Exponential Family Entropy and KL divergence

Philip Schulz and Wilker Aziz

last modified: November 6, 2017

Abstract

This note shows how to compute the entropy and KL divergence (and by extension many other information-theoretic quantities) of exponential family distributions. We use the univariate Gaussian distribution as a running example and encourage the reader to apply the learned techniques to distributions in exercises. We expect the reader to be familiar with exponential families and some of their properties (such as moment computation).

1 Exponential Family Form

Most exponential family distributions are standardly written in what is called their canonical form. In order to apply results proved for exponential families to them, it is often convenient to rewrite them in their natural parametrisation. The general way of doing this is to apply the identity function $\exp(\log(\cdot))$ to them and reorder terms. Before we do so for the Gaussian, let us introduce the notation we will use for exponential families throughout.

$$p(x|\overbrace{\theta}^{\substack{\text{canonical}\\\text{parameters}}}) = \overbrace{h(x)}^{\substack{\text{base}\\\text{measure}}} \exp\left(\underbrace{\eta(\theta)^\top}_{\substack{\text{natural}\\\text{parameters}}} \times \overbrace{t(x)}^{\substack{\text{sufficient}\\\text{statistics}}} - \underbrace{A(\eta(\theta))}_{\text{log-normalizer}}\right) \tag{1}$$

By default we assume that all terms above are vector-valued. Notice that the canonical parameters may be a function of any conditioning context and thus Equation (1) covers marginal as well as conditional distributions.

Gaussian Distribution We first write the Gaussian distribution with mean μ and variance σ^2 in its canonical form. In that case we have $\theta^\top = \begin{bmatrix}\mu & \sigma^2\end{bmatrix}$.

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

Now we apply the identity function to transform this density into its exponential family form.

$$
\begin{aligned}
p(x|\mu,\sigma^2) &= \exp\left(\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\right)\right) && \text{(3a)}\\
&= \exp\left(\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log\sigma - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) && \text{(3b)}\\
&= \frac{1}{\sqrt{2\pi}}\exp\left(-\log\sigma - \frac{x^2-2x\mu+\mu^2}{2\sigma^2}\right) && \text{(3c)}\\
&= \frac{1}{\sqrt{2\pi}}\exp\left(\frac{2x\mu-x^2}{2\sigma^2} - \log\sigma - \frac{\mu^2}{2\sigma^2}\right) && \text{(3d)}\\
&= \frac{1}{\sqrt{2\pi}}\exp\left(\begin{bmatrix}\frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2}\end{bmatrix}\times\begin{bmatrix}x\\ x^2\end{bmatrix} - \log\sigma - \frac{\mu^2}{2\sigma^2}\right) && \text{(3e)}
\end{aligned}
$$

We recognise this as the exponential family form of the Gaussian with

$$
\begin{aligned}
h(x) &= \frac{1}{\sqrt{2\pi}} \quad & t(x) &= \begin{bmatrix}x\\ x^2\end{bmatrix}\\
A(\eta(\theta)) &= \log\sigma + \frac{\mu^2}{2\sigma^2} \quad & \eta(\theta) &= \begin{bmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{bmatrix}\ .
\end{aligned} \tag{4}
$$

Exercise Derive the exponential family form of the exponential and binomial distribution.

2 Entropy of Exponential Families

First, recall the definition of entropy:

$$
\mathbb{H}(X) = -\mathbb{E}\left[\log(X)\right] \tag{5}
$$

A convenient property of exponential families is that the only terms depending on X are the base measure and the sufficient statistics. This means we can split up the expectation term.

$$
\begin{aligned}
\mathbb{H}(X) &= -\mathbb{E}\left[\log(h(X)) + \eta(\theta))^\top t(X) - A(\eta(\theta))\right] && \text{(6a)}\\
&= -\mathbb{E}\left[\log(h(X))\right] + \eta(\theta))^\top\mathbb{E}\left[t(X)\right] - A(\eta(\theta)) && \text{(6b)}
\end{aligned}
$$

For most distributions the expectation of the log base measure is a constant because $h(\cdot)$ does not actually depend on X (see the Gaussian example above). In order to compute the expected sufficient statistics we exploit another convenient property of exponential families.

$$
\mathbb{E}\left[t(X)\right] = \frac{d}{d\eta(\theta))}A(\eta(\theta)) \tag{7}
$$