Chapter 1

Introduction

> "I only work on intractable nonparametrics - Gaussian processes don't count."
>
> Sinead Williamson, personal communication

Regression is usually stated mathematically as the problem learning a function $f(x)$ mapping from some inputs to some outputs. From a human's point of view, the problem is closer to: Which method should I use? What model class or software will probably work well? Is my current method silently failing?

We would like an expressive language which can represent both simple parametric forms of $f(x)$ such as linear, polynomial, etc. and also complex nonparametric functions specified in terms of properties such as smoothness, periodicity, etc. Fortunately, Gaussian processes (GPs) provide a very general and analytically tractable way of capturing both simple and complex functions.

1.1 Gaussian process models

Gaussian processes are a flexible and tractable prior over functions, useful for solving regression and classification tasks (Rasmussen and Williams, 2006). The kind of structure which can be captured by a GP model is mainly determined by its *kernel*: the covariance function. One of the main difficulties in specifying a Gaussian process model is in choosing a kernel which can represent the structure present in the data. For small to medium-sized datasets, the kernel has a large impact on modeling efficacy.

Figure 1.1 shows a Gaussian process distribution, as it is conditioned on more and more observations.

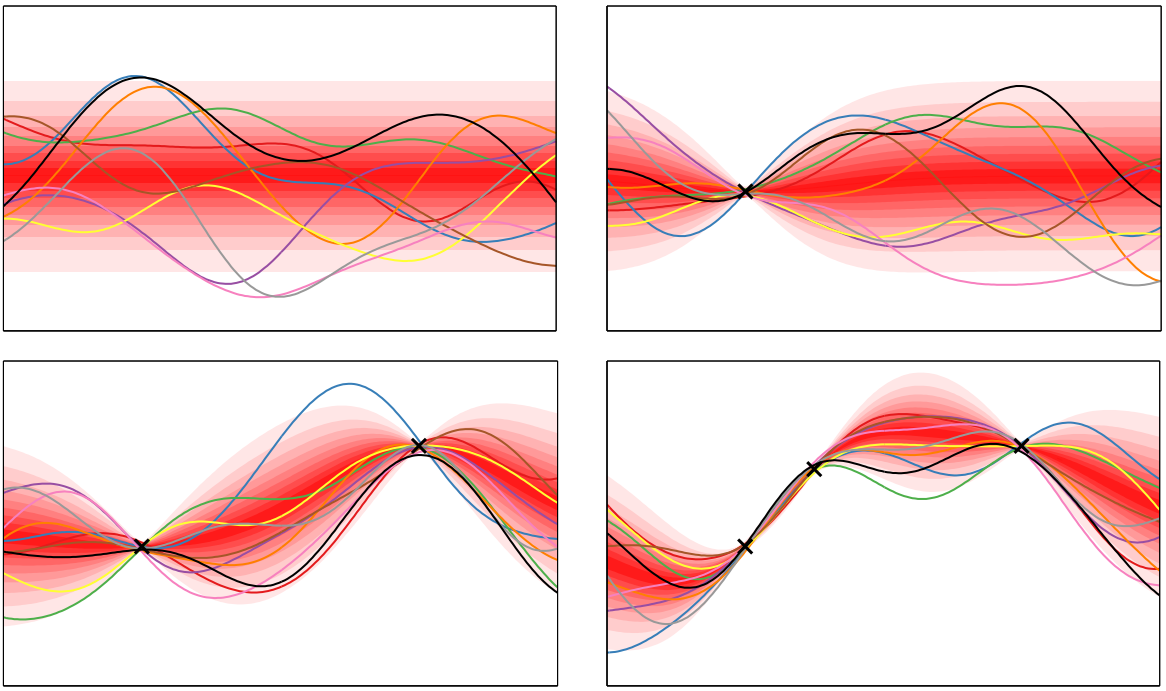

Fig. 1.1 A visual representation of a one-dimensional Gaussian process posterior. Red isocountours show the marginal density at each input location. Coloured lines are samples from the posterior.

Gaussian processes are distributions over functions such that any finite subset of function evaluations, $(f(x_1), f(x_2), \dots f(x_N))$, have a joint Gaussian distribution (Rasmussen and Williams, 2006). A GP is completely specified by its mean function, $\mu(x) = \mathbb{E}(f(x))$ and kernel (or covariance) function $k(x, x') = \mathrm{Cov}(f(x), f(x'))$. It is common practice to assume zero mean, since marginalizing over an unknown mean function can be equivalently expressed as a zero-mean GP with a new kernel. The structure of the kernel captures high-level properties of the unknown function, f, which in turn determines how the model generalizes or extrapolates to new data. We can therefore define a language of regression models by specifying a language of kernels.

For concreteness, we give the marginal likelihood of the data given a GP prior:

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mu(x), k(x,x)) &= \mathcal{N}(\mathbf{y}|\mu(\mathbf{X}), k(\mathbf{X},\mathbf{X})) \\ &= (2\pi)^{-\frac{N}{2}} |k(\mathbf{X},\mathbf{X})|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y} - \boldsymbol{\mu}(\mathbf{X})\right)^{\mathsf{T}} k(\mathbf{X},\mathbf{X})^{-1}\left(\mathbf{y} - \boldsymbol{\mu}(\mathbf{X})\right)\right\} \end{aligned} \tag{1.1}$$

$$p(y(x^*)|\mathbf{y}, \mathbf{X}, \mu(x), k(x,x)) = \mathcal{N}(y(x^*)|\mu(y(x^*)) + k(x^*, \mathbf{X})k(\mathbf{X}, \mathbf{X})^{-1}(\mathbf{y} - \mu(\mathbf{X}))$$
$$k(x^*, x^*) - k(x^*, \mathbf{X})k(\mathbf{X}, \mathbf{X})^{-1}k(\mathbf{X}, x^*)) \quad (1.2)$$

1.1.1 Useful properties of Gaussian process models

- **Tractable inference** Given a kernel function, the posterior distribution can be computed exactly in closed form. This is a rare property for nonparametric models to have.
- **Expressivity** By choosing different covariance functions, we can express a very wide range of modeling assumptions.
- **Integration over hypotheses** The fact that a GP posterior lets us exactly integrate over a wide range of hypotheses means that overfitting is less of an issue than in comparable model classes - for example, neural nets.
- **Marginal likelihood** A side benefit of being able to integrate over all hyoptheses is that we compute the *marginal likelihood* of the data given the model. This gives us a principled way of comparing different Gaussian process models.
- **Closed-form posterior** The posterior predictive distribution of a GP is another GP. This means that GPs can easily be composed with other models or decision procedures. For example, (⋆) Carl's reinforcement learning work.

The class of models that could be called Gaussian processes is extremely broad. Examples of commonly used models not usually cast as GPs are linear regression, splines, some forms of generalized additive models, and Kalman filters.

1.2 Limitations of Gaussian process models

- **Intractable inference** Computing the matrix inverse in (1.1) and (1.2) takes $\mathcal{O}(N^3)$ time.This problems can be addressed by approximate inference schemes, and most GP software packages implement some subset of these.
- **Gaussian likelihoods** Another limitation in practice is that we may wish to consider noise models other than Gaussian, in order to be robust to noise. When

making predictions, we might also want a heavy-tailed predictive distribution, to ensure our decisions take into account the possibility of extreme events. Using non-Gaussian noise models requires approximate inference schemes; fortunately, mature software packages exist which can automatically perform approximate inference for a wide variety of likelihoods.

- **Stationarity**

- **Symmetric** The predictive distribution of a GP is always symmetric about its mean function. This means that GP models are inappropriate for modeling non-negative functions, such as likelihoods. However, this problem can be addressed by simply exponentiating a GP model, giving rise to a *log-Gaussian process*.

1.2.1 Why stick to such a limited model class?

It may seem unsatisfying to restrict ourselves to a limited model class. Shouldn't we instead learn to use some more flexible model class, such as the set of all computible functions? The answer is: simple models can be used as well-understood building blocks for constructing more interesting models in diverse settings.

Consider linear models. Although they form an extremely limited model class, they are fast, simple, and easy to analyze. This makes them easy to incorporate into other models or procedures. Linear models may seem like a hopelessly simple model class, but they're arguably the most useful modeling tools in existence.

Gaussian processes can be seen as a dual representation of Bayesian linear regression. By moving into this dual space, we pay a price in computational complexity, but gain the ability to model functions to any desired level of detail. Crucially, the marginal likelihood allows us to automatically discover the appropriate amount of detail to use, by Bayesian Occam's razor (**??**).

1.3 Outline and Contributions

This thesis presents a set of related results about how the probabilistic nature of Gaussian process models allows them to be easily extended or composed with other models. Furthermore, the fact that the marginal likelihood is often available (or easily approximable) means that we can evaluate how much evidence the data provides for one structure over another.

Chapter ?? contains a wide-ranging overview of many of the types of structured priors on functions that can be easily expressed by constructing appropriate covariance functions. For example, in chapter **??**, we'll see how GPs can be combined with latent variable models to produce models of nonparametric manifolds. By introducing structure into the kernels of those GPs, we can create manifolds with diverse topological structures, such as cylinders, torii and Möbius strips.

Given a wide variety of structures, plus the ability to evaluate the suitability of each one, it's straightforward to automatically search over models. **Chapter ??** shows how to construct a general, open-ended language over kernels - which implies a corresponding language over models. In chapter **??**, we'll see how the marginal likelihood can guide us into automatically building structured models, and how capturing structure allows us to extrapolate, rather than simply interpolating.

Another benefit of using a relatively simple model class is that the resulting models are easy to understand. **Chapter ??** demonstrates how easy-to-understand the resulting models are, by demonstrating a simple system which automatically describes the structure discovered in a dataset by a search over GP models. This system automatically generates reports with graphs and english-language descriptions of GP models. Chapter **??** shows that, for the particular language of models constructed in chapter **??**, it's relatively easy to automatically generate english-language descriptions of the models discovered. Augmented with interpretable plots decomposing the predictive posterior, we demonstrate how to automatically generate useful analyses of time-series. Combined with the automatic model search developed in chapter **??**, this system represents the beginnings of an "automatic statistician". We discuss the advantages and potential pitfalls of automating the modeling process in this way.

Chapter ?? examines the model class obtained by performing dropout in GPs, finding them to have equivalent covariance to *additive Gaussian processes*, a model summing over exponentially-many GP models, each depending on a different subset of the input variables. An polynomial-time algorithm for doing inference in this model class is given, and the resulting model class is characterized and related to existing model classes.

Chapter ?? develops an extension of the GP-LVM in which the latent distribution is a mixture of Gaussians. This model gives rise to a Bayesian clustering model in the clusters have nonparametric shapes. Like the density manifolds learned by the GP-LVM, the shapes of the clusters learned by the iWMM follow the contours of the data density.

Chapter ?? examines the prior over functions obtained by composing GP priors

to form *deep Gaussian processes*, and relates them to existing neep neural network architecures. We find that, as the number of layers in such models increases, the amount of information retained about the original input diminshes to a single degree of freedom. We show that a simple change to the network architecture fixes this pathology.

1.4 Attribution

This thesis was made possible (and enjoyable to produce) by the substantial contributions of the many co-authors I was fortunate to work with. In this section, I attempt to give proper credit to my tireless co-authors.

Structure through kernels Section **??** of chapter **??**, describing how symmetries in the kernel of a GP-LVM give rise to priors on manifolds with interesting topologies, is based on a collaboration with David Reshef, Roger Grosse, Josh Tenenbaum, and Zoubin Ghahramani.

Structure Search The research upon which Chapter **??** is based was done in collaboration with James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani, and was published in (Duvenaud et al., 2013). [Joint first author] Myself, James Lloyd and Roger Grosse jointly developed the idea of searching through a grammar-based language of GP models, inspired by Grosse et al. (2012), and wrote the first versions of the code together. James Lloyd ran almost all of the experiments. Carl Rasmussen, Zoubin Ghahramani and Josh Tenenbaum provided many conceptual insights, as well as suggestions about how the resulting procedure could be most fruitfully applied.

Automatic Statistician The work appearing in chapter **??** was written in collaboration with James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, Zoubin Ghahramani, and was published in (Lloyd et al., 2014), in which James Lloyd is joint first author. The idea of the correspondence between kernels and adjectives grew out of discussions between James and myself. James Lloyd wrote most of the code to automatically generate reports, and ran all of the experiments. The text was written mainly by myself, James Lloyd, and Zoubin Ghahramani, with many helpful contributions and suggestions from Roger Grosse and Josh Tenenbaum.

Additive Gaussian processes The work in chapter **??** discussing additive GPs was done in collaboration with Hannes Nickisch and Carl Rasmussen, who developed a richly parameterized kernel which efficiently sums all possible products of input dimensions. My role in the project was to examine the properties of the resulting model, clarify the connections to existing methods, and to create all figures and run all experiments. This work was previously published in (Duvenaud et al., 2011).

Warped Mixtures The work comprising the bulk of chapter **??** was done in collaboration with Tomoharu Iwata and Zoubin Ghahramani, and appeared in (Iwata et al., 2013). Specifically, the main idea was borne out of a conversation between Tomo and myself, and together we wrote almost all of the code together as well as the paper. Tomo ran most of the experiments. Zoubin Ghahramani provided initial guidance, as well as many helpful suggestions throughout the project.

Deep Gaussian Processes The ideas contained in chapter **??** were developed through discussions with Oren Rippel, Ryan Adams and Zoubin Ghahramani, and appear in (Duvenaud et al., 2014). The derivations, experiments and writing were done mainly by myself, with many helpful suggestions by my co-authors.

References

David Duvenaud, Hannes Nickisch, and Carl Edward Rasmussen. Additive Gaussian processes. In *Advances in Neural Information Processing Systems 24*, pages 226–234, Granada, Spain, 2011. (page 7)

David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, June 2013. (page 6)

David Duvenaud, Oren Rippel, Ryan P. Adams, and Zoubin Ghahramani. Avoiding pathologies in very deep networks. Reykjavik, Iceland, April 2014. URL `http://arxiv.org/pdf/1402.5836.pdf`. (page 7)

Roger B. Grosse, Ruslan Salakhutdinov, William T. Freeman, and Joshua B. Tenenbaum. Exploiting compositionality to explore a large space of model structures. In *Uncertainty in Artificial Intelligence*, 2012. (page 6)

Tomoharu Iwata, David Duvenaud, and Zoubin Ghahramani. Warped mixtures for nonparametric cluster shapes. Bellevue, Washington, July 2013. URL `http://arxiv.org/pdf/1206.1846`. (page 7)

James Robert Lloyd, David Duvenaud, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Automatic construction and natural-language description of nonparametric regression models. Technical Report arXiv:1402.4304 [stat.ML], 2014. (page 6)

C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*, volume 38. The MIT Press, Cambridge, MA, USA, 2006. (pages 1 and 2)