Learning Exponential Families

Anonymous Author(s)
Affiliation
Address
email

Abstract

[SLOPPY NOTES STAGE JUST TO GET THOUGHTS DOWN]
Recently much attention has been paid to probabilistic models defined by a deep neural network transformation of a simpler random variable; these implicit generative models have been used to great success across variational inference, generative modeling of complex data types, and more. In essentially all of these settings, the model is specified by the network architecture, and a particular member of that model is chosen to minimize some loss (be it adversarial or information divergence)
We treat the problem of learning an exponential family – the model itself, rather than the typical setting of learning a particular member of that model.
Many intractable distributions encountered in machine learning belong to exponential families. In rare cases these distributions are tractable due to either known conjugacy in the problem setup (such as the normal-inverse-Wishart), or due to careful numerical work historically that has made these distributions computationally indistinguishable from tractable (eg the Dirichlet).

1 Introduction

People use lots of implicit generative models:

Across machine learning, including ABC [**?**] , GANs [1], VAEs [2, 3], and their many follow-ons (too numerous to cite in any detail), models that specify a distribution via the nonlinear transformation of latent random variable. We prefer and use the terminology of [4], calling such a distribution an *implicit generative model*, defined as:

something like eq 1 and 2 in Mohamed:2016aa, defining$q_\theta(z)$

Also use the proper notation of the density implied by the pushforward measure of the function $f_{\theta\sharp}$ if useful. Also reference to this being super standard and widespread [5]. The two central uses are at present generative distributions of interesting data types (as in GANs), and for variational inference Regardless, all of these use cases specify a *model* (or variational family) $\mathcal{Q} = \{q_\theta : \theta \in \Theta\}$, and then minimize a suitable loss $\mathcal{L}(q, p)$ over $q \in \mathcal{Q}$. In the case of VI p is the posterior (or the unnormalized log joint) and $\mathcal{L}$is the KL divergence (or so called ELBO), in GAN p is the sample density of a (large) dataset and $\mathcal{L}$ is the adversarial objective whose details do not matter here.

All these learn a single member of a family

Inherent in all the above approaches is an algorithmic procedure to select a *single* distribution $q_\theta(z)$ from among the *model* $\mathcal{Q}$. Implicit in this effort is the belief that $\mathcal{Q}$ is suitably general to contain the true distribution of interest, or at least an adequately close approximation.

Here we learn the family

We leverage the natural parameterization of exponential families to derive a novel objective that is amenable to stochastic optimization.

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

A note on amortization

Several have pointed out that these IGMs are in fact strictly less expressive than a mean field, at least in the conventional VI setting. See for example http://dustintran.com/blog/variational-auto-encoders-do-not-train-complex-generative-models (here I like the line "The neural network used in the encoder (variational distribution) does not lead to any richer approximating distribution. It is a way to amortize inference such that the number of parameters does not grow with the size of the data (an incredible feat, but not one for expressivity!) (Stuhlmuller et al., 2013)"). You have to optimize for every data point individually, or instead you get to do so in aggregate once in advance (at a much higher cost) and then recover that cost over future data points within that distribution (and hence the term amortization, though perhaps there is shared statistical power as well) Etc etc what we are doing here is *amortized* amortized inference, in the sense that we are amortizing not the data points, but the distribution itself.

Our contributions include:

...

Our results demonstrate

...

2 Learning exponential families

Why this is important

Exp fams are awesome and fundamental []. Also [**?**] rightly point out that many many inference problems can be cast as exponential families. Can we cast the VAE encoder network as a suitable exp fam... sure I think that's right; the network parameters of z form the statistics, and then the observations are eta's.

Why this is coherent

Θ defines quite a big $\mathcal{Q}$, and indeed the subject of compressibility, generalization, etc is of keen interest to many [**?**]. So actually the space of distributions is quite large, and in many cases certainly larger than it needs be. Why? Well, we know precisely the parameter space of the exponential family; it is defined by the *natural* parameters $\eta \in \mathbb{R}^p$ (or whatever we choose there).

Figure 1

Figure of model space. Yeah that's good. Then graphical model. Note that perhaps $\mathcal{Q}$ is too big, and a simpler model space (the $\|\eta\|$ dimensional subspace of Θ) would be better for the usual robustness/generalization reasons.

Aside

A neat idea is to ask if learning the $\theta(\eta)$ network leads to better VI in terms of inference networks, since it is apparently appropriately regularized and can just take suff stats. That's testable if we have time.

In many situations, statistical inference attempts to learn, at least approximately, a member of an exponential family. We often consider this exponential family intractable in the sense that we don't know how to normalize or sample from it. Approximate inference, such as variational

3 To Do

3.1 SRB

- set up submission at `https://cmt.research.microsoft.com/NIPS2018/`
- review and conform to style requirements (see website with template); 8 pages not including refs and acks and appendices.

3.2 JPC

- Outline

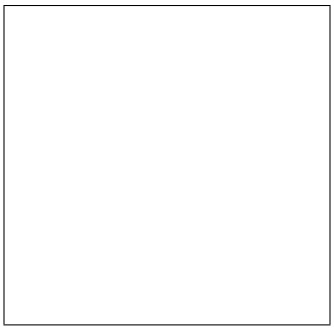

Figure 1: Sample figure caption.

Table 1: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

- Write

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. 12 2013.

[3] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[4] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. 10 2016.

[5] Luc Devroye. *Non-uniform random variate generation*. Springer-Verlag, New York, 1986.