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.

REparameterization trick (Kingma and Welling (2013), Rezende et al. (2014) and Titsias and Lazaro-Gredilla 2014).. See also Archer 2015 / Gao 2016 for clean explanation.

Key for obvious norm flow connection but also a good bibliography and some good historical views to Dayan and Gershman and other people who did norm flows. https://arxiv.org/pdf/1505.05770.pdf

Our contributions include:

...

This should not be confused with "Learning to learn by gradient descent by gradient descent" (Andrycowicz et. al. 2016) and similar works.

...

Important to distinguish carefully from VI. In a sense VI does parameterize a family: given data, you get local variational parameters and that parmaterizes a density (like a regular VAE). Inference networks are exclusively used to data to amortize with a global set of parameters a variational distribution, not a model. Of course it is in a sense a model, but that's a bunch of normals. The sampling mechanism is easy (Guassian).

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.

Why Flow Networks

We choose flow networks [] and [] because duh. And "implicit generative models aka density networks" (or rather, density networks are the instantiation of an IGM with deep nets, which is effectively synonymous these days. Gibbs and MacKay Density Networks 1997! And invertible networks In that vein probably definitely cite invertible deep nets in general: Baird et al IJCAI 2005, Ripple and Adams 2013 . Note that what norm flows (the Rezende/Mohamed stuff specifically) did is make it tractable and scalable and in the modern VAE style. That makes these comparisons legitimate and apples to apples. Any generalization of this is also dandy though, so could use a mean field approach (standard) or any of the things that go beyond mean field, either classically *(Saul and Jordan, 1996; Barber and Wiegerinck, 1999); this is called structured variational inference. Another way to expand the family is to consider mixtures of variational densities, i.e., additional latent variables within the variational family (Bishop et al., 1998).* or newer stuff [] [Tran Copula VI, Hoffman and Blei 2015].

As noted in norm flows paper: "The true posterior distribution will be more com- plex than this assumption allows for, and defining multi- modal and constrained posterior approximations in a scalable manner remains a significant open problem in varia- tional inference."

Couch this in terms of normalizing flows though point out this is not strictly necessary. Note in particular Tabak, E. G. and Turner, C. V. A family of nonparametric density estimation algorithms. Communications on Pure and Applied Mathematics, 66(2):145?164, 2013. Tabak, E. G and Vanden-Eijnden, E. Density estimation by dual ascent of the log-likelihood. Communications in Mathematical Sciences, 8(1):217?233, 2010. A nice line from Rezende and Mohamed is: Thus, an ideal family of variational distributions q(z|x) is one that is highly flexible, preferably flexible enough to contain the true posterior as one solution. One path towards this ideal is based on the principle of nor- malizing flows (Tabak Turner, 2013; Tabak VandenEijnden, 2010).

Related work / How close is this to norm flows or VAE

In a restricted technical sense, rather close: VAE and other black box VI that uses reparameterization results in a conditional density $q_\phi(z|x)$. If we consider η as x, then sure yes the previous stuff specifies a model $\mathcal{Q}_{VAE} = \{q_\phi(z|x) : x \in X\}$. But that's a little pedantic, and any way that is very often a normal family with variational parameters specified by (a deep function of) x. Much closer is Figure 2 in Rezende and Mohamed, where like here they use a network to index the *parameters* of the normalizing flow. In that case it's a function of x the observation, and as such that network is an inference network; here it's a function of η and as such is a parameter network. That's just nomenclature, so naturally the next question is do they differ at some other level. Yes, distinctly. The other term implied in a VI (or norm flow VAE style as they use) is the expected log joint $E_{q_{\phi(x)}}(\log p_\theta(x,z))$. Now sure that's a loss function on x, z, so then when we look at that same term in EFN we see $E_{q_{\phi(\eta)}}\left(\eta^\top t(z)\right)$, which sure also looks like a loss function on η, z. And yes, they are both unnormalized (in the sense that VI is an ELBO / joint $p(x,z)$ and EFN lacks the normalizer because it's constant, so we're not getting a KL estimate). A picky difference is that the exp family doesn't really correspond to a proper unnormalized log joint (though I suppose it could), as there is not a prior on η in the objective (but is that just ignoring $p(\eta)$ in our sampling scheme)? But yes in general we could see this as a specific case where $x = \eta$ and thus we are learning a family just as in the inference case. And for example we are giving in the bayesian logistic regression example full datasets for inference instead of single data points. First, then we need to produce really compelling results focusing on when learning an exp fam is key. Second we need some very tight language to draw this distinction without seeming a small tweak on normalizing flows. One way to do this is the restricted model class argument, a la Fig 7.2 in Hastie and Tibshirani. Another is to actually produce a conditional exp fam, as in something indexed on both x and η. Third, possible novelties in norm flows, like triple spinners or other better choices than planar flows (yuck).

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.

Figure 1: Figure 1: possibly Fig 7.2 bias-variance tradeoff and then benefit of a restricted model from Hastie Book, or similar from W4400 (ask PO for batman permission).

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

3.2 JPC

- Outline
- 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.

Stuff on wake sleep and the Helmholtz machine

Stuff on sampling from Gibbs distributions (max ent models), and sampling from exp fams generally, with MCMC and such.

Flow networks

Devroye's book.

Hoffman et al 2013 SVI

From Blei review on VI. ThedevelopmentofvariationaltechniquesforBayesian inference followed two parallel, yet separate, tracks. Peterson and Anderson (1987) is arguably the first variational procedure for a particular model: a neural network. This paper, along with insights from statistical mechanics (Parisi, 1988), led to a flurry of variational inference procedures for a wide class of models (Saul et al., 1996; Jaakkola and Jordan, 1996, 1997; Ghahramani and Jordan, 1997; Jordan et al., 1999). In parallel, Hinton and Van Camp (1993) proposed a variational algorithm for a similar neural network model. Neal and Hinton (1999) (first published in 1993) made important connections to the expectation maximization (EM) algorithm (Dempster et al., 1977), which then led to a variety of variational inference algorithms for other types of models (Waterhouse et al., 1996; MacKay, 1997).

Salimans, T. and Knowles, D. (2014). On using control variates with stochastic approximation for variational Bayes. arXiv preprint arXiv:1401.1022.

Salimans, T., Kingma, D., and Welling, M. (2015). Markov chain Monte Carlo and variational inference: Bridging the gap. In International Conference on Machine Learning, pages 1218? 1226.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In Artificial Intelligence and Statistics.

Hoffman, M. D. and Blei, D. M. (2015). Structured stochastic variational inference. In Artificial Intelligence and Statistics.

Possibly some of Burda, Y., Grosse, R., & Salakhutdinov, R. (2016). Importance Weighted Autoencoders. In International Conference on Learning Representations. Damianou, A. C., & Lawrence, N. D. (2013). Deep Gaussian Processes. In Artificial Intelligence and Statistics. Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz Machine. Neural Computation, 7(5), 889?904. http://doi.org/10.1162/neco.1995.7.5.889 Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2016). Density estimation using Real NVP. arXiv.org. Harville, D. A (1977). Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association, 72(358):320?338. Hinton, G. and Van Camp, D (1993). Keeping the neural networks simple by minimizing the description length of the weights. In Computational Learning Theory, pp. 5?13. ACM. Johnson, M. J., Duvenaud, D., Wiltschko, A. B., Datta, S. R., & Adams, R. P. (2016). Composing graphical models with neural networks for structured representations and fast inference. arXiv.org. Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. In International Conference on Learning Representations. Kingma, D. P., Salimans, T., & Welling, M. (2016). Improving Variational Inference with Inverse Autoregressive Flow. arXiv.org. Louizos, C., & Welling, M. (2016). Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors. In International Conference on Machine Learning. Maaloe, L., Sonderby, C. K., Sonderby, S. K., & Winther, O. (2016). Auxiliary Deep Generative Models. In International Conference on Machine Learning. MacKay, D. J., & Gibbs, M. N. (1999). Density networks. Statistics and neural networks: advances at the interface. Oxford University Press, Oxford, 129-144. Mnih, A., & Rezende, D. J. (2016). Variational inference for Monte Carlo objectives. In International Conference on Machine Learning. Ranganath, R., Tran, D., & Blei, D. M. (2016). Hierarchical Variational Models. In International Conference on Machine Learning. Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In International Conference on Machine Learning. Salimans, T., Kingma, D. P., & Welling, M. (2015). Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. In International Conference on Machine Learning. Salakhutdinov, R., Tenenbaum, J. B., and Torralba, A (2013). Learning with hierarchical-deep models. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35 (8):1958?1971. Stuhlmuller, A., Taylor, J., & Goodman, N. (2013). Learning Stochastic Inverses. In Neural Information Processing Systems. Tran, D., Blei, D. M., & Airoldi, E. M. (2015). Copula variational inference. In Neural Information Processing Systems. Tran, D., Ranganath, R., & Blei, D. M. (2016). The Variational Gaussian Process. International Conference on Learning Representations. Waterhouse, S., MacKay, D., and Robinson, T (1996). Bayesian methods for mixtures of experts. In Neural Information Processing Systems.