# Probabilistic Programming with Gaussian Process Memoization

**Ulrich Schaechtle** ulrich.schaechtle@rhul.ac.uk
*Department of Computer Science*
*Royal Holloway, University of London*

**Ben Zinberg** bzinberg@mit.edu
*Computer Science and Artificial Intelligence Laboratory*
*Massachusetts Institute of Technology*

**Alexey Radul** axofch@gmail.com
*Computer Science and Artificial Intelligence Laboratory*
*Massachusetts Institute of Technology*

**Kostas Stathis** kostas.stathis@rhul.ac.uk
*Department of Computer Science*
*Royal Holloway, University of London*

**Vikash K. Mansinghka** vkm@mit.edu
*Computer Science and Artificial Intelligence Laboratory*
*Massachusetts Institute of Technology*

**Editor:** N.A.

## 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

Gaussian Processes (GP) are widely used tools in statistics (Barry, 1986), machine learning (Neal, 1995; Williams and Barber, 1998; Kuss and Rasmussen, 2005; Rasmussen and Williams, 2006; Damianou and Lawrence, 2013), robotics (Ferris et al., 2006), computer vision (Kemmler et al., 2013), and scientific computation (Kennedy and O'Hagan, 2001; Schneider et al., 2008; Kwan et al., 2013). They are also central to probabilistic numerics,

an emerging effort to develop more computationally efficient numerical procedures, and to Bayesian optimization, a family of meta-optimization techniques that are widely used to tune parameters for deep learning algorithms (Snoek et al., 2012; Gelbart et al., 2014). They have even seen use in artificial intelligence; for example, they provide the key technology behind a project that produces qualitative natural language descriptions of time series (Duvenaud et al., 2013; Lloyd et al., 2014).

This paper describes Gaussian process memoization, a technique for integrating GPs into a probabilistic programming language, and demonstrates its utility by re-implementing and extending state-of-the-art applications of the GP. Memoization is a classic programming technique in which a procedure is augmented with an input-output cache that is checked each time before the function is invoked. This prevents unnecessary recomputation, potentially saving time at the cost of increased storage requirements. Gaussian process memoization, implemented by the `gpmem()` procedure, generalizes this idea to include a statistical emulator that uses previously computed values as data in a statistical model that can accurately forecast probable outputs. The covariance function for the Gaussian process is also allowed to be an arbitrary probabilistic program.

This paper presents three applications of gpmem: (i) a replication of results Neal (1997) on outlier rejection via hyper-parameter inference; (ii) a fully Bayesian extension to the Automated Statistician project; and (iii) an implementation of Bayesian optimization via Thompson sampling. The first application can in principle be replicated in several other probabilistic languages embedding the proposal that is described in this paper. The remaining two applications rely on distinctive capabilities of Venture (Mansinghka et al., 2014): support for fully Bayesian structure learning and language constructs for inference programming. All applications share a single 50-line Python library and require fewer than 20 lines of probabilistic code each.

## 2. Background on Gaussian Processes

GPs are a Bayesian method for regression. We consider the regression input to be real-valued scalars $x_i$ and the regression output as the value of a function $f$ at $x_i$. The complete training data will be denoted by column vectors $\mathbf{x}$ and $\mathbf{f}$. Unseen test data is denoted with $\hat{\mathbf{x}}$ and $\hat{\mathbf{f}}$. GPs present a non-parametric way to express prior knowledge on the space of all possible functions $f$ modeling a regression relationship. Formally, a GP is an infinite-dimensional extension of the multivariate Gaussian distribution.

The collection of random variables $\{f(x_i)\}$ (indexed by $i$) represents the values of the function $f$ at each location $x_i$. We drop the index in the following for readability. We write $f \sim \mathrm{GP}(m, k)$, where $m$ is the *mean function* and $k$ is the *covariance function* or *kernel*. That is, $m(x)$ is the prior mean of the random variable $f(x)$, and $k(x, x')$ is the prior covariance of the random variables $f(x)$ and $f(x')$. Throughout, we write $\mathbf{K}(\mathbf{x}, \mathbf{x}')$ for the prior covariance matrix determined by $\mathbf{x}$ and $\mathbf{x}'$, that is, the covariance between the random vectors $\{f(x)\}_{x \in \mathbf{x}}$ and $\{f(x')\}_{x' \in \mathbf{x}'}$. We differentiate three different situations, of how $f$ can be generated with a GP:

1. $\mathbf{f}_*$ - we have not seen any training data yet. The GP samples at any input vector $\mathbf{x}_*$ from the prior with $\mathbf{f}_* \sim \mathcal{N}\big(0, K(\mathbf{x}_*, \mathbf{x}_*)\big)$;

2. $\mathbf{f}$ - the observed values for $\mathbf{f} := f(\mathbf{x})$. This is the target value of the value pairs supplied in the training data; and

3. $\hat{\mathbf{f}}$ - the predictive posterior distribution of test output $\hat{\mathbf{f}} := f(\hat{\mathbf{x}})$ conditioned on training data $\mathbf{f} := f(\mathbf{x})$.

We now compute the predictive posterior distribution of test output $\hat{\mathbf{f}} := f(\hat{\mathbf{x}})$ conditioned on training data $\mathbf{f} := f(\mathbf{x})$. (Here $\mathbf{x}$ and $\hat{\mathbf{x}}$ are known constant vectors, and we are conditioning on an observed value of $\mathbf{f}$.) To simplify the calculation, we will assume the prior mean $m$ is identically zero; once the derivation is done, this assumption can be easily relaxed via translation.

The predictive posterior can be computed by first forming the joint density when both training and test data are treated as randomly chosen from the prior, then fixing the value of $\mathbf{f}$ to a constant. To start, let

$$\Sigma := \begin{bmatrix} K(\mathbf{x},\mathbf{x}) & K(\mathbf{x},\hat{\mathbf{x}}) \\ K(\hat{\mathbf{x}},\mathbf{x}) & K(\hat{\mathbf{x}},\hat{\mathbf{x}}) \end{bmatrix} \text{ and } \Sigma^{-1} =: \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}.$$

We then have

$$P(\mathbf{f},\hat{\mathbf{f}}) \propto \exp\left\{ -\frac{1}{2} \begin{bmatrix} \mathbf{f}^\top & \hat{\mathbf{f}}^\top \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ \hat{\mathbf{f}} \end{bmatrix} \right\}.$$

Treating $\mathbf{f}$ as a fixed constant, we obtain

$$P\left(\hat{\mathbf{f}} \;\middle|\; \mathbf{f}\right) \propto P(\mathbf{f},\hat{\mathbf{f}}) \propto \exp\left\{ -\frac{1}{2}\hat{\mathbf{f}}^\top M_{22}\hat{\mathbf{f}} - \mathbf{h}^\top \hat{\mathbf{f}} \right\},$$

where $\mathbf{h} = M_{21}\mathbf{f}$ is a constant vector. Thus $P(\hat{\mathbf{f}}|\mathbf{f})$ is Gaussian,

$$P\left(\hat{\mathbf{f}} \;\middle|\; \mathbf{f}\right) \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{K}}), \tag{1}$$

with covariance matrix $\hat{\mathbf{K}} = M_{22}^{-1}$. To find its mean $\hat{\boldsymbol{\mu}}$, we note that $P_{\hat{\mathbf{f}}|\mathbf{f}}(\hat{\mathbf{f}} + \hat{\boldsymbol{\mu}})$ is Gaussian with the same covariance as $P(\hat{\mathbf{f}}|\mathbf{f})$, but its exponent has no linear term:

$$\begin{aligned} P_{\hat{\mathbf{f}}|\mathbf{f}}\left(\hat{\mathbf{f}} + \hat{\boldsymbol{\mu}} \;\middle|\; \mathbf{f}\right) &\propto \exp\left\{ -\frac{1}{2}(\hat{\mathbf{f}} + \hat{\boldsymbol{\mu}})^\top M_{22}(\hat{\mathbf{f}} + \hat{\boldsymbol{\mu}}) - \mathbf{h}^\top(\hat{\mathbf{f}} + \hat{\boldsymbol{\mu}}) \right\} \\ &\propto \exp\left\{ -\frac{1}{2}\hat{\mathbf{f}}^\top M_{22}\hat{\mathbf{f}} - \underbrace{(\mathbf{h} + M_{22}\hat{\boldsymbol{\mu}})^\top}_{\text{must be 0}} \hat{\mathbf{f}} \right\}. \end{aligned}$$

Thus $\mathbf{h} = -M_{22}\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}} = -M_{22}^{-1}\mathbf{h} = -M_{22}^{-1}M_{21}\mathbf{f}$.

The partioned inverse equations (Barnett and Barnett, 1979 following MacKay, 1998) give

$$\begin{aligned} M_{22}^{-1} &= K(\hat{\mathbf{x}},\hat{\mathbf{x}}) - K(\hat{\mathbf{x}},\mathbf{x})K(\mathbf{x},\mathbf{x})^{-1}K(\mathbf{x},\hat{\mathbf{x}}), \\ M_{21} &= -M_{22}K(\hat{\mathbf{x}},\mathbf{x})K(\mathbf{x},\mathbf{x})^{-1}. \end{aligned}$$

Substituting these in the above gives

$$\hat{\mathbf{K}} = K(\hat{\mathbf{x}}, \hat{\mathbf{x}}) - K(\hat{\mathbf{x}}, \mathbf{x})K(\mathbf{x}, \mathbf{x})^{-1}K(\mathbf{x}, \hat{\mathbf{x}}), \tag{2}$$

$$\hat{\boldsymbol{\mu}} = K(\hat{\mathbf{x}}, \mathbf{x})K(\mathbf{x}, \mathbf{x})^{-1}\mathbf{f}. \tag{3}$$

Together, $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{K}}$ determine the computation of the predictive posterior with unseen input data (1).

Often one assumes the observed regression output is noisily measured, that is, one only sees the values of $\mathbf{y}_{\text{noisy}} = \mathbf{f} + \mathbf{w}$ where $\mathbf{w}$ is Gaussian white noise with variance $\sigma^2_{\text{noise}}$. This noise term can be absorbed into the covariance matrix $\mathbf{K}(\mathbf{x}, \mathbf{x})$ which in the following, we will write as $\mathbf{K}$ for readability. The log-likelihood of a GP can then be written as:

$$\log P(\mathbf{f} \mid \mathbf{x}) = -\frac{1}{2}\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi \tag{4}$$

where $n$ is the number of data points. Both log-likelihood and predictive posterior can be computed efficiently using a Stochastic Process (SP) in Venture (Mansinghka et al., 2014) with an algorithm that resorts to Cholesky factorization(Rasmussen and Williams, 2006, chap. 2). We write the Cholesky factorization as $\mathbf{L} \coloneqq \text{chol}(\mathbf{K})$ when :

$$\mathbf{K} = LL^\top \tag{5}$$

where L is a lower triangular matrix. This allows us to compute the inverse of a covariance matrix as

$$\mathbf{K}^{-1} = (\mathbf{L}^{-1})^\top(\mathbf{L}^{-1}) \tag{6}$$

and its determinant as

$$det(\mathbf{K}) = det(\mathbf{L})^2 \tag{7}$$

We compute (4) as

$$\log(P(\mathbf{f} \mid \mathbf{x}) \coloneqq -\frac{1}{2}\mathbf{f}^\top\boldsymbol{\alpha} - \sum_i \log \mathbf{L}_{ii} - \frac{n}{2}\log 2\pi \tag{8}$$

where

$$\mathbf{L} \coloneqq \text{chol}(\mathbf{K}) \tag{9}$$

and

$$\boldsymbol{\alpha} \coloneqq \mathbf{L}^\top \backslash (\mathbf{L} \backslash \mathbf{f}). \tag{10}$$

This results in a computational complexity for sampling in the number of data points of $O(n^3/6)$ for (9) an $O(n^2/2)$ for (10).

Above, we defined the GP prior as $\mathbf{f}_* \sim \mathcal{N}\big(0, K(\mathbf{x}_*, \mathbf{x}_*)\big)$. We see that this prior is fully determined by its covariance function.

### 2.1 Covariance Functions

The covariance function (or kernel) of a GP governs high-level properties of the observed data such as smoothness or linearity. A linear covariance can be written as:

$$\mathrm{LIN} = \sigma_1^2(xx'). \tag{11}$$

We can also express periodicity:

$$\mathrm{PER} = \sigma_2^2 \exp\left(\frac{2\sin^2(\pi(x-x')/p}{\ell^2}\right). \tag{12}$$

By changing these properties we get completely different prior behavior for sampling $\mathbf{f}_*$ from a GP with a linear kernel

$$\mathbf{f}_* \sim \mathcal{N}\big(0, \mathrm{LIN}(\mathbf{x}, \mathbf{x})\big)$$

as compared to sampling from the prior predictive with a periodic kernel (as depicted in Fig. 1 (c) and (d))

$$\mathbf{f}_* \sim \mathcal{N}\big(0, \mathrm{PER}(\mathbf{x}, \mathbf{x})\big).$$

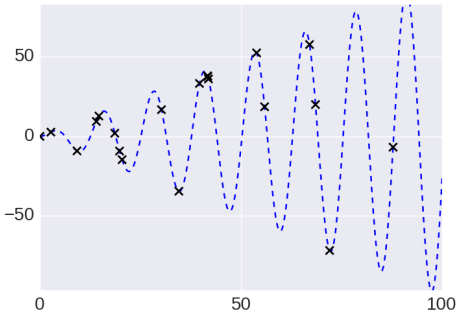
These high-level properties are compositional via addition and multiplication of different covariance functions. That means that we can also combine these properties. By using multiplication of kernels we can model a local interaction of two components, for example

$$\mathrm{LIN} \times \mathrm{PER} = \sigma_1^2(xx')\,\sigma_2^2 \exp\left(\frac{2\sin^2(\pi(x-x')/p}{\ell^2}\right) \tag{13}$$

This results in a combination of the higher level properties of linearity and periodicity. In Fig 1 (e) we depict samples for $\mathbf{f}_*$ that are periodic with linearly increasing amplitude. We consider this a local interaction because the actual interaction depends on the similarity of two data points. An addition of covariance functions models a global interaction, that is an interaction of two high-level components that is qualitatively not dependent on the input space. An example for this a periodic function with a linear trend.

Covariance functions come with free parameters that we call hyper-parameters which we will refer to as $\boldsymbol{\theta}$. For each kernel type, each $\boldsymbol{\theta}$ is different, that is, in (11) we have $\theta = \{\sigma_1\}$, in (12) we have $\boldsymbol{\theta} = \{\sigma_2, p, \ell\}$ and in (13) we have $\boldsymbol{\theta} = \{\sigma_1, \sigma_2, p, \ell\}$. Adjusting these hyper-parameters changes lower level qualitative attributes such as length scales ($\ell$) while preserving the higher level qualitative properties of the distribution such as linearity. We write the parameterized covariance function as $\mathbf{K}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x})$ and the covariance matrix determined by a parameterized covariance function as $\mathbf{K}_{\boldsymbol{\theta}}$.

When we observe the data, that is we condition on the input-output pairs $\{\mathbf{x}, \mathbf{f}\}$ we can sample from the predictive posterior $P(\hat{\mathbf{f}} \mid \mathbf{f})$ given $\mathbf{K}_{\boldsymbol{\theta}}$ and respectively $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{K}}_{\boldsymbol{\theta}}$. If we choose suitable parameters, for example by performing inference, we can capture the underlying dynamics of the data well (see Fig. 1 (f-h)) while sampling $\hat{\mathbf{f}}$. Note that goodness of fit is not only limited to the parameters. A too simple qualitative structure implies unsuitable behaviour, as for example in (Fig. 1 (g)) where additional recurring spikes are introduced to account for the changing amplitude of the true function that generated the data.




(a) Raw Data

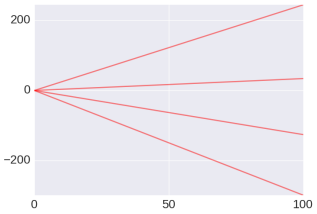
$$\text{LIN} = \sigma_1^2(xx')$$

$$\text{PER} = \sigma_2^2 \exp\left(\frac{2\sin^2(\pi(x-x')/p}{\ell^2}\right)$$

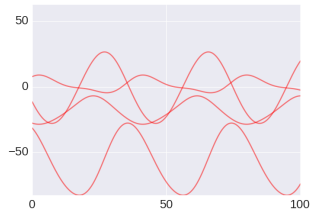
$$\text{LIN} \times \text{PER} = \sigma_1^2(xx')\,\sigma_2^2 \exp\left(\frac{2\sin^2(\pi(x-x')/p}{\ell^2}\right)$$

(b) Kernels

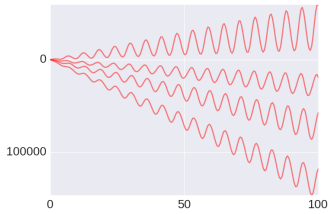
Prior predictive:




(c) LIN

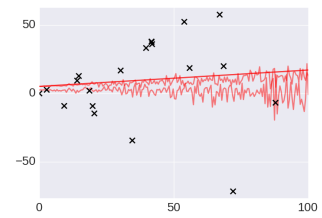



(d) PER

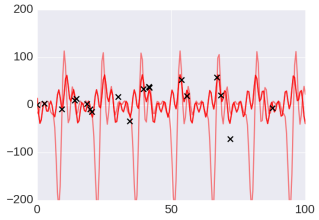



(e) LIN × PER

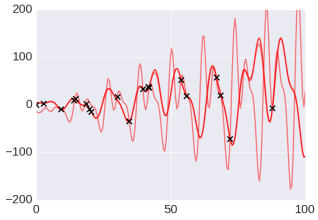
Posterior predictive:




(f) LIN




(g) PER




(h) LIN × PER

Figure 1: We depict kernel composition. (a) shows raw data (black) generated with a sine function with linearly growing amplitude (blue). This data is used for all the plots (c-h). (b) shows the linear and the periodic base kernel as well as a composition of both. The multiplication of the two kernels indicates local interaction. The local interaction we account for in this case is the growing amplitude (a). For each column (c-h) $\boldsymbol{\theta}$ is different.(c-e) show samples from the prior prior predictive $\mathbf{f}_*$ where random parameters are used, that is, we sample before any data points are observed. (f-h) show samples from the predictive posterior $\hat{\mathbf{f}}$, after the data has been observed.

## 3. Gaussian Process Memoization in Venture

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 re-computation. To transfer this idea to probabilistic programming, we now introduce a language construct called a *statistical memoizer*. Suppose we have a function `f` which can be evaluated but we wish to learn about the behavior of `f` using as few evaluations as possible. The statistical memoizer, which here we give the name `gpmem`, was motivated by this purpose. It produces two outputs:

$$\mathtt{f} \xrightarrow{\mathtt{gpmem}} (\mathtt{f}_{\text{compute}}, \mathtt{f}_{\text{emu}}).$$

The function $\mathtt{f}_{compute}$ calls `f` and stores the output in a memo table, just as traditional memoization does. The function $\mathtt{f}_{\text{emu}}$ is an online statistical emulator which uses the memo table as its training data. A fully Bayesian emulator, modelling the true function `f` as a random function $f \sim P(f)$, would satisfy

$$(\mathtt{f}_{\text{emu}}\ \mathtt{x}_1\ \ldots\ \mathtt{x}_k) \sim P\left(f(\mathtt{x}_1), \ldots, f(\mathtt{x}_k) \mid f(\mathtt{x}) = (\mathtt{f}\ \mathtt{x}) \text{ for each } \mathtt{x} \text{ in memo table}\right).$$

Different implementations of the statistical memoizer can have different prior distributions $P(f)$; in this paper, we deploy a GP prior (implemented as `gpmem` below). Note that we require the ability to sample $\mathtt{f}_{\text{emu}}$ jointly at multiple inputs because the values of $f(\mathtt{x}_1), \ldots, f(\mathtt{x}_k)$ will in general be dependent.

We explain how `gpmem`, the statistical memoizer with GP-prior, works using a simple tutorial (Fig. 2). The top panel (Fig. 2, (a)) of this figure sketches the schematic of `gpmem`. `f` is the external process that we memoize. It can be evaluated using resources that potentially come from outside of Venture. We feed this function into `gpmem` alongside a parameterised kernel $\mathbf{K}_{\boldsymbol{\theta}}$. In this example, we make the qualitative assumption of $f$ being smooth, and define $\mathbf{K}_{\boldsymbol{\theta}}$ to be a squared-exponential covariance function:

$$\mathbf{K}_{\boldsymbol{\theta}} = \text{SE} = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right).$$

The hyper-parameters $\boldsymbol{\theta}$ for this kernel are sampled from a prior distribution which is depicted in the top right box. Note that we annotate $\boldsymbol{\theta} = \{\mathtt{sf}, \mathtt{l}\}$ for subsequent inference as belonging to (i) the scope "hyper-parameter" and (ii) blocks 0 and 1 respectively.

`gpmem` implements a memoization table, where all previously computed function evaluations are stored. We also initialize a GP-prior that will serve as our statistical emulator:

$$P(f_{emu}(x) \mid \mathbf{x}_{past}, \mathbf{f}) \sim \mathcal{N}(\mu(\mathbf{x}), \mathbf{K}_\theta(\mathbf{x}, \mathbf{x}))$$

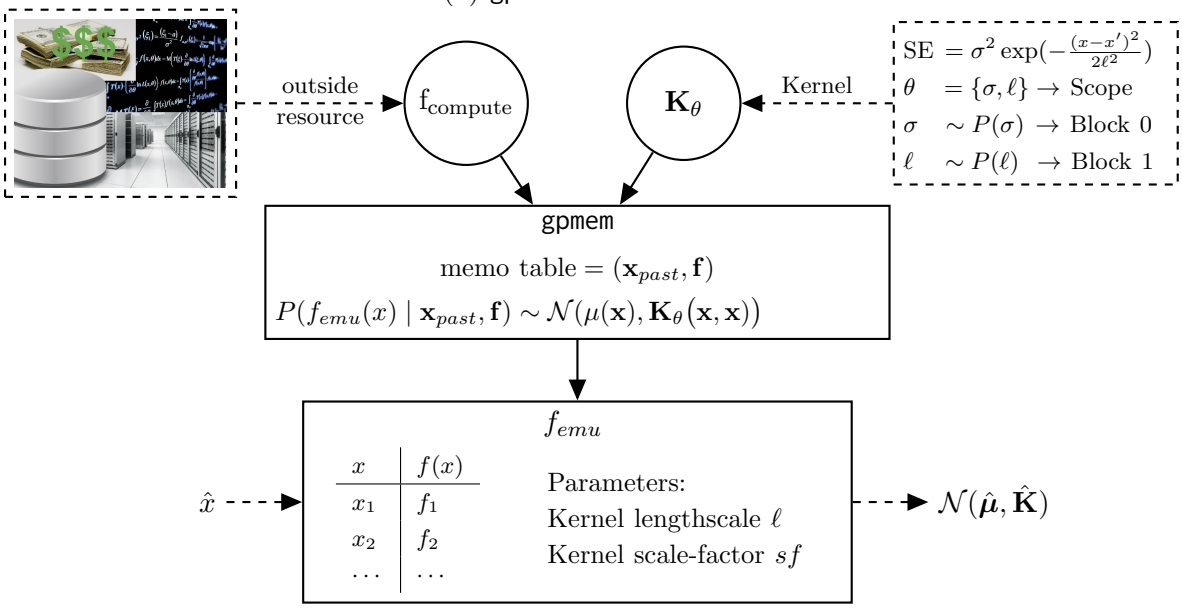
where

$$P(f_{emu}(x) \mid \mathbf{x}_{past}, \mathbf{f}) = \hat{\mathbf{f}}$$

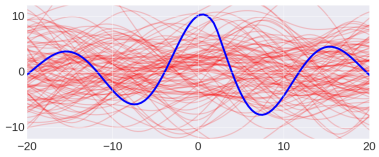
under the traditional GP perspective[1]. All value pairs stored in the memoization table (memo table $= (\mathbf{x}_{past}, \mathbf{f})$) are incorporated as observations of the GP. The emulator allows us to make probabilistic predictions on function evaluations, estimating $P(\hat{\mathbf{f}} \mid \mathbf{f})$ with

1. Note that we write $\mathbf{f}_*$ instead of $\hat{\mathbf{f}}$ if $\mathbf{x}_{past} = \{\}$ and $\mathbf{f} = \{\}$.

(a) gpmem: Schematic

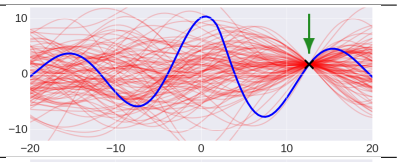



(b)
```
define f = proc(x) {
    exp(-0.1*abs(x-2)) * 10 * cos(0.4*x) + 0.2}
assume (f_compute, f_emu) = gpmem(f, K)
sample f_emu(array(-20, ···, 20))
```



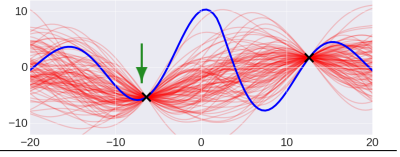
(c)
```
predict f_compute(12.6)

sample f_emu(array(-20, ···, 20))
```

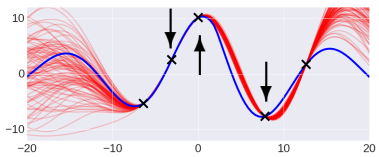


(d)
```
predict f_compute(-6.4)

sample f_emu(array(-20, ···, 20))
```



(e)
```
observe f_emu(-3.1) =  2.60
observe f_emu(7.8)  = -7.60
observe f_emu(0.0)  = 10.19
sample f_emu(array(-20, ···, 20))
```



(f)
```
infer mh("hyper_parameter", one, 50)

sample f_emu(array(-20, ···, 20))
```

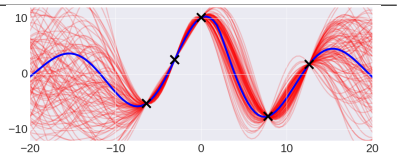


Figure 2: gpmem tutorial. The top shows a schematic of gpmem. f_compute probes an outside resource. This can be expensive (top left). Every probe is memoized and improves the emulator. Below the schematic we see the evolution of gpmem's state of believe of the world given certain Venture directives. On the right, we depict the true function (blue), samples from the emulator (red) and incorporated observations (black).

$\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{K}})$. We simply feed the regression input into the emulator and output a predictive posterior Gaussian distribution determined by the GP and the memoization table.

We can either define the function f that serves as as input for `gpmem` natively in Venture (as shown in the Fig. 2 (b)) or we interleave Venture with foreign code. This can be useful when `f` is computed with the help of outside resources. Before making any observations or calls to `f` we can sample from the prior at the inputs from -20 to 20 using the emulator:

```
assume (f_compute, f_emu) = gpmem(f, K))

sample f_emu(array(-20, ..., 20))
```

where the second line corresponds to:

$$\mathbf{f}_* \sim \mathcal{N}\left(0, K\left(\begin{bmatrix}-20\\ \cdots \\ 20\end{bmatrix}, \begin{bmatrix}-20\\ \cdots \\ 20\end{bmatrix}\right)\right).$$

In Fig. 2 (c), we probe the external function `f` at point 12.6 and memoize it's result by calling

```
predict f_compute (12.6).
```

When we subsequently sample from the emulator, that is compute the $\hat{\mathbf{f}}$ at the input $\hat{\mathbf{x}} = \left[-20, \cdots, 20\right]^\top$. We see how the posterior which shifts from uncertainty to near certainty close to the input 12.6.

We can repeat the process at a different point (probing point -6.4 in Fig. 2 (d)) to see that we gain certainty about another part of the curve.

We can add information to $\mathtt{f}_{\text{emu}}$ about presumable value pairs of `f` without calling $\mathtt{f}_{\text{compute}}$ (Fig. 2 (e)). If a friend tells us the value of `f` we can call observe to store this information in the incorporated observations for $\mathtt{f}_{\text{emu}}$ only:

```
observe f_emu( -3.1) = 2.60.
```

We have this value pair now available for the computation $\hat{\mathbf{f}}$. For sampling with the emulator, the effect is the same as calling predict with the $\mathtt{f}_{\text{compute}}$. However, we can imagine at least one scenario where such as distinction in the treatment of observations is beneficial. Let us say we do not only have the real function available but also a domain expert with knowledge about this function. This expert could tell us what the value is at a given input. Potentially, the value provided by the expert could disagree with the value computed with `f` for example due to different levels of observation noise.

Finally, we can update our posterior by inferring the posterior over hyper-parameter values $\boldsymbol{\theta}$. We can take 50 Metropolis-Hastings (MH) steps by calling that where we take an MH steps for either of the two blocks for the scope "hyper-parameter":

```
infer mh( 'hyper-parameter, one, 50).
```

The newly inferred hyper-parameters allow us now adequately reflect uncertainty about the curve given all incorporated observations (compare Fig. 2, bottom panel (f) on the right with the samples before inference, one panel above (e)).

## 4. Applications

This paper illustrates the flexibility of `gpmem` by showing how it can concisely encode three different applications of GPs. The first is a standard example from hierarchical Bayesian statistics, where Bayesian inference over a hierarchical hyper-prior is used to provide a curve-fitting methodology that is robust to outliers. The second is a structure learning application from probabilistic artificial intelligence, where GPs are used to discover qualitative structure in time series data. The third is a reinforcement learning application, where GPs are used as part of a Thompson sampling formulation of Bayesian optimization for general real-valued objective functions with real inputs.

### 4.1 Nonlinear regression in the presence of outliers

We can apply `gpmem` for regression in a hierarchical Bayesian setting (Fig. 3). In a Bayesian treatment of hyper-parameter learning for GPs, we can write the posterior probability of the hyper-parameters of a GP (Fig. 3, (a)) given covariance function $\mathbf{K}$ as:

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

where $\mathbf{D} = \{\mathbf{x}, \mathbf{f}\}$ is a training data set. Since we can apply `gpmem` to any process or procedure, it can be used in situations where only a data set is available via a look-up function `f_look_up`. In fact, we demonstrate `gpmem`'s application to regression using an example where the data was generated by a function which is not available (Fig. 3 (b)). This function, $f_{\text{true}}$, is taken from a paper on the treatment of outliers with hierarchical Bayesian hyper-priors for GPs (Neal, 1997):

$$f_{\text{true}}(x) = 0.3 + 0.4x + 0.5\sin(2.7x) + \frac{1.1}{(1+x^2)} + \eta \quad with \;\; \eta \sim \mathcal{N}(0, \sigma). \tag{15}$$

We synthetically generate outliers by setting $\sigma = 0.1$ in 95% of the cases and to $\sigma = 1$ in the remaining cases. Instead of accessing the $f_{\text{true}}$ directly, we are accessing the `data` in form of a a two dimensional `array` with `f_look_up`.

We parameterize $\mathbf{K}_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} = \{sf, \ell, \sigma\}$. For these hyper-parameters, Neals work suggests a hierarchical system for for hyper-parameterization. Here, we draw hyper-parameters from $\Gamma$ distributions:

$$\ell^{(t)} \sim \Gamma(\alpha_1, \beta_1), \;\; \sigma^{(t)} \sim \Gamma(\alpha_2, \beta_2) \tag{16}$$

and in turn sample the $\alpha$ and $\beta$ from $\Gamma$ distributions as well:

$$\alpha_1^{(t)} \sim \Gamma(\alpha_\alpha^1, \beta_\alpha^1), \;\; \alpha_2^{(t)} \sim \Gamma(\alpha_\alpha^2, \beta_\alpha^2), \cdots \tag{17}$$

We model this in Venture as illustrated in Fig. 3 (c), using the build-in SP `gamma`. Note that we omit the prior distributions of (17) due limited space in the tutorial.

(a) $P(\boldsymbol{\theta} = \{\ell, sf, \sigma\} \mid \mathbf{D}, \mathbf{K})$


(b)

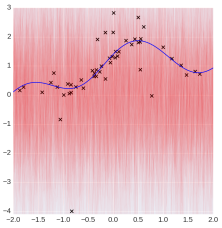
```
// Data and look-up function
define data = array(array(-1.87, 0.13), ..., array(1.67, 0.81))
assume f_look_up = proc(index) {lookup(data, index)}
```

(c)

```
assume sf = tag("hyper", 0, gamma(alpha_sf, beta_sf)))
assume l = tag("hyper", 1, gamma(alpha_l, beta_l)))
assume sigma = tag("hyper", 2, uniform_continuous(0, 2))
```

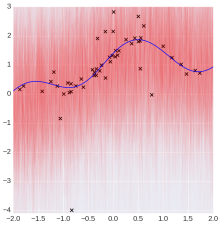
(d)

```
// The covariance function
assume se = make_squaredexp(sf, l)
assume wn = make_whitenoise(sigma)
assume composite_covariance = add_funcs(se, wn)
// Create a prober and emulator using gpmem
assume (f_compute, f_emu) =
    gpmem(f_look_up, composite_covariance)
sample f_emu(array(-2, ···, 2))
```



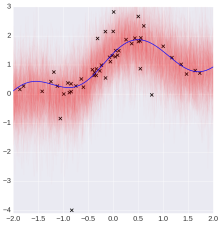
(e)

```
// Observe all data points
for n ... N
    observe f_emu(first(lookup(data,n))) =
        second(lookup(data,n))
// Or: probe all data points
for n ... N
    predict f_compute(first(lookup(data,n)))
sample f_emu(array(-2, ···, 2))
```



(f)

```
// Metropolis-Hastings
infer repeat(100, do(
    mh("hyperhyper", one, 2),
    mh("hyper",      one, 1)))
sample f_emu(array(-2, ···, 2))
```



(g)

```
// Optimization
infer map("hyper", all, 0.01, 15)

sample f_emu(array(-2, ···, 2))
```

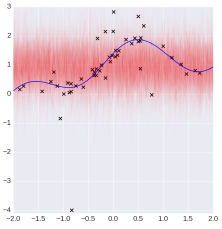


Figure 3: Regression with outliers and hierarchical prior structure

In Fig. 3, panel (d), we see that $\mathbf{K}_{\boldsymbol{\theta}}$ is defined as a composite covariance function. It is the sum (`add_funcs`) of an SE kernel (`make_squaredexp`) and a white noise (WN, Appendix A) kernel which is implemented with `make_whitenoise`[2]. We then initialize `gpmem` feeding it with `composite_covariance` and the data look-up function `f_look_up`. We sample from the prior $\mathbf{f}_*$ with random parameters `sf,l` and `sigma` and without any observations available. We depict those samples on the right (red), alongside the true function that generated the data (blue) and the data points we have available in the data set (black).

We can incorporate observations using both `observe` and `predict` (Fig. 3 (e)). When we subsequently sample $\hat{\mathbf{f}}$ from the emulator with $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{K}})$, we can see that the GP posterior incorporates knowledge about the `data`. Yet, the hyper-parameters `sf,l` and `sigma` are still random, so the emulator does not capture the true underlying dynamics ($f_{\text{true}}$) of the `data` correctly.

Next, we demonstrate how we can capture these underlying dynamics within only 100 nested MH steps on the hyper-parameters to get a good approximation for their posterior $\hat{\mathbf{f}}$ (Fig. 3 (f)). We say nested because we first take two sweeps in the scope `hyperhyper` which characterizes (17) and then one sweep on the scope `hyper` which characterizes (16). This is repeated 100 using `repeat( 100, do(`$\cdots$ . Note that Neal devises an additional noise model and performs a large number of Hybrid-Monte Carlo and Gibbs steps to achieve this, whereas inference in Venture with `gpmem` is merely one line of code.

Finally, we can change our inference strategy altogether. If we decide that instead of following a Bayesian sampling approach, we would like to perform empirical optimization, we do this by only changing one line of code, deploying `map` instead of `mh` (Fig. 3 (g)).

### 4.2 Discovering qualitative structure from time series data

Inductive learning of symbolic expressions for continuous-valued time series data is a hard task which has recently been tackled using a greedy search over the approximate posterior of the possible kernel compositions for GPs (Duvenaud et al., 2013; Lloyd et al., 2014)[3].

With `gpmem` we can provide a fully Bayesian treatment of this, previously unavaible, using a stochastic grammar (see Fig. 4). This allows us to read an unstructured time series and automatically output a high-level, qualitative description of it. The stochastic grammar takes two inputs:

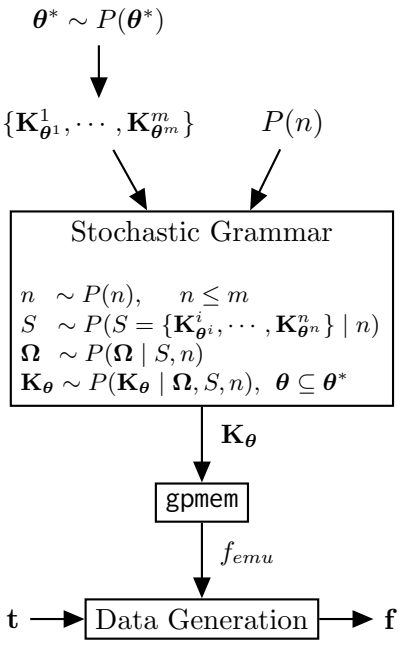
1. a set of base kernels $\{\mathbf{K}^1_{\boldsymbol{\theta}^1}, \cdots, \mathbf{K}^m_{\boldsymbol{\theta}^m}\}$ of size $m$, including their corresponding parametrization $\boldsymbol{\theta}^* = \{\boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^m\}$ (Fig. 4 (a) and (b)); and

2. a prior distribution on the number $n \leq m$ that determines how many of the base kernels are used for the composition. This can be uniform over the size, that is each possible $n$ is equally likely, or over the compositions, that is larger $n$ are more likely since they can yield a larger number of structures.

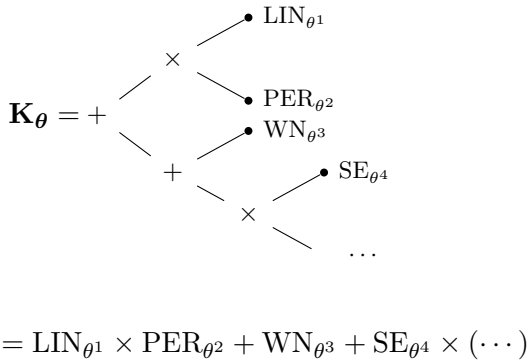
Internally, the stochastic grammar first samples the number of base kernels that it is going to combine into a compositional kernel,

$$n \sim P(n),$$

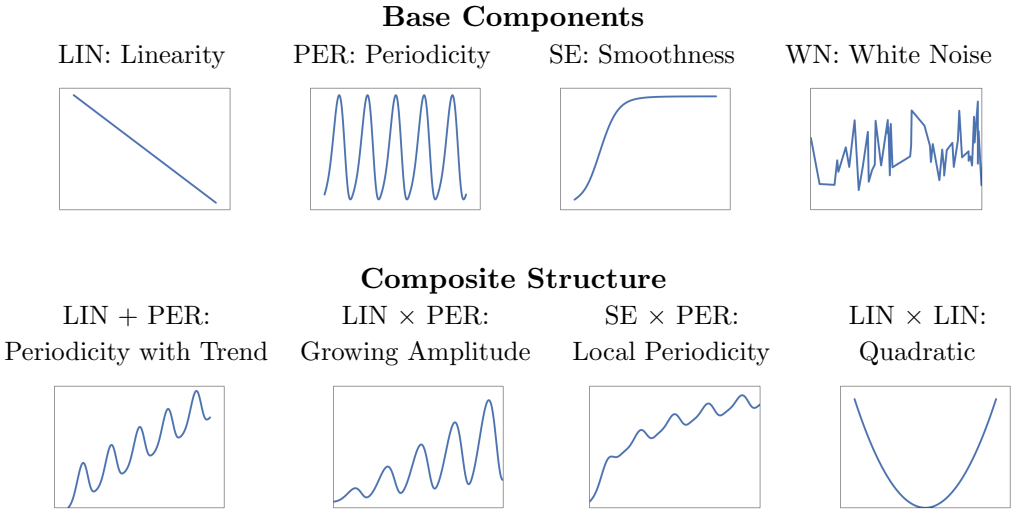
2. Note that in Neal's work (1997) the sum of an SE plus a constant kernel is used. We use a WN kernel for illustrative purposes instead.
3. `http://www.automaticstatistician.com/`




(a)




(b)




(c)

Figure 4: (a) Graphical description of Bayesian GP structure learning. A stochastic grammar generates samples of kernels that are used as input for `gpmem`. We supply a set of base kernels that serve as hypothesis space as well as a distribution of the possible size of the composite kernel. We observe value pairs of unstructured time series data on the bottom of the schematic. We can also generate predictions for future points in time. (b) An example of composite kernel structure. We use addition and multiplication to combine base kernels. Base kernels and compositional kernels are shown in (c) alongside the natural language interpretation of the structure.

using the supplied prior. We use this number to sample a random subset S of size $n$ of the set of supplied base kernels. We write

$$S = \{\mathbf{K}^i_{\boldsymbol{\theta}^i}, \cdots, \mathbf{K}^n_{\boldsymbol{\theta}^n}\} \sim P(S = \{\mathbf{K}^i_{\boldsymbol{\theta}^i}, \cdots, \mathbf{K}^n_{\boldsymbol{\theta}^n}\} \mid n)$$

with

$$P(S = \{\mathbf{K}^i_{\boldsymbol{\theta}^i}, \cdots, \mathbf{K}^n_{\boldsymbol{\theta}^n}\} \mid n) = \frac{n!}{\mid S = \{\mathbf{K}^i_{\boldsymbol{\theta}^i}, \cdots, \mathbf{K}^n_{\boldsymbol{\theta}^n}\} \mid !}.$$

The only building block that we are now missing is how to combine the sampled base kernels into a compositional covariance function (see Fig. 4 (b)). For each interaction $i$, we have to infer whether the data supports a local interaction or a global interaction, chosing between one out of two algebraic operators $\boldsymbol{\Omega}_i = \{+, \times\}$. The probability for all such decisions is given by a binomial distribution:

$$P(\boldsymbol{\Omega} \mid S, n) = \binom{n}{r} p^r_{+\times}(1 - p_{+\times})^{n-r}. \tag{18}$$

We can write the marginal probability of a composite structure as

$$P(\mathbf{K}_{\boldsymbol{\theta}} \mid \boldsymbol{\Omega}, S, n) = P(\boldsymbol{\Omega} \mid S, n) \times P(S \mid n) \times P(n), \tag{19}$$

with $\boldsymbol{\theta} \subseteq \boldsymbol{\theta}^*$. This composite kernel is fed into `gpmem`. The emulator generated by `gpmem` observes unstructured time series data. We show an implementation of our approach with `gpmem` in Listing 1.

Many equivalent covariance structures can be sampled due to covariance function algebra and equivalent representations with different parameterization (Lloyd et al., 2014). To inspect the posterior of these equivalent structures we convert each kernel expression into a sum of products and subsequently simplify. All base kernels can be found in Appendix A, rules for this simplification can be found in appendix B. The code for learning of kernel structure is as follows:

Listing 1: Structure Learning

```
// Initialize the set of base kernels
assume kernels = list(se, wn, lin, per) // defined as above

// Assign prior on the number of kernels. e.g., P(n=1)=0.5, P(n=2)=...
assume prior_num_kernels = uniform_structure(n)

// Generate a random subset of possible base kernels
assume subset_kernels = tag("grammar", 0,
                            subset(kernels, prior_num_kernels))

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

assume K = tag("grammar", 1, composition(subset_kernels))

// Apply gpmem
assume (f_compute, f_emu) = gpmem(f_look_up, K)

// Probe all data points
for n ... N
  predict f_compute(get_data_xs(n))

// Perform inference
infer repeat(200, do(
  mh("grammar", one, 1),
  mh("parameters", one, 2)))
```

We defined a simple space of covariance structures in a way that allows us to produce results coherent with work presented in Automatic Statistician. We will illustrate our results with two data sets.

**Mauna Loa $CO_2$ data**

We illustrate results in Fig 5. In Fig 5 (a) we depict the raw data. We see mean centered $CO_2$ measurements of the Mauna Loa Observatory, an atmospheric baseline station on Mauna Loa, on the island of Hawaii. A description of the data set can be found in Rasmussen and Williams, 2006, chapter 5. We use those raw data to compute a posterior on structure, parameters and GP samples. The latter are shown in Fig 5 (b) where we zoom in to show

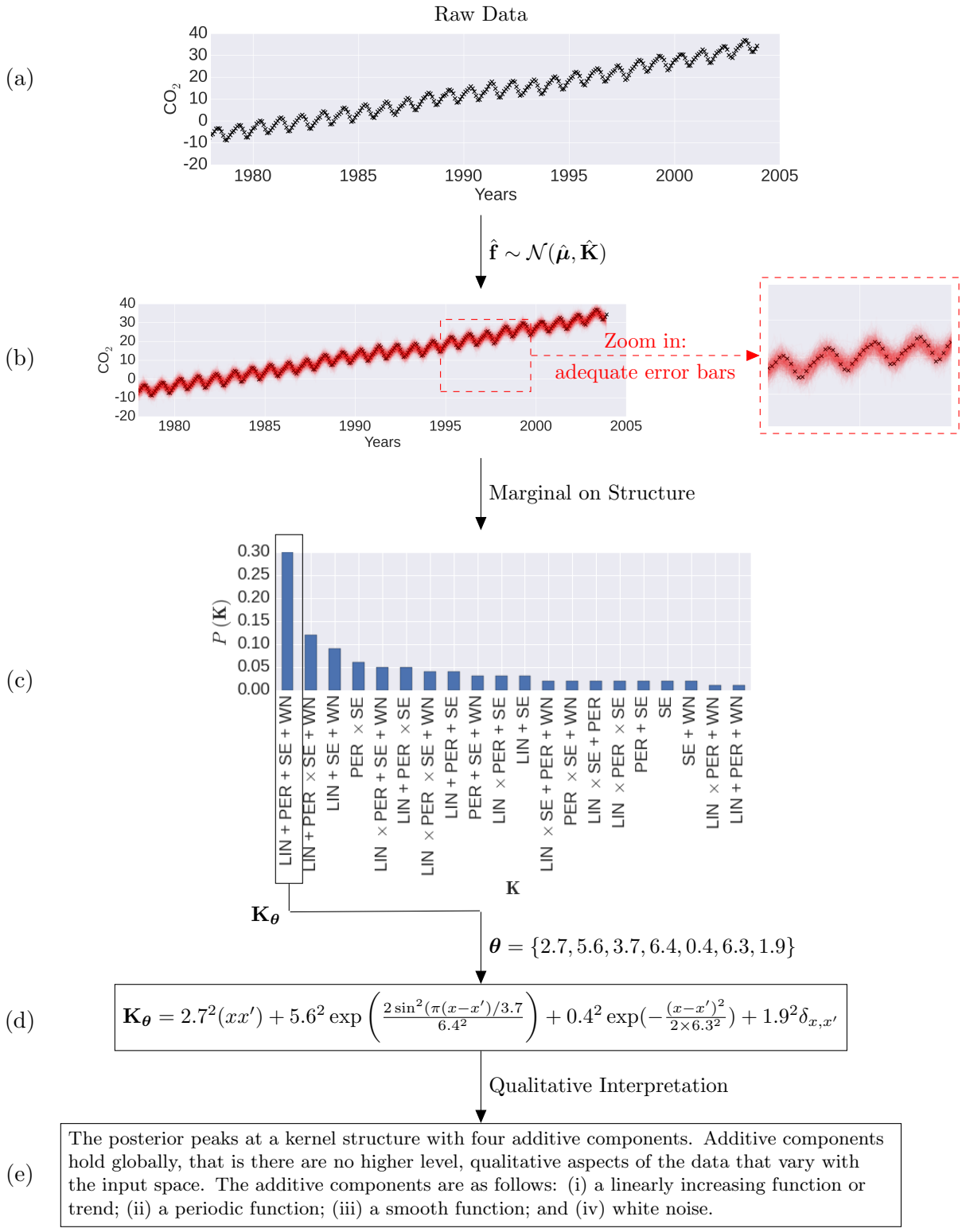



Figure 5: Structure Learning. Starting with raw data (a), we fit a GP (b) and compute the posterior distribution on structures (c). We take a sample of the peak of this distribution (LIN + PER + SE + WN) including its parameters and write it in functional form (d). We depict the human readable interpretation (e). We used (d) to plot (b).

how the posterior captures the error bars adequately. This posterior of the GP is generated with a random sample from the parameters of the peak of the distribution on structure (Fig 5 (c)). The distribution peaks at:

$$\mathbf{K} = \text{LIN} + \text{PER} + \text{SE} + \text{WN}. \tag{20}$$

We write this Kernel equation out in Fig 5 (d). This kernel structure has a natural language interpretation that we spell out in Fig 5 (e), explaining that the posterior peaks at a kernel structure with four additive components. Each of which holds globally, that is there are no higher level, qualitative aspects of the data that vary with the input space. The additive components for this result are as follows:

- a linearly increasing function or trend;
- a periodic function;
- a smooth function; and
- white noise.

Previous work on automated kernel discovery (Duvenaud et al., 2013) illustrated the Mauna Loa data using an RQ kernel. We resort to the white noise kernel instead of RQ (similar to (Lloyd et al., 2014)).

**Airline Data**
The second data set (Fig. 6) we depict results for is the airline data set describing monthly totals of international airline passengers (Box et al., 1997, according to Duvenaud et al., 2013).

We illustrate results for this data set in Fig 6. In Fig 6 (a) we depict the raw data. Again, the data is mean centered and we use it to compute a posterior on structure, parameters and GP samples. The latter are shown in Fig 6 (b). This posterior of the GP is generated with a random sample from the parameters of the peak of the distribution on structure (Fig 6 (c)). The distribution peaks at:

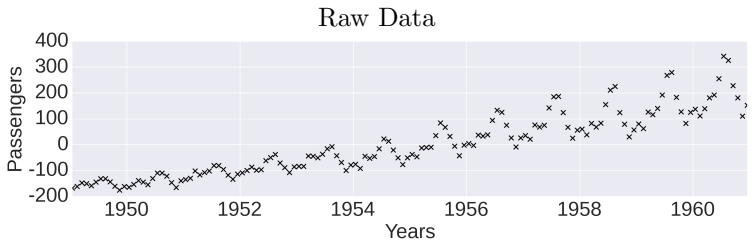
$$\mathbf{K} = \text{LIN} + \text{SE} \times \text{PER} + \text{WN}. \tag{21}$$

We write this Kernel equation out in Fig 6 (d). This kernel structure has a natural language interpretation that we spell out in Fig 6 (e), explaining that the posterior peaks at a kernel structure with three additive components. Additive components hold globally, that is there are no higher level, qualitative aspects of the data that vary with the input space. The additive components are as follows:

- a linearly increasing function or trend;
- a approximate periodic function; and
- white noise.

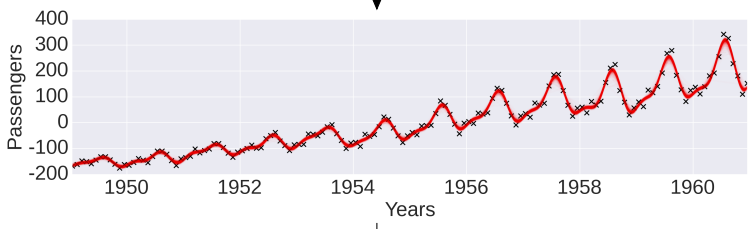
Both datasets served as illustrations in the Automatic Statistician project.

(a)



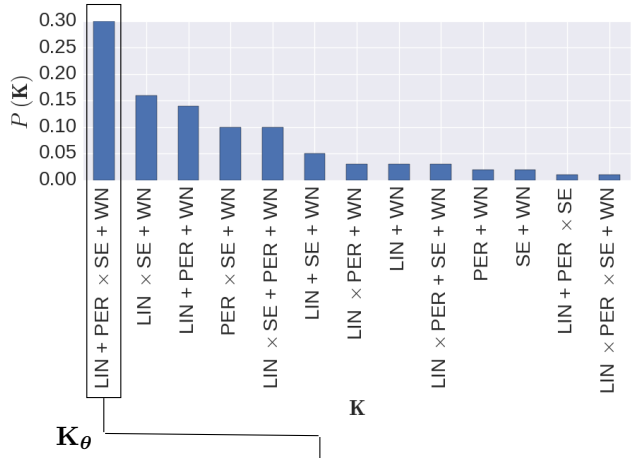

$\hat{\mathbf{f}} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{K}})$

(b)




Marginal on Structure

(c)




$\mathbf{K}_{\boldsymbol{\theta}}$

$\boldsymbol{\theta} = \{7.47, 0.27, 4.63, 7.34, 4.4, 4.55, 2.93\}$

(d)

$$= 7.47^2(xx') + \left(0.27^2 \exp(-\tfrac{(x-x')^2}{2\times 4.63^2}) \times 7.34^2 \exp\left(\tfrac{2\sin^2(\pi(x-x')/4.4}{4.55^2}\right)\right) + 2.93^2\delta_{x,x'}$$

Qualitative Interpretation

(e)

The posterior peaks at a kernel structure with three additive components. Additive components hold globally, that is there are no higher level, qualitative aspects of the data that vary with the input space. The additive components are as follows: (i) a linearly increasing function or trend; (ii) a approximate periodic function; and (iv) white noise.

Figure 6: Structure Learning. Starting with raw data (a), we fit a GP (b) and compute the posterior distribution on structures (c). We take a sample of the peak of this distribution (LIN + PER × SE + WN) including its parameters and write it in functional form (d). We depict the human readable interpretation (e). We used (d) to plot (b).

**Querying time series**

With our Bayesian approach to structure learning we can gain valuable insights into time series data that were previously unavailable. This is due to our ability to estimate posterior marginal probabilities over the kernel structure. Over this marginal, we define boolean search operations that allow us to query the data for the probability of certain structures to hold true globally. We write

$$P(\mathbf{K}) = \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{K}^n) \;\; \text{where } f(\mathbf{K}^n) = \begin{cases} 1, & \text{if } \mathbf{K} \underset{global}{\in} \mathbf{K}^n, \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

to ask whether it is true that a global structure $\mathbf{K}$ is present. $N$ is the number of all posterior samples for $\mathbf{K}_{\boldsymbol{\theta}}$ and $\mathbf{K}^n$ is one such sample. We omit $\boldsymbol{\theta}$ in the following to indicate that the qualitative type of structure does not depend on a specific parameterization. We can now ask simple questions, for example:

Is there white noise in the data?

where we set $\mathbf{K}$ =WN in (22). We can also formulate more sophisticated search operations using Boolean operators such as AND ($\wedge$) and OR ($\vee$). The AND operator is defined as follows:

$$P(\mathbf{K}^a \wedge \mathbf{K}^b) = \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{K}^n) \;\; \text{where } f(\mathbf{K}^n) = \begin{cases} 1, & \text{if } \mathbf{K}^a \text{ and } \mathbf{K}^b \underset{global}{\in} \mathbf{K}^n, \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

By estimating $P(\text{LIN} \wedge \text{WN})$ we can use this operator to ask questions such as

Is there a Linear component AND a white noise in the data?

Finally, we define the logical OR as

$$P(\mathbf{K}^a \vee \mathbf{K}^b) = P(\mathbf{K}^a) + P(\mathbf{K}^b) - P(\mathbf{K}^a \wedge \mathbf{K}^b) \tag{24}$$

which allows us to ask questions about structures that are logically connected with OR, such as:

Is there white noise or heteroskedastic noise?

by estimating $P(\text{LIN} \times \text{WN} \;\vee\; \text{WN})$. We know that noise can either be heteroskedastic or white, and we also know due to simple manipulations using kernel algebra that LIN $\times$ WN and WN are the only possible ways to construct noise with kernel composition, we see that we can generalize the question above to:

Is there noise in the data?

where we write the marginal posterior on qualitative structure for noise:

$$P(\mathbf{K}^{\text{noise}}) = P(\text{LIN} \times \text{WN} \;\vee\; \text{WN}). \tag{25}$$

Note that this allows us to start with general queries and subsequently formulate follow up queries that go into more detail. For example, we could start with a general query, such as:

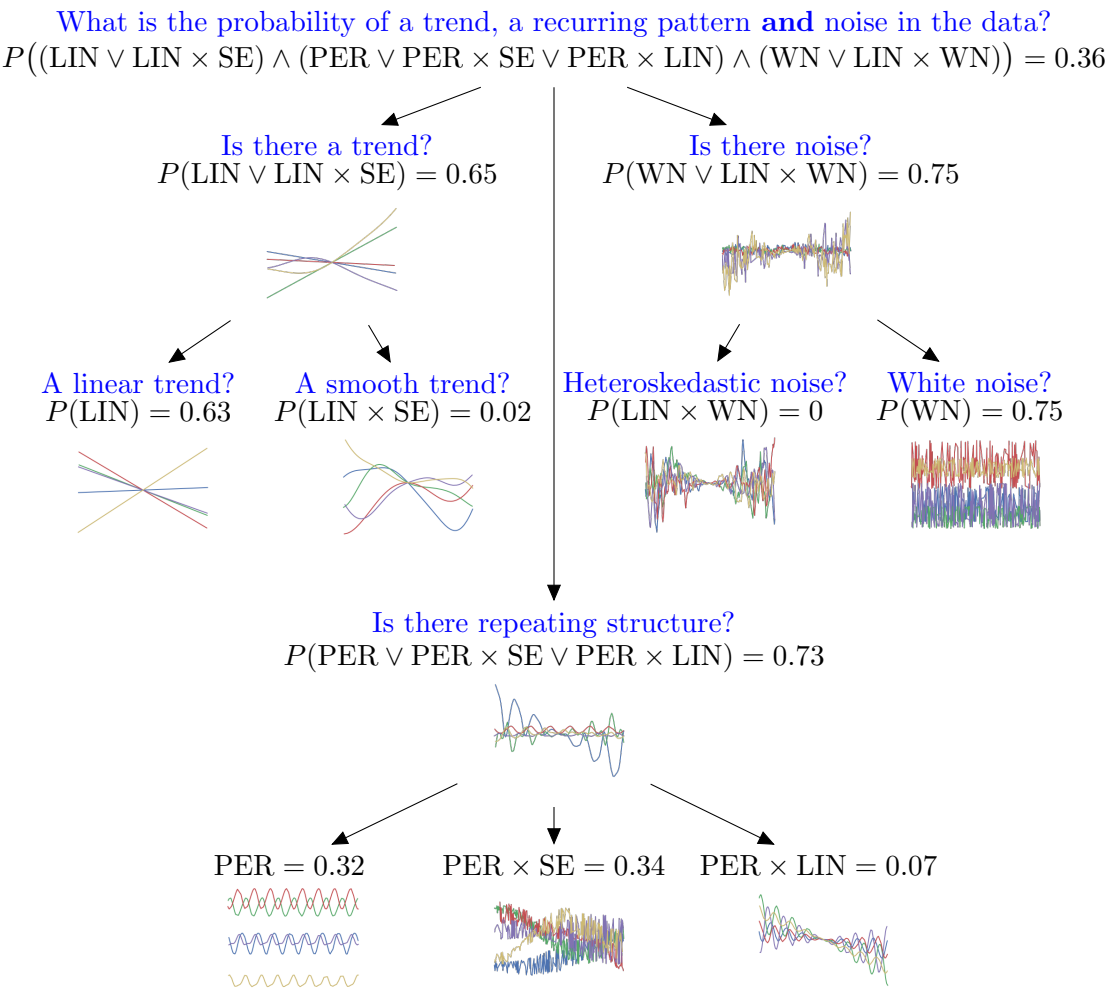



Figure 7: We can query the data to see if some logical statements are probable to be true. We start at the top with a fairly general question. We do this with the help of the logical operators defined above. We then ask more specific follow-up questions, for example, is it true that there is noise in the data? If we find it probable that there is noise than we can drill even deeper asking whether or not the data supports the hypothesis that there is heteroskedastic noise.

What is the probability of a trend, a recurring pattern **and** noise in the data?

and then follow up with more detailed questions (Fig 7). This way of querying your data for their statistical implications is in stark contrast to what previous research in automatic kernel construction was able to provide. We could view our approach as a time series search engine which allows us to test whether or not certain structures can be found in an available time series. Another way to view this approach is as a new language to interact with the world. Real-world observations often come with time-stamps and in form of continuous valued sensor measurements. We provide the toolbox to query such observations in a similar manner as one would query a knowledge base in a logic programming language.

### 4.3 Bayesian optimization

The final application demonstrating the power of `gpmem` illustrates its use in Bayesian optimization. We introduce Thompson sampling, the basic solution strategy underlying the Bayesian optimization with `gpmem`. Thompson sampling Thompson (1933) is a widely used Bayesian framework for addressing the trade-off between exploration and exploitation in multi-armed (or continuum-armed) bandit problems. We cast the multi-armed bandit problem as a one-state Markov decision process (MDP), and describe how Thompson sampling can be used to choose actions for that MDP.

The MDP is as follows: 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 $P_{\text{true}}(r \mid a)$. The agent's goal is to maximize the total reward received for all actions in an online manner. In Thompson sampling, the agent accomplishes this by placing a prior distribution $P(\vartheta)$ on the possible "contexts" $\vartheta \in \Theta$. Here a context is a believed model of the conditional distributions $\{P(r \mid a)\}_{a \in \mathcal{A}}$, or at least, a believed statistic of these conditional distributions which is sufficient for deciding an action $a$. If actions are chosen so as to maximize expected reward, then one such sufficient statistic is the believed conditional mean $V(a \mid \vartheta) = \mathbb{E}[r \mid a; \vartheta]$, which can be viewed as a believed value function. For consistency with what follows, we will assume our context $\vartheta$ takes the form $(\boldsymbol{\theta}, \mathbf{a}, \mathbf{r})$ where $\mathbf{a}$ is the vector of past actions, $\mathbf{r}$ is the vector of their rewards, and $\boldsymbol{\theta}$ (the "semicontext") contains any other information that is included in the context.

In this setup, Thompson sampling has the following steps:

**Algorithm 1** Thompson sampling.

Repeat as long as desired:

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

Note that when $\mathbb{E}[r|a;\vartheta]$ (under the sampled value of $\boldsymbol{\theta}$ for some points $a$) is far from the true value $\mathbb{E}_{P_{\text{true}}}[r \mid a]$, the chosen action $a$ may be far from optimal, but the information gained by probing action $a$ will improve the belief $\vartheta$. This amounts to "exploration." When $\mathbb{E}[r \mid a;\vartheta]$ is close to the true value except at points $a$ for which $\mathbb{E}[r \mid a;\vartheta]$ is low, exploration will be less likely to occur, but the chosen actions $a$ will tend to receive high rewards. This amounts to "exploitation." The trade-off between exploration and exploitation is illustrated in Figure 8. Roughly speaking, exploration will happen until the context $\vartheta$ is reasonably sure that the unexplored actions are probably not optimal, at which time the Thompson sampler will exploit by choosing actions in regions it knows to have high value.

Typically, when Thompson sampling is implemented, the search over contexts $\vartheta \in \Theta$ is limited by the choice of representation. In traditional programming environments, $\boldsymbol{\theta}$ 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 $\Theta$ 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 $\{P(r \mid a)\}_{a\in\mathcal{A}}$ 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 $\Theta$ has a natural embedding as a subset of $\widehat{\Theta}$.

**A Mathematical Specification**

We now describe a particular case of Thompson sampling with the following properties:

- The regression function has a Gaussian process prior.
- The actions $a_1, a_2, \ldots \in \mathcal{A}$ are chosen by a Metropolis-like search strategy with Gaussian drift proposals.
- The hyperparameters of the Gaussian process are inferred using Metropolis–Hastings sampling after each action.

In this version of Thompson sampling, the contexts $\vartheta$ are Gaussian processes over the action space $\mathcal{A} = [-20, 20] \subseteq \mathbb{R}$. That is,

$$V \sim \mathcal{GP}\big(\boldsymbol{\mu}(a), \mathbf{K}(a,a)\big),$$

where the mean $\boldsymbol{\mu}$ is a computable function $\mathcal{A} \to \mathbb{R}$ and the covariance $\mathbf{K}$ 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 $a$. We write past actions as $\mathbf{a}$ and past rewards as $\mathbf{r}$. Computationally, we represent a context as a data structure

$$\vartheta = (\boldsymbol{\theta}, \mathbf{a}, \mathbf{r}) = (\boldsymbol{\mu}, \mathbf{K}, \boldsymbol{\theta}, \mathbf{a}, \mathbf{r}),$$

where $\boldsymbol{\mu}$ is a procedure to be used as the prior mean function and $\mathbf{K}$ is a procedure to be used as the prior covariance function, parameterized by $\boldsymbol{\theta}$. As above set $\boldsymbol{\mu} \equiv 0$.

Note that the context space $\Theta$ is not a finite-dimensional parametric family, since the vectors $\mathbf{a}$ and $\mathbf{r}$ grow as more samples are taken. $\Theta$ is, however, representable as a computational procedure together with parameters and past samples, as we do in the representation $\vartheta = (\boldsymbol{\mu}, \mathbf{K}, \eta, \mathbf{a}, \mathbf{r})$.

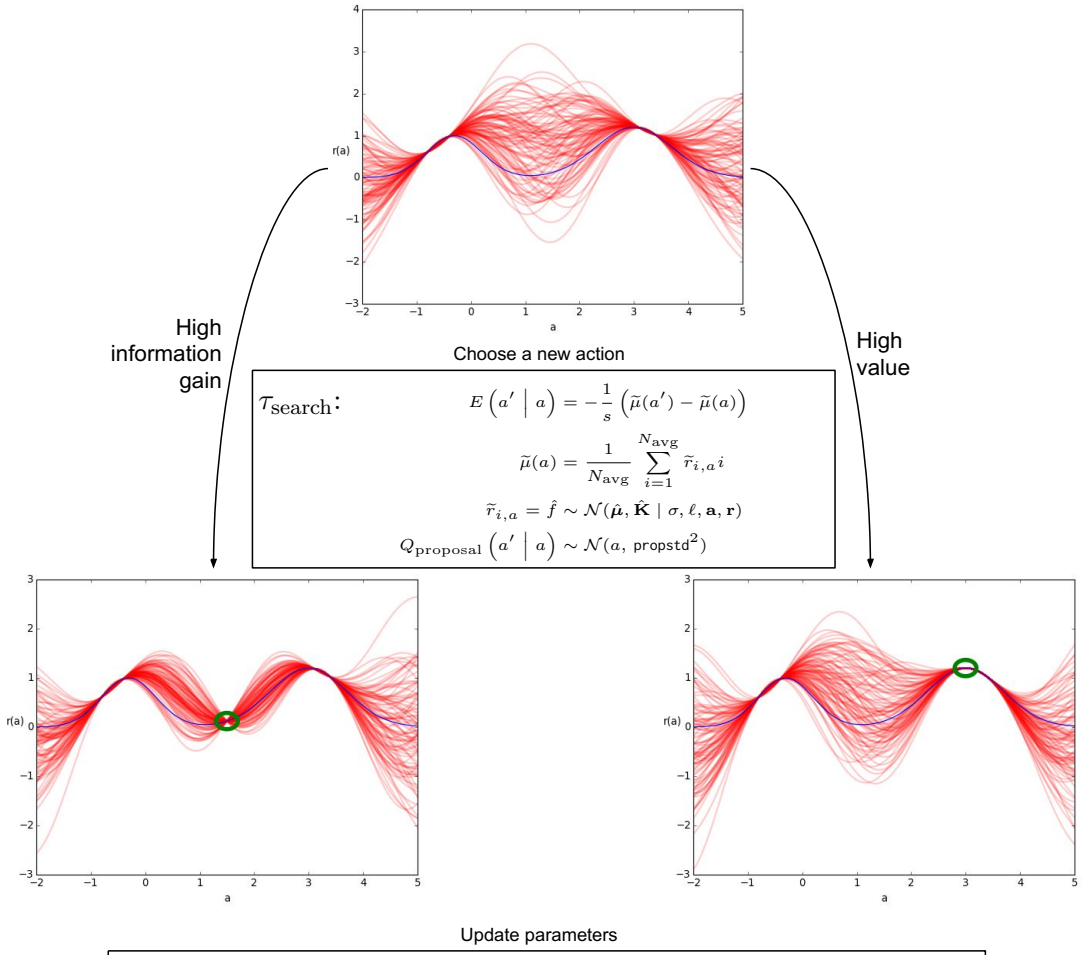



Figure 8: Two possible actions (in green) for an iteration of Thompson sampling. The believed distribution on the value function $V$ is depicted in red. In this example, the true reward function is deterministic, and is drawn in blue. The action on the right receives a high reward, while the action on the left receives a low reward but greatly improves the accuracy of the believed distribution on $V$. The transition operators $\tau_{\text{search}}$ and $\tau_{\text{update}}$ are described in Section 4.3.

We combine the Update and Sample steps of Algorithm 1 by running a Metropolis–Hastings (MH) sampler whose stationary distribution is the posterior $P(\boldsymbol{\theta} \mid \mathbf{a}, \mathbf{r})$. The functional forms of $\boldsymbol{\mu}$ and $\mathbf{K}$ are fixed in our case, so inference is only done over the parameters $\boldsymbol{\theta} = \{\sigma, \ell\}$; hence we equivalently write $P(\sigma, \ell \mid \mathbf{a}, \mathbf{r})$ for the stationary distribution. We make MH proposals to one variable at a time, using the prior as proposal distribution:

$$Q_{\text{proposal}}\left(\sigma', \ell \;\middle|\; \sigma, \ell\right) = P(\sigma')$$

and

$$Q_{\text{proposal}}\left(\sigma, \ell' \;\middle|\; \sigma, \ell\right) = P(\ell').$$

The MH acceptance probability for such a proposal is

$$P_{\text{accept}}\left(\sigma', \ell' \;\middle|\; \sigma, \ell\right) = \min\left\{1,\ \frac{Q_{\text{proposal}}\left(\sigma, \ell \mid \sigma', \ell'\right)}{Q_{\text{proposal}}\left(\sigma', \ell' \mid \sigma, \ell\right)} \cdot \frac{P\left(\mathbf{a}, \mathbf{r} \mid \sigma', \ell'\right)}{P\left(\mathbf{a}, \mathbf{r} \mid \sigma, \ell\right)}\right\}$$

Because the priors on $\sigma$ and $\ell$ are uniform in our case, the term involving $Q_{\text{proposal}}$ equals 1 and we have simply

$$\begin{aligned} P_{\text{accept}}\left(\sigma', \ell' \;\middle|\; \sigma, \ell\right) &= \min\left\{1,\ \frac{P\left(\mathbf{a}, \mathbf{r} \mid \sigma', \ell'\right)}{P\left(\mathbf{a}, \mathbf{r} \mid \sigma, \ell\right)}\right\} \\ &= \min\left\{1,\ \exp\left(-\frac{1}{2}\left(\mathbf{r}^T \mathbf{K}\left(\mathbf{a}, \mathbf{a} \mid \sigma', \ell'\right)^{-1} \mathbf{r} \right.\right.\right. \\ &\qquad \left.\left.\left. - \mathbf{r}^T \mathbf{K}\left(\mathbf{a}, \mathbf{a} \mid \sigma, \ell\right)^{-1} \mathbf{r}\right)\right)\right\}. \end{aligned}$$

The proposal and acceptance/rejection process described above define a transition operator $\tau_{\text{update}}$ which is iterated a specified number of times; the resulting state of the MH Markov chain is taken as the sampled semicontext $\boldsymbol{\theta}$ in Step 1 of Algorithm 1.

For Step 2 (Search) of Thompson sampling, we explore the action space using an MH-like transition operator $\tau_{\text{search}}$. As in MH, each iteration of $\tau_{\text{search}}$ produces a proposal which is either accepted or rejected, and the state of this Markov chain after a specified number of steps is the new action $a$. The Markov chain's initial state is the most recent action, and the proposal distribution is Gaussian drift:

$$Q_{\text{proposal}}\left(a' \;\middle|\; a\right) \sim \mathcal{N}(a,\, \texttt{propstd}^2),$$

where the drift width `propstd` is specified ahead of time. The acceptance probability of such a proposal is

$$P_{\text{accept}}\left(a' \;\middle|\; a\right) = \min\left\{1,\ \exp\left(-E\left(a' \;\middle|\; a\right)\right)\right\},$$

where the energy function $E(\bullet \mid a)$ is given by a Monte Carlo estimate of the difference in value from the current action:

$$E\left(a' \;\middle|\; a\right) = -\frac{1}{s}\left(\widetilde{\mu}(a') - \widetilde{\mu}(a)\right)$$

where

$$\widetilde{\mu}(a) = \frac{1}{N_{\text{avg}}} \sum_{i=1}^{N_{\text{avg}}} \widetilde{r}_{i,a}$$

and

$$\widetilde{r}_{i,a} = \hat{f} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{K}}))$$

and $\{\widetilde{r}_{i,a}\}_{i=1}^{N_{\text{avg}}}$ are i.i.d. for a fixed $a$. Here the temperature parameter $s \geq 0$ and the population size $N_{\text{avg}}$ are specified ahead of time. Proposals of estimated value higher than that of the current action are always accepted, while proposals of estimated value lower than that of the current action are accepted with a probability that decays exponentially with respect to the difference in value. The rate of the decay is determined by the temperature parameter $s$, where high temperature corresponds to generous acceptance probabilities. For $s = 0$, all proposals of lower value are rejected; for $s = \infty$, all proposals are accepted. For points $a$ at which the posterior mean $\hat{\boldsymbol{\mu}}$ is low but the posterior variance $\hat{\mathbf{K}}$ is high, it is possible (especially when $N_{\text{avg}}$ is small) to draw a "wild" value of $\widetilde{\mu}(a)$, resulting in a favorable acceptance probability.

Indeed, taking an action $a$ with low estimated value but high uncertainty serves the useful function of improving the accuracy of the estimated value function at points near $a$ (see Figure 8).[4,5] We see a comlete probabilistic program with `gpmem` implementing Bayesian optimization with Thompson Sampling below (Listing 2).

4. At least, this is true when we use a smoothing prior covariance function such as the squared exponential.
5. For this reason, we consider the sensitivity of $\widehat{\mu}$ to uncertainty to be a desirable property; indeed, this is why we use $\widehat{\mu}$ rather than the exact posterior mean $\mu$.

Listing 2: Bayesian optimization using gpmem

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

// A naive estimate of the argmax of the given function
define mc_argmax = proc(func) {
  candidate_xs = mapv(proc(i) {uniform_continuous(-20, 20)},
                      arange(20));
  candidate_ys = mapv(func, candidate_xs);
  lookup(candidate_xs, argmax_of_array(candidate_ys))
};

// Shortcut to sample the emulator at a single point without packing
// and unpacking arrays
define emulate_pointwise = proc(x) {
  run(sample(lookup(f_emulate(array(unquote(x))), 0)))
};

// Main inference loop
infer repeat(15, do(pass,
  // Probe V at the point mc_argmax(emulate_pointwise)
  predict(f_compute(unquote(mc_argmax(emulate_pointwise)))),
  // Infer hyperparameters
  mh("hyper", one, 50)));
```

In Fig. 9 we show results for our implementation of Bayesian Optimization with Thompson sampling. We compare two different proposal distributions, namely uniform proposals and Gaussian drift proposals. We see that in this experiment, Gaussian drift is starting near the global optimum and drifts quickly towards it. (red curve, top panel of Fig. 9). Uniform proposals take longer to find the global optimum (blue curve, top panel of Fig. 9) but we see that it can surpass the local optima of the curve[6]. The bottom panel of Fig. 9 depicts a sequence of actions using uniform proposals. The sequence illustrates the exploitation exploration trade-off that the implementation overcomes. We start with complete uncertainty ($i = 2$). The Bayesian agent performs exploration until it gets a (wrong!) idea of where the optimum could be (exploiting the local optima $i = 5$ to $i = 10$). $i = 11$ shows a change in tactic. The Bayesian agent, having exploited the local optima in previous steps, is now reducing uncertainty in area it knows nothing about, eventually finding the global optimum.

6. In fact, repeated experiments have shown that when the Gaussian drift proposals starts near a local optimum, it gets stuck there. Uniform proposals do not.

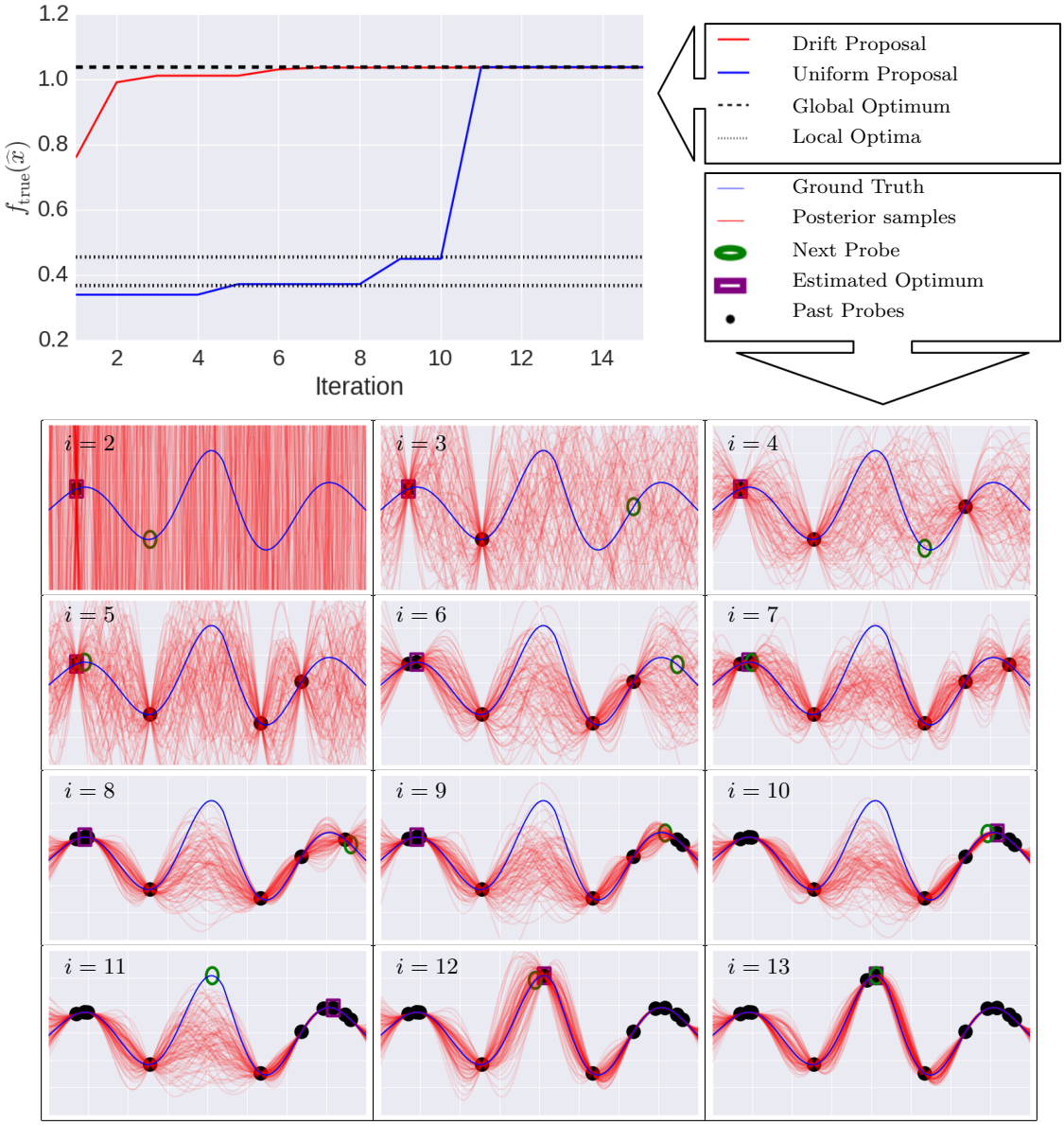



Figure 9: Top: the estimated optimum over time. Blue and Red represent optimization with uniform and Gaussian drift proposals. Black lines indicate the local optima of the true functions. Bottom: a sequence of actions. Depicted are iterations 7-12 with uniform proposals.

## 5. Discussion

We provided `gpmem`, an elegant linguistic framework for function learning-related tasks such as Bayesian optimization and GP kernel structure learning. We highlighted how `gpmem` overcomes shortcomings of the notations currently used in statistics, and how language constructs from programming allow the expression of models which would be cumbersome (prohibitively so, in some cases) to express in statistics notation. We evaluated our contribution on a range of hard problems for state-of-the-art Bayesian nonparametrics.

# Appendix

## A Covariance Functions

$$\mathrm{SE} = \sigma^2 \exp(-\frac{(x-x')^2}{2\ell^2}) \tag{26}$$

$$\mathrm{LIN} = \sigma^2(xx') \tag{27}$$

$$\mathrm{C} = \sigma^2 \tag{28}$$

$$\mathrm{WN} = \sigma^2\delta_{x,x'} \tag{29}$$

$$\mathrm{RQ} = \sigma^2\left(1 + \frac{(x-x')^2}{2\alpha\ell^2}\right)^{-\alpha} \tag{30}$$

$$\mathrm{PER} = \sigma^2 \exp\left(\frac{2\sin^2(\pi(x-x')/p}{\ell^2}\right). \tag{31}$$

## B Covariance Simplification

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

Rule 1 is derived as follows:

$$\begin{aligned}
\sigma_c^2 \exp(-\frac{(x-x')^2}{2\ell_c^2}) &= \sigma_a^2 \exp(-\frac{(x-x')^2}{2\ell_a^2}) \times \sigma_b^2 \exp(-\frac{(x-x')^2}{2\ell_b^2}) \\
&= \sigma_c^2 \exp(-\frac{(x-x')^2}{2\ell_a^2}) \times \exp(-\frac{(x-x')^2}{2\ell_b^2}) \\
&= \sigma_c^2 \exp\left(-\frac{(x-x')^2}{2\ell_a^2} - \frac{(x-x')^2}{2\ell_b^2}\right) \\
&= \sigma_c^2 \exp\left(-\frac{(x-x')^2}{2\ell_c^2}\right)
\end{aligned} \tag{32}$$

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 (Rule 2). Take for example the SE kernel:

$$\sigma_a^2 \exp\left(-\frac{(x-x')^2}{2\ell_c^2}\right) \times \sigma_b\delta_{x,x'} = \sigma_a\sigma_b\delta_{x,x'} \tag{33}$$

Rule 3 is derived as follows:

$$\theta_c(x \times x') = \theta_a(x \times x') + \theta_b(x \times x') \tag{34}$$

Multiplying any kernel with a constant obviously changes only the scale parameter of a kernel (Rule 4).

## C Glossary

| | |
|---|---|
| $x$ | Scalar |
| $X$ | Random variable |
| $\mathbf{x}$ | Column vector |
| $\mathbf{X}$ | Matrix |
| $\mathcal{D}$ | Data matrix |
| $\mathbb{E}$ | Expectation |
| $\mathcal{N}$ | (Multivariate-) Gaussian |
| $\mathcal{GP}$ | Gaussian Process |
| $\mu$ | Mean |
| $\boldsymbol{\mu}$ | Mean vector |
| $\sigma$ | Standard deviation |
| $m(x)$ | Mean function |
| $k(x, x')$ | Kernel or covariance function |
| $\mathbf{K} = \mathbf{K}(\mathbf{x}, \mathbf{x})$ | Prior Kernel or covariance function given a data vector |
| $\hat{\mathbf{x}}$ | unseen input data |
| $\mathbf{K}(\mathbf{x}, \hat{\mathbf{x}})$ | Cross covariance |
| $f = f(x)$ | One-dimensional sample from a GP |
| $\mathbf{f}$ | Sample from a GP prior \ regression target, i.e. observed output |
| $\hat{\mathbf{f}}$ | Sample from a predictive posterior, that is regression output for unseen data $\hat{\mathbf{x}}$ |
| $\hat{\boldsymbol{\mu}}$ | Posterior mean for $\hat{\mathbf{f}}$ |
| $\hat{\mathbf{K}}$ | Posterior covariance for $\hat{\mathbf{f}}$ |
| SE | Squared exponential covariance function on a data vector |
| LIN | Linear covariance function on a data vector |
| PER | Periodic covariance function on a data vector |
| RQ | Rational quadratic covariance function on a data vector |
| C | Constant covariance function on a data vector |
| WN | White noise covariance function on a data vector |
| $\wedge$ | Logical and |
| $\vee$ | Logical or |

## References

Stephen Barnett and Stephen Barnett. *Matrix methods for engineers and scientists*. McGraw-Hill, 1979.

D. Barry. Nonparametric bayesian regression. *The Annals of Statistics*, 14(3):934–953, 1986.

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

A. Damianou and N. Lawrence. Deep gaussian processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 207–215, 2013.

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 International Conference on Machine Learning (ICML)*, pages 1166–1174, 2013.

B. Ferris, D. Haehnel, and D. Fox. Gaussian processes for signal strength-based location estimation. In *Proceedings of the Conference on Robotics Science and Systems*. Citeseer, 2006.

M. A. Gelbart, J. Snoek, and R. P. Adams. Bayesian optimization with unknown constraints. *arXiv preprint arXiv:1403.5607*, 2014.

M. Kemmler, E. Rodner, E. Wacker, and J. Denzler. One-class classification with gaussian processes. *Pattern Recognition*, 46(12):3507–3518, 2013.

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 425–464, 2001.

M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary gaussian process classification. *The Journal of Machine Learning Research*, 6:1679–1704, 2005.

J. Kwan, S. Bhattacharya, K. Heitmann, and S. Habib. Cosmic emulation: The concentration-mass relation for wcdm universes. *The Astrophysical Journal*, 768(2):123, 2013.

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

David JC MacKay. Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166, 1998.

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

R. M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1995.

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

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

M. D. Schneider, L. Knox, S. Habib, K. Heitmann, D. Higdon, and C. Nakhleh. Simulations and cosmological inference: A statistical model for power spectra means and covariances. *Physical Review D*, 78(6):063529, 2008.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 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.

C. K. I. Williams and D. Barber. Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.