BreGMN: scaled-Bregman Generative Modeling Networks

Akash Srivastava
MIT-IBM Watson AI Lab
IBM Research
Cambridge, MA
akash.srivastava@ibm.com

Kristjan Greenewald
MIT-IBM Watson AI Lab
IBM Research
Cambridge, MA
kristjan.h.greenewald@ibm.com

Farzaneh Mirzazadeh
MIT-IBM Watson AI Lab
IBM Research
Cambridge, MA
farzaneh@ibm.com

Abstract

The family of f-divergences is ubiquitously applied to generative modeling in order to adapt the distribution of the model to that of the data. Well-definedness of f-divergences, however, requires the distributions of the data and model to overlap completely in every time step of training. As a result, as soon as the support of distributions of data and model contain non-overlapping portions, gradient-based training of the corresponding model becomes hopeless. Recent advances in generative modeling are full of remedies for handling this support mismatch problem: key ideas include either modifying the objective function to integral probability measures (IPMs) that are well-behaved even on disjoint probabilities, or optimizing a well-behaved variational lower bound instead of the true objective. We, on the other hand, establish that a complete change of the objective function is unnecessary, and instead an augmentation of the *base measure* of the problematic divergence can resolve the issue. Based on this observation, we propose a generative model which leverages the class of *Scaled Bregman Divergences* and generalizes both f-divergences and Bregman divergences. We analyze this class of divergences and show that with the appropriate choice of base measure it can resolve the support mismatch problem and incorporate geometric information. Finally, we study the performance of the proposed method and demonstrate promising results on MNIST, CelebA and CIFAR-10 datasets.

1 Introduction

Modern deep generative modeling paradigms offer a powerful approach for learning data distributions. Pioneering models in this family such as generative adversarial networks (GANs) (Goodfellow et al., 2014) and variational autoencoders (VAEs) (Kingma & Welling, 2014) propose elegant solutions to generate high quality photo-realistic images, which were later evolved to generate other modalities of data. Much of the success of attaining photo-realism in generated images is attributed to the *adversarial* nature of training in GANs. Essentially, GANs are neural samplers in which a deep neural network G_ϕ is trained to generate high dimensional samples from some low dimensional noise input. During the training, the generator is pitched against a classifier: the classifier is trained to distinguish the generated from the true data samples and the generator is simultaneously trained to generate samples that look like true data. Upon successful training, the classifier fails to distinguish

Preprint. Under review.

between the generated and actual samples. Unlike VAE, GAN is an implicit generative model since its likelihood function is implicitly defined and is in general intractable. Therefore training and inference are carried out using likelihood-free techniques such as the one described above.

In its original formulation, GANs can be shown to approximately minimize an f-divergence measure between the true data distribution p_x and the distribution q_ϕ induced by its generator G_ϕ. The difficulty in training the generator using the f-divergence criterion is that the supports of data and model distributions need to perfectly match. If at any time in the training phase, the supports have non-overlapping portions, the divergence either maxes out or becomes undefined. If the divergence or its gradient cannot be evaluated, it cannot, in turn, direct the weights of model towards matching distributions (Arjovsky et al., 2017) and training fails.

In this work, we present a novel method, BreGMN, for implicit adversarial and non-adversarial generative models that is based on *scaled Bregman divergences* (Stummer & Vajda, 2012) and does not suffer from the aforementioned problem of support mismatch. Unlike f-divergences, scaled Bregman divergences can be defined with respect to a base measure such that they stay well-defined even when the data and the model distributions do not have matching support. Such an observation leads to a key contribution of our work, which is to identify base measures that can play such a useful role. We find that measures whose support include the supports of data and model are the ones applicable. In particular, we leverage Gaussian distributions to augment distributions of data and model into a base measure that guarantees the desired behavior. Finally we propose training algorithms for both adversarial and non-adversarial versions of the proposed model.

The proposed method facilitates a steady decrease of the objective function and hence progress of training. We empirically evaluate the advantage of the proposed model for generation of synthetic and real image data. First, we study simulated data in a simple 2D setting with mismatched supports and show the advantage of our method in terms of convergence. Further, we evaluate BreGMN when used to train both adversarial and non-adversarial generative models. For this purpose, we provide illustrative results on the MNIST, CIFAR10, and CelebA datasets, that show comparable performance to the sample quality of the state-of-art methods. In particular, our quantitative results on generative real datasets also demonstrate the effectiveness of the proposed method in terms of sample quality.

The remainder of this document is organized as follows. Section 2 outlines related work. We introduce the scaled Bregman divergence in Section 3, demonstrate how it generalizes a wide variety of popular discrepancy measures, and show that with the right choice of base measure it can eliminate the support mismatch issue. Our application of the scaled Bregman divergence to generative modeling networks is described in Section 4, with empirical results presented in Section 5. Section 6 concludes the paper.

2 Related work

Since the genesis of adversarial generative modeling, there has been a flurry of work in this domain, e.g. (Nowozin et al., 2016; Srivastava et al., 2017; Li et al., 2017; Arjovsky et al., 2017) covering both practical and theoretical challenges in the field. Within this, a line of research addresses the serious problem of *support mismatch* that makes training hopeless if not remedied. One proposed way to alleviate this problem and stabilize training is to match the distributions of the data and the model based on a different, well-behaved discrepancy measure that can be evaluated even if the distributions are not equally supported. Examples of this approach include Wasserstein GANs (Arjovsky et al., 2017) that replace the f-divergence with Wasserstein distance between distributions and other integral probability metric (IPM) based methods such as MMD GANs (Li et al., 2017), Fisher GAN (Mroueh & Sercu, 2017), etc. While IPM based methods are better behaved with respect to the non-overlapping support issue, they have their own issues. For example, MMD-GAN requires several additional penalties such as feasible set reduction in order to successfully train the generator. Similarly, WGAN requires some ad-hoc method for ensuring the Lipschitz constraint on the critic via gradient clipping, etc. Another approach to remedy the support mismatch issue comes from Nowozin et al. (2016). They showed how GANs can be trained by optimizing a variational lowerbound to the actual f-divergence the original GAN formulation proposed. They also showed how the original GAN loss minimizes a Jenson-Shannon divergence and how it can be modified to train the generator using a f-divergence of choice.

In parallel, works such as (Amari & Cichocki, 2010) have studied the relation between the many different divergences available in the literature. An important extension to *Bregman* divergences, namely *scaled Bregman divergences*, was proposed in the works of Stummer & Vajda (2012); Kißlinger & Stummer (2013) and generalizes both f-divergences and Bregman divergences. The Bregman divergence in its various forms has long been used as the objective function for *training* machine learning models. Supervised learning based on least squares (a Bregman divergence) is perhaps the earliest example. Helmbold et al. (1995); Auer et al. (1996); Kivinen & Warmuth (1998) study the use of Bregman divergences as the objective function for training single-layer neural networks for univariate and multivariate regression, along with elegant methods for matching the Bregman divergence with the network's nonlinear transfer function via the so-called *matching loss* construct. In unsupervised learning, Bregman divergences are unified as the objective for clustering in Banerjee et al. (2005), while convex relaxations of Bregman clustering models are proposed in Cheng et al. (2013). Generative modeling based on Bregman divergences is explored in Uehara et al. (2016a,b), which relies on a duality relationship between Bregman and f divergences. These works retain the f-divergence based f-GAN objective, but use a Bregman divergence as a distance measure for estimating the needed density ratios in the f-divergence estimator. This contrasts with our approach which uses the scaled Bregman divergence as the overall training objective itself.

3 Generative modeling via discrepancy measures

The choice of distance measure between the data and the model distribution is critical, as the success of the training procedure largely depends on the ability of these distance measures to provide meaningful gradients to the optimizer. Common choices for distances include the Jensen-Shannon divergence (vanilla GAN) f-divergence (f-GAN) (Nowozin et al., 2016) and various integral probability metrics (IPM, e.g. in Wasserstein-GAN, MMD-GAN) (Arjovsky et al., 2017; Li et al., 2017). In this section, we consider a generalization of the Bregman divergence that also subsumes the Jensen-Shannon and f-divergences as special cases, and can be shown to incorporate some geometric information in a way analogous to IPMs.

3.1 Scaled Bregman divergence

The **Bregman divergence** (Bregman, 1967) forms a measure of distance between two vectors $p, q \in \mathbb{R}^d$ using a convex function $F : \mathbb{R}^d \to \mathbb{R}$ as

$$B_F(p, q) = F(p) - F(q) - \nabla F(q) \cdot (p - q),$$

which includes a variety of distances, such as the squared Euclidean distance and the KL divergence between finite-cardinality probability mass functions, as special cases.

More useful in our setting is the class of **separable** Bregman divergences of the form

$$B_f(P, Q) = \int_{\mathcal{X}} f(p(x)) - f(q(x)) - f'(q(x))(p(x) - q(x))dx \tag{1}$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ is a convex function, f' is its right derivative and P and Q are measures on $\mathcal{X}$ with densities p and q respectively. In this form the divergence is a discrepancy measure for distributions as desired. In general, as the name divergence implies, the quantity is non-symmetric. It does not satisfy the triangle inequality either (Acharyya et al., 2013).

While this is a valid discrepancy measure, the Bregman divergence does not yield meaningful gradients for training when the two distributions in question have non-overlapping portions in their support, similar to the case of f-divergences (Arjovsky & Bottou, 2017). We thus propose to use the *scaled* Bregman divergence, which introduces a third measure M with density m that can depend on P and Q and uses it as a base measure for the Bregman divergence. Specifically, the **scaled Bregman divergence** (Stummer & Vajda, 2012) is given by

$$B_f(P, Q|M) = \int_{\mathcal{X}} f\left(\frac{p(x)}{m(x)}\right) - f\left(\frac{q(x)}{m(x)}\right) - f'\left(\frac{q(x)}{m(x)}\right)\left(\frac{p(x)}{m(x)} - \frac{q(x)}{m(x)}\right) dM. \tag{2}$$

This expression is equal to the separable Bregman divergence (1) when M is equal to the Lebesgue measure.

As shown in (Stummer & Vajda, 2012), the scaled Bregman divergence (2) contains many popular discrepancy measures as special cases. In particular, when $f(t) = t \log t$ it reduces to the **KL divergence** for any choice of M (as does the vanilla Bregman divergence).

Many classical criteria (including the KL and Jensen-Shannon divergences) belong to the family of f**-divergences**, defined as

$$D_f(P, Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx.$$

where the function $f : R_+ \to R$ is a convex, lower-semi-continuous function satisfying $f(1) = 0$, where the densities p and q are absolutely continuous with respect to each other. The scaled Bregman divergence with choice of $M = Q$ reduces to the f divergence family as:

$$B_f(P, Q|Q) = \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) - f'(1)\left(\frac{p(x)}{q(x)} - 1\right) dQ = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx,$$

which shows all f-divergences are special cases of the scaled Bregman divergence. A more complete list of discrepancy measures included in the class of scaled Bregman divergences is found in Stummer & Vajda (2012).

3.2 Noisy base measures and support mismatch

A widely-known weakness of f-divergence measures is that when the supports of p and q are disjoint, the value of the divergence is trivial or undefined. In the context of generative models, this issue is often tackled by adding noise to the model distribution which extends its support over the entire observed space such as in VAEs. However, adding noise to the observed space is not particularly well-suited for tasks such as image generation as it results in blurry images. In this work we propose choosing a base measure M that in some sense incorporates geometric information in such a way that the gradients in the disjoint setting become informative without compromising the image quality.

For the scaled Bregman $B_f(P, Q|M)$, we propose choosing a "noisy" base measure M, specifically one that is formed by convolving some other measure with the Gaussian measure $\mathcal{N}(0, \Sigma)$. Recall that convolution of two distributions corresponds to the addition of the associated random variables, hence in this case we are in affect adding Gaussian noise to the variable generated by M. In addition to adding noise, we require a base measure $\tilde{M}$ that depends on P and Q to avoid the vanilla Bregman divergence's lack of informative gradients (see Section 4.1 below). By analogy to the Jensen-Shannon divergence, we choose

$$\tilde{M} = \alpha(P * \mathcal{N}(0, \Sigma_1)) + (1 - \alpha)(Q * \mathcal{N}(0, \Sigma_2)) \tag{3}$$

for $0 \leq \alpha \leq 1$ and some covariances Σ_1 and Σ_2, where $*$ denotes the convolution of two distributions. Denote the density of $\tilde{M}$ as $\tilde{m}$.

Importantly, observe that each term of the corresponding scaled Bregman $B_f(P, Q|\tilde{M})$ is always well defined and finite (with the exception of certain choices of f such as $-\log$ that require numerical stabilization similar to the case of f-divergence) since $\tilde{M}$ has full support. Furthermore, since $\tilde{M}$ is a noisy copy of $\alpha P + (1 - \alpha)Q$, the ratio $\frac{p}{\tilde{m}}$ will be affected by q even outside the support of q, and vice versa. This ensures that a training signal remains in the support mismatch case.

The presence of this training signal seems to indicate that geometric information is being used, since it varies with the distance between the supports. To further explore this intuitive connection between noisy base measures and geometric information, we attempt to relate $B_f(P, Q|\tilde{M})$ to the W_p distance. In what follows, for simplicity we focus on the case of $f(t) = t \log t$; analysis for more general choices of f is left for future work. For the KL divergence for example, Pinsker's inequality states that

$$D_{KL}(p||q) \geq 2(W_1(p, q))^2.$$

A similar lower bound for the W_2 distance and certain log-concave q follows from Talagrand's inequality (Bobkov & Ledoux, 2000). These lower bounds are not surprising, since the KL divergence can go to infinity when Wasserstein-p is finite. However, lower bounds of this type are not sufficient to imply that a divergence is using geometric information, since it can increase very quickly while W_p increases only slightly.

Our use of a noisy M_0, however, allows us to obtain an upper bound for a symmetrized version of $B_f(P, Q|\tilde{M})$, which implies a continuity with respect to geometric information. While we found in our generative modeling experiments that a symmetrized version is unnecessary to use in practice, it is useful for comparison to IPMs. Recall that the Jensen-Shannon divergence constructs a symmetric measure by symmetrizing the KL divergence around $(P + Q)/2$. Any Bregman divergence can be similarly symmetrized (Eq. 16 in Nielsen & Nock (2011)). For simplicity, we consider the special case of $\tilde{M}$, namely $M_0 = \frac{P+Q}{2} * \mathcal{N}_\sigma$ with density m_0, and use it to both scale and symmetrize the scaled Bregman divergence, obtaining the measure $B_f(P, M_0|M_0) + B_f(Q, M_0|M_0) = D_f(P||M_0) + D_f(Q||M_0)$. In Section A of the Supplement we prove:

Proposition 1. *Assume that $\mathbb{E}_{U\sim P}\|U\|$ and $\mathbb{E}_{V\sim P}\|V\|$ are bounded. Then*

$$|B_{t\log t}(P, M_0|M_0) - B_{t\log t}(Q, M_0|M_0)| \leq cW_2(P, Q) + |h(Q) - h(P)|,$$

where c is a constant given in the proof and $h(P)$ is the Shannon entropy of P.

While an $h(P) - h(Q)$ term remains, it is simple to rescale Q to match the entropy of P, eliminating that term and leaving the Wasserstein distance.[1]

While not fully characterizing the geometric information in $B_f(P, Q|M_0)$, these observations seem to imply that the use of the noisy $\tilde{M}$ is capable of incorporating some geometric information without having to resort to IPMs with their associated training difficulties in the GAN context such as gradient clipping and feasible set reduction (Arjovsky & Bottou, 2017; Li et al., 2017).

4 Model

Let $\{x_i|x_i \in \mathbb{R}^d\}_{i=1}^N$ be a set of N samples drawn from the data generating distribution p_x that we are interested in learning through a parametric model G_ϕ. The goal of generative modeling is to train G_ϕ, generally implemented as a deep neural network, to map samples from a k-dimensional easy-to-sample distribution to the ambient d dimensional data space, i.e. $G_\phi : \mathbb{R}^k \mapsto \mathbb{R}^d$. Letting q_ϕ be the distribution induced by the generator function G_ϕ, almost all training criteria are of the form

$$\min_\phi D(p_x\|q_\phi) \tag{4}$$

where $\mathbb{D}(.\|.)$ is a measure of discrepancy between the data and the model distributions. We propose to use the scaled-Bregman divergence as D in Equation (4). We will show that unlike f-divergences, scaled-Bregman divergences can be easily estimated with respect to a base measure using only samples from the distributions. This is important when we aim to match distributions in very high dimensional spaces where they may not have any overlapping support (Arjovsky et al., 2017).

In order to compute the divergence between data and model distributions, it is not required that both densities are known or can be evaluated on realizations from distributions. Instead, being able to evaluate the ratio between them, i.e. *density ratio estimation*, is typically all that is needed. For example, generative models based on f-divergences only need density ratio estimation. Importantly, similar to the case of f-divergences, scaled-Bregman divergence estimation requires estimates of the density ratios only.

Below, we describe two methods of density ratio estimation (DRE) between two distributions. In what follows, suppose $r = \frac{p_x}{q_\phi}$ is the density ratio.

Discriminator-based DRE: This family of models uses a discriminator to estimate the density ratio. Let $y = 1$ if $x \sim p_x$ and $y = 0$ if $x \sim q_\phi$. Further, let $\sigma(C(x)) = p(y = 1|x)$, namely the discriminator, be a trained binary classifier on samples from p_x and q_ϕ where σ is the Sigmoid function. It is then easy to show that $C(x) = -\log \frac{p_x(x)}{q_\phi(x)} = -\log r(x)$ (Sugiyama et al., 2012), so C is a function of density ratio $r(x)$. In fact, this is the underlying principle in adversarial generative models (Goodfellow et al., 2014). As such, most discriminator-based DREs result in adversarial training procedures when used in generative models.

[1] Under certain smoothness conditions on P and Q $|h(P) - h(Q)|$ can itself be upper bounded by the Wasserstein distance (see Polyanskiy & Wu (2016) for details).

Algorithm 1: Training Algorithm of BreGMN

while *not converged* **do**
  Step 1 Estimate the density ratios $r_{p/m}$ and $r_{q_\phi/m}$, using either the adversarial discriminator-based (GAN-like) or
  non-adversarial two-sample-test-based (MMD-like) method.
  Step 2 Train the generator by optimizing
  $$\min_\phi \hat{B}_f(p_x, q_\phi|\tilde{M}).$$
end

MMD-based DRE: This family of models estimate the density ratio without the use of a discriminator. In order to estimate the density ratio r without training a classifier, thereby avoiding adversarial training of the generator later, we can employ the maximum mean discrepancy (MMD) (Gretton et al., 2012) criterion as in Sugiyama et al. (2012). By solving for r in the RKHS in

$$\min_{r\in\mathcal{H}} \left\| \int k(x;.)p_x(x)dx - \int k(x;.)r(x)q_\phi(x)dx \right\|_{\mathcal{H}}^2, \tag{5}$$

where k is a kernel function, we obtain a closed form estimator of the density ratio as

$$\hat{\mathbf{r}}_{p/q} = K_{q,q}^{-1} K_{q,p}\mathbf{1}. \tag{6}$$

Here $K_{q,q}$ and $K_{q,p}$ denote the Gram matrices corresponding to kernel k.

4.1 Empirical estimation

Using the DRE estimators introduced above we create empirical estimators of the scaled-Bregman divergence (2) as

$$\begin{aligned}\hat{B}_f(p_x, q_\phi|M) = \frac{1}{N}\sum_{i=1}^{N} & f\left(r_{p/m}(x_i)\right) - f\left(r_{q_\phi/m}(x_i)\right) \\ & -f'\left(r_{q_\phi/m}(x_i)\right)\left(r_{p/m}(x_i) - r_{q_\phi/m}(x_i)\right)\end{aligned} \tag{7}$$

where $r_{p/m}$ denotes a DRE of p/m and the x_i are i.i.d. samples from the base distribution m with measure M. Note that this empirical estimator $\hat{B}_f$ does not have gradients with respect to ϕ if we only evaluate the DRE estimators on samples from the base measure m. Choices of M that depend on p and q, however, including our choice of $\tilde{M}$ (3) as well as the choice $M = Q$ (f-divergences), have informative gradients, allowing us to train the generator.

4.2 Training

Training the generator function G_ϕ using scaled-Bregman divergence (shown in Algorithm 1) alternates the following two steps until convergence.

Step 1: Estimate the density ratios $r_{p/m}$ and $r_{q_\phi/m}$ using either the adversarial discriminator-based method (as in a GAN) or the non-adversarial MMD-based method.

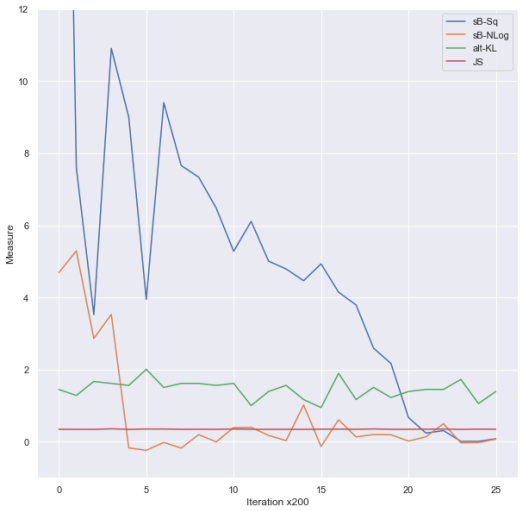
Step 2: Train the generator by optimizing

$$\min_\phi \hat{B}_f(p_x, q_\phi|\tilde{M}). \tag{8}$$

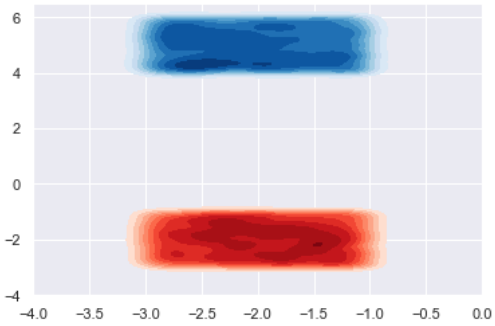
5 Experiments

In this section we present a detailed evaluation of our proposed scaled-Bregman divergence based method for training generative models. Since most generative models aim to learn the data generating distribution, our method can be generically applied to a large number of simple or complex and deep generative models for training. We demonstrate this by training a range of simple to complex models with our method.

Figure 1: f-divergence and scaled-Bregman divergence based training on synthetic dataset of two disjoint, non-overlapping 2D distributions.



(a) Measure vs Training



(b) 2D distributions.

5.1 Synthetic data: support mismatch

In this experiment, we evaluate our method in the regime where p and q_ϕ have mismatched support, in order to validate the intuition that the noisy base measure $\tilde{M}$ aids learning in this setting. As shown in Figure 1(b), we start by training a simple probabilistic model (blue) to match the data distribution (red). The data distribution is a simple uniform distribution with finite support. Our model is a therefore parameterized as a uniform distribution with one trainable parameter.

Figure 1(a) shows the effect of training this model with f-divergence and with our method. Clearly, neither the KL nor JS divergences are able to provide any meaningful gradients for the training of this simple model. Our scaled-Bregman based training method, however, is indeed able to learn the model. Interestingly, as Figure 1 shows, the choice of the function f matters in the empirical convergence rate of our method, with the convergence of $f(t) = -\log t$ much faster than that of $f(t) = t^2$.

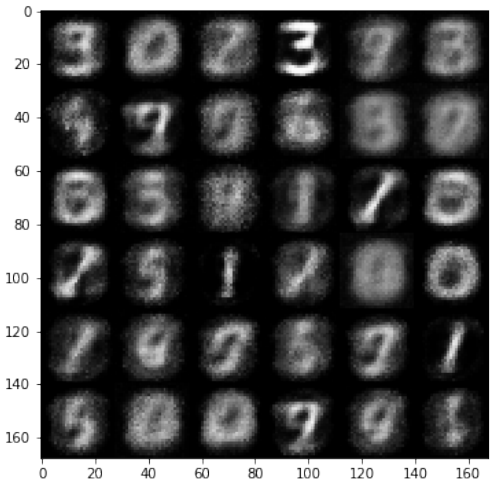
5.2 Non-adversarial generative model

Our training procedure is not intrinsically adversarial, i.e. it is not a saddle-point problem when the MMD-based DRE is used. To demonstrate the capability of the proposed model in training non-adversarial models, in this section, we apply the MMD-based DRE to train a generative model on the MNIST dataset in a non-adversarial fashion. As shown in Figure 3(a), our method can be used to successfully train generative models of a simple dataset without using adversarial techniques. While the sample quality is not optimal (better sample quality may be achievable by carefully tuning the kernel in the MMD criterion), the training procedure is remarkably stable as shown in Figure 3(b).

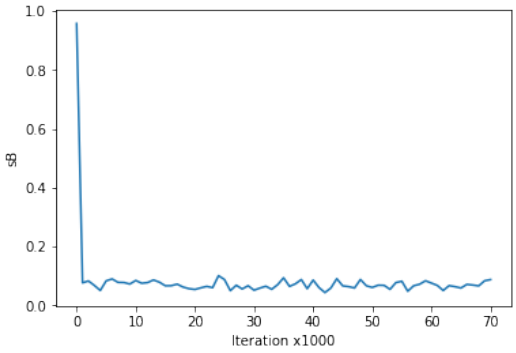
5.3 Adversarial generative model

Training generative models on complicated high-dimensional datasets such as those of natural images is done preferably with adversarial techniques since they tend to lead to better sample quality. One straightforward way to assign adversarial advantage to our method is to use a discriminator based DRE. To evaluate our training method on adversarial generation, in this section, we compare the Frechet Inception Distance (FID) (Heusel et al., 2017) of MMD-GAN (Li et al., 2017), GAN (Goodfellow et al., 2014) against BreGMN on CIFAR10 and CelebA dataset. FID measures the distance between the data and the model distributions by embedding their samples into a certain higher layer of a pre-trained Inception Net. We used a 4-layer DCGAN (Radford et al., 2015) architecture for all the experiments and averaged the FID over multiple runs. $\mathcal{N}(0, 0.001)$ is used as the noise level across all the experiments. MMD-GAN trains a generator network using the maximum mean discrepency

Figure 2: Non-adversarial Training using scaled-Bregman Divergence and MMD based DRE.

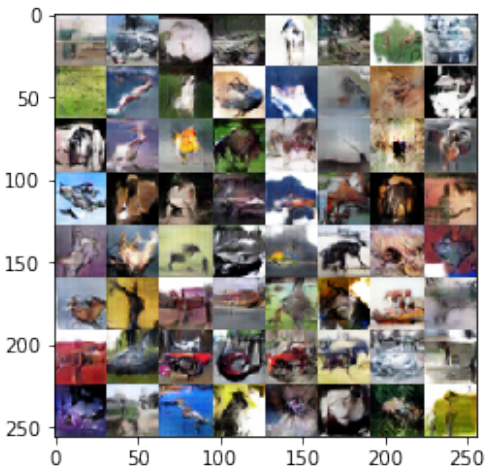


(a) Samples from the generator.

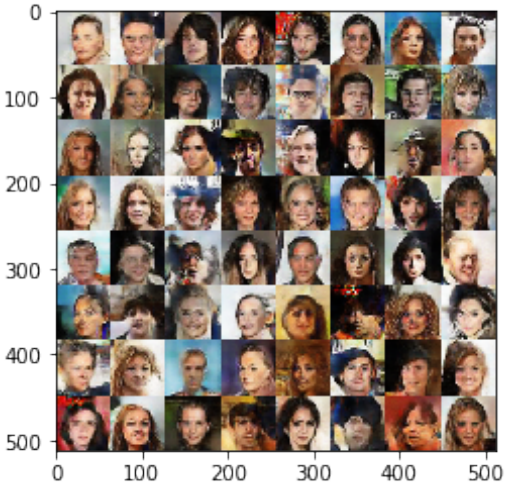


(b) Generator loss steadily decreases.

Figure 3: Random samples from Adversarial BreGMN models (after 5 Epochs)



(a) CIFAR10



(b) CELEB A

Table 1: Sample quality (measured by FID; lower is better) of BreGMN compared to GANs.

Archtitecture	Dataset	MMD-GAN	GAN	BreGMN
DCGAN	**Cifar10**	40	26.82	**26.62**
DCGAN	**CelebA**	41.10	30.97	**30.84**

(Gretton et al., 2012) where the kernel is trained in an adversarial fashion. As shown in Table 1, both BreGMN and GANs performs better than MMD-GAN in terms of sample quality. While BreGMN performs slightly better than GAN on average, their sample qualities are comparable.

6 Conclusions

In this work, we proposed scaled-Bregman divergence based generative models and identified base measures for them to facilitate effective training. We showed that the proposed approach provides a certifiably advantageous criterion to model the data distribution using deep generative networks in comparison to the f-divergence based training methods. We clearly established that unlike f-divergence based training our method does not fail to train even when the model and the data distributions do not have any overlapping support to start with. A future direction of research addresses the choice of the base measure and the effect of noise level on the optimization. Another, more theoretical direction is to study and establish the relationship between scaled-Bregman divergence and other IPMs.

References

Acharyya, S., Banerjee, A., and Boley, D. Bregman divergences and triangle inequality. In *SIAM International Conference on Data Mining*. SIAM, 2013.

Amari, S.-i. and Cichocki, A. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010.

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations, ICLR*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Auer, P., Herbster, M., and Warmuth, M. K. Exponentially many local minima for single neurons. In *Neural Information Processing Systems*, 1996.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

Bobkov, S. and Ledoux, M. From Brunn-Minkowski to Brascamp-Lieb and to logarithmic Sobolev inequalities. *Geometric & Functional Analysis GAFA*, 10(5):1028–1052, 2000.

Bregman, L. M. 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.

Cheng, H., Zhang, X., and Schuurmans, D. Convex relaxations of Bregman divergence clustering. In *Uncertainty in Artificial Intelligence, UAI*, 2013.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *Neural Information Processing Systems*, 2014.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Helmbold, D. P., Kivinen, J., and Warmuth, M. K. Worst-case loss bounds for single neurons. In *Neural Information Processing Systems*. MIT Press, 1995.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANS trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, 2017.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR*, 2014.

Kißlinger, A. and Stummer, W. Some decision procedures based on scaled Bregman distance surfaces. In *Geometric Science of Information - First International Conference, GSI 2013, Paris, France, August 28-30, 2013. Proceedings*, 2013.

Kivinen, J. and Warmuth, M. K. Relative loss bounds for multidimensional regression problems. In *Neural Information Processing Systems*, 1998.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. MMD GAN: Towards deeper understanding of moment matching network. In *Neural Information Processing Systems*, 2017.

Mroueh, Y. and Sercu, T. Fisher gan. In *Advances in Neural Information Processing Systems*, pp. 2513–2523, 2017.

Nielsen, F. and Nock, R. Skew Jensen-Bregman Voronoi diagrams. In *Transactions on Computational Science XIV*, pp. 102–128. Springer, 2011.

Nowozin, S., Cseke, B., and Tomioka, R. f-GAN: Training generative neural samplers using variational divergence minimization. In *Neural Information Processing Systems*, 2016.

Polyanskiy, Y. and Wu, Y. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Trans. Information Theory*, 62(7):3992–4002, 2016.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., and Sutton, C. A. VEEGAN: reducing mode collapse in GANs using implicit variational learning. In *Neural Information Processing Systems*, 2017.

Stummer, W. and Vajda, I. On Bregman distances and divergences of probability measures. *IEEE Trans. Information Theory*, 58(3):1277–1288, 2012.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Uehara, M., Sato, I., Suzuki, M., Nakayama, K., and Matsuo, Y. b-GAN: Unified framework of generative adversarial networks. 2016a.

Uehara, M., Sato, I., Suzuki, M., Nakayama, K., and Matsuo, Y. Generative adversarial nets from a density ratio estimation perspective. *arXiv preprint arXiv:1610.02920*, 2016b.

Supplementary material for: BreGMN: scaled-Bregman Generative Modeling Networks

A Proof of Proposition 1

Observe that

$$\begin{aligned}
&B_{t\log t}(P, M_0|M_0) - B_{t\log t}(Q, M_0|M_0) = D_{KL}(P||M_0) - D_{KL}(Q||M_0) \\
&= \int_{\mathcal{X}} \log\left(\frac{p(x)}{m_0(x)}\right) dP - \int_{\mathcal{X}} \log\left(\frac{q(x)}{m_0(x)}\right) dQ \\
&= \int_{\mathcal{X}} \log\left(m_0(x)\right) dQ - \int_{\mathcal{X}} \log\left(m_0(x)\right) dP + h(Q) - h(P) \\
&= \mathbb{E}_{V\sim Q} \log\left(m_0(V)\right) - \mathbb{E}_{U\sim P} \log\left(m_0(U)\right) + h(Q) - h(P)
\end{aligned}$$

where we denote the Shannon entropy as $h(P) = -\int_{\mathcal{X}} \log(p(x))dP$. Note that

$$\begin{aligned}
&|\log\left(m_0(V)\right) - \log\left(m_0(V)\right)| = \left|\int_0^1 \langle \nabla \log m_0(tv + (1-t)u), u - v\rangle dt\right| \\
&\leq \int_0^1 \left(\frac{3}{\sigma^2}(t\|v\| + (1-t)\|u\|) + \frac{4}{\sigma^2}\left(\mathbb{E}_{U\sim P}\|U\| + \mathbb{E}_{V\sim Q}\|V\|\right)\right)\|u - v\| dt \\
&= \left(\frac{3}{2\sigma^2}(\|v\| + \|u\|) + \frac{4}{\sigma^2}\left(\mathbb{E}_{U\sim P}\|U\| + \mathbb{E}_{V\sim Q}\|V\|\right)\right)\|u - v\| \qquad (9)
\end{aligned}$$

where we have used Cauchy-Schwartz inequality and have noted that

$$\|\nabla \log m_0(x)\| \leq \frac{3}{\sigma^2}\|x\| + \frac{4}{\sigma^2}\left(\mathbb{E}_{U\sim P}\|U\| + \mathbb{E}_{V\sim Q}\|V\|\right), \qquad \forall x \in \mathbb{R}^d,$$

by Proposition 2 of Polyanskiy & Wu (2016).

Let $W_p(\cdot,\cdot)$ denote the Wasserstein-p distance

$$W_p(\mu,\nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|^p d\pi(x,y)\right)^{\frac{1}{p}},$$

where $\Pi(\mu,\nu)$ denotes the set of *couplings* of μ and ν, i.e. the set of measures on $\mathcal{X}\times\mathcal{X}$ with marginals μ and ν.

Now, taking the expectation of (9) with respect to the W_2-optimal coupling π between P and Q, we have

$$\begin{aligned}
&|B_{t\log t}(P, M_0|M_0) - B_{t\log t}(Q, M_0|M_0)| \\
&\leq \mathbb{E}_{(u,v)\sim\pi}\left[\left(\frac{3}{2\sigma^2}(\|v\| + \|u\|) + \frac{4}{\sigma^2}\left(\mathbb{E}_{U\sim P}\|U\| + \mathbb{E}_{V\sim Q}\|V\|\right)\right)\|u - v\|\right] + |h(Q) - h(P)| \\
&\leq \sqrt{\left(\mathbb{E}_\pi\left(\frac{3}{2\sigma^2}(\|v\| + \|u\|) + \frac{4}{\sigma^2}\left(\mathbb{E}_{U\sim P}\|U\| + \mathbb{E}_{V\sim Q}\|V\|\right)\right)\right)\left(\mathbb{E}_\pi\|u - v\|^2\right)} + |h(Q) - h(P)| \\
&= cW_2(P,Q) + |h(Q) - h(P)|,
\end{aligned}$$

where we have again used the Cauchy-Schwarz inequality and have set the constant $c = \frac{11}{2\sigma^2}\left(\mathbb{E}_{U\sim P}\|U\| + \mathbb{E}_{V\sim Q}\|V\|\right)$. □