# Deep Exponential Families (AISTATS 2015)

Rajesh Ranganath, Linpeng Tang, Laurent Charlin, David Blei

Presented by
Wittawat Jitkrittum
Gatsby machine learning journal club

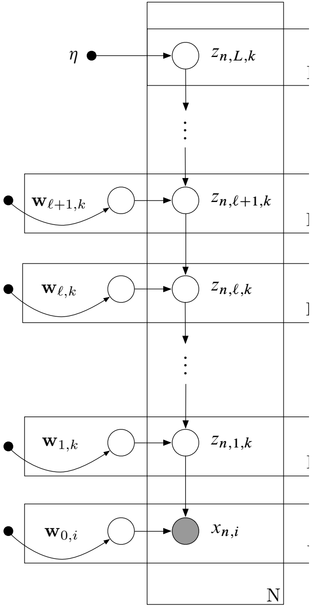
18 May 2015

# Exponential families

$$p(x|\eta) = \exp(\eta^\top T(x) - a(\eta))$$

- $\eta$ = natural parameter. $T(x)$ = sufficient statistic. $a(\eta)$ = log partition.
- $\mathbb{E}[T(x)] = \nabla_\eta a(\eta)$.
- Poisson: $p(z|\eta) = z!^{-1} \exp(\eta z - \exp(\eta))$
- Gamma: $p(z|\alpha, \beta) = z^{-1} \exp(\alpha \log(z) - \beta z - \log \Gamma(\alpha) - \alpha \log \beta)$
- To be more flexible, propose deep exponential families (DEF).

# Deep exponential families [Ranganath et al., 2015]




- For each observation $x_n$, $L$ layers of hidden variables $\{\boldsymbol{z}_{n,1}, \ldots, \boldsymbol{z}_{n,L}\}$.
- $K_\ell$-dimensional $\boldsymbol{z}_{n,\ell} = (z_{n,\ell,1}, \ldots, z_{n,\ell,K_l})^\top$.
- $L-1$ layers of weights $\{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{L-1}\}$.
- $\boldsymbol{W}_\ell = (\boldsymbol{w}_{\ell,1}, \ldots, \boldsymbol{w}_{\ell,K_l}) \in \mathbb{R}^{K_{\ell+1} \times K_\ell}$.
  Prior $p(\boldsymbol{W}_\ell)$.
- Dropping subscript $n$:

$$p(z_{\ell,k} \mid \boldsymbol{z}_{\ell+1}, \boldsymbol{w}_{\ell,k}) = \mathrm{ExpFam}_\ell(g_\ell(\boldsymbol{z}_{\ell+1}^\top \boldsymbol{w}_{\ell,k})).$$

- Link function $g_\ell : \boldsymbol{z}_{\ell+1}^\top \boldsymbol{w}_{\ell,k} \mapsto$ natural param.
- Likelihood $p(x_{n,i} \mid \boldsymbol{z}_{n,1})$.
- Sigmoid belief net = Bernoulli layers + identity link function

# $\nabla a$ and non-linearity

- Expected sufficient statistics = gradient of the log parition

$$\mathbb{E}[T(z_{\ell,k})] = \nabla_{\eta_{\ell,k}} a(g_\ell(\boldsymbol{z}_{\ell+1}^\top \boldsymbol{w}_{\ell,k})),$$

where $\eta_{\ell,k} := g_\ell(\boldsymbol{z}_{\ell+1}^\top \boldsymbol{w}_{\ell,k})$.

$$\begin{tikzcd} \mathbf{w}_{\ell,k} \arrow[dr] & & \\ & g_\ell \arrow[r, "\nabla a"] & \mathrm{E}[T(z_{n,\ell,k})] \\ \mathbf{z}_{n,\ell+1} \arrow[ur] & & \end{tikzcd}$$

- Consider $g_\ell(x) = x$ and $T(z_{\ell,k}) = z_{\ell,k}$.
- Then $\mathbb{E}[z_{\ell,k}] =$ linear function of $\boldsymbol{w}_{\ell,k}$ tranformed by $\nabla_{\eta_{\ell,k}} a(\cdot)$.
- This is one source of non-linearity.

# Inference with mean field

- Log partition function $a$ intractable.
- $N$ observations. Mean field approximation:

$$q(z, W) = q(\boldsymbol{W}_0) \prod_{l=1}^{L} q(\boldsymbol{W}_l) \prod_{n=1}^{N} q(\boldsymbol{z}_{n,l}),$$

  and $q(\boldsymbol{W}_l), q(\boldsymbol{z}_{n,l})$ fully factorized.
- Maximize evidence lower bound (ELBO)

$$\mathcal{L}(q) = \mathbb{E}_{q(z,W)} [\log p(x, z, W) - \log q(z, W)] \leq \log p(X).$$

- $q(W \mid \xi)$ in the same family as $p(W)$.
- $q(z_{n,l,k} \mid \lambda_{n,l,k}) = \mathrm{ExpFam}_l(z_{n,l,k} \mid \lambda_{n,l,k})$, same family as $p$.
- $\mathbb{E}_{q(z,W)}[\cdots]$ will not have a simple analytic form.
- Use blackbox variational inference (BBVI).

# Blackbox variational inference (BBVI) [Ranganath et al., 2013]

- Stochastic optimization. Follow noisy unbiased gradients.
- Gradient:

$$\begin{aligned}\nabla_{\lambda_{n,\ell,k}} \mathcal{L}(q) &= \nabla_{\lambda_{n,\ell,k}} \mathbb{E}_{q(z,W)} \left[\log p(x, z, W) - \log q(z, W)\right] \\ &= \mathbb{E}_q \left\{\nabla_{\lambda_{n,\ell,k}} \log q(z_{n,\ell,k}) \left[\log p_{n,\ell,k}(x, z, W) - \log q(z_{n,\ell,k})\right]\right\}.\end{aligned}$$

where $p_{n,\ell,k}(x, z, W)$ = terms in the joint containing $z_{n,\ell,k}$ (its Markov blanket).

- Markov blanket terms:

$$\log p_{n,\ell,k}(x, z, W) = \log p(z_{n,\ell,k} \mid \boldsymbol{z}_{n,\ell+1}, \boldsymbol{w}_{\ell,k}) + \log p(\boldsymbol{z}_{n,\ell-1} \mid \boldsymbol{z}_{n,\ell}, \boldsymbol{W}_{\ell-1}).$$

- Draw from $q$. Monte Carlo estimate of $\nabla_{\lambda_{n,\ell,k}} \mathcal{L}(q)$.
- Can parallelize $n = 1, \ldots, N$.
- Gradients of $\xi$ (for $W$) are similar.

# BBVI Algorithm

**Algorithm 1** BBVI for DEFs

**Input:** data $X$, model $p$, $L$ layers.
**Initialize** $\lambda, \xi$ randomly, $t = 1$.
**repeat**
  Sample a datapoint $x$
  **for** s = 1 to $S$ **do**
    $z_x[s], W[s] \sim q$
    $p[s] = \log p(z_x[s], W[s], x)$
    $q[s] = \log q(z_x[s], W[s])$
    $g[s] = \nabla \log q(z_x[s], W[s])$
  **end for**
  Compute gradient using BBVI
  Update variational parameters for $z$ and $W$
**until** change in validation likelihood is small

- Gradient = average $g[s]$ for $s = 1, \ldots, S$.

# Example: sparse gamma DEF

- Gamma distributed layers.

$$p(z|\alpha, \beta) = z^{-1} \exp(\alpha \log(z) - \beta z - \log \Gamma(\alpha) - \alpha \log \beta).$$

- Link functions

$$g_\alpha = \alpha_\ell,$$
$$g_\beta = \frac{\alpha_\ell}{\boldsymbol{z}_{\ell+1}^\top \boldsymbol{w}_{\ell,k}}.$$

- $p(W)$ = Gamma distribution so $\boldsymbol{z}_{\ell+1}^\top \boldsymbol{w}_{\ell,k} > 0$.
- $\alpha_\ell$ and shape parameters of $p(W)$ are set to be less than $1$.
- Probability mass near 0 $\Rightarrow$ sparse gamma.

# Example: Poisson DEF

$$p(z|\eta) = z!^{-1} \exp(\eta z - \exp(\eta))$$

with mean $\exp(\eta)$.

- Poisson DEF = Poisson latent + log-link function.

$$p(z_{\ell,k} \mid \boldsymbol{z}_{\ell+1}, \boldsymbol{w}_{\ell,k}) = (z_{\ell,k}!)^{-1} \exp(\log(\boldsymbol{z}_{\ell+1}^\top \boldsymbol{w}_{\ell,k}) z_{\ell,k} - \boldsymbol{z}_{\ell+1}^\top \boldsymbol{w}_{\ell,k}).$$

- So, the mean of $z_{\ell,k}$ is $\boldsymbol{z}_{\ell+1}^\top \boldsymbol{w}_{\ell,k}$.
- $p(\boldsymbol{W}_\ell)$ a factorized gamma distribution.

# Experiment: text modelling

- Datasets: The New York Times (NYT), and Science.
- Multinomial likelihood: $p(\text{count of } w \mid \text{latent})$.
- 3 cases for latent: Poisson, gamma and Bernoulli.
- All methods see 10% of the words in each doc. 90% held-out.
- Perplexity on a held out set of 1000 documents.

$$\exp\left(\frac{-\sum_{d \in \text{docs}} \sum_{w \in d} \log p(w \mid \#\text{held out words in } d)}{N_{\text{held out words}}}\right).$$

Lower is better.

# Text modelling results

| Model | DEF $\mathbf{W}$ | *NYT* | *Science* |
|---|---|---|---|
| LDA [6] | | 2717 | 1711 |
| DocNADE [19] | | 2496 | 1725 |
| Sparse Gamma 100 | $\emptyset$ | 2525 | 1652 |
| Sparse Gamma 100-30 | $\Gamma$ | 2303 | 1539 |
| Sparse Gamma 100-30-15 | $\Gamma$ | 2251 | 1542 |
| Sigmoid 100 | $\emptyset$ | 2343 | 1633 |
| Sigmoid 100-30 | $\mathcal{N}$ | 2653 | 1665 |
| Sigmoid 100-30-15 | $\mathcal{N}$ | 2507 | 1653 |
| Poisson 100 | $\emptyset$ | 2590 | 1620 |
| Poisson 100-30 | $\mathcal{N}$ | 2423 | 1560 |
| Poisson 100-30-15 | $\mathcal{N}$ | 2416 | 1576 |
| Poisson log-link 100-30 | $\Gamma$ | 2288 | 1523 |
| Poisson log-link 100-30-15 | $\Gamma$ | 2366 | 1545 |

- 100-30-15 indicates sizes of the layers.
- DEFs outperform the baselines (LDA and DocNADE).
- Deeper layers help.
- Sigmoid DEFs difficult to train.

# Experiment: matrix factorization

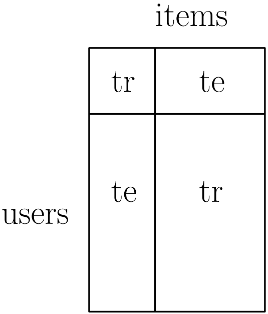
- Consider user-item matrices containing ratings.

$$p(x_{n,i} \mid \boldsymbol{z}^c_{n,1}, \boldsymbol{z}^r_{i,1}) = \text{Poisson}(\boldsymbol{z}^{c\top}_{n,1} \boldsymbol{z}^r_{i,1}).$$

- $\boldsymbol{z}^c_{n,1}$ = hidden representation of user $n$.
- $\boldsymbol{z}^r_{i,1}$ = hidden representation of item $i$.
- Put hierarchies on both $\boldsymbol{z}^c$ and $\boldsymbol{z}^r$.
- Datasets:
  1. **Netflix movie ratings**. 50K users. 17.7K movies.
  2. **ArXiv click data**.
     - Viewers $\times$ papers matrix containing click counts.
     - 18K users. 20K docs.

# Matrix factorization results

| Model | Netflix Perplexity | Netflix NDCG | ArXiv Perplexity | ArXiv NDCG |
|---|---|---|---|---|
| Gaussian MF [32] | – | 0.008 | – | 0.013 |
| 1 layer Double DEF | 2319 | 0.031 | 2138 | 0.049 |
| 2 layer Double DEF | 2299 | 0.022 | 1893 | 0.050 |
| 3 layer Double DEF | 2296 | 0.037 | 1940 | 0.053 |




- Layer sizes: 100-30-15.
- Report perpexity on the test set as before.
- 1000 users in the test set.
- Claim: deeper is better (sort of).
- Gaussian MF = $\ell_2$-regularized Gaussian matrix factorization.
- NDCG = multi-level ranking measure.

# References I


Ranganath, R., Gerrish, S., and Blei, D. M. (2013).
Black Box Variational Inference.
arXiv:1401.0118 [cs, stat].
arXiv: 1401.0118.

Ranganath, R., Tang, L., Charlin, L., and Blei, D. (2015).
{Deep Exponential Families}.
pages 762–771.

# All DEFs in the paper

| z-Dist | $\mathbf{z}_{\ell+1}$ | W-dist | $\mathbf{w}_{\ell,k}$ | $g_\ell$ | $\mathrm{E}[T(z_{\ell,k})]$ |
|---|---|---|---|---|---|
| Gamma | $R_+^{K_{\ell+1}}$ | Gamma | $R_+^{K_{\ell+1}}$ | [constant; inverse] | $[z_{\ell+1}^\top \mathbf{w}_{\ell,k}; \Psi(\alpha_\ell) - \log(\alpha) + \log(z_{\ell+1}^\top \mathbf{w}_{\ell,k})]$ |
| Bernoulli | $\{0,1\}^{K_{\ell+1}}$ | Normal | $R^{K_{\ell+1}}$ | identity | $\sigma(z_{\ell+1}^\top \mathbf{w}_{\ell,k})$ |
| Poisson | $N^{K_{\ell+1}}$ | Gamma | $R_+^{K_{\ell+1}}$ | log | $z_{\ell+1}^\top \mathbf{w}_{\ell,k}$ |
| Poisson | $N^{K_{\ell+1}}$ | Normal | $R^{K_{\ell+1}}$ | log-softmax | $\log(1+\exp(z_{\ell+1}^\top \mathbf{w}_{\ell,k}))$ |

Table 1: A summary of all the DEFs we present in terms of their layer distributions, weight distributions, and link functions.

- Focus on document (bag of words) modelling .