Chapter 1

Conclusions and Futue Work

1.1 Future Work

How far can we push the kernel structure discovery procedure outlined in this thesis?

Multiple Dimensions

The experiments in section 1.4 showed that the structure search had very good performance in moderate dimensions.

Input and Output Warpings

1.2 Structured versus Unstructured GP Models

One question left unanswered by this thesis is when to prefer the structured kernel based models described sections 1.4 to 1.4 to the relatively unstructured deep GP models described in section 1.4.

The warped mixture model of section 1.4 represents a compromise between these two approaches, combining a discrete clustering model with an unstructured warping function. However, the results of Damianou and Lawrence (2013) suggest that clustering can be automatically accomplished by a sufficiently deep, unstructured GP.

Difficulty of Optimization

The discrete nature of the search over composite kernel structures can be seen as a blessing and a curse. Certainly, a mixed discrete and continuous optimization requires more complex procedures compared to the continuous-only optimization possible in deep GPs.

However, the discrete nature of the space of composite kernels offers the possibility of learning heuristics to suggest which types of structure to add. For example, finding periodic structure or growing variance in the residuals of a model suggests adding periodic or linear components to the kernel, respectively. It is not clear whether such heuristics can easily be found for optimizing the variational parameters of a deep GP.

Extrapolation

Another question is whether, and how, an equally rich inductive bias can be encoded into relatively unstructured models such as deep GPs. As an example, consider the problem of extrapolating a periodic function. A deep GP could learn a latent representation similar to that of the periodic kernel, projecting into a basis eqeuivalent to $[\sin(x), \cos(x)]$ in the first hidden layer. However, to extrapolate a periodic function, the sin and cos functions would have to continue to repeat beyond the range of the training data, which would not happen if each layer assumed only local smoothness.

One obvious possibility is to marry the two approaches, learning deep GPs with structured kernels. However, we may lose some of the advantages of interpretability by this approach.

Another point to consider is that, in high dimensions, the line between interpolation and extrapolation is blurred, and that learning a suitable representation of the data manifold may be sufficient for most purposes.

Ease of Interpretation

Section 1.4 showed that composite kernels allow automatic visualization and desription of low-dimensional structure. On the other hand, Damianou and Lawrence (2013) showed that deep GP-LVMs allow summarization of high-dimensional structure through showing samples from the posterior, examining the dimension of each latent layer, visualizing latent coordinates, or examining how the predictive distribution changes as one moves in the latent space.

1.3 Automating Model Construction

This thesis is part of a push to automate the practice of model building and inference. Broadly speaking, this problem is being attacked from two directions.

From the top-down, the probabilistic programming community is developing automatic inference engines for extremely broad classes of models (Goodman et al., 2008; Liang et al., 2010; **?**) such as the class of all computable distributions (Li and Vitányi, 2009; Solomonoff, 1964). As discussed in **??**, model construction can be seen as a search through such open-ended model classes. Universal search strategies have been constructed for these very general model classes, (Hutter, 2002; Levin, 1973; Schmidhuber, 2002), but they remain impractically slow.

The bottom-up approach is to design procedures which extend and combine existing model classes, for which relatively efficient inference algorithms are known. For example, Grosse (2014) built an open-ended language of matrix decomposition models and a corresponding compositional language of relatively efficient approximate inference algorithms. This approach makes inference feasible for models in the language, but extending the language requires developing new inference algorithms. The language of models proposed in section 1.4 is an example of this bottom-up approach, and has the same benefits and limitations.

We might consider deep learning an example of this bottom-up approach, although the type of composition used is usually limited to stacking layers, and inference in general is still difficult in these models.

It seems clear that, one way or another, large parts of the existing practice of model-building will eventually be automated. Historically, the statistics community has put much more emphasis on the interpretabilty and meaning of models than the machine learning community, which has focused more on predictive performance. To automate the practice of statistics, developing model-description procedures for powerful model classes seems like the direction with the most low-hanging fruit.

1.4 Summary of Contributions

The main contribution of this thesis is to show how to automate the construction of interpretable nonparametric models using Gaussian processes. This was done in several parts. First, section 1.4 systematically described many different kernel construction techniques, and demonstrated properties of the corresponding GP priors. Next, section 1.4 showed how to automatically search over an open-ended space of GP models, and that those models could be automatically decomposed into diverse parts showing the structure found in the data. Section 1.4 showed that the effect each part of a kernel can be described modularly, allowing automatically written text to be included in detailed reports describing GP models. An example report is included in **??**. Together, these chapters describe the beginnings of an "automatic statistician", capable of the sort of model construction and description currently performed only by experts.

The second half of this thesis develops several extensions to Gaussian processes that can automatically determine modeling choices that were previously set by trial and error or cross-validation. Section 1.4 contains theorems and visualizations characterizing deep Gaussian processes, relates them to existing deep neural networks, and derives novel deep kernels. Section 1.4 investigates additive GPs, a family of structured models consisting of sums of functions of all combinations of input variables, which are shown to have the same covariance as a GP using dropout regularization. Section 1.4 extends the GP latent variable model into a Bayesian clustering model which automatically infers nonparametric shape of each cluster, as well as the number of clusters.

References

Andreas Damianou and Neil D. Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013. (pages 1 and 2)

Noah D. Goodman, Vikash K. Mansinghka, Daniel M. Roy, K. Bonawitz, and Joshua B. Tenenbaum. Church: A language for generative models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 220–229, 2008. (page 3)

Roger B. Grosse. *Model Selection in Compositional Spaces*. PhD thesis, Massachusetts Institute of Technology, 2014. (page 3)

Marcus Hutter. The fastest and shortest algorithm for all well-defined problems. *International Journal of Foundations of Computer Science*, 13(03):431–443, 2002. (page 3)

Leonid A. Levin. Universal sequential search problems. *Problemy Peredachi Informatsii*, 9(3):115–116, 1973. (page 3)

Ming Li and Paul M.B. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 2009. (page 3)

Percy Liang, Michael I. Jordan, and Dan Klein. Learning programs: A hierarchical Bayesian approach. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 639–646, 2010. (page 3)

Jürgen Schmidhuber. The speed prior: a new simplicity measure yielding near-optimal computable predictions. In *Computational Learning Theory*, pages 216–228. Springer, 2002. (page 3)

Ray J. Solomonoff. A formal theory of inductive inference. Part I. *Information and control*, 7(1):1–22, 1964. (page 3)