Information Geometry: Applications on Dually Flat Manifolds*

Maximilian Herzog

July 18, 2022

Abstract

Dually flat manifolds, also called Bregmann geometries, represent a perfect setting to apply information geometric algorithms, as the primal/ dual geodesics are straight lines in the respective coordinate system and are therefore computationally attractive. In this work, we will study two possible applications on the dually flat statistical manifolds generated by exponential families and mixture families.

Dually flat manifolds – Bregmann geometry

We will start of by reviewing some core concepts of dually flat manifolds(DFM). For a full discussion of DFMs we refer to nielsens work. [2, 5, 4] At the center of Information Geometry is the divergence $D(\cdot:\cdot)$, a smooth and potentially asymmetric distance measure. It generates a dualistic geometry on the manifold. [2] For dually flat manifolds, it is given by the *Bregmann divergence*

$$B_F(\theta_1:\theta_2) := F(\theta_1) - F(\theta_2) - \langle \theta_1-\theta_2, \nabla F(\theta_2)\rangle \tag{1}$$

where $F \in C^2$ is a strictly convex, D-dimensional real-valued function, the *Bregmann generator*.

It is then very interesting to look at the dual of the divergence, which is given by swapping its parameters $D^*(\theta_1:\theta_2) = D(\theta_2:\theta_1)$. In the case of a Bregmann divergence, this swapping is equivalent to a *Legendre-Fenchel transformation* on F, i.e. $B^*_F(\theta_1:\theta_2) = B_{F^*}(\theta_1:\theta_2) = B_F(\theta_2:\theta_1)$, with $F^*(\eta) = \sup_{\theta\in\Theta}\{\langle\theta,\eta\rangle - F(\theta)\}$. One introduces *dual coordinates* $\eta(\theta) = \nabla F(\theta)$ for which the supremum is reached. It then follows that $\theta(\eta) = \nabla F^*(\eta) = (\nabla F)^{-1}(\eta)$. Notice that as F is strictly convex, ∇F and $(\nabla F)^{-1} = \nabla F^*$ are defined.

We can then build up the information manifold with metric g and dual affine connections $(M, {}^F g, {}^F\nabla, {}^F\nabla^*)$

$${}^F g_{ij} = -\partial_i\partial_j B_F(\theta:\theta')|_{\theta=\theta'} = \partial_i\partial_j F(\theta) \quad \text{and} \quad {}^{F^*}g^{ij} = \partial^i\partial^j F^*(\eta) \tag{2}$$

$${}^F\Gamma_{ij}{}^k = -\partial_i\partial_j\partial^k B_F(\theta:\theta')|_{\theta=\theta'} = 0 = {}^{F^*}\Gamma_{ij}{}^k, \tag{3}$$

where $\partial_l :=: \frac{\partial}{\partial\theta^l}$ and $\partial^l :=: \frac{\partial}{\partial\eta_l}$. The dual flatness is the main point here. Note that single flatness would be forbidden by the fundamental theorem of information geometry.[2] For us the most important consequence is that both of the geodesics are straight lines in their respective coordinate systems: In primal(exponential) $\gamma_e(\theta,\theta') = \{(1-\lambda)\theta + \lambda\theta'\}_{\lambda\in[0,1]}$ and analogously for the dual(mixture) geodesics $\gamma^*(\eta,\eta')$.

As Bregmann divergences are the *canonical divergences* of dually flat spaces [1], i.e. one can for every DFM find a corresponding generator F such that $(M, {}^D g, {}^D\nabla, {}^D\nabla^*) \equiv (M, \nabla^2 F, {}^F\nabla, {}^F\nabla^*) =: (M, F)$, one refers to the geometry of dually flat spaces as Bregmann geometry.

We will now discuss the statistical mixture and exponential manifolds, as these have the computationally nice property of being dually flat. There B_F will turn out to be given by the *Kullback-Leibler divergence* $\mathrm{KL}[p(x):q(x)] = \int_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}\,\mathrm{d}\mu(x)$ and its dual KL^*.

*Manuscript to a talk given on July 15, 2022

Hypothesis testing in exponential families

An exponential family is given by the *sufficient statistics vector* $t(x)$ and a measure $\mu(x)$ as

$$\varepsilon_{t,\mu} = \{p(x;\theta) \propto \exp(\langle t(x), \theta\rangle)\}_\theta \tag{4}$$

with $p(x;\theta) := \frac{\mathrm{d}P(\theta)}{\mathrm{d}\mu}(x)$. The densities are then normalized by the *cumulant function*

$$F(\theta) := \ln\left(\int_{x\in\mathcal{X}} \exp(\langle t(x),\theta\rangle)\,\mathrm{d}\mu(x)\right). \tag{5}$$

To describe even more probability distribution families, like Poisson distributions, one can also include an extra *carrier term* $k(x)$ with the adjusted measure $\nu(x) := \frac{\mu(x)}{\exp(k(x))}$: $\varepsilon_{t,k,\mu} = \{p(x;\theta) \propto \exp(\langle t(x),\theta\rangle + k(x))\}_\theta$.

Interestingly, as we mentioned above, one gets that using the cumulant function F in (5) as a Bregmann generator one obtains:

$$\begin{aligned}(\varepsilon_{t,\mu}, \mathrm{KL}*, \nabla^*_{\mathrm{KL}}, \nabla_{\mathrm{KL}}) &\cong (\Theta, F)\\ \mathrm{KL}[P_{\theta_1} : P_{\theta_2}] &= B_F(\theta_2 : \theta_1)\\ &= B_{F^*}(\eta_1 : \eta_2)\end{aligned} \tag{6}$$

In the case of exponential families without a carrier term, η is terned the *expectation parameter*, as it holds that $\eta = \nabla F(\theta) = E_{p(x;\theta)}[t(x)] = \int_{x\in\mathcal{X}} p(x;\theta)t(x)\,\mathrm{d}\mu(x)$. Also the Fisher-information metric is equal to the variance of $t(x)$, i.e. $g = \mathrm{Var}_\theta[t(x)]$. We are now set up to study our first application.

In *binary hypothesis testing*, the task is, given two probability distributions, $P_0 \sim p_0(x)$ and $P_1 \sim p_1(x)$, to classify whether n given data points $X_{1:n} = \{x_1,\ldots,x_n\} \subset \mathcal{X}$ are sampled from P_0 or P_1, denoted as hypothesis H_0 and H_1 respectively. The *error exponent* describes how the probability of an error occurring (either a false positive $P(\text{error}|H_0)$ or false negative $P(\text{error}|H_1)$) decays with a rate of α when increasing the sample size n, i.e. $P_{\text{error}} \propto e^{-n\alpha}$. It is therefore defined as

$$\alpha := \lim_{n\to\infty} \frac{-\ln P_{\text{error}}}{n}. \tag{7}$$

When minimizing the Bayesian average error, then the optimal error exponent is given as $\alpha = \lim_{n\to\infty} \frac{-\ln P_{\text{error}}}{n} = C[P_0 : P_1]$. Here, $C[P_0 : P_1]$ is the *Chernoff-information*, which maximizes the *Bhattacharyya-distance* $B_\alpha[P_0 : P_1]$

$$C[P_0 : P_1] = \max_{\alpha\in(0,1)} B_\alpha[P_0 : P_1] = -\ln \min_{\alpha\in(0,1)} \int_{x\in\mathcal{X}} p_0^\alpha(x) p_1^{\alpha-1}(x)\,\mathrm{d}\mu(x) \in (0,\infty) \tag{8}$$

$$B_\alpha[P_0 : P_1] = -\ln \underbrace{\int_{x\in\mathcal{X}} p_0^\alpha(x) p_1^{\alpha-1}(x)\,\mathrm{d}\mu(x)}_{\in[0,1)} \tag{9}$$

Intuitively, the Chernoff-information quantifies the unavoidable overlap between both of the probability distributions, and therefore yields the best error exponent. Note that by construction, the Chernoff-information is symmetric, $C[P : Q] = C[P, Q]$ as we can redefine $\alpha \to 1-\alpha$. Interestingly, we managed to project this idea of the best exponent error to the realm of distance measures between distributions, handled by information geometry.

If we now assume that the P_0 and P_1 are obtained from the same exponential family $\varepsilon_{t,\mu}$, with its corresponding cumulant function F (see (5)) at hand, the Bhattacharyya-distance can be shown to be equivalent to the *Jensen-skew parameter divergence*[3]

$$B_\alpha[P_{\theta_0} : P_{\theta_1}] = J_F^\alpha(\theta_0 : \theta_1) = \alpha F(\theta_0) + (1-\alpha)F(\theta_1) - F(\theta_{01}^\alpha) \tag{10}$$

where $\theta_{01}^\alpha := \alpha\theta_0 + (1-\alpha)\theta_1$ is the linear interpolation between both distributions. Consequently, finding the Chernoff-information amounts to maximizing $J_F^\alpha(\theta_0 : \theta_1)$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} J_F^\alpha(\theta_0 : \theta_1) = F(\theta_0) - F(\theta_1) - \langle \theta_0 - \theta_1, \nabla F(\theta_{01}^\alpha) \rangle \stackrel{!}{=} 0 \tag{11}$$

A solution for α exists, because $J_F'^\alpha(\theta_0 : \theta_1) = -\langle (\theta_0 - \theta_1)^T \nabla^2 F(\theta_{01}^\alpha), \theta_0 - \theta_1 \rangle < 0$ for $\theta_0 \neq \theta_1$ as $\nabla^2 F \succ 0$. We will call it the best exponent error α^*.[2]

Consider now again the Bregmann divergence

$$\begin{aligned} B_F(\theta_0 : \theta_{01}^{\alpha^*}) &= F(\theta_0) - F(\theta_{01}^{\alpha^*}) - \langle \theta_0 - \theta_{01}^{\alpha^*}, \nabla F(\theta_{01}^{\alpha^*}) \rangle \\ &= F(\theta_0) - F(\theta_{01}^{\alpha^*}) - (1-\alpha)\langle \theta_0 - \theta_1, \nabla F(\theta_{01}^{\alpha^*}) \rangle \\ &= F(\theta_0) - F(\theta_{01}^{\alpha^*}) - (1-\alpha)(F(\theta_0) - F(\theta_1)) \\ &= \alpha F(\theta_0) + (1-\alpha) F(\theta_1) - F(\theta_{01}^\alpha) \\ &= J_F^{\alpha^*}(\theta_0 : \theta_1) = C[P_{\theta_0}, P_{\theta_1}] \end{aligned}$$

We conclude that the Chernoff-information can also be calculated using the Bregmann divergence. Further, a similar calculation can be done for $B_F(\theta_1 : \theta_{01}^{\alpha^*})$:

$$C[P_{\theta_0}, P_{\theta_1}] = B_F(\theta_0 : \theta_{01}^{\alpha^*}) = B_F(\theta_1 : \theta_{01}^{\alpha^*}) \tag{12}$$

$$= \mathrm{KL}[\theta_{01}^{\alpha^*} : \theta_0] = \mathrm{KL}[\theta_{01}^{\alpha^*} : \theta_1]. \tag{13}$$

Geometrically, this can be understood as $P^* \sim \theta_{01}^{\alpha^*}$ laying on the m-bisector between P_0 and P_1, i.e. the mixture bisector, which is a hyperplane in η coordinates, and a *'curved'* hypersurface in θ coordinates. It is given as

$$\mathrm{Bi}_m(P_{\theta_0}, P_{\theta_1}) = \{P_\theta : B_F(\theta_0 : \theta) - B_F(\theta_1 : \theta) = F(\theta_0) - F(\theta_1) + \langle \nabla F(\theta), \theta_1 - \theta_0 \rangle = 0\}. \tag{14}$$

Alternatively one can construct the exponential bisectors Bi_e, where the variable θ would occur in the first argument of B_F. See also [3].

Also, as $\theta_{01}^{\alpha^*}$ is just the interpolation between both points, it should lay on the e-geodesic $\gamma_e(P_1, P_2) = \{\lambda P_1 + (1-\lambda)P_2\}_{\lambda \in [0,1]}$(primal geodesic). Therefore the Chernoff distribution is the point of intersection

$$P_{\theta_{01}^{\alpha^*}} = \gamma_e(\theta_0, \theta_1) \cap \mathrm{Bi}_m(\theta_0, \theta_1). \tag{15}$$

In practice, the problem of finding α^* still remains a standard optimization problem. One could for example try to bracket in α^* by choosing $\alpha_- = 0$, $\alpha_+ = 1$ and then defining $\alpha = (\alpha_+ + \alpha_-)/2$ and testing which point is closer in order to adjust α to go away from that point, i.e. if $B_F(\theta_0 : \theta_{01}^\alpha) < B_F(\theta_1 : \theta_{01}^\alpha)$ set $\alpha_+ = \alpha$, otherwise set $\alpha_- = \alpha$ and reiterate to arbitrary precision. This procedure can be similarly be performed in dual coordinates η.[3]

What we achieved is finding the best error exponent α^* and the corresponding Chernoff distribution $P_{\theta_{01}^{\alpha^*}}$ which is useful in information fusion, trying to find a distribution *between* two distributions. Please find the code for this manuscript, where this has been done numerically. The result can be seen in Figure 1.

As another step one could consider more possible distributions that the data could instead be sampled from. If we include those in our consideration, we obtain a *Bregmann Voronoi diagram*, where the Chernoff-information α between each pair is computed.[2, 4]

Mixture clustering in mixture families and Monte Carlo information geometry

Given a set of $D+1$ probability densities $p_1(x), \ldots, p_D(x)$ all sharing the same support $\mathcal{X}$, a mixture family consist of all strictly convex combinations of these densities:

$$\mathcal{M}_{\{p_1,\ldots,p_D\},\mu} := \left\{ m(x;\eta) = \sum_{i=1}^{D} \eta_i p_i(x) + (1 - \sum_{i=1}^{D} \eta_i) p_0(x) : \eta_i > 0 \, \forall i, \sum_{i=1}^{D} \eta_i < 1 \right\} \tag{16}$$

Funnily enough, applying the Kullback-Leibler divergence KL to this family again yields a dually flat manifold

$$\begin{aligned} (\mathcal{M}, \mathrm{KL}, \nabla_{\mathrm{KL}}, \nabla^*_{\mathrm{KL}}) &\cong (\mathcal{M}, F) \\ \mathrm{KL}[P_{\eta_1} : P_{\eta_2}] &= B_F(\eta_1 : \eta_2) \\ &= B_{F^*}(\theta_2 : \theta_1), \end{aligned}$$

where the Bregmann generator is given by the negative Shannon entropy

$$F(\eta) = -h(m(x;\eta)) = \int_{x \in \mathcal{X}} m(x;\eta) \ln m(x;\eta) \, \mathrm{d}\mu(x). \tag{17}$$

Note that naming the parameter coordinates η is arbitrary, and is just to ensure the same ordering in the Bregmann divergence as for the exponential family case (6).

Most of the time $F(\eta)$ is not available in closed form and is computationally intractable. In this section we will present Monte Carlo information geometry, an idea put forward by [5] to sample the Bregmann generator $F_S(\eta) \approx F(\eta)$ on some set of supports $S = x_1, \ldots, x_m \subset \mathcal{X}$

$$F(\eta) \approx F_S(\eta) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{q(x_i)} m(x_i;\eta) \ln m(x_i;\eta), \tag{18}$$

where $q(x)$ is the *proposal distribution*, which in practice can be chosen as the uniform mixture $q(x) = m(x; \eta_i = \frac{1}{D})$.

We then can use this approximated Bregmann generator to apply our favorite information geometric algorithms. It can be shown that as putting more and more points into our set of samples $S_m \subset S_{m+1}$, the geometry, i.e. the divergence converges to the true Bregmann generator pointwise.

$$\lim_{m \to \infty} F_{S_m} = F \text{ pointwise} \tag{19}$$

One interesting application is *mixture clustering*. Consider the task of clustering $m_{\eta_0}, \ldots, m_{\eta_N}$ living in the same mixture family $\mathcal{M}$ by their KL distance to each other. Correspondingly we can use the Monte Carlo Bregmann divergence B_{F_S} on the parameter space η. To group the N mixtures into k clusters, one may use the Bregmann version of the k-Means algorithm[6]: Start by choosing k randomly chosen means. Then, add all the points that are – by means of Bregmann divergence in the first argument – *closest to* each mean to the corresponding cluster. Recalculate the means and repeat the distance considerations. When nothing changes the algorithm is considered complete.

In the python notebooks, this has been implemented, and the results can be seen in Figure 2. Also a Bregmann-Voronoi diagram has been computed, where we used a Laguerre-Voronoi diagram, which is equivalent considering a specific circle radius.[4]

Monte Carlo information geometry of course has implications beyond mixture clustering and the estimation of the Bregmann generator F is useful in many cases, where F is not computable. This includes exponential families with higher order sufficient statistics vectors $t(x)$ using a suited approximation for the generator.[5]

Conclusion

We have seen two applications on dually flat manifolds. The key point was that we were able to use the straight line geodesics in order to simplify the computations. Notice, that there are many more possible choices for DFMs.[1] Furthermore, we refer to [1] for more applications of information geometry.

Please find the python notebooks to reproduce the plots shown in this article at
https://github.com/maxiherzog/infogeo-applications

References

[1] Shun-ichi Amari. *Information Geometry and Its Applications*. 1st. Springer Publishing Company, Incorporated, 2016. ISBN: 4431559779.

[2] Frank Nielsen. "An elementary introduction to information geometry". In: *CoRR* abs/1808.08271 (2018). arXiv: 1808.08271. URL: http://arxiv.org/abs/1808.08271.

[3] Frank Nielsen. "An Information-Geometric Characterization of Chernoff Information". In: *IEEE Signal Processing Letters* 20 (Mar. 2013), pp. 269–272. DOI: 10.1109/LSP.2013.2243726.

[4] Frank Nielsen, Jean-Daniel Boissonnat, and Richard Nock. "Bregman Voronoi Diagrams: Properties, Algorithms and Applications". In: *CoRR* abs/0709.2196 (2007). arXiv: 0709.2196. URL: http://arxiv.org/abs/0709.2196.

[5] Frank Nielsen and Gaëtan Hadjeres. "Monte Carlo Information Geometry: The dually flat case". In: *CoRR* abs/1803.07225 (2018). arXiv: 1803.07225. URL: http://arxiv.org/abs/1803.07225.

[6] Julie Zhang. "A Generalization of K-Means Clustering Using Bregman Divergences". In: 2020.

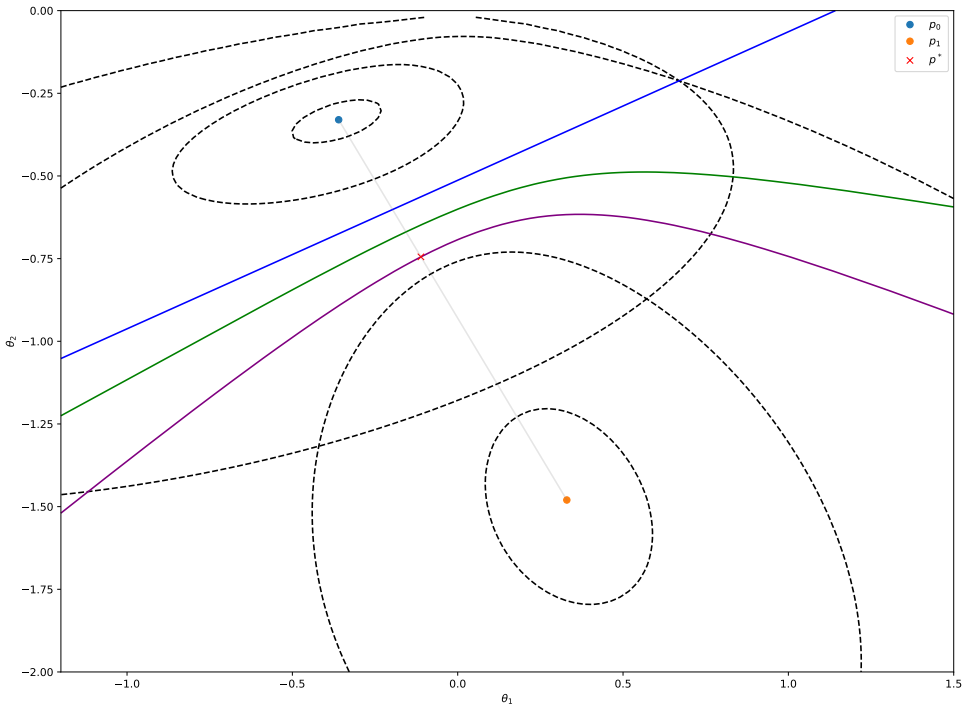

Figure 1: Determination of α depicted in parameter space θ. The primal geodesic γ_e is depicted in gray, Bi_e in blue, Bi_m in purple, and the bisector of the symmetrized divergence $S(\theta_1, \theta_2) = \frac{1}{2}(B_F(\theta_1 : \theta_2) + B_F(\theta_2 : \theta_1))$ in green. Furthermore, some Bregmann balls of the first type around P_0 and P_1 are shown in dashed lines. $\alpha \approx 0.638$ is determined as the intersection of γ_e and Bi_m, see (15).

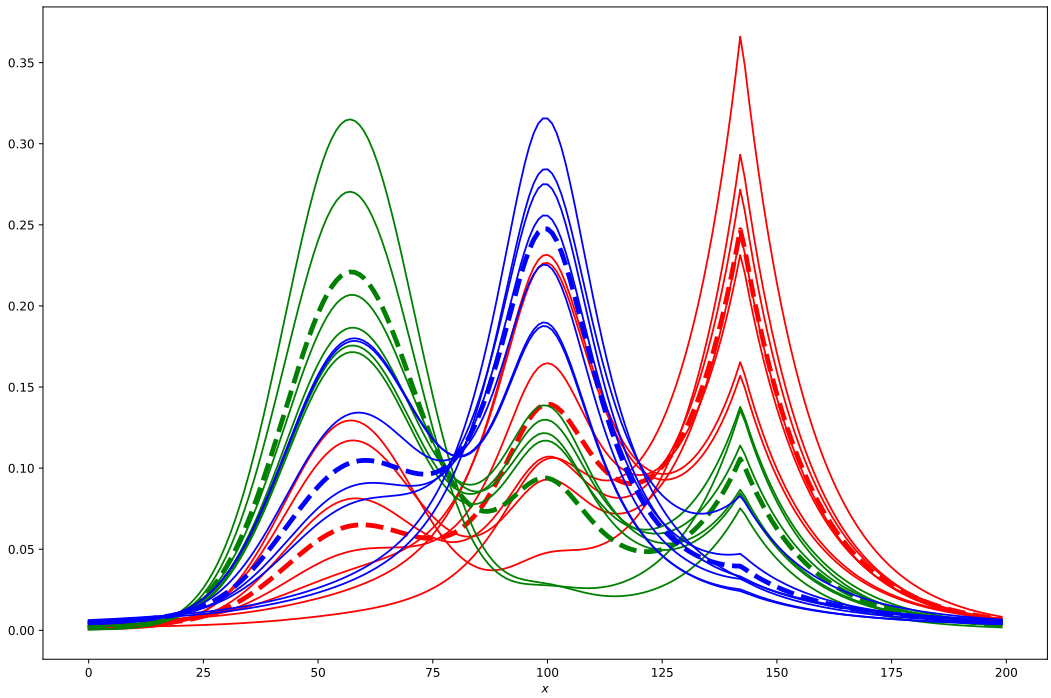

Figure 2: Mixture clustering with $k = 3, N = 16$ for the mixture family combining a Gaussian, a Cauchy, and a Laplace distribution defined over $\mathbb{R}$.