# 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 $(\chi, \Sigma, \mu)$ 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 $\phi$, we denote its Fréchet derivative as $\delta\phi(p)$ and the epigraph of $\phi$ as *epi* $\phi$; with their definitions briefly given below (Gelfand et al., 2000) :

**Fréchet derivative of** $\phi$. 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 $\phi$ at $f$.

**Directional Fréchet derivative of** $\phi$. The derivative of a functional $\phi$ 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** $\phi$. The epigraph of a functional $\phi$ 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 $\phi$ to ensure the first property, and we can subtract $\int p(x)\delta\phi(0_f)dx$ from $\phi$ 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

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

[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. Copyright 2020 by the author(s).

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

Letting $p = 0_f$ and simplifying the above equation (and using $\phi(0_f) = 0$ and $\delta\phi(0_f) = 0$), we obtain the following:

$$2\phi(q) = \int q(x)\delta\phi(q)(x)dx.$$

Note that this equation holds for any $q$. Plugging this equation (along with the same equation where $p$ has replaced $q$) into (1), we obtain the following identity:

$$\int p(x)\delta\phi(q)(x)dx = \int q(x)\delta\phi(p)(x)dx. \tag{2}$$

This can be used to establish that $\delta\phi$ is linear. For example, to establish that $\delta\phi$ is homogeneous, we must show that $\delta\phi(\alpha p) = \alpha\delta\phi(p)$, for non-zero $\alpha$. Using (2) twice (first and third line), we can establish the following for any $p$ and $q$:

$$\begin{aligned}\int q(z)\delta\phi(\alpha p)(z)dz &= \int \alpha p(z)\delta\phi(q)(z)dz \\ &= \alpha\int p(z)\delta\phi(q)(z)dz \\ &= \alpha\int q(z)\delta\phi(p)(z)dz.\end{aligned}$$

This can then be used to show that $\delta\phi(\alpha p) = \alpha\delta\phi(p)$: for any point $x$, suppose $p$ is a Dirac delta function at $x$. Then the above equation establishes that $\delta\phi(\alpha p)$ equals $\alpha\delta\phi(p)$ at $x$. Since the equation is true for all $p$, then $\delta\phi(\alpha p)$ equals $\alpha\delta\phi(p)$ for all points.

A similar argument can be used to establish that $\delta\phi(p+q) = \delta\phi(p) + \delta\phi(q)$. In particular, $\int r(z)\delta\phi(p+q)(z)dz$

$$\begin{aligned}&= \int (p(z)+q(z))\delta\phi(r)(z) \\ &= \int p(z)\delta\phi(r)(z) + \int q(z)\delta\phi(r)(z)dz \\ &= \int r(z)\delta\phi(p)(z)dz + \int r(z)\delta\phi(q)(z)dz\end{aligned}$$

for all $r$, establishes that $\delta\phi(p+q) = \delta\phi(p) + \delta\phi(q)$ and choosing $r$ as Dirac delta functions ensures this equality for all points.

In the case of functions, if a gradient function $\delta\phi$ is linear, then the function $\phi$ must be quadratic; this is because we take an anti-derivative of a linear function and obtain a quadratic function. In the functional case, this means that $\phi$ must have the following form:

$$\phi(p) = \iint p(x)p(y)\psi(x,y)dxdy,$$

where $\psi$ is a symmetric, positive semi-definite function. (In the vector setting, $\phi(x) = \boldsymbol{x}^T A\boldsymbol{x}$ for a positive semi-definite matrix $A$, so this is a generalization to the functional setting.) One can verify that the gradient $\delta\phi$ is of the form

$$\delta\phi(p)(y) = 2\int p(x)\psi(x,y)dx,$$

which is indeed a linear function. Given this form for $\phi$, the final step is to plug $\phi$ into the definition for the functional divergence and to simplify the resulting divergence. After simplification using the definition of $\phi$ and its derivative, along with the fact that $\psi(x,y) = \psi(y,x)$, we obtain

$$D_\phi(p,q) = \iint (p(x)-q(x))(p(y)-q(y))\psi(x,y)dxdy.$$

Now that we have established one direction of the theorem, we can establish the other. This direction is considerably simpler. We must show that a divergence that has the form

$$D_\phi(p,q) = \iint (p(x)-q(x))(p(y)-q(y))\psi(x,y)dxdy$$

is in fact a symmetric functional Bregman divergence. The fact that it is symmetric follows directly. The fact that it is a functional Bregman divergence follows from the fact that choosing the strictly convex functional $\phi(p) = \iint p(x)p(y)\psi(x,y)dxdy$ yields the resulting divergence.

□

## C. Proof of Theorem 3.2

In this section, we show that our convex generating functional form is justified in that any convex functional can be represented as a supremum over linear functionals.

Up to this point, we notated convex functionals as $\phi(p)$, in terms of only their input functions. Here we will use the notation $\phi(x; p(x))$ for convex functionals, where $x$ refers to the input of the function $p$.

*Proof.* ($\supseteq$) We first show that the right hand side is indeed a convex functional.

We will use the standard definition of convexity since it directly extends to the functional case. The domain of the functionals is a convex subset since for all $\lambda \in [0,1]$, and $p, q \in L^p$, $||\lambda p + (1-\lambda q)||_p < \infty$, so $\lambda p + (1-\lambda)q \in L^p$ naturally.

For an arbitrary pair $(w, b_w)$, and $p, q \in F$ we have:

$$\begin{aligned}&\int(\lambda p(x) + (1-\lambda)q(x))w(x)dx + b_w \leq \qquad (3)\\ &\lambda\Big[\int p(x)w(x)dx + b_w\Big] + (1-\lambda)\Big[\int q(x)w(x)dx + b_w\Big] \\ &= \lambda\phi_p(x;p(x)) + (1-\lambda)\phi_q(x;q(x)) \\ &\leq \lambda\phi^*(x;p(x)) + (1-\lambda)\phi^*(x;q(x)),\end{aligned}$$

where $\phi^*$ represents the supremum attained for the right-hand side of (3), and $\phi_a(b) = \int b(x)a(x)dx + b_a$ . Since the inequalities hold for all $(w, b_w)$, we can take the sup of the first line and obtain:

$$\phi^*(\lambda p + (1-\lambda)q(x)) \leq \lambda\phi^*(x;p(x)) \\ + (1-\lambda)\phi^*(x;q(x)).$$

$(\subseteq)$ We now show that for a given convex functional $\phi(x;p(x))$, we can find a set of affine functionals to write it as (3).

Assume $\delta\phi(x;p(x))$ is the Frechet derivative of $\phi$ at function $p \in W$. Then $\delta\phi(x;p(x))$ is a linear operator. Since $\phi$ is continuous and bounded, we can find $r(x) = \arg\inf_f \int \delta\phi(x;p(x))f(x)dx$[2]. Define $\delta^{'}_{\phi}(x;p(x)) := \delta\phi(x;p(x)) + \phi(x;r(x)) \geq 0$. This positive functional can be represented in an integral form by Riesz-Markov-Kakutani representation theorem on the measure $d(\delta^{'}_{\phi}(x))$ (Fréchet, 1907). Note that we can always add or substract properly scaled constant terms and preserve the information since these transformations are linear. For a given $p'$, applying Riesz theorem gives us the representation below:

$$\delta\phi(x;p'(x)) = \int \delta\phi'(x;p'(x))p'(x)dx,$$

with a support function

$$l_{\phi_{p'}}(x;p(x)) = \int \delta\phi'(x;p'(x))p(x)dx \\ + \phi(x;p'(x)) - \delta\phi(x;p'(x)).$$

We also have $l_{\phi_{p'}}(x;p(x)) \leq \textit{epi}\ \phi(x;p(x))$ for all $p, p' \in W$, since $\phi$ is a convex functional. $l_{p'}(x;p(x)) \leq \phi(x;p(x))$ for all $p$ since we are on a compact domain and $\phi$ is continuous and convex. Now for a given convex functional $\phi(x;p(x))$, consider $\bigcup_{p' \in W} l_{\phi_{p'}}(x;p(x))$ as a set of affine functionals, further denoted by $L_\phi$ for convenience. Define:

$$\psi(x;p(x)) = \sup_{l \in L_\phi} \int l(x)p(x)dx + b_l.$$

$\psi$ is a convex functional by the first part of the proof. Since $\phi$ is convex, for all $p \neq p'$ we have $\phi(x;p^{'}(x)) - l_\phi(x;p^{'}(x)) \geq 0$, so for an arbitrary $p$, $\psi(x;p(x)) = l_\phi(x;p(x))$. This concludes $\psi(x;p(x))$ forms a set of functionals to construct $\phi$. □

Note that if we restrict our space to a set of $n$ distributions, then all we need to know is at most $n$ corresponding maximizing affine functionals; in this case the supremum can be replaced by the maximum as we did in the paper.

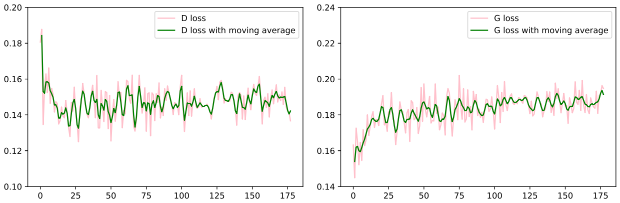

*Figure 1.* The discriminator (left) and generator (right) losses during training for CelebA. Each epoch is split into 25 averaged batch losses. The window size is 2 for the moving average.

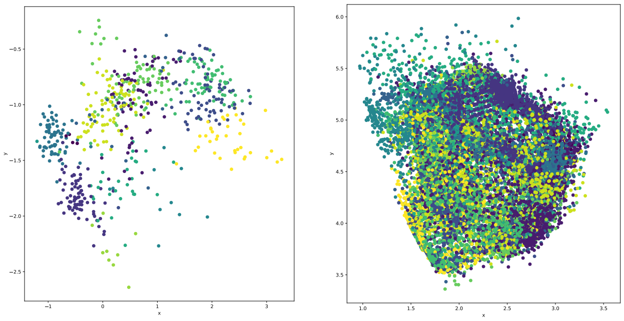

*Figure 2.* (Left) Embedding of the WHARF dataset learned by our method using contrastive loss with a moment-matching function. (Right) Embedding learned by the baseline deep learning approach.

# D. Applications Details

## D.1. Additional GAN Model Details

In this section, we present more training details related to our GAN model. We use RMSprop optimizer with a momentum value of 0.99, and set the learning rates to $10^{-3}$ for the discriminator and $3 \times 10^{-3}$ for the generator. The minibatch size is chosen as 64. Our main model has convolutional layers with kernel sizes equal to 5 and filter sizes equal to 64. The strides are halved towards the final layers. Stride sizes are determined based on the input image dimensions.

In our experiments, we incorporate contrastive loss into deep Bregman learning in order to supervise the discriminator. However, our distributional loss formula has the potential to be directly used in the GAN setting, which we leave as a future work.

We provide the loss plots for the generator and the discriminator through the training phase below in Figure 1. We observe that the discriminator first learns the metric, then the training preserves the balance between the two networks. We note that image quality still improves for a while after the losses become saturated, due to the nature of contrastive loss.

[2] $r(x)$ also can be constructed from the $\epsilon$-subdifferentials of $\phi$ to ensure existence.

| $k$ | 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.