

Relationship between Kullback-Leibler Divergence, Fisher Information, and Jeffreys' Prior

Jean-Michel Marin

20 september 2026

1 Kullback-Leibler divergence and Fisher information

Let p and q be two probability densities defined with respect to the same dominating measure. The Kullback-Leibler divergence from q to p is defined by

$$D_{\text{KL}}(p(\cdot) \| q(\cdot)) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx,$$

provided that p is absolutely continuous with respect to q .

Although it is often used as a measure of discrepancy between probability distributions, the Kullback-Leibler divergence is not a distance in the strict mathematical sense, since it is generally not symmetric :

$$D_{\text{KL}}(p(\cdot) \| q(\cdot)) \neq D_{\text{KL}}(q(\cdot) \| p(\cdot)).$$

Moreover,

$$D_{\text{KL}}(p(\cdot) \| q(\cdot)) \geq 0,$$

with equality if and only if $p = q$ almost everywhere.

Consider now a regular parametric family of densities

$$\{f(x|\theta) | \theta \in \Theta\},$$

and two nearby parameter values θ and $\theta + d\theta$. Their Kullback-Leibler divergence is

$$D_{\text{KL}}(f(\cdot|\theta) \| f(\cdot|\theta + d\theta)) = \mathbb{E}_{x|\theta} [\log f(x|\theta) - \log f(x|\theta + d\theta)].$$

We now show how Fisher information appears in the local expansion of this divergence.

Let

$$\ell(\theta; x) = \log f(x|\theta)$$

denote the log-likelihood for one observation.

A second-order Taylor expansion around θ gives

$$\ell(\theta + d\theta; x) = \ell(\theta; x) + d\theta \frac{\partial}{\partial \theta} \ell(\theta; x) + \frac{1}{2} (d\theta)^2 \frac{\partial^2}{\partial \theta^2} \ell(\theta; x) + o((d\theta)^2).$$

Therefore,

$$\ell(\theta; x) - \ell(\theta + d\theta; x) = -d\theta \frac{\partial}{\partial \theta} \ell(\theta; x) - \frac{1}{2} (d\theta)^2 \frac{\partial^2}{\partial \theta^2} \ell(\theta; x) + o((d\theta)^2).$$

Taking the expectation under f_θ yields

$$D_{\text{KL}}(f(\cdot|\theta) \| f(\cdot|\theta + d\theta)) = -d\theta \mathbb{E}_{x|\theta} \left[\frac{\partial}{\partial \theta} \log f(x|\theta) \right] - \frac{1}{2} (d\theta)^2 \mathbb{E}_{x|\theta} \left[\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right] + o((d\theta)^2).$$

Under the usual regularity conditions, the expectation of the score function is zero :

$$\mathbb{E}_{x|\theta} \left[\frac{\partial}{\partial \theta} \log f(x|\theta) \right] = 0.$$

Indeed,

$$\begin{aligned} \mathbb{E}_{x|\theta} \left[\frac{\partial}{\partial \theta} \log f(x|\theta) \right] &= \int f(x|\theta) \frac{\partial}{\partial \theta} \log f(x|\theta) dx \\ &= \int \frac{\partial}{\partial \theta} f(x|\theta) dx \\ &= \frac{\partial}{\partial \theta} \int f(x|\theta) dx \\ &= \frac{\partial}{\partial \theta} 1 = 0. \end{aligned}$$

The Fisher information about θ provided by x can be written equivalently as

$$I_x^\theta(\theta) = \mathbb{E}_{x|\theta} \left[\left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right)^2 \right] = -\mathbb{E}_{x|\theta} \left[\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right].$$

Remark. Fisher introduced his notion of information in 1922, through the likelihood and its curvature. The Kullback–Leibler divergence was introduced later, by Kullback and Leibler in 1951. Therefore, Fisher did not derive Fisher information from the KL divergence.

Consequently, the first-order term vanishes and the second-order term is determined by Fisher information :

$$D_{\text{KL}}(f(\cdot|\theta) \| f(\cdot|\theta + d\theta)) = \frac{1}{2} I_x^\theta(\theta) (d\theta)^2 + o((d\theta)^2).$$

Thus, Fisher information describes the local curvature of the Kullback-Leibler divergence. It quantifies how rapidly the probability distribution changes when the parameter is slightly modified.

Therefore, the same numerical variation in θ does not necessarily have the same statistical significance at every point of the parameter space. A small change in a region where $I_x^\theta(\theta)$ is large produces a greater change in the distribution than the same numerical change in a region where $I_x^\theta(\theta)$ is small.

2 Local Statistical Distance

The previous local expansion,

$$D_{\text{KL}}(f(\cdot|\theta) \| f(\cdot|\theta + d\theta)) = \frac{1}{2} I_x^\theta(\theta) (d\theta)^2 + o((d\theta)^2),$$

suggests defining the squared infinitesimal statistical length by

$$ds^2 = I_x^\theta(\theta) (d\theta)^2,$$

or, equivalently,

$$ds = \sqrt{I_x^\theta(\theta)} |d\theta|.$$

Remark. Although ds and $d\theta$ are both differentials, they have distinct interpretations. The quantity $d\theta$ represents an oriented infinitesimal displacement and may therefore be positive or negative. In contrast, ds represents an infinitesimal length and is nonnegative by definition. Hence,

$$ds = \sqrt{I_x^\theta(\theta)} |d\theta|.$$

Equivalently, one could write

$$|ds| = \sqrt{I_x^\theta(\theta)} |d\theta|,$$

but the absolute value on ds is unnecessary because $ds \geq 0$ by convention.

This length measures the local change in the probability distribution induced by a variation of the parameter, rather than merely the numerical change in the parameter itself. Thus, two parameter increments of the same numerical size may correspond to different statistical changes when they occur at points where the Fisher information differs.

The statistical length does not depend on the particular parameter used to describe the model

Consider a smooth one-to-one reparameterization $\eta = g(\theta)$ with inverse $\theta = g^{-1}(\eta) = h(\eta)$. By the chain rule, the score with respect to η is

$$\frac{\partial}{\partial \eta} \log f(x|h(\eta)) = \frac{\partial}{\partial \theta} \log f(x|\theta) \frac{dh}{d\eta}.$$

Therefore, the Fisher information in the new parameterization is

$$\begin{aligned} I_x^\eta(\eta) &= \mathbb{E}_{x|\eta} \left[\left(\frac{\partial}{\partial \eta} \log f(x|h(\eta)) \right)^2 \right] \\ &= \mathbb{E}_{x|\theta} \left[\left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right)^2 \right] \left(\frac{dh}{d\eta} \right)^2 \\ &= I_x^\theta(\theta) \left(\frac{dh}{d\eta} \right)^2. \end{aligned}$$

Then,

$$\begin{aligned} \sqrt{I_x^\eta(\eta)} |d\eta| &= \sqrt{I_x^\theta(\theta) \left(\frac{dh}{d\eta} \right)^2} |d\eta| \\ &= \sqrt{I_x^\theta(\theta)} \frac{|dh|}{|d\eta|} |d\eta| \\ &= \sqrt{I_x^\theta(\theta)} |dh| \\ &= \sqrt{I_x^\theta(\theta)} |d\theta| \\ &= ds \end{aligned}$$

The quantity ds therefore measures an intrinsic statistical distance between distributions, independently of the choice of the parametrization.

3 The Jeffreys prior distribution

The pdf of the Jeffreys prior distribution is defined by

$$\pi_J(\theta) \propto \sqrt{I_x^\theta(\theta)} \mathbf{1}_\Theta(\theta).$$

The Jeffreys prior distribution can be interpreted as being uniform with respect to the intrinsic statistical length ds , rather than uniform with respect to the numerical coordinate $d\theta$. In other words, equal intervals of statistical length receive equal prior weight.

If the normalizing constant is finite, namely if

$$\int_{\Theta} \sqrt{I_x^\theta(\theta)} d\theta < \infty,$$

The density of the Jeffrey prior distribution can be normalized as

$$\pi_J(\theta) = \frac{\sqrt{I_x^\theta(\theta)}}{\int_{\Theta} \sqrt{I(t)} dt}.$$

Otherwise, the Jeffreys prior distribution is improper, although it may still produce a proper posterior distribution.

Interpretation

Jeffreys' prior does not necessarily assign the same prior mass to two intervals having the same numerical width in the parameter θ . Instead, it assigns equal prior mass to intervals having the same statistical length, whose infinitesimal element is

$$ds = \sqrt{I_x^\theta(\theta)} |d\theta|.$$

Clearly, if $I_x^\theta(\theta)$ is large, a small change in θ produces a substantial change in the corresponding probability distribution and if $I_x^\theta(\theta)$ is small, the same numerical change in θ produces only a small change in the distribution. The Jeffreys prior mass assigned to a small parameter interval is proportional to its statistical length.

The Bernoulli model

Consider the model

$$x \mid \theta \sim \text{Bernoulli}(\theta), \quad 0 < \theta < 1.$$

The probability mass function of one observation is

$$f(x \mid \theta) = \theta^x (1 - \theta)^{1-x}, \quad x \in \{0, 1\}.$$

The log-likelihood is

$$\ell(\theta; x) = x \log \theta + (1 - x) \log(1 - \theta),$$

and the score function is

$$\frac{\partial}{\partial \theta} \ell(\theta; x) = \frac{x}{\theta} - \frac{1-x}{1-\theta}.$$

Therefore, the Fisher information for one observation is

$$I_x^\theta(\theta) = \mathbb{E}_{x|\theta} \left[\left(\frac{\partial}{\partial \theta} \ell(\theta; x) \right)^2 \right] = \frac{1}{\theta(1-\theta)}.$$

The Jeffreys prior distribution is such that

$$\pi_J(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}} \mathbf{1}_{]0,1[}(\theta).$$

After normalization,

$$\pi_J(\theta) = \frac{1}{\pi \sqrt{\theta(1-\theta)}} \mathbf{1}_{]0,1[}(\theta).$$

or, equivalently,

$$\theta \sim \text{Beta} \left(\frac{1}{2}, \frac{1}{2} \right).$$

4 Kullback-Leibler Divergence Between Prior and Posterior

The information provided by the data can be measured by comparing the posterior distribution with the prior distribution. After observing $x^{(n)} = (x_1, \dots, x_n)$, define the information gain by

$$D_{\text{KL}} \left(\pi(\cdot|x^{(n)}) \parallel \pi(\cdot) \right) = \int \pi(\theta|x^{(n)}) \log \left(\frac{\pi(\theta|x^{(n)})}{\pi(\theta)} \right) d\theta.$$

The direction posterior relative to prior is used because this quantity measures how much our knowledge changes after observing the data.

Using Bayes' formula,

$$\pi(\theta|x^{(n)}) = \frac{f(x^{(n)}|\theta)\pi(\theta)}{m(x^{(n)})},$$

we obtain

$$D_{\text{KL}} \left(\pi(\cdot|x^{(n)}) \parallel \pi(\cdot) \right) = \mathbb{E}_{\theta|x^{(n)}} \left[\log f(x^{(n)}|\theta) \right] - \log m(x^{(n)}).$$

Bernstein-von Mises Theorem

Suppose that $x_1, \dots, x_n$ is an n -sample from $x \sim f(x|\theta^*)$. The Bernstein-von Mises theorem provides an asymptotic approximation of the posterior distribution. Under the usual regularity conditions,

$$\pi(\theta|x^{(n)}) \approx \mathcal{N} \left(\hat{\theta}_n, \frac{1}{nI_x^\theta(\theta^*)} \right),$$

where $\hat{\theta}_n$ is the maximum likelihood estimator of θ .

Expected Information Gain and the Jeffreys prior distribution

The expected Kullback-Leibler divergence is

$$\mathbb{E}_{x^{(n)}} \left[D_{\text{KL}} \left(\pi(\cdot|x^{(n)}) \parallel \pi(\cdot) \right) \right].$$

Using the Bernstein-von Mises theorem, it can be shown the the Jeffreys prior distribution is the prior that asymptotically maximizes the expected Kullback-Leibler divergence from the prior to the posterior.