Exponential Family of Distributions

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, Indiana, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

September 6, 2018

References

- C P Robert, The Bayesian Choice: From Decision-Theoretic Motivations to Compulational Implementation, Springer-Verlag, NY, 2001 (online resource)
- A Gelman, JB Carlin, HS Stern and DB Rubin, Bayesian Data Analysis, Chapman and Hall CRC Press, 2nd Edition, 2003.
- J M Marin and C P Robert, The Bayesian Core, Spring Verlag, 2007 (online resource)
- D. Sivia and J Skilling, Data Analysis: A Bayesian Tutorial, Oxford University Press, 2006.
- Bayesian Statistics for Engineering, Online Course at Georgia Tech, B. Vidakovic.
- Chris Bishops' PRML book, Chapter 2
- M. Jordan, An introduction to Probabilistic Graphical Models, Chapter 8 (pre-print)
- Kevin Murphy's, Machine Learning: A probablistic perspective, Chapters 2 and 4

Contents

Exponential Family

- Large family of useful distributions with common properties
  - Bernoulli, beta, binomial, chi-square, Dirichlet, gamma, Gaussian, geometric, multinomial, Poisson, Weibull, . .
- Not in the family:
  - ✓ Uniform,
  - ✓ Student's T,
  - ✓ Cauchy,
  - ✓ Laplace,
  - ✓ Mixture of Gaussians,
  - ✓ . . .
- Variable can be discrete/continuous (or vectors thereof)

Exponential Family

- The exponential family of distributions over $\boldsymbol{x}$, given parameters $\boldsymbol{\eta}$, is defined to be the set of distributions of the form

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} \; or$$

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\}, \; where: A(\boldsymbol{\eta}) = -\log g(\boldsymbol{\eta})$$

  $\boldsymbol{x}$ is scalar/vector, discrete/continuous. ***η are the natural parameters* and *$u(x)$ is referred to as a sufficient statistic.***

- *$g(\boldsymbol{\eta})$ ensures that the distribution is normalized* and satisfies

$$g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}d\boldsymbol{x} = 1$$

- The normalization factor Z and the log of it A are defined as:

$$Z(\boldsymbol{\eta}) = \frac{1}{g(\boldsymbol{\eta})}, \; A(\boldsymbol{\eta}) = -\ln g(\boldsymbol{\eta}) = \ln Z(\boldsymbol{\eta}) = \ln\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}d\boldsymbol{x}$$

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\Big/Z(\boldsymbol{\eta})$$

- The space of $\boldsymbol{\eta}$ for which $\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}d\boldsymbol{x} < \infty$ is the ***natural parameter space***.

Canonical or Natural Parameters

- When the parameter $\boldsymbol{\theta}$ enters the exponential family as $\eta(\theta)$, we write the probability density of the exponential family as follows:

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x}) g\left(\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)\right) \exp\left\{\boldsymbol{\eta}^{T}\left(\boldsymbol{\theta}\right) u(\boldsymbol{x})\right\} \; or$$

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^{T}\left(\boldsymbol{\theta}\right) u(\boldsymbol{x}) - A\left(\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)\right)\right\},$$

$$where : A\left(\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)\right) = -\log g\left(\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)\right)$$

- *$\boldsymbol{\eta}(\boldsymbol{\theta})$ are the canonical or natural parameters,*

- *$\boldsymbol{\theta}$ is the parameter vector of some distribution that can be written in the exponential family format*

Exponential Family: The Bernoulli Distribution

- Consider the Bernoulli distribution:

$$p(x\mid\mu)=\mathcal{B}ern(x\mid\mu)=\mu^{x}(1-\mu)^{1-x}=\exp\left\{x\ln\mu+(1-x)\ln(1-\mu)\right\}=$$

$$=\underbrace{(1-\mu)}_{g(\eta)}\exp\left\{\underbrace{\ln\left(\frac{\mu}{1-\mu}\right)}_{\eta}x\right\}$$

$$p(\boldsymbol{x}\mid\boldsymbol{\eta})=h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{T}u(\boldsymbol{x})\right\}$$
$$=h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^{T}u(\boldsymbol{x})-A(\boldsymbol{\eta})\right\}$$

- From this we see that (note that *the relation $\mu(\eta)$ is invertible*)

$$\eta=\ln\left(\frac{\mu}{1-\mu}\right)\Rightarrow \qquad \mu=\sigma(\eta)=\frac{1}{1+e^{-\eta}}$$

Logistic sigmoid function

and

$$g(\eta)=1-\mu=1-\sigma(\eta)=\sigma(-\eta)$$

- Finally:

$$p(x\mid\eta)=g(\eta)\exp\left\{\eta x\right\},\ u(x)=x,\ h(x)=1,\ g(\eta)=\sigma(-\eta),$$
$$A(\eta)=-\ln g(\boldsymbol{\eta})=-\log(1-\mu)=\log(1+e^{\eta})$$

Exponential Family: The Beta Distribution

- Consider the Beta distribution

$$\mathcal{B}eta(\mu \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \exp\left[(a-1)\ln\mu + (b-1)\ln(1-\mu)\right]$$

- Comparing this with our exponential family:

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{\eta}) &= h(\boldsymbol{x}) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} \\ &= h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\} \end{aligned}$$

we can easily indentify:

$$u(\mu) = (\ln\mu, \ln(1-\mu))^T, \boldsymbol{\eta} = (a-1, b-1)^T, h(\mu) = 1, \ g(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)},$$

$$A(a,b) = -\ln g(\boldsymbol{\eta}) = \ln\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Exponential Family: The Gaussian

- Consider the univariate Gaussian

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

- Comparing this with our exponential family:

$$p(\boldsymbol{x}|\boldsymbol{\eta})=h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}=h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})-A(\boldsymbol{\eta})\right\}$$

we can indentify (this is a two parameter distribution):

$$u(x)=(x,x^2)^T,\ \boldsymbol{\eta}=\left(\frac{\mu}{\sigma^2},-\frac{1}{2\sigma^2}\right)^T,\ h(x)=\frac{1}{\sqrt{2\pi}},\ g(\boldsymbol{\eta})=\left(-2\eta_2\right)^{1/2}\exp\frac{\eta_1^2}{4\eta_2}$$

$$A(\eta)=-\ln g(\boldsymbol{\eta})=-\frac{1}{2}\ln\left(-2\eta_2\right)-\frac{\eta_1^2}{4\eta_2}$$

Conjugate Priors

- In general, for a given probability distribution $p(\boldsymbol{x}|\boldsymbol{\eta})$, we can seek a prior $p(\boldsymbol{\eta})$ that is conjugate to the likelihood function, so that the posterior distribution has the same functional form as the prior.

  - For the Bernoulli, the conjugate prior is the Beta distribution

  - For the Gaussian, the conjugate prior for the mean is a Gaussian, and the conjugate prior for the precision is the Wishart distribution

Conjugate Priors

- For any member of the exponential family with likelihood

$$p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = h(\boldsymbol{x})\,g(\boldsymbol{\eta}(\boldsymbol{\theta}))\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\,u(\boldsymbol{x})\right\}$$

there exists a conjugate prior that can be written in the form

$$p(\boldsymbol{\theta}|\nu_0,\boldsymbol{\tau}_0) \propto g\big(\boldsymbol{\eta}(\boldsymbol{\theta})\big)^{\nu_0}\exp\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{\tau}_0\} = \exp\{\nu_0\boldsymbol{\eta}^T(\boldsymbol{\theta})\bar{\tau}_0 - A\big(\boldsymbol{\eta}(\boldsymbol{\theta})\big)\nu_0\},\ where\colon \boldsymbol{\tau}_0 \equiv \nu_0\bar{\boldsymbol{\tau}}_0$$

- In normalized form, we write:

$$p(\boldsymbol{\theta}|\nu_0,\boldsymbol{\tau}_0) = \frac{1}{Z(\nu_0,\boldsymbol{\tau}_0)}g\big(\boldsymbol{\eta}(\boldsymbol{\theta})\big)^{\nu_0}\exp\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{\tau}_0\} = \frac{1}{Z(\nu_0,\boldsymbol{\tau}_0)}\exp\{\nu_0\boldsymbol{\eta}^T(\boldsymbol{\theta})\bar{\boldsymbol{\tau}}_0 - A\big(\boldsymbol{\eta}(\boldsymbol{\theta})\big)\nu_0\}$$

$$where\colon Z(\nu_0,\boldsymbol{\tau}_0) = \int \exp\{\nu_0\boldsymbol{\eta}^T(\boldsymbol{\theta})\bar{\boldsymbol{\tau}}_0 - A\big(\boldsymbol{\eta}(\boldsymbol{\theta})\big)\nu_0\}d\boldsymbol{\theta}$$

Conjugate Priors

$$p(\boldsymbol{X}\mid\boldsymbol{\theta}) = \prod_{n=1}^{N}\left(h(\boldsymbol{x}_n)g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})u(\boldsymbol{x}_n)\right\}\right) = \prod_{n=1}^{N}\left(h(\boldsymbol{x}_n)\right)g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^N \exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\sum_{n=1}^{N}u(\boldsymbol{x}_n)\right\}$$

$$p(\boldsymbol{\theta}|\nu_0,\boldsymbol{\tau_0}) = \frac{1}{Z(\nu_0,\boldsymbol{\tau}_0)}g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_0}\exp\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{\tau}_0\} = \frac{1}{Z(\nu_0,\boldsymbol{\tau}_0)}\exp\left\{\nu_0\boldsymbol{\eta}^T(\boldsymbol{\theta})\overline{\boldsymbol{\tau}}_0 - A\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)\nu_0\right\}$$

- Using $\overline{\boldsymbol{u}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u}(\boldsymbol{x}_n)$, the posterior becomes (this form justifies $\overline{\boldsymbol{\tau}}_0$):

$$p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{\chi},\nu) \propto g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_0+N}\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\left(\sum_{n=1}^{N}\boldsymbol{u}(\boldsymbol{x}_n)+\nu_0\overline{\boldsymbol{\tau}}_0\right)\right\} = g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_0+N}\exp\{\boldsymbol{\eta}^{\boldsymbol{T}}(\boldsymbol{\theta})(N\overline{\boldsymbol{u}}+\nu_0\overline{\boldsymbol{\tau}}_0)\}$$

- The parameter ν_0 can be interpreted as *effective number of fictitious observations* in the prior each of which has a value for the sufficient statistic equal to $\overline{\boldsymbol{\tau}}_0$.

$$p(\boldsymbol{\theta}|\boldsymbol{X},\nu_N,\boldsymbol{\tau_N}) = \frac{1}{Z(\nu_N,\boldsymbol{\tau}_N)}g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_N}\exp\left\{(N+\nu_0)\boldsymbol{\eta}^T(\boldsymbol{\theta})\frac{N\overline{\boldsymbol{u}}+\nu_0\overline{\boldsymbol{\tau}}_0}{N+\nu_0}\right\} = \frac{1}{Z(\nu_N,\tau_N)}g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_N}\exp\{\nu_N\boldsymbol{\eta}^T(\theta)\overline{\boldsymbol{\tau}}_N\},$$

$$where\ \nu_N = \nu_0 + N,\ \overline{\boldsymbol{\tau}}_N = \frac{N\overline{\boldsymbol{u}}+\nu_0\overline{\boldsymbol{\tau}}_0}{N+\nu_0},\boldsymbol{\tau}_N = \nu_N\overline{\boldsymbol{\tau}}_N = N\overline{\boldsymbol{u}}+\nu_0\overline{\boldsymbol{\tau}}_0 = \sum_{i=1}^{N}\boldsymbol{u}(x_i)+\boldsymbol{\tau}_0$$

Posterior Predictive

- Let $\boldsymbol{u}(\boldsymbol{X})=\sum_{i=1}^{N}\boldsymbol{u}(\boldsymbol{x}_i),\ \boldsymbol{u}(\boldsymbol{X}^{'})=\sum_{i=1}^{N'}\boldsymbol{u}(\boldsymbol{x}_i^{'}),$ the posterior predictive is then:

$$p(\boldsymbol{X}^{'}\mid\boldsymbol{X})=\int p(\boldsymbol{X}^{'}\mid\boldsymbol{\theta})p(\boldsymbol{\theta}\mid\boldsymbol{X})d\boldsymbol{\theta}$$

$$=\prod_{i=1}^{N'}h(\boldsymbol{x}_i^{'})\int g(\boldsymbol{\eta})^{N'}\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{X}^{'})\right\}\frac{1}{Z(\nu_0+N,\boldsymbol{u}(\boldsymbol{X})+\boldsymbol{\tau}_0)}g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_N}\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\left(\boldsymbol{u}(\boldsymbol{X})+\boldsymbol{\tau}_0\right)\right\}d\boldsymbol{\theta}$$

- This is simplified as follows:

$$p(\boldsymbol{X}^{'}\mid\boldsymbol{X})=\prod_{i=1}^{N'}h(\boldsymbol{x}_i^{'})\frac{1}{Z\left(\nu_0+N,\boldsymbol{u}(\boldsymbol{X})+\boldsymbol{\tau}_0\right)}\int g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{N'+\nu_N}\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\left(\boldsymbol{u}(\boldsymbol{X}^{'})+\boldsymbol{u}(\boldsymbol{X})+\boldsymbol{\tau}_0\right)\right\}d\boldsymbol{\theta}$$

$$=\prod_{i=1}^{N'}h(\boldsymbol{x}_i^{'})\frac{Z\left(\nu_0+N+N',\boldsymbol{u}(\boldsymbol{X}^{'})+\boldsymbol{u}(\boldsymbol{X})+\boldsymbol{\tau}_0\right)}{Z\left(\nu_0+N,\boldsymbol{u}(\boldsymbol{X})+\boldsymbol{\tau}_0\right)}$$

- If $N=0$, this becomes the marginal likelihood of X', which reduces to the normalizer of the posterior divided by the normalizer of the prior multiplied by a constant.

Beta/Bernoulli: Posterior Predictive

- Consider a Bernoulli likelihood with a Beta prior. The likelihood takes the familiar exponential distribution form:

$$p(\mathcal{D}|\theta) = \theta^{\sum_i x_i}(1-\theta)^{N-\sum_i x_i} = (1-\theta)^N \exp\left(\log\frac{\theta}{1-\theta}\underbrace{\sum_i x_i}_{s=\sum_i \mathbb{I}(x_i=1)}\right)$$

- The conjugate prior is a Beta: $p(\theta|\nu_0,\tau_0) = \theta^{\tau_0}(1-\theta)^{\nu_0-\tau_0} \propto (1-\theta)^{\nu_0}\exp\left(\log\left(\frac{\theta}{1-\theta}\right)\tau_0\right)$

$$p(\theta|\nu_0,\tau_0) = \mathcal{Beta}(\alpha,\beta), \alpha = \tau_0+1, \beta = \nu_0-\tau_0+1,$$

- Thus the posterior becomes: $p(\theta|\mathcal{D}) \propto \theta^{\tau_0+s}(1-\theta)^{\nu_0-\tau_0+N-s} \Rightarrow$

$$p(\theta|\mathcal{D}) = \mathcal{Beta}(\alpha_N,\beta_N), \alpha_N = \alpha+s, \beta_N = \beta+(N-s), s = \sum_i \mathbb{I}(x_i=1)$$

- Let s the number of heads in the past data. The probability of $s' = \sum_{i=1}^m \mathbb{I}(x_i'=1)$ future heads in m trials is then:

$$p(s'|\mathcal{D},m) = \int \theta^{s'}(1-\theta)^{m-s'}\frac{\Gamma(\alpha_N+\beta_N)}{\Gamma(\alpha_N)\Gamma(\beta_N)}\theta^{\alpha_N-1}(1-\theta)^{\beta_N-1}d\theta = \frac{\Gamma(\alpha_N+\beta_N)}{\Gamma(\alpha_N)\Gamma(\beta_N)}\frac{\Gamma(\alpha_{N+m})\Gamma(\beta_{N+m})}{\Gamma(\alpha_{N+m}+\beta_{N+m})}$$

$$\alpha_{N+m} = \alpha_N+s', \beta_{N+m} = \beta_N+(m-s')$$

Computing Moments of Sufficient Statistics u(x)

- Differentiate wrt η the $\int p(\boldsymbol{x}|\boldsymbol{\eta})d\boldsymbol{x}=1$ for the exponential family:

$$\int p(\boldsymbol{x}|\boldsymbol{\eta})d\boldsymbol{x}=\int h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}d\boldsymbol{x}=1$$

$$\nabla g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right\}d\boldsymbol{x}+g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})d\boldsymbol{x}=0\Rightarrow$$

$$-\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}=g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})d\boldsymbol{x}=\int p(\boldsymbol{x}|\boldsymbol{\eta})\boldsymbol{u}(\boldsymbol{x})d\boldsymbol{x}=\mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right]$$

- The above equation can be further simplified if written in terms of the partition function $Z=1/g(\eta)$ or $A=logZ=-logg(\eta)$:

$$\nabla A(\boldsymbol{\eta})=\mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right]$$

- Let us re-write explicitly the above equation as:

$$\nabla A(\boldsymbol{\eta})=g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})d\boldsymbol{x}$$

- We can compute the variance of $\boldsymbol{u}(\boldsymbol{x})$ by differentiating the Eq. above with respect to $\boldsymbol{\eta}$.

Computing Moments of Sufficient Statistics u(x)

$$\nabla A(\boldsymbol{\eta}) = g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})d\boldsymbol{x}$$

$$\nabla^2 A(\boldsymbol{\eta}) = \underbrace{\nabla g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})d\boldsymbol{x}}_{-\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})]\mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})^T\right]} + \underbrace{g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x})^T d\boldsymbol{x}}_{\mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x})^T\right]}$$

- Thus the covariance of $u(\boldsymbol{x})$ can be expressed in terms of the 2nd derivatives of $A(\boldsymbol{\eta})$ and similarly for higher order moments.

$$\nabla^2 A(\boldsymbol{\eta}) = \operatorname{cov}\left[\boldsymbol{u}(\boldsymbol{x})\right] \textit{ so } A(\boldsymbol{\eta}) \textit{ is convex}$$

- Provided we can normalize a distribution from the exponential family, we can always find its moments by simple differentiation.

Computing Moments of Sufficient Statistics u(x)

$$\nabla A(\boldsymbol{\eta}) = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right]$$

$$\nabla^2 A(\boldsymbol{\eta}) = \operatorname{cov}\left[\boldsymbol{u}(\boldsymbol{x})\right] \; \textit{so } A(\boldsymbol{\eta}) \textit{ is convex}$$

- Let us check these relations for the Univariate Gaussian:

$$g(\eta) = (-2\eta_2)^{1/2} \exp \frac{\eta_1^2}{4\eta_2}$$

$$A(\boldsymbol{\eta}) = -\frac{1}{2}\ln\left(-2\eta_2\right) - \frac{\eta_1^2}{4\eta_2}, \boldsymbol{\eta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T, u(x) = (x, x^2)^T$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta_1} = -\frac{\eta_1}{2\eta_2} = \mu = \mathbb{E}\left[X\right], \frac{\partial A(\boldsymbol{\eta})}{\partial \eta_2} = -\frac{1}{2\eta_2} + \frac{\eta_1^2}{4\eta_2^2} = \mu^2 + \sigma^2 = \mathbb{E}\left[X^2\right]$$

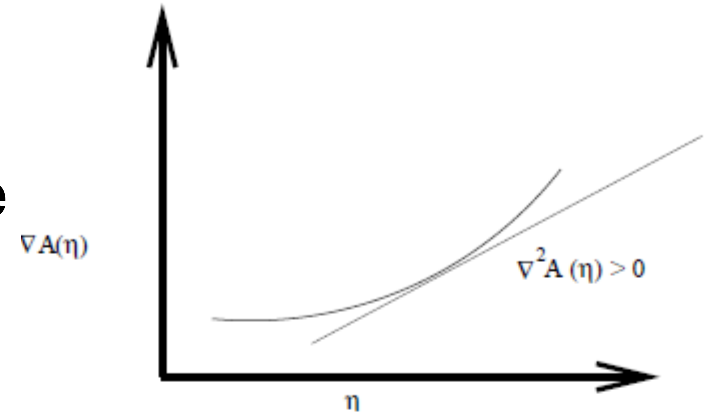
$$\frac{\partial^2 A(\boldsymbol{\eta})}{\partial \eta_1^2} = -\frac{1}{2\eta_2} = \sigma^2 = \operatorname{var}\left[X\right], etc.$$

Moment Parametrization

- We have shown that we can compute the mean of the distribution $\mu = E[u(\boldsymbol{x})]$ in terms of the canonical parameter η:

$$\mu = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right] = \nabla A(\boldsymbol{\eta})$$

- We have also shown that $A(\eta)$ is a convex function. Since for a convex function there is one-to-one relation between the argument of the function and its derivative, the mapping $\mu(\eta)$ is invertable.



- Thus the exponential family of distributions can also be parameterized in terms of μ (*moment parametrization*) exactly as we started this course.

MLE for the Exponential Family

- The joint density for a data $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ *is itself an exp. distribution with sufficient statistics* $\sum_{n=1}^{N} u(\boldsymbol{x}_n)$

$$p(\boldsymbol{X}|\boldsymbol{\eta}) = \prod_{n=1}^{N}\left(h(\boldsymbol{x}_n)g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}_n)\right\}\right) = \prod_{n=1}^{N}\left(h(\boldsymbol{x}_n)\right)g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^T \sum_{n=1}^{N} u(\boldsymbol{x}_n)\right\} \Rightarrow$$

$$\ln p(\boldsymbol{X}|\boldsymbol{\eta}) = \sum_{n=1}^{N} h(\boldsymbol{x}_n) + N\ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^T\sum_{n=1}^{N} u(\boldsymbol{x}_n) = \sum_{n=1}^{N} h(\boldsymbol{x}_n) - NA(\boldsymbol{\eta}) + \boldsymbol{\eta}^T\sum_{n=1}^{N} u(\boldsymbol{x}_n)$$

- *The exponential family is the only family of distributions* ***with finite sufficient statistics*** *(size independent of the data set size).*

- *The log likelihood is concave (*A *convex) and has a unique maximum.*

- Maximizing wrt $\boldsymbol{\eta}$ gives: $\nabla A(\boldsymbol{\eta}_{ML}) = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u}(\boldsymbol{x}_n) \Rightarrow \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right] = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u}(\boldsymbol{x}_n)$

- *At the MLE, the empirical average of the sufficient statistic is equal the model's theoretical expected sufficient statistics (moment matching).*

- *Thus to find the expected value of the sufficient statistics, one can use directly the data without having to estimate* $\boldsymbol{\eta}$. When $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{x}$, the above allows us to compute the expectation of $\boldsymbol{x}$ directly from the data.

MLE for the Exponential Family

$$\nabla A(\boldsymbol{\eta}_{ML}) = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right] = \frac{1}{N}\sum_{n=1}^{N} u(\boldsymbol{x}_n)$$

- Using the sufficient statistic, one can in principle invert the above equ. to compute $\boldsymbol{\eta}_{MLE}$. For example, for the Bernoulli distribution,

$$p(x \mid \eta) = g(\eta)\exp\left\{\eta x\right\}, u(x) = x, h(x) = 1,$$

$$\mu = \frac{1}{1+e^{-\eta}}, \ g(\eta) = \frac{1}{1+e^{\eta}}, \eta = \ln\left(\frac{\mu}{1-\mu}\right)$$

and thus:

$$\mathbb{E}[X] = p(X=1) = \bar{\mu} \equiv \mu_{MLE} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(x_n = 1)$$

and

$$\boxed{\eta_{MLE} = \ln\left(\frac{\bar{\mu}}{1-\bar{\mu}}\right)}$$

MLE and Kullback-Leibler Distance

- A useful property for the MLE (and not just a property for the exponential family of distributions) is the following:
- Minimizing the KL distance to the empirical distribution is equivalent to maximizing the likelihood.
- Indeed, let us consider the model $log p(x|\theta)$ and the empirical distribution:

$$p_{emp}(x) = \frac{1}{N}\sum_{n=1}^{N}\delta\left(x, x_n\right)$$

- We can then derive the following:

$$\sum_x p_{emp}(x)\log p(x|\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_x \delta\left(x, x_n\right)\log p(x|\theta) = \frac{1}{N}\sum_{n=1}^{N}\log p(x_n|\theta) = \frac{1}{N}\ell(\theta|\mathscr{D})$$

and from this:

$$KL\left(p_{emp}(x), p(x|\theta)\right) = \sum_x p_{emp}(x)\log\frac{p_{emp}(x)}{p(x|\theta)} = \sum_x p_{emp}(x)\log p_{emp}(x) - \sum_x p_{emp}(x)\log p(x|\theta)$$

$$= \sum_x p_{emp}(x)\log p_{emp}(x) - \frac{1}{N}\ell(\theta|\mathscr{D})$$

- Since the 1st term is independent of θ, the assertion is proved.

Maximum Entropy and Exponential Family

- If nothing is known about a distribution except that it belongs to a certain class, then the distribution with the largest entropy should be chosen as the default.[a]
- The entropy is defined as
  - discrete case $$\mathbb{H}(\pi) = -\sum_k \pi(\theta_k)\log\left(\pi(\theta_k)\right)$$
- When some statistics (moments) of the distribution are known,

$$\mathbb{E}_\pi\left[g_k\left(\theta\right)\right] = w_k,\, k = 1,..,K$$

the maximum entropy distribution is of the form (λ's are the Lagrange multipliers enforcing the constraints):

$$\pi(\theta_i) = \frac{\exp\left(-\sum_{k=1}^{K}\lambda_k g_k(\theta_i)\right)}{\sum_j \exp\left(-\sum_{k=1}^{K}\lambda_k g_k(\theta_j)\right)},\ \lambda_k = Lagrange\ multipliers$$

- *Thus the MaxEnt distribution has the form of the exponential family.*

[a] C. P. Robert, *The Bayesian Choice*, *Springer, 2nd edition,* chapter 3 (full text available)

Generalized Linear Models

- We now study regression given data X and Y and a GLIM.
- We choose a particular conditional expectation of Y. We denote the modeled value of conditional expectation as $m = f(\boldsymbol{\theta}^T\boldsymbol{x}), \xi = \boldsymbol{\theta}^T\boldsymbol{x}$.
- For linear regression, *GLIM extends these ideas beyond the Gaussian, Bernoulli and multinomial setting to the more general exponential family.*
- X enters linearly as $\boldsymbol{\theta}^T\boldsymbol{x}$ and f *is called a response function.* Ψ is a one-to-one map of μ to η.

$$\begin{tikzcd}
\theta \arrow[dr] & & & \\
& \xi \arrow[r, "f"] & \mu \arrow[r, "\Psi"] & \eta \\
x \arrow[ur] & & &
\end{tikzcd}$$

- To specify a GLIM we need (a) a choice of exponential family distribution, and (b) a choice of the response function $f(\cdot)$.
- Choosing the exponential family distribution is strongly constrained by the nature of the data.
- Note that $f(\cdot)$ needs to be both monotonic and differentiable. However, *there is a particular response function (canonical response function) that is uniquely associated with a given exponential family distribution.*

Canonical Response Function

- Canonical response function:

$$f(\cdot) = \Psi^{-1}(\cdot)$$
$$\xi = \eta$$

$$\begin{array}{l}\theta \searrow \\ \quad\quad \xi \xrightarrow{f} \mu \xrightarrow{\Psi} \eta \\ x \nearrow\end{array}$$

- *If we decide to use the canonical response function, the choice of the exponential family density completely determines the GLIM.*

$$\xi = f^{-1}(\mu) = \Psi(\mu) = \eta$$

MLE & Canonical Response Function

- Consider a regression problem with data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i\}, i = 1, \ldots N\}$. *The log likelihood for a GLIM is as:*

$$\ell(\theta, \mathcal{D}) = \sum_{n=1}^{N} \log h(y_n) + \sum_{n=1}^{N} \left(\underbrace{\eta_n}_{\psi(\mu_n)} y_n - A(\eta_n) \right), where: \ \mu_n = f(\xi_n) \ with \ \xi_n = \boldsymbol{\theta}^T \boldsymbol{x}_n$$

- *For a canonical response,* $\eta = \xi = \boldsymbol{\theta}^T \boldsymbol{x}$ *, and this is simplified as:*

$$\ell(\theta, \mathcal{D}) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{\theta}^T \underbrace{\sum_{n=1}^{N} \boldsymbol{x}_n y_n}_{Sufficient\ statistic\ for\ \boldsymbol{\theta}} - \sum_{n=1}^{N} A(\eta_n)$$

expectation

- Regardless of N, the size of the sufficient statistic is fixed: the dimension of $\boldsymbol{x}_n$ - important reason for using canonical response.

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \mathcal{D}) = \sum_{n=1}^{N} (y_n - A'(\eta_n)) \nabla_{\boldsymbol{\theta}} \eta_n = \sum_{n=1}^{N} (y_n - \mu_n) \underbrace{\nabla_{\boldsymbol{\theta}} \eta_n}_{\boldsymbol{x}_n} = \sum_{n=1}^{N} (y_n - \mu_n) \boldsymbol{x}_n \ or \ \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \mathcal{D}) = \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{\mu})$$

- *This is a general expression for GLM with exponential family distributions and the canonical response function*.

Iterative Reweighted Least Squares (IRLS)

- The Hessian can now be computed from

$$\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta},\mathcal{D})=\sum_{n=1}^{N}(y_n-\mu_n)\boldsymbol{x}_n \; or \; \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta},\mathcal{D})=\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{\mu})$$

as:

$$\boldsymbol{H}=\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta},\mathcal{D})=-\sum_{n=1}^{N}\frac{d\mu_n}{d\eta_n}\boldsymbol{x}_n\boldsymbol{x}_n^T \; or \; \nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta},\mathcal{D})=-\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\,, where: \boldsymbol{W}=\left\{\frac{d\mu_1}{d\eta_1},\dots,\frac{d\mu_n}{d\eta_n}\right\}$$

- To estimate parameters in the canonical response function choice, one can use the iteratively reweighted least squares (IRLS) algorithm

- *The batch Newton algorithm now takes the familiar IRLS form:*

$$\boldsymbol{\theta}^{t+1}=\boldsymbol{\theta}^t+\left(\boldsymbol{X}^T\boldsymbol{W}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\left(\boldsymbol{y}-\boldsymbol{\mu}^t\right)=\left(\boldsymbol{X}^T\boldsymbol{W}^t\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{W}^t\boldsymbol{X}\boldsymbol{\theta}^t+\boldsymbol{X}^T\left(\boldsymbol{y}-\boldsymbol{\mu}^t\right)\right)$$

$$=\left(\boldsymbol{X}^T\boldsymbol{W}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^t\left(\underbrace{\boldsymbol{X}\boldsymbol{\theta}^t}_{\boldsymbol{\eta}^t}+\boldsymbol{W}^{t^{-1}}\left(\boldsymbol{y}-\boldsymbol{\mu}^t\right)\right)=\left(\boldsymbol{X}^T\boldsymbol{W}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^t\left(\boldsymbol{\eta}^t+\boldsymbol{W}^{t^{-1}}\left(\boldsymbol{y}-\boldsymbol{\mu}^t\right)\right)$$

- For non-canonical response functions, the Hessian has an extra term that contains the factor $(\boldsymbol{y}-\boldsymbol{\mu})$. When we take expectations this term vanishes! So using the expected Hessian in the Newton method the algorithm looks essentially the same (Fisher Scoring algorithm).

Sequential Estimation - LMS

- An on-line estimation algorithm can be obtained by following the stochastic gradient of the log likelihood function.

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} + \rho\left(y_n - \mu_n^t\right)x_n, \mu_n^t = f\left(\boldsymbol{\theta}^{t^T} x_n\right)$$

- If we do not use the canonical response function, then the gradient also includes the derivatives of $f(\cdot)$ and $\Psi(\cdot)$. These can be viewed as scaling coefficients that alter the step size, but otherwise leave the general LMS form intact.

- *The LMS algorithm is the generic stochastic gradient algorithm for models throughout the GLIM family.*