
BreGMN: scaled-Bregman Generative Modeling Networks

Abstract

Recent advances in f -divergence based generative modeling are full of remedies for handling support mismatch problem. Key ideas include modifying the objective function to integral probability measures (IPMs) that are well-behaved even on disjoint probabilities and 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* which 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. We study the behavior of the proposed model in theory and in practice, and demonstrate promising results on MNIST, CelebA and CIFAR-10 datasets.

1 Introduction

Modern deep generative modeling paradigms offer a powerful approach to learn data distributions. Pioneering models in this family such as generative adversarial networks (GANs) [1] and variational autoencoders (VAEs) [2] propose elegant solutions to generate high quality photo-realistic images, which were later evolved to generate other modalities of data. In GANs, much of the success of attaining photo-realism in generated images is attributed to the *adversarial* nature of training. 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 adversarial training, the generator is

pitched against a classifier: the classifier is trained to distinguish the generated from the true data samples, while the generator is simultaneously trained to generate samples that appear to come from the true data distribution. Upon successful training, the classifier fails to distinguish between the generated and true samples. Unlike the VAE, GAN is an implicit generative model since its likelihood function is implicitly defined and is intractable to express explicitly in general. Therefore training and inference are carried out using likelihood-free techniques such as the adversarial approach described above.

In their original formulation, GANs can be shown to approximately minimize an f -divergence measure between the true data distribution P and the distribution Q_ϕ induced by its generator G_ϕ [3]. The difficulty in training the generator using the f -divergence criterion is that the supports of data and model distributions need to perfectly match throughout training, since if the supports have non-overlapping portions at any time in the training phase, 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 [4] and training fails.

In this work, we present a novel method, BreGMN, for training both implicit adversarial and non-adversarial generative models that is based on *scaled Bregman divergences* (sBD) [5] 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 remain well-defined even when the data and generated distributions do not have matching supports. This well-definedness holds so long as the base measure has support that includes the supports of both the data and model distributions. In this work, we choose to leverage Gaussian distributions to augment distributions of the data and model to form a base measure that satisfies this constraint while also providing a training signal containing important geometric information. Despite using a Gaussian-smoothed base measure, we will show, in Section 3.2.1, that our method successfully learns generative models whose generated distribution matches the true data distribution, in contrast to noise adding regularization methods [6, 7] that have similar stability properties but converge to a noisy corruption

of the data distribution. Using this proposed base measure facilitates a stable decrease of the objective function and hence a more stable training behavior. Finally, we propose practical training algorithms for both adversarial and non-adversarial training of the proposed BreGMN model, and test it on simulated and real data. We study simulated data in a 2D setting with mismatched supports and show the advantage of our method in terms of convergence compared to f -divergences. We then evaluate BreGMN on real data when used to train both adversarial and non-adversarial generative models. We provide illustrative results on the MNIST, CIFAR10, and CelebA datasets that show comparable sample quality performance to the state-of-art.

The remainder of this paper 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

The genesis of adversarial generative modeling has ignited a flurry of research covering both practical and theoretical challenges in the field, e.g. [3, 8, 9, 4]. Within this body of work, 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 (i.e. not an f -divergence) that can be evaluated even if the distributions are not equally supported. Examples of this approach include Wasserstein GANs [4] that replace the f -divergence with Wasserstein distance between distributions, and other integral probability metric (IPM) based methods such as MMD GANs [9], Fisher GAN [10], etc. While IPM based methods are better behaved under non-overlapping support, 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 [11]. Similarly, WGAN requires ad-hoc methods for enforcing the Lipschitz constraint on the critic, e.g. via gradient clipping or penalization. The second class of approaches to remedy the support mismatch issue comes from [3], which showed how GANs can be trained by optimizing a *variational lower bound* to the f -divergence. The third class of methods to cope with instability and

support mismatch in generative models, and maybe the most closely related to our method is the class of *noise inducing regularization methods* [6, 7] that add full-support noise to the real or generated data and then run the model with this new full-support distribution. Similar to us, such methods solve the support mismatch problem by adding noise. There is a crucial difference, however. In their case, they must fit the model to a noisy version of data as opposed to a clean version. To cope with this new issue, it is proposed to decrease the noise over time with an annealing method. We show in Section 3.2.1, how our proposed method is superior to noise inducing regularization methods in this regard, as our method directly fits to the true distribution without any annealing.

In parallel, works such as [12] have studied the relation between the many different divergences available in the literature. An important extension to both Bregman and f -divergences, namely *scaled Bregman divergences*, is proposed in [5, 13]. 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. [14, 15, 16] study the use of Bregman divergences as the objective function for training single-layer neural networks for univariate and multivariate regression. In unsupervised learning, Bregman divergences have been used as the objective for clustering, in [17], while Bregman divergences for clustering models are relaxed to convex forms in [18]. [19, 20] have explored generative modeling based on Bregman divergences, 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) [1], f -divergence (f -GAN) [3], and various integral probability metrics (IPM, e.g. in Wasserstein-GAN, MMD-GAN) [4, 9]. 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** [21] forms a measure of distance between two vectors $\vec{p}, \vec{q} \in \mathbb{R}^d$ using a convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ as

$$B_f(\vec{p}, \vec{q}) = f(\vec{p}) - f(\vec{q}) - \nabla f(\vec{q}) \cdot (\vec{p} - \vec{q}),$$

which subsumes 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

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

where $f : \mathbb{R}^+ \rightarrow \mathbb{R}$ 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. The integration in (1) makes it possible to compute the discrepancy between two distributions using samples drawn from those distributions. In general, as the name divergence implies, the quantity is non-symmetric. It does not satisfy the triangle inequality either [22].

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 [6]. 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 M as a **base measure** for the Bregman divergence. Specifically, the **scaled Bregman divergence** [5] is given by

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

Notably, this expression is equal to the separable Bregman divergence of (1) when M is equal to the Lebesgue measure.

As shown in [5], the scaled Bregman divergence of (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). On the other hand, many classical criteria (including the KL and

Jensen-Shannon divergences) also 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 $f : \mathbb{R}^+ \rightarrow \mathbb{R}$ 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 -divergences family as:

$$\begin{aligned} B_f(P, Q|Q) &= \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) - f'(1)\left(\frac{p(x)}{q(x)} - 1\right) dQ \\ &= \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx, \end{aligned} \quad (3)$$

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 can be found in [5].

3.2 Scaled Bregman divergences with noisy base measures

As noted above, 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. To address this problem, for the scaled Bregman $B_f(P, Q|M)$, we are thus motivated to choose a “noisy” base measure M , specifically one that is formed by convolving some other measure with the Gaussian measure $\mathcal{N}(0, \Sigma)$, which has full support. Recall that convolution of two distributions corresponds to the addition of the associated random variables, i.e. from random realizations drawn from the distributions, hence in this case we are in effect 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 4.2 below). By analogy to the Jensen-Shannon divergence, in the scaled Bregman divergence,

$$B_f(P, Q|\tilde{M}) = \int_{\mathcal{X}} f\left(\frac{p(x)}{\tilde{m}(x)}\right) - f\left(\frac{q(x)}{\tilde{m}(x)}\right) - f'\left(\frac{q(x)}{\tilde{m}(x)}\right)\left(\frac{p(x)}{\tilde{m}(x)} - \frac{q(x)}{\tilde{m}(x)}\right) d\tilde{M}, \quad (4)$$

we choose \tilde{M} as,

$$\tilde{M} = \alpha(P * \mathcal{N}(0, \Sigma_1)) + (1 - \alpha)(Q * \mathcal{N}(0, \Sigma_2)) \quad (5)$$

for $0 \leq \alpha \leq 1$ and some covariances Σ_1 and Σ_2 , where

* denotes the convolution of two distributions.¹ Denote the density of \tilde{M} as \tilde{m} .

This choice prevents the terms $\frac{p(x)}{\tilde{m}(x)}$ and $\frac{q(x)}{\tilde{m}(x)}$ from going to infinity at any x . Importantly, observe that each term of the corresponding scaled Bregman $B_f(P, Q|\tilde{M})$ is always well defined and finite so long as $f(0)$ is since \tilde{M} has full support. This holds for any value of $0 \leq \alpha \leq 1$. 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 . Similarly, the ratio $\frac{q}{\tilde{m}}$ will be affected by P even outside the support of P . These ensure that training signals remain in the support mismatch case.

Interestingly, the presence of this training signal seems to indicate that geometric information is contained in it, since it varies with the distance between the supports of P and Q , which is explored in Section ??.

3.2.1 Distribution matching behavior

In this section, we analyze whether an optimizer based on scaled Bregman divergence with noisy base measure is consistent. To do this we identify a globally optimal solution to the ‘‘infinite data’’ problem

$$Q^* = \arg \min_Q B_f(P, Q|M),$$

where P is the true data distribution of training data. For this purpose, observe that P is in the feasible set so $B_f(P, P|M)$ upper bounds the optimal value, i.e. $B_f(P, P|M) \geq B_f(Q^*, P|M)$. Now, notice that for the convex function f , the scaled Bregman divergence is always *nonnegative*, $B_f(P, Q^*|M) \geq 0$. Finally, observe that

$$B_f(P, P|M) = \int f\left(\frac{p(x)}{m(x)}\right) - f\left(\frac{p(x)}{m(x)}\right) - f'\left(\frac{p(x)}{m(x)}\right) \left(\frac{p(x)}{m(x)} - \frac{p(x)}{m(x)}\right) dM = 0.$$

Hence the upper bound is tight and P is a global minimizer, i.e. $Q^* = P$. Hence in BregMN, the optimal distribution matches the *clean data* distribution, despite the base measure being noisy.

This is not the case for the very seemingly similar class of noise inducing regularization methods. Recall from Section 2, that these methods add full support noise to the data distribution to combat support mismatch problem and then run the model with this new full-

¹In practice we choose symmetric $\alpha = 0.5$ and isotropic $\Sigma_1 = \Sigma_2 = \sigma^2 I$ where we tune σ . Tuning α and Σ_i further based on data may yield small performance gains at the cost of additional computation.

support distribution.² In fact, they optimize for

$$Q^* = \arg \min_Q D_f(P_{X+\epsilon}, Q),$$

in which the optimal solution is clearly $Q^* = P_{X+\epsilon}$. In other words, in the class of noise inducing methods, the optimal distribution matches to the *noise corrupted* distribution yielding to a generator that must inherently fail to correctly describe the original data distribution. As a result, these methods must introduce ad-hoc patches such as annealing to address this issue, which adds training complexity and yields a new suite of parameters that must be tuned. Our approach instead successfully matches the correct distributions by itself.

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 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 ,

We propose to use the scaled-Bregman divergence (4) with noisy base measure (5) of Section 3.2 as the measure of the discrepancy between true and generated data distributions yielding

$$\begin{aligned} \min_\phi B_f(P, Q_\phi|\tilde{M}_\phi) &= \min_\phi \int_{\mathcal{X}} f\left(\frac{p(x)}{\tilde{m}_\phi(x)}\right) - \\ & f\left(\frac{q_\phi(x)}{\tilde{m}_\phi(x)}\right) - f'\left(\frac{q_\phi(x)}{\tilde{m}_\phi(x)}\right) \left(\frac{p(x)}{\tilde{m}_\phi(x)} - \frac{q_\phi(x)}{\tilde{m}_\phi(x)}\right) d\tilde{M}_\phi, \\ \text{s.t. } \tilde{M}_\phi &= \alpha(P * \mathcal{N}(0, \Sigma_1)) + (1 - \alpha)(Q_\phi * \mathcal{N}(0, \Sigma_2)) \end{aligned} \quad (6)$$

where p and q_ϕ are densities of true and generated distributions P, Q_ϕ respectively.

We will show in Section 4.1 that scaled-Bregman divergences can be easily estimated with respect to a base measure using only samples from the distributions. Unlike for f -divergences, the procedure here works even in the case of support mismatching distributions. This is particularly important when we aim to match distributions in very high dimensional spaces where there may not be any overlapping support [4].

4.1 Density ratio estimation (DRE)

In order to compute the divergence between data and model distributions, it is not always required that the

²See for example Section 3 of [6].

densities are known or can be evaluated on the realizations from those distributions. Instead, being able to evaluate the ratio between them, i.e. *density ratio estimation*, is typically all that is needed. For example, it is known that generative models based on f -divergences only need density ratio estimation.

Now, observe that scaled-Bregman divergences can be estimated via evaluating the ratio of the densities without needing to evaluate the densities themselves, since we can express the training objective of (6) as

$$\min_{\phi} \int f(r_{p/\tilde{m}}(x_i)) - f(r_{q_{\phi}/\tilde{m}}(x_i)) - f'(r_{q_{\phi}/\tilde{m}}(x_i)) (r_{p/\tilde{m}}(x_i) - r_{q_{\phi}/\tilde{m}}(x_i)) d\tilde{M}, \quad (7)$$

where $r_{a/b}$ denotes the ratio of the density function a to the density function b . Below, we describe two methods of density ratio estimation (DRE) between two distributions.

Discriminator-based DRE One possibility is to deploy a discriminator to estimate the density ratio. Let $y = 1$ if $x \sim P$ and $y = 0$ if $x \sim Q_{\phi}$. Further, let $\sigma(C(x)) = p(y = 1|x)$ be a discriminator, i.e. binary classifier, trained on samples from P and Q_{ϕ} where σ is the sigmoid function. It is then easy to show that $C(x) = -\log \frac{p(x)}{q_{\phi}(x)} = -\log r(x)$ [23], so C is represented as an invertible function of density ratio $r(x)$, from which $r(x)$ can be uniquely computed and used in (4). In fact, this method is the underlying principle in adversarial generative models. As such, most discriminator-based DREs result in adversarial training procedures when used in generative models.

MMD-based DRE The second method of estimating the density ratio does not use a discriminator, and thereby avoids potentially costly adversarial training of the generator. It instead uses methods based on kernelized statistical two-sample tests such as maximum mean discrepancy (MMD) of [24]. For this purpose, we can employ the MMD criterion as in [23] by solving for r in the RKHS in

$$\hat{r} = \arg \min_{r \in \mathcal{H}} \left\| \int k(x; \cdot) p(x) dx - \int k(x; \cdot) r(x) q_{\phi}(x) dx \right\|_{\mathcal{H}}^2,$$

where k is a kernel function and K is the corresponding Gram matrices of reproducing Hilbert space \mathcal{H} . We obtain a closed form estimator of the density ratio as

$$\hat{\mathbf{r}}_{p/q} = K_{q,q}^{-1} K_{q,p} \mathbf{1}. \quad (8)$$

4.2 Empirical estimation

Using the DRE estimators introduced above we create empirical estimators of the scaled-Bregman divergence

of (7) from finite training samples as

$$\min_{\phi} \hat{B}_f(P, Q_{\phi} | \tilde{M}_{\phi}) = \min_{\phi} \frac{1}{N} \sum_{i=1}^N f(r_{p/\tilde{m}_{\phi}}(x_i)) - f(r_{q_{\phi}/\tilde{m}_{\phi}}(x_i)) - f'(r_{q_{\phi}/\tilde{m}_{\phi}}(x_i)) (r_{p/\tilde{m}_{\phi}}(x_i) - r_{q_{\phi}/\tilde{m}_{\phi}}(x_i))$$

where the x_i are i.i.d. samples from the base distribution \tilde{m} with measure \tilde{M} .

Note that this empirical estimator \hat{B}_f does not have informative gradients with respect to ϕ if we only evaluate the DRE estimators on samples from an arbitrary base measure M which is not dependent on the generator parameters. However, choices of M that depend on Q , including our choice of \tilde{M} in (5) as well as the choice $M = Q$ (f -divergences), have informative gradients, allowing us to train the generator.

4.3 Training

Training the generator function G_{ϕ} using scaled-Bregman divergence is shown in Algorithm 1.

Algorithm 1: Training Algorithm of BreGMN

- 1 **while** *not converged* **do**
 - 2 **Step 1** Estimate the density ratios $r_{p/\tilde{m}}$ and $r_{q_{\phi}/\tilde{m}}$ using either (a) the adversarial discriminator-based (GAN-like) method or (b) the non-adversarial two-sample-test-based (MMD-like) method.
 - 3 **Step 2** Train the generator by optimizing
 - 4
$$\min_{\phi} \hat{B}_f(P, Q_{\phi} | \tilde{M}(P, Q_{\phi})) \quad \text{subject to}$$

$$\tilde{M}(P, Q_{\phi}) = 0.5 P * \mathcal{N}(0, \Sigma_1) + 0.5 Q_{\phi} * \mathcal{N}(0, \Sigma_2).$$
 - 5 **end**
-

5 Experiments

In this section, we present a detailed empirical evaluation of our proposed method. We first present a battery of controlled experiments on synthetic datasets to compare and contrast sBD with f -divergence measures and study its empirical convergence properties. Next, we show how sBD can be employed to train neural network based deep generative models of image data in both, non-adversarial and adversarial settings. Our results demonstrate that models trained using sBD are at par or better than the state-of-the-art generative models on generative quality. For network configurations and hyper-parameter setting used for experiments in this section, please see the reference code provided at <https://anonymous.4open.science/r/5ea4417b-15e9-49a0-b9c9-89ad46e5911f/>.

Figure 1: Non-adversarial training using scaled-Bregman divergence and MMD based DRE. While the quality is lower than GAN, it is on par with other non-adversarial methods such as MMDnet [25](Figure 2, top-right), while achieving an order of magnitude shorter training time.



5.1 Synthetic data: support mismatch

In these experiments, we demonstrate that in the regime where P and Q_ϕ have mismatched support, unlike f -divergence, sBD stays well defined and can be used to minimize the distance between P and Q_ϕ by adapting ϕ . Figure 2(a) shows two truncated Gaussian distributions that do not have overlapping supports ($P = \text{Red}$, $Q_\phi = \text{Green}$) and Figures 2(c-e) show the change in Q_ϕ as sBD is minimized ($M = \text{Blue}$) and Figures 2(f-h) denote the change in the divergence value as optimization progresses under different divergences. The competing divergences are KL, PD and sBD.

We know, by definition it is not possible to measure f -divergence between distributions with support mismatch. We empirically study this for two f -divergences and compare it with the case of sBD. Importantly, as shown in Figure 2(f) it is not only possible to measure sBD in this setting but it can also be minimized with respect to the parameters of the distributions as shown in Figures 2(b-e). On the other hand, minimizing KL divergence (KL) or Pearson divergence (PD) is not possible, see Figures 2(g-h). Notice, while KL is inf for distributions with non-overlapping supports, in our experiments we have clipped its value to keep it defined when the density ratio is zero.

Next, note that our empirical estimator for sBD depends on two important factors: the variance of the base distribution and number of samples (mini-batch size) used in the Monte-Carlo (MC) approximation. In Figure 3 we show how the empirical convergence rate observed above changes depending on these two factors. Figure 3(a) shows that the convergence rate is optimal when the variance of the base (Gaussian) distribution is set to six. Higher or lower values than this, nega-

tively affect the convergence. This is because at six, the base measure optimally overlaps with both P and Q_ϕ . Sample size is very critical for faster convergence, as shown in Figure 3(b) the method converges faster for higher value of samples size used in the MC step.

5.2 Non-adversarial generative model

The sBD estimator, as proposed in this work, depends on the two density ratio estimators $r_{p/m}$ and $r_{q_\phi/m}$. Since, like MMD and unlike f -divergences, sBD is defined for distributions with mismatched support, the MMD based density ratio estimator can be used in sBD estimation leading to a non-adversarial training similar to [25]. Figure 1 shows samples from a generator trained on MNIST dataset using our sBD estimator with MMD-based DRE estimators. As is typical with non-adversarial methods [25], the generative quality is not at par with adversarial methods but the training is significantly more stable.

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) [26] of MMD-GAN [9], GAN [1] 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 [27] 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 discrepancy [24] where the kernel is trained in an adversarial fashion. As shown in Table 1, both BreGMN and GAN perform better than MMD-GAN in terms of sample quality. While in our experiments BreGMN performs only slightly better than GAN on average, it has been shown in a large scale study of GANs [28] that the original GAN is at par with (and sometimes better than) most of the state of art variants including WGAN [4] on most typical datasets. Indeed, in general no single GAN-variant is better when evaluated on FID across all the datasets.

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

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

Figure 2: Figures (a-e): sBD can be used to minimize the divergence measure between two truncated-Gaussian distributions (in green and red) that do not have overlapping support. In Figures (b-d) we use the kernel density plots and blue points represent the samples from the base measure. Figure (f) shows how KL and PD change as the sBD is minimized. Figures (g,h) show that minimizing KL or PD is not possible since the two distributions have disjoint support.

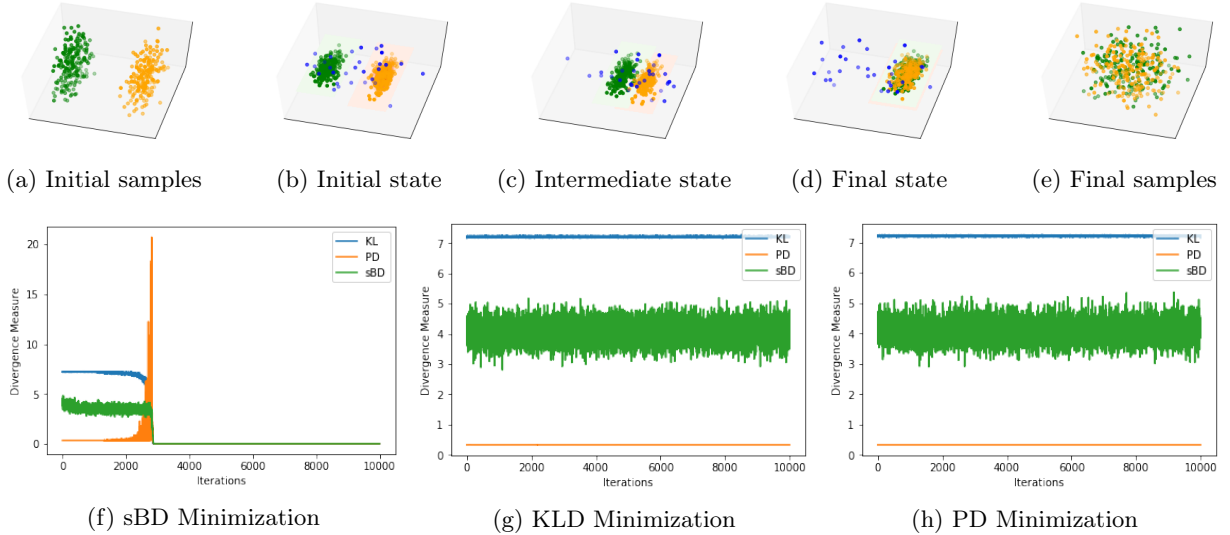
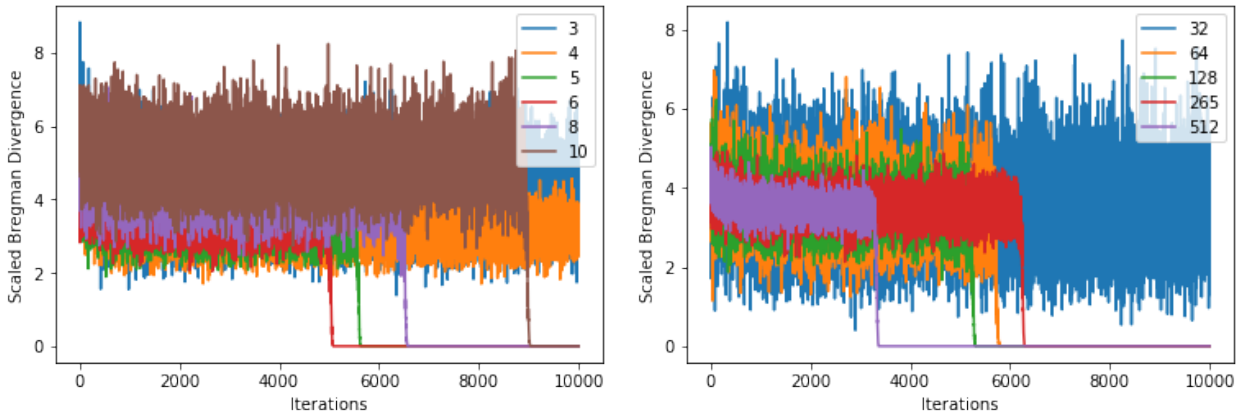


Figure 3: Empirical convergence analysis for the experiment in Figure 2 (we keep one of the distributions fixed). Figure (a) shows the effect of the variance of the base distribution on the empirical convergence and Figure (b) shows the effect of the batch size (number of samples) on the empirical convergence.



(a) Effect of the variance of the base distribution on the empirical rate of convergence (b) Effect of the batch size on the empirical rate of convergence

6 Conclusion

In this work, we proposed scaled-Bregman divergence based generative models and identified base measures

for them that facilitate effective training. We showed that the proposed approach provides a divergence-based criterion to model the data distribution using deep generative networks that is robust to support mismatch, in contrast to the f -divergence based training methods. Future research directions include the study of alternative base measures in the scaled Bregman divergence objective, and further analysis of the relationship between scaled-Bregman divergences and IPMs.

References

- [1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems*, 2014.
- [2] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *International Conference on Learning Representations, ICLR*, 2014.
- [3] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in *Neural Information Processing Systems*, 2016.
- [4] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.
- [5] W. Stummer and I. Vajda, "On Bregman distances and divergences of probability measures," *IEEE Trans. Information Theory*, vol. 58, no. 3, pp. 1277–1288, 2012.
- [6] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *International Conference on Learning Representations, ICLR*, 2017.
- [7] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann, "Stabilizing training of generative adversarial networks through regularization," in *Neural Information Processing Systems*, 2017.
- [8] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. A. Sutton, "VEEGAN: reducing mode collapse in GANs using implicit variational learning," in *Neural Information Processing Systems*, 2017.
- [9] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos, "MMD GAN: Towards deeper understanding of moment matching network," in *Neural Information Processing Systems*, 2017.
- [10] Y. Mroueh and T. Sercu, "Fisher GAN," in *Advances in Neural Information Processing Systems*, 2017.
- [11] A. Srivastava, K. Xu, M. U. Gutmann, and C. A. Sutton, "Ratio matching MMD nets: Low dimensional projections for effective deep generative models," *CoRR*, vol. abs/1806.00101, 2018.
- [12] S.-i. Amari and A. Cichocki, "Information geometry of divergence functions," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 58, no. 1, pp. 183–195, 2010.
- [13] A. Kißlinger and W. Stummer, "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.
- [14] D. P. Helmbold, J. Kivinen, and M. K. Warmuth, "Worst-case loss bounds for single neurons," in *Neural Information Processing Systems*. MIT Press, 1995.
- [15] P. Auer, M. Herbster, and M. K. Warmuth, "Exponentially many local minima for single neurons," in *Neural Information Processing Systems*, 1996.
- [16] J. Kivinen and M. K. Warmuth, "Relative loss bounds for multidimensional regression problems," in *Neural Information Processing Systems*, 1998.
- [17] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.
- [18] H. Cheng, X. Zhang, and D. Schuurmans, "Convex relaxations of Bregman divergence clustering," in *Uncertainty in Artificial Intelligence, UAI*, 2013.
- [19] M. Uehara, I. Sato, M. Suzuki, K. Nakayama, and Y. Matsuo, "b-GAN: Unified framework of generative adversarial networks," 2016.
- [20] —, "Generative adversarial nets from a density ratio estimation perspective," *arXiv preprint arXiv:1610.02920*, 2016.
- [21] L. 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*, vol. 7, no. 3, pp. 200–217, 1967.
- [22] S. Acharyya, A. Banerjee, and D. Boley, "Bregman divergences and triangle inequality," in *SIAM International Conference on Data Mining*. SIAM, 2013.
- [23] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

-
- [24] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.
- [25] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, “Training generative neural networks via maximum mean discrepancy optimization,” *arXiv preprint arXiv:1505.03906*, 2015.
- [26] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANS trained by a two time-scale update rule converge to a local nash equilibrium,” in *Neural Information Processing Systems*, 2017.
- [27] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.
- [28] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, “Are gans created equal? a large-scale study,” in *Advances in neural information processing systems*, 2018, pp. 700–709.
- [29] Y. Polyanskiy and Y. Wu, “Wasserstein continuity of entropy and outer bounds for interference channels,” *IEEE Trans. Information Theory*, vol. 62, no. 7, pp. 3992–4002, 2016.

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

A Proof of Theorem ??

We have

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

Note that for fixed \tilde{m} of full support and not dependent on ϕ , the first two claims in Theorem ?? are immediate. We thus focus on the case where \tilde{m} depends on ϕ .

For $\tilde{m} = 1/2(P * \mathcal{N}_\sigma + Q_\phi * \mathcal{N}_\sigma)$ and fixed σ , the density ratios can be zero but never infinity on more than a measure zero set (measure zero w.r.t \tilde{M}). More specifically, the density ratios are continuous and differentiable with respect to Q_ϕ .

Claim 1. Since f and f' are continuous everywhere including at zero, and Q_ϕ is (pointwise) continuous with respect to ϕ , $B_f(P, Q|\tilde{M})$ must also be.

Claim 2. Since f and f' are differentiable everywhere including at zero, and Q_ϕ is (pointwise) differentiable with respect to ϕ , $B_f(P, Q|\tilde{M})$ must also be.

Claim 3. The claim regarding the Wasserstein distance follows immediately from [4]. For f -divergences with unbounded $f(t)/t$, since f is convex the f -divergence integrand $q(x)f(p(x)/q(x))$ must be infinite whenever $q(x) = 0$ and $p(x) > 0$, i.e. whenever support of p is not contained in that of q . Hence the corresponding f -divergence is not continuous everywhere.

B Proof of Proposition ??

Observe that

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

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

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

where we have used Cauchy-Schwartz inequality and have noted that for all $x \in \mathbb{R}^d$,

$$\|\nabla \log \tilde{m}(x)\| \leq \frac{3}{\sigma^2} \|x\| + \frac{4}{\sigma^2} (\mathbb{E}_{U \sim P} \|U\| + \mathbb{E}_{V \sim Q} \|V\|),$$

by Proposition 2 of [29].

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 .

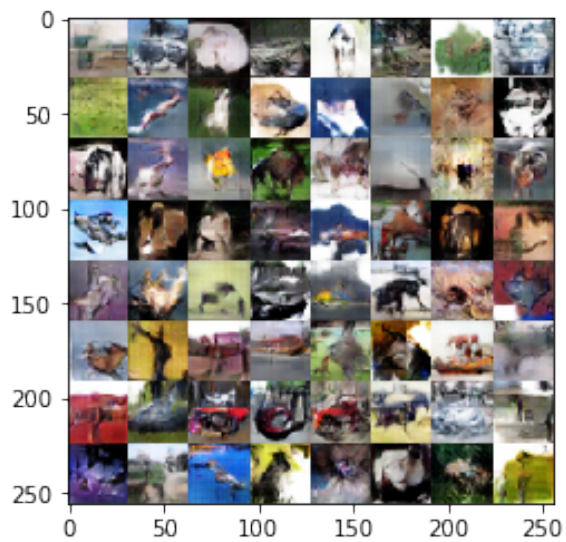
Now, taking the expectation of (9) with respect to the W_2 -optimal coupling π between P and Q , we have

$$\begin{aligned} & |B_{t \log t}(P, \tilde{M}|\tilde{M}) - B_{t \log t}(Q, \tilde{M}|\tilde{M})| \\ &\leq \mathbb{E}_{(u, v) \sim \pi} \left[\left(\frac{3}{2\sigma^2} (\|v\| + \|u\|) \right. \right. \\ &\quad \left. \left. + \frac{4}{\sigma^2} (\mathbb{E}_{U \sim P} \|U\| + \mathbb{E}_{V \sim Q} \|V\|) \right) \|u - v\| \right] \\ &\quad + |h(Q) - h(P)| \\ &\leq (\mathbb{E}_\pi \|u - v\|^2)^{\frac{1}{2}} \left(\mathbb{E}_\pi \left(\frac{3}{2\sigma^2} (\|v\| + \|u\|) \right. \right. \\ &\quad \left. \left. + \frac{4}{\sigma^2} (\mathbb{E}_{U \sim P} \|U\| + \mathbb{E}_{V \sim Q} \|V\|) \right) \right)^{\frac{1}{2}} \\ &\quad + |h(Q) - h(P)| \\ &= cW_2(P, Q) + |h(Q) - h(P)|, \end{aligned}$$

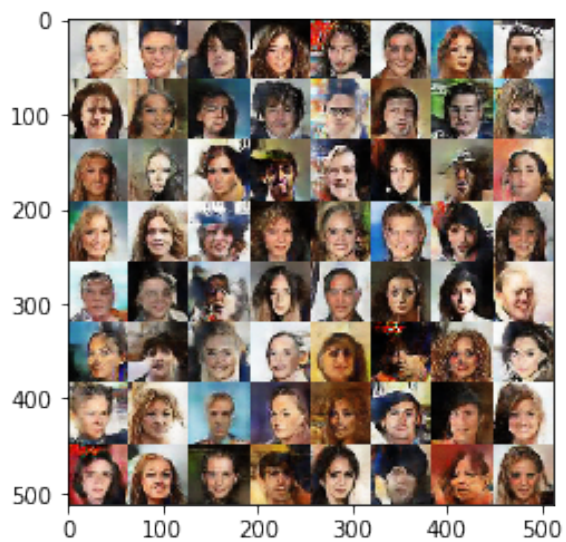
where we have again used the Cauchy-Schwarz inequality and have set the constant $c = \frac{11}{2\sigma^2} (\mathbb{E}_{U \sim P} \|U\| + \mathbb{E}_{V \sim Q} \|V\|)$. \square

C Generated Samples

Figure 4: Random samples from Adversarial BreGMN models (after 5 Epochs)



(a) CIFAR10



(b) CELEB A