Deep Exponential Families

Rajesh Ranganath Linpeng Tang Laurent Charlin David Blei

Paper Presentation by
Evangelos Chatzipantazis
for CIS620

October 2020

Table of Contents

Table of Contents

Exponential Families

- A family $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ of probability measures on $\mathcal{X} \subset \mathbb{R}^d$

Exponential Families

- A family $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ of probability measures on $\mathcal{X} \subset \mathbb{R}^d$ that have densities $p_\theta(x)$ of the form:

$$
\begin{aligned}
p_\theta(x) &= h(x) \exp\left(\sum_{i=1}^{m} \eta_i(\theta) T_i(x) - A(\theta)\right) \\
&= \frac{h(x)}{Z(\theta)} \exp\left(\eta(\theta)^T T(x)\right)
\end{aligned}
$$

forms an m-parametric exponential family.

Exponential Families

- A family $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ of probability measures on $\mathcal{X} \subset \mathbb{R}^d$ that have densities $p_\theta(x)$ of the form:

$$
\begin{aligned}
p_\theta(x) &= h(x) \exp\big(\sum_{i=1}^{m} \eta_i(\theta) T_i(x) - A(\theta)\big) \\
&= \frac{h(x)}{Z(\theta)} \exp\big(\eta(\theta)^T T(x)\big)
\end{aligned}
$$

  forms an m-parametric exponential family.

- Disclaimer: The definition above can be extended to the Lebesque measure (eg. the counting measure for pmfs) or any other ($\sigma-$ finite) measure, but we stick to the case where $\mathcal{X} \subset \mathbb{R}^d$ and a pdf exists for clarity

Exponential Families

- A family $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ of distributions on $\mathcal{X} \subset \mathbb{R}^d$ that have densities $p_\theta(x)$ of the form:

$$
\begin{aligned}
p_\theta(x) &= h(x) \exp\left(\sum_{i=1}^{m} \eta_i(\theta) T_i(x) - A(\theta)\right) \\
&= \frac{h(x)}{Z(\theta)} \exp\left(\eta(\theta)^T T(x)\right)
\end{aligned}
$$

forms an m-parametric exponential family.

$$
\begin{aligned}
h &: \mathbb{R}^d \to \mathbb{R}^+ && \text{(support or base measure)} \\
\eta &: \Theta \to \mathbb{R}^m && \text{(natural parameter)} \\
T &: \mathbb{R}^d \to \mathbb{R}^m && \text{(sufficient statistic)} \\
A &: \Theta \to \mathbb{R} && \text{(log partition)}
\end{aligned}
$$

Exponential Families

- (natural parametrization) $\mathcal{P} = \{p_\eta : \eta \in \eta(\Theta)\}$

$$p_\eta(x) = h(x)\exp\left(\eta^\top T(x) - A(\eta)\right)$$

Exponential Families

- (natural parametrization) $\mathcal{P} = \{p_\eta : \eta \in \eta(\Theta)\}$

$$p_\eta(x) = h(x) \exp\left(\eta^\top T(x) - A(\eta)\right)$$

  - $\eta = (\eta_1, \cdots, \eta_m) \in \mathcal{H}$ (natural parameter space)

Exponential Families

- (natural parametrization) $\mathcal{P} = \{p_\eta : \eta \in \eta(\Theta)\}$

$$p_\eta(x) = h(x) \exp\left(\eta^\top T(x) - A(\eta)\right)$$

  - $\eta = (\eta_1, \cdots, \eta_m) \in \mathcal{H}$ (natural parameter space)
  - θ's indicate members of the family. η's may be redundant.

Exponential Families

- (natural parametrization) $\mathcal{P} = \{p_\eta : \eta \in \eta(\Theta)\}$

$$p_\eta(x) = h(x) \exp\left(\eta^\top T(x) - A(\eta)\right)$$

  - $\eta = (\eta_1, \cdots, \eta_m) \in \mathcal{H}$ (natural parameter space)
  - θ's indicate members of the family. η's may be redundant.
  - $A(\eta) = \ln \int_{\mathbb{R}^d} h(x) \exp\left(\eta^\top T(x)\right) dx \in \mathbb{R}$, if $\eta \in \mathcal{H}$

Exponential Families

- (natural parametrization) $\mathcal{P} = \{p_\eta : \eta \in \eta(\Theta)\}$

$$p_\eta(x) = h(x) \exp\left(\eta^\top T(x) - A(\eta)\right)$$

  - $\eta = (\eta_1, \cdots, \eta_m) \in \mathcal{H}$ (natural parameter space)
  - θ's indicate members of the family. η's may be redundant.
  - $A(\eta) = \ln \int_{\mathbb{R}^d} h(x) \exp\left(\eta^\top T(x)\right) dx \in \mathbb{R}$, if $\eta \in \mathcal{H}$
  - for a specific distribution η, T, h are not uniquely defined.

Exponential Families

- (natural parametrization) $\mathcal{P} = \{p_\eta : \eta \in \eta(\Theta)\}$
$$p_\eta(x) = h(x) \exp\left(\eta^\top T(x) - A(\eta)\right)$$
  - $\eta = (\eta_1, \cdots, \eta_m) \in \mathcal{H}$ (natural parameter space)
  - θ's indicate members of the family. η's may be redundant.
  - $A(\eta) = \ln \int_{\mathbb{R}^d} h(x) \exp\left(\eta^\top T(x)\right) dx \in \mathbb{R}$, if $\eta \in \mathcal{H}$
  - for a specific distribution η, T, h are not uniquely defined.
  - $T(x)$ is the m-dimensional sufficient statistic. Important to find minimal m ($\Longleftarrow$ *linearly independent* T).

Exponential Families

- (natural parametrization) $\mathcal{P} = \{p_\eta : \eta \in \eta(\Theta)\}$

$$p_\eta(x) = h(x) \exp\left(\eta^\top T(x) - A(\eta)\right)$$

  - $\eta = (\eta_1, \cdots, \eta_m) \in \mathcal{H}$ (natural parameter space)
  - θ's indicate members of the family. η's may be redundant.
  - $A(\eta) = \ln \int_{\mathbb{R}^d} h(x) \exp\left(\eta^\top T(x)\right) dx \in \mathbb{R}$, if $\eta \in \mathcal{H}$
  - for a specific distribution η, T, h are not uniquely defined.
  - $T(x)$ is the m-dimensional sufficient statistic. Important to find minimal m ($\Longleftarrow$ *linearly independent* T).
  - (*minimal*) $k <$ (*minimal*) $m \implies$ *curved family* $\implies$ k independent parameters embedded in a m dimensional parameter space.
    - For at least 1 member of a family (at least 1 θ)

Exponential Families

- (natural parametrization) $\mathcal{P} = \{p_\eta : \eta \in \eta(\Theta)\}$

$$p_\eta(x) = h(x) \exp\left(\eta^\top T(x) - A(\eta)\right)$$

- Misconception:
  - There is no such thing as THE exponential family.
  - $T(,h)$ fixed and η varies $\implies$ *same family* of distributions.
  - For example, $\mathcal{N}(x|\mu,\sigma)$ is a *family* of gaussian distributions. This family is actually AN exponential family.

How to detect an exponential family

- Takeaway: Observations and parameters must factorize
  - Either directly.
  - Or both in base and the exponent

$$\mathrm{p}(\mathrm{x}) = Z(\theta)^{-1} h(x) \exp\left(\eta(\theta)^{\top} T(x)\right)$$

Examples of Exponential Families

$$p(x) = Z(\theta)^{-1} h(x) \exp\left(\eta(\theta)^\top T(x)\right)$$

1) Exponential (1 parametric):

$$p(x;\theta) = \theta \exp\left(-\theta x\right) I(x \geq 0), \quad \Theta = \mathbb{R}^{*+}$$
$$\eta(\theta) = \theta \text{ "canonical/natural form"}$$

Examples of Exponential Families

$$p(x) = Z(\theta)^{-1} h(x) \exp\left(\eta(\theta)^\top T(x)\right)$$

1) Exponential (1 parametric):

$$p(x;\theta) = \theta \exp\left(-\theta x\right) I(x \geq 0), \quad \Theta = \mathbb{R}^{*+}$$
$$\eta(\theta) = \theta \text{ ''canonical/natural form''}$$

2) Binomial (n fixed) (1 parametric):

- $$p(k;\theta) = \binom{n}{k} \theta^k (1-\theta)^{(n-k)}$$
$$= (1-\theta)^n \binom{n}{k} I(k \in [n]) \exp\left(\ln \frac{\theta}{1-\theta} k\right)$$

- $\Theta = (0,1)$
- $\eta = \ln \frac{\theta}{1-\theta}$ (logits) $\implies A(\eta) = n \log\left(e^\eta + 1\right),\ \mathcal{H} = \mathbb{R}$

Examples of Exponential Families

$$p(x) = Z(\theta)^{-1} h(x) \exp\left(\eta(\theta)^\top T(x)\right)$$

3) Gaussian (2-parametric)

$$
\begin{aligned}
p(x;\mu,\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(-\left(\frac{1}{2\sigma^2}, \frac{-\mu}{\sigma^2}\right)^\top (x^2, x)\right), \ \Theta = \mathbb{R} \times \mathbb{R}^{*+}
\end{aligned}
$$

$$p(x) = Z(\theta)^{-1} h(x) \exp\left(\eta(\theta)^\top T(x)\right)$$

3) Gaussian (2-parametric)

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(-\left(\frac{1}{2\sigma^2}, \frac{-\mu}{\sigma^2}\right)^\top (x^2, x)\right), \quad \Theta = \mathbb{R} \times \mathbb{R}^{*+}$$

- $\mathcal{N}(x|\mu, \mu^2)$ is a *curved family* ($dim(\theta) < dim(\eta)$).
  Statistics are again 2 dimensional $(T_1(x), T_2(x)) = (x^2, x)$
  - Check $Cov_\theta(T(X)) \succ 0$ for some θ (for minimality) ✓
  - Strict 2-dimensional family (Stonger). ($\Longleftarrow$)
    $\{1, T_1(x), T_2(x)\}$ linearly independent for all θ (w.r.t the measure).
    And $\{1, \eta_1(\theta), \eta_2(\theta)\}$ linearly independent.
  - η's could/should still have non-linear (curved) dependencies!

How to get away with Curved

- Multinomial: M independent trials, K events, $X_i = \#$(event i occured in M trials), $\pi_i = \mathbb{P}[X^{(n)} = i], X = (X_1, \cdots, X_K)$

How to get away with Curved

- Multinomial: M independent trials, K events, $X_i = \#(\text{event i occured in M trials})$, $\pi_i = \mathbb{P}[X^{(n)} = i], X = (X_1, \cdots, X_K)$
- $$p(x; \{\pi_i\}_{i=1}^{K}) = \frac{M!}{x_1! \cdots x_K!} \pi_1^{x_1} \cdots \pi_K^{x_K} = \frac{M!}{x_1! \cdots x_K!} e^{(\sum_{i=1}^{K} x_i \ln \pi_i)}$$

How to get away with Curved

- Multinomial: M independent trials, K events, $X_i = \#$(event i occured in M trials), $\pi_i = \mathbb{P}[X^{(n)} = i], X = (X_1, \cdots, X_K)$
-
$$p(x; \{\pi_i\}_{i=1}^K) = \frac{M!}{x_1! \cdots x_K!} \pi_1^{x_1} \cdots \pi_K^{x_K} = \frac{M!}{x_1! \cdots x_K!} e^{(\sum_{i=1}^K x_i \ln \pi_i)}$$
- $h(x) = \frac{M!}{x_1! \cdots x_K!} I(x_i \in \mathbb{Z}, \sum_{i=1}^K x_i = M)$ (absorbed in dom. measure)

How to get away with Curved

- Multinomial: M independent trials, K events, $X_i = \#$(event i occured in M trials), $\pi_i = \mathbb{P}[X^{(n)} = i], X = (X_1, \cdots, X_K)$
- $$p(x; \{\pi_i\}_{i=1}^K) = \frac{M!}{x_1! \cdots x_K!}\pi_1^{x_1} \cdots \pi_K^{x_K} = \frac{M!}{x_1! \cdots x_K!} e^{(\sum_{i=1}^K x_i \ln \pi_i)}$$
- $h(x) = \frac{M!}{x_1! \cdots x_K!} I(x_i \in \mathbb{Z}, \sum_{i=1}^K x_i = M)$ (absorbed in dom. measure)
- $T(X) = X$, $\eta = \{\ln \pi_i\}_{i=1}^K$, $A(\eta) = 0$ (?).

How to get away with Curved

- Multinomial: M independent trials, K events, $X_i = \#$(event i occured in M trials), $\pi_i = \mathbb{P}[X^{(n)} = i], X = (X_1, \cdots, X_K)$
- $$p(x; \{\pi_i\}_{i=1}^K) = \frac{M!}{x_1! \cdots x_K!} \pi_1^{x_1} \cdots \pi_K^{x_K} = \frac{M!}{x_1! \cdots x_K!} e^{(\sum_{i=1}^K x_i \ln \pi_i)}$$
- $h(x) = \frac{M!}{x_1! \cdots x_K!} I(x_i \in \mathbb{Z}, \sum_{i=1}^K x_i = M)$ (absorbed in dom. measure)
- $T(X) = X$, $\eta = \{\ln \pi_i\}_{i=1}^K$, $A(\eta) = 0$ (?).
- Is this family curved? or full-rank? Is it *strictly* K-dimensional?

How to get away with Curved

- Intuitively there are only $K-1$ independent parameters but η has dimension K. $\Theta = \{\pi : \pi_i \in (0,1), \sum_{i=1}^{K} \pi_i = 1\}$
- $\mathcal{H} = \mathbb{R}^K$ and not only the subspace $\mathcal{H}^- = \{\eta : \sum_{i=1}^{K} e_i^{\eta} = 1\}$.
  - $A(\eta) = 0$ if $\eta \in \mathcal{H}^-$
  - $A(\eta) \neq 0$ if $\eta \in \mathcal{H}$!
  - Inconvenient to work on the ambient space. Redundant representation!
- $\mathcal{X} = \{x_j \in \mathbb{N}, \sum_{i=1}^{K} x_i = M\} \implies \sum_{i=1}^{K} T_i(x) = M \implies \{1, T_1(x), \cdots, T_K(x)\}$ not linear independent $\implies$ not minimal representation! (More than sufficient)

Multinomial Version 2

- $x_K = M - x_{K-1} - \cdots - x_1$
- $\pi_K = 1 - \sum_{i=1}^{K-1} \pi_i$

$$p(x;\pi) = \binom{M}{x_1, \cdots, x_K} e^{\left(\sum_{i=1}^{K-1} x_i \ln \pi_i + (M - \sum_{i=1}^{K-1} x_i) \ln \pi_K\right)}$$

$$= \binom{M}{x_1, \cdots, x_K} \pi_K^M e^{\sum_{i=1}^{K-1} x_i \ln \frac{\pi_i}{\pi_K}}$$

- Representation is minimal. Not curved, full-rank family, of order K-1.
- $\pi_k = \frac{e^{\eta_k}}{\sum_{i=1}^{K-1} e^{\eta_i} + 1}, \; k \in [K-1]$ (softmax)
- $A(\eta) = M \ln \left(\sum_{i=1}^{K-1} e^{\eta_i} + 1\right)$
- $\mathbb{E}[T_i(X)] = \mathbb{E}[X_i] = M\pi_i = \nabla_{\eta_i} A(\eta)$
- $Cov[T_i(X), T_j(X)] = Cov(X_i, X_j) = -M\pi_i\pi_j = \frac{\partial^2 A(\eta)}{\partial \eta_i \eta_j}$

More exponential families

Name	sufficient stats	domain	use case
Bernoulli	$\phi(x) = [x]$	$\mathbb{X} = \{0; 1\}$	coin toss
Poisson	$\phi(x) = [x]$	$\mathbb{X} = \mathbb{R}_+$	emails per day
Laplace	$\phi(x) = [1, x]^\intercal$	$\mathbb{X} = \mathbb{R}$	floods
Helmert (χ^2)	$\phi(x) = [x, -\log x]$	$\mathbb{X} = \mathbb{R}$	variances
Dirichlet	$\phi(x) = [\log x]$	$\mathbb{X} = \mathbb{R}_+$	class probabilities
Euler (Γ)	$\phi(x) = [x, \log x]$	$\mathbb{X} = \mathbb{R}_+$	variances
Wishart	$\phi(X) = [X, \log \lvert X \rvert]$	$\mathbb{X} = \{X \in \mathbb{R}^{N \times N} \mid v^\intercal X v \geq 0 \forall v \in \mathbb{R}^N\}$	covariances
Gauss	$\phi(X) = [X, XX^\intercal]$	$\mathbb{X} = \mathbb{R}^N$	functions
Boltzmann	$\phi(X) = [X, \mathrm{triag}(XX^\intercal)]$	$\mathbb{X} = \{0; 1\}^N$	thermodynamics

Figure: More Exponential Families. Figure taken from P.Hennig

Counter Examples

- As we saw many exponential families: Beta,Gamma,Poisson, Laplace, Chi-squared, Wishart.

Counter Examples

- As we saw many exponential families: Beta,Gamma,Poisson, Laplace, Chi-squared, Wishart.
- Do we know any parametric families that are not exponential?

Counter Examples

- As we saw many exponential families: Beta,Gamma,Poisson, Laplace, Chi-squared, Wishart.
- Do we know any parametric families that are not exponential?
- Remember the parameters and observations must *factorize*.

Counter Examples

- As we saw many exponential families: Beta,Gamma,Poisson, Laplace, Chi-squared, Wishart.
- Do we know any parametric families that are not exponential?
- Remember the parameters and observations must *factorize*.
- A non Example: $p_\theta(x) = U(0, \theta) = \frac{1}{\theta} I(x \in [0, \theta])$

Counter Examples

- As we saw many exponential families: Beta,Gamma,Poisson, Laplace, Chi-squared, Wishart.
- Do we know any parametric families that are not exponential?
- Remember the parameters and observations must *factorize*.
- A non Example: $p_\theta(x) = U(0,\theta) = \frac{1}{\theta} I(x \in [0,\theta])$
- A note: Of course, for any fixed θ we get a uniform distribution, which is a trivial exponential family. We cannot gather all those trivial exponential families into a parametric family that is still exponential!

More Counter Examples

- Terms $1 + f(x)g(\theta)$ do not factorize:
    - Cauchy: $\frac{1}{\pi\theta}\frac{1}{(x-x_0)^2/\theta^2+1}$
    - Student't, etc.
- Note: $h(x)$ is fixed for all θ. In exponential families the support is fixed.
    - A non example: $p(x; \{\theta, n\}) = Bin(x; \{\theta, n\})$
- Most *mixtures* are not exponential families: Mixture of Gaussians.
- Most *Compound Probability Distributions* are not exponential!
    - Except Conjugacy!
    - Relevant to today's paper.

Sufficient Statistic

- A *Statistic* is a function of the sample (only).

$$T : x \in \mathcal{X} \to T(x) = t \in \mathcal{T}$$

Since X is a random variable, so is T(X).

Sufficient Statistic

- A *Statistic* is a function of the sample (only).

$$T : x \in \mathcal{X} \to T(x) = t \in \mathcal{T}$$

Since X is a random variable, so is T(X).

Sufficient Statistic

A Statistic T is *Sufficient* for the statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of X if the conditional distribution of X given $T = t$ is independent of θ.

-
  - How frequenists think about it: $p(x|T(x), \theta) = p(x|T(x))$
  - How bayesians think about it: $p(\theta|x, T(x)) = p(\theta|T(x))$

Sufficient Statistic

- A *Statistic* is a function of the sample (only).

$$T : x \in \mathcal{X} \to T(x) = t \in \mathcal{T}$$

  Since X is a random variable, so is T(X).

Sufficient Statistic

A Statistic T is *Sufficient* for the statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of X if the conditional distribution of X given $T = t$ is independent of θ.

-
  - How frequenists think about it: $p(x|T(x), \theta) = p(x|T(x))$
  - How bayesians think about it: $p(\theta|x, T(x)) = p(\theta|T(x))$
  - Creates an information bottleneck between our data and our parameters.

Sufficient Statistic

- A *Statistic* is a function of the sample (only).

$$T : x \in \mathcal{X} \to T(x) = t \in \mathcal{T}$$

Since X is a random variable, so is T(X).

Sufficient Statistic

A Statistic T is *Sufficient* for the statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of X if the conditional distribution of X given $T = t$ is independent of θ.

-
  - How frequenists think about it: $p(x|T(x), \theta) = p(x|T(x))$
  - How bayesians think about it: $p(\theta|x, T(x)) = p(\theta|T(x))$
  - Creates an information bottleneck between our data and our parameters.
  - Data reduction technique. You can throw away the data as long as the statistics are known. No more information for inference on θ

Sufficiency

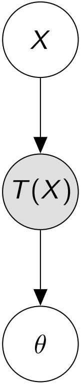


Figure: a) Bayesian

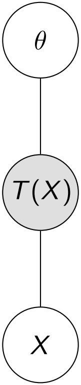


Figure: c) Neyman Factorization $p(x, T(x), \theta) = \psi_1(T(x), \theta)\psi_2(T(x), x)$

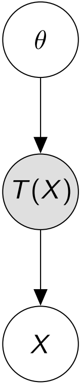


Figure: b) Frequenist

Factorization Theorem

Theorem (Neyman Factorization Theorem)

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model, with probability function $p(\cdot;\theta)$. A statistic T is sufficient for $\mathcal{P}$ if and only if there exist non-negative functions $g(\cdot;\theta)$ and h such that the probability function satisfies:

$$p(\mathrm{x};\theta) = g(T(\mathrm{x});\theta)h(x)$$

- *θ is connected with X only through $T(X)$.*
- *We can see why $T(x)$ is a sufficient statistic in exponential families.*

Factorization Theorem

Theorem (Neyman Factorization Theorem)

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model, with probability function $p(\cdot;\theta)$. A statistic T is sufficient for $\mathcal{P}$ if and only if there exist non-negative functions $g(\cdot;\theta)$ and h such that the probability function satisfies:

$$p(\mathrm{x};\theta) = g(T(\mathrm{x});\theta)h(x)$$

- *θ is connected with X only through $T(X)$.*
- *We can see why $T(x)$ is a sufficient statistic in exponential families.*

Example ($X \sim \mathcal{N}(\mu, \sigma)$ Draw n i.i.d. samples)

$$p(\mathrm{x};\theta) = (\frac{1}{2\pi\sigma^2})^{n/2} \exp\left(-\frac{1}{2\sigma^2}(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2)\right) \implies$$

$$T(\mathrm{x}) = (\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2)$$ are *sufficient statistics*. h=1, g rest.

- With n i.i.d gaussian samples, the most we can infer about θ is in the *empirical mean* and the *empirical variance*

Minimal Sufficient Statistics

- A Sufficient Statistic *partitions* the sample space.

Minimal Sufficient Statistics

- A Sufficient Statistic *partitions* the sample space.
- We would like to find the *coarsest* partition possible. We call that a *minimal sufficient* statistic. (Assume strictly m-parametric family)

Minimal Sufficient Statistics

- A Sufficient Statistic *partitions* the sample space.
- We would like to find the *coarsest* partition possible. We call that a *minimal sufficient* statistic. (Assume strictly m-parametric family)
- It is essentially a function of any other sufficient statistic $(T'(x) = T'(y) \implies T(x) = T(y), \forall x, y \in \mathcal{X})$

Minimal Sufficient Statistics

- A Sufficient Statistic *partitions* the sample space.
- We would like to find the *coarsest* partition possible. We call that a *minimal sufficient* statistic. (Assume strictly m-parametric family)
- It is essentially a function of any other sufficient statistic $(T'(x) = T'(y) \implies T(x) = T(y), \forall x, y \in \mathcal{X})$
- Criterion: $\frac{p(x;\theta)}{p(y;\theta)}$ independent of $\theta, \forall x, y \in \mathcal{X} \implies T(x) = T(y)$

Minimal Sufficient Statistics

- A Sufficient Statistic *partitions* the sample space.
- We would like to find the *coarsest* partition possible. We call that a *minimal sufficient* statistic. (Assume strictly m-parametric family)
- It is essentially a function of any other sufficient statistic $(T'(x) = T'(y) \implies T(x) = T(y), \forall x, y \in \mathcal{X})$
- Criterion: $\frac{p(x;\theta)}{p(y;\theta)}$ independent of $\theta, \forall x, y \in \mathcal{X} \implies T(x) = T(y)$
- For the example above $T(x) = (\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2)$ is a minimal sufficient statistic.

Minimal Sufficient Statistics

- A Sufficient Statistic *partitions* the sample space.
- We would like to find the *coarsest* partition possible. We call that a *minimal sufficient* statistic. (Assume strictly m-parametric family)
- It is essentially a function of any other sufficient statistic $(T'(x) = T'(y) \implies T(x) = T(y), \forall x, y \in \mathcal{X})$
- Criterion: $\frac{p(x;\theta)}{p(y;\theta)}$ independent of $\theta, \forall x, y \in \mathcal{X} \implies T(x) = T(y)$
- For the example above $T(x) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$ is a minimal sufficient statistic.
- Note: Different than the *strictly m-parametric* condition or the minimality in m, we discussed above.

Minimal Sufficiency in Exponential Family

- For a sample, of i.i.d. random variables from a strictly m-parametric exponential family it holds:

$$T_{(n)}(x) = (\sum_{i=1}^{n} T_1(x_i), \cdots, \sum_{i=1}^{n} T_m(x_i))$$

  is a minimal sufficient statistic.
- Note: Dimensions of T fixed; independent of n!
- Sufficiency still holds for non-strict exponential families, minimality not. So curved families still have sufficient statistics.
- In some sense, exponential family is the *only* family with minimal, finite-dimensional, sufficient statistics (Pitman–Koopman–Darmois theorem)
  - $U(0, \theta)(?)$: non fixed domain!

Conjugacy

- Conjugate Priors: Important for Bayesian Statistics, inference.

Conjugacy

- Conjugate Priors: Important for Bayesian Statistics, inference.
- For a likelihood function in exponential *the conjugate prior is again an exponential family*.

Conjugacy

- Conjugate Priors: Important for Bayesian Statistics, inference.
- For a likelihood function in exponential *the conjugate prior is again an exponential family*.
- Conjugate prior p_π of the parameter η of an exponential family.

Conjugacy

- Consider the exponential family:

$$p(x|\eta) = h(x) \exp\left(\eta^\top T(x) - \ln A(\eta)\right)$$

Conjugacy

- Consider the exponential family:

$$p(x|\eta) = h(x) \exp\left(\eta^\top T(x) - \ln A(\eta)\right)$$

- Its conjugate is the exponential family:

$$p_\pi(\eta|\alpha, \nu) = \exp\left(\begin{pmatrix}\alpha \\ \nu\end{pmatrix}^\top \begin{pmatrix}\eta \\ A(\eta)\end{pmatrix} - \ln F(\alpha, \nu)\right)$$

Conjugacy

- Consider the exponential family:

$$p(x|\eta) = h(x)\exp\left(\eta^\top T(x) - \ln A(\eta)\right)$$

- Its conjugate is the exponential family:

$$p_\pi(\eta|\alpha,\nu) = \exp\left(\begin{pmatrix}\alpha \\ \nu\end{pmatrix}^\top \begin{pmatrix}\eta \\ A(\eta)\end{pmatrix} - \ln F(\alpha,\nu)\right)$$

  - $F(\alpha,\nu) = \int_{\mathcal{H}} \exp\left(\begin{pmatrix}\alpha \\ \nu\end{pmatrix}^\top \begin{pmatrix}\eta \\ A(\eta)\end{pmatrix}\right) d\eta$
  - ν : pseudo-observations from prior.
  - α: effective amount these pseudo-observations contribute to the sufficient statistic (vector).

Posterior and Bayesian Inference

- Posterior (n iid samples):

$$p(\eta|X,\alpha,\nu) = p_\pi(\eta|\alpha + \sum_{i=1}^{n} T(x_i), \nu + n)$$

- Same family as the prior (T,h)
- Predictive (1 sample):

$$p(x) = \int_\eta p(x|\eta)p_\pi(\eta|\alpha,\nu)d\eta = h(x)\frac{F(T(x)+\alpha,\nu+1)}{F(\alpha,\nu)}$$

- Calculation of $F(\alpha,\nu)$, often intractable! Even if $A(\eta)$ is known.

Parameter Estimation in Exponential Families

- for N iid datapoints, parameter estimation in $\mathcal{O}(N)$ from sufficient statistics.
- $\mathcal{H}$ is a convex set.
- As we saw in the examples above (for full rank families):
  - $\mathbb{E}[T(X)] = \nabla_\eta A(\eta)$
  - $[Cov(T_i(X), T_j(X))] = Hessian[A(\eta)]$
  - Thus $A(\eta)$ is convex (strictly for minimal families).
- $ll(\eta) = \ln\left(\prod_{n=1}^{N} h(x_n)\right) + \eta^\top\left(\sum_{n=1}^{N} T(x_n)\right) - NA(\eta)$
  - $\nabla_\eta ll(\eta) = 0 \iff \mathbb{E}[T(X)] = \nabla_\eta A(\eta) = \sum_{n=1}^{N} T(x_n)$
  - $Hessian[ll(\eta)] = Hessian[A(\eta)] = [Cov(T_i, T_j)] \succcurlyeq 0$
  - Data appear only through the sufficient stat. for η estimation.
  - Not always closed form! But iterative algorithms converge!
- Note: $ll(\eta)$ concave (in η), does not mean pdf is unimodal (in x). See Beta(1/2,1/2).

Cramer-Rao, Rao-Blackwell and more

- Unbiased MLE estimator: $\mathbb{E}[\sum_{n=1}^{N} T(x_n)] = \mathbb{E}[T(X)]$ (not for η, but there exists an equivalent *mean parametrization*)

Cramer-Rao, Rao-Blackwell and more

- Unbiased MLE estimator: $\mathbb{E}[\sum_{n=1}^{N} T(x_n)] = \mathbb{E}[T(X)]$ (not for η, but there exists an equivalent *mean parametrization*)
- Attains Cramer-Rao bound: $I(\eta) = -\mathbb{E}[\frac{\partial^2 \ln p(X|\eta)}{d\eta^2}] = Cov(T(X))$ (optimal)

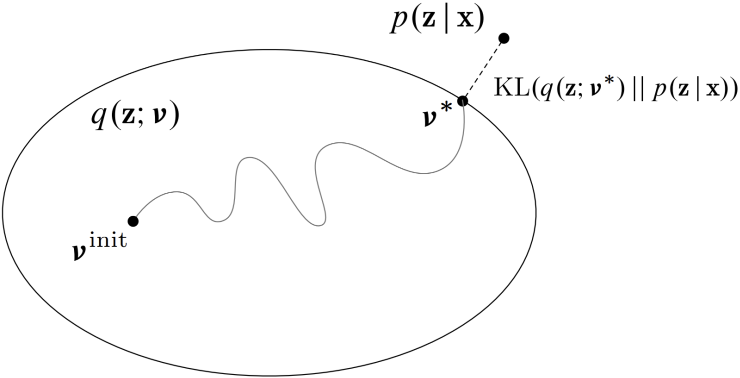
Cramer-Rao, Rao-Blackwell and more

- Unbiased MLE estimator: $\mathbb{E}[\sum_{n=1}^{N} T(x_n)] = \mathbb{E}[T(X)]$ (not for η, but there exists an equivalent *mean parametrization*)
- Attains Cramer-Rao bound: $I(\eta) = -\mathbb{E}[\frac{\partial^2 \ln p(X|\eta)}{d\eta^2}] = Cov(T(X))$ (optimal)
- Other interesting properties:
    - Solutions to Maximum Entropy problems, consistent with constraints in expected values. For example, Normal is the maximum entropy distribution, out of all those with a bounded variance.
    - Bregman Divergence between parameters (wrt log partition) is the KL divergence between the distributions with those parameters.
    - Method of Moments converge to the maximum likelihood estimator.
    - Full rank exponential family $\implies$ Complete and Sufficient stat. $\implies$ UMVUE for mean-value param. (Rao-Blackwell, Lehmann-Scheffe)
    - Conditional Conjugacy

Table of Contents

Black-box VI [Ranganath, Gerrish, and Blei 2014]

- Variational Inference:
  - Recasts posterior inference as an optimization problem.
  - Approximate only if the variational family does not include the posterior; or the optimization algorithm cannot find it.
  - Always strive for richer variational families. No notion of "overfitting".



Black-box VI

- Variational Family does not need to be parametric. Calculus of Variations is functional optimization! Still waiting for mathematicians to step up...
- For now, we posit a parametric posterior $q(z;\nu) \in \mathcal{Q}$ over the latent variables and optimize over ν:
  - (Typical) mean-field family: $q(z) = \prod_{i=1}^{K} q_i(z_i)$
  - Amortization: $q(z_i|x_i;\nu)$, shared ν
  - (Typical) Algorithms: SGD performs VI. [Chaudhari and Soatto 2018]
  - (Typical) Objective: $KL(q(z;\nu)||p(z|x)) \to \min_\nu$
    - KL divergence is problematic (∞ if no overlap; in high dim. spaces, no overlap $\implies$ no gradient.
    - Mode seeking behavior of $KL(q||p)$ plus mean-field assumptions underestimate the variance!

Black-box VI

- Optimizing ν is one task, we also need to optimize θ, the parameters of our model!
- ELBO in Reverse:
  - $ELBO(\theta, q) = \mathbb{E}_{x_i \in p_{data}} \mathbb{E}_{z_i \sim q_i(z_i|x_i)}[\ln(p(z_i)p(x_i|z_i;\theta)) - \ln q_i(z_i|x_i)]$
  - $ELBO(\theta, q) = -\mathbb{E}_{x_i \sim p_{data}}[KL(q_i(z_i|x_i)||p(z_i|x_i;\theta))] + \ln p(D;\theta)$
  - Maximization: $\max_\theta \max_{q \in \mathcal{Q}} ELBO(\theta, q) \leq \max_\theta \ln p(D;\theta)$
  - $" = "$ if $p(z|x) \in \mathcal{Q}$

Black-Box VI

Motivation behind BBVI:

- In need of scalable variational inference (massive data). In need of generic variational inference, no model specific bounds.
- Follow noisy, unbiased estimates of the gradient! Double stochasticity in sampling dataset and approximate posterior.
- Reformulate ELBO's gradient as an expectation. Use Monte Carlo sampling.

Black-box VI

- How to optimize functions of the form:

$$\mathcal{L}(\nu, \theta) = \mathbb{E}_{q_\nu(z)}[f(z; \theta)]$$

- "Easy" w.r.t θ, Difficult w.r.t ν (Stochastic Optimization!)
- REINFORCE Estimator (or log-derivative trick, or score-function estimator)

$$\nabla_\nu \mathcal{L}(\nu, \theta) = \mathbb{E}_{q_\nu(z)}[\nabla_\nu \ln q_\nu(z) f(z)]$$

- Note: $E_q[\nabla_\nu \ln q_\nu(z)] = 0 \implies$ oscillates. around 0.

Black-box VI

About the log-derivative trick:

- The estimator has HUGE variance, not much better than random search ([Mania, Guy, and Recht 2018]). HUGE variance in stochastic optimization theory $\implies$ slow convergence.
- Pretends to be a 1-st order estimator, it is actually a 0-th order estimator (no $f(z;\theta)$ gradient)
- Control Variate Techniques can reduce it but not enough to be practical (if not careful, can magnify it).
- We have seen another estimator (Reparametrization trick) ([Kingma and Welling 2014]).
  - Very small variance when it is applicable (location-scale distributions; $z = T(\epsilon, \nu)$).
  - Latent has to be continuous! Otherwise no differentiable reparametrization exists! (Gumbel-Softmax relaxation (?)).

Black-box VI

Black-box Variational Inference Necessary Criteria:

- Sample $q(z;\nu)$ easily.
- Evaluate $\ln p(x,z), \ln q(z)$ easily.
- Compute $\nabla_\nu \ln q_\nu(z)$ easily.

Table of Contents

Model Specification:

$$
\begin{aligned}
p(t_i) &= \mathcal{N}(0, I) \\
p(x_i|t_i) &= \mathcal{N}(Wt_i + b, \Sigma) \\
\implies p(x_i) &= {\color{red}(?)}
\end{aligned}
$$

Deep Exponential Families

- Model Specification :

$$
\begin{aligned}
p(t_i) &= \mathcal{N}(0, I) \\
p(x_i|t_i) &= \mathcal{N}(Wt_i + b, \Sigma) \\
\implies p(x_i) &= \mathcal{N}(b, WW^\top + \Sigma)
\end{aligned}
$$

Deep Exponential Families

- Model Specification :

$$
\begin{aligned}
p(t_i) &= \mathcal{N}(0, I) \\
p(x_i|t_i) &= \mathcal{N}(Wt_i + b, \Sigma) \\
\implies p(x_i) &= \mathcal{N}(b, WW^\top + \Sigma)
\end{aligned}
$$

- Model Specification:

$$
\begin{aligned}
p(t_i) &= \mathcal{N}(0, I) \\
p(x_i|t_i) &= \mathcal{N}(ResNet(t_i), \Sigma) \\
\implies p(x_i) &= {\color{red}(?)}
\end{aligned}
$$

Deep Exponential Families

- Model Specification :

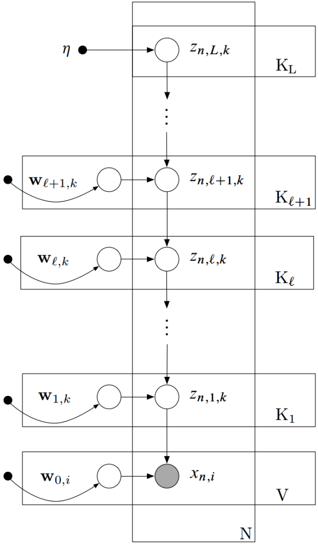
$$p(t_i) = \mathcal{N}(0, I)$$
$$p(x_i|t_i) = \mathcal{N}(Wt_i + b, \Sigma)$$
$$\implies p(x_i) = \mathcal{N}(b, WW^\top + \Sigma)$$

- Model Specification:

$$p(t_i) = \mathcal{N}(0, I)$$
$$p(x_i|t_i) = \mathcal{N}(ResNet(t_i), \Sigma)$$
$$\implies p(x_i) = (?)$$

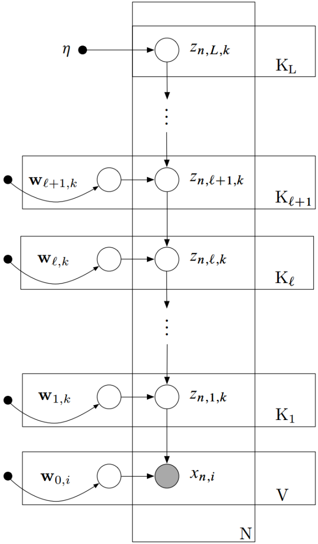
- Indicates that Deep Exponential Families will be more expressive than shallow ones.

Generative Model: The story of our data



- Chain Exponential Family in a hierarchy.
- Each layer's draw is input to the natural parameters of the next.
- Plate notation.
  - η, hyperparameters of $w_{l \in L}$: Shared (for data and z-nodes)
  - $w_{l,k}$ shared across data, local for each z-node.

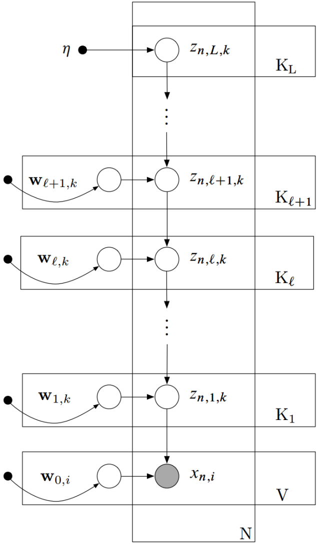
Generative Model: The story of our data



- Model specification (for each x_n, $n \in [N]$):

$$\begin{aligned}
p(z_{n,L,k}) &= ExpFam_L(\eta) \ \forall k \in [K_L] \\
p(z_{n,l,k} | \boldsymbol{z}_{n,l+1}, \boldsymbol{w}_{l,k}) &= ExpFam_l(g_l(\boldsymbol{z}_{n,l+1}^\top \boldsymbol{w}_{l,k})) \\
p(x_{n,i} | \boldsymbol{z}_{n,1}, \boldsymbol{w}_{0,i}) &= ExpFam_0(g_0(\boldsymbol{z}_{n,1}^\top \boldsymbol{w}_{0,i})) \\
W_l &\sim p(W_l)
\end{aligned}$$

Generative Model: The story of our data

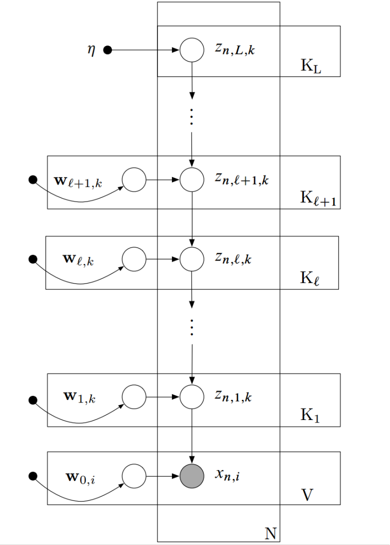


- Model specification (for each x_n, $n \in [N]$):

$$\begin{aligned} p(z_{n,L,k}) &= ExpFam_L(\eta)\ \forall k \in [K_L] \\ p(z_{n,l,k}|\boldsymbol{z}_{n,l+1}, \boldsymbol{w}_{l,k}) &= ExpFam_l(g_l(\boldsymbol{z}_{n,l+1}^\top \boldsymbol{w}_{l,k})) \\ p(x_{n,i}|\boldsymbol{z}_{n,1}, \boldsymbol{w}_{0,i}) &= ExpFam_0(g_0(\boldsymbol{z}_{n,1}^\top \boldsymbol{w}_{0,i})) \\ W_l &\sim p(W_l) \end{aligned}$$

- $z_{n,l,k}$: scalar.
  $\boldsymbol{z}_{n,l+1}$: K_{l+1}-vector.
  $\boldsymbol{w}_{l,k}$: K_{l+1}-vector.
  K_l vectors like $\boldsymbol{w}_{l,k}$ in layer l.
  $\boldsymbol{W}_l \in \mathbb{R}^{K_l \times K_{l+1}}$

Generative Model: The story of our data



- Model specification (for each x_n, $n \in [N]$):

$$p(z_{n,L,k}) = ExpFam_L(\eta)\ \forall k \in [K_L]$$
$$p(z_{n,l,k}|\boldsymbol{z}_{n,l+1}, \boldsymbol{w}_{l,k}) = ExpFam_l(g_l(\boldsymbol{z}_{n,l+1}^\top \boldsymbol{w}_{l,k}))$$
$$p(x_{n,i}|\boldsymbol{z}_{n,1}, \boldsymbol{w}_{0,i}) = ExpFam_0(g_0(\boldsymbol{z}_{n,1}^\top \boldsymbol{w}_{0,i}))$$
$$W_l \sim p(W_l)$$

- $z_{n,l,k}$: scalar.
  $\boldsymbol{z}_{n,l+1}$: K_{l+1}-vector.
  $\boldsymbol{w}_{l,k}$: K_{l+1}-vector.
  K_l vectors like $\boldsymbol{w}_{l,k}$ in layer l.
  $\boldsymbol{W}_l \in \mathbb{R}^{K_l \times K_{l+1}}$
- $ExpFam_l()$ can be different across layers. But it has to be defined a priory.

Link Function

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

As we have already seen:

$$\mathbb{E}_{z_{n,l,k}}[T(z_{n,l,k})] = \nabla_\eta a(\eta)\Big|_{g_l(\boldsymbol{w}_{l,k}^\top \boldsymbol{z}_{n,l+1})}$$

- Misconception: Sufficient Statistics are fixed a priori. Expected Sufficient Statistics are functions of the parameters.
- Two kinds of non-linearities (from $\boldsymbol{w}$ to z): 1) link function g(x), 2) log-normalizer's gradient $\nabla_\eta a(\eta)$.

Example Link Function

- We reformulated the Binomial into its natural parametrization, above.
- If we do the same for Bernoulli we get:

$$p(z_{n,l,k}|\eta) = \exp\left(\eta z_{n,l,k} - \ln\left(1 + \exp\eta\right)\right)$$

where $z_{n,l,k} \in \{0, 1\},\ \eta = |_{a.s.} g_l(\boldsymbol{z}_{n,l+1}^{\top} \boldsymbol{w}_{l,k})$

- With identity link function:

$$\mathbb{E}[z_{n,l,k}|\eta] = \frac{1}{1 + \exp\left(-\eta\right)} = \sigma(\boldsymbol{z}_{n,l+1}^{\top} \boldsymbol{w}_{l,k})$$

Sigmoid Belief Network!

Sparse Gamma DEF

- Gamma PDF:

$$Z \sim Gamma(\alpha, \beta) \iff p(z) = \frac{1}{\Gamma(a)} \beta^{\alpha} z^{\alpha-1} e^{-\beta z}, \; z, \alpha, \beta > 0$$

Sparse Gamma DEF

- Gamma PDF:

$$Z \sim Gamma(\alpha, \beta) \iff p(z) = \frac{1}{\Gamma(a)} \beta^{\alpha} z^{\alpha - 1} e^{-\beta z}, \; z, \alpha, \beta > 0$$

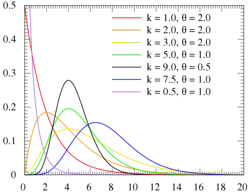
- Shape parameter α, Rate parameter β.

Sparse Gamma DEF

- Gamma PDF:

$$Z \sim Gamma(\alpha, \beta) \iff p(z) = \frac{1}{\Gamma(a)} \beta^{\alpha} z^{\alpha-1} e^{-\beta z}, \ z, \alpha, \beta > 0$$

- Shape parameter α, Rate parameter β.

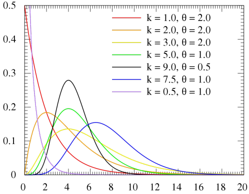


Sparse Gamma DEF

- Gamma PDF:

$$Z \sim Gamma(\alpha, \beta) \iff p(z) = \frac{1}{\Gamma(a)} \beta^{\alpha} z^{\alpha-1} e^{-\beta z}, \; z, \alpha, \beta > 0$$

- Shape parameter α, Rate parameter β.



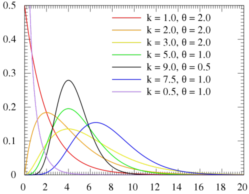
- $\mathbb{E}[z] = \alpha/\beta, \; Var[z] = \alpha/\beta^2$

Sparse Gamma DEF

- Gamma PDF:

$$Z \sim Gamma(\alpha, \beta) \iff p(z) = \frac{1}{\Gamma(a)} \beta^{\alpha} z^{\alpha-1} e^{-\beta z}, \; z, \alpha, \beta > 0$$

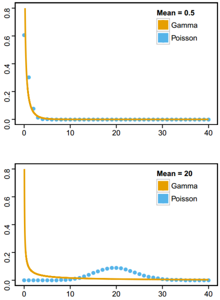
- Shape parameter α, Rate parameter β.



- $\mathbb{E}[z] = \alpha/\beta, \; Var[z] = \alpha/\beta^2$
- $p(z) = z^{-1} \exp\left((\alpha, \beta) \cdot (\ln z, -z) - \ln \Gamma(\alpha) - \alpha \ln \beta\right)$

Sparse Gamma DEF

- $p(W_l) \sim Gamma(\alpha_{W_l}, \beta_{W_l})$ to ensure $z_{l+1}^{\top} w_{l,k} > 0$
- Set $\alpha_l, \alpha_{W_l} < 1 \implies$ soft spike-slab prior (sparse gamma).
- Sparse Gamma DEFs for documents means that an observable does not need to express every "super-topic" in the "concept" it expresses.
- $g_\alpha(\cdot) = \alpha_l = const,\ g_\beta(\cdot) = \frac{\alpha_l}{\mathbf{z}_{n,l+1}^{\top} \mathbf{w}_{l,k}} \implies$
- $\mathbb{E}[z_{n,l,k}] = \mathbf{z}_{n,l+1}^{\top} \mathbf{w}_{l,k}$



z-Dist	$\mathbf{z}_{\ell+1}$	W-dist	$\mathbf{w}_{\ell,k}$	g_ℓ	$\mathrm{E}[T(z_{\ell,k})]$
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.

Inference BBVI

- Compounding Exponential Families is not an Exponential Family (Except Conjugacy!)
- Black-box Variational Inference to train DEFs.
  - Monte Carlo approximation of variational objective (and gradient).
  - Stochastic Optimization routine, follows noisy unbiased gradients.
- Mean-field family:

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

- $q(\boldsymbol{W}_l|\xi_l), q(\boldsymbol{z}_{n,l})$ fully factorized. Same family as the model distribution p. In actual model, $\boldsymbol{z}_{n,l} \not\perp\!\!\!\perp \boldsymbol{z}_{m,l}|\{x_m, x_n\}$ (common cause W's)
- $z_{n,l,k} \sim ExpFam_l(\lambda_{n,l,k})$

Inference BBVI

- $\mathcal{L}(\lambda, \xi) = \mathbb{E}_{q(z,W;\xi,\lambda)}[\ln p(x, z, W) - \ln q(z, W)]$

Inference BBVI

- $\mathcal{L}(\lambda, \xi) = \mathbb{E}_{q(z,W;\xi,\lambda)}[\ln p(x, z, W) - \ln q(z, W)]$
- Intractable to compute estimation! But we "only" need to compute derivative!

Inference BBVI

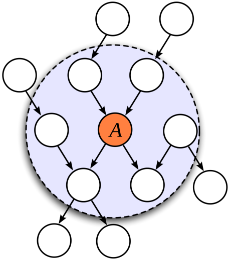
- $\mathcal{L}(\lambda, \xi) = \mathbb{E}_{q(z,W;\xi,\lambda)}[\ln p(x, z, W) - \ln q(z, W)]$
- Intractable to compute estimation! But we "only" need to compute derivative!
- $\nabla_{\lambda,\xi}\mathcal{L}(\lambda, \xi) = \mathbb{E}_q[\nabla_{\lambda,\xi} \ln q \ (\ln p(x, W, z) - \ln q(W, z))]$
  $\approx \frac{1}{S}\sum_{i=1}^{S} \nabla_{\lambda,\xi} \ln q(z_s, W_s|\lambda, \xi) \ (\ln p(x, W_s, z_s) - \ln q(W_s, z_s|\lambda, \xi))$
  $z_s, W_s \sim q(z, W|\lambda, \xi)$

Inference BBVI

- $\mathcal{L}(\lambda, \xi) = \mathbb{E}_{q(z,W;\xi,\lambda)}[\ln p(x, z, W) - \ln q(z, W)]$
- Intractable to compute estimation! But we "only" need to compute derivative!
- $\nabla_{\lambda,\xi}\mathcal{L}(\lambda, \xi) = \mathbb{E}_q[\nabla_{\lambda,\xi} \ln q \ (\ln p(x, W, z) - \ln q(W, z))]$
  $\approx \frac{1}{S}\sum_{i=1}^{S} \nabla_{\lambda,\xi} \ln q(z_s, W_s|\lambda, \xi) \ (\ln p(x, W_s, z_s) - \ln q(W_s, z_s|\lambda, \xi))$
  $z_s, W_s \sim q(z, W|\lambda, \xi)$
- Each row in the jacobian is actually sparse (for this graphical model). The dependency on $z_{n,l,k}$ is only through its markov blanket. (Computational graph resolves it).

Inference BBVI

- $\mathcal{L}(\lambda, \xi) = \mathbb{E}_{q(z,W;\xi,\lambda)}[\ln p(x, z, W) - \ln q(z, W)]$
- Intractable to compute estimation! But we "only" need to compute derivative!
- $\nabla_{\lambda,\xi}\mathcal{L}(\lambda, \xi) = \mathbb{E}_q[\nabla_{\lambda,\xi} \ln q \ (\ln p(x, W, z) - \ln q(W, z))]$
  $\approx \frac{1}{S}\sum_{i=1}^{S} \nabla_{\lambda,\xi} \ln q(z_s, W_s|\lambda, \xi) \ (\ln p(x, W_s, z_s) - \ln q(W_s, z_s|\lambda, \xi))$
  $z_s, W_s \sim q(z, W|\lambda, \xi)$
- Each row in the jacobian is actually sparse (for this graphical model). The dependency on $z_{n,l,k}$ is only through its markov blanket. (Computational graph resolves it).



Inference BBVI on DEFs

- Coordinate gradients:

$$\nabla_{\lambda_{n,l,k}} \mathcal{L} = \mathbb{E}_q[\nabla_{\lambda_{n,l,k}} \ln q(z_{n,l,k})(\ln p_{n,l,k}(x, z, W) - \ln q(z_{n,l,k})]$$

- Markov Blanket of $z_{n,l,k}$:

$$\ln p_{n,l,k}(x, z, W) = \ln p(z_{n,l,k}|z_{n,l+1}, W_{l,k}) + \ln p(z_{n,l-1}|z_{n,l}, W_{l-1})$$

Algorithm 1 BBVI for DEFs

Input: data X, model p, L layers.
Initialize λ, ξ 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

Text Modelling

- Hierarchical Clustering of words into topics, groups of topics etc.

Text Modelling

- Hierarchical Clustering of words into topics, groups of topics etc.
- N Documents $\{x_n\}_{n=1}^N$.
  V-dimensional $x_{n,i} = \#$ term i in doc n. (Observable)

Text Modelling

- Hierarchical Clustering of words into topics, groups of topics etc.
- N Documents $\{x_n\}_{n=1}^{N}$.
  V-dimensional $x_{n,i} = \#$ term i in doc n. (Observable)
- Observation Likelihood: $p(x_{n,i}|\boldsymbol{z}_{n,1}, \boldsymbol{w}_{0,i}) = Poisson(g(\boldsymbol{z}_{n,1}^{\top}\boldsymbol{w}_{0,i}))$

Text Modelling

- Hierarchical Clustering of words into topics, groups of topics etc.
- N Documents $\{x_n\}_{n=1}^{N}$.
  V-dimensional $x_{n,i} = \#$ term i in doc n. (Observable)
- Observation Likelihood: $p(x_{n,i}|\boldsymbol{z}_{n,1}, \boldsymbol{w}_{0,i}) = Poisson(g(\boldsymbol{z}_{n,1}^{\top}\boldsymbol{w}_{0,i}))$
- $[\boldsymbol{W}_0]_{\boldsymbol{i,j}} \sim Gamma(\alpha, \beta) \implies \boldsymbol{W}_0$ puts positive mass on groups of terms : "topics"!

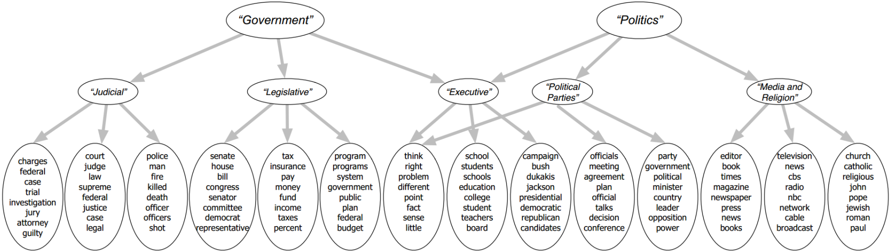
Text Modelling

- Hierarchical Clustering of words into topics, groups of topics etc.
- N Documents $\{x_n\}_{n=1}^N$.
  V-dimensional $x_{n,i} = \#$ term i in doc n. (Observable)
- Observation Likelihood: $p(x_{n,i}|\boldsymbol{z}_{n,1}, \boldsymbol{w}_{0,i}) = Poisson(g(\boldsymbol{z}_{n,1}^\top \boldsymbol{w}_{0,i}))$
- $[\boldsymbol{W}_0]_{\boldsymbol{i},\boldsymbol{j}} \sim Gamma(\alpha, \beta) \implies \boldsymbol{W}_0$ puts positive mass on groups of terms : "topics"!
- "topics" $z_{n,1,k} = \#$ topic k in document n
- "super topics" $z_{n,2,k} = \#$ super topic k in document n.
- "concepts" $z_{n,3,k}$, etc.

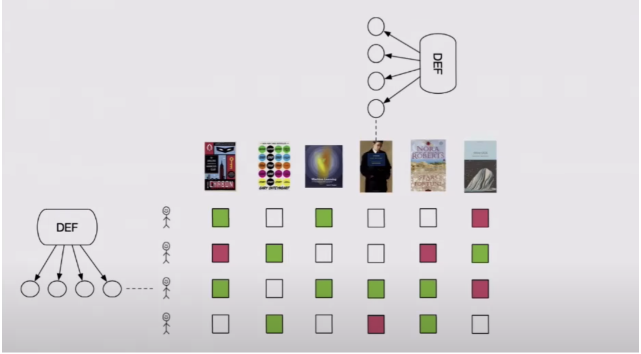
Text Modelling

- Hierarchical Clustering of words into topics, groups of topics etc.
- N Documents $\{x_n\}_{n=1}^N$.
  V-dimensional $x_{n,i} = \#$ term i in doc n. (Observable)
- Observation Likelihood: $p(x_{n,i}|\boldsymbol{z}_{n,1}, \boldsymbol{w}_{0,i}) = Poisson(g(\boldsymbol{z}_{n,1}^\top \boldsymbol{w}_{0,i}))$
- $[\boldsymbol{W}_0]_{\boldsymbol{i,j}} \sim Gamma(\alpha, \beta) \implies \boldsymbol{W}_0$ puts positive mass on groups of terms : "topics"!
- "topics" $z_{n,1,k} = \#$ topic k in document n
- "super topics" $z_{n,2,k} = \#$ super topic k in document n.
- "concepts" $z_{n,3,k}$, etc.
- $p(\boldsymbol{z}_{n,1}|\boldsymbol{z}_{n,2}, \boldsymbol{W}_1) =$ "distribution of topics in a document given the super-topics in the same document". Bernoulli,Sparse Gamma, Poisson etc.

Text Modelling



Double DEF



- Draw user preferences: $\theta_i \sim DEF()$
- Draw item attributes: $\beta_j \sim DEF()$
- Draw rating: $y_{i,j} = f(\theta_i^\top \beta_j)$

Table of Contents

Discussion

- Differences with Deep GPs
- BBVI in Probabilistic Programming.
- Directed versus Undirected Models (Explaining away)
- Survival Analysis

Bibliography

Pratik Chaudhari and Stefano Soatto. *Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks*. 2018. arXiv: 1710.11029 [cs.LG].

Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2014. arXiv: 1312.6114 [stat.ML].

Horia Mania, Aurelia Guy, and Benjamin Recht. *Simple random search provides a competitive approach to reinforcement learning*. 2018. arXiv: 1803.07055 [cs.LG].

Rajesh Ranganath, Sean Gerrish, and David Blei. "Black Box Variational Inference". In: ed. by Samuel Kaski and Jukka Corander. Vol. 33. Proceedings of Machine Learning Research. Reykjavik, Iceland: PMLR, 22–25 Apr 2014, pp. 814–822. URL: http://proceedings.mlr.press/v33/ranganath14.html.