Contents

Deep GPs (Lawrence)

Full lectore notes: http://inverseprobability.com/talks/notes/deep-gps.html. Includes fully-written out text and references to MacKay tutorials.

At the heart of deep GPs is a variational approximation to $p(f|u)$, where f are the observed data and u are the variational parameters.

Deep GPs allow nonparametric discontinuities. "Nonparametric" because GPs can have discontinuities: for example, just introduce a step function parametrized by a location parameter in the mean function. Deep GPs are able to sidestep the parametrization issue.

You can't write down the likelihood of deep GPs but you can bound them, which is why we're interested in them. This approximation allows you to compose GPs together: composite GPs.

Parametric bottleneck

Assume a general linear model $y = \sum_i w_i \phi_i$

The parameters w of a model are a bottleneck for next predictions.

Fundamental goal of ML: make predictions on test data:

$$p(y'|X', y, X) = \int p(y'|w, X')p(w'|y, X)dw$$

Q: Is the bottleneck idea still valid when the likelihood does not factorize over data points (i.e. non-independent data)?

A: Kind of, but I did not understand his answer.

This parametric bottleneck can be removed with non-parametric models: GP

Degenerate covariance matrices (not full rank): when you include a parametric model (like $y = cx$; the linear kernel). This is the way the bottleneck problem manifests in the general linear model case: overdetermination.

Non-parametric models are the general case: parametric is a special case when conditional prediction can be summarized in *fixed* number of parameters.

Neural networks solve this problem by introducing an enormous amount of parameters, way more than needed, but they lack intrinsic regularization. It turns out that once these neural nets are fitted, the weights W between connected layers tend to be low rank: only a small number of (nodes? weights?) are being actually used.

Augment variable space

Notation: f is what is *fundamental*, *u* is what *you* are going to store about the world.

You can change the size of u (length of the vector) during inference runtime. This is allowed by the fact that u are variational parameters.

Also allows SVI (Hoffman et al. 2013).

Other techniques as GPs

Many ML techniques can be seen as GPs applied in a certain way (certain assumptions) which makes the inference tractable.

- Generalized linear regression: the covariance matrix is low rank since a linear model is being used.
- Kalman filter: exploits sparse precision matrix (which specifies the conditional dependencies).

Approximations to GPs: getting the error bars right

Don't compromise fit to data to get error bars right! GPs work well here.

KL divergence: once you go outside the region where there is observed data you revert to prior. But the KL term in the ELBO contributes only little: “getting error bars right is at best like getting 1 data point right”.

Introduction to Bayesian optimisation (Gonzalez)

Can be seen as intelligently sampling locations x where to evaluate an utility function $f(x)$ to be optimized based on a surrogate model of f denoted $s(x)$ which is less expensive to evaluate than f. For example, $s(x) \sim GP$.

Thompson sampling is a way of balancing exploration/exploitation. Different problems require different utility functions. For example, Thompson sampling performs well for some functions and worse for others.

Connection to multi-armed bandits: if the arms are correlated and infinite in amount, you can translate this to Bayesian optimisation.

Round table with Gonzalez

Q: Has BO been applied to the problem of likelihood-constrained sampling from a likelihood function? In other words, given L, *sample x uniformly from a region L(x) >= L*.

A: Not (really), it seems. This is like a generalization: find a region in x rather than an optimal x. New approach to nested sampling in small dimensions?

Q: How does the (assumed) Lipschitz constant for the utility function affect the lengthscale of the surrogate kernel?

A: There is a definite connection: Rasmussen and Williams (2006).

Q: Is the distribution of x_min on slide 20 available in closed form? Or did you use a density estimator?

A: Unanswered.

Q: How big can the dimensionality of x be?

A: In practice: about 10 or less. Going further than about 10, you need to compensate with more prior information or parallelization.

Q: Good book for BO?

A: There are a couple of tutorials but not a comprehensive book yet. He recommends https://ieeexplore.ieee.org/document/7352306.

Bayesian Neural Networks from a GP Perspective (Wilson)

Q: Example of when infinite (nonparametric) models are undesirable?

A: Dirichlet process: sometimes too many clusters. But in general they are desirable.

Q: Should we change our model as a function of the amount of data we receive?

A: Typically: no: it should probably not depend on it. This is an argument favoring nonparametric distributions.

Loss landscapes

Amazingly, the best basins are often

- geometrically similar
- have similar densities

but may contain different amounts of mass, which is how the contributions should weighed, **yet contain complementary solutions**. This means that the solutions they represent are not similar at all and do represent complementary approaches to the problem.

https://losslandscape.com/

Tempering and VI

Q: It seems weird to "mess" with the product rule by inserting this T power. Is it a form of VI, where T is a variational parameter?

A: A tempering term is often included in variational approximations; it is being used there.

Hoffman, Matthew D, David M Blei, Chong Wang, and John Paisley. 2013. "Stochastic Variational Inference." *The Journal of Machine Learning Research* 14 (1): 1303–47.

Rasmussen, Carl Edward, and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, Mass: MIT Press.