The Kullback-Leibler Divergence

Viktor Glombik

Seminar Optimal Transport
Institute for Mathematics, TU Berlin
30.04.2021
Last edited on May 31, 2022

Introduction: What is the KL divergence and why do we need it?

- KL divergence (or relative entropy) measures discrepancy between measures (but is not a distance).
- Regularises optimal transport problems ("Entropic Regularization"[3, Section 4][1]).
- Is a Bregman distance and a φ-divergence.
- Connected to total variation norm: for $\mu, \nu \in \mathcal{M}^+(X)$

$$\|\mu - \nu\|_{\mathrm{TV}}^2 \leq 2\,\mathrm{KL}(\mu, \nu). \qquad \text{(Pinsker's inequality)}$$

[1] Gabriel Peyré, Marco Cuturi: "Computational Optimal Transport", *Foundations and Trends in Machine Learning*, 2019

Let X be a polish space.

Definition (Kullback-Leibler divergence)

The Kullback-Leibler (KL) divergence is

$$\mathrm{KL}\colon \mathcal{M}^+(X) \times \mathcal{M}^+(X) \to [0, \infty],$$
$$(\mu, \nu) \mapsto \begin{cases} \int_X \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu + \nu(X) - \mu(X), & \text{if } \nu \ll \mu, \\ +\infty, & \text{else,} \end{cases}$$

where $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \in L^1(X, \nu)$ is the Radon-Nykodym derivative.[a]

[a] Beier et al.: *Unbalanced Multi-Marginal Optimal Transport*,, 2021

Special Case: Probability measures

For distributions U and V of an absolutely continuous random variable, with densities u and v (with respect to the Lebesgue measure m on $\mathbb{R}$) this simplifies to

$$\mathrm{KL}(U, V) := \int_{\mathbb{R}} u(x) \log\left(\frac{u(x)}{v(x)}\right) \mathrm{d}m(x),$$

as the last two terms cancel for probability measures.

Consider two random variables $P \sim \mathcal{N}(m, \sigma_1^2)$ and $Q \sim \mathcal{N}(m, \sigma_2^2)$ with $\sigma_1^2, \sigma_2^2 > 0$. Then[2]

$$2\,\mathrm{KL}(P,Q) = \frac{\sigma_1^2}{\sigma_2^2} - 1 + \log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) \xrightarrow{\sigma_2^2 \searrow 0} \infty.$$

As $\sigma_2^2 \searrow 0$, $N(m, \sigma_2^2)$ degenerates to a Dirac measure δ_m.

$\rightsquigarrow$ "Singular Gaussians are infinitely far away from all other Gaussians" [3].

[2] Costa, Santos, and Strapasson: "Fisher information distance: A geometrical reading", *Discrete Applied Mathematics*, 2015

[3] Gabriel Peyré, Marco Cuturi: "Computational Optimal Transport", *Foundations and Trends in Machine Learning*, 2019

KL divergence of Gaussians: Animation

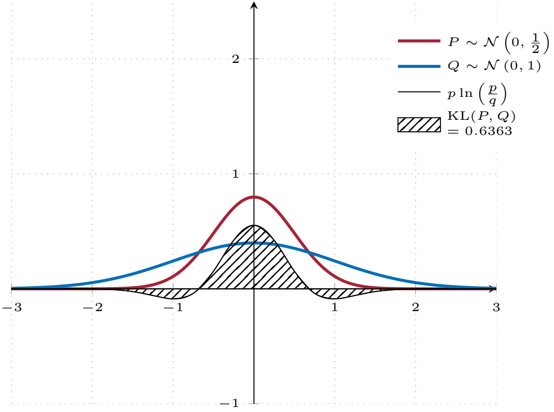


Fig. 1: The KL divergence of two Gaussians, where the variance of one varies.

- Nonnegative: $\mathrm{KL}(\mu,\nu) \geq 0$, (φ-div, Bregman)
- Positive definite: $\mathrm{KL}(\mu,\nu) = 0 \iff \mu = \nu$ almost everywhere, (φ-div)
- Non-symmetric: $\mathrm{KL}(\mu,\nu) \neq \mathrm{KL}(\nu,\mu)$, (Bregman)
- $\mathrm{KL}(\mu,\nu) \not\leq \mathrm{KL}(\mu,\xi) + \mathrm{KL}(\xi,\nu)$, (Bregman)
- jointly convex, strictly convex in first entry.[4] (φ-div)

[4] Beier et al.: *Unbalanced Multi-Marginal Optimal Transport*,, 2021

Thank you for your attention!

References I

[1] Florian Beier et al. *Unbalanced Multi-Marginal Optimal Transport*. 2021. URL: https://arxiv.org/abs/2103.10854.

[2] Sueli Irene R. Costa, Sandra Augusta Santos, and João Eloir Strapasson. "Fisher information distance: A geometrical reading". In: *Discrete Applied Mathematics* 197 (2015). Distance Geometry and Applications, pp. 59–69. URL: https://doi.org/10.1016/j.dam.2014.10.004.

[3] Gabriel Peyré, Marco Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607. URL: https://optimaltransport.github.io/book/.

KL divergence of Gaussians: Animation

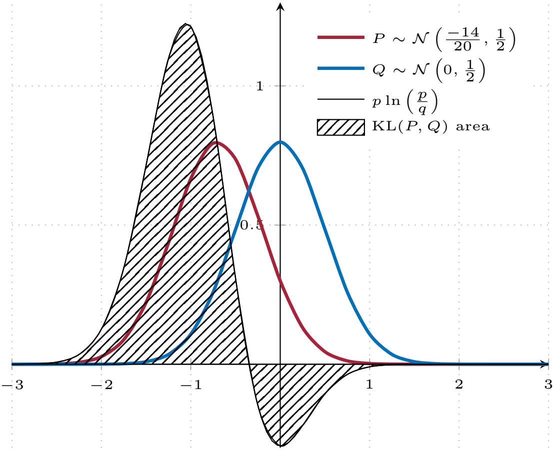


Fig. 2: The KL divergence of two Gaussians, where the mean of one varies.

For $\mu, \nu \in \mathcal{M}^+(X)$ with $\nu \ll \mu$ we have [3]

$$\mathrm{KL}(\mu, \nu) = D_{\varphi_{\mathrm{KL}}}(P, Q) := \int_X \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu,$$

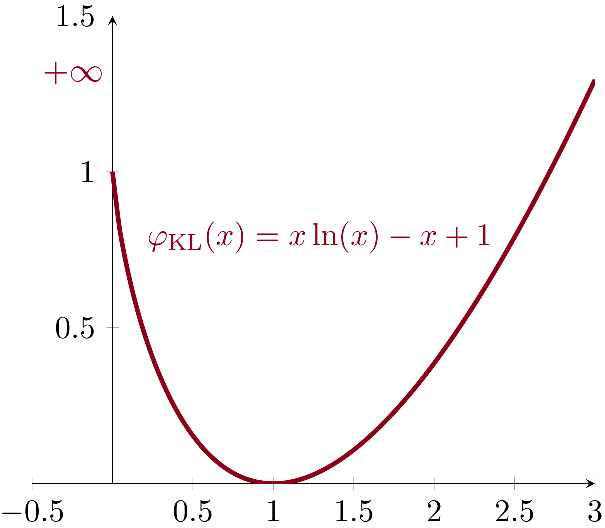
where

$$\varphi_{\mathrm{KL}}(s) := \begin{cases} s\log(s) - s + 1, & \text{for } s > 0, \\ 1, & \text{for } s = 0, \\ +\infty, & \text{otherwise,} \end{cases}$$

as $\varphi_{\mathrm{KL}} \in \Gamma_0(\mathbb{R})$ with $\varphi_{\mathrm{KL}}(1) = 0$.

$$\int_X \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \frac{\mathrm{d}\mu}{\cancel{\mathrm{d}\nu}} \cancel{\mathrm{d}\nu} - \int_X \frac{\mathrm{d}\mu}{\cancel{\mathrm{d}\nu}} \cancel{\mathrm{d}\nu} + \int_X 1 \, \mathrm{d}\nu = \int_X \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu - \mu(X) + \nu(X).$$

The Shannon-Boltzmann entropy function



First, recall that the KL divergence between two distributions P and Q is defined as

$$D_{KL}(P||Q) = \mathrm{E}_P \left[\log \frac{P}{Q}\right].$$

Also, the density function for a multivariate Gaussian (normal) distribution with mean μ and covariance matrix Σ is

$$p(x) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right).$$

Now, consider two multivariate Gaussians in $\mathbb{R}^n$, P_1 and P_2. We have

$$\begin{aligned}
&D(P_1||P_2) \\
&= \mathrm{E}_{P_1}[\log P_1 - \log P_2] \\
&= \frac{1}{2}\mathrm{E}_{P_1}\left[-\log\det\Sigma_1 - (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \log\det\Sigma_2 + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right] \\
&= \frac{1}{2}\log\frac{\det\Sigma_2}{\det\Sigma_1} + \frac{1}{2}\mathrm{E}_{P_1}\left[-(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right] \\
&= \frac{1}{2}\log\frac{\det\Sigma_2}{\det\Sigma_1} + \frac{1}{2}\mathrm{E}_{P_1}\left[-\mathrm{tr}(\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^T) + \mathrm{tr}(\Sigma_2^{-1}(x-\mu_2)(x-\mu_2)^T)\right] \\
&= \frac{1}{2}\log\frac{\det\Sigma_2}{\det\Sigma_1} + \frac{1}{2}\mathrm{E}_{P_1}\left[-\mathrm{tr}(\Sigma_1^{-1}\Sigma_1) + \mathrm{tr}(\Sigma_2^{-1}(xx^T - 2x\mu_2^T + \mu_2\mu_2^T))\right] \\
&= \frac{1}{2}\log\frac{\det\Sigma_2}{\det\Sigma_1} - \frac{1}{2}n + \frac{1}{2}\mathrm{tr}(\Sigma_2^{-1}(\Sigma_1 + \mu_1\mu_1^T - 2\mu_2\mu_1^T + \mu_2\mu_2^T)) \\
&= \frac{1}{2}\left(\log\frac{\det\Sigma_2}{\det\Sigma_1} - n + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + \mathrm{tr}(\mu_1^T\Sigma_2^{-1}\mu_1 - 2\mu_1^T\Sigma_2^{-1}\mu_2 + \mu_2^T\Sigma_2^{-1}\mu_2)\right) \\
&= \frac{1}{2}\left(\log\frac{\det\Sigma_2}{\det\Sigma_1} - n + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right)
\end{aligned}$$