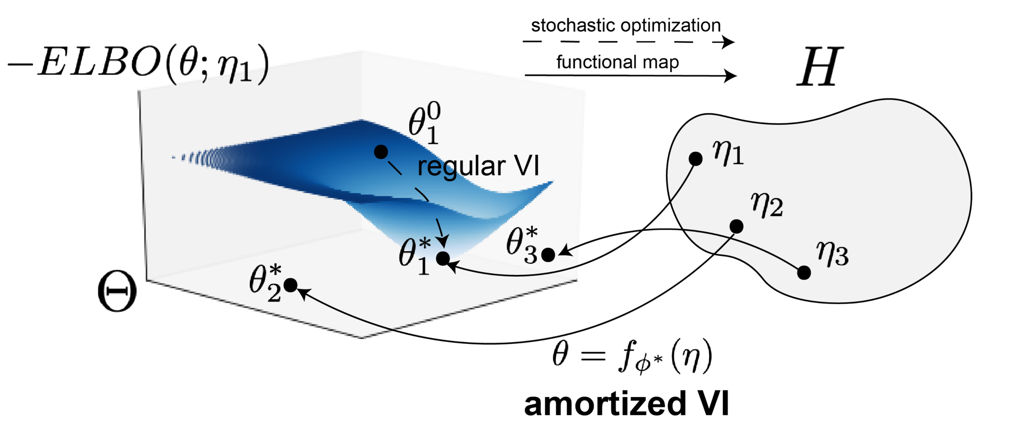
Approximating exponential family models (not single distributions) with a two-network architecture

Sean R. Bittner[1], John P. Cunningham[2]

[1]Department of Neuroscience and [2]Department of Statistics, Columbia University Medical Center

Motivation

- Many models used in machine learning are intractable exponential families.
- Variational inference (VI) on intractable exponential families incurs a cost of optimization.
- We introduce a deep generative two-network architecture called exponential family networks (EFNs) for learning intractable exponential family *models* (not single distributions).
- EFNs learn a smooth function $f_{\phi^*}: H \rightarrow \Theta$ mapping natty p's η (i.e. ) to optimal variational parameters θ^*.



- EFNs afford substantial computational savings through amortized VI.

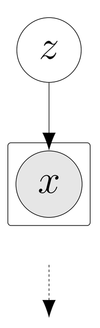
Exp fams as target models $\mathcal{P}$

- Exponential family models $\mathcal{P}$ have the form

$$\mathcal{P} = \left\{ \frac{h(\cdot)}{A(\eta)} \exp\left\{\eta^\top t(\cdot)\right\} : \eta \in H \right\}$$

with natural parameter η, sufficient statistics $t(\cdot)$, base measure $h(\cdot)$, and log normalizer $A(\eta)$.

- We focus on the fundamental problem setup of probabilistic inference: N conditionally independent observations x_i given latent variable z.



$p(z|x_1, ..., x_N)$

- With exponential family prior and likelihood,

$$p_0(z) = \frac{1}{A_0(\alpha)} \exp\left\{\alpha^\top t_0(z)\right\}$$

$$p(x_i|z) = \frac{1}{A(z)} \exp\left\{\nu(z)^\top t(x_i)\right\}$$

our posterior is the following exp fam

$$p(z|x_1,...,x_N) \propto \exp\left\{ \begin{bmatrix} \alpha \\ \sum_i t(x_i) \\ -N \end{bmatrix}^\top \begin{bmatrix} t_0(z) \\ \nu(z) \\ \log A(z) \end{bmatrix} \right\}$$

which is intractable for nonconjugate priors, requiring computation for inference.

Deep approximating families $\mathcal{M}$

- Deep generative models are commonly used as approximating famillies to single distributions.
- The density network (vertical) of our two-network architecture is a cascade of normalizing flows, which maps a base random variable ω to sample $z = g_\theta(\omega)$ with variational parameters θ.

$$\omega \sim q_0(\omega), z = g_\theta(\omega) = g_L \circ ... \circ g_1(\omega)$$

- Bijective normalizing flows allow us to calculate the density for each sample.

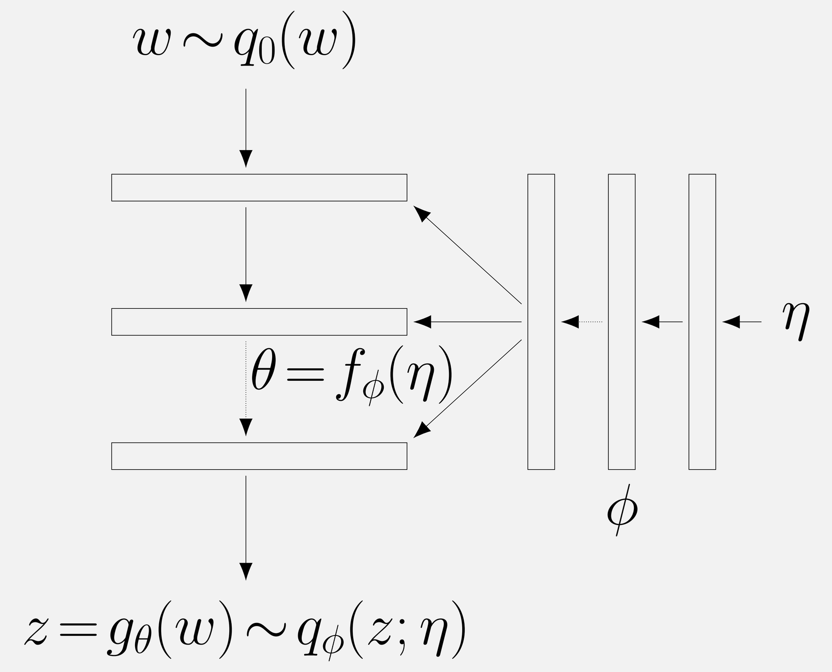
$$q_\theta(z) = q_0\left(g_1^{-1} \circ ... \circ g_L^{-1}(z)\right) \prod_{\ell=1}^{L} \frac{1}{|J_\theta^\ell(z)|}$$

- The density network induces a model

$$\mathcal{M} = \{q(g_\theta(\omega)) : \theta \in \Theta\}$$

Exponential family networks (EFNs)

- EFNs are comprised of two networks :
  - density network: $z = g_\theta(\omega)$
  - parameter network: $\theta = f_\phi(\eta)$
- The parameter network (horizontal) is a fully connected neural network mapping $f_\phi : H \rightarrow \Theta$.
- EFNs learn approximations $\mathcal{Q}_\phi \subset \mathcal{M}$ of exponential family models $\mathcal{P}$, so that $\mathcal{Q}_\phi \approx \mathcal{P}$, where $Q_\phi = \{q_{f_\phi}(z; \eta) : \eta \in H\}$.



- For a given η, we minimize the KL divergence $\mathcal{D}$ between the indexed distribution of $\mathcal{P}$ and $\mathcal{Q}_\phi$.

$$D\left(q_\phi(z;\eta) || p(z;\eta)\right) = \mathbb{E}_{q_\phi}\left(\log q_\phi(z;\eta) - \eta^\top t(z) + \log(A(\eta))\right)$$

We do this over a desired prior distribution $p(\eta)$,

$$\underset{\phi}{\text{argmin}}\, \mathbb{E}_{p(\eta)}\left(D\left(q_\phi(z;\eta)||p(z;\eta)\right)\right) = \underset{\phi}{\text{argmin}}\, D\left(q_\phi(z;\eta)p(\eta)||p(z;\eta)p(\eta)\right)$$

which corresponds to the loss below.

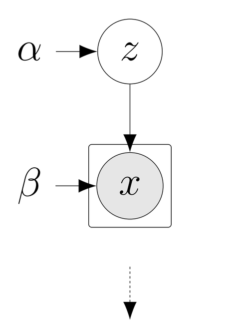
$$\mathbb{L}(\phi) = \frac{1}{K}\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M}\left(\log q_0\left(g_{\theta^k}^{-1}(z^m)\right) + \sum_{\ell=1}^{L}\log|J_{\theta^k}^\ell(z^m)| - \eta_k^\top t(z^m)\right)$$

Intractable exponential families

- Nonconjugate priors yield intractable exponential families requiring computation for inference.

Example: Hierarchical Dirichlet

- Dirichlet prior: $z \sim Dir(\alpha)$
- iid Dirichlet draws: $x_i \mid z \sim Dir(\beta z)$

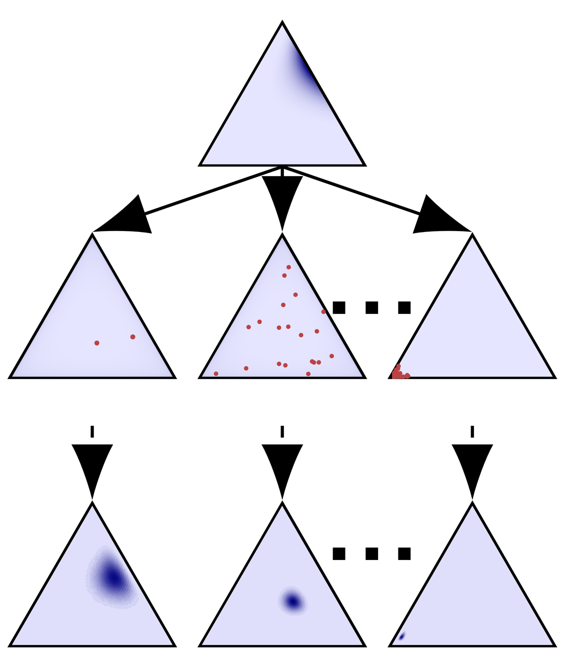


$p(z|x_1, ..., x_N)$

$$p(z \mid X) \propto \exp\left\{\eta^T t(z)\right\}$$

$$= \exp\left\{ \begin{bmatrix} \alpha - 1 \\ \sum_i \log(x_i) \\ -N \end{bmatrix}^\top \begin{bmatrix} \log(z) \\ \beta z \\ \log(B(\beta a)) \end{bmatrix} \right\}$$

- Consider a situation in which we want to do inference on the hierarchical Dirichlet model for three seprate data sets (red dots, middle triangles) given a constant prior (top triangle).



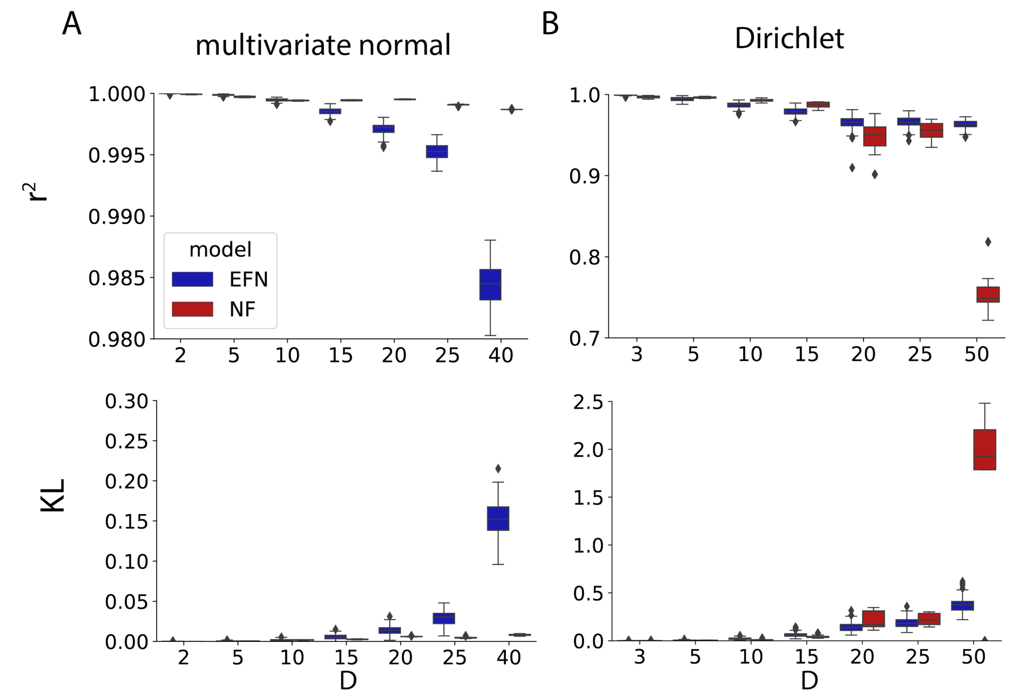
- Rather than running variational inference independently for each dataset, with an EFN, we can

1. Compute $\eta = \begin{bmatrix} \alpha - 1 \\ \sum_i \log(x_i) \\ -N \end{bmatrix}$ for each dataset.
2. Compute $\theta = f_{\phi^*}(\eta)$ using the parameter network.
3. Sample from the posterior using density network $z = g_{f_\phi^*(\eta)}(\omega)$ (bottom triangles).

- By training an EFN for the hierarchical Dirichlet intractable exponential family, we are amortizing variational inference.

Scaling dimensionality

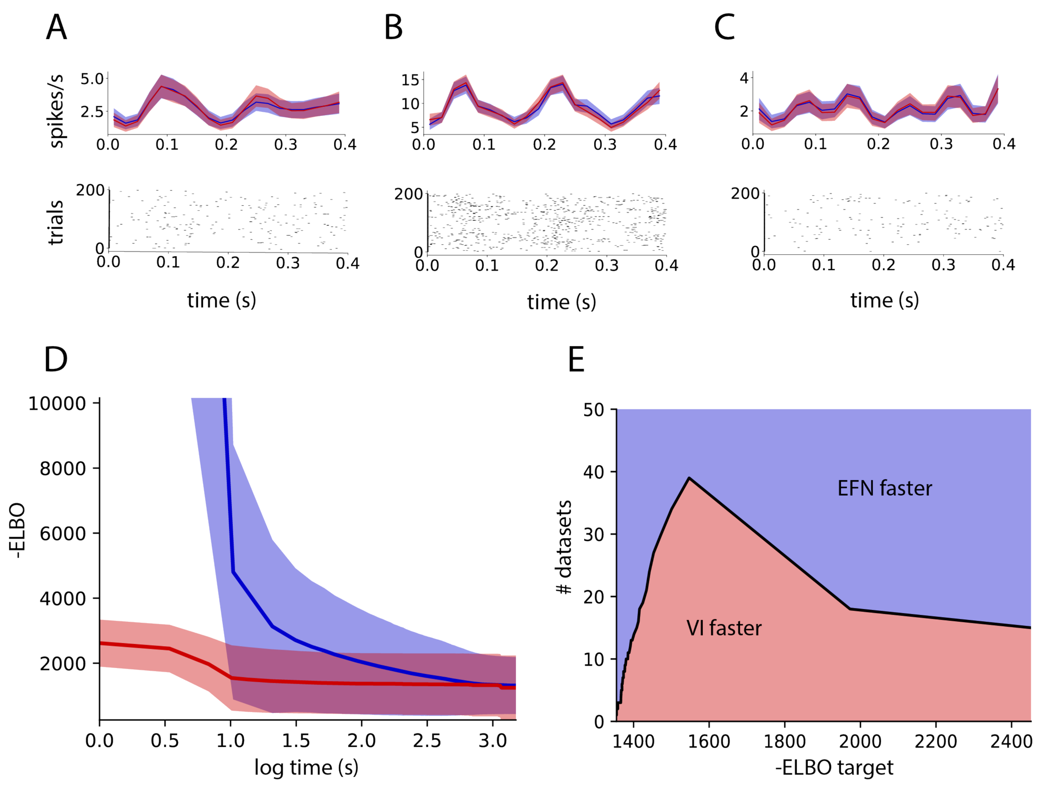
- EFN scales to 40-50 dimensions for ground truth exponential families.



- The regularization afforded by learning in a restricted model class $\mathcal{Q}_\phi$,results in better Dirichlet approximations in high dimensions.

Lookup posterior inference with neural data

- The log-Gaussian Poisson model is a popular intractable exponential family model used in neuroscience to represent neural spike emissions over a trial.
- If we want to run inference on the 2,964 datasets of 6.25Hz drift grating responses from [1], we can save a tremendous amount of computation using EFN (red) compared to independent runs of VI (blue).



Summary

- We learn exponential family models using a deep generative network, the parameters of which are the image of the natural parameters under another deep neural network.
- We demonstrated high quality empirical performance across a range of dimensionalities with the potential for better approximations when learning in a restricted model space.
- Finally, we show computational savings afforded by immediate posterior inference lookup.

References 1. Smith, Matthew A., and Adam Kohn. "Spatial and temporal scales of neuronal correlation in primary visual cortex." Journal of Neuroscience 28.48 (2008): 12591-12603.

Acknowledgements

NSF Graduate Research Fellowship, DGE-1644869, McKnight Endowment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF 1707398, The Gatsby Charitable Foundation