Deep Divergence Learning: Supplementary Material

Kubra Cilingir [1] **Rachel Manzelli** [1] **Brian Kulis** [1]

Appendices

A. Notation and Definitions

In this section, we introduce basic concepts from functional analysis and the notation used for extending vector spaces to function spaces, which will be used for our proofs.

A.1. Assumptions and definitions from functional analysis

We first present basic notation from functional analysis, since we extend vector spaces to function spaces to derive this formulation.

Assume we have a finite measure space (χ, Σ, μ) which is Lebesgue-measurable, and $\chi \in \mathbb{R}^d$. Note that we mainly consider a set of distributions in this paper, which is a special case that uses a Radon measure and a bounded Borel set. Consider a set of measurable functions $F \subseteq L^p$, defined as $F = \{f \in F \mid f : \chi \to R, ||f||_p \leq C_1 < \infty \text{ and } f \geq 0\}$, where C_1 is a constant and $1 \leq p \leq \infty$. The restriction that $f \geq 0$ is not limiting, since it can be easily satisfied by using its equivalence class obtained by only applying an affine transformation (Frigyik et al., 2008).

Assume $W \subseteq L^p$ is a compact set of functions. All bounded continuous linear functionals have an integral representation with respect to our focus of measure space (Gierz, 1987), with a corresponding function $w \in W,\ w : \chi \to \mathbb{R}$. Similarly, we can characterize affine functionals by their function and constant pairs $A = \{(w, b_w) \mid w \in W, b_w \in \mathbb{R} \text{ and } |b_w| \leq C_2\}$, with C_2 a constant.

For a convex functional ϕ, we denote its Fréchet derivative as $\delta\phi(p)$ and the epigraph of ϕ as $epi\ \phi$; with their definitions briefly given below (Gelfand et al., 2000) :

Fréchet derivative of ϕ. If for every $h \in W$, there exists $\delta\phi(f)$ s.t.

$$\lim_{||h||_p \to 0} \frac{\phi(f+h) - \phi(f) - \delta\phi(f)[h]}{||h||_p} = 0,$$

then $\phi(f)$ is Fréchet differentiable and $\delta\phi(f)$ is the Fréchet derivative of ϕ at f.

Directional Fréchet derivative of ϕ. The derivative of a functional ϕ at f in the direction of a function g is defined as:

$$\delta\phi[f; g] = \int \delta\phi(f)(x)g(x)dx.$$

Epigraph of ϕ. The epigraph of a functional ϕ is defined as:

$$epi\ \phi := \{(c, f) \in \mathbb{R} \times F \mid \phi(f) \leq c\}.$$

B. Proof of Theorem 3.1

Proof. To prove the result, we can generalize a known symmetry result for standard Bregman divergences seen in Bauschke & Borwein, Lemma 3.16 (Bauschke & Borwein, 2001), or this Mathematics Stack Exchange discussion[1].

We start by establishing that any symmetric functional Bregman divergence has the form given in the statement of the theorem. Let 0_f be the zero-function (given, for example by the function $p - p$ for any p). We can assume without loss of generality that $\phi(0_f) = 0$ and $\delta\phi(0_f) = 0$—we can always add a constant to ϕ to ensure the first property, and we can subtract $\int p(x)\delta\phi(0_f)dx$ from ϕ to ensure the second property, both without changing the resulting Bregman divergence.

Next, if $D_\phi(p, q) = D_\phi(q, p)$ for all p, q, then writing out the Bregman divergences and equating them yields

$$\begin{aligned}&\phi(p) - \phi(q) - \int (p(x) - q(x))\delta\phi(q)(x)dx \\ &= \phi(q) - \phi(p) - \int (q(x) - p(x))\delta\phi(p)(x)dx. \qquad (1)\end{aligned}$$

[1]Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts, USA. Correspondence to: Kubra Cilingir <kubra@bu.edu>, Rachel Manzelli <manzelli@bu.edu>, Brian Kulis <bkulis@bu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, Austria, PMLR 119, 2020.

[1]https://math.stackexchange.com/questions/2242980/bregman-divergence-symmetric-iff-function-is-quadratic

k	5	20	50	100	200	500	1000
acc	71.9	77.8	79.4	80.0	77.4	74.1	70.8

Table 1. Accuracy on Cifar10 when varying the number of subnetworks (k).

D.2. Additional K-nn Classification Details

Here we provide more details regarding our K-nn experiments between deep Bregman and Euclidean cases. All factors in our experimental settings are created by very standard choices for a fair comparison. The batches are chosen randomly from the relevant dataset, and then the pairs are created within that batch at each iteration. We use a validation set ratio of 20%. Once the training is complete, we obtain test embeddings and run the K-nn algorithm on these embeddings.

We choose k, the number of subnetworks, to be equal to the number of classes. Additionally, we run a small experiment over varying k from 5 to 1000 and reported the results in Table 1. The results indicate that performance improves to a point, and then the model starts to overfit. This suggests that an optimal k can be found by adding it as a hyperparameter in the experiments.

References

Bauschke, H. H. and Borwein, J. M. Joint and separate convexity of the Bregman distance. *Studies in Computational Mathematics*, 8:23–36, 2001.

Fréchet, M. Sur les ensembles de fonctions et les opérations linéaires. *CR Acad. Sci. Paris*, 144:1414–1416, 1907.

Frigyik, B. A., Srivastava, S., and Gupta, M. R. Functional Bregman divergences and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54 (11):5130–5139, 2008.

Gelfand, I. M., Silverman, R. A., et al. *Calculus of variations*. Courier Corporation, 2000.

Gierz, G. Integral representations of linear functionals on function modules. *The Rocky Mountain Journal of Mathematics*, pp. 545–554, 1987.