Deep Exponential Families (Appendix)

Rajesh Ranganath
Princeton University

Linpeng Tang
Princeton University

Laurent Charlin
Columbia University

David M. Blei
Columbia University

{rajeshr,linpengt}@cs.princeton.edu
{lcharlin,blei}@cs.columbia.edu

Appendix

Gamma Distribution Figure 1 visually demonstrates how the sparse gamma distrubution and Poisson distribution change when moving from low to high mean. We plot both the Poisson and sparse gamma distribution in both settings. The mode of the Poisson distribution moves, while the slab grows for the gamma.

General Algorithm. Following the notation from the main paper, the general algorithm for mean field variational inference in deep exponential families is given in (Alg. 1).

Properties of q. In our experiments, we use four variational families (Poisson, gamma, Bernoulli, and normal). We detail the necessary score functions here. For the Poisson, the distribution is given by:

$$q(z) = e^{-\lambda}\frac{\lambda^z}{z!}.$$

The score function is

$$\frac{\partial \log q(z)}{\partial \lambda} = -1 + \frac{z}{\lambda}.$$

For the gamma, we use the shape α and scale θ as variational parameters. The distribution is given by

$$q(z) = \frac{1}{\Gamma(\alpha)\theta^\alpha} z^{\alpha-1} e^{-z/\theta}.$$

The score function is

$$\frac{\partial \log q(z)}{\partial \alpha} = -\Psi(\alpha) - \log\theta, + \log z$$
$$\frac{\partial \log q(z)}{\partial \theta} = -\alpha/\theta + z/\theta^2,$$

where Ψ is the digamma function.

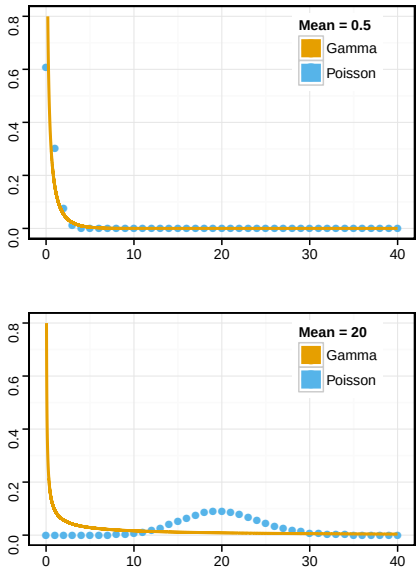

Figure 1: Draws from the Poisson (blue) and sparse gamma distribution (orange) with low and high mean. The shape of the sparse gamma is held fixed. Note the high mean shifts the Poisson, while does not shift the sparse gamma. Notice the spike-slab appearance of the sparse gamma distribution.

For the Bernoulli distribution, we use the natural parameterization with parameter η to form the variational approximation. The distribution is

$$q(z) = \frac{1}{1 + e^{-(2z-1)\eta}}.$$

The score function is

$$\frac{\partial \log q(z)}{\partial \eta} = (2z - 1)\frac{e^{-(2z-1)\eta}}{1 + e^{-(2z-1)\eta}}.$$

For the normal variational approximation, we use the standard parameterization of mean μ and variance σ^2. The distribution is

$$q(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

The score function is

$$\frac{\partial \log q(z)}{\partial \mu} = \frac{x - \mu}{\sigma^2}$$
$$\frac{\partial \log q(z)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}.$$

Parameterizations of Variational Distributions. Several of our variational parameters like the variance of the normal have positive constraints To enforce positivity constraints, we transform an unconstrained variable by $\log(1 + \exp(x))$. To avoid numerical issues when sampling, we truncate values when appropriate. Gradients of the unconstrained parameters are obtained with the chain rule from the score function and the derivative of softmax: $\exp(x)/(1 + \exp(x))$.

Optimization We perform gradient ascent step on the ELBO using

$$\Delta\theta = \rho\Gamma\nabla_\theta \text{ELBO} \tag{1}$$

ρ is a fixed scalar set to 0.2 in our experiments. ∇_θELBO is a noisy gradient estimated using BBVI. Γ is a diagonal preconditioning matrix estimated using the RMSProp heuristic. A diagonal element of Γ is the reciprocal of the squared root of a running average of the squares of historical gradients of that component. We used a window size of 10 in our experiments.

Hyperparameters and Convergence We use the same hyperparameters on Gamma distributions on each layer with shape and rate 0.3. For the sigmoid belief network we use a prior of 0.1 to achieve some sparsity as well. We fix the Poisson prior rate to be 0.1. For gamma W's we use shape 0.1 and rate 0.3. For Gaussian W's we use a prior mean of 0 and variance of 1. We let the experiments run for 10,000 iterations at which point the validation likelihood is stable.

Algorithm 1 BBVI for DEFs

Input: data X, model p, L layers .
Initialize λ, ξ randomly, $t = 1$.
repeat
  // **Draw a single data point from X**
  $n =$Unif(D)
  // **Get S samples in parallel**
  for $s = 1$ **to** S **do**
    $z_1[s] \sim q(z_1; \lambda_{n,1})$
    $W_0[s] \sim q(W_0 \mid \xi_0)$
    $p_0[s] = \log p(x_n \mid z_1[s], W_0[s])$
    $q_1[s] = \log q(z_1[s]; \lambda_{n,1})$
    $g_1[s] = \nabla_{\lambda_{n,1}} \log q(z_1[s]; \lambda_{n,1})$
    $g_{W_0}[s] = \nabla_{\xi_1} \log q(W_0)$
    $p_{W_0}[s] = \log p(W_0; \xi_1)$
    $q_{W_0}[s] = \log q(W_0; \xi_1)$
    for $l = 2$ **to** L **do**
      $z_l[s] \sim q(z_l; \lambda_{n,l})$
      $W_{l-1}[s] \sim q(W_{l-1} \mid \xi_{l-1})$
      $p_l[s] = \log p(z_{l-1} \mid z_l, W_{l-1}[s])$
      $q_l[s] = \log q(z_l; \lambda_{n,l})$
      $g_l[s] = \nabla_{\lambda_{n,l}} \log q(z_l; \lambda_{n,l})$
      $g_{W_{l-1}}[s] = \nabla_{\xi_{l-1}} \log q(W_{l-1})$
      $p_{W_{l-1}}[s] = \log p(W_{l-1}; \xi_{l-1})$
      $q_{W_{l-1}}[s] = \log q(W_{l-1}; \xi_{l-1})$
    end for
    $p_L[s] = \log p(z_L)$
  end for
  // **Update parameters**
  for $l = 1$ **to** L **do**
    for $k = 1$ **to** K_l **do**
      $S = g_{l,k}(p_{l-1} + p_{l,k} - q_{l,k})$
      $\lambda_{n,1,k} = \lambda_{n,1,k} + \rho$ mean(S)
    end for
    $T = g_{W_{l-1}}(p_{W_{l-1}} - q_{W_{l-l}} + p_{l-1})$
    $\xi_{l-1} = \xi_{l-1} + \rho$ mean(T)
  end for
until change of val likelihood is less than ϵ.

For the double DEF, we set all shapes to 0.1 and rates to 0.3. We let the Double DEF experiment run for about 10,000 iterations. The validation likelihood had converged for all models by this point.

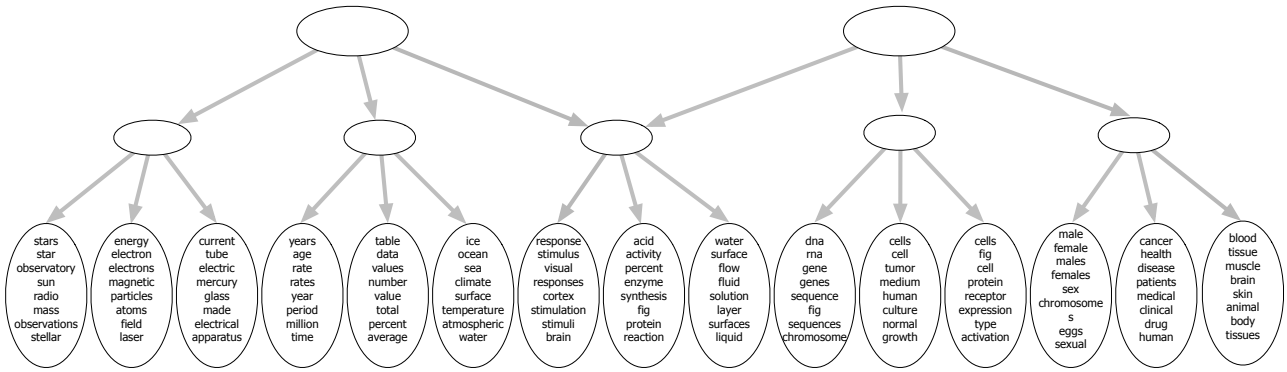

Figure 2: A fraction of the three layer topic hierarchy of the *Science* corpus. The top words are shown for each "topic." The arrows represent hierarchical groupings. We choose top three components at each layer. Similar "topics" are grouped into "super topics." The two "concepts" share a "super topic."