

---

# 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  $f_{\theta\#}$  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.

## 31 *A note on amortization*

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

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

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

47 *Our contributions include:*

48 ...

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

51 ...

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

57 *Our results demonstrate*

58 ...

## 59 **2 Learning exponential families**

60 We are interested in perhaps the most classic inference problem:

$$p(z|x) \propto p(z) \prod_{i=1}^n p(x_i|z)$$

61 shown with the attached plate model (not local latents). Supposing as is often the case that the  
 62 likelihood is a member of the  $s$  exp fam, we have:

$$p(z|x) \propto \exp \left\{ \left[ \sum_{i=1}^n s(x_i) \right]^\top [t(z)] + g_0(\alpha, z) \right\}$$

63 where the natural parameters of the sampling distribution are indexed by the latent parameter on  
 64 which we want to inference ( $z$ ). Here I've written the prior as arbitrary, and possibly not exp fam,  
 65 which is fine, since this is still an exp fam in the sense of, for a fixed  $\alpha$ , the function  $g_0$  can just be  
 66 viewed as a sufficient statistic. Even if  $\alpha$  is not fixed though, we can sample over that too to learn the  
 67 whole fam (but maybe not if we want to infer it?). Regardless, life is simpler to make sense of if we  
 68 take an exp fam prior  $g_0(\alpha, z) = \alpha^\top t_0(z)$ , and then the desired posterior is an intractable exp fam,  
 69 but still just an exp fam.

70 Note: consider changing all  $z$  to  $\theta$  to remind the average reader that we're doing real bayesian  
 71 inference and not just run of the mill VI with local latents in a nonlinear dimension reduction setting.

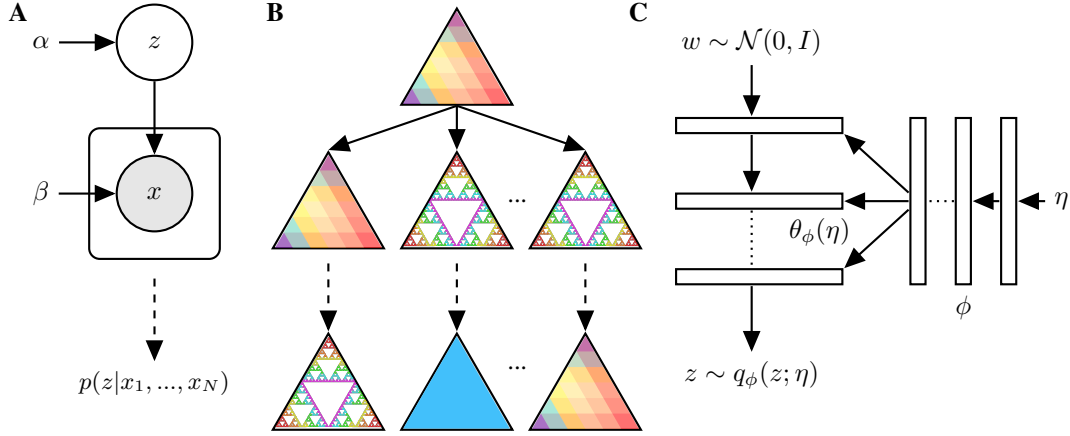


Figure 1: Learning exponential families. A shows the graphical model, emphasizing conditional iid sampling. B shows Dirichlet prior (a density), conditional Dirichlet observations (some observed points in the simplex), and then the posteriors learned by an EFN. SRB to fill in these triangles. C shows the EFN network schematic.

72 Perhaps an important reminder that most all of VAE and such are for inference of local latents, and  
73 that's a little bit too bad. We fix that.

74 *Why this is important*

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

79 *Why this is coherent*

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

84 Note somewhere that the natural parameter space needs to be considered in general. That is, not all  $\eta$   
85 lead to a valid distribution (standard fact, see for example Wainwright and Jordan 08). In practice  
86 that's not often a problem, as the space is known for most distributions one uses, and when one  
87 composes them in a posterior scheme (for example), this is inherited (eg the normal covariance...  
88 So we skip that here. But yes in general that needs to be considered.

89 *Figure 1*

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

93 *Aside*

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

97 *Why Flow Networks*

98 We choose flow networks [] and [] because duh. And "implicit generative models aka density  
99 networks" (or rather, density networks are the instantiation of an IGM with deep nets, which is  
100 effectively synonymous these days. Gibbs and MacKay Density Networks 1997! And invertible  
101 networks In that vein probably definitely cite invertible deep nets in general: Baird et al IJCAI 2005,  
102 Ripple and Adams 2013 . Note that what norm flows (the Rezende/Mohamed stuff specifically)

103 did is make it tractable and scalable and in the modern VAE style. That makes these comparisons  
104 legitimate and apples to apples. Any generalization of this is also dandy though, so could use a mean  
105 field approach (standard) or any of the things that go beyond mean field, either classically (*Saul*  
106 *and Jordan, 1996; Barber and Wiering, 1999*); this is called *structured variational inference*.  
107 Another way to expand the family is to consider mixtures of variational densities, i.e., additional  
108 latent variables within the variational family (*Bishop et al., 1998*). or newer stuff [] [Tran Copula  
109 VI, Hoffman and Blei 2015].

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

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

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

122 In a restricted technical sense, rather close: VAE and other black box VI that uses reparameterization  
123 results in a conditional density  $q_\phi(z|x)$ . If we consider  $\eta$  as  $x$ , then sure yes the previous stuff  
124 specifies a model  $\mathcal{Q}_{VAE} = \{q_\phi(z|x) : x \in X\}$ . But that's a little silly, and any way that is very  
125 often a normal family with variational parameters specified by (a deep function of)  $x$ . Much closer  
126 is Figure 2 in Rezende and Mohamed, where like here they use a network to index the *parameters*  
127 of the normalizing flow. In that case it's a function of  $x$  the observation, and as such that network  
128 is an inference network; here it's a function of  $\eta$  and as such is a parameter network. That's just  
129 nomenclature, so naturally the next question is do they differ at some other level. Yes, distinctly.  
130 The other term implied in a VI (or norm flow VAE style as they use) is the expected log joint  
131  $E_{q_\phi(x)}(\log p_\theta(x, z))$ . Now sure that's a loss function on  $x, z$ , so then when we look at that same  
132 term in EFN we see  $E_{q_\phi(\eta)}(\eta^\top t(z))$ , which sure also looks like a loss function on  $\eta, z$ . And yes,  
133 they are both unnormalized (in the sense that VI is an ELBO / joint  $p(x, z)$  and EFN lacks the  
134 normalizer because it's constant, so we're not getting a KL estimate). A picky difference is that  
135 the exp family doesn't really correspond to a proper unnormalized log joint (though I suppose it  
136 could), as there is not a prior on  $\eta$  in the objective (but is that just ignoring  $p(\eta)$  in our sampling  
137 scheme?). But yes if we want to be reductionist and pedantic [use nicer words] in general we could  
138 see this as a specific case where  $x = \eta$  and thus we are learning a family just as in the inference  
139 case. Or rather, we are putting the data in as sufficient stat (computation of natural parameters),  
140 but that's nonobvious. And for example we are giving in the bayesian logistic regression example  
141 full datasets for inference instead of single data points. To make this as close as possible, we write  
142  $p(\eta|z) = \frac{1}{A(t(z))} \exp\{\eta^\top t(z)\}$ . That's the "likelihood" of an EFN in some wonky sense. So this  
143 reveals the mechanical differences: first,  $t(z)$  is not a deep generative model with parameters  $\theta$ , but  
144 rather it is a fixed set of sufficient statistics that define the exp fam. Next, there is no clear prior  $p(z)$ ,  
145 which is critical to understanding how VI behaves (see Hoffman and Johnson ELBO surgery paper,  
146 also Duvenaud's <https://arxiv.org/pdf/1801.03558.pdf>). So yes there is a hand wavy sense in which  
147 EFN is a specific case of norm flow, but of course it is. And anyway norm flow is a specific case of a  
148 DNN architecture or Helmholtz machine or deep density network (Ripple and Adams). This is just  
149 rambling but good to have all perspective here. Ok so what to do? First, then we need to produce  
150 really compelling results focusing on when learning an exp fam is key. Second we need some very  
151 tight language to draw this distinction without seeming a small tweak on normalizing flows. One way  
152 to do this is the restricted model class argument, a la Fig 7.2 in Hastie and Tibshirani. Another is to  
153 actually produce a conditional exp fam, as in something indexed on both  $x$  and  $\eta$ . Third, possible  
154 novelties in norm flows, like triple spinners or other better choices than planar flows (yuck).

155 Another related work is that this is somehow the dual of MEFN, or a generalization of the dual  
156 problem. In the wainwright and jordan sense of forward and backward mappings.

### 3 Results

#### Chapter 1, Fig 1

Toy figure that demonstrates what we are doing and a simple example. Note this should probably not be in Results but in the EFN section or similar. Ideas:

- value of a restricted model, see hastie tibshirani fig 7.2, or porbanz's batman version from 4400 slides. ... well that's a bit off topic. At least worth a mention in motivation.
- graphical model. yeah probably needed.
- network model. yeah probably needed.
- cartoon example three sets of natural parameters in, three dirichlet distributions out. Or similar.

#### Chapter 2: Fig 2 and 3 and 4 Ground truth toy examples, etc.

Figure 2.

Single EFN:

Panel A:  $r^2$  throughout training

Panel B: KL throughout training

Panel C: Distributin of MMD p values

Figure 3:

EFN performance by dimensionality

Panel A: Dir KL for NF1 and EFN

Panel B NIW KL for NF1 and EFN

Panel C: Gaussian KL for NF1 and EFN

Note Number of panar flows is always D (intrinsic dimensionality of flows), units per layer ramping is always the same function of D. The number of layers in the theta network is always a function of D - will probably just always use 8 layers.

Fig 4. [This idea was Fig 5 in disguise; see below. Currently no need for this figure].

#### Chapter 3: Fig 5 and 6

Fig 5. The intractable posterior inference example. **Key real data result.** Learn the full posterior family for some problem (see ideas below). Then get some data  $X$ . Then find the posterior distribution for that data by indexing the natural parameters (as in, just plugging in the correct choice of  $\eta$ , which is after all some function of the prior and  $X$ ). That gives the EFN posterior  $q(z|X)$ . (Possible preceding figure: show its properties, show a low-d picture, show its non-Gaussianity). Now, as Alternative 1 do full norm flow variational inference (explore all of  $\phi$  space with the full flow network model  $\mathcal{Q}$ ), which is to say  $\arg \min_{\phi} KL(q_{\phi}||p)$ : the key difference here is that, while you have the *same exact* flow network architecture, now you have to optimize over  $\phi$  with a limited single dataset. As Alternative 2, be literal to Figure 2 of the Norm Flow VI paper, give the sufficient statistics of that K=1 dataset, and learn an EFN from scratch. This alternative is important because it is the most specific (but kind of annoying, hence alternative 1) interpretation of norm flow VI paper.

Now, PANEL A of this figure shows performance as a size of the dataset. This will likely show that when the dataset gets small, this "traditional VI" will get arbitrarily bad (can't learn a network); eventually, there will be so much data that the VI will match or outperform the EFN... outperform because VI can focus specifically on this distribution rather than over the whole family, so the EFN has less effective data for this  $\eta$  (but not because it has a broader range of models, since we believe the EFN contains the closest member). Alternative 2 should do shittier across the board than alt 1, I think? Performance metric should be ELBO on some held out data or something like that (it's a posterior, so log likelihood doesn't really make sense). Test data anyway. Check VI papers for usual metrics. PANEL B of this figure shows performance as dimension of the problem grows. Pick some middle dataset size, then repeat same performance metric as in Panel A for a range of dimensionalities of the exponential family. VI will generalize to test data worse and worse as dimensionality grows,

but EFN will learn the family less well on its computational budget. This could go either way but will be interesting regardless. I suppose we should also have those panels for training data. A key point to make here is that one great virtue of EFNs is learning a restricted model, which should demonstrate the usual bias-variance tradeoff (see for example Hastie and Tibshirani book, Fig 7.2). Maybe that's Panel A. Or Figure 4 is bias-variance and some sample posteriors in 2-d (showing how nicely it works), and then Fig 5 is the above performance, with both train and test. Notice one pain here is that Panel B requires training a new EFN at every dimensionality. Sorry.

This will be for one real example  $X$ . As such, to get error bars, just take a big dataset and randomly subsample. Then the posterior performance is really for that very dataset, so the sem is coherent and the right thing to calculate/show. Important to clarify that doing so *does not* test how well this does across the entire exp fam, but just this one posterior. To test that, we do it in simulation: generate *many datasets*  $X$ , then do the above for every one of them. Same computation for EFN (since its just plugging in a dataset), but VI alternatives 1 and 2 now need to be rerun for every dataset. And it's still simulated data, not really offering something fundamentally more than Fig 3 (well ok it's an intractable model, but I'm not sure that offers so much).

Fig 5. Heaps of examples with conditional iid exp fams. Math details of that pending. Some cool examples:

- Censored data. normal prior, censored normal observations, what is posterior distribution on mean? Lots of work in that.
- Truncated data. truncated mvn prior, with some observations thereafter, what is posterior? (Does this work?...)
- Poisson/Bern "process" data. Phony process like in neuro, normal prior on log intensity (ooh maybe that's not an exp fam prior), then a "spike train" of bern or poisson count observations
- multivariate t with inverse wishart prior or something like that. That's neat but doesn't have great "oh yeah people do care about that problem" recognition. Seems contrived.
- check MKB book for other cool MV distributions. (Marshall-Olkin)... seems contrived.
- Elliptically contoured prior with some conditionally iid exp fam observations. People in ML like elliptical distributions.
- von Mises-Fisher distribution, eg <http://www.jmlr.org/papers/volume6/banerjee05a/banerjee05a.pdf> or <https://arxiv.org/pdf/1605.00316.pdf>, but again not clustering (see below), since it's a local latent variable problem then.s
- Note: a whole heap of models don't quite fit comfortably here.
  - Bayesian Logistic Regression. This is an intractable exp fam in the desired sense, but the natural parameter (when parameterized) depends on  $x_i$ . Thus, it grows with every datapoint, or put differently it's a diff exp fam for every dataset. No bueno. This is then true of GLMs, so those are out too.
  - Latent Dirichlet Allocation. Local variational parameters mean that the exp fam grows with datasize. That means that the posterior is already too big for uninteresting sizes of LDA. This is then true of hierarchical models with local latent variables in general.

Fig 6. The Killer real data. Perhaps Gibbs or Markov Random Field. Learn it, then pick some  $\eta$ , then show samples from it. Can this look interesting? Some thoughts...

Criteria:

- Needs to be an exp fam.
- Needs to be a forward exp fam. As in, not fit to data, because we don't have  $\mu$  parameters, we have  $\eta$  parameters.
- "real data" is a misnomer, since we are not doing VI or similar. Really we want an exp fam that is real and somehow useful in its own right, and that people want to sample from.
- Reminder: we will *always* be comparing to "well normally you can do this with learning a *single* distribution in the  $\min KL(q||p)$  sense. That's fine. The point is we can learn the whole family, then choose and sample, vs just one by one.

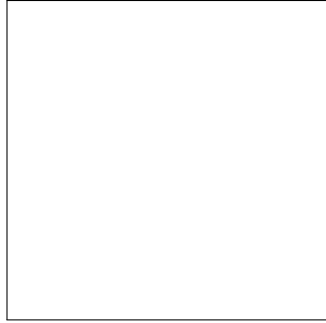


Figure 2: Figure 1: possibly Fig 7.2 bias-variance tradeoff and then benefit of a restricted model from Hastie Book, or similar from W4400 (ask PO for batman permission).

Table 1: Sample table title

Part		
Name	Description	Size ( $\mu\text{m}$ )
Dendrite	Input terminal	$\sim 100$
Axon	Output terminal	$\sim 10$
Soma	Cell body	up to $10^6$

- something hard to sample will be key, since the "toy" results will have used things we already "know" how to sample, like NIW or Dirichlet.

#### Ideas:

- Fancy Exp Fam like Marshall-Olkin. Yeah but who really cares about this esoteric distribution? It doesn't look cool visually either.
- Ising models: classic, bw images, but gross NP-Hard Cooper 1990.
- Potts model: great because failure of MCMC (Gibbs sampling) here is at least locally well known from Geman and Geman 1984 through Sudderth correcting this (see Gibbs sampler slides from Advanced ML, Peter's part). But that is kind of a failure example, not an interesting one (MRFs are smoothness prior, not segmentation prior). Also both Potts and Ising are NP-hard Cooper 1990 The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks
- Markov Random Fields / Gibbs Random Fields (same, by Hammersley Clifford theorem). Yes this is cool: image distributions, texture distributions. Can show wild diff sets of textures, none of which require any sampling or any such thing. Can we make this super intractable from an MCMC perspective? Need to read on how sampling is done there. Erik Sudderth and his phd thesis are likely good resources.
- Gatys and Simoncelli texture stuff (see for example MEFN paper for refs); those are interesting distributions on textures, or specified moments. Can then just sample from this family.

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

## 280 4 Appendix

281 Exponential form of posterior for Dirichlet-Dirichlet

282  $\mathbf{z} \sim \text{Dir}(\boldsymbol{\alpha}_0)$

283  $\mathbf{x}_i \sim \text{Dir}(\beta \mathbf{z})$

284  $p(\mathbf{z}) \propto \exp(\boldsymbol{\alpha}_0^T \log(\mathbf{z}) - \sum_{d=1}^D \log(z_d))$

285  $p(\mathbf{x}_i | \mathbf{z}) \propto \exp(\beta \mathbf{z}^T \log(\mathbf{x}_i) - \sum_{d=1}^D \log(x_{i,d}) - (\sum_{d=1}^D \log(\Gamma(\beta z_d)) - \log(\Gamma(\beta \sum_{d=1}^D z_d))))$

286  $p(X | \mathbf{z}) \propto \exp(\beta \mathbf{z}^T [\sum_{i=1}^N \log(\mathbf{x}_i)] - \sum_{i,d=1}^{N,D} \log(x_{i,d}) - N(\sum_{d=1}^D \log(\Gamma(\beta z_d)) - \log(\Gamma(\beta \sum_{d=1}^D z_d))))$

287

288  $p(\mathbf{z} | X) \propto p(\mathbf{z})p(X | \mathbf{z})$

289  $\propto \exp(\boldsymbol{\alpha}_0^T \log(\mathbf{z}) - \sum_{d=1}^D \log(z_d))$

290  $\exp(\beta \mathbf{z}^T [\sum_{i=1}^N \log(\mathbf{x}_i)] - \sum_{i,d=1}^{N,D} \log(x_{i,d}) - N(\sum_{d=1}^D \log(\Gamma(\beta z_d)) - \log(\Gamma(\beta \sum_{d=1}^D z_d))))$

291 We don't care about the term that just has  $x$  in it.

292  $p(\mathbf{z} | X) \propto \exp(\boldsymbol{\alpha}_0^T \log(\mathbf{z}) + \beta [\sum_{i=1}^N \log(\mathbf{x}_i)]^T \mathbf{z} - \sum_{d=1}^D \log(z_d) - N(\sum_{d=1}^D \log(\Gamma(\beta z_d)) - \log(\Gamma(\beta \sum_{d=1}^D z_d))))$

293  $p(\mathbf{z} | X) \propto \exp\left(\begin{pmatrix} \boldsymbol{\alpha}_0 - \mathbf{1} \\ \sum_{i=1}^N \log(\mathbf{x}_i) \\ -N \end{pmatrix}^T \begin{pmatrix} \log(\mathbf{z}) \\ \beta \mathbf{z} \\ \log(\Gamma(\beta \mathbf{z})) \\ \log(\Gamma(\beta \sum_{d=1}^D z_d)) \end{pmatrix}\right)$

294 This seems right to me. I moved  $\beta$  for the second element of the natural parameters to be over with  
295 his other  $\beta$ -friends in the sufficient statistics.

296 Here's a more cleaned up version:

$$p(\mathbf{z} | X) \propto \exp \left\{ \begin{bmatrix} \boldsymbol{\alpha}_0 - \mathbf{1} \\ \sum_{i=1}^N \log(\mathbf{x}_i) \\ -N \mathbf{1} \\ -N \end{bmatrix}^\top \begin{bmatrix} \log(\mathbf{z}) \\ \beta \mathbf{z} \\ \log(\Gamma(\beta \mathbf{z})) \\ \log(\Gamma(\beta \mathbf{1}^\top \mathbf{z})) \end{bmatrix} \right\} \triangleq \exp \{ \boldsymbol{\eta}^\top t(\mathbf{z}) \}$$



## 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. The development of variational techniques for Bayesian 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*,

344 pp. 5713. ACM. Johnson, M. J., Duvenaud, D., Wiltchko, A. B., Datta, S. R., & Adams, R. P.  
 345 (2016). Composing graphical models with neural networks for structured representations and fast  
 346 inference. arXiv.org. Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. In  
 347 International Conference on Learning Representations. Kingma, D. P., Salimans, T., & Welling, M.  
 348 (2016). Improving Variational Inference with Inverse Autoregressive Flow. arXiv.org. Louizos, C.,  
 349 & Welling, M. (2016). Structured and Efficient Variational Deep Learning with Matrix Gaussian  
 350 Posteriors. In International Conference on Machine Learning. Maaloe, L., Sonderby, C. K., Sonderby,  
 351 S. K., & Winther, O. (2016). Auxiliary Deep Generative Models. In International Conference  
 352 on Machine Learning. MacKay, D. J., & Gibbs, M. N. (1999). Density networks. Statistics and  
 353 neural networks: advances at the interface. Oxford University Press, Oxford, 129-144. Mnih, A., &  
 354 Rezende, D. J. (2016). Variational inference for Monte Carlo objectives. In International Conference  
 355 on Machine Learning. Ranganath, R., Tran, D., & Blei, D. M. (2016). Hierarchical Variational  
 356 Models. In International Conference on Machine Learning. Rezende, D. J., Mohamed, S., & Wierstra,  
 357 D. (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In  
 358 International Conference on Machine Learning. Salimans, T., Kingma, D. P., & Welling, M. (2015).  
 359 Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. In International Conference  
 360 on Machine Learning. Salakhutdinov, R., Tenenbaum, J. B., and Torralba, A (2013). Learning with  
 361 hierarchical-deep models. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35  
 362 (8):1958-1971. Stuhlmüller, A., Taylor, J., & Goodman, N. (2013). Learning Stochastic Inverses.  
 363 In Neural Information Processing Systems. Tran, D., Blei, D. M., & Airoldi, E. M. (2015). Copula  
 364 variational inference. In Neural Information Processing Systems. Tran, D., Ranganath, R., & Blei, D.  
 365 M. (2016). The Variational Gaussian Process. International Conference on Learning Representations.  
 366 Waterhouse, S., MacKay, D., and Robinson, T (1996). Bayesian methods for mixtures of experts. In  
 367 Neural Information Processing Systems.