χ^2 Generative Adversarial Network

Chenyang Tao [1] **Liqun Chen** [1] **Ricardo Henao** [1] **Jianfeng Feng** [2] **Lawrence Carin** [1]

Abstract

To assess the difference between real and synthetic data, Generative Adversarial Networks (GANs) are trained using a distribution discrepancy measure. Three widely employed measures are information-theoretic divergences, integral probability metrics, and Hilbert space discrepancy metrics. We elucidate the theoretical connections between these three popular GAN training criteria and propose a novel procedure, called χ^2-GAN, that is conceptually simple, stable at training and resistant to mode collapse. Our procedure naturally generalizes to address the problem of simultaneous matching of multiple distributions. Further, we propose a resampling strategy that significantly improves sample quality, by repurposing the trained critic function via an importance weighting mechanism. Experiments show that the proposed procedure improves stability and convergence, and yields state-of-art results on a wide range of generative modeling tasks.

1. Introduction

Learning to sample from complicated distributions has attracted considerable recent interest, with many important applications (Zhu et al., 2017; Ledig et al., 2017; Yu et al., 2017; Hu et al., 2017). Likelihood-free models avoid the need to explicitly assume a particular parametrization of the data-generating distribution $p_G(x)$. Such models implicitly define a distribution via *generator* $G(z;\theta) : \mathcal{Z} \to \mathcal{X}$, and a latent random variable Z with pre-specified distribution $q(z)$. Samples from the generator are produced by first drawing $Z \sim q(z)$ and then feeding it through the generator.

To match $p_G(x)$ to the true data distribution $p_d(x)$, one estimates a discrepancy measure, $d(p_d, p_G)$. In the GAN framework, the discrepancy is first estimated by maximizing an auxiliary variational functional $V(p_d, p_G; D) : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ between distributions $p_d(x)$ and $p_G(x)$ satisfying $d(p_d, p_G) = \max_D V(p_d, p_G; D)$, where $\mathcal{P}$ is the space of probability distributions and $V(p_d, p_G; D)$ is estimated using samples from the two distributions. Function $D(x;\omega)$, parameterized by ω and known as the *critic* function, is intended to maximally discriminate between samples of the two distributions. One seeks to match the generator distribution $p_G(x)$ to the unknown true distribution $p_d(x)$ by solving a minimax game between the critic and the generator: $\min_G \max_D V(p_d, p_G; D)$.

Following ideas from the original GAN (Goodfellow et al., 2014), which optimizes the Jensen-Shannon divergence, much recent work has focused on information-theoretic divergences, such as the KL-divergence (Sønderby et al., 2017). Many other studies have investigated the generalized f-divergence (Csiszár, 1963), $\text{Div}_f(p_d \parallel p_G) \triangleq \int f\left(\frac{p_d(x)}{p_G(x)}\right) p_G(x)\, \mathrm{d}x$, where $f(\cdot) : \mathbb{R} \to \mathbb{R}$ is a convex function satisfying $f(1) = 0$, that summarizes the local discrepancy between $p_d(x)$ and $p_G(x)$. Nowozin et al. (2016) proposed an algorithm based on the variational formulation of $\text{Div}_f(p_d \parallel p_G)$, Uehara et al. (2016) explored in depth its density-ratio formulation, and Nock et al. (2017) further generalized it from an information-geometric perspective. Interestingly, Mao et al. (2017) showed that a specific type of f-divergence, namely the χ^2-divergence, can be directly optimized for GAN learning, by recasting it as a least-squares regression problem.

However, Arjovsky & Bottou (2017) showed that when using divergence-based objectives, the parameter updates for the generator can be either uninformative or numerically unstable, and divergence-based objectives may not be continuous wrt the generator parameters. These issues motivated development of GAN formulations based on Integral Probability Metrics (IPMs) (Müller, 1997). IPM models seek to optimize an objective of the form $V_{\text{IPM}}(p_d, p_G; D) = \mathbb{E}_{X\sim p_d}[D(X;\omega)] - \mathbb{E}_{X'\sim p_G}[D(X';\omega)]$, where $\mathbb{E}_{X\sim p_d}[\cdot]$ denotes the expectation wrt to distributions $p_d(x)$. When the critic $D(x;\omega)$ is chosen from a unit ball of Lipschitz-1 functions ($\|D\|_{\text{Lip}} \leq 1$), the IPM reduces to the Wasserstein-1 or earth mover's distance (Rubner et al., 2000). In this case, the challenge with $d_{\text{IPM}}(p_d, p_G)$ lies in constraining the critic function Lipschitz constant (Arjovsky et al., 2017; Gulrajani et al., 2017; Miyato et al., 2017).

[1]Electrical & Computer Engineering, Duke University, Durham, NC 27708, USA [2]ISTBI, Fudan University, Shanghai, China. Correspondence to: Chenyang Tao <chenyang.tao@duke.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Separately, Reproducing Kernel Hilbert Space (RKHS) theory has motivated development of a powerful set of methods to handle probability problems (Muandet et al., 2017). In particular, the embedding of probability measures via kernels (Sriperumbudur et al., 2010) has attracted significant interest. Let $\kappa(\cdot,\cdot)$ be a positive definite function known as the *kernel function*. The kernel embedding of distribution $p(x)$ is given by $\nu_p(x) \triangleq \mathbb{E}_{X\sim p}[\kappa(x,X)]$. The Maximal Mean Discrepancy $\text{MMD}(p_d,p_G) \triangleq \|\nu_{p_d}-\nu_{p_G}\|_{\mathcal{H}}$, defines a distance metric on distributions $p_d(x)$ and $p_G(x)$, where $\|\cdot\|_{\mathcal{H}}$ is the kernel-induced norm in $\mathcal{H}$, a Hilbert space. MMD readily translates into an algorithm that does not require the adversarial game for generative modeling (Li et al., 2015; Dziugaite et al., 2015). However, RKHS-based generative models have high computational cost, while in the case of generative models, also struggling when dealing with complex distributions (Bińkowski et al., 2018). In practice, good performance can be achieved with careful hyperparameter tuning and by introducing auxiliary loss terms to the objective (Zhang et al., 2017; Li et al., 2017b).

From a pragmatic perspective, GANs rarely converge to the desired equilibrium (Arora & Zhang, 2018), instead settling for a sub-optimal local solution, where samples produced by the trained generative model often lack diversity (Salimans et al., 2016). To alleviate these issues, most existing studies focus on seeking more stable architectures (Radford et al., 2016), enforcing heuristically-derived or theoretically-inspired regularized objectives (Salimans et al., 2016; Warde-Farley & Bengio, 2017; Mescheder et al., 2017b; Roth et al., 2017), and procedures that leverage carefully designed optimization paths (Karras et al., 2018).

We present new theoretical insights on GAN-based generative modeling, the cause of some of its difficulties, and principled solutions to address associated challenges. Our key contributions include: (i) We present theory connecting three major generative modeling frameworks: divergence-, IPM- and kernel-based approaches. (ii) A novel, conceptually simple procedure is introduced, termed χ^2-GAN, that is stable at training and embraces sample diversity during generation. (iii) It is demonstrated that our formulation naturally generalizes to problems requiring simultaneous matching of multiple distributions. (iv) We propose to fully exploit the learned critic function, by repurposing it as a weighting mechanism in a resampling procedure, leveraging useful information from the critic to improve sample quality.

2. Learning χ^2 GANs

2.1. Distribution Mixture and Generative Modeling

Consider joint random variables (X,Y), where X is drawn from mixture distribution $[p(x)+q(x)]/2$, and Y is a random variable identifying the mixture component from which X is drawn; $Y=1$ if X is drawn from $p(x)$, with $Y=-1$ if X is drawn from $q(x)$. We denote the joint density for (X,Y) as $\mu(x,y;p,q)$; to avoid notational clutter, we often omit its dependency on (p,q) when the context is clear. Further, let $\mu(x)$ and $\mu(y)$ be the marginals of $\mu(x,y)$.

It can be readily verified that X and Y from $\mu(x,y)$ are statistically independent if and only if $p(x)$ and $q(x)$ match. For subsequent generative modeling, $p(x)$ is the true data distribution, $p_d(x)$, and $q(x)$ is the generator distribution, $p_G(x)$. We can therefore cast the problem of learning the parameters of a generative model $G(x;\theta)$ as seeking to match our generator distribution to that of the data, by minimizing the statistical dependency between the data variable, X, and its label, Y.

More generally, random variables (X,Y) are manifested with X drawn from mixture model $\frac{1}{M}\sum_{m=1}^{M} p_m(x)$ and with $Y\in\{1,\dots,M\}$ identifying the mixture component from which X is drawn. Extending the ideas discussed above, we can jointly match $M>2$ distributions $\{p_m(x)\}$, by minimizing the statistical dependency between data X and mixture-component label Y.

2.2. Covariance Operator and Statistical Dependency

Let $f(x):\mathcal{X}\rightarrow\mathbb{R}$ and $g(y):\mathcal{Y}\rightarrow\mathbb{R}$ be two square-integrable functions defined over the domains of X and Y, respectively. The covariance wrt the joint density $\mu(x,y)$ is

$$\text{cov}(f,g)=\mathbb{E}_{X,Y\sim\mu}[(f(X)-\mathbb{E}[f])(g(Y)-\mathbb{E}[g])],$$

where $\mathbb{E}[f]=\mathbb{E}_X[f(X)]$ is the marginal mean of $f(x)$, similarly defined for $\mathbb{E}[g]$. When X and Y are statistically independent, *i.e.* $\mu(x,y)=\mu(x)\mu(y)$, $\text{cov}(f,g)=0$ for every choice of $f(x)$ and $g(y)$. We show below that the statistical independency between X and Y is implied when $\text{cov}(f,g)=0$ holds for every $f(x)$ and $g(y)$ chosen from a sufficiently rich function space.

Let $\mathcal{H}_X$ and $\mathcal{H}_Y$ be two Hilbert spaces for functions defined on $\mathcal{X}$ and $\mathcal{Y}$, equipped with inner products $\langle\cdot,\cdot\rangle_{\mathcal{H}_X}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}_Y}$ respectively. In an RKHS, the kernel function defines the inner product, *i.e.*, $\langle\cdot,\cdot\rangle_{\mathcal{H}}=\kappa(\cdot,\cdot)$. The *cross-covariance operator* wrt the triplet $\{\mu(x,y),\mathcal{H}_X,\mathcal{H}_Y\}$ is defined as the operator $C_{XY}:\mathcal{H}_X\rightarrow\mathcal{H}_Y$ satisfying

$$\langle g,C_{XY}f\rangle_{\mathcal{H}_Y}=\text{cov}(f(X),g(Y)),$$

for all $f(x)\in\mathcal{H}_X$ and $g(y)\in\mathcal{H}_Y$. The existence of C_{XY} is a direct result of the Riesz representation theorem (Yoshida, 1974). One can further define the covariance operator $C_{XX}:\mathcal{H}_X\rightarrow\mathcal{H}_X$ as

$$\langle f_1,C_{XX}f_2\rangle_{\mathcal{H}_X}=\text{cov}(f_1(x),f_2(x)), \tag{1}$$

and similarly for C_{YY}.

Further assume $\mathcal{H}_X$ and $\mathcal{H}_Y$ are separable and let $\mathcal{E} = \{e_i\}_{i=1}^{\infty}$ and $\mathcal{F} = \{f_i\}_{i=1}^{\infty}$ be their respective Complete Orthonormal Systems (CONS). The *Hilbert-Schmidt norm* $\|\cdot\|_{\text{HS}}$ of an operator $\mathcal{A} : \mathcal{H}_X \to \mathcal{H}_Y$ is defined as

$$\|\mathcal{A}\|_{\text{HS}}^2 := \sum_{i,j=1}^{\infty} \langle \mathcal{A}e_i, f_j \rangle_{\mathcal{H}_Y}^2, \tag{2}$$

which is independent of the choice of the CONS (Adams & Fournier, 2003). It can be readily verified that for the cross-covariance operator $\|C_{XY}\|_{\text{HS}}^2 = 0$ implies $\text{cov}(f, g) = 0$ for $f(x) \in \mathcal{H}_X$ and $g(y) \in \mathcal{H}_Y$. To understand when does it imply independence, let us further denote the space of square integrable functions wrt random variable X as L_X^2. The following result, adapted from Theorem 4 of Gretton et al. (2005a), states that when $\mathcal{H}_X$ and $\mathcal{H}_Y$ are sufficiently rich, a vanishing $\|C_{XY}\|_{\text{HS}}^2$ implies the statistical independence of X and Y, and *vice versa*.

Theorem 1. *If $\mathcal{H}_X$ and $\mathcal{H}_Y$ are dense in L_X^2 and L_Y^2, respectively, then $\|C_{XY}\|_{HS}^2 = 0$ if and only if X and Y are statistically independent.*

Proofs for all theoretical results are found in the Supplementary Material (SM).

The cross-covariance operator norm $\|C_{XY}\|_{\text{HS}}^2$ in (2) is at the core of the Hilbert-Schmidt Independence Criteria (HSIC) (Gretton et al., 2012); we are interested in such criteria to assess the independence of X and Y discussed in Sec. 2.1. In an RKHS, let $\tilde{K}_X^{(n)}$ be the Gram matrix of samples $\{x_i\}_{i=1}^n$, whose entries are defined as $[\tilde{K}_X^{(n)}]_{ij} = \kappa(x_i, x_j)$, then the centralized Gram matrix is given by $K_X^{(n)} = H_n \tilde{K}_X^{(n)} H_n$, where $H_n \triangleq I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is the centralizing matrix, with a similar definition for $K_Y^{(n)}$; $\mathbf{1}_n$ is an n-dimensional vector of all ones, and I_n is the $n \times n$ identity matrix). The empirical estimator for HSIC results in an elegant population-wise expression:

$$\|\hat{C}_{XY}^{(n)}\|_{\text{HS}}^2 \triangleq \frac{1}{n^2}\text{tr}\left(K_X^{(n)} K_Y^{(n)}\right), \tag{3}$$

where $\text{tr}(\cdot)$ is the trace operation. Despite its simple mathematical expression, it is difficult to use HSIC as an optimization objective, because computations require quadratic time, $\mathcal{O}(n^2)$, and (3) depends on the inner product of the Hilbert space, which is not invariant to the kernel function space used. As a result, HSIC is known to be highly sensitive to the choice of the kernel Hilbert space.

2.3. Normalization of Cross-covariance Operator

To circumvent the dependence on the form of the inner product, we consider the *normalized cross-covariance operator* $V_{XY} : \mathcal{H}_X \to \mathcal{H}_Y$ (Baker, 1973), with slight abuse of notation, given by

$$V_{XY} = C_{YY}^{-1/2} C_{XY} C_{XX}^{-1/2}.$$

Intuitively, operators $C_{XX}^{-1/2}$ and $C_{YY}^{-1/2}$ defined as (3) via (1) normalize the covariances in function space, thus V_{XY} can be understood as the *cross-correlation operator*. The following Lemma formalizes this intuition.

Lemma 2. *For $\|f\|_{\mathcal{H}_X}^2 = \|g\|_{\mathcal{H}_Y}^2 = 1$, we have*

$$\langle g, V_{XY} f \rangle_{\mathcal{H}_Y} = corr\left(C_{XX}^{-1/2} f(X), C_{YY}^{-1/2} g(Y)\right).$$

The next two theoretical results expand on the invariance of $\|V_{XY}\|_{\text{HS}}^2$ to the choice of the function space.

Proposition 3. *If $\mathcal{E} = \{e_i\}_i^{\infty}$ is a complete orthonormal system in $\mathcal{H}_X$, then $C_{XX}^{-1/2}\mathcal{E} = \{C_{XX}^{-1/2} e_i\}_i^{\infty}$ is a complete orthonormal system in L_X^2.*

Corollary 4. *Let $\mathcal{E} = \{e_i\}_i^{\infty}$ and $\mathcal{F} = \{f_j\}_j^{\infty}$ be the respective CONS for $\mathcal{H}_X$ and $\mathcal{H}_Y$, we have*

$$\|V_{XY}\|_{HS}^2 = \sum_{i,j} \left\{ corr\left(C_{XX}^{-1/2} e_i, C_{YY}^{-1/2} f_j\right)\right\}^2. \tag{4}$$

The above result shows that $\|V_{XY}\|_{\text{HS}}^2$ only depends on $C_{XX}^{-1/2}\mathcal{E}$ and $C_{YY}^{-1/2}\mathcal{F}$, and not directly on the inner product of the Hilbert space. Further, in RKHS it is known that $\|V_{XY}\|_{\text{HS}}^2$ is invariant when the kernel functions are characteristic *e.g.*, Gaussian kernels.

Fukumizu et al. (2007) proposed a regularized empirical estimator for $\|V_{XY}\|_{\text{HS}}^2$ in (4), given by

$$\|\hat{V}_{XY}^{(n)}\|_{\text{HS}}^2 = \text{tr}(R_X^{(n)} R_Y^{(n)}), \tag{5}$$

where $R_X^{(n)} = K_X^{(n)}(K_X^{(n)} + \epsilon_n I_n)^{-1}$ (similarly defined for $R_Y^{(n)}$) and $\epsilon_n > 0$ is the regularization parameter.

Use of a metric like $\|\hat{V}_{XY}^{(n)}\|_{\text{HS}}^2$ appears promising as a means of assessing the independence of X and Y. However, in addition to the challenge of selecting kernels for the RKHSs, (5) comes with significant computational overhead; the matrix inversion in $R_X^{(n)}$ requires $\mathcal{O}(n^3)$ operations. While this estimator is shown to be consistent, provided $\epsilon_n \to 0$ and $\epsilon_n^3 n \to \infty$, empirical results indicate that ϵ_n must be carefully tuned to avoid degenerate solutions. Further, despite its theoretical invariance, empirical estimates vary significantly wrt the choice of kernels, while also not performing well in high-dimensional settings.

2.4. χ^2 Generative Adversarial Net

Key to circumvent the excessive computational burden of (5) is that rather than explicitly computing $C_{XX}^{-1/2}\mathcal{E}$ and

$C_{YY}^{-1/2}\mathcal{F}$ in (4), we can instead use some pre-specified function spaces $\tilde{\mathcal{E}}$ and $\tilde{\mathcal{F}}$ to estimate $\|V_{XY}\|_{\text{HS}}^2$. Next, we derive specific results on evaluating $\|V_{XY}\|_{\text{HS}}^2$ for the mixture distribution $\mu(x, y)$ defined in Sec. 2.1.

Let $\sigma_\mu^2(f)$ denote the variance of a function $f(x)$ wrt random variable X with marginal $\mu(x)$. Let $\phi(x) = \mathbb{E}[Y|X = x]$ be the conditional expectation of Y given X, $g_\chi(x) = \frac{\phi(x)}{\sigma_\mu(\phi)}$ is the variance normalized $\phi(x)$, and $\psi(x) = \Pr(Y = +1|X = x)$ is the conditional probability of $Y = +1$ given X, *i.e.*, the critic used in the original GAN, $D(x; \omega)$. The χ^2 mutual information of $\mu(x, y)$ is defined as

$$\text{MI}_\chi(\mu) = \iint \left(\frac{\mu(x, y)}{\mu(x)\mu(y)} - 1\right)^2 \mu(x)\mu(y)\,\mathrm{d}x\,\mathrm{d}y,$$

where $\mu(x)$ and $\mu(y)$ denote the marginals. The χ^2 distance between $p(x)$ and $q(x)$ is thus defined as

$$\text{Dis}_\chi(p, q) = \sqrt{\int \frac{(p(x) - q(x))^2}{\frac{p(x)+q(x)}{2}}\,\mathrm{d}x}.$$

The next proposition, connecting divergence, IPM and Hilbert spaced based discrepancies constitutes our main result.

Proposition 5. *The following quantities are identical:*

(i) $\|V_{YX}\|_{HS}^2$, *(ii)* $MI_\chi(\mu)$,
(iii) $\{Dis_\chi(p, q)\}^4/(16\|\phi(X)\|_\mu^2)$, *(iv)* $\{corr(\phi(X), Y)\}^2$,
(v) $\{corr(\psi(X), Y)\}^2$,
(vi) $(\mathbb{E}_{X_1\sim p}[g_\chi(X_1)] - \mathbb{E}_{X_2\sim q}[g_\chi(X_2)])^2/4$.

The equivalence between (i)-(iii) unveils the connections between the RKHS independence metric $\|V_{YX}\|_{\text{HS}}^2$, the information theoretic divergence metric $\text{MI}_\chi(\mu)$ and the variance-constrained IPM metric $\text{Dis}_\chi(p, q)$ (Mroueh & Sercu, 2017). On the other hand, (iv)-(vi) provide us with practical empirical estimators for these metrics. The key insight from Proposition 5 is that we only need to compute the critic $\psi(x)$ or $\phi(x)$ to estimate $\|V_{YX}\|_{\text{HS}}^2$, or equivalently, the χ^2 mutual information of the mixture distribution in Sec. 2.1. Consequently, to formulate an optimization procedure for generative modeling, we can minimize $\|V_{XY}\|_{\text{HS}}^2$ wrt the generator. We call the framework χ^2-GAN due to its inherent connection to the χ^2 metric.

We now detail the construction of χ^2-GAN based on estimator (iv) from Proposition 5. Let $p_G(x)$ be the data generating distribution implicitly defined by (a deep neural) generator $G(z; \theta)$ parametrized by θ, $Z \sim q(z)$, and let $\mathcal{D} = \{D(x; \omega); \omega \in \Omega\}$ be a parameterized family from which we choose the critic $D(x; \omega)$. The (least-squares) loss functions for the critic and generator are given by

$$L_D(\omega|\theta) = \mathbb{E}_{X,Y\sim\mu}[(D(X; \omega) - Y)^2],$$
$$L_G(\theta|\omega) = \left(\frac{\mathbb{E}_{X\sim p_d}[D(X; \omega)] - \mathbb{E}_{X\sim p_G}[D(X; \omega)]}{\sigma_\mu(D(X; \omega))}\right)^2.$$

where $\mu(x, y) = \frac{1}{2}p_d(x) + \frac{1}{2}p_G(x)$ as in Sec. 2.1. Note that $\mu(x, y)$ and $\sigma_\mu(\cdot)$ are implicitly dependent on θ via $p_G(x)$.

We further denote $\omega^*(\theta) \triangleq \arg\min_\omega L_D(\omega|\theta)$, the optimal critic parameters conditioned on the generator. When $\mathcal{D}$ is dense in L_X^2, we have

$$D(x; \omega^*(\theta)) = \phi(x), \text{ and } L_G(\theta|\omega^*(\theta)) = 4\|V_{XY}\|_{\text{HS}}^2,$$

which follows from Proposition 5. To match $p_G(x)$ to $p_d(x)$, we solve $\min_\theta\ L_G(\theta|\omega^*(\theta))$. We propose to decouple the above optimization scheme into the following GAN-like iterations:

$$\omega_t \leftarrow \arg\min_\omega L_D(\omega|\theta_{t-1}),$$
$$\theta_t \leftarrow \arg\min_{\theta\in\Delta(\theta_{t-1})} L_G(\theta|\omega_t, \theta_{t-1}),$$

where $\Delta(\theta_{t-1})$ denotes the trust region for the generator update. In $L_G(\theta|\omega_t, \theta_{t-1})$, we have replaced $\sigma_\mu(D(x; \omega))$ with its stale estimate $\sigma_{\mu_{t-1}}(D(x; \omega))$. We regularize the update of the generator to make sure $L_G(\theta|\omega_t)$ remains a good approximation to $\|V_{XY}\|_{\text{HS}}^2$. This can be implemented with proximal gradient descent, or simply via gradient clipping as summarized in Algorithm 1.

Algorithm 1 χ^2 GAN.

Input: data $\{x_i\}$, batchsize b, decay ρ, learning rate δ.
for $t = 1, 2, 3, \ldots$ **do**
1. Sample minibatch $\{x_i \sim p_d(x), z_i \sim p(z)\}_{i=1}^b$
2. Update the critic $D(x; \omega)$ to improve (minimize) $\sum_i (D(x_i; \omega) - 1)^2 + (D(G(z_i; \theta_{t-1}); \omega) + 1)^2$
3. Update the variance σ_μ^2 for $D(x; \omega_t)$
4. Update the correlation estimate $c_t(\theta_{t-1}) = (1 - \rho)c_{t-1} + \frac{\rho}{2b}\sum_i \left(\frac{D(x_i;\omega_t) - D(G(z_i;\theta_{t-1});\omega_t)}{\sigma_\mu(D(x;\omega_t))}\right)$
5. Update the generator $G(z; \theta)$ $\theta_t = \theta_{t-1} + \delta\,\text{GradClip}\left(\nabla_\theta(c_t(\theta_{t-1}))^2\right)$

end for

2.5. Joint Matching of Multiple Distributions

We now discuss the generalization to multi-component mixtures. Since the M components in $\frac{1}{M}\sum_{m=1}^M p_m(x)$ are mutually exclusive, the space of L_Y^2 has dimension $M - 1$. We use the following $M - 1$ empirical basis functions to

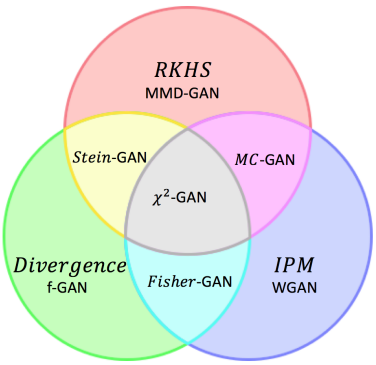

Figure 1. Relations between likelihood-free generative models.

span L^2_Y:

$$\gamma_m(y_l) = \begin{cases} 1, & m = l \\ -(M-1)^{-1}, & m \neq l \end{cases},$$

for $m \in \{1, \ldots, M-1\}$ and $l \in \{1, \ldots, M\}$. We collect $\{\gamma_m(y)\}_{m=1}^{M-1}$ into a vector function $\Gamma(y) = [\gamma_1(y), \cdots, \gamma_{M-1}(y)]^T$, with corresponding $(M-1) \times (M-1)$ covariance matrix Σ_Y with elements $[\Sigma_Y]_{ij} = 1/(M-1)$ if $i = j$ and $[\Sigma_Y]_{ij} = -1/(M-1)^2$ otherwise. Note that $\mathbb{E}_Y[\Gamma(Y)] = 0$, and this construction exactly recovers our $Y \in \{\pm 1\}$ binary labeling when $M = 2$. For the data side, we use critics $\psi_m(x) = \Pr(Y = y_m | X = x)$ as the empirical basis, and similarly use a compact $M-1$ dimensional vector representation $\tilde{\Psi}(x) = [\psi_1(x), \ldots, \psi_{(M-1)}(x)]^T$. We use $\Psi(x) = \tilde{\Psi}(x) - \mathbb{E}_X[\tilde{\Psi}(X)]$ for mean centering and denote its covariance matrix as Σ_X. The objective for multi-distribution matching is then

$$V_{\chi^2} = \left\| \mathbb{E}_{X,Y \sim \mu} \left[\Sigma_X^{-1/2} \Psi(X)^T \Gamma(Y) \Sigma_Y^{-1/2} \right] \right\|^2_{\text{Fro}},$$

where $\|\cdot\|_{\text{Fro}}$ denotes the Frobenius norm. In practice, we use cross-entropy loss to estimate $\tilde{\Psi}(x)$, and leverage a moving average estimator to track the expectation and covariance of $\Psi(x)$, otherwise it is similar to Algorithm 1. The complete algorithm and additional remarks are found in the SM.

2.6. Importance Resampling

Current practice in modeling with GANs discards the critic after training, with the learned generator used for sampling. However, the generator distribution rarely reaches the desired target distribution on real-world complex datasets, while the trained critic contains useful local information that does not get incorporated into the generator during training. Consider two cases: i) the generator does not have enough capacity to characterize the target distribution, and

Algorithm 2 Importance resampling.

Input: Generator $G(z;\theta)$, critic $D(x;\omega)$, sample size n.

1. Sample candidates: $\{x_i = G(z_i;\theta)\}_{i=1}^n$, $z_i \sim p(z)$.
2. Compute importance weights:

$$w_i = \tilde{w}_i (\textstyle\sum_{j=1}^n \tilde{w}_j)^{-1},$$

where $\tilde{w}_i = \zeta(D(x_i;\omega))$.

3. Sample $j \sim \text{Cat}(w_1, \ldots, w_n)$.

Return: $x_j = G(z_j;\theta)$

approaches a solution that covers the support of the target without properly capturing its topology; ii) the target has disjoint support regions and the generator covers each of them, but with inconsistent amounts of probability mass. In either case, as described below, the critic can be repurposed to provide additional information to assist sample generation (after training).

To fully harness the information from the critic to improve sample quality, we propose to resample the generator. Recall that in importance sampling, one uses a proposal distribution $q(x)$ to sample, and then one reweights samples by their importance weights $w(x) \triangleq p(x)/q(x)$, to compute the expectation wrt to the target distribution $p(x)$, *i.e.*,

$$\mathbb{E}_{X \sim p}[f(X)] = \mathbb{E}_{X \sim q}[w(X) f(X)] \approx \frac{1}{n} \sum_{i=1}^n \frac{p(x_i)}{q(x_i)} f(x_i),$$

where $\{x_i\}_{i=1}^n$ are iid samples from $q(x)$. We propose treating the data-generating distribution $p_G(x)$ as the proposal distribution, and collecting the importance weights $w(x) = p_d(x)/p_G(x)$ from the critic. For divergence-based generative models (f-GAN, χ^2-GAN, *etc.*), importance weights can often be directly computed from the critic via some simple transformation $\zeta(\cdot)$ (see SM for details). For other generative models, an auxiliary log-density ratio critic can be trained with cross-entropy loss to track the importance weights. We summarize the importance resampling procedure in Algorithm 2.

3. Related Work

The proposed χ^2 GAN connects three popular likelihood-free generative modeling frameworks (see Figure 1). It is derived from the theory of RKHS independence analysis and it can be shown that the popular MMD objective (Gretton et al., 2012) is an unnormalized version of our χ^2 objective (see SM). χ^2 GAN optimizes a divergence criteria with an IPM loss, using a critic trained with a stable least-squares loss, similar to that of LS-GAN (Mao et al., 2017). Regular divergence-based GANs directly optimize the divergence between $p_d(x)$ and $p_G(x)$, while χ^2-GAN instead optimizes the divergence between $\mu(x, y)$ and $\mu(x)\mu(y)$, which characterizes the independence between the sample

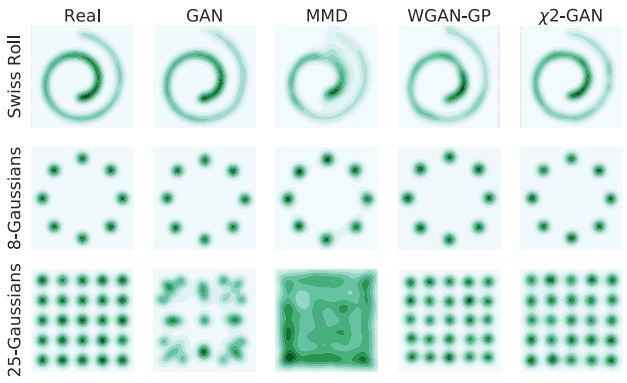

Figure 2. Toy model comparison for GAN, MMD-GAN, WGAN-GP and χ^2-GAN. Distributions are visualized with KDE plots.

Table 1. Quantitative results on MNIST. † is estimated using AIS; ‡ is reported in (Hu et al., 2018); § are likelihood-based models.

Model	$\log p(x) \geq$	IS
NF§ (Rezende & Mohamed, 2015)	-85.1	-
PixelRNN§ (Oord et al., 2016)	-79.2	-
AVB§ (Mescheder et al., 2017a)	-79.5	-
ASVAE§ (Pu et al., 2017)	-81.14	-
sVAE-r§ (Pu et al., 2018a)	-79.26 †	9.12
GAN (Goodfellow et al., 2014)	-114.25 †	8.34 ‡
WGAN-GP (Gulrajani et al., 2017)	-79.92 †	8.45 ‡
DCGAN (Radford et al., 2016)	-79.47 †	8.93
χ^2 GAN (ours)	-78.85 †	9.01

and associated label, *i.e.* the mutual information. This allows easy generalization beyond matching two distributions. Fisher GAN (Mroueh & Sercu, 2017) is closest to our χ^2-GAN, which builds on the IPM framework and like ours, also normalizes the critic with its second moment. The key differences are that Fisher GAN relies on a more sophisticated augmented Lagrangian to optimize the same objective for both the critic and generator, while χ^2 GAN decouples the critic and generator objectives, requiring simpler (unconstrained) stochastic-gradient-descent-type updates.

For RKHS-based generative modeling the choice of kernel is crucial. Theoretical properties and empirical performance have been analyzed for a number of popular kernels, such as inverse multi-quadratic (Gorham & Mackey, 2017), Plummer (Unterthiner et al., 2018), rational quadratic (Bińkowski et al., 2018) and energy distance (Liu, 2017). Although derived from a kernel formulation, our training procedure is in practice kernel-free.

Modern generative modeling has deep roots in statistical testing. Prior studies have primarily focused on two sample tests (Gretton et al., 2012). Our study builds on the work of independence testing (Gretton et al., 2005b), which generalizes two sample tests and can be extended to settings with multiple generators and critics (Durugkar et al., 2017).

Simultaneous matching of multiple distributions is a key technique needed in many machine learning applications (Zhao et al., 2017). In the GAN context, generalization to the JSD metric have been explored to address this challenge (Gan et al., 2017; Li et al., 2017a; Pu et al., 2018b). However for RKHS and IPM-based generative models, currently there is no generalization, and one has to build $\frac{M(M-1)}{2}$ pairwise critics, which can be prohibitively expensive when M grows large. Our χ^2 GAN represents the first attempt to bridge this gap. Our theory implies using a quadratic number of critics is unnecessary, and instead χ^2 GAN computes an $M-1$ dimensional critic.

Importance sampling is a classic technique used in Monte Carlo methods (Liu, 2008). One of its key applications is to evaluate the quality of statistical models (Neal, 2001; Wu et al., 2017). Recently, this idea has been used in likelihood-based generative models to sharpen the variational bound (Burda et al., 2016), and has been extended to improve the training of likelihood-assisted GAN variants (Hu et al., 2018). To the best of the authors' knowledge, resampling the generator as proposed here has not yet been explored before and, not being exclusive to our formulation, can be easily used with other methods.

4. Experiments

We consider a wide range of synthetic and real-world tasks to experimentally validate χ^2 GAN, and benchmark it against other state-of-the-art solutions. All experiments are implemented with Tensorflow and run on a single NVIDIA TITAN X GPU. Details of the experimental setup are in the SM, and code for our experiments are available from `https://www.github.com/chenyang-tao/chi2gan`.

4.1. Toy Distributions

We compare χ^2 GAN with one representative model from each category, namely the original GAN, WGAN-GP and MMD GAN, on three toy distributions. The same model architecture is used for all models, with the exception of MMD GAN which does not need an explicit critic function. The generation results are summarized in Figure 2.

All models except MMD perform well on the baseline Swiss roll experiment. The mixture-of-Gaussians experiments test algorithm robustness to mode collapse. The original GAN demonstrates its vulnerability dealing with distributions with disjoint modes, and MMD learns an overly smoothed distribution, even with carefully tuned kernel hyperparameter. Both WGAN-GP and χ^2 GAN successfully learn good approximations to the target distribution, with the latter showing faster convergence and more-stable training.

Table 2. Unsupervised Inception Score on CIFAR-10

Model	IS
ALI (Dumoulin et al., 2017)	$5.34 \pm .05$
DCGAN (Radford et al., 2016)	$6.16 \pm .07$
MMD-GAN (Li et al., 2017b)	$6.17 \pm .07$
WGAN-GP (Gulrajani et al., 2017)	$6.56 \pm .05$
ASVAE (Pu et al., 2017)	$6.89 \pm .05$
sVAE-r (Pu et al., 2018a)	$6.96 \pm .066$
χ^2-GAN (ours)	7.47 ± 0.105
WGAN-GP ResNet	$7.86 \pm .07$
Fisher-GAN ResNet (Mroueh & Sercu, 2017)	$7.90 \pm .05$
χ^2-GAN ResNet (ours)	$7.88 \pm .10$

Table 3. Supervised Inception Score on CIFAR-10

Model	IS
SteinGAN (Wang & Liu, 2016)	$6.16 \pm .07$
DCGAN (Radford et al., 2016)	$6.58 \pm .05$
AC-GAN (Odena et al., 2017)	$8.25 \pm .07$
SGAN (Huang et al., 2017)	$8.59 \pm .12$
Fisher-GAN ResNet (Mroueh & Sercu, 2017)	$8.16 \pm .12$
WGAN-GP ResNet (Gulrajani et al., 2017)	$8.42 \pm .10$
χ^2-GAN ResNet (ours)	$8.44 \pm .10$

Table 4. Unsupervised Inception Score on ImageNet

Model	IS
DCGAN (Radford et al., 2016)	5.965
PixelCNN++ (Salimans et al., 2017)	7.65
ASVAE (Pu et al., 2017)	11.14
χ^2-GAN (ours)	11.34

4.2. Image Datasets

We trained χ^2 GAN on a number of popular image datasets to demonstrate its ability to learn complex distributions for real-world applications. For supervised generation tasks, we condition the generator on the label of an image. To quantitatively evaluate model performance, we consider the following metrics in our experiments: (1) *Inception score (IS)* (Salimans et al., 2016) for datasets associated with one-hot labels; (2) *AIS score* (Wu et al., 2017) to estimate the log likelihood.

We only report the quantitative results for the raw generator distribution in the main text, and results for the importance resampled generator are found separately in SM. Two network architectures were considered in these experiments: DCGAN and ResNet. In all experiments we have used Xaiver initialization and Adam optimizer. All images shown are random samples and not cherry picked. We note better quantitative results have been reported in the literature using specific techniques orthogonal to our main contributions (Karras et al., 2018; Warde-Farley & Bengio, 2017; Miyato et al., 2017; Salimans et al., 2018), see SM for a discussion.

MNIST We used the binarized MNIST in this experiment and compared with the results from prior results in Table 1. Our χ^2 GAN achieves an AIS score of -78.85 nats and an IS score of 9.01. These results lead the chart for all likelihood-free generative models we considered, and they are comparable to, or even better than those from the best-performing likelihood-based models.

Cifar10 For this dataset, we experimented with both unsupervised and supervised generation tasks. Quantitative results are summarized in Tables 2 and 3. For both tasks, χ^2 GAN consistently achieved state-of-the-art results obtained with the network architectures considered. Most notably, our χ^2 GAN significantly outperformed DCGAN, MMD GAN and WGAN-GP in the unsupervised generation task with the DCGAN architecture. We also provide quantitative results with the Fréchet Inception Distance (FID) (Heusel et al., 2017) in SM. See Figure 3 for qualitative assessment.

CelebA We provide a comparison of DCGAN, Fisher GAN and our χ^2 GAN on the face generation task in Figure 4. We trained DCGAN and χ^2 GAN to generate the face samples, and collected Fisher GAN's samples from the original paper. All models used the DCGAN architecture. We observe that χ^2 GAN produced (subjectively) more compelling samples, capturing facial details, illumination and more realistic textures compared with its counterparts. Additional samples are shown in the SM, and we find no evidence of mode collapse. We also provide additional experimental evidence in the SM to verify χ^2-GAN learns generalizable features rather than remembering training examples.

We also use the face-generation task to demonstrate the efficacy of importance resampling. In Figure 4 we compare the accepted and rejected samples[1]. Those less-compelling samples produced by the generator are immediately identified based on the importance score. Quantitive results for importance resampling can be found in the SM.

ImageNet We also considered ImageNet to evaluate the scalability of the models on large datasets. All images are resized to 64×64, and the quantitative results are reported in Table 4. With the simple DCGAN architecture, our χ^2 GAN significantly outperformed more-sophisticated PixelCNN++ GAN, even surpassing the performance of likelihood-based ASVAE. See the SM for generated samples.

Stability, robustness, convergence and sample diversity In all our experiments on the image datasets, χ^2 GAN demonstrated stable training dynamics. It showed a similar convergence rate [2] compared with WGAN-GP in terms of iterations, but much cheaper per-iteration cost. χ^2 GAN also demonstrated robustness as we varied model architectures, network normalization schemes and hyperparameters

[1] We used a less well-trained model and picked our samples based on the importance weights to highlight the difference.

[2] Wrt IS score and visual inspection, see SM.

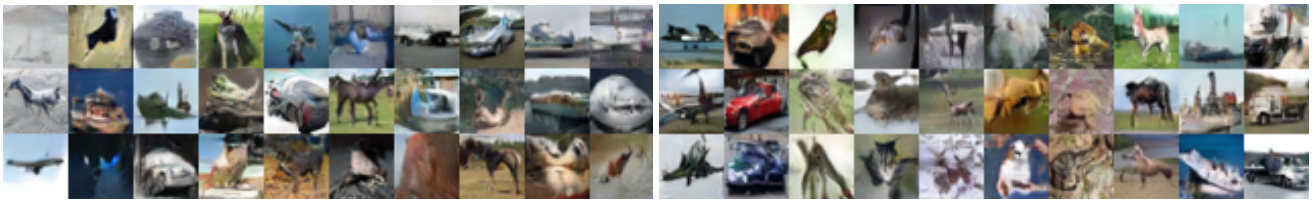

Figure 3. Generated images on Cifar-10. Left: unsupervised generation; right: supervised generation.

Figure 4. Face generation quality comparison. From left to right: DCGAN, Fisher GAN, χ^2 GAN, high importance weight samples and low importance weight samples.

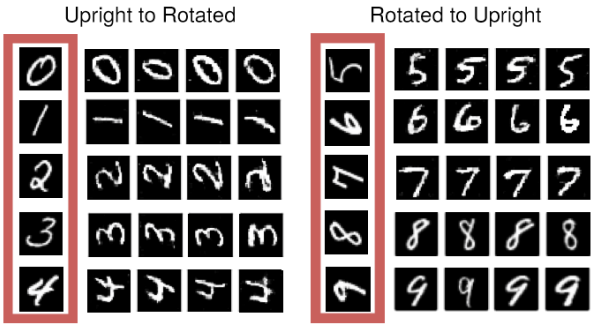

Figure 5. MNIST image translation. First column is the input image from test set, subsequent columns are sampled translations.

Figure 6. Edges2shoes translation.

(results not shown). We also do not find any evidence for mode collapse in our experiments.

4.3. Matching Multiple Distributions

To demonstrate the flexibility of the χ^2 GAN framework to match multiple distributions, we consider the following translation task: Given paired examples $\{(x_i, z_i)\}_{i=1}^n$, sample from all possible translations z for a new observation x, and *vice versa*. More explicitly, consider the distribution triplet: $p_0(x, z) = p_d(x, z)$, $p_1(x, z) = p_d(x)q_1(z|x; \theta_1)$, and $p_2(x, z) = p_d(z)q_2(x|z; \theta_2)$, where $q_1(z|x; \theta_1)$ and $q_2(x|z; \theta_2)$ are the translation models. When translations are faithful to the data distribution, we have $p_0(x, z) = p_1(x, z) = p_2(x, z)$. Here (x, z) are paired data, and we consider the problem of image-to-image translation.

Rotated MNIST In this experiment we pair each MNIST digit with a random sample of the same type, rotated by 90°. Our translation results are presented in Figure 5. It is observed that χ^2 GAN translations achieved both fidelity and diversity for this task.

Edges-to-shoes We evaluate the performance on the more-realistic edges-to-shoes dataset, where the model learns to translate between shoes and sketches. As shown in Figure 6, χ^2 GAN learned to produce faithful translations.

5. Conclusions

We have developed a framework that unifies prior theoretical frameworks for likelihood-free generative modelling, and based on this we proposed a novel algorithm named χ^2 GAN. Our approach is conceptually simple, and can be readily generalized to match multiple distributions. Empirical evidence verified that this new method offers competitive performance on a wide range of generation tasks. For future work, we intend to investigate its connections to the likelihood-based generative models, and to seek novel applications by integrating it with other machine learning algorithms.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments. This research was supported in part by DARPA, DOE, NIH, ONR and NSF. J Feng is partially supported by the key project of Shanghai Science & Technology Innovation Plan (No. 16JC1420402). The authors would also like to thank Prof. C Leng, Dr. Y Zhang, Dr. Y Pu and K Bai for fruitful discussions.

References

Adams, R. A. and Fournier, J. J. *Sobolev spaces*, volume 140. Academic press, 2003.

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. In *NIPS Workshop*. 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *ICML*, 2017.

Arora, S. and Zhang, Y. Do GANs actually learn the distribution? an empirical study. In *ICLR*, 2018.

Baker, C. R. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186: 273–289, 1973.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans. In *ICLR*, 2018.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. In *ICLR*, 2016.

Csiszár, I. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad*, 8: 85–108, 1963.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A. Adversarially learned inference. In *ICLR*, 2017.

Durugkar, I., Gemp, I., and Mahadevan, S. Generative multi-adversarial networks. In *ICLR*, 2017.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *NIPS*, 2007.

Gan, Z., Chen, L., Wang, W., Pu, Y., Zhang, Y., Liu, H., Li, C., and Carin, L. Triangle generative adversarial networks. In *NIPS*, 2017.

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

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *ICML*, 2017.

Gretton, A., Bousquet, O., Smola, A., and Scholkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, 2005a.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. Kernel methods for measuring independence. *The Journal of Machine Learning Research*, 6: 2075–2129, 2005b.

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.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*, 2017.

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 *NIPS*, 2017.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., and Xing, E. P. Toward controlled generation of text. In *ICML*, 2017.

Hu, Z., Yang, Z., Salakhutdinov, R., and Xing, E. P. On unifying deep generative models. In *ICLR*, 2018.

Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., and Belongie, S. Stacked generative adversarial networks. In *CVPR*, 2017.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

Li, C., Liu, H., Chen, C., Pu, Y., Chen, L., Henao, R., and Carin, L. Alice: Towards understanding adversarial learning for joint distribution matching. In *NIPS*, 2017a.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. Mmd gan: Towards deeper understanding of moment matching network. In *NIPS*, 2017b.

Li, Y., Swersky, K., and Zemel, R. Generative moment matching networks. In *ICML*, 2015.

Liu, J. S. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

Liu, L. On the two-sample statistic approach to generative adversarial networks. Master's thesis, Princeton University, 2017.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. Least squares generative adversarial networks. In *ICCV*. 2017.

Mescheder, L., Nowozin, S., and Geiger, A. Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. In *ICML*, 2017a.

Mescheder, L., Nowozin, S., and Geiger, A. The numerics

of gans. In *NIPS*, 2017b.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *ICML Implicit Models Workshop*, 2017.

Mroueh, Y. and Sercu, T. Fisher GAN. In *NIPS*. 2017.

Mroueh, Y., Sercu, T., and Goel, V. McGAN: Mean and covariance feature matching gan. In *ICML*, 2017.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29 (2):429–443, 1997.

Neal, R. M. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Nock, R., Cranko, Z., Menon, A. K., Qu, L., and Williamson, R. C. f-gans in an information geometric nutshell. In *NIPS*, 2017.

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

Odena, A., Olah, C., and Shlens, J. Conditional image synthesis with auxiliary classifier GANs. In *ICML*, 2017.

Oord, A. v. d., Kalchbrenner, N., and Kavukcuoglu, K. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

Pu, Y., Wang, W., Henao, R., Chen, L., Gan, Z., Li, C., and Carin, L. Adversarial symmetric variational autoencoder. In *NIPS*, 2017.

Pu, Y., Chen, L., Dai, S., Wang, W., Li, C., and Carin, L. Symmetric variational autoencoder and connections to adversarial learning. In *AISTATS*, 2018a.

Pu, Y., S., D., Gan, Z., Wang, W., G., W., Y., Z., Henao, R., and Carin, L. Joint gan: Multi-domain joint distribution learning with generative adversarial nets. In *ICML*, 2018b.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. In *ICML*, 2015.

Roth, K., Lucchi, A., Nowozin, S., and Hofmann, T. Stabilizing training of generative adversarial networks through regularization. In *NIPS*, 2017.

Rubner, Y., Tomasi, C., and Guibas, L. J. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training GANs. In *NIPS*, 2016.

Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *ICLR*, 2017.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. Improving gans using optimal transport. In *ICLR*, 2018.

Sønderby, C. K., Caballero, J., Theis, L., Shi, W., and Huszár, F. Amortised map inference for image super-resolution. In *ICLR*, 2017.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

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*, 2016.

Unterthiner, T., Nessler, B., Klambauer, G., Heusel, M., Ramsauer, H., and Hochreiter, S. Coulomb GANs: Provably optimal nash equilibria via potential fields. In *ICLR*, 2018.

Wang, D. and Liu, Q. Learning to draw samples: With application to amortized MLE for generative adversarial learning. *arXiv:1611.01722*, 2016.

Warde-Farley, D. and Bengio, Y. Improving generative adversarial networks with denoising feature matching. In *ICLR*, 2017.

Wu, Y., Burda, Y., Salakhutdinov, R., and Grosse, R. On the quantitative analysis of decoder-based generative models. In *ICLR*. 2017.

Yoshida, K. *Functional analysis*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, 1974.

Yu, L., Zhang, W., Wang, J., and Yu, Y. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2017.

Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D., and Carin, L. Adversarial feature matching for text generation. In *ICML*, 2017.

Zhao, H., Zhang, S., Wu, G., Costeira, J. P., Moura, J. M., and Gordon, G. J. Multiple source domain adaptation with adversarial training of neural networks. *arXiv preprint arXiv:1705.09684*, 2017.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.