Unifying Theory of GLMs (Part II)

Rebecca C. Steorts

Reading

These slides provide some background that will be helpful in tying together generalized linear models completely.

https://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf

Agenda

- Introduce alternative parameterization of exponential family that is widely used.
- Applies beyond the one-parameter setting for generalized linear models
- Using tricks and the formulation of the exponential family for GLMs, we can calculate certain summary statistics, such as the mean and variance without relying on integration

Exponential Families

Consider the following:

$$p(y \mid \eta) = h(y) \exp\{\eta^T t(y) - A(\eta)\},$$

where

- η is the natural parameter
- $h(y)$ is the underlying measure
- $A(\eta)$ is the log normalizer (ensuring the distribution integrates to one) or **log partition function**.

Note that $Z(\eta) = \exp\{A(\eta)\}$ or $A(\eta) = \log Z(\eta)$.

Exponential Families

We can alternatively write the exponential family as follows:

$$p(y \mid \eta) = \frac{1}{Z(\eta)} h(y) \exp\{\eta^T u(y)\},$$

where

- η are the natural parameters
- $u(y)$ are the sufficient statistics
- $Z(\eta)$ is a partition function that ensures the density is normalized.

$$Z(\eta) = \int h(y) \exp\{\eta^T u(y)\} \, dy.$$

Bernouli

$$p(y \mid \eta) = \frac{1}{Z(\eta)} h(y) \exp\{\eta^T u(y)\}$$

Consider

$$\begin{aligned} \text{Bern}(y \mid \mu) &= \mu^y (1-\mu)^{1-y} && (1) \\ &= (1-\mu)(\frac{\mu}{1-\mu})^y && (2) \\ &= (1-\mu) \times 1 \times \exp\{\log(\frac{\mu}{1-\mu}) \times y\} && (3) \end{aligned}$$

- $Z(\eta) = \frac{1}{1-\mu}$
- $h(y) = 1$
- $\eta = \log(\frac{\mu}{1-\mu})$ (canonical link or natural parameter)
- $u(y) = y$

Bernoulli (continued)

Solving for μ, we find that

$$\mu = \frac{e^{\eta}}{1 + e^{\eta}}.$$

Then

$$1 - \mu = \frac{1}{1 + e^{\eta}} \implies Z(\eta) = 1 + e^{\eta}.$$

Later in the lecture, we will work with the partition function to find the mean and variance of the density.

Bernoulli continued

1. $$\eta = \beta_0 + \sum_{i=1}^{p} \beta_j x_j$$

2. $$\eta = g(\mu) \implies g^{-1}(\eta) = \mu.$$

3. Putting this together,

$$\eta = g(\mu) = \beta_0 + \sum_{i=1}^{p} \beta_j x_j.$$

For the Bernoulli, we thus have the following:

$$\eta = \log(\frac{\mu}{1-\mu}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}.$$

Summarize GLM

We can summarize GLMs in two ways:

1. We must specify the response variable as belonging to an exponential family. (Random component).
2. The systematic component, which specifies a linear predictor $\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}$ and a link function $g(\mu) = \eta$ that is known, monotonic and differentiable.

Specification

To specify this for $y_i \sim \text{Bernoulli}(\mu_i)$, we proceed as follows:

1. $y_i \mid \mu_i \sim \text{Bernoulli}(\mu_i)$ (random component)
2. $\log(\frac{\mu_i}{1-\mu_i}) = \beta_0 + \beta_1 x_i$ (systematic component)

where x_i refers to one explanatory variable.

We have already verified that the Bernoulli distribution is a member of the exponential family.

Why are we studying the exponential family?

Many convenient properties such as:

- Sufficient statistics for maximum likelihood
- Many convenient identities for the partition function that allow for quick computation directly using the partition function (or functions of it) instead of having to work with the likelihood function.
- For instance, we can quickly compute the moments of an exponential family using the partition function.

Maximum Likelihood and sufficient statistics

Suppose we have observations that are i.i.d $y_1, \ldots, y_n$.

Find η that maximizes $p(Y \mid \eta)$.

$$p(Y \mid \eta) = \prod_{i=1}^{n} \frac{1}{Z(\eta)} \left(h(y_i) \exp\{\eta^T u(y_i)\} \right) \tag{4}$$

$$= (\frac{1}{Z(\eta)})^n \prod_{i=1}^{n} h(y_i) \exp\{\eta^T \sum_{i=1}^{n} u(y_i)\} \tag{5}$$

Maximum Likelihood and sufficient statistics

It follows that

$$\log p(Y \mid \eta) = -n \log(Z(\eta)) + \eta^T \sum_{i=1}^{n} u(y_i) + \log(\prod_{i=1}^{n} h(y_i))$$

$$\nabla_\eta \log p(Y \mid \eta) = -n \nabla_\eta \log Z(\eta) + \sum_{i=1}^{n} u(y_i) := 0$$

This implies that

$$\nabla_\eta \log Z(\eta) = \frac{1}{n} \sum_{i=1}^{n} u(y_i).$$

The maximum likelihood solution only depends on $\sum_{i=1}^{n} u(y_i)$.

Thus, $u(y)$ is called the **sufficient statistic.**

Maximum Likelihood and sufficient statistics

We showed that the maximum likelihood estimator (that belongs to an exponential family) only depends on the sufficient statistic.

Sufficiency

The notion of a sufficient statistic is a fundamental one in statistical theory and its applications.

Fisher (1922) introduced this into the literature in the attempt to formalize the notion of no loss of information.

A sufficient statistic should solely contain all of the information about the unknown parameters of the underlying distribution that the entire sample could have provided.

In that sense, there is nothing to lose by restricting attention to just a sufficient statistic.

Bernoulli (continued)

Bernoulli:

- Recall $u(y) = y$.
- Only need to store $\sum_i y_i$.

Distributions not in the Exponential family

Can you think of some distributions that would not be in the exponential family and why?

Identities

1.

$$\nabla_\eta \log Z(\eta) = E_{y\sim p(y|\eta)}[u(y)] =: \xi \quad \text{(moments)}$$

2.

$$\nabla_\eta \log p(Y \mid \eta) = u(y) - E_{y\sim p(y|\eta)}[u(y)].$$

3.

$$\nabla^2_\eta \log Z(\eta) = Cov(u(x)) = -\nabla^2_\eta \log p(y \mid \eta).$$

These show that to calculate the mean and the variance for a distribution in the exponential family, we can work directly with the partition function.

Bernoulli (continued)

Find the mean and variance using the derivation of the Bernoulli being in the exponential family.

Specifically, show that the mean is μ and the variance is $\mu(1-\mu)$ using the identities above.

Bernoulii (continued)

Recall that $Z(\eta) = 1 + e^{\eta}$.

Consider

$$\nabla_{\eta} \log Z(\eta) = \nabla_{\eta}(1 + e^{\eta}) = \frac{e^{\eta}}{1 + e^{\eta}} = \mu.$$

Consider

$$\nabla_{\eta}^2 \log Z(\eta) = \nabla_{\eta}(\frac{e^{\eta}}{1 + e^{\eta}}) = \frac{e^{\eta^2}}{(1 + e^{\eta})^2} = \mu(1 - \mu).$$

Lemma 1

$$Z(\eta) = \int h(y) \exp\{\eta^T u(y)\}\, dy.$$

$$\nabla_\eta \log Z(\eta) = E_{y \sim p(y|\eta)}[u(y)] =: \xi \quad \text{(moments)}$$

Proof of Lemma 1

$$Z(\eta) = \int h(y) \exp\{\eta^T u(y)\}\, dy.$$

$$\nabla_\eta \log Z(\eta) = E_{y \sim p(y|\eta)}[u(y)] =: \xi$$

By the chain rule,

$$\nabla_\eta \log Z(\eta) = \frac{\nabla_\eta Z(\eta)}{Z(\eta)}.$$

Proof of Lemma 1 (continued)

It then follows

$$\nabla_\eta \log Z(\eta) = \frac{1}{Z(\eta)} \nabla_\eta \int h(y) \exp\{\eta^T u(y)\}\, dy \tag{6}$$

$$= \frac{1}{Z(\eta)} \int h(y) \exp\{\eta^T u(y)\} u(y)\, dy \tag{7}$$

$$= E_{y \sim p(y|\eta)}[u(y)] =: \xi \tag{8}$$

1. Observe there is a one-to-one mapping between η and ξ.
2. ξ provides an alternate parameterization for the exponential family.

Lemma 2

$$p(y \mid \eta) = \frac{1}{Z(\eta)} h(y) \exp\{\eta^T u(y)\}$$

$$\nabla_\eta \log p(Y \mid \eta) = u(y) - E_{y \sim p(y|\eta)}[u(y)].$$

Proof of Lemma 2

$$p(y \mid \eta) = \frac{1}{Z(\eta)} h(y) \exp\{\eta^T u(y)\}$$

$$\begin{aligned}
\nabla_\eta \log p(Y \mid \eta) &= \nabla_\eta \left[-logZ(\eta) + \log h(y) + \eta^T u(y) \right] && (9) \\
&= -\nabla_\eta \log Z(\eta) + \nabla_\eta [\eta^T u(y)] && (10) \\
&= -E_{y \sim p(y|\eta)}[u(y)] + u(y) && (11)
\end{aligned}$$

Connection with Maximum likelihood and Moments

Recall in maximum likelihood, we have

$$\nabla_\eta \sum_{i=1}^{n} \log p(y_i \mid \eta) = n[\underbrace{\frac{1}{n}\sum_{i=1}^{n} u(y_i)}_{\text{empirical moments } \hat{\xi}} - \underbrace{\nabla_\eta \log Z(\eta))}_{\text{moments } \xi}] \tag{12}$$

$$= n(\hat{\xi} - \xi). \tag{13}$$

This shows that maximum likelihood leads to moment matching.

Moment Matching

The moment matching method (MME) is a widely used method of estimation of parameters.

The idea is to find values of the unknown parameters that result in a match between the theoretical (or population) and sample moments evaluated from data.

Lemma 3

$$\nabla_\eta^2 \log Z(\eta) = Cov(u(x)) = -\nabla_\eta^2 \log p(y \mid \eta).$$

Proof of Lemma 3

$$p(y \mid \eta) = \frac{1}{Z(\eta)} h(y) \exp\{\eta^T u(y)\}$$

Show

$$\nabla_\eta^2 \log p(y \mid \eta) = -\nabla_\eta^2 \log Z(\eta)$$

Recall that

$$\begin{aligned}
\nabla_\eta \log p(Y \mid \eta) &= \nabla_\eta \left[-logZ(\eta) + \log h(y) + \eta^T u(y) \right] && (14) \\
&= -\nabla_\eta \log Z(\eta) + \nabla_\eta [\eta^T u(y)] && (15) \\
&= -\nabla_\eta \log Z(\eta) + u(y), \implies && (16)
\end{aligned}$$

$$\nabla_\eta^2 \log p(y \mid \eta) = -\nabla_\eta^2 \log Z(\eta).$$

Proof of Lemma 3

Show

$$\nabla_\eta^2 \log Z(\eta) = Cov(u(y)).$$

Other relationships

$$\nabla_\eta^2 \log Z(\eta) = F_\eta := E[-\nabla_\eta^2 \log p(y \mid \eta)]$$

(Fisher information)

$$F_\eta = \nabla_\eta \xi = J_{\xi,\eta}.$$

(Jacobian mapping $\eta \to \xi$).

Exercise: Normal (μ random)

$$p(y \mid \eta) = \frac{1}{Z(\eta)} h(y) \exp\{\eta^T u(y)\}$$

$$p(y \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{\frac{-1}{2\sigma^2}(y-\mu)^2\} \tag{17}$$

$$= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{\frac{-1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{\mu^2}{2\sigma^2}\} \tag{18}$$

$$= \underbrace{\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{\frac{-1}{2\sigma^2}y^2\}}_{h(y)} \underbrace{\exp\{-\frac{\mu^2}{2\sigma^2}\}}_{-Z(\eta)} \exp\{\underbrace{\frac{\mu}{\sigma^2}}_{\eta^T} \underbrace{y}_{u(y)}\} \tag{19}$$

Exercise (continued)

The natural parameter is $\eta = \frac{\mu}{\sigma^2} \implies \mu = \eta\sigma^2$ and the partition function is

$$Z(\eta) = \exp\{\frac{\mu^2}{2\sigma^2}\} = \exp\{\frac{\eta^2\sigma^4}{2\sigma^2}\}\exp\{\frac{\eta^2\sigma^2}{2}\}$$

This implies that

$$\log Z(\eta) = \frac{\eta^2\sigma^2}{2}.$$

Using our identities from earlier, it follows that

$$\nabla_\eta \log Z(\eta) = \frac{2\eta\sigma^2}{2} = \frac{2\mu\sigma^2}{2\sigma^2} = \mu$$

and

$$\nabla_\eta^2 \log Z(\eta) = \nabla_\eta \eta\sigma^2 = \sigma^2.$$

Exercise: Normal (μ and σ^2)

$$p(y \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{\frac{-1}{2\sigma^2}(y-\mu)^2\} \tag{20}$$

$$= \frac{1}{(2\pi)^{1/2}} \sigma^{-1} \exp\{\frac{-1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{\mu^2}{2\sigma^2}\} \tag{21}$$

$$= \underbrace{\frac{1}{(2\pi)^{1/2}}}_{h(y)} \underbrace{\sigma^{-1} \exp\{-\frac{\mu^2}{2\sigma^2}\}}_{-Z(\eta)} \exp\{\underbrace{\begin{bmatrix} \frac{\mu}{\sigma^2} \\ \frac{-1}{2\sigma^2} \end{bmatrix}^T}_{\eta^T} \underbrace{\begin{bmatrix} y \\ y^2 \end{bmatrix}}_{u(y)}\} \tag{22}$$

$$\tag{23}$$

Normal (continued)

The partition function can be written as:

$$Z(\mu, \sigma^2) = \exp\{\frac{\mu^2}{2\sigma^2} + \log(\sigma)\} \implies \log Z(\mu, \sigma^2) = \frac{\mu^2}{2\sigma^2} + \log(\sigma).$$

From the natural parameters $\eta_1 = \frac{\mu}{\sigma^2}$ and $\eta_2 = -\frac{1}{2\sigma^2}$, we solve for μ and σ^2 in terms of η_1 and η_2:

$$\mu = \eta_1 \sigma^2, \quad \sigma^2 = -\frac{1}{2\eta_2}$$

Normal (continued)

Substituting these into the log-partition function, we compute the terms:

$$\frac{\mu^2}{2\sigma^2} = \frac{(\eta_1\sigma^2)^2}{2\sigma^2} = \frac{\eta_1^2\sigma^2}{2}$$

$$\frac{\eta_1^2\sigma^2}{2} = \frac{\eta_1^2}{2} \cdot \left(-\frac{1}{2\eta_2}\right) = -\frac{\eta_1^2}{4\eta_2}$$

Thus, the term $\log(\sigma)$ simplifies

$$\log(\sigma) = \log\left(\sqrt{-\frac{1}{2\eta_2}}\right) = \frac{1}{2}\log\left(-\frac{1}{2\eta_2}\right)$$

$$\log(\sigma) = -\frac{1}{2}\log(2) - \frac{1}{2}\log(\eta_2)$$

Normal (continued)

Combining these terms, the log-partition function in terms of η_1 and η_2 is:

$$\log Z(\eta_1, \eta_2) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(2) - \frac{1}{2}\log(\eta_2)$$

Normal (continued)

$$\log Z(\eta_1, \eta_2) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(2) - \frac{1}{2}\log(\eta_2)$$

Consider

$$\nabla_{\eta_1} \log Z(\eta_1, \eta_2) = -\frac{\eta_1}{2\eta_2}$$

and

$$\nabla^2_{\eta_1} \log Z(\eta_1, \eta_2) = -\frac{1}{2\eta_2}$$

Normal (continued)

Recall $\eta_1 = \frac{\mu}{\sigma^2}$ and $\eta_2 = -\frac{1}{2\sigma^2}$.

It follows that

$$\nabla_{\eta_1} \log Z(\eta_1, \eta_2) = -\frac{\mu}{2\sigma^2}(-2\sigma^2) = \mu.$$

and

$$\nabla^2_{\eta_1} \log Z(\eta_1, \eta_2) = -\frac{1}{2\eta_2} = \sigma^2.$$

Think about why this derivation gives the mean and the variance of y.