ALPHA DIVERGENCE

Thursday, September 18, 2008
Rice University
STAT 631 / ELEC 633: **Graphical Models**

Scribe:
Andrew WATERS

Instructor:
Dr. Volkan CEVHER

Edited by Ahmad Beirami and David Kahle

1. REVIEW FROM PREVIOUS CLASS

In the previous lecture, we discussed the difficulties of carrying out exact inference when the posterior density $p_{\boldsymbol{Z}|\boldsymbol{X}}$ is either intractable or overly complex.[1] The graph we had in mind is contained in Figure 1; it is this same graph which continues to be of interest in this article. In the situation where the posterior is unwieldy, we noted that is often easier to perform approximate inference over a more tractable density $q_{\boldsymbol{Z}}$ which approximates $p_{\boldsymbol{Z}|\boldsymbol{X}}$.

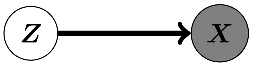

FIGURE 1. Simple Observed Directed Graph

In our pursuit, we discovered that the logarithm of $p_{\boldsymbol{X}}$ admits the expansion

$$\log p_{\boldsymbol{X}}(\boldsymbol{x}) = \mathcal{L}\left(q_{\boldsymbol{Z}}\right) + \mathrm{KL}\left(q_{\boldsymbol{Z}} || p_{\boldsymbol{Z}|\boldsymbol{X}}\right), \tag{1}$$

where

$$\begin{aligned} \mathcal{L}\left(q_{\boldsymbol{Z}}\right) &= \mathbb{E}_{q_{\boldsymbol{Z}}}\left[\log \frac{p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{Z})}{q_{\boldsymbol{Z}}(\boldsymbol{Z})}\right] \\ \mathrm{KL}\left(q_{\boldsymbol{Z}} || p_{\boldsymbol{Z}|\boldsymbol{X}}\right) &= \mathbb{E}_{q_{\boldsymbol{Z}}}\left[\log \frac{q_{\boldsymbol{Z}}(\boldsymbol{Z})}{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{x})}\right]. \end{aligned}$$

The relation turned out to be fundamental in the sense that it provides a means by which we can determine optimal approximations. The optimality requires finding the $q_{\boldsymbol{Z}}$ which minimizes the KL term, the Kullback-Leibler divergence. In this article, we discuss a generalization of the KL divergence, the so-called "α-divergence," and various properties which it exhibits. In addition, our considerations elucidate the meaning of optimality not only for solutions obtained via the minimization of the Kullback-Leibler divergence but also the α divergences as well.

2. α-DIVERGENCES

In previous notes we introduced the KL-divergence and discussed the consequences of using both $\mathrm{KL}\left(q_{\boldsymbol{Z}} || p_{\boldsymbol{Z}|\boldsymbol{X}}\right)$ and $\mathrm{KL}\left(p_{\boldsymbol{Z}|\boldsymbol{X}} || q_{\boldsymbol{Z}}\right)$. In order to obtain more freedom in choosing the metric according to which we are approximating a density, we introduce a more general class of metrics called α-divergences, which can be used to obtain distributions for $q_{\boldsymbol{Z}}$ which approximate $p_{\boldsymbol{Z}|\boldsymbol{X}}$. For ease of notation,

[1] For a more detailed description of notation, see the related article on variational Bayes.

since there is no ambiguity for future reference we will refer to these two densities simply as q and p.

We define the α-divergence as follows:

$$D_\alpha(p||q) = \frac{\int \alpha p(x) + (1-\alpha)q(x) - [p(x)]^\alpha [q(x)]^{1-\alpha} dx}{\alpha(1-\alpha)}, \alpha \in [-\infty, +\infty], \tag{2}$$

Some of the properties of the α-divergence are

(1) $D_\alpha(p||q)$ is convex with respect to both p and q.
(2) $D_\alpha(p||q) \geq 0$
(3) $D_\alpha(p||q) = 0$ when $p = q$ a.e.

Similar to the KL divergence, these properties allow us to minimize the α-divergence in order to find the best approximating distribution $q(x)$ in some class of potential approximations. There are several special cases for various values of α that are of interest to us. The most important cases are

$$\lim_{\alpha \to 0} D_\alpha(p||q) = KL(q||p) \tag{3}$$

$$\lim_{\alpha \to 1} D_\alpha(p||q) = KL(p||q). \tag{4}$$

Hence the α-divergences include the KL divergences as a special case. Other special cases are

$$D_{-1}(p||q) = \frac{1}{2} \int \frac{(q(x) - p(x))^2}{q(x)} dx \tag{5}$$

$$D_2(p||q) = \frac{1}{2} \int \frac{(q(x) - p(x))^2}{p(x)} dx \tag{6}$$

$$D_{\frac{1}{2}}(p||q) = 2 \int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx \tag{7}$$

$D_{\frac{1}{2}}$ is known as the Hellinger distance. $\sqrt{D_{\frac{1}{2}}}$ is a valid distance metric (it satisfies both the traingle inequality and symmetric properties).

3. Investigating the behavior of D_α for different values of α

The definition of the α-divergence in (2) exhibits the reduced representation

$$D_\alpha(p||q) = \frac{1 - \int [p(x)]^\alpha [q(x)]^{1-\alpha} dx}{\alpha(1-\alpha)}. \tag{8}$$

As discussed before, we are interested in approximating an intractable probability distribution $p(x)$ with a tractable distribution $q(x)$. In this process, we introduced α-divergence, $D_\alpha(p||q)$, as a class of pseudo-metrics which measure the accuracy of our approximation. Thus, we minimize $D_\alpha(p||q)$ over a tractable family of approximating distributions $q(x)$ in order to find the best approximation of $p(x)$ in that family.

In this section, we investigate the behavior of the α-divergence for different values of α. Sweeping α from $-\infty$ to ∞ will result in different properties for the resulting approximation $q(x)$. We will start with an example and will discuss the main results. These results will be proven in the next lecture.

We will describe the properties of the approximation using an example. Suppose the density $p(x)$ is given by

$$p(x) = |x-1|^2 e^{-(x-0.5)^2}, \qquad x \in \mathbb{R}. \tag{9}$$

Here, we are interested in finding an approximation to $p(x)$ using a Gaussian distribution. Moreover, we are interested in understanding the resulting "optimal" approximation $q(x)$ as a result of different choices of α. These approximations are demonstrated in Figure 2.

First, assume that α is a very large negative value. In this case, the minimization of $D_\alpha(p||q)$ will force $q(x)$ to be an exclusive approximation, i.e., the mass of $q(x)$ will lie within $p(x)$, as shown in Figure 2 for $\alpha = -11$.

When α is increased toward zero, the approximation starts to lose the exclusivity property. The approximation for $\alpha \leq 0$ satisfies the following: if $p(x) = 0$ at some point x_0, the approximation will be such that $q(x) = 0$ at x_0, as demonstrated in Figure 2.

When $\alpha = 0$, $D_0(p||q) = KL(q||p)$, the Kullback-Leibler divergence. We discussed the properties of KL-divergences in the previous lectures. In particular, $q(x)$ is such that $p(x) = 0 \Rightarrow q(x) = 0$, i.e., the mechanism by which we obtain solutions is zero forcing.

From $\alpha = 0$ to $\alpha = 1$, the approximation changes from $D_0(p||q) = KL(q||p)$ to $D_1(p||q) = KL(p||q)$. For $\alpha \geq 1$, the approximation satisfies the following property: $p(x) > 0 \Rightarrow q(x) > 0$, as demonstrated in Figure 2.

As α grows from 1 to infinity, the approximation becomes inclusive, i.e., the mass of $q(x)$ includes all the mass of $p(x)$.

As discussed above and demonstrated in Figure 2, the value of α plays a significant role in determining the properties of the approximation by minimizing the α-Divergence.

References

1. C.M. Bishop and SpringerLink (Online service), *Pattern recognition and machine learning*, Springer, 2006.

Rice University, Department of Electrical and Computer Engineering, 6100 Main St., Houston, Texas, 77005.

E-mail address: andrew.e.waters@rice.edu, beirami@rice.edu, dkahle@rice.edu, volkan@rice.edu

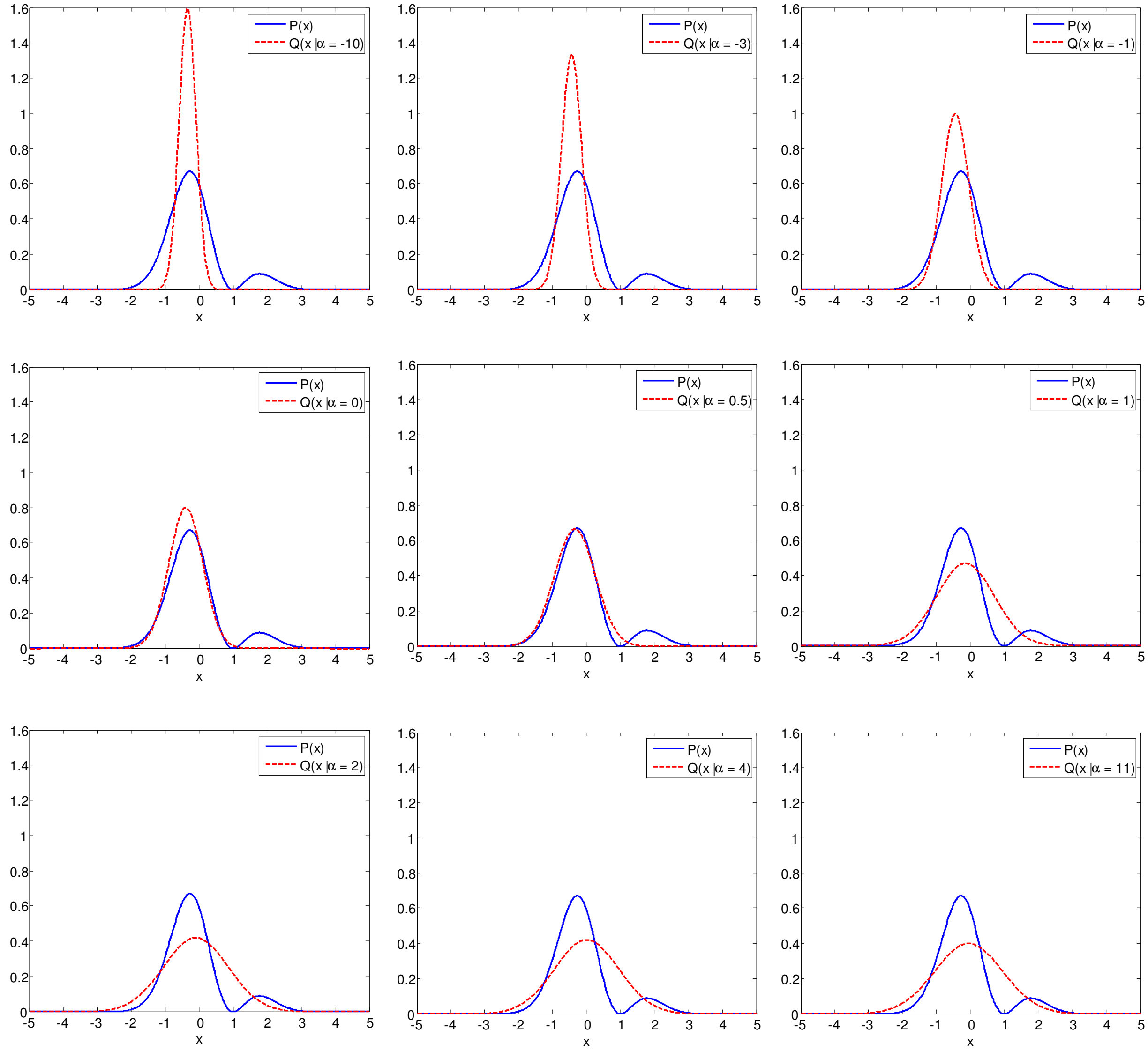

Figure 2. Optimal Gaussian approximation q(x) for p(x) for different values of α