The weird behaviour of the Kullback-Leibler divergence

Jan van Waaij

October 18, 2022

Suppose $Q \ll P$. P and Q have a common dominating measure, say μ, e.g. $P+Q$. Let p and q be the densities of P and Q with respect to μ, respectively. The Kullback-Leibler divergence of P from Q is defined as $K(P;Q) = K_\mu(p;q) = \mathbb{E}_P \log(p/q) = P\log(p/q) = \int \log(p/q)p d\mu$, provided the integral exist. As p/q is almost surely independent of the choice of μ, the definition of the Kullback-Leibler divergence is independent of μ. Recall that the Hellinger distance is defined by $\sqrt{\int(\sqrt{p}-\sqrt{q})^2 d\mu}$.

Theorem 1. *The Kullback-Leibler divergence is positive definite and not necessarily symmetric nor transitive. Furthermore it is bounded below by the squared Hellinger distance.*

In order to prove that the KL divergence is neither symmetric nor transitive, we need to give a counterexample for which we use the Poisson distribution.

Example 2. *Let* $P = \text{Poisson}(\lambda)$ *and* $Q = \text{Poisson}(\mu)$. *Then*

$$K(P;Q) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \left[\mu - \lambda + k \log(\lambda/\mu)\right]$$
$$= \mu - \lambda + \lambda \log(\lambda/\mu).$$

Proof of theorem 1. Note that

$$\int (\sqrt{p}-\sqrt{q})^2 d\mu$$
$$= 2 - 2\int \sqrt{pq}\, d\mu$$
$$= 2\int (1-\sqrt{q/p}) p d\mu.$$

Note that $\log x \le x-1$ for all $x>0$, so $1-x \le \log(x^{-1})$ for all $x>0$ and we can write

$$\int (\sqrt{p}-\sqrt{q})^2 d\mu$$
$$\le 2\int \log(\sqrt{p/q}) p d\mu$$
$$= \int \log(p/q) p d\mu.$$

So the KL-divergence is lower bounded by the Hellinger distance, and is therefore in particular positive definite.

Note that for $P = \text{Poisson}(1), Q = \text{Poisson}(2)$, $K(P,Q) \approx 0.31$, but $K(Q,P) \approx 0.39$, so K is not symmetric. If we define $R = \text{Poisson}(3)$, we have

$$\max\{K(P,Q) + K(R,Q),\ K(P,Q) + K(Q,R),\ K(Q,P) + K(R,Q),\ K(Q,P) + K(Q,R)\}$$
$$\approx 0.60 < 0.90 \approx \min\{K(P,R), K(R,P)\}.$$

So the triangle inequality does not hold. □

The KL paradox

In variational inference (VI), one approximates the posterior P by a simpler function Q. One ought to minimise $K(P,Q)$, however, as this is intractable, one minimises $K(Q,P)$ instead. So one might wonder, that when $K(P,Q) < K(P,R)$, is $K(Q,P) < K(R,P)$ as well? That is not the case, as the following example shows.

Example 3 (**Example with** $K(P,Q) < K(P,R)$**, but** $K(Q,P) > K(R,P)$**).** *Take* $P = \text{Poisson}(3)$*,* $Q = \text{Poisson}(6)$*, and* $R = \text{Poisson}(1)$*. Then* $K(P,Q) \approx 0.92 < 1.30 \approx K(P,R)$*, but* $K(Q,P) \approx 1.16 > 0.90 \approx K(R,P)$*.*

Remark 4. This example shows that when R is a better approximation of P than Q with respect to $K(\cdot\,, P)$, it might be a worse approximation with respect to $K(P, \cdot\,)$.

In VI one searches for a measure Q in a family of probability measures $\mathcal{Q}$ that minimises $K(Q;P)$, where P is the posterior. However (McCulloch, 1989) $K(P;Q)$ measures how good Q approximates P. Choosing a Q that makes $K(Q;P)$ smaller, might make $K(P;Q)$ larger. So this is an argument against the use of VI. The following two examples illustrate this further.

Example 1

Consider $P_m = \text{Poisson}(1/m)$ and $Q_n = \text{Poisson}(e^{-n})$, with $n \in \{m, \dots, 2m\}$. Consider approximating P_m with $Q_n, m \leq n \leq 2m$. Then using that $f(x) = xe^{-x}$ is decreasing for $x > 1$, and $\frac{1}{x}\log x$ is decreasing for $x > e$, we see that

$$\begin{aligned}
0 \leq K(Q_n, P_m) =& \frac{1}{m} - e^{-n} + e^{-n}\log\left(\frac{e^{-n}}{1/m}\right) \\
=& \frac{1}{m} - e^{-n} + e^{-n}\log m - ne^{-n} \\
\leq& \frac{1}{m} - e^{-2m} + e^{-m}\log m - 2me^{-2m} \to 0, \text{ as } m \to \infty
\end{aligned}$$

But

$$\begin{aligned}
K(P_n, Q_n) =& e^{-n} - \frac{1}{m} + \frac{1}{m}\log\left(\frac{1/m}{e^{-n}}\right) \\
=& e^{-n} - \frac{1}{m} - \frac{1}{m}\log m + \frac{n}{m}.
\end{aligned}$$

So

$$e^{-2m} - \frac{1}{m} + 1 - \frac{1}{m}\log m \leq K(P_n, Q_n) \leq e^{-m} - \frac{1}{m} + 2 - \frac{1}{2m}\log(2m).$$

So for $m \geq 3$, and $n \in \{m, \dots, 2m\}$,

$$0.3 \leq K(P_m, Q_n) \leq 2.05.$$

So $Q = \text{argmin}\{Q_n : K(Q_n, P_n), m \leq n \leq 2m\}$ satisfies $K(Q, P_m) \to 0$ as $m \to \infty$, but $K(P_m, Q) \geq 0.3$ for all m.

Example 2

Consider $P = N^+(0, \sigma^2)$ and $Q = \text{Exp}(\lambda)$

Then P has density

$$f_{\sigma^2}(x) = \frac{2}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad x \geq 0,$$

and Q has density

$$g_\lambda(x) = \lambda e^{-\lambda x}.$$

Note that P has mean $\frac{2}{\sqrt{2\pi}}\sigma$ and Q has mean $\frac{1}{\lambda}$.

Then

$$\begin{aligned}
K(P,Q) = &\int_0^\infty \left(\log 2 - \log\sigma - \frac{1}{2}\log(2\pi) - \frac{x^2}{2\sigma^2} - \log\lambda + \lambda x\right) f_{\sigma^2}(x)dx \\
=& \log 2 - \log\sigma - \frac{1}{2}\log(2\pi) - \frac{1}{2} - \log\lambda + \frac{2\sigma\lambda}{\sqrt{2\pi}} \\
=& C_1 - \log(\sigma\lambda) + \frac{2\sigma\lambda}{\sqrt{2\pi}}.
\end{aligned}$$

and

$$\begin{aligned}
K(Q,P) = &\int_0^\infty \left(\log\lambda - \lambda x - \log 2 + \log\sigma + \frac{1}{2}\log(2\pi) + \frac{x^2}{2\sigma^2}\right) g_\lambda(x)dx \\
=& \log\lambda - 1 - \log 2 + \log\sigma + \frac{1}{2}\log(2\pi) + \frac{1}{\sigma^2\lambda^2} \\
=& C_2 + \log(\sigma\lambda) + \frac{1}{\sigma^2\lambda^2}.
\end{aligned}$$

Suppose P is fixed, and first we optimise $K(P,Q)$ over $\lambda \in (0,\infty)$. Then $K(P,Q)$ is minimised for $\lambda = \frac{\sqrt{2\pi}}{2\sigma}$ and $K(Q,P)$ is minimised for $\lambda = \frac{\sqrt{2}}{\sigma}$. So the estimates differ by a factor $\sqrt{\pi}/2 \approx 0.89$.

References

McCulloch, R.E. (1989). "Local Model Influence". In: *Journal of the American Statistical Association* 84.406, pp. 473–478. DOI: 10.1080/01621459.1989.10478793.