LEARNING IN IMPLICIT GENERATIVE MODELS

Shakir Mohamed and Balaji Lakshminarayanan
DeepMind, London
{shakir,balajiln}@google.com

ABSTRACT

Generative adversarial networks (GANs) provide an algorithmic framework for constructing generative models with several appealing properties: they do not require a likelihood function to be specified, only a generating procedure; they provide samples that are sharp and compelling; and they allow us to harness our knowledge of building highly accurate neural network classifiers. Here, we develop our understanding of GANs with the aim of forming a rich view of this growing area of machine learning—to build connections to the diverse set of statistical thinking on this topic, of which much can be gained by a mutual exchange of ideas. We frame GANs within the wider landscape of algorithms for learning in implicit generative models—models that only specify a stochastic procedure with which to generate data—and relate these ideas to modelling problems in related fields, such as econometrics and approximate Bayesian computation. We develop likelihood-free inference methods and highlight hypothesis testing as a principle for learning in implicit generative models, using which we are able to derive the objective function used by GANs, and many other related objectives. The testing viewpoint directs our focus to the general problem of density ratio estimation. There are four approaches for density ratio estimation, one of which is a solution using classifiers to distinguish real from generated data. Other approaches such as divergence minimisation and moment matching have also been explored in the GAN literature, and we synthesise these views to form an understanding in terms of the relationships between them and the wider literature, highlighting avenues for future exploration and cross-pollination.

1 IMPLICIT GENERATIVE MODELS

It is useful to make a distinction between two types of probabilistic models: prescribed and implicit models (Diggle and Gratton, 1984). *Prescribed probabilistic models* are those that provide an explicit parametric specification of the distribution of an observed random variable $\mathbf{x}$, specifying a log-likelihood function $\log q_\theta(\mathbf{x})$ with parameters $\boldsymbol{\theta}$. Most models in machine learning and statistics are of this form, whether they be state-of-the-art classifiers for object recognition, complex sequence models for machine translation, or fine-grained spatio-temporal models tracking the spread of disease. Alternatively, we can specify *implicit probabilistic models* that define a stochastic procedure that directly generates data. Such models are the natural approach for problems in climate and weather, population genetics, and ecology, since the mechanistic understanding of such systems can be used to directly create a data simulator, and hence the model. It is exactly because implicit models are more natural for many problems that they are of interest and importance.

Implicit generative models use a latent variable $\mathbf{z}$ and transform it using a deterministic function $\mathcal{G}_\theta$ that maps from $\mathbb{R}^m \to \mathbb{R}^d$ using parameters $\boldsymbol{\theta}$. Such models are amongst the most fundamental of models, e.g., many of the basic methods for generating non-uniform random variates are based on simple implicit models and one-line transformations (Devroye, 2006). In general, implicit generative models specify a valid density on the output space that forms an effective likelihood function:

$$\mathbf{x} = \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{z}'); \qquad \mathbf{z}' \sim q(\mathbf{z}) \tag{1}$$

$$q_\theta(\mathbf{x}) = \frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_d} \int_{\{\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{z}) \leq \mathbf{x}\}} q(\mathbf{z}) d\mathbf{z}, \tag{2}$$

where $q(\mathbf{z})$ is a latent variable that provides the external source of randomness and equation (2) is the definition of the transformed density as the derivative of the cumulative distribution function. When the function $\mathcal{G}$ is well-defined, such as when the function is invertible, or has dimensions $m = d$ with easily characterised roots, we recover the familiar rule for transformations of probability distributions.

We are interested in developing more general and flexible implicit generative models where the function $\mathcal{G}$ is a non-linear function with $d > m$, specified by deep networks. The integral (2) is intractable in this case: we will be unable to determine the set $\{\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{z}) \leq \mathbf{x}\}$, the integral will often be unknown even when the integration regions are known and, the derivative is high-dimensional and difficult to compute. Intractability is also a challenge for prescribed models, but the lack of a likelihood term significantly reduces the tools available for learning. In implicit models, this difficulty motivates the need for methods that side-step the intractability of the likelihood (2), or are likelihood-free.

Generative adversarial networks (GANs) (Goodfellow et al., 2014) provide a solution for exactly this type of problem. This connection makes it possible to export GAN-type solutions to other areas, and to import new approaches from areas addressing the same research problem. GANs specify an algorithmic framework for learning in implicit generative models, also referred to as *generator networks* or *generative neural samplers* or *differentiable simulators* in the literature. The framework relies on a learning principle based on discriminating real from generated data, which we shall show instantiates a core principle of likelihood-free inference, that of hypothesis and two-sample testing.

Note on notation. We denote data by the random variable $\mathbf{x}$, the (unknown) true data density by $p^*(\mathbf{x})$, our (intractable) model density by $q_\theta(\mathbf{x})$. $q(\mathbf{z})$ is a density over latent variables $\mathbf{z}$. Parameters of the model are $\boldsymbol{\theta}$, and parameters of the ratio and discriminator functions are ϕ.

2 Hypothesis Testing and Density Ratios

2.1 Likelihood-free Inference

Without a likelihood function, many of the widely-used tools for inference and parameter learning become unavailable. But there are tools that remain, including the method-of-moments (Hall, 2005), the empirical likelihood (Owen, 1988), Monte Carlo sampling (Marin et al., 2012), and mean-shift estimation (Fukunaga and Hostetler, 1975). Since we can easily draw samples from the model, we can use any method that compares two sets of samples—one from the true data distribution and one from the model distribution—to drive learning. This is a process of *density estimation-by-comparison* that tests the hypothesis that the true data distribution $p^*(\mathbf{x})$ and our model distribution $q(\mathbf{x})$ are equal, using the *density ratio function* $r(\mathbf{x}) = {}^{p^*(\mathbf{x})}/_{q(\mathbf{x})}$. The density ratio provides information about the departure of our model distribution from the true distribution, and our aim is to see this ratio be one. The density ratio $r(\mathbf{x})$ is the core quantity for hypothesis testing, motivated by either the Neyman-Pearson lemma or the Bayesian posterior evidence, appearing as the likelihood ratio or the Bayes factor (Kass and Raftery, 1995), respectively. This is the focus of our approach for likelihood-free inference: estimating density ratios and using them as the driving principle for learning in implicit generative models.

The direct approach of computing the density ratio by first computing the individual marginals is not possible with implicit models. By directly estimating the ratio and exploiting knowledge of the probabilities involved, it will turn out that computing the ratio can be a much easier problem than computing the marginal likelihoods, and is what will make this approach appealing. There are four general approaches to consider (Sugiyama et al., 2012a): 1) class-probability estimation, 2) divergence minimisation, 3) ratio matching, and 4) moment matching. These are highly developed research areas in their own right, but their role in providing a learning principle for density estimation is under-appreciated and opens up an exciting range of approaches for learning in implicit generative models. Figure 1 summarises these approaches by showing pathways available for learning, which follow from the choice of inference driven by hypothesis testing and comparison.

2.2 Class Probability Estimation

The density ratio can be computed by building a classifier to distinguish observed data from that generated by the model. This is the most popular approach for density ratio estimation and the first port of call for learning in implicit models. Hastie et al. (2013) call this approach unsupervised-as-supervised learning, Qin (1998) explore this for analysis of case-control in observational studies, both Neal (2008) and Cranmer et al. (2015) explore this approach for high-energy physics applications, Gutmann and Hyvärinen (2012) exploit it for learning un-normalised models, Lopez-Paz and Oquab

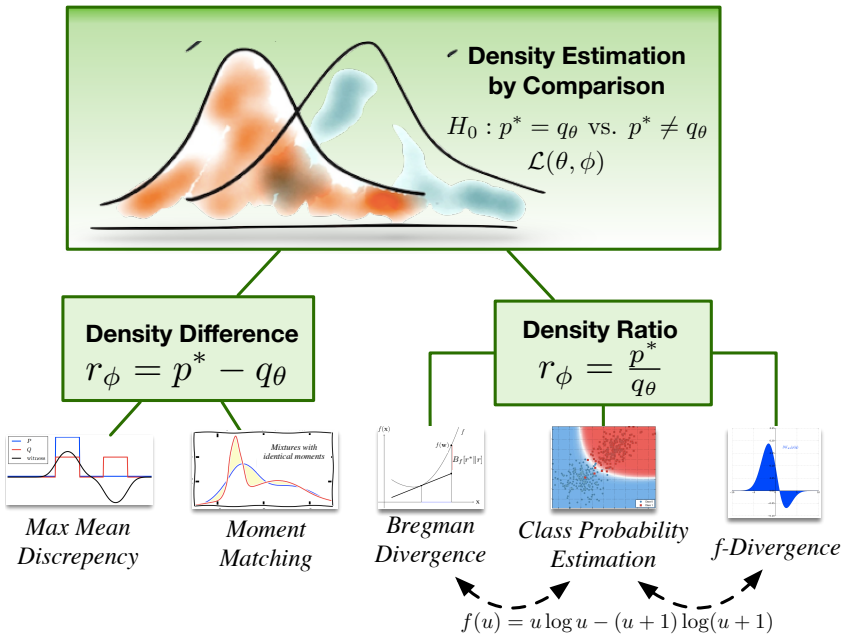

Figure 1: Summary of approaches for learning in implicit models. We define a joint function $\mathcal{L}(\phi, \theta)$ and alternate between optimising the loss w.r.t. ϕ and θ.

(2016) for causal discovery, and Goodfellow et al. (2014) for learning in implicit generative models specified by neural networks.

We denote the domain of our data by $\mathcal{X} \subset \mathbb{R}^d$. The true data distribution has a density $p^*(\mathbf{x})$ and our model has density $q_\theta(\mathbf{x})$, both defined on $\mathcal{X}$. We also have access to a set of n samples $\mathcal{X}_p = \{\mathbf{x}_1^{(p)}, \ldots, \mathbf{x}_n^{(p)}\}$ from the true data distribution, and a set of n' samples $\mathcal{X}_q = \{\mathbf{x}_1^{(q)}, \ldots, \mathbf{x}_{n'}^{(q)}\}$ from our model. We introduce a random variable y, and assign a label $y = 1$ to all samples in $\mathcal{X}_p$ and $y = 0$ to all samples in $\mathcal{X}_q$. We can now represent $p^*(\mathbf{x}) = p(\mathbf{x}|y = 1)$ and $q_\theta(\mathbf{x}) = p(\mathbf{x}|y = 0)$. By application of Bayes' rule, we can compute the ratio $r(\mathbf{x})$ as:

$$\frac{p^*(\mathbf{x})}{q_\theta(\mathbf{x})} = \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} = \frac{p(y=1|\mathbf{x})p(\mathbf{x})}{p(y=1)} \Big/ \frac{p(y=0|\mathbf{x})p(\mathbf{x})}{p(y=0)} = \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} \cdot \frac{1-\pi}{\pi}, \tag{3}$$

which indicates that the problem of density ratio estimation is equivalent to that of class probability estimation, since the problem is reduced to computing the probability $p(y = 1|\mathbf{x})$. We assume that the marginal probability over classes is $p(y = 1) = \pi$, which allows the relative proportion of data from the two classes to be adjusted if they are imbalanced; in most formulations $\pi = ½$ for the balanced case, and in imbalanced cases $\frac{1-\pi}{\pi} \approx {}^{n'}/_{n}$.

Our task is now to specify a scoring function, or discriminator, $\mathcal{D}(\mathbf{x}; \boldsymbol{\phi}) = p(\mathbf{y} = 1|\mathbf{x})$: a function bounded in [0,1] with parameters $\boldsymbol{\phi}$ that computes the probability of data belonging to the positive (real data) class. This discriminator is related to the density ratio through the mapping $\mathcal{D} = {}^{r}/_{(r+1)}$; $r = {}^{\mathcal{D}}/_{(1-\mathcal{D})}$. Conveniently, we can use our knowledge of building classifiers and specify these functions using deep neural networks. Given this scoring function, we must specify a proper scoring rule (Gneiting and Raftery, 2007; Buja et al., 2005) for binary discrimination to allow for parameter learning, such as those in Table 1. A natural choice is to use the Bernoulli (logarithmic) loss:

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{x}|y)p(y)}[-y \log \mathcal{D}(\mathbf{x}; \boldsymbol{\phi}) - (1-y)\log(1 - \mathcal{D}(\mathbf{x}; \boldsymbol{\phi}))] \tag{4}$$

$$= \pi \mathbb{E}_{p^*(\mathbf{x})}[-\log \mathcal{D}(\mathbf{x}; \boldsymbol{\phi})] + (1-\pi)\mathbb{E}_{q_\theta(\mathbf{x})}[-\log(1 - \mathcal{D}(\mathbf{x}; \boldsymbol{\phi}))]. \tag{5}$$

Since we know the underlying generative process for $q_\theta(\mathbf{x})$, using a change of variables, we can express the loss in term of an expectation over the latent variable $\mathbf{z}$ and the generative model $\mathcal{G}(\mathbf{z}; \boldsymbol{\theta})$:

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \pi \mathbb{E}_{p^*(\mathbf{x})}[-\log \mathcal{D}(\mathbf{x}; \boldsymbol{\phi})] + (1-\pi)\mathbb{E}_{q(\mathbf{z})}[-\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z}; \boldsymbol{\theta}); \boldsymbol{\phi}))]. \tag{6}$$

Table 1: Proper scoring rules that can be minimised in class probability-based learning of implicit generative models.

Loss	**Objective Function** ($\mathcal{D} := \mathcal{D}(\mathbf{x}; \boldsymbol{\phi})$)
Bernoulli loss	$\pi \mathbb{E}_{p^*(\mathbf{x})}[-\log \mathcal{D}] + (1-\pi)\mathbb{E}_{q_\theta(\mathbf{x})}[-\log(1-\mathcal{D})]$
Brier score	$\pi \mathbb{E}_{p^*(\mathbf{x})}[(1-\mathcal{D})^2] + (1-\pi)\mathbb{E}_{q_\theta(\mathbf{x})}[\mathcal{D}^2]$
Exponential loss	$\pi \mathbb{E}_{p^*(\mathbf{x})}\left[\left(\frac{1-\mathcal{D}}{\mathcal{D}}\right)^{\frac{1}{2}}\right] + (1-\pi)\mathbb{E}_{q_\theta(\mathbf{x})}\left[\left(\frac{\mathcal{D}}{1-\mathcal{D}}\right)^{\frac{1}{2}}\right]$
Misclassification	$\pi \mathbb{E}_{p^*(\mathbf{x})}[\mathbb{I}[\mathcal{D} \leq 0.5]] + (1-\pi)\mathbb{E}_{q_\theta(\mathbf{x})}[\mathbb{I}[\mathcal{D} > 0.5]]$
Hinge loss	$\pi \mathbb{E}_{p^*(\mathbf{x})}\left[\max\left(0, 1 - \log \frac{\mathcal{D}}{1-\mathcal{D}}\right)\right] + (1-\pi)\mathbb{E}_{q_\theta(\mathbf{x})}\left[\max\left(0, 1 + \log \frac{\mathcal{D}}{1-\mathcal{D}}\right)\right]$
Spherical	$\pi \mathbb{E}_{p^*(\mathbf{x})}\left[-\alpha\mathcal{D}\right] + (1-\pi)\mathbb{E}_{q_\theta(\mathbf{x})}\left[-\alpha(1-\mathcal{D})\right]; \quad \alpha = (1 - 2\mathcal{D} + 2\mathcal{D}^2)^{-1/2}$

The final form of this objective (6) is exactly that used in generative adversarial networks (GANs) (Goodfellow et al., 2014). In practice, the expectations are computed by Monte Carlo integration using samples from p^* and q_θ. Equation (6) allows us to specify a bi-level optimisation (Colson et al., 2007) by forming a *ratio loss* and a *generative loss*, using which we perform an alternating optimisation. Our convention throughout the paper will be to always form the ratio loss by extracting all terms in $\mathcal{L}$ related to the ratio function parameters $\boldsymbol{\phi}$, and minimise the resulting objective. For the generative loss, we will similarly extract all terms related to the model parameters $\boldsymbol{\theta}$, flip the sign, and minimise the resulting objective. For equation (6), the bi-level optimisation is:

$$\textbf{Ratio loss:} \quad \min_{\phi} \pi \mathbb{E}_{p^*(\mathbf{x})}[-\log \mathcal{D}(\mathbf{x}; \boldsymbol{\phi})] + (1-\pi)\mathbb{E}_{q_\theta(\mathbf{x})}[-\log(1 - \mathcal{D}(\mathbf{x}; \boldsymbol{\phi}))] \tag{7}$$

$$\textbf{Generative loss:} \quad \min_{\theta} \mathbb{E}_{q(\mathbf{z})}[\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z}; \boldsymbol{\theta})))]. \tag{8}$$

The ratio loss is minimised since it acts as a surrogate negative log-likelihood; the generative loss is minimised since we wish to minimise the probability of the negative (generated-data) class. We explicitly write out these two stages to emphasise that the objectives used are separable. While we can derive the generative loss from the ratio loss as we have done, any generative loss that drives q_θ to p^*, such as minimising the widely-used $\mathbb{E}_{q(\mathbf{z})}[-\log \mathcal{D}(\mathcal{G}(\mathbf{z}; \boldsymbol{\theta}))]$ (Goodfellow et al., 2014; Nowozin et al., 2016) or $\mathbb{E}_{q(\mathbf{z})}[-\log \frac{\mathcal{D}(\mathcal{G}(\mathbf{z};\boldsymbol{\theta}))}{1-\mathcal{D}(\mathcal{G}(\mathbf{z};\boldsymbol{\theta}))}] = \mathbb{E}_{q(\mathbf{z})}[-\log r(\mathcal{G}(\mathbf{z}; \boldsymbol{\theta}))$ (Sønderby et al., 2016), is possible.

Any scoring rule from Table 1 can be used to give a loss function for optimisation. These rules are amenable to stochastic approximation and alternating optimisation, as described by Goodfellow et al. (2014), along with many of the insights for optimisation that have been developed since (Salimans et al., 2016; Radford et al., 2015; Zhao et al., 2016; Sønderby et al., 2016). The Bernoulli loss can be criticised in a number of ways and makes other scoring rules interesting to explore. The Brier loss provides a similar decision rule, and its use in calibrated regression makes it appealing; the motivations behind many of these scoring rules are discussed in (Gneiting and Raftery, 2007). Finally, while we have focussed on the two sample hypothesis test, we believe it could be advantageous to extend this reasoning to the case of multiple testing, where we simultaneously test several sets of data (Bickel et al., 2008).

The advantage of using a proper scoring rule is that the global optimum is achieved iff $q_\theta = p^*$ (cf. the proof in (Goodfellow et al., 2014) for the Bernoulli loss); however there are no convergence guarantees since the optimisation is non-convex. Goodfellow et al. (2014) discuss the relationship to maximum likelihood estimation, which minimises the divergence $\mathrm{KL}[p^* \| q]$, and show that the GAN objective with Bernoulli loss is instead related to the Jensen Shannon divergence $\mathrm{JS}\,[p^* || q]$. In the objective (6), π denotes the marginal probability of the positive class; however several authors have proposed choosing π depending on the problem. In particular, Huszár (2015) showed that varying π is related to optimizing a generalised Jensen-Shannon divergence $\mathrm{JS}_\pi\,[p^* || q]$. Creswell and Bharath (2016) presented results showing that different values of π are desirable, depending on whether we wish to fit one of the modes (a 'high precision, low recall' task such as generation) or explain all of the modes (a 'high recall, low precision' task such as retrieval).

2.3 DIVERGENCE MINIMISATION

A second approach to testing is to use the divergence between the true density p^* and our model q, and use this as an objective to drive learning of the generative model. A natural class of divergences to use are the f-divergences (or Ali-Silvey (Ali and Silvey, 1966) or Csiszar's ϕ-divergence (Csiszàr, 1967)) since they are fundamentally linked to the problem of two-sample hypothesis testing (Liese and Vajda, 2008): f-divergences represent an integrated Bayes risk since they are an expectation of the density ratio. Nowozin et al. (2016) develop f-GANs using this view. The f-divergences contain the KL divergence as a special case and are equipped with an exploitable variational formulation:

$$D_f\left[p^*(\mathbf{x})\|q_\theta(\mathbf{x})\right] = \int q_\theta(\mathbf{x}) f\left(\frac{p^*(\mathbf{x})}{q_\theta(\mathbf{x})}\right) d\mathbf{x} = \mathbb{E}_{q_\theta(\mathbf{x})}[f(r(\mathbf{x}))] \geq \sup_t \mathbb{E}_{p^*(\mathbf{x})}[t(\mathbf{x})] - \mathbb{E}_{q_\theta(\mathbf{x})}[f^\dagger(t(\mathbf{x}))] \quad (9)$$

where f is a convex function with derivative f' and Fenchel conjugate $f^\dagger$; this divergence class instantiates many familiar divergences, such as the KL and Jensen-Shannon divergence. The variational formulation introduces the functions $t(\mathbf{x})$ whose optimum is related to the density ratio since $t^*(\mathbf{x}) = f'(r(\mathbf{x}))$. Substituting t^* in (9), we transform the objective (10) into supremum over r_ϕ (which is attained when $r_\phi = r^* = {}^{p^*}\!/_{q_\theta}$). For self-consistency, we flip the sign to make it a minimisation problem in r_ϕ, leading to the bi-level optimisation:

$$\mathcal{L} = \mathbb{E}_{p^*(\mathbf{x})}[-f'(r_\phi(\mathbf{x}))] + \mathbb{E}_{q_\theta(\mathbf{x})}[f^\dagger(f'(r_\phi(\mathbf{x}))] \quad (10)$$

$$\textbf{Ratio loss:} \quad \min_\phi \mathbb{E}_{p^*(\mathbf{x})}[-f'(r_\phi(\mathbf{x}))] + \mathbb{E}_{q_\theta(\mathbf{x})}[f^\dagger(f'(r_\phi(\mathbf{x}))] \quad (11)$$

$$\textbf{Generative loss:} \quad \min_\theta \mathbb{E}_{q(\mathbf{z})}[-f^\dagger(f'(r(\mathcal{G}(\mathbf{z};\boldsymbol{\theta})))], \quad (12)$$

where we derived equation (12) by extracting all the terms involving $q_\theta(\mathbf{x})$ in equation (10), used the change of variables to express it in terms of the underlying generative model and flipping the sign to obtain a minimisation. There is no discriminator in this formulation, and this role is taken by the ratio function. We minimise the ratio loss, since we wish to minimise the negative of the variational lower bound; we minimise the generative loss since we wish to drive the ratio to one. By using the function $f(u) = u\log u$, we recover an objective using the KL divergence; when $f(u) = u\log u - (u+1)\log(u+1)$, we recover the objective function in equation (6) from the previous section using the Bernoulli loss, and hence the objective for GANs.

The density ratio implies that $p^*(\mathbf{x}) \approx \tilde{p} = r_\phi(\mathbf{x})q_\theta(\mathbf{x})$, since it is the amount by which we must correct our model $q_\theta(\mathbf{x})$ to match the true distribution. This led us to a divergence formulation that evaluates the divergence between the distributions p^* and $\tilde{p}$, using the KL divergence:

$$D_{KL}[p^*(\mathbf{x})\|\tilde{p}(\mathbf{x})] = \int p^*(\mathbf{x}) \log \frac{p^*(\mathbf{x})}{r_\phi(\mathbf{x})q_\theta(\mathbf{x})} d\mathbf{x} + \int (r_\phi(\mathbf{x})q_\theta(\mathbf{x}) - p^*(\mathbf{x}))d\mathbf{x} \quad (13)$$

$$\mathcal{L} = \mathbb{E}_{p^*(\mathbf{x})}[-\log r_\phi(\mathbf{x})] + \mathbb{E}_{q_\theta(\mathbf{x})}[r_\phi(\mathbf{x}) - 1] - \mathbb{E}_{p^*(\mathbf{x})}[\log q_\theta(\mathbf{x})] + \mathbb{E}_{p^*(\mathbf{x})}[\log p^*(\mathbf{x})], \quad (14)$$

where the first equation is the KL for un-normalised densities (Minka, 2005). This leads to a convenient and valid ratio loss since all terms independent of r can be ignored. But we are unable to derive a useful generative loss from this expression (14) , since the third term with $\log q$ cannot be ignored, and is unavailable. Since the generative loss and ratio losses need not be coupled, any other generative loss can be used, e.g., equation (8). But this is not ideal, since we would prefer to derive valid ratio losses from the same principle to make reasoning about correctness and optimality easier. We include this discussion to point out that while the formulation of equation (9) is generally applicable, the formulation (14), while useful for ratio estimation, is not useful for generative learning. The difficulty of specifying stable and correct generative losses is a common theme of work in GANs; we will see a similar difficulty in the next section.

The equivalence between divergence minimisation and class probability estimation is also widely discussed, and most notably developed by Reid and Williamson (2011) using the tools of weighted integral measures, and more recently by Menon and Ong (2016). This divergence minimisation viewpoint (9) was used in f-GANs and explored in depth by Nowozin et al. (2016) who provide a detailed description, and explore many of the objectives that become available and practical guidance.

2.4 RATIO MATCHING

A third approach is to directly minimise the error between the true density ratio and an estimate of it. If we denote the true density ratio as $r^*(\mathbf{x}) = {}^{p^*(\mathbf{x})}/_{q_\theta(\mathbf{x})}$ and its approximation as $r_\phi(\mathbf{x})$, we can define a loss using the squared error:

$$\mathcal{L} = \frac{1}{2}\int q_\theta(\mathbf{x})(r(\mathbf{x}) - r^*(\mathbf{x}))^2 d\mathbf{x} = \tfrac{1}{2}\mathbb{E}_{q_\theta(\mathbf{x})}[r_\phi(\mathbf{x})^2] - \mathbb{E}_{p^*(\mathbf{x})}[r_\phi(\mathbf{x})] + \tfrac{1}{2}\mathbb{E}_{p^*(\mathbf{x})}[r^*(\mathbf{x})] \quad (15)$$

$$= \tfrac{1}{2}\mathbb{E}_{q_\theta(\mathbf{x})}[r_\phi(\mathbf{x})^2] - \mathbb{E}_{p^*(\mathbf{x})}[r_\phi(\mathbf{x})] \quad s.t.\ r_\phi(\mathbf{x}) \geq 0, \quad (16)$$

where the final objective is obtained by ignoring terms independent of $r_\phi(\mathbf{x})$ and is used to derive ratio and generative losses. When used to learn the ratio function, Sugiyama et al. (2012b) refer to this objective as KL importance estimation procedure (KLIEP). Concurrently with this work, Uehara et al. (2016) recognised the centrality of the density ratio, the approach for learning by ratio matching, and its connections to GANs (Goodfellow et al., 2014) and f-GANs (Nowozin et al., 2016), and provide useful guidance on practical use of ratio matching.

We can generalise (16) to loss functions beyond the squared error using the Bregman divergence for density ratio estimation (Sugiyama et al., 2012a; Uehara et al., 2016), and is the unifying tool exploited in previous work (Reid and Williamson, 2011; Sriperumbudur et al., 2009; Sugiyama et al., 2012a; Menon and Ong, 2016). This leads to a minimisation of the Bregman divergence B_f between ratios:

$$B_f(r^*(\mathbf{x})\|r_\phi(\mathbf{x})) = \int \left(f(r^*(\mathbf{x})) - f(r_\phi(\mathbf{x})) - f'(r_\phi(\mathbf{x}))\left[r^*(\mathbf{x}) - r_\phi(\mathbf{x})\right]\right) q_\theta(\mathbf{x}) d\mathbf{x} \quad (17)$$

$$= \mathbb{E}_{q_\theta(\mathbf{x})}\left[r_\phi(\mathbf{x}) f'(r_\phi(\mathbf{x})) - f(r_\phi(\mathbf{x}))\right] - \mathbb{E}_{p^*}[f'(r_\phi(\mathbf{x}))] + D_f[p^*(\mathbf{x})\|q_\theta(\mathbf{x})] \quad (18)$$

$$= \mathcal{L}_B(r_\phi(\mathbf{x})) + D_f[p^*(\mathbf{x})\|q_\theta(\mathbf{x})], \quad (19)$$

where we have used $p^* = r^* q_\theta$, and D_f is the f-divergence defined in equation (9). We can derive a ratio loss from (19) by extracting all the terms in r_ϕ, leading to the minimisation of $\mathcal{L}_B(r_\phi)$. The role of the discriminator in GANs is again taken by the ratio r that provides information about the relationship between the two distributions. This ratio loss, as pointed out by Uehara et al. (2016), is *equivalent* to the ratio loss we derived using divergence minimisation (10), since:

$$\mathcal{L}_B(r_\phi(\mathbf{x})) = E_{p^*}[-f'(r_\phi(\mathbf{x}))] + \mathbb{E}_{q_\theta(\mathbf{x})}[r_\phi(\mathbf{x}) f'(r_\phi(\mathbf{x})) - f(r_\phi(\mathbf{x}))] \quad (20)$$

$$= E_{p^*}[-f'(r_\phi(\mathbf{x}))] + \mathbb{E}_{q_\theta(\mathbf{x})}[f^\dagger(f'(r_\phi(\mathbf{x})))]. \quad (21)$$

The equivalence of the second terms in (20) and (21) can be derived by using the definition of the dual function: $f^\dagger(f'(x)) = \max_r r f'(x) - f(r)$. The maximum is attained at $x = r$, leading to the identity $f^\dagger(f'(r_\phi(\mathbf{x}))) = r_\phi(\mathbf{x}) f'(r_\phi(\mathbf{x})) - f(r_\phi(\mathbf{x}))$.

If we follow the strategy we used in previous sections to obtain a generative loss, by collecting the terms in equation (19) dependent on q_θ, we obtain:

$$\mathcal{L}(q_\theta) = \mathbb{E}_{q_\theta(\mathbf{x})}[r_\phi(\mathbf{x}) f'(r_\phi(\mathbf{x}))] - \mathbb{E}_{q_\theta(\mathbf{x})}[f(r_\phi(\mathbf{x}))] + D_f[p^*(\mathbf{x})||q_\theta(\mathbf{x})]. \quad (22)$$

But this does not lead to a useful generative loss since there are terms involving q_θ whose density is always unknown, similar to the difficulty we encountered with divergence minimisation in equation (14). Since we can use any generative loss that drives the ratio to one, we can employ other losses, like the alternatives described in section 2.2 . One approximation suggested by Uehara et al. (2016) is to assume $p^* \approx r_\phi q_\theta$, i.e. assume a near-optimal ratio, which reduces the f-divergence to:

$$D_f[p^*(\mathbf{x})\|q_\theta(\mathbf{x})] = \mathbb{E}_{q_\theta(\mathbf{x})}\left[f\left(\frac{p^*}{q_\theta(\mathbf{x})}\right)\right] \approx \mathbb{E}_{q_\theta(\mathbf{x})}\left[f\left(\frac{q_\theta(\mathbf{x}) r_\phi(\mathbf{x})}{q_\theta(\mathbf{x})}\right)\right] = \mathbb{E}_{q_\theta(\mathbf{x})}[f(r_\phi(\mathbf{x}))] \quad (23)$$

As a result, the last two terms in equation (22) cancel, leaving a generative loss that can be used for learning. Using ratio matching under the Bregman divergence we obtain the bi-level optimisation:

$$\textbf{Ratio loss:} \quad \min_\phi \mathbb{E}_{q_\theta(\mathbf{x})}[r_\phi(\mathbf{x}) f'(r_\phi(\mathbf{x})) - f(r_\phi(\mathbf{x}))] - \mathbb{E}_{p^*}[f'(r_\phi(\mathbf{x}))] \quad (24)$$

$$\textbf{Generative loss:} \quad \min_\theta \mathbb{E}_{q_\theta(\mathbf{x})}[r_\phi(\mathbf{x}) f'(r_\phi(\mathbf{x}))] \quad (25)$$

2.5 MOMENT MATCHING

A final approach for testing is to evaluate whether the moments of the distributions p^* and q are the same, i.e. by moment matching. We compare the moments of the two distributions by minimising their distance, using test statistics $s(\mathbf{x})$ that provides the moments of interest:

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) = (\mathbb{E}_{p^*(\mathbf{x})}[s(\mathbf{x})] - \mathbb{E}_{q_\theta(\mathbf{x})}[s(\mathbf{x})])^2 = (\mathbb{E}_{p^*(\mathbf{x})}[s(\mathbf{x})] - \mathbb{E}_{q(\mathbf{z})}[s(\mathcal{G}(\mathbf{z}; \boldsymbol{\theta}))])^2 \tag{26}$$

The choice of test statistics $s(\mathbf{x})$ is critical, since ideally, we wish to match all moments of the two distributions. When the functions $s(\mathbf{x})$ are defined within a reproducing kernel Hilbert space, we obtain kernel-based forms of these objectives, which are highly flexible and allow easy handling of data such as images, strings and graphs. The objective (26) can then be re-expressed in closed-form in terms of kernel functions, leading to the maximum mean discrepancy criterion (MMD). The role of the MMD for learning in implicit generative models defined by neural networks has been explored by both Li et al. (2015) and Dziugaite et al. (2015). While typically expensive to compute, there are now highly efficient approaches for computing the MMD (Chwialkowski et al., 2015). The objective using feature matching by Salimans et al. (2016) is a form of moment matching that defines the test statistics using intermediate layers of the discriminator function. And a further set of objective functions is possible by framing this problem as one of *density difference* (rather than ratio) estimation (Sugiyama et al., 2013).

The other significant body of research that focusses on learning in implicit generative models using the lens of moment matching is approximate Bayesian computation (ABC) (Marin et al., 2012). The models in ABC are often related to problems in population genetics and ecology, where knowledge of these systems leads to ABC being easily exploited to learn simulator parameters. The ABC literature has widely-explored the use of test statistics, including fixed functions and kernels, and Markov chain Monte Carlo methods to learn the posterior distribution of the parameters. There is a great deal of opportunity for exchange between GANs, ABC and ratio estimation in aspects of scalability, applications, and theoretical understanding. A further insight obtained from a moment-matching approach is that other empirical measures such as the Dudley and Wasserstein distances can easily be considered, establishing even further connections to other growing areas, such as optimal transport (Frogner et al., 2015; Sriperumbudur et al., 2009).

The moment matching approach can be generalised by representing them as an integral probability metric (Sriperumbudur et al., 2009), which is what makes it possible to describe the relationships and ways of transforming from density ratio estimation using using moment-matching, f-divergences, and class-probability estimation. Sriperumbudur et al. (2009) showed that f-divergences and integral probability metrics (MMD, Wasserstein) intersect only at the total variation distance. As we described previously, by cleverly formulating our problem as a Bregman divergence minimisation, we show that all the methods we described for density ratio estimation are very closely related (Sugiyama et al., 2012a; Reid and Williamson, 2011; Sriperumbudur et al., 2009; Sugiyama et al., 2012b; Uehara et al., 2016).

3 DISCUSSION

By using an inferential principle driven by hypothesis testing, we have been able to develop a number of indirect methods for learning the parameters of generative models. These methods do not compute the probability of the data or posterior distributions over latent variables, but instead only involve relative statements of probability by comparing populations of data from the generative model to observed data. This view allows us to better understand how algorithms such as generative adversarial networks, approximate Bayesian computation, noise-contrastive estimation, and density ratio estimation are related. Ultimately, these techniques make it possible for us to make contributions to applications in climate and weather, economics, population genetics, and epidemiology, all areas whose principal tools are implicit generative models.

Distinction between implicit and prescribed models. The distinction between implicit and prescribed models is useful to keep in mind for at least two reasons: the choice of model has direct implications on the types of learning and inferential principles that can be called upon; and it makes explicit that there are many different ways in which to specify a model that captures our beliefs about data generating processes. Any implicit model can be easily turned into a prescribed model by adding a simple likelihood function (noise model) on the generated outputs, so the distinction is

not essential. And models with likelihood functions also regularly face the problem of intractable marginal likelihoods. But the specification of a likelihood function provides knowledge of p^* that leads to different algorithms by exploiting this knowledge, e.g., NCE resulting from class-probability based testing in un-normalised models (Gutmann and Hyvärinen, 2012), or variational lower bounds for directed graphical models. We strive to maintain a clear distinction between the choice of model, choice of inference, and the resulting algorithm, since it is through such a structured view that we can best recognise the connections between research areas that rely on the same sets of tools.

Model misspecification and non-maximum likelihood methods. Once we have made the choice of an implicit generative model, we cannot use likelihood-based techniques, which then makes testing and estimation-by-comparison appealing. What is striking, is that this leads us to principles for parameter learning that do not require inference of any underlying latent variables, side-stepping one of the major challenges in statistical practice. This piques our interest in more general approaches for non-maximum likelihood and likelihood-free estimation methods, of which there is much work (Lyu, 2011; Gutmann and Hyvärinen, 2012; Hall, 2005; Marin et al., 2012; Frogner et al., 2015). We often deal with misspecified models where q_θ cannot represent p^*, and non maximum likelihood methods could be a more robust choice depending on the task (see figure 1 in (Huszár, 2015) for an illustrative example).

Choice of loss function. This density ratio viewpoint has led us to derive multiple different objective functions and an understanding of how they are related. But it has left us unsatisfied since we have not gained the insight needed to choose between them. Sugiyama et al. (2012a) make statements on this choice in the simpler setting where only the density ratio is to be estimated. But when we wish to learn implicit generative models, this choice, at present, remains unclear.

Perceptual losses. Several authors have also proposed using pre-trained discriminative networks to define the test functions since the difference in activations (of say a pre-trained VGG classifier) can better capture perceptual similarity than the reconstruction error in pixel space. This provides a strong motivation for further research into *joint* models of images and labels. However, it is not completely unsupervised as the pre-trained discriminative network contains information about labels and invariances. This makes evaluation difficult since we lack good metrics and can only fairly compare to other joint models that use both label and image information.

Bayesian inference. We have mainly discussed point estimation methods for parameter learning. It is also desirable to perform Bayesian inference in implicit models, where we learn the posterior distribution over the model parameters $p(\boldsymbol{\theta}|\mathbf{x})$, allowing knowledge of parameter uncertainty to be used in risk minimisation and other decision-making tasks. This is the aim of approximate Bayesian computation (ABC) (Marin et al., 2012). The most common approach for thinking about ABC is through moment-matching, but as we explored, there are other approaches available. An approach through class-probability estimation is appealing and leads to classifier ABC (Gutmann et al., 2014). We have highly diverse approaches for Bayesian reasoning in prescribed models, and it is desirable to develop a similar breadth of choice for implicit models.

Evaluation. Our view suggests that value of the density ratio, a quantity we can always compute using the approaches described, can be a useful metric to report. But density ratio estimation has not yet illuminated a method for consistent evaluation of the generative model that is learnt. Without such a measure, it is hard to make fair comparisons; but this is a problem faced by all related fields, and establishing such measures is an important contribution to be made. Furthermore, we have also not yet gained the tools to make theoretical statements that allow us to assess the correctness of the model learning framework, although theoretical developments in the literature on approximate Bayesian computation may help in this regard, e.g., Frazier et al. (2016).

Non-differentiable models. We have restricted our development to implicit models that are differentiable. In many practical applications, the implicit model (or simulator) will be non-differentiable, discrete or defined in other ways, such as through a stochastic differential equation . The stochastic optimisation problem we are generally faced with (for differentiable and non-differentiable models), is to compute $\Delta = \nabla_\theta \mathbb{E}_{q_\theta(\mathbf{x})}[f(\mathbf{x})]$, the gradient of the expectation of a function. As our exposition followed, when the implicit model is differentiable, the pathwise derivative estimator can be used, i.e. $\Delta = \mathbb{E}_{q(\mathbf{z})}[\nabla_\theta f(\mathcal{G}_\theta(\mathbf{z}))]$ by rewriting the expectation in terms of the known and easy to sample distribution $q(\mathbf{z})$. It is commonly assumed that when we encounter non-differentiable functions that the score function estimator (or likelihood ratio or reinforce estimator) can be used; the score-function

estimator is $\Delta = \mathbb{E}_{q_\theta(\mathbf{x})}[f(\mathbf{x})\nabla_\theta \log q_\theta(\mathbf{x})]$. For implicit models, we do not have knowledge of the density $q(\mathbf{x})$ whose log derivative we require, making this estimator inapplicable. This leads to the first of three tools available for non-differentiable models: the use of the *weak derivative and related stochastic finite difference* estimators, which require forward-simulation only and compute gradients by perturbation of the parameters (Glasserman, 2003; Fu, 2005).

The two other approaches are: *moment matching and ABC-MCMC* (Marjoram et al., 2003), which has been successful for many problems with moderate dimension; and the natural choice of *gradient-free optimisation* methods, which include familiar tools such as Bayesian optimisation (Gutmann and Corander, 2016), evolutionary search like CMA-ES, and the Nelder-Mead method, amongst others (Conn et al., 2009). For all three approaches, new insights will be needed to help scale to high-dimensional data with complex dependency structures.

Ultimately, these concerns serve to highlight the many opportunities that remain for advancing our understanding of inference and parameter learning in implicit generative models.

REFERENCES

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.

S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer. Multi-task learning for HIV therapy screening. In *Proceedings of the 25th international conference on Machine learning*, pages 56–63. ACM, 2008.

A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*, 2005.

K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

B. Colson, P. Marcotte, and G. Savard. An overview of bilevel optimization. *Annals of operations research*, 153(1):235–256, 2007.

A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to derivative-free optimization*, volume 8. Siam, 2009.

K. Cranmer, J. Pavez, and G. Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

A. Creswell and A. A. Bharath. Task specific adversarial cost function. *arXiv preprint arXiv:1609.08661*, 2016.

I. Csiszàr. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

L. Devroye. Non-uniform random variate generation. *Handbooks in operations research and management science*, 13:83–121, 2006.

P. J. Diggle and R. J. Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 193–227, 1984.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

D. T. Frazier, G. M. Martin, C. P. Robert, and J. Rousseau. Asymptotic properties of approximate Bayesian computation. *arXiv preprint arXiv:1607.06903*, 2016.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

M. C. Fu. Stochastic gradient estimation. Technical report, DTIC Document, 2005.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1):32–40, 1975.

P. Glasserman. *Monte Carlo methods in financial engineering*, volume 53. Springer Science & Business Media, 2003.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

M. U. Gutmann and J. Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125):1–47, 2016.

M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(Feb): 307–361, 2012.

M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. Statistical inference of intractable generative models via classification. *arXiv preprint arXiv:1407.4981*, 2014.

A. R. Hall. *Generalized method of moments*. Oxford University Press, 2005.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, pp 495–497, 10th printing, 2nd edition, 2013.

F. Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430): 773–795, 1995.

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015.

F. Liese and I. Vajda. f-divergences: Sufficiency, deficiency and testing of hypotheses. *Advances in Inequalities from Probability Theory and Statistics*, pages 113–158, 2008.

D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests for GAN evaluation and causal discovery. *arXiv preprint arXiv:1610.06545*, 2016.

S. Lyu. Unifying non-maximum likelihood learning objectives with minimum KL contraction. In *Advances in Neural Information Processing Systems*, pages 64–72, 2011.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

A. Menon and C. S. Ong. Linking losses for density ratio and class-probability estimation. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 304–313, 2016.

T. Minka. Divergence measures and message passing. Technical report, MSR, 2005.

R. M. Neal. Computing likelihood functions for high-energy physics experiments when distributions are defined by simulators with nuisance parameters. In *PHYSTAT LHC Workshop on Statistical Issues for LHC Physics*, page 111, 2008.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016.

A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75 (2):237–249, 1988.

J. Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

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

M. D. Reid and R. C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(Mar):731–817, 2011.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. *arXiv preprint arXiv:1606.03498*, 2016.

C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár. Amortised MAP inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet. On integral probability metrics, φ-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

M. Sugiyama, T. Suzuki, and T. Kanamori. Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012a.

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

M. Sugiyama, T. Kanamori, T. Suzuki, M. C. du Plessis, S. Liu, and I. Takeuchi. Density-difference estimation. *Neural Computation*, 25(10):2734–2775, 2013.

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

J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.