Hölder divergence (HD)

Consider two monotone embeddings[17] $\rho(\cdot)$ and $\tau(\cdot)$ of positive densities $\tilde{p}, \tilde{q} \ll \nu$. The *Hölder divergence* (HD) for $\alpha > 1$ and conjugate exponents $\frac{1}{\alpha} + \frac{1}{\beta} = 1$ ($\beta = \frac{\alpha}{\alpha - 1} > 1$) between two positive densities $\tilde{p}$ and $\tilde{q}$ is defined by:

$$\mathrm{HD}_{\alpha,\rho,\tau}(\tilde{p} : \tilde{q}) = -\log \left(\frac{\int \rho(\tilde{p}(x))\tau(\tilde{q}(x))\mathrm{d}\nu(x)}{\left(\int \rho(\tilde{p}(x))^\alpha \mathrm{d}\nu(x)\right)^{\frac{1}{\alpha}} \left(\int \tau(\tilde{q}(x))^\beta \mathrm{d}\nu(x)\right)^{\frac{1}{\beta}}} \right). \tag{18}$$

When the default identity embedding $\rho(x) = \tau(x) = x$ is considered, we write concisely $\mathrm{HD}_\alpha(\tilde{p} : \tilde{q})$ for $\mathrm{HD}_{\alpha,\rho,\tau}(\tilde{p} : \tilde{q})$. We may also denote by (ρ, τ)-HD the HD divergence taken wrt. the ρ and τ monotone embeddings.

By construction from Rodgers-Hölder inequality[18], the Hölder divergence is zero when:

$$\rho(\tilde{p})^\alpha \propto \tau(\tilde{q})^\beta, \quad \text{ae.} \tag{19}$$

Thus the (ρ, τ)-HD defines a *proper* divergence on probability densities (ie., satisfying the law of indiscernability $\mathrm{HD}_{\alpha,\rho,\tau}(p : q) = 0$ iff. $p(x) = q(x)$ ae.) when:

$$\rho^{-1}(\tau(x)^{\frac{\beta}{\alpha}}) = \rho^{-1}(\tau(x)^{\frac{1}{\alpha-1}}) = x, \alpha > 1, \tag{20}$$

or equivalently when the inverse transformation is the identity function:

$$\tau^{-1}(\rho(x)^{\alpha-1}) = x, \alpha > 1. \tag{21}$$

Key properties

- The (ρ, τ)-HD is a *projective divergence* when the embeddings are *homogeneous functions* (like power functions):

$$\forall \lambda, \lambda' > 0, \quad \mathrm{HD}_{\alpha,\rho,\tau}(\lambda\tilde{p} : \lambda'\tilde{q}) = \mathrm{HD}_{\alpha,\rho,\tau}(\tilde{p} : \tilde{q}), \quad \rho(\gamma x) = \gamma^a \rho(x), \tau(\gamma x) = \gamma^b \tau(x). \tag{22}$$

Notable members

- The Bhattacharyya distance[19] is a Hölder divergence in disguise for the *square root embedding* $\rho(x) = \tau(x) = \sqrt{x}$ of densities:

$$\mathrm{Bhat}(p : q) = -\log \int \sqrt{p(x)q(x)}\mathrm{d}\nu(x) = -\log \frac{\int \sqrt{p(x)q(x)}\mathrm{d}\nu(x)}{(\int \sqrt{p(x)}^2 \mathrm{d}\nu(x))^{\frac{1}{2}} (\int \sqrt{q(x)}^2 \mathrm{d}\nu(x))^{\frac{1}{2}}}, \tag{23}$$

since $\int \sqrt{p(x)}^2 \mathrm{d}\nu(x) = \int \sqrt{q(x)}^2 \mathrm{d}\nu(x) = 1$.

It follows that the Bhattacharyya distance can be extended to positive measures as follows:

$$\mathrm{Bhat}(\tilde{p} : \tilde{q}) = -\log \frac{\int \sqrt{\tilde{p}(x)\tilde{q}(x)}\mathrm{d}\nu(x)}{(\int \sqrt{\tilde{p}(x)}^2 \mathrm{d}\nu(x))^{\frac{1}{2}} (\int \sqrt{\tilde{q}(x)}^2 \mathrm{d}\nu(x))^{\frac{1}{2}}} = -\log \frac{\int \sqrt{\tilde{p}(x)\tilde{q}(x)}\mathrm{d}\nu(x)}{(\int \tilde{p}(x)\mathrm{d}\nu(x))^{\frac{1}{2}} (\int \tilde{q}(x)\mathrm{d}\nu(x))^{\frac{1}{2}}}. \tag{24}$$

A similar interpretation holds for the skewed Bhattacharyya distance[20].

- The Cauchy-Schwarz divergence[21] is the only *symmetric* Hölder divergence obtained for $\alpha = \beta = 2$:

$$\mathrm{CSD}(\tilde{p}, \tilde{q}) = -\log \frac{\int \tilde{p}(x)\tilde{q}(x)\mathrm{d}\nu(x)}{\sqrt{\int \tilde{p}(x)^2 \mathrm{d}\nu(x)) \int \tilde{q}(x)^2 \mathrm{d}\nu(x)}}. \tag{25}$$

The Cauchy-Schwarz divergence is a log-ratio gap divergence derived from the Cauchy-Buniakovski-Schwarz inequality. This divergence admits closed-form formulas for mixtures of exponential families with conic natural parameter space[22].

[17] Jun Zhang. "On monotone embedding in information geometry". In: *Entropy* 17.7 (2015), pp. 4485–4499.
[18] Frank Nielsen, Ke Sun, and Stéphane Marchand-Maillet. "On Hölder projective divergences". In: *CoRR* abs/1701.03916 (2017). URL: `https://arxiv.org/abs/1701.03916`.
[19] A. Bhattacharyya. "On a measure of divergence between two statistical populations defined by their probability distributions". In: *Bulletin of the Calcutta Mathematical Society* 35 (1943), pp. 99–109.
[20] Frank Nielsen and Sylvain Boltz. "The Burbea-Rao and Bhattacharyya centroids". In: *IEEE Transactions on Information Theory* 57.8 (2011), pp. 5455–5466.
[21] Marcin Budka, Bogdan Gabrys, and Katarzyna Musial. "On accuracy of PDF divergence estimators and their applicability to representative data sampling". In: *Entropy* 13.7 (2011), pp. 1229–1266.
[22] Frank Nielsen. "Closed-form information-theoretic divergences for statistical mixtures". In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. Nov. 2012, pp. 1723–1726.

Notes

The Hölder divergence is derived from the Rodgers-Hölder inequality[23] (log-ratio gap). Hölder divergences are studied in[24] where closed-form formulas are reported for distributions belonging to the exponential families with conic natural parameter spaces. This *inequality-induced Hölder divergence* shall not to be confused with the *score-induced Hölder divergence*[25].

---

[23] Leonhard James Rogers. "An extension of a certain theorem in inequalities". In: *Messenger of Math* 17 (1888), pp. 145–150; Otto Ludwig Holder. "Über einen Mittelwertssatz". In: *Nachr. Akad. Wiss. Gottingen Math.-Phys. Kl.* (1889).

[24] Nielsen, Sun, and Marchand-Maillet, "On Hölder projective divergences".

[25] Takafumi Kanamori, Hironori Fujisawa, et al. "Affine invariant divergences associated with proper composite scoring rules and their applications". In: *Bernoulli* 20.4 (2014), pp. 2278–2304; Takafumi Kanamori. "Scale-invariant divergences for density functions". In: *Entropy* 16.5 (2014), pp. 2611–2628.