LECTURE 11: EXPONENTIAL FAMILY AND GENERALIZED LINEAR MODELS

HANI GOODARZI AND SINA JAFARPOUR

1. EXPONENTIAL FAMILY.

Exponential family comprises a set of flexible distribution ranging both continuous and discrete random variables. The members of this family have many important properties which merits discussing them in some general format. Many of the probability distributions that we have studied so far are specific members of this family:

- Gaussian: $\mathbb{R}^p$
- Multinomial: $categorical$
- Bernoulli: binary $\{0, 1\}$
- Binomial: counts of success/failure
- Von mises: sphere
- Gamma: $\mathbb{R}^+$
- Poisson: $\mathbb{N}^+$
- Laplace: $\mathbb{R}^+$
- Exponential: $\mathbb{R}^+$
- Beta: $(0, 1)$
- Dirichlet: Δ (Simplex)
- Weibull: $\mathbb{R}^+$
- Weishart: symmetric positive-definite matrices

All these distributions follow the general format:

$$p(x|\eta) \;=\; h(x)\ \exp\left(\eta^\top\, t(x) \;-\; a(\eta)\right); \tag{1}$$

where, η is called "natural parameter", $t(x)$ is "sufficient statistic" (a statistic is a function of data), $h(x)$ is the "underlying measure" and $a(\eta)$ is called "log normalizer", which ensures that the distribution integrates to one. Hence,

$$a(\eta) = \log \int h(x)\ \exp\left(\eta^\top t(x)\right) dx.$$

We start by showcasing a number of known distributions and illustrate that they are indeed members of the exponential family.

1.1. **Bernoulli.** Bernoulli distribution is defined on a binary (0 or 1) random variable using parameter π where $\pi = \Pr(x = 1)$. The Bernoulli distribution can be written as:

$$p(x|\pi) = \pi^x(1-\pi)^{1-x}. \tag{2}$$

In order to convert Equation (2) to the general exponential format (Equation (1)), we rewrite it as,

$$\begin{aligned} p(x|\pi) &= \exp\{\log\left(\pi^x(1-\pi)^{1-x}\right)\} \\ &= \exp\{x\log\pi + (1-x)\log(1-\pi)\} \\ &= \exp\left\{x\log\frac{\pi}{1-\pi} + \log(1-\pi)\right\} \end{aligned} \tag{3}$$

In Equation (3),

- $\eta = \log\frac{\pi}{1-\pi}$,
- $t(x) = x$,
- $a(\eta) = -\log(1-\pi)$,
- and $h(x) = 1$.

To put $a(\eta)$ in its correct form, we use the relationship between η and π:

$$\begin{aligned} \eta &= \log\frac{\pi}{1-\pi} \Rightarrow \\ -\eta &= \log\frac{1-\pi}{\pi} = \log\left(\frac{1}{\pi}-1\right) \Rightarrow \\ e^{-\eta} &= \frac{1}{\pi}-1 \Rightarrow \\ \pi &= \frac{1}{1+e^{-\eta}} = \sigma(\eta) \end{aligned} \tag{4}$$

Consequently,

$$a(\eta) = \log(1+e^{\eta}), \tag{5}$$

and

$$p(x|\eta) = \sigma(-\eta)e^{-\eta x}. \tag{6}$$

1.2. **Multinomial.** Although not discussed in the class, it is important to see this process for the multinomial distribution as well. While the Bernoulli is defined with the parameter π, multinomial has a vector of parameters μ_k where k goes from 1 to M:

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M}\mu_k^{x_k} = \exp\{\sum_{k=1}^{M} x_k\,\log\mu_k\},$$

where $\boldsymbol{x} = (x_1, x_2, \cdots, x_N)^\top$ and $\sum_{k=1}^{M} \mu_k = 1$. Following the same process as Bernoulli, we have:

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \exp\{\boldsymbol{\eta}^\top \boldsymbol{x} + \log(1 + \sum_{k=1}^{M} \eta_k)^{-1}\},$$

where

$$\mu_k = \frac{\exp(\eta_k)}{1 + \sum_j \exp(\eta_j)} = \text{softmax}(k, \boldsymbol{\eta}). \tag{7}$$

1.3. **Poisson.** Poisson is a discrete distribution defined to express the number events that occur in a unit of time or space. This distribution, which is similar to Gaussian distribution but for count data, is given by

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} \exp\{x \log \lambda - \lambda\}, \tag{8}$$

where

- $\eta = \lambda$,
- $t(x) = x$,
- $a(\eta) = \lambda = e^\eta$,
- and $h(x) = \frac{1}{x!}$.

1.4. **Univariate Gaussian.** Similarly, the Gaussian distribution can be also rewritten in terms of the general exponential format;

$$\begin{aligned} P(x|\mu, \sigma^2) &= \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{\mu}{\sigma^2}.x - \frac{1}{2\sigma^2}.x^2 - \frac{1}{2\sigma^2}\mu^2 - \log(\sigma)\right\}, \end{aligned} \tag{9}$$

where

- $\eta = \langle \frac{\mu}{\sigma^2} \quad \frac{-1}{2\sigma^2} \rangle$,
- $t(x) = \langle x \quad x^2 \rangle$,
- $h(x) = \frac{1}{\sqrt{2\pi}}$,
- and $a(\eta) = \frac{\mu^2}{2\sigma^2} + \log(\sigma) = \frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)$.

2. MOMENTS OF EXPONENTIAL FAMILY.

In the family of exponential distributions, the $a(\eta)$ function is in fact the generating function. We show this by derivatizing this term:

$$\begin{aligned} \frac{d\,a(\eta)}{d\eta} &= \frac{d}{d\eta}\left\{\log\left(\int \exp\{\eta^\top t(x)\}h(x)dx\right)\right\} \\ &= \frac{\frac{d}{d\eta}\int \exp\left\{\eta^\top t(x)\right\}h(x)dx}{\int \exp\{\eta^\top t(x)\}h(x)dx} \\ &= \frac{\int t(x)h(x)\exp\{\eta^\top t(x)\}dx}{\int \exp\{\eta^\top t(x)\}h(x)dx} \\ &= \frac{\int t(x)\exp\{\eta^\top t(x)\}h(x)dx}{\exp\{-a(\eta)\}} \\ &= \int t(x)\exp\{\eta^\top t(x) - a(\eta)\}h(x)dx \\ &= \mathbb{E}\left[t(x)\right]. \end{aligned} \tag{10}$$

Likewise, it can be shown that:

$$\frac{d^2\,a(\eta)}{d\eta^2} = \text{Var}\left(t(x)\right) = \mathbb{E}\left[t(x)^2\right] - \mathbb{E}\left[t(x)\right]^2. \tag{11}$$

For example, in Bernoulli distribution we have, $a(\eta) = \log(1+e^\eta)$ is the moment generating function. The first derivative of this function is given by

$$\frac{d\,a(\eta)}{d\eta} = \frac{\frac{d}{d\eta}(1+e^\eta)}{1+e^\eta} = \frac{1}{1+e^{-\eta}} = \pi = \mathbb{E}[X]. \tag{12}$$

In this context, μ defined as $\mathbb{E}\left[t(X)\right]$ can be computed from $\frac{da(\eta)}{d\eta}$ which is solely a function η. This relationship connects μ and η and since the function is convex (i.e. the second derivative is greater than 0), this relationship is invertible. Thus we can define

$$\eta = \Psi(\mu). \tag{13}$$

where Ψ is a function which maps the **natural (canonical) parameters** to the **mean parameter**.

3. GENERALIZED LINEAR MODELS

The *generalized linear model* (GLM) is a powerful generalization of linear regression to more general exponential family. Figure 3 demonstrates the graphical model representation of a generalized linear model. The model is based on the following assumptions:

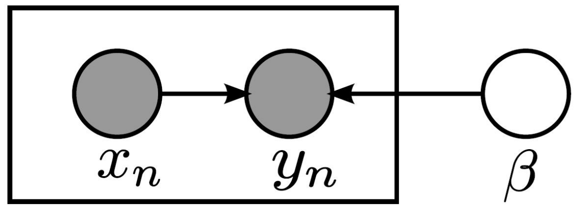

FIGURE 1. Representation of a generalized linear model

- The observed input enters the model through a linear function ($\beta^\top X$).
- The conditional mean of response, is represented as a function of the linear combination:

$$\mathbb{E}[Y|X] \doteq u = f(\beta^\top X). \tag{14}$$

- The observed response is drawn from an exponential family distribution with conditional mean μ, as explained in Equation (13).

Figure 3 summarizes the relationships between the variables in a GLM. It

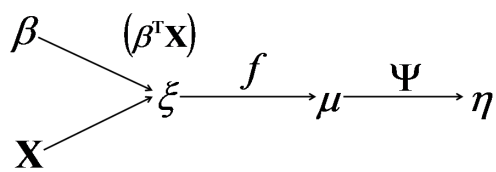

FIGURE 2. Relationship between the variables in a generalized linear model

is usually convenient to work with **overdispersed** exponential families. We assume that the observed response comes from the following probability distribution:

$$p(y|\eta) = h(y, \eta) \exp\left\{\frac{\eta^\top y - a(\eta)}{\sigma}\right\}. \tag{15}$$

For a fixed σ, Equation (15) is an exponential family, but as a function of σ, it is not an exponential family since h is a function of both y and σ.

As a simple example, in the case of linear regression:

- $h(y, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{\frac{-y^2}{2}\right\}$,
- $a(\eta) = \frac{\eta^2}{2}$,
- f : identity,
- Ψ : identity.

Consequently,

$$p(y|\eta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{\frac{-y^2}{2}\right\} \exp\left\{\frac{\eta y - \eta^2/2}{\sigma}\right\} \tag{16}$$
$$= p(y|\eta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{\frac{-(y-\eta)^2}{2\sigma}\right\}.$$

Generally, we have two choice points in order to specify the generalized linear model. The choice of the response function f, or how to treat the linear combination of the observed input, and the choice of the exponential family distribution of the observed output y. Note that Ψ is completely determined by choosing the exponential family. As a result, choosing appropriate response function and exponential family is one of the major tasks in probabilistic modeling, and once the choices are made, the general framework of the exponential family can be applied to the modeled data.