The symmetrized Bregman divergence as dual geodesic energy functionals*

Frank Nielsen
Frank.Nielsen@acm.org

June 2022

The Bregman divergence [2] between two vector parameters θ_1 and θ_2 induced by strictly convex function $F(\theta)$ is $B_F(\theta_1 : \theta_2) := F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2)$. Let $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$ denote the dual parameterizations obtained by the Legendre-Fenchel convex conjugate $F^*(\eta)$ of $F(\theta)$. The Jeffreys-type symmetrization[1] (i.e., arithmetic symmetrization of the sided Bregman divergences [5]) of the Bregman divergence is defined by

$$S_F(\theta_1; \theta_2) := B_F(\theta_1 : \theta_2) + B_F(\theta_2 : \theta_1) = (\theta_2 - \theta_1)^\top (\eta_2 - \eta_1) = S_{F^*}(\eta_1; \eta_2),$$

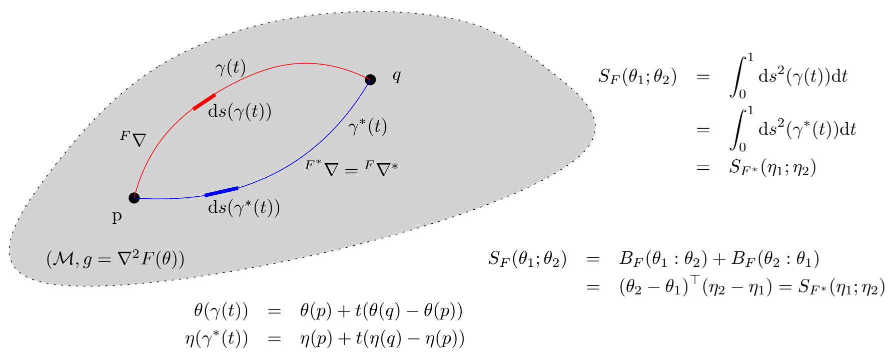

Figure 1: The symmetrized Bregman divergence $S_F(\theta_p; \theta_q)$ can be interpreted as the energy of the Hessian metric along the primal or dual geodesics linking p to q.

Proposition 1 (Theorem 3.2 of [1], illustration in Figure 1) *The Jeffreys-Bregman divergence $S_F(\theta_1; \theta_2)$ is interpreted as the energy induced by the Hessian metric $\nabla^2 F(\theta)$ on the dual geodesics:*

$$S_F(\theta_1; \theta_2) = \int_0^1 \mathrm{d}s^2(\gamma(t))\mathrm{d}t = \int_0^1 \mathrm{d}s^2(\gamma^*(t))\mathrm{d}t.$$

*Extracted from ongoing working notes [4].

[1] To distinguish it with the Jensen-Shannon type symmetrization [3].

Proof: The proof is based on the first-order and second-order directional derivatives. The first-order directional derivative $\nabla_u F(\theta)$ with respect to vector u is defined by

$$\nabla_u F(\theta) = \lim_{t \to 0} \frac{F(\theta + tv) - F(\theta)}{t} = v^\top \nabla F(\theta).$$

The second-order directional derivatives $\nabla^2_{u,v} F(\theta)$ is

$$\begin{aligned} \nabla^2_{u,v} F(\theta) &= \nabla_u \nabla_v F(\theta), \\ &= \lim_{t \to 0} \frac{v^\top \nabla F(\theta + tu) - v^\top \nabla F(\theta)}{t}, \\ &= u^\top \nabla^2 F(\theta) v. \end{aligned}$$

Now consider the squared length element $\mathrm{d}s^2(\gamma(t))$ on the primal geodesic $\gamma(t)$ expressed using the primal coordinate system θ: $\mathrm{d}s^2(\gamma(t)) = \mathrm{d}\theta(t)^\top \nabla^2 F(\theta(t)) \mathrm{d}\theta(t)$ with $\theta(\gamma(t)) = \theta_1 + t(\theta_2 - \theta_1)$ and $\mathrm{d}\theta(t) = \theta_2 - \theta_1$. Let us express the $\mathrm{d}s^2(\gamma(t))$ using the second-order directional derivative:

$$\mathrm{d}s^2(\gamma(t)) = \nabla^2_{\theta_2 - \theta_1} F(\theta(t)).$$

Thus we have $\int_0^1 \mathrm{d}s^2(\gamma(t))\mathrm{d}t = [\nabla_{\theta_2-\theta_1} F(\theta(t))]_0^1$, where the first-order directional derivative is $\nabla_{\theta_2-\theta_1} F(\theta(t)) = (\theta_2 - \theta_1)^\top \nabla F(\theta(t))$. Therefore we get $\int_0^1 \mathrm{d}s^2(\gamma(t))\mathrm{d}t = (\theta_2 - \theta_1)^\top (\nabla F(\theta_2) - \nabla F(\theta_1)) = S_F(\theta_1; \theta_2)$.

Similarly, we express the squared length element $\mathrm{d}s^2(\gamma^*(t))$ using the dual coordinate system η as the second-order directional derivative of $F^*(\eta(t))$ with $\eta(\gamma^*(t)) = \eta_1 + t(\eta_2 - \eta_1)$:

$$\mathrm{d}s^2(\gamma^*(t)) = \nabla^2_{\eta_2 - \eta_1} F^*(\eta(t)).$$

Therefore, we have $\int_0^1 \mathrm{d}s^2(\gamma^*(t))\mathrm{d}t = [\nabla_{\eta_2-\eta_1} F^*(\eta(t))]_0^1 = S_{F^*}(\eta_1; \eta 2)$. Since $S_{F^*}(\eta_1; \eta_2) = S_F(\theta_1; \theta_2)$, we conclude that

$$S_F(\theta_1; \theta_2) = \int_0^1 \mathrm{d}s^2(\gamma(t))\mathrm{d}t = \int_0^1 \mathrm{d}s^2(\gamma^*(t))\mathrm{d}t$$

In 1D, both pregeodesics $\gamma(t)$ and $\gamma^*(t)$ coincide. We have $\mathrm{d}s^2(t) = (\theta_2 - \theta_1)^2 f''(\theta(t)) = (\eta_2 - \eta_1) f^{*\prime\prime}(\eta(t))$ so that we check that $S_F(\theta_1; \theta_2) = \int_0^1 \mathrm{d}s^2(\gamma(t))\mathrm{d}t = (\theta_2 - \theta_1)[f'(\theta(t))]_0^1 = (\eta_2 - \eta_1)[f^{*\prime}(\eta(t))]_0^1 = (\eta_2 - \eta_1)(\theta_2 - \theta_2)$. □

Remark 1 *In Riemannian geometry, a curve γ minimizes the energy $E(\gamma) = \int_0^1 |\dot\gamma(t)|^2 \mathrm{d}t$ if it minimizes the length $L(\gamma) = \int_0^1 \|\dot\gamma(t)\| \mathrm{d}t$ and $\|\dot\gamma(t)\|$ is constant. Using Cauchy-Schwartz inequality, we can show that $L(\gamma) \leq E(\gamma)$.*

References

[1] Shun-ichi Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan, 2016.

[2] Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[3] Frank Nielsen. On the Jensen–Shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019.

[4] Frank Nielsen. *Information Geometry for Machine Learning*. 2022. ongoing working notes.

[5] Frank Nielsen and Richard Nock. Sided and symmetrized Bregman centroids. *IEEE transactions on Information Theory*, 55(6):2882–2904, 2009.