McGan: Mean and Covariance Feature Matching GAN

Youssef Mroueh [* 1 2] **Tom Sercu** [* 1 2] **Vaibhava Goel** [2]

Abstract

We introduce new families of Integral Probability Metrics (IPM) for training Generative Adversarial Networks (GAN). Our IPMs are based on matching statistics of distributions embedded in a finite dimensional feature space. Mean and covariance feature matching IPMs allow for stable training of GANs, which we will call McGan. McGan minimizes a meaningful loss between distributions.

1. Introduction

Unsupervised learning of distributions is an important problem, in which we aim to learn underlying features that unveil the hidden the structure in the data. The classic approach to learning distributions is by explicitly parametrizing the data likelihood and fitting this model by maximizing the likelihood of the real data. An alternative recent approach is to learn a generative model of the data without explicit parametrization of the likelihood. Variational Auto-Encoders (VAE) (Kingma & Welling, 2013) and Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) fall under this category.

We focus on the GAN approach. In a nutshell GANs learn a generator of the data via a min-max game between the generator and a discriminator, which learns to distinguish between "real" and "fake" samples. In this work we focus on the objective function that is being minimized between the learned generator distribution $\mathbb{P}_\theta$ and the real data distribution $\mathbb{P}_r$.

The original work of (Goodfellow et al., 2014) showed that in GAN this objective is the Jensen-Shannon divergence. (Nowozin et al., 2016) showed that other φ-divergences can be successfully used. The Maximum Mean Discrepancy objective (MMD) for GAN training was proposed in (Li

et al., 2015; Dziugaite et al., 2015). As shown empirically in (Salimans et al., 2016), one can train the GAN discriminator using the objective of (Goodfellow et al., 2014) while training the generator using mean feature matching. An energy based objective for GANs was also developed recently (Zhao et al., 2017). Finally, closely related to our paper, the recent work *Wasserstein GAN* (WGAN) of (Arjovsky et al., 2017) proposed to use the Earth Moving distance (EM) as an objective for training GANs. Furthermore (Arjovsky et al., 2017) show that the EM objective has many advantages as the loss function correlates with the quality of the generated samples and the *mode dropping* problem is reduced in WGAN.

In this paper, inspired by the MMD distance and the kernel mean embedding of distributions (Muandet et al., 2016) we propose to embed distributions in a finite dimensional feature space and to match them based on their **mean and covariance feature statistics**. Incorporating first and second order statistics has a better chance to capture the various modes of the distribution. While mean feature matching was empirically used in (Salimans et al., 2016), we show in this work that it is theoretically grounded: similarly to the EM distance in (Arjovsky et al., 2017), mean and covariance feature matching of two distributions can be written as a distance in the framework of Integral Probability Metrics (IPM) (Muller, 1997). To match the means, we can use any ℓ_q norm, hence we refer to mean matching IPM, as $\text{IPM}_{\mu,q}$. For matching covariances, in this paper we consider the Ky-Fan norm, which can be computed cheaply without explicitly constructing the full covariance matrices, and refer to the corresponding IPM as IPM_Σ.

Our technical contributions can be summarized as follows:

a) We show in Section 3 that the ℓ_q mean feature matching $\text{IPM}_{\mu,q}$ has two equivalent primal and dual formulations and can be used as an objective for GAN training in both formulations.

b) We show in Section 3.3 that the parametrization used in *Wasserstein GAN* corresponds to ℓ_1 mean feature matching GAN ($\text{IPM}_{\mu,1}$ GAN in our framework).

c) We show in Section 4.2 that the covariance feature matching IPM_Σ admits also two dual formulations, and can be used as an objective for GAN training.

---

[*]Equal contribution [1]AI Foundations. IBM T.J. Watson Research Center, NY, USA [2]Watson Multimodal Algorithms and Engines Group. IBM T.J. Watson Research Center, NY, USA. Correspondence to: Youssef Mroueh <mroueh@us.ibm.com>.

d) Similar to *Wasserstein GAN*, we show that mean feature matching and covariance matching GANs (McGan) are stable to train, have a reduced mode dropping and the IPM loss correlates with the quality of the generated samples.

2. Integral Probability Metrics

We define in this Section IPMs as a distance between distribution. Intuitively each IPM finds a "critic" f (Arjovsky et al., 2017) which maximally discriminates between the distributions.

2.1. IPM Definition

Consider a compact space $\mathcal{X}$ in $\mathbb{R}^d$. Let $\mathscr{F}$ be a set of measurable and bounded real valued functions on $\mathcal{X}$. Let $\mathscr{P}(\mathcal{X})$ be the set of measurable probability distributions on $\mathcal{X}$. Given two probability distributions $\mathbb{P}, \mathbb{Q} \in \mathscr{P}(\mathcal{X})$, the Integral probability metric (IPM) indexed by the function space $\mathscr{F}$ is defined as follows (Muller, 1997):

$$d_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathscr{F}} \left| \mathbb{E}_{x \sim \mathbb{P}} f(x) - \mathbb{E}_{x \sim \mathbb{Q}} f(x) \right|.$$

In this paper we are interested in symmetric function spaces $\mathscr{F}$, i.e $\forall f \in \mathscr{F}, -f \in \mathscr{F}$, hence we can write the IPM in that case without the absolute value:

$$d_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathscr{F}} \left\{ \mathbb{E}_{x \sim \mathbb{P}} f(x) - \mathbb{E}_{x \sim \mathbb{Q}} f(x) \right\}. \quad (1)$$

It is easy to see that $d_{\mathscr{F}}$ defines a pseudo-metric over $\mathscr{P}(X)$. ($d_{\mathscr{F}}$ non-negative, symmetric and satisfies the triangle inequality. A pseudo metric means that $d_{\mathscr{F}}(\mathbb{P}, \mathbb{P}) = 0$ but $d_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) = 0$ does not necessarily imply $\mathbb{P} = \mathbb{Q}$.

By choosing $\mathscr{F}$ appropriately (Sriperumbudur et al., 2012; 2009), various distances between probability measures can be defined. In the next subsection following (Arjovsky et al., 2017; Li et al., 2015; Dziugaite et al., 2015) we show how to use IPM to learn generative models of distributions, we then specify a special set of functions $\mathscr{F}$ that makes the learning tractable.

2.2. Learning Generative Models with IPM

In order to learn a generative model of a distribution $\mathbb{P}_r \in \mathscr{P}(\mathcal{X})$, we learn a function

$$g_\theta : \mathcal{Z} \subset \mathbb{R}^{n_z} \to \mathcal{X},$$

such that for $z \sim p_z$, the distribution of $g_\theta(z)$ is close to the real data distribution $\mathbb{P}_r$, where p_z is a fixed distribution on $\mathcal{Z}$ (for instance $z \sim \mathcal{N}(0, I_{n_z})$). Let $\mathbb{P}_\theta$ be the distribution of $g_\theta(z), z \sim p_z$. Using an IPM indexed by a function class $\mathscr{F}$ we shall solve therefore the following problem:

$$\min_{g_\theta} d_{\mathscr{F}}(\mathbb{P}_r, \mathbb{P}_\theta) \quad (2)$$

Hence this amounts to solving the following min-max problem:

$$\min_{g_\theta} \sup_{f \in \mathscr{F}} \mathbb{E}_{x \sim \mathbb{P}_r} f(x) - \mathbb{E}_{z \sim p_z} f(g_\theta(z))$$

Given samples $\{x_i, 1 \dots N\}$ from $\mathbb{P}_r$ and samples $\{z_i, 1 \dots M\}$ from p_z we shall solve the following empirical problem:

$$\min_{g_\theta} \sup_{f \in \mathscr{F}} \frac{1}{N} \sum_{i=1}^{N} f(x_i) - \frac{1}{M} \sum_{j=1}^{M} f(g_\theta(z_j)),$$

in the following we consider for simplicity $M = N$.

3. Mean Feature Matching GAN

In this Section we introduce a class of functions $\mathscr{F}$ having the form $\langle v, \Phi_\omega(x) \rangle$, where vector $v \in \mathbb{R}^m$ and $\Phi_\omega : \mathcal{X} \to \mathbb{R}^m$ a non linear feature map (typically parametrized by a neural network). We show in this Section that the IPM defined by this function class corresponds to the distance between the mean of the distribution in the Φ_ω space.

3.1. IPM$_{\mu,q}$: Mean Matching IPM

More formally consider the following function space:

$$\mathscr{F}_{v,\omega,p} = \{ f(x) = \langle v, \Phi_\omega(x) \rangle \big| v \in \mathbb{R}^m, \|v\|_p \leq 1,$$
$$\Phi_\omega : \mathcal{X} \to \mathbb{R}^m, \omega \in \Omega \},$$

where $\|.\|_p$ is the ℓ_p norm. $\mathscr{F}_{v,\omega,p}$ is the space of bounded linear functions defined in the non linear feature space induced by the parametric feature map Φ_ω. Φ_ω is typically a multi-layer neural network. The parameter space Ω is chosen so that the function space $\mathscr{F}$ is bounded. Note that for a given ω, $\mathscr{F}_{v,\omega,p}$ is a finite dimensional Hilbert space.

We recall here simple definitions on dual norms that will be necessary for the analysis in this Section. Let $p, q \in [1, \infty]$, such that $\frac{1}{p} + \frac{1}{q} = 1$. By duality of norms we have: $\|x\|_q = \max_{v, \|v\|_p \leq 1} \langle v, x \rangle$ and the Holder inequality: $\left| \langle x, y \rangle \right| \leq \|x\|_p \|y\|_q$.

From Holder inequality we obtain the following bound:

$$\left| f(x) \right| = \left| \langle v, \Phi_\omega x \rangle \right| \leq \|v\|_p \|\Phi_\omega(x)\|_q \leq \|\Phi_\omega(x)\|_q.$$

To ensure that f is bounded, it is enough to consider Ω such that $\|\Phi_\omega(x)\|_q \leq B, \forall x \in \mathcal{X}$. Given that the space $\mathcal{X}$ is bounded it is sufficient to control the norm of the weights and biases of the neural network Φ_ω by regularizing the ℓ_∞ (clamping) or ℓ_2 norms (weight decay) to ensure the boundedness of $\mathscr{F}_{v,\omega,p}$.

a) IPM $_{\mu,2}$: Level sets of $f(x) = \langle v^*, \Phi_\omega(x) \rangle$
$$v^* = \frac{\mu_w(\mathbb{P}) - \mu_w(\mathbb{Q})}{\|\mu_w(\mathbb{P}) - \mu_w(\mathbb{Q})\|_2}.$$

b) IPM $_\Sigma$: Level sets of $f(x) = \sum_{j=1}^k \langle u_j, \Phi_\omega(x) \rangle \langle v_j, \Phi_\omega(x) \rangle$
$k = 3, u_j, v_j$ left and right singular vectors of $\Sigma_w(\mathbb{P}) - \Sigma_w(\mathbb{Q})$.

Level Sets of c) $\langle u_1, \Phi_\omega(x) \rangle \langle v_1, \Phi_\omega(x) \rangle$ d) $\langle u_2, \Phi_\omega(x) \rangle \langle v_2, \Phi_\omega(x) \rangle$ e) $\langle u_3, \Phi_\omega(x) \rangle \langle v_3, \Phi_\omega(x) \rangle$
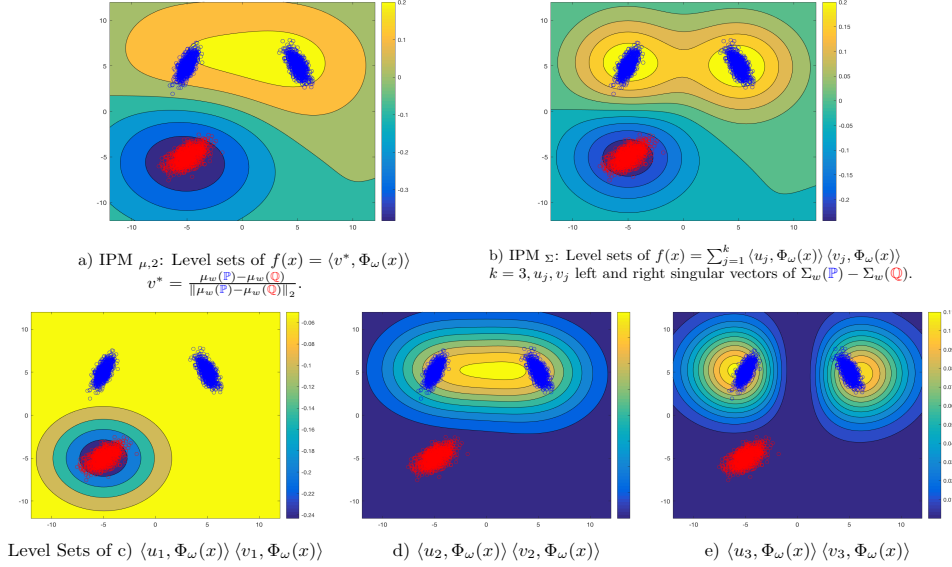
Figure 1. Motivating example on synthetic data in 2D, showing how different components in covariance matching can target different regions of the input space. Mean matching (a) is not able to capture the two modes of the bimodal "real" distribution $\mathbb{P}$ and assigns higher values to one of the modes. Covariance matching (b) is composed of the sum of three components (c)+(d)+(e), corresponding to the top three "critic directions". Interestingly, the first direction (c) focuses on the "fake" data $\mathbb{Q}$, the second direction (d) focuses on the "real" data, while the third direction (e) is mode selective. This suggests that using covariance matching would help reduce mode dropping in GAN. In this toy example Φ_ω is a fixed random Fourier feature map (Rahimi & Recht, 2008) of a Gaussian kernel (i.e. a finite dimensional approximation).

Now that we ensured the boundedness of $\mathscr{F}_{v,\omega,p}$, we look at its corresponding IPM:

$$d_{\mathscr{F}_{v,\omega,p}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathscr{F}_{v,\omega,p}} \mathbb{E}_{x \sim \mathbb{P}} f(x) - \mathbb{E}_{x \sim \mathbb{Q}} f(x)$$

$$= \max_{\omega \in \Omega, v, \|v\|_p \leq 1} \left\langle v, \mathbb{E}_{x \sim \mathbb{P}} \Phi_\omega(x) - \mathbb{E}_{x \sim \mathbb{Q}} \Phi_\omega(x) \right\rangle$$

$$= \max_{\omega \in \Omega} \left[\max_{v, \|v\|_p \leq 1} \left\langle v, \mathbb{E}_{x \sim \mathbb{P}} \Phi_\omega(x) - \mathbb{E}_{x \sim \mathbb{Q}} \Phi_\omega(x) \right\rangle \right]$$

$$= \max_{\omega \in \Omega} \|\mu_\omega(\mathbb{P}) - \mu_\omega(\mathbb{Q})\|_q,$$

where we used the linearity of the function class and expectation in the first equality and the definition of the dual norm $\|.\|_q$ in the last equality and our definition of the mean feature embedding of a distribution $\mathbb{P} \in \mathscr{P}(\mathcal{X})$:

$$\mu_\omega(\mathbb{P}) = \mathbb{E}_{x \sim \mathbb{P}} \left[\Phi_\omega(x) \right] \in \mathbb{R}^m.$$

We see that the IPM indexed by $\mathscr{F}_{v,\omega,p}$, corresponds to the Maximum mean feature Discrepancy between the two distributions. Where the maximum is taken over the parameter set Ω, and the discrepancy is measured in the ℓ_q sense between the mean feature embedding of $\mathbb{P}$ and $\mathbb{Q}$. In other words this IPM is equal to the worst case ℓ_q distance between mean feature embeddings of distributions. We refer in what follows to $d_{\mathscr{F}_{v,\omega,p}}$ as IPM$_{\mu,q}$.

3.2. Mean Feature Matching GAN

We turn now to the problem of learning generative models with IPM$_{\mu,q}$. Setting $\mathscr{F}$ to $\mathscr{F}_{v,\omega,p}$ in Equation (2) yields to the following min-max problem for learning generative models:

$$\min_{g_\theta} \max_{\omega \in \Omega} \max_{v, \|v\|_p \leq 1} \mathscr{L}_\mu(v, \omega, \theta), \tag{3}$$

where

$$\mathscr{L}_\mu(v, \omega, \theta) = \left\langle v, \mathbb{E}_{x \sim \mathbb{P}_r} \Phi_\omega(x) - \mathbb{E}_{z \sim p_z} \Phi_\omega(g_\theta(z)) \right\rangle,$$

or equivalently using the dual norm:

$$\min_{g_\theta} \max_{\omega \in \Omega} \|\mu_\omega(\mathbb{P}_r) - \mu_\omega(\mathbb{P}_\theta)\|_q, \tag{4}$$

where $\mu_\omega(\mathbb{P}_\theta) = \mathbb{E}_{z \sim p_z} \Phi_\omega(g_\theta(z))$.

We refer to formulations (3) and (4) as primal and dual formulation respectively.

The dual formulation in Equation (4) has a simple interpretation as an adversarial learning game: while the feature space Φ_ω tries to map the mean feature embeddings of the *real* distribution $\mathbb{P}_r$ and the *fake* distribution $\mathbb{P}_\theta$ to be far apart (maximize the ℓ_q distance between the mean embeddings), the generator g_θ tries to put them close one to another. Hence we refer to this IPM as *mean matching* IPM.

We devise empirical estimates of both formulations in Equations (3) and (4), given samples $\{x_i, i = 1 \ldots N\}$ from $\mathbb{P}_r$, and $\{z_i, i = 1 \ldots N\}$ from p_z. The primal formulation (3) is more amenable to stochastic gradient descent since the expectation operation appears in a linear way in the cost function of Equation (3), while it is non linear in the cost function of the dual formulation (4) (inside the norm). We give here the empirical estimate of the primal formulation by giving empirical estimates $\hat{\mathscr{L}}_\mu(v, \omega, \theta)$ of the primal cost function:

$$(P_\mu) : \min_{g_\theta} \max_{\substack{\omega \in \Omega \\ v, ||v||_p \leq 1}} \left\langle v, \frac{1}{N} \sum_{i=1}^{N} \Phi_\omega(x_i) - \frac{1}{N} \sum_{i=1}^{N} \Phi_\omega(g_\theta(z_i)) \right\rangle$$

An empirical estimate of the dual formulation can be also given as follows:

$$(D_\mu) : \min_{g_\theta} \max_{\omega \in \Omega} \left\| \frac{1}{N} \sum_{i=1}^{N} \Phi_\omega(x_i) - \frac{1}{N} \sum_{i=1}^{N} \Phi_\omega(g_\theta(z_i)) \right\|_q$$

In what follows we refer to the problem given in (P_μ) and (D_μ) as ℓ_q *Mean Feature Matching* GAN. Note that while (P_μ) does not need real samples for optimizing the generator, (D_μ) does need samples from real and fake. Furthermore we will need a large minibatch of real data in order to get a good estimate of the expectation. This makes the primal formulation more appealing computationally.

3.3. Related Work

We show in this Section that several previous works on GAN, can be written within the ℓ_q mean feature matching IPM ($\text{IPM}_{\mu,q}$) minimization framework:

a) *Wasserstein GAN* (WGAN): (Arjovsky et al., 2017) recently introduced Wasserstein GAN. While the main motivation of this paper is to consider the IPM indexed by Lipchitz functions on $\mathcal{X}$, we show that the particular parametrization considered in (Arjovsky et al., 2017) corresponds to a mean feature matching IPM.
Indeed (Arjovsky et al., 2017) consider the function set parametrized by a convolutional neural network with a linear output layer and weight clipping. Written in our notation, the last linear layer corresponds to v, and the convolutional neural network below corresponds to Φ_ω. Since v and ω are simultaneously clamped, this corresponds to restricting v to be in the ℓ_∞ unit ball, and to define in Ω constraints on the ℓ_∞ norms of ω. In other words (Arjovsky et al., 2017) consider functions in $\mathscr{F}_{v,\omega,p}$, where $p = \infty$. Setting $p = \infty$ in Equation (3), and $q = 1$ in Equation (4), we see that in WGAN we are minimizing $d_{\mathscr{F}_{v,\omega,\infty}}$, that corresponds to ℓ_1 mean feature matching GAN.

b) *MMD GAN*: Let $\mathscr{H}$ be a Reproducing Kernel Hilbert Space (RKHS) with k its reproducing kernel. For any valid

PSD kernel k there exists an *infinite dimensional feature map* $\Phi : \mathcal{X} \to \mathscr{H}$ such that: $k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathscr{H}}$. For an RKHS Φ is noted usually $k(x, .)$ and satisfies the reproducing proprety:

$$f(x) = \langle f, \Phi(x) \rangle_{\mathscr{H}}, \text{ for all } f \in \mathscr{H}.$$

Setting $\mathscr{F} = \{ f \, | \, ||f||_{\mathscr{H}} \leq 1 \}$ in Equation (1) the IPM $d_{\mathscr{F}}$ has a simple expression:

$$
\begin{aligned}
d_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) &= \sup_{f, ||f||_{\mathscr{H}} \leq 1} \left\{ \left\langle f, \mathop{\mathbb{E}}_{x \sim \mathbb{P}} \Phi(x) - \mathop{\mathbb{E}}_{x \sim \mathbb{Q}} \Phi(x) \right\rangle \right\} \\
&= \left\| \mu(\mathbb{P}) - \mu(\mathbb{Q}) \right\|_{\mathscr{H}}, \quad (5)
\end{aligned}
$$

where $\mu(\mathbb{P}) = \mathop{\mathbb{E}}_{x \sim \mathbb{P}} \Phi(x) \in \mathscr{H}$ is the so called kernel mean embedding (Muandet et al., 2016). $d_{\mathscr{F}}$ in this case is the so called Maximum kernel Mean Discrepancy (MMD) (Gretton et al., 2012) . Using the reproducing property MMD has a closed form in term of the kernel k. Note that $\text{IPM}_{\mu,2}$ is a special case of MMD when the feature map is finite dimensional, with the main difference that the feature map is fixed in case of MMD and learned in the case of $\text{IPM}_{\mu,2}$. (Li et al., 2015; Dziugaite et al., 2015) showed that GANs can be learned using MMD with a fixed gaussian kernel.

c) *Improved GAN*: Building on the pioneering work of (Goodfellow et al., 2014), (Salimans et al., 2016) suggested to learn the discriminator with the binary cross entropy criterion of GAN while learning the generator with ℓ_2 mean feature matching. The main difference of our $\text{IPM}_{\mu,2}$ GAN is that both "discriminator" and "generator" are learned using the mean feature matching criterium, with additional constraints on Φ_ω.

4. Covariance Feature Matching GAN

4.1. IPM_Σ: Covariance Matching IPM

As follows from our discussion of mean matching IPM comparing two distributions amounts to comparing a first order statistics, the mean of their feature embeddings. Here we ask the question how to incorporate second order statistics, i.e covariance information of feature embeddings.

In this Section we will provide a function space $\mathscr{F}$ such that the IPM in Equation (1) captures second order information. Intuitively a distribution of points represented in a feature space can be approximately captured by its mean and its covariance. Commonly in unsupervised learning, this covariance is approximated by its first k principal components (PCA directions), which capture the directions of maximal variance in the data. Similarly, the metric we define in this Section will find k directions that maximize the discrimination between the two covariances. Adding second order information would enrich the discrimination power of the feature space (See Figure 1).

This intuition motivates the following function space of bilinear functions in Φ_ω :

$$\mathscr{F}_{U,V,\omega} = \{f(x) = \sum_{j=1}^{k} \langle u_j, \Phi_\omega(x) \rangle \langle v_j, \Phi_\omega(x) \rangle$$

$$\{u_j\}, \{v_j\} \in \mathbb{R}^m \text{ orthonormal } j = 1 \dots k, \omega \in \Omega\}.$$

Note that the set $\mathscr{F}_{U,V,\omega}$ is symmetric and hence the IPM indexed by this set (Equation (1)) is well defined. It is easy to see that $\mathscr{F}_{U,V,\omega}$ can be written as:

$$\mathscr{F}_{U,V,\omega} = \{f(x) = \langle U^\top \Phi_\omega(x), V^\top \Phi_\omega(x) \rangle) \Big|$$

$$U, V \in \mathbb{R}^{m \times k}, U^\top U = I_k, V^\top V = I_k, \omega \in \Omega\}$$

the parameter set Ω is such that the function space remains bounded. Let

$$\Sigma_\omega(\mathbb{P}) = \underset{x \sim \mathbb{P}}{\mathbb{E}} \Phi_\omega(x) \Phi_\omega(x)^\top,$$

be the uncentered feature covariance embedding of $\mathbb{P}$. It is easy to see that $\underset{x \sim \mathbb{P}}{\mathbb{E}} f(x)$ can be written in terms of U, V, and $\Sigma_\omega(\mathbb{P})$:

$$\underset{x \sim \mathbb{P}}{\mathbb{E}} f(x) = \underset{x \sim \mathbb{P}}{\mathbb{E}} \langle U^\top \Phi(x), V^\top \Phi(x) \rangle = Trace(U^\top \Sigma_\omega(\mathbb{P})V).$$

For a matrix $A \in \mathbb{R}^{m \times m}$, we note by $\sigma_j(A)$ the singular value of A, $j = 1 \dots m$ in descending order. The 1-schatten norm or the nuclear norm is defined as the sum of singular values, $\|A\|_* = \sum_{j=1}^{m} \sigma_j$. We note by $[A]_k$ the k-th rank approximation of A. We note $\mathcal{O}_{m,k} = \{M \in \mathbb{R}^{m \times k} | M^\top M = I_k\}$. Consider the IPM induced by this function set. Let $\mathbb{P}, \mathbb{Q} \in \mathscr{P}(\mathcal{X})$ we have:

$$d_{\mathscr{F}_{U,V,\omega}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathscr{F}_{U,V,\omega}} \underset{x \sim \mathbb{P}}{\mathbb{E}} f(x) - \underset{x \sim \mathbb{Q}}{\mathbb{E}} f(x)$$

$$= \max_{\substack{\omega \in \Omega \\ U,V \in \mathcal{O}_{m,k}}} \underset{x \sim \mathbb{P}}{\mathbb{E}} f(x) - \underset{x \sim \mathbb{Q}}{\mathbb{E}} f(x)$$

$$= \max_{\omega \in \Omega} \max_{U,V \in \mathcal{O}_{m,k}} Trace \left[U^\top (\Sigma_\omega(\mathbb{P}) - \Sigma_\omega(\mathbb{Q}))V \right]$$

$$= \max_{\omega \in \Omega} \sum_{j=1}^{k} \sigma_j \left(\Sigma_\omega(\mathbb{P}) - \Sigma_\omega(\mathbb{Q}) \right)$$

$$= \max_{\omega \in \Omega} \| [\Sigma_\omega(\mathbb{P}) - \Sigma_\omega(\mathbb{Q})]_k \|_* ,$$

where we used the variational definition of singular values and the definition of the nuclear norm. Note that U, V are the left and right singular vectors of $\Sigma_\omega(\mathbb{P}) - \Sigma_\omega(\mathbb{Q})$. Hence $d_{\mathscr{F}_{U,V,\omega}}$ measures the worst case distance between the covariance feature embeddings of the two distributions, this distance is measured with the *Ky Fan k-norm* (nuclear norm of truncated covariance difference). Hence we call this IPM covariance matching IPM, IPM$_\Sigma$.

4.2. Covariance Matching GAN

Turning now to the problem of learning a generative model g_θ of $\mathbb{P}_r \in \mathscr{P}(\mathcal{X})$ using IPM$_\Sigma$ we shall solve:

$$\min_{g_\theta} d_{\mathscr{F}_{U,V,\omega}}(\mathbb{P}_r, \mathbb{P}_\theta),$$

this has the following primal formulation:

$$\min_{g_\theta} \max_{\omega \in \Omega, U,V \in \mathcal{O}_{m,k}} \mathscr{L}_\sigma(U, V, \omega, \theta), \qquad (6)$$

where $\mathscr{L}_\sigma(U, V, \omega, \theta) = \underset{x \sim \mathbb{P}_r}{\mathbb{E}} \langle U^\top \Phi_\omega(x), V^\top \Phi_\omega(x)) \rangle$

$$- \underset{z \sim p_z}{\mathbb{E}} \langle U^\top \Phi_\omega(g_\theta(z)), V^\top \Phi_\omega(g_\theta(z)) \rangle,$$

or equivalently the following dual formulation:

$$\min_{g_\theta} \max_{\omega \in \Omega} \| [\Sigma_\omega(\mathbb{P}_r) - \Sigma_\omega(\mathbb{P}_\theta)]_k \|_* , \qquad (7)$$

where $\Sigma_\omega(\mathbb{P}_\theta) = \mathbb{E}_{z \sim p_z} \Phi_\omega(g_\theta(z)) \Phi_\omega(g_\theta(z))^\top$.

The dual formulation in Equation (7) shows that learning generative models with IPM$_\Sigma$, consists in an adversarial game between the feature map and the generator, when the feature maps tries to maximize the distance between the feature covariance embeddings of the distributions, the generator tries to minimize this distance. Hence we call learning with IPM$_\Sigma$, covariance matching GAN.

We give here an empirical estimate of the primal formulation in Equation (6) which is amenable to stochastic gradient. The dual requires nuclear norm minimization and is more involved. Given $\{x_i, x_i \sim \mathbb{P}_r\}$, and $\{z_j, z_j \sim p_z\}$, the *covariance matching GAN* can be written as follows:

$$\min_{g_\theta} \max_{\omega \in \Omega, U,V \in \mathcal{O}_{m,k}} \hat{\mathscr{L}}_\sigma(U, V, \omega, \theta), \qquad (8)$$

where $\hat{\mathscr{L}}_\sigma(U, V, \omega, \theta) = \frac{1}{N} \sum_{i=1}^{N} \langle U^\top \Phi_\omega(x_i), V^\top \Phi_\omega(x_i) \rangle$

$$- \frac{1}{N} \sum_{j=1}^{N} \langle U^\top \Phi_\omega(g_\theta(z_j)), V^\top \Phi_\omega(g_\theta(z_j)) \rangle.$$

4.3. Mean and Covariance Matching GAN

In order to match first and second order statistics we propose the following simple extension:

$$\min_{g_\theta} \max_{\substack{\omega \in \Omega, v, \|v\|_p \leq 1 \\ U,V \in \mathcal{O}_{m,k}}} \mathscr{L}_\mu(v, \omega, \theta) + \mathscr{L}_\sigma(U, V, \omega, \theta),$$

that has a simple dual adversarial game interpretation

$$\min_{g_\theta} \max_{\omega \in \Omega} \| \mu_\omega(\mathbb{P}) - \mu_\omega(\mathbb{P}_\theta) \|_q + \| [\Sigma_\omega(\mathbb{P}_r) - \Sigma_\omega(\mathbb{P}_\theta)]_k \|_* ,$$

where the discriminator finds a feature space that discriminates between means and variances of real and fake, and the generator tries to match the real statistics. We can also give empirical estimates of the primal formulation similar to expressions given in the paper.

5. Algorithms

We present in this Section our algorithms for mean and covariance feature matching GAN (McGan) with $\text{IPM}_{\mu,q}$ and IPM_Σ.

Mean Matching GAN. *Primal P_μ:* We give in Algorithm 1 an algorithm for solving the primal $\text{IPM}_{\mu,q}$ GAN (P_μ). Algorithm 1 is adapted from (Arjovsky et al., 2017) and corresponds to their algorithm for $p = \infty$. The main difference is that we allow projection of v on different ℓ_p balls, and we maintain the clipping of ω to ensure boundedness of Φ_ω. For example for $p = 2$, $\text{proj}_{B_{\ell_2}}(v) = \min(1, \frac{1}{\|v\|_2})v$. For $p = \infty$ we obtain the same clipping in (Arjovsky et al., 2017) $\text{proj}_{B_{\ell_\infty}}(v) = \text{clip}(v, -c, c)$ for $c = 1$.

Dual D_μ: We give in Algorithm 2 an algorithm for solving the dual formulation $\text{IPM}_{\mu,q}$ GAN (D_μ). As mentioned earlier we need samples from "real" and "fake" for training both generator and the "critic" feature space.

Covariance Matching GAN. *Primal P_Σ:* We give in Algorithm 3 an algorithm for solving the primal of IPM_Σ GAN (Equation (8)). The algorithm performs a stochastic gradient ascent on (ω, U, V) and a descent on θ. We maintain clipping on ω to ensure boundedness of Φ_ω, and perform a QR retraction on the Stiefel manifold $\mathcal{O}_{m,k}$ (Absil et al., 2007), maintaining orthonormality of U and V.

Algorithm 1 Mean Matching GAN - Primal (P_μ)

Input: p to define the ball of v , η Learning rate, n_c number of iterations for training the critic, c clipping or weight decay parameter, N batch size
Initialize v, ω, θ
repeat
  for $j = 1$ **to** n_c **do**
    Sample a minibatch $x_i, i = 1 \ldots N, x_i \sim \mathbb{P}_r$
    Sample a minibatch $z_i, i = 1 \ldots N, z_i \sim p_z$
    $(g_v, g_\omega) \leftarrow (\nabla_v \hat{\mathscr{L}}_\mu(v, \omega, \theta), \nabla_\omega \hat{\mathscr{L}}_\mu(v, \omega, \theta))$
    $(v, \omega) \leftarrow (v, \omega) + \eta \text{ RMSProp} ((v, \omega), (g_v, g_\omega))$
    {Project v on ℓ_p ball, $B_{\ell_p} = \{x, \|x\|_p \leq 1\}$}
    $v \leftarrow \text{proj}_{B_{\ell_p}}(v)$
    $\omega \leftarrow \text{clip}(\omega, -c, c)$ {Ensure Φ_ω is bounded}
  end for
  Sample $z_i, i = 1 \ldots N, z_i \sim p_z$
  $d_\theta \leftarrow -\nabla_\theta \left\langle v, \frac{1}{N} \sum_{i=1}^{N} \Phi_\omega(g_\theta(z_i)) \right\rangle$
  $\theta \leftarrow \theta - \eta \text{ RMSProp} (\theta, d_\theta)$
until θ converges

Algorithm 2 Mean Matching GAN - Dual (D_μ)

Input: q the matching ℓ_q norm , η Learning rate, n_c number of iterations for training the critic, c clipping or weight decay parameter, N batch size
Initialize v, ω, θ
repeat
  for $j = 1$ **to** n_c **do**
    Sample a minibatch $x_i, i = 1 \ldots N, x_i \sim \mathbb{P}_r$
    Sample a minibatch $z_i, i = 1 \ldots N, z_i \sim p_z$
    $\Delta_{\omega,\theta} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \Phi_\omega(x_i) - \frac{1}{N} \sum_{i=1}^{N} \Phi_\omega(g_\theta(z_i))$
    $g_\omega \leftarrow \nabla_\omega \|\Delta_{\omega,\theta}\|_q$
    $\omega \leftarrow \omega + \eta \text{ RMSProp} (\omega, g_\omega)$
    $\omega \leftarrow \text{clip}(\omega, -c, c)$ {Ensure Φ_ω is bounded}
  end for
  Sample $z_i, i = 1 \ldots N, z_i \sim p_z$
  Sample $x_i, i = 1 \ldots M, x_i \sim \mathbb{P}_r \; (M > N)$
  $\Delta_{\omega,\theta} \leftarrow \frac{1}{M} \sum_{i=1}^{M} \Phi_\omega(x_i) - \frac{1}{N} \sum_{i=1}^{N} \Phi_\omega(g_\theta(z_i))$
  $d_\theta \leftarrow \nabla_\theta \|\Delta_{\omega,\theta}\|_q$
  $\theta \leftarrow \theta - \eta \text{ RMSProp} (\theta, d_\theta)$
until θ converges

Algorithm 3 Covariance Matching GAN - Primal (P_Σ)

Input: k the number of components , η Learning rate, n_c number of iterations for training the critic, c clipping or weight decay parameter, N batch size
Initialize U, V, ω, θ
repeat
  for $j = 1$ **to** n_c **do**
    Sample a minibatch $x_i, i = 1 \ldots N, x_i \sim \mathbb{P}_r$
    Sample a minibatch $z_i, i = 1 \ldots N, z_i \sim p_z$
    $G \leftarrow (\nabla_U, \nabla_V, \nabla_\omega) \hat{\mathscr{L}}_\sigma(U, V, \omega, \theta)$
    $(U, V, \omega) \leftarrow (U, V, \omega) + \eta \text{ RMSProp} ((U, V, \omega), G)$
    { Project U and V on the Stiefel manifold $\mathcal{O}_{m,k}$ }
    $Q_u, R_u \leftarrow QR(U) \;\; s_u \leftarrow \text{sign}(\text{diag}(R_u))$
    $Q_v, R_v \leftarrow QR(V) \;\; s_v \leftarrow \text{sign}(\text{diag}(R_v))$
    $U \leftarrow Q_u \text{Diag}(s_u)$
    $V \leftarrow Q_v \text{Diag}(s_v)$
    $\omega \leftarrow \text{clip}(\omega, -c, c)$ {Ensure Φ_ω is bounded}
  end for
  Sample $z_i, i = 1 \ldots N, z_i \sim p_z$
  $d_\theta \leftarrow -\nabla_\theta \frac{1}{N} \sum_{j=1}^{N} \langle U \Phi_\omega(g_\theta(z_j)), V \Phi_\omega(g_\theta(z_j)) \rangle$
  $\theta \leftarrow \theta - \eta \text{ RMSProp} (\theta, d_\theta)$
until θ converges

---

6. Experiments

We train McGan for image generation with both Mean Matching and Covariance Matching objectives. We show generated images on the labeled faces in the wild (lfw) (Huang et al., 2007), LSUN bedrooms (Yu et al., 2015), and cifar-10 (Krizhevsky & Hinton, 2009) datasets.

It is well-established that evaluating generative models is hard (Theis et al., 2016). Many GAN papers rely on a combination of samples for quality evaluation, supplemented by a number of heuristic quantitative measures. We will mostly focus on training stability by showing plots of the

loss function, and will provide generated samples to claim comparable sample quality between methods, but we will avoid claiming better sample quality. These samples are all generated at random and are not cherry-picked.

The design of g_θ and Φ_ω are following DCGAN principles (Radford et al., 2015), with both g_θ and Φ_ω being a convolutional network with batch normalization (Ioffe & Szegedy, 2015) and ReLU activations. Φ_ω has output size $bs \times F \times 4 \times 4$. The inner product can then equivalently be implemented as `conv(4x4, F->1)` or `flatten + Linear(4*4*F -> 1)`. We generate 64×64 images for lfw and LSUN and 32×32 images on cifar, and train with minibatches of size 64. We follow the experimental framework and implementation of (Arjovsky et al., 2017), where we ensure the boundedness of Φ_ω by clipping the weights pointwise to the range $[-0.01, 0.01]$.

Primal versus dual form of mean matching. To illustrate the validity of both the primal and dual formulation, we trained mean matching GANs both in the primal and dual form, see respectively Algorithm 1 and 2. Samples are shown in Figure 2. Note that optimizing the dual form is less efficient and only feasible for mean matching, not for covariance matching. The primal formulation of $\text{IPM}_{\mu,1}$ GAN corresponds to clipping v, i.e. the original WGAN, while for $\text{IPM}_{\mu,2}$ we divide v by its ℓ_2 norm if it becomes larger than 1. In the dual, for $q = 2$ we noticed little difference between maximizing the ℓ_2 norm or its square.

We observed that the default learning rates from WGAN (5e-5) are optimal for both primal and dual formulation. Figure 3 shows the loss (i.e. IPM estimate) dropping steadily for both the primal and dual formulation independently of the choice of the ℓ_q norm. We also observed that during the whole training process, samples generated from the same noise vector across iterations, remain similar in nature (face identity, bedroom style), while details and background will evolve. This qualitative observation indicates valuable stability of the training process.

For the dual formulation (Algorithm 2), we confirmed the hypothesis that we need a good estimate of $\mu_\omega(\mathbb{P}_r)$ in order to compute the gradient of the generator ∇_θ: we needed to increase the minibatch size of real threefold to 3×64.

Covariance GAN. We now experimentally investigate the IPM defined by covariance matching. For this section and the following, we use only the primal formulation, i.e. with explicit u_j and v_j orthonormal (Algorithm 3). Figure 4 and 5 show samples and loss from lfw and LSUN training respectively. We use Algorithm 3 with $k = 16$ components. We obtain samples of comparable quality to the mean matching formulations (Figure 2), and we found training to be stable independent of hyperparameters like number of components k varying between 4 and 64.
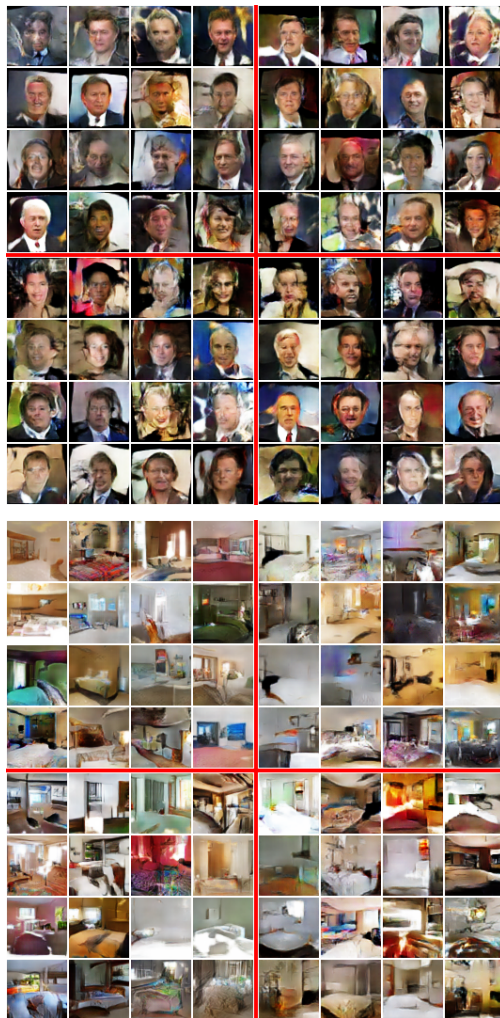

Figure 2. Samples generated with primal (left) and dual (right) formulation, in ℓ_1 (top) and ℓ_2 (bottom) norm. (A) lfw (B) LSUN.
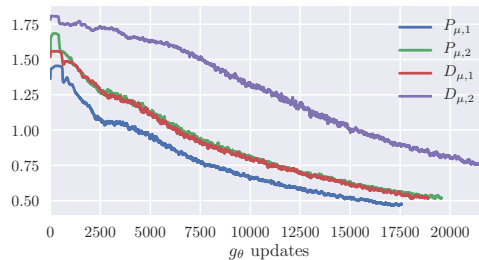

Figure 3. Plot of the loss of $P_{\mu,1}$ (i.e. WGAN), $P_{\mu,2}$ $D_{\mu,1}$ $D_{\mu,2}$ during training of lfw, as a function of number of updates to g_θ. Similar to the observation in (Arjovsky et al., 2017), training is stable and the loss is a useful metric of progress, across the different formulations.
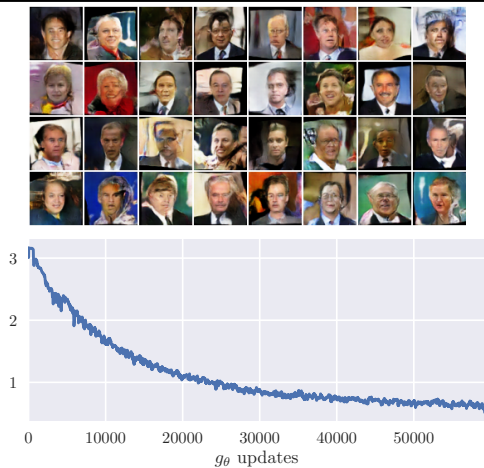
Figure 4. lfw samples generated with covariance matching and plot of loss function (IPM estimate) $\hat{\mathscr{L}}_\sigma(U, V, \omega, \theta)$.
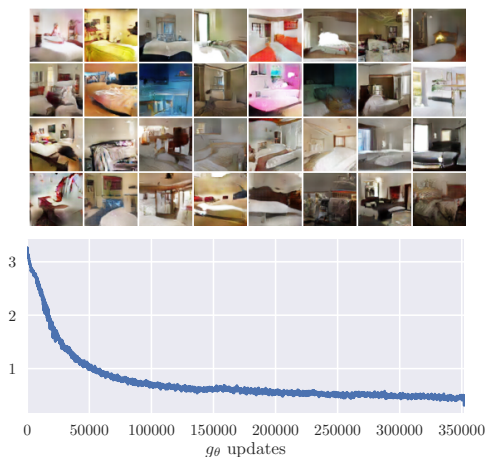

Figure 5. LSUN samples generated with covariance matching and plot of loss function (IPM estimate) $\hat{\mathscr{L}}_\sigma(U, V, \omega, \theta)$.

Covariance GAN with labels and conditioning.

Finally, we conduct experiments on the cifar-10 dataset, where we will leverage the additional label information by training a GAN with conditional generator $g_\theta(z, y)$ with label $y \in [1, K]$ suppplied as one-hot vector concatenated with noise z. Similar to Infogan (Chen et al., 2016) and AC-GAN (Odena et al., 2016), we add a new output layer, $S \in \mathbb{R}^{K \times m}$ and will write the logits $\langle S, \Phi_\omega(x) \rangle \in \mathbb{R}^K$. We now optimize a combination of the IPM loss and the cross-entropy loss $CE(x, y; S, \Phi_\omega) = -\log [\text{Softmax}(\langle S, \Phi_\omega(x) \rangle)_y]$. The critic loss becomes $\mathscr{L}_D = \hat{\mathscr{L}}_\sigma - \lambda_D \frac{1}{N} \sum_{(x_i, y_i) \in \text{lab}} CE(x_i, y_i; S, \Phi_\omega)$, with hyper-parameter λ_D. We now sample three minibatches for each critic update: a labeled batch for the CE term, and for the IPM a real unlabeled + generated batch.

The generator loss (with hyper-param λ_G) becomes: $\mathscr{L}_G = \hat{\mathscr{L}}_\sigma + \lambda_G \frac{1}{N} \sum_{z_i \sim p_z, y_i \sim p_y} CE(g_\theta(z_i, y_i), y_i; S, \Phi_\omega)$ which still only requires a single minibatch to compute.
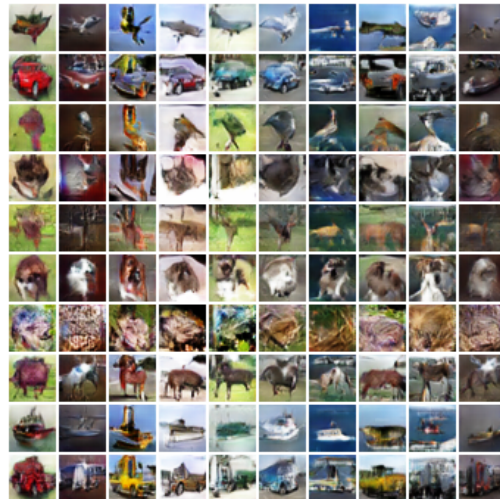

Figure 6. Cifar-10: Class-conditioned generated samples. Within each column, the random noise z is shared, while within the rows the GAN is conditioned on the same class: from top to bottom *airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck.*

Table 1. Cifar-10: inception score of our models and baselines.

	Cond (+L)	Uncond (+L)	Uncond (-L)
L1+Sigma	7.11 ± 0.04	6.93 ± 0.07	6.42 ± 0.09
L2+Sigma	7.27 ± 0.04	6.69 ± 0.08	6.35 ± 0.04
Sigma	$\mathbf{7.29 \pm 0.06}$	$\mathbf{6.97 \pm 0.10}$	$\mathbf{6.73 \pm 0.04}$
WGAN	3.24 ± 0.02	5.21 ± 0.07	6.39 ± 0.07
BEGAN (Berthelot et al., 2017)			5.62
Impr. GAN "-LS" (Salimans et al., 2016)		6.83 ± 0.06	
Impr. GAN Best (Salimans et al., 2016)		8.09 ± 0.07	

We confirm the improved stability and sample quality of objectives including covariance matching with inception scores (Salimans et al., 2016) in Table 1. Samples corresponding to the best inception score (Sigma) are given in Figure 6. Using the code released with WGAN (Arjovsky et al., 2017), these scores come from the DCGAN model with `n_extra_layers=3` (deeper generator and discriminator) . More samples are in appendix with combinations of Mean and Covariance Matching. Notice rows corresponding to recognizable classes, while the noise z (shared within each column) clearly determines other elements of the visual style like dominant color, across label conditioning.

7. Discussion

We noticed the influence of clipping on the capacity of the critic: a higher number of feature maps was needed to compensate for clipping. The question remains what alternatives to clipping of Φ_ω can ensure the boundedness. For example, we succesfully used an ℓ_2 penalty on the weights of Φ_ω. Other directions are to explore geodesic distances between the covariances (Arsigny et al., 2006), and extensions of the IPM framework to the multimodal setting (Isola et al., 2017).

References

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2007.

Arjovsky, Martin, Chintala, Soumith, and Bottou, Leon. Wasserstein gan. *ICML*, 2017.

Arsigny, Vincent, Fillard, Pierre, Pennec, Xavier, and Ayache, Nicholas. Log-euclidean metrics for fast and simple calculus on diffusion tensors. In *Magnetic Resonance in Medicine*, 2006.

Berthelot, David, Schumm, Tom, and Metz, Luke. Began: Boundary equilibrium generative adversarial networks. *arXiv:1703.10717*, 2017.

Chen, Xi, Duan, Yan, Houthooft, Rein, Schulman, John, Sutskever, Ilya, and Abbeel, Pieter. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

Dziugaite, Gintare Karolina, Roy, Daniel M., and Ghahramani, Zoubin. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *NIPS*. 2014.

Gretton, Arthur, Borgwardt, Karsten M., Rasch, Malte J., Schölkopf, Bernhard, and Smola, Alexander. A kernel two-sample test. *JMLR*, 2012.

Huang, Gary B., Ramesh, Manu, Berg, Tamara, and Learned-Miller, Erik. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, 2007.

Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proc. ICML*, 2015.

Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

Kingma, Diederik P. and Welling, Max. Auto-encoding variational bayes. *NIPS*, 2013.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Master's thesis*, 2009.

Li, Yujia, Swersky, Kevin, and Zemel, Richard. Generative moment matching networks. In *ICML*, 2015.

Mirza, Mehdi and Osindero, Simon. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

Muandet, Krikamol, Fukumizu, Kenji, Sriperumbudur, Bharath, and Schlkopf, Bernhard. Kernel mean embedding of distributions: A review and beyond. *arXiv:1605.09522*, 2016.

Muller, Alfred. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 1997.

Nowozin, Sebastian, Cseke, Botond, and Tomioka, Ryota. f-gan: Training generative neural samplers using variational divergence minimization. In *NIPS*, 2016.

Odena, Augustus, Olah, Christopher, and Shlens, Jonathon. Conditional image synthesis with auxiliary classifier gans. *arXiv:1610.09585*, 2016.

Radford, Alec, Metz, Luke, and Chintala, Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015.

Rahimi, Ali and Recht, Benjamin. Random features for large-scale kernel machines. In *NIPS*. 2008.

Salimans, Tim, Goodfellow, Ian, Zaremba, Wojciech, Cheung, Vicki, Radford, Alec, Chen, Xi, and Chen, Xi. Improved techniques for training gans. In *NIPS*. 2016.

Sriperumbudur, Bharath K., Fukumizu, Kenji, Gretton, Arthur, Schlkopf, Bernhard, and Lanckriet, Gert R. G. On integral probability metrics, phi -divergences and binary classification, 2009.

Sriperumbudur, Bharath K., Fukumizu, Kenji, Gretton, Arthur, Schlkopf, Bernhard, and Lanckriet, Gert R. G. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 2012.

Theis, Lucas, Oord, Aäron van den, and Bethge, Matthias. A note on the evaluation of generative models. *ICLR*, 2016.

Yu, Fisher, Zhang, Yinda, Song, Shuran, Seff, Ari, and Xiao, Jianxiong. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv:1506.03365*, 2015.

Zhao, Junbo, Mathieu, Michael, and Lecun, Yann. Energy based generative adversarial networks. *ICLR*, 2017.