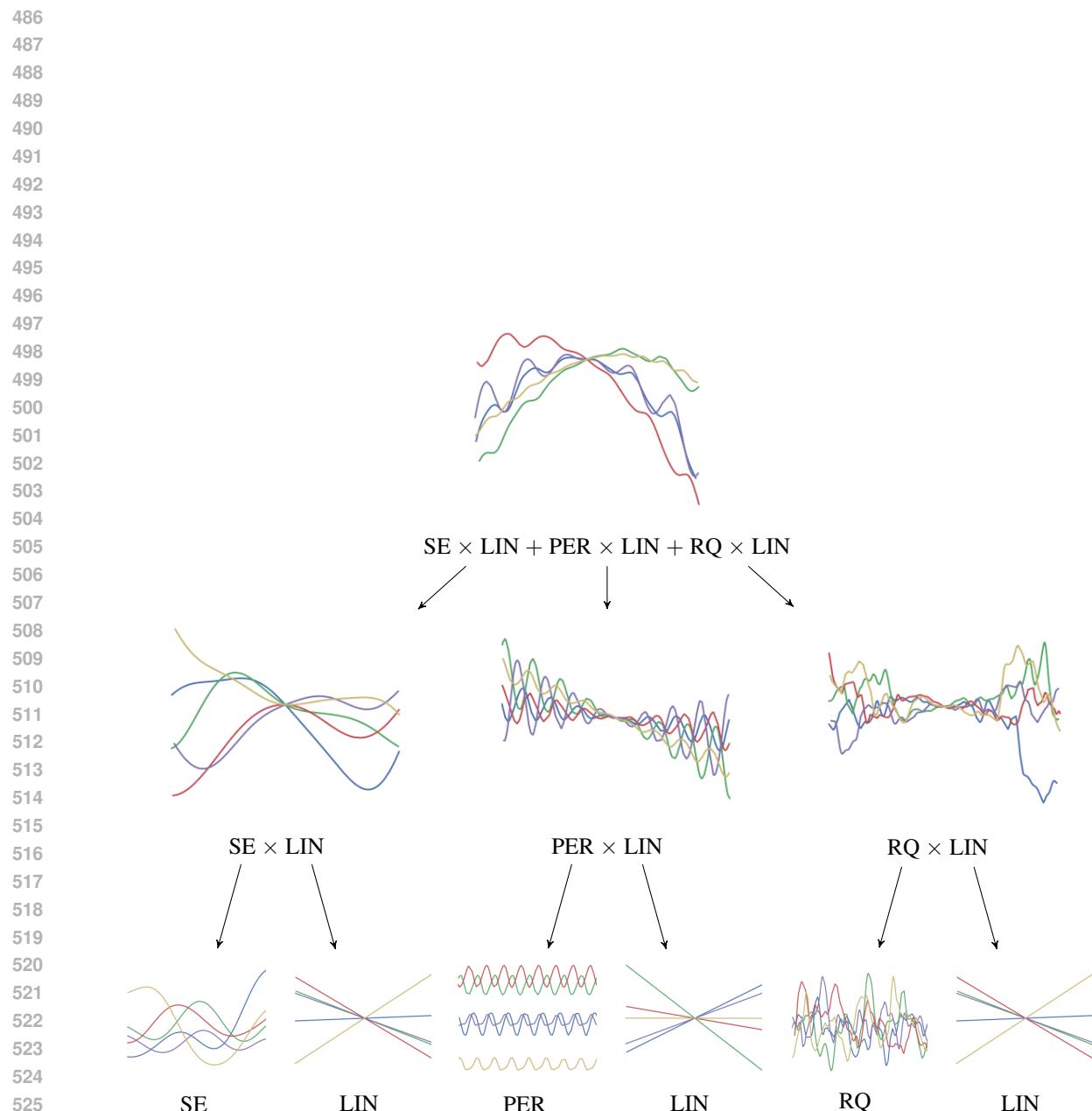


Figure 4: Composition of covariance function parsed using the example $\text{SE} \times \text{LIN} + \text{PER} \times \text{LIN} + \text{RQ} \times \text{LIN}$

540 Here, we start with a set of possible kernels and draw a random subset. For this subset of size n , we
 541 sample a set of possible operators that operate on the base kernels.
 542

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

$$P(\mathbf{K} \mid \boldsymbol{\Omega}, s, n) = P(\boldsymbol{\Omega} \mid s, n) \times P(s \mid n) \times P(n), \quad (18)$$

546 with
 547

$$P(\boldsymbol{\Omega} \mid s, n) = \binom{n}{r} p_{+ \times}^k (1 - p_{+ \times})^{n-k} \quad (19)$$

548 and
 549

$$P(s \mid n) = \frac{n!}{|s|!} \quad (20)$$

550 where $P(n)$ is a prior on the number of base kernels used which can sample from a discrete uniform
 551 distribution. This will strongly prefer simple covariance structures with few base kernels since
 552 individual base kernels are more likely to be sampled in this case due to (20). Alternatively, we
 553 can approximate a uniform prior over structures by weighting $P(n)$ towards higher numbers. It is
 554 possible to also assign a prior for the probability to sample global or local structures, however, we
 555 have assigned complete uncertainty to this with the probability of a flip $p = 0.5$.
 556

557 Many equivalent covariance structures can be sampled due to covariance function algebra and equiv-
 558 alent representations with different parameterization (Lloyd et al., 2014). Certain covariance func-
 559 tions can differ in terms of the hyper-parameterization but can be absorbed into a single covariance
 560 function with a different parameterization. To inspect the posterior of these equivalent structures
 561 we convert each kernel expression into a sum of products and subsequently simplify. Rules for this
 562 simplification can be found in appendix B.

563 For reproducing results from the Automated Statistician Project in a Bayesian fashion we first define
 564 a prior on the hypothesis space. Note that, as in the implementation of the Automated Statistician,
 565 we upper-bound the complexity of the space of covariance functions we want to explore. We also
 566 put vague priors on hyper-parameters.
 567

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

// prior on the number of kernels
assume p_number_k = tag('number_kernels, uniform_structure(n))
assume s = tag('choice_subset, 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('composit, 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('number_kernel, one, 1),
    mh('choice_subset, one, 1),
    mh('composit, one, 1),
    mh('hyper, one, 10)))
```

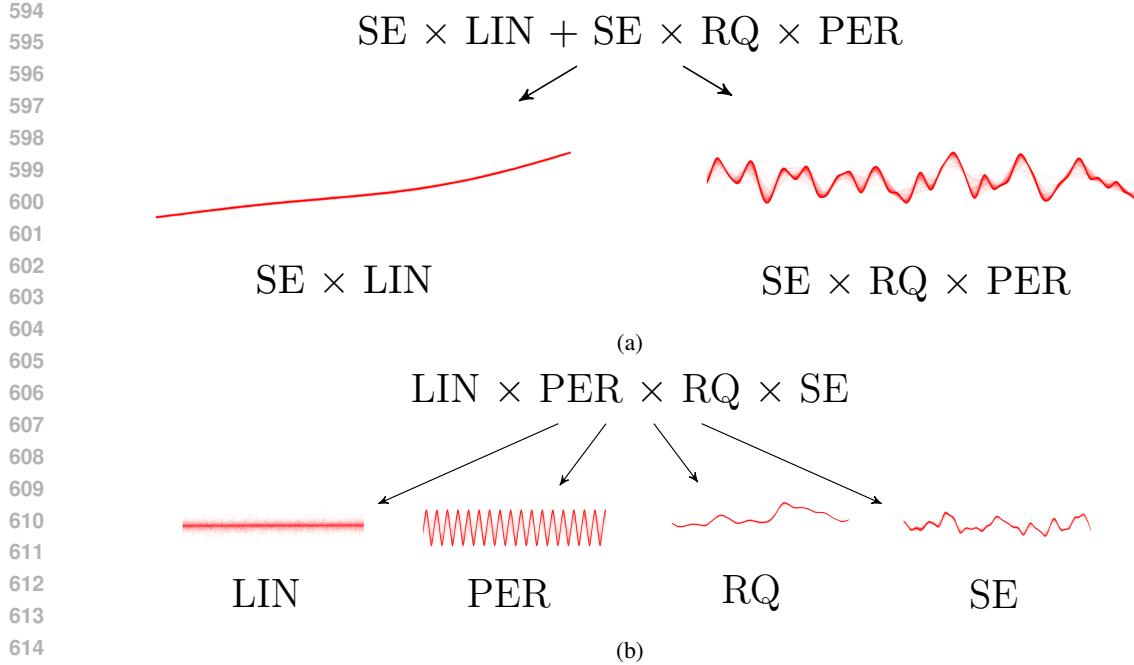


Figure 5: (a) most frequent sample drawn from the posterior on structure. We have found two global components. First, A smooth trend (LINxSE) with a non-linear increasing slope. Second, a periodic component with increasing variation and noise. (b) second most frequent sample drawn from the posterior on structure. We found one global component. It is comprised of local changes that are periodic and with changing variation.

We 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). We will elaborate the result using a sample from the posterior (Fig. ??). 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 CO₂ 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 further investigated the quality of our stochastic processes by running a leave one out cross-validation to gain confidence on the posterior. This resulted in 545 independent runs of the Markov chain that produced a coherent posterior: our Bayesian interpretation of GP structure and GPs produced a posterior of structures that is in line with previous results on this data set (Duvenaud et al., 2013; see Fig. ??).

We ran similar evaluation on the airline data set resulting in a similar structure to what was previously reported (Fig. ??, residuals and log-score along the Markov chain see Fig. ??).

We found the final sample of multiple runs to be most informative. This kind of Markov Chain seems to produce samples that are highly auto-correlated.

Given two additive components $\mathbf{K} = \mathbf{K}_a + \mathbf{K}_b$, one can compute the marginal of a global component of a composite kernel structure (Benavoli and Mangili, 2015) with a gaussian posterior $\mathcal{N}(f_a | \hat{\mu}_a, \hat{\mathbf{K}}_a)$ where:

$$\hat{\mu}_a = \mathbf{K}_a(\mathbf{x}, \mathbf{x}^*) \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*)^{-1} \mathbf{y} \quad (21)$$

and covariance matrix

$$\hat{\mathbf{K}}_a = \mathbf{K}_a(\mathbf{x}, \mathbf{x}) - \mathbf{K}_a(\mathbf{x}, \mathbf{x}^*) \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*)^{-1} \mathbf{K}_a(\mathbf{x}^*, \mathbf{x}). \quad (22)$$

648 **4 Bayesian Optimization**
649
650

651 Bayesian optimization casts the problem of finding the global maximum of an unknown function as
652 a hierarchical decision problem (Ghahramani, 2015). Evaluating the actual function may be very
653 expensive, either in computation time or in some other resource. For one example, when searching
654 for the best configuration for the learning algorithm of a large convolutional neural network, a large
655 amount of computational work is required to evaluate a candidate configuration, and the space of
656 possible configurations is high-dimensional. Another common example, alluded to in Section 3.2.3,
657 is data acquisition: for machine learning problems in which a large body of data is available, it is
658 often desirable to choose the right queries to produce a data set on which learning will be most
659 effective. In continuous settings, many Bayesian optimization methods employ GPs (e.g. Snoek
660 et al., 2012).

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

666 **4.1 Thompson sampling framework**
667

668 We now lay out the setup of Thompson sampling for Markov decision processes (MDPs). An agent
669 is to take a sequence of actions a_1, a_2, \dots from a (possibly infinite) set of possible actions \mathcal{A} . After
670 each action, a reward $r \in \mathbb{R}$ is received, according to an unknown conditional distribution $P_{\text{true}}(r|a)$.
671 The agent's goal is to maximize the total reward received for all actions. In Thompson sampling,
672 the Bayesian agent accomplishes this by placing a prior distribution $P(\theta)$ on the possible "contexts"
673 $\theta \in \Theta$. Here a context is a believed model of the conditional distributions $\{P(r|a)\}_{a \in \mathcal{A}}$, or at least,
674 a believed statistic of these conditional distributions which is sufficient for deciding an action a .
675 One example of such a sufficient statistic is the conditional mean $V(a|\theta) = \mathbb{E}[r|a, \theta]$, which can be
676 thought of as a value function. Thompson sampling thus has the following steps, repeated as long
677 as desired:

- 678 1. Sample a context $\theta \sim P(\theta)$.
679
680 2. Choose an action $a \in \mathcal{A}$ which (approximately) maximizes $V(a|\theta) = \mathbb{E}[r|a, \theta]$.
681
682 3. Let r_{true} be the reward received for action a . Update the believed distribution on θ , i.e.,
683 $P(\theta) \leftarrow P_{\text{new}}(\theta)$ where $P_{\text{new}}(\theta) = P(\theta | a \mapsto r_{\text{true}})$.

685 Note that when $\mathbb{E}[r|a, \theta]$ (under the sampled value of θ for some points a) is far from the true value
686 $\mathbb{E}_{P_{\text{true}}}[r|a]$, the chosen action a may be far from optimal, but the information gained by probing
687 action a will improve the belief θ . This amounts to "exploration." When $\mathbb{E}[r|a, \theta]$ is close to the
688 true value except at points a for which $\mathbb{E}[r|a, \theta]$ is low, exploration will be less likely to occur, but
689 the chosen actions a will tend to receive high rewards. This amounts to "exploitation." Roughly
690 speaking, exploration will happen until the context θ is reasonably sure that the unexplored actions
691 are probably not optimal, at which time the sampler will exploit by choosing actions in regions it
692 knows to have high value.

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

702 **4.2 Thompson sampling in Venture**

703
 704 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 $\mathcal{A} \rightarrow \mathbb{R}$ and the covariance K is a computable (symmetric, positive-semidefinite) function $\mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}$. 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

- 714 • $K_{\text{prior}} = K_{\text{prior}, \sigma, \ell}$ is a procedure, with parameters σ, ℓ , to be used as the prior covariance
 715 function: $K_{\text{prior}}(a, a') = \sigma^2 \exp\left(-\frac{(a-a')^2}{2\ell^2}\right)$
- 717
 718 • σ and ℓ are (hyper)parameters for K_{prior}
- 719
 720 • $\mathbf{a}_{\text{past}} = (a_i)_{i=1}^n$ are the previously probed actions
- 722
 723 • $\mathbf{r}_{\text{past}} = (r_i)_{i=1}^n$ are the corresponding rewards
- 724

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

$$\begin{aligned} 729 \quad \mu(\mathbf{a}) &= \mu(\mathbf{a} \mid \mathbf{a}_{\text{past}}, \mathbf{r}_{\text{past}}) \\ 730 \quad &= K_{\text{prior}}(\mathbf{a}, \mathbf{a}_{\text{past}}) K_{\text{prior}}(\mathbf{a}_{\text{past}}, \mathbf{a}_{\text{past}})^{-1} \mathbf{r}_{\text{past}} \\ 732 \quad K(\mathbf{a}, \mathbf{a}) &= K(\mathbf{a}, \mathbf{a} \mid \mathbf{a}_{\text{past}}, \mathbf{r}_{\text{past}}) \\ 733 \quad &= K_{\text{prior}}(\mathbf{a}, \mathbf{a}) - K_{\text{prior}}(\mathbf{a}, \mathbf{a}_{\text{past}}) K_{\text{prior}}(\mathbf{a}_{\text{past}}, \mathbf{a}_{\text{past}})^{-1} K_{\text{prior}}(\mathbf{a}_{\text{past}}, \mathbf{a}). \end{aligned}$$

735
 736 Note that even in this simple example, the context space Θ is not a finite-dimensional parametric
 737 family, since the vectors \mathbf{a}_{past} and \mathbf{r}_{past} grow as more samples are taken. Θ is, however, quite easily
 738 representable as a computational procedure together with parameters and past samples, as we do in
 739 the representation $\theta = (K_{\text{prior}}, \sigma, \ell, \mathbf{a}_{\text{past}}, \mathbf{r}_{\text{past}})$.

741
 742 **4.3 Implementation with gpmem**

744 As a demonstration, we use Thompson sampling to optimize an unknown function $V(x)$ (the value
 745 function) using gpmem. (TODO we should not assume V is deterministic, it would be easy enough
 746 to make it random or have it give noisy samples.) We assume V is made available to Venture as a
 747 black-box. The code for optimizing V is given in Listing 1. For step 3 of Thompson sampling, the
 748 Bayesian update, we not only condition on the new data (the chosen action a and the received reward
 749 r), but also perform inference on the hyperparameters σ, ℓ using a Metropolis–Hastings sampler.
 750 These two inference steps take 1 line of code: 0 lines to condition on the new data (as this is done
 751 automatically by gpmem), and 1 line to call Venture’s built-in MH operator. The results are shown
 752 in Figure 6. We can see from the figure that, roughly speaking, each successive probe point a is
 753 chosen either because the current model V_{emu} thinks it will have a high reward, or because the value
 754 of $V_{\text{emu}}(a)$ has high uncertainty. In the latter case, probing at a decreases this uncertainty and, due to
 755 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.

756

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

```

766 assume log_sf = tag('hyper, log(uniform_continuous(0, 10)))
767 assume log_l = tag('hyper, log(uniform_continuous(0, 10)))
768 assume se = make_squaredexp(log_sf, log_l)
769 assume blackbox_f = get_bayesopt_blackbox()
770 assume (f_compute, f_emu) = gpmem(blackbox_f, se)

771 define get_uniform_candidate = proc(prev_xs) {
772   uniform_continuous(-20, 20)
773 }

774 define mc_argmax = proc(func, prev_xs) {
775   // Monte Carlo estimator for the argmax of func.
776   run(do(
777     candidate_xs <- mapv(proc(i) {get_uniform_candidate(prev_xs)},
778                             linspace(0, 19, 20)),
779     candidate_ys <- mapv(func, candidate_xs),
780     lookup(candidate_xs, argmax_of_array(candidate_ys))))
781 }

782 define emulator_point_sample = proc(x) {
783   run(sample(lookup(
784     f_emu(array(x)),
785     0)))
786 }

787 infer repeat(15, do(pass,
788   // Phase 2: Call f_compute on the next probe point
789   predict f_compute(
790     mc_argmax(emulator_point_sample, '_)),
791   // Phase 1: Hyperparameter inference
792   mh('hyper, one, 50)))

```

793

794

795

796

797

798

799

800

801

802

803

804

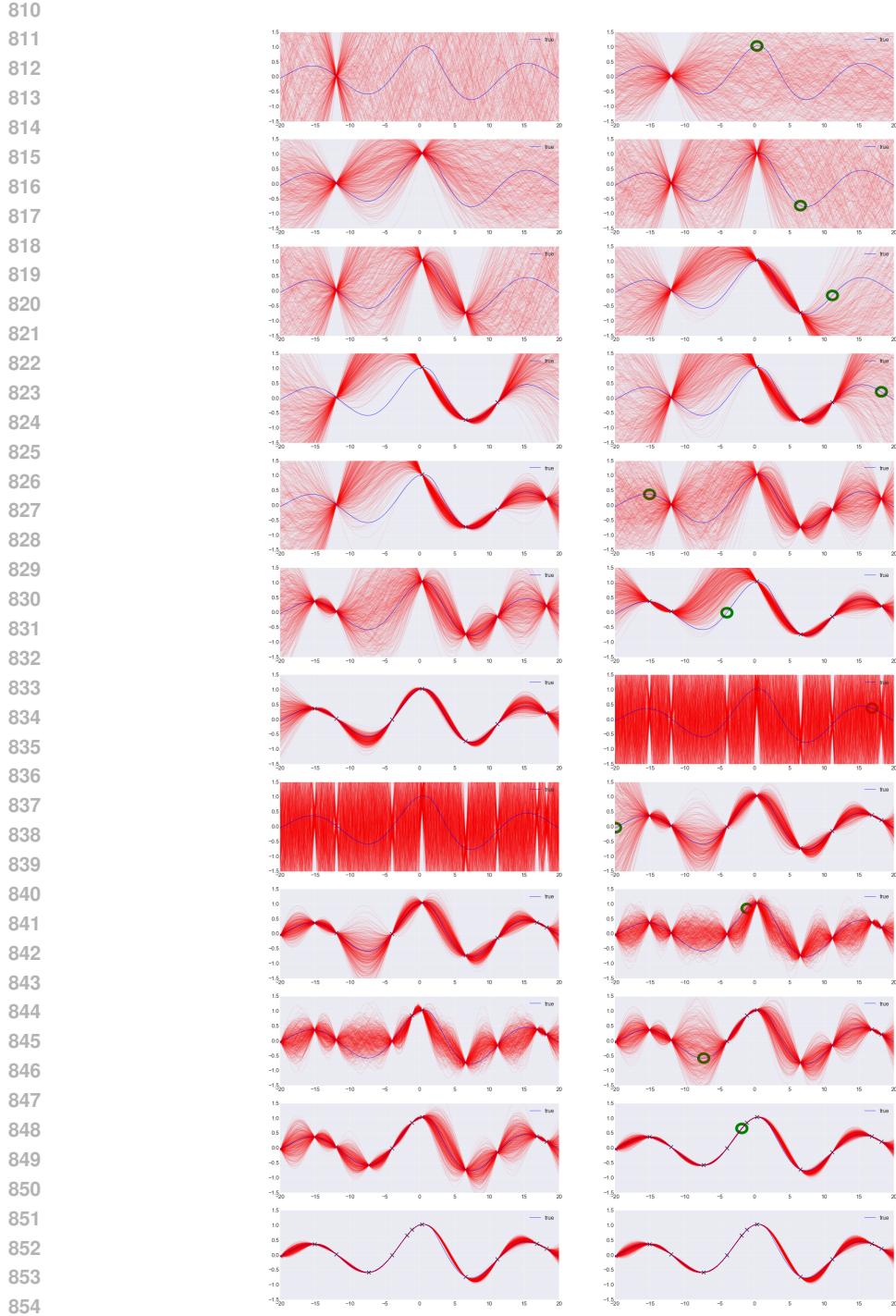
805

806

807

808

809



855
856
857
858
859
860
861
862
863
Figure 6: 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.

864 **5 Conclusion**
 865

866 We have shown Venture GPs. We have introduced novel stochastic processes for a probabilistic
 867 programming language. We showed how flexible non-parametric models can be treated in Venture
 868 in only a few lines of code. We evaluated our contribution on a range of hard problems for state-of-
 869 the-art Bayesian non-parametrics. Venture GPs showed competitive performance in all of them.
 870

871 **Appendix**
 872

873 **A Covariance Functions**
 874

875 SE and WN are defined in the text above, for completeness we will introduce the covariance:
 876

$$k_{LIN}(x, x') = \theta(x x') \quad (23)$$

$$k_{PER}(x, x') = \theta \exp\left(\frac{2 \sin^2(\pi(x - x')/p)}{\ell^2}\right) \quad (24)$$

$$k_{RQ}(x, x') = \theta \left(1 + \frac{(x - x')^2}{2\alpha\ell^2}\right)^{-\alpha} \quad (25)$$

884 **B Covariance Simplification**
 885

SE × SE	→ SE
{SE, PER, C, WN} × WN	→ WN
LIN + LIN	→ LIN
{SE, PER, C, WN, LIN} × C	→ {SE, PER, C, WN, LIN}

891 Rule 1 is derived as follows:

$$\begin{aligned} \sigma_c^2 \exp(-\frac{(x - x')^2}{2\ell_c^2}) &= \sigma_a^2 \exp(-\frac{(x - x')^2}{2\ell_a^2}) \times \sigma_b^2 \exp(-\frac{(x - x')^2}{2\ell_b^2}) \\ &= \sigma_c^2 \exp(-\frac{(x - x')^2}{2\ell_a^2}) \times \exp(-\frac{(x - x')^2}{2\ell_b^2}) \\ &= \sigma_c^2 \exp\left(-\frac{(x - x')^2}{2\ell_a^2} - \frac{(x - x')^2}{2\ell_b^2}\right) \\ &= \sigma_c^2 \exp\left(-\frac{(x - x')^2}{2\ell_c^2}\right) \end{aligned} \quad (26)$$

902 Rule 3 is derived as follows:

$$\theta_c(x \times x') = \theta_a(x \times x') + \theta_b(x \times x') \quad (27)$$

903 For stationary kernels that only depend on the lag vector between x and x' it holds that multiplying
 904 such a kernel with a WN kernel we get another WN kernel. Take for example the SE kernel:
 905

$$\sigma_a^2 \exp\left(-\frac{(x - x')^2}{2\ell_c^2}\right) \times \sigma_b \delta_{x,x'} = \sigma_a \sigma_b \delta_{x,x'} \quad (28)$$

906 Multiplying any kernel with a constant obviously changes only the scale parameter of a kernel.
 907

911
 912
 913
 914
 915
 916
 917

918 **References**
919

- 920 Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to MCMC for
921 machine learning. *Machine learning*, 50(1-2):5–43.
- 922 Benavoli, A. and Mangili, F. (2015). Gaussian processes for bayesian hypothesis tests on regression
923 functions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence
924 and Statistics*, pages 74–82.
- 925 Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1997). *Time series analysis: forecasting and
926 control*.
- 927 Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J., and Ghahramani, Z. (2013). Structure
928 discovery in nonparametric regression through compositional kernel search. In *Proceedings of
929 the 30th International Conference on Machine Learning (ICML-13)*, pages 1166–1174.
- 930 Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*,
931 521(7553):452–459.
- 932 Goodman, N. D .and Mansinghka, V. K., Roy, D., Bonawitz, K., and Tenenbaum, J. (2008). Church:
933 A language for generative models. In *Proceedings of the 24th Conference on Uncertainty in
934 Artificial Intelligence, UAI 2008*, pages 220–229.
- 935 Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J., and Ghahramani, Z. (2014). Automatic
936 construction and natural-language description of nonparametric regression models. In *Twenty-
937 Eighth AAAI Conference on Artificial Intelligence*.
- 938 Mansinghka, V. K., Selsam, D., and Perov, Y. (2014). Venture: a higher-order probabilistic pro-
939 gramming platform with programmable inference. *arXiv preprint arXiv:1404.0099*.
- 940 McAllester, D., Milch, B., and Goodman, N. D. (2008). Random-world semantics and syntactic
941 independence for expressive languages. Technical report.
- 942 Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation
943 of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–
944 1092.
- 945 Neal, R. M. (1997). Monte carlo implementation of gaussian process models for bayesian regression
946 and classification. *arXiv preprint physics/9701026*.
- 947 Poole, D. (1993). Probabilistic horn abduction and bayesian networks. *Artificial Intelligence*,
948 64(1):81–129.
- 949 Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning (Adap-
950 tive Computation and Machine Learning)*. The MIT Press.
- 951 Sato, T. (1995). A statistical learning method for logic programs with distribution semantics. In *In
952 Proceedings of the 12th International Conference on Logic Programming*. Citeseer.
- 953 Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine
954 learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959.
- 955 Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view
956 of the evidence of two samples. *Biometrika*, pages 285–294.
- 957
958
959
960
961
962
963
964
965
966
967
968
969
970
971