Generative neural samplers are probabilistic models that implement sampling using feedforward neural networks: they take a random input vector and produce a sample from a probability distribution defined by the network weights. These models are expressive and allow efficient computation of samples and derivatives, but cannot be used for computing likelihoods or for marginalization. The generative-adversarial training method allows to train such models through the use of an auxiliary discriminative neural network. We show that the generative-adversarial approach is a special case of an existing more general variational divergence estimation approach. We show that any $f$-divergence can be used for training generative neural samplers. We discuss the benefits of various choices of divergence functions on training complexity and the quality of the obtained generative models.
Because $D_{BS}(P\|Q)$ is a proper divergence measure between distributions this implies that the true distribution $P$ can be approximated well in case there are sufficient training samples and the model class $\mathcal{Q}$ is rich enough to represent $P$.

In this work we show that the principle of GANs is more general and we can extend the variational divergence estimation framework proposed by Nguyen et al. \cite{nguyen2018variational} to recover the GAN training objective and generalize it to arbitrary f-divergences.

More concretely, we make the following contributions over the state-of-the-art:

- We derive the GAN training objectives for all f-divergences and provide as example additional divergence functions, including the Kullback-Leibler and Pearson divergences.
- We simplify the saddle-point optimization procedure of Goodfellow et al. \cite{goodfellow2014generative} and provide a theoretical justification.
- We provide experimental insight into which divergence function is suitable for estimating generative neural samplers for natural images.

## 2 Method

We first review the divergence estimation framework of Nguyen et al. \cite{nguyen2018variational} which is based on f-divergences. We then extend this framework from divergence estimation to model estimation.

### 2.1 The f-Divergence Family

Statistical divergences such as the well-known Kullback-Leibler divergence measure the difference between two given probability distributions. A large class of different divergences are the so-called f-divergences \cite{kulikov1995generalized, craven1958extension}, also known as the Ali-Silvey distances \cite{ali1966statistical}. Given two distributions $P$ and $Q$ that possess, respectively, an absolutely continuous density function $p$ and $q$ with respect to a base measure $dx$ defined on the domain $\mathcal{X}$, we define the f-divergence,

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

where the generator function $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ is a convex, lower-semicontinuous function satisfying $f(1) = 0$. Different choices of $f$ recover popular divergences as special cases in \cite{nguyen2018variational}. We illustrate common choices in Table \ref{table:divergences}. See supplementary material for more divergences and plots.

### 2.2 Variational Estimation of f-Divergences

Nguyen et al. \cite{nguyen2018variational} derive a general variational method to estimate f-divergences given only samples from $P$ and $Q$. An equivalent result has also been derived by Reid and Williamson \cite{reid2009variational}. We will extend these results from merely estimating a divergence for a fixed model to estimating model parameters. We call this new method variational divergence minimization (VDM) and show that generative-adversarial training is a special case of our VDM framework.

For completeness, we first provide a self-contained derivation of Nguyen et al.’s divergence estimation procedure. Every convex, lower-semicontinuous function $f$ has a convex conjugate function $f^*$, also known as Fenchel conjugate \cite{rockafellar1970convex}. This function is defined as

$$f^*(t) = \sup_{u \in \text{dom}^* f} \{ ut - f(u) \}. \quad (3)$$

The function $f^*$ is again convex and lower-semicontinuous and the pair $(f, f^*)$ is dual to another in the sense that $f^{**} = f$. Therefore, we can also represent $f$ as $f(u) = \sup_{t \in \text{dom}^* f} \{ tu - f^*(t) \}$.

Nguyen et al. leverage the above variational representation of $f$ in the definition of the f-divergence to obtain a lower bound on the divergence,

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) \sup_{t \in \text{dom}^* f} \left\{ t \frac{p(x)}{q(x)} - f^*(t) \right\} \, dx$$

$$\geq \sup_{T \in \mathcal{T}} \left( \int_{\mathcal{X}} p(x) T(x) \, dx - \int_{\mathcal{X}} q(x) f^*(T(x)) \, dx \right)$$

$$= \sup_{T \in \mathcal{T}} \left( \mathbb{E}_{x \sim P} [T(x)] - \mathbb{E}_{x \sim Q} [f^*(T(x))] \right), \quad (4)$$

2
We now use the variational lower bound (4) on the 

We list common 

where 

To apply the variational objective (6) for different 

To optimize (6) on a given finite training data set, we approximate the expectations using minibatch 

variation of the lower bound in (4) w.r.t. 

We approximate 

E 

... 

2.4 Representation for the Variational Function

To apply the variational objective (6) for different f-divergences, we need to respect the domain dom\( f^* \) of the conjugate functions f\(^*\). To this end, we assume that variational function \( T_\omega \) is represented in the form \( T_\omega(x) = g_f(V_\omega(x)) \) and rewrite the saddle objective (6) as follows:

\[
F(\theta, \omega) = E_{x \sim \mathcal{P}} \left[ g_f(V_\omega(x)) \right] + E_{x \sim Q_\theta} \left[ -f^*(g_f(V_\omega(x))) \right],
\]

where \( V_\omega : \mathcal{X} \rightarrow \mathbb{R} \) without any range constraints on the output, and \( g_f : \mathbb{R} \rightarrow \text{dom}\ f^* \) is an output activation function specific to the f-divergence used. In Table 2 we propose suitable output activation functions for the various conjugate functions f\(^*\) and their domains. Although the choice of \( g_f \) is somewhat arbitrary, we choose all of them to be monotone increasing functions so that a large output

1Note that for numerical implementation we recommend directly implementing the scalar function f\(^*(g_f(\cdot))\) robustly instead of evaluating the two functions in sequence; see Figure 2.
We can see the GAN objective, $Q_{\theta}$ w.r.t. $\theta$ where

$$
\frac{\partial}{\partial \theta} \left( -\log(1 + e^{-v}) \right) = \frac{1}{1 + e^{-v}} - 1
$$

The critical value $f'(1)$ can be interpreted as a classification threshold applied to $T(x)$ to distinguish between true and generated samples.

### Table 2

<table>
<thead>
<tr>
<th>Name</th>
<th>Output activation $g_f$</th>
<th>$\text{dom}_f$</th>
<th>Conjugate $f^*(t)$</th>
<th>$f'(1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kullback-Leibler (KL)</td>
<td>$v$</td>
<td>$\mathbb{R}$</td>
<td>$\exp(t - 1)$</td>
<td>1</td>
</tr>
<tr>
<td>Reverse KL</td>
<td>$- \exp(-v)$</td>
<td>$\mathbb{R}_+$</td>
<td>$-1 - \log(-t)$</td>
<td>$-1$</td>
</tr>
<tr>
<td>Pearson $\chi^2$</td>
<td>$v$</td>
<td>$\mathbb{R}$</td>
<td>$\frac{4}{t^2} + t$</td>
<td>0</td>
</tr>
<tr>
<td>Squared Hellinger</td>
<td>$1 - \exp(-v)$</td>
<td>$t &lt; 1$</td>
<td>$\frac{1}{t^2}$</td>
<td>0</td>
</tr>
<tr>
<td>Jensen-Shannon</td>
<td>$\log(2) - \log(1 + \exp(-v))$</td>
<td>$t &lt; \log(2)$</td>
<td>$-\log(2 - \exp(t))$</td>
<td>0</td>
</tr>
<tr>
<td>GAN</td>
<td>$-\log(1 + \exp(-v))$</td>
<td>$\mathbb{R}_+$</td>
<td>$-\log(1 - \exp(t))$</td>
<td>$-\log(2)$</td>
</tr>
</tbody>
</table>

Table 2: Recommended final layer activation functions and critical variational function level defined by $f'(1)$. The critical value $f'(1)$ can be interpreted as a classification threshold applied to $T(x)$ to distinguish between true and generated samples.

![Figure 1](image.png)

Figure 1: The two terms in the saddle objective (7) are plotted as a function of the variational function $V_\omega(x)$.

$V_\omega(x)$ corresponds to the belief of the variational function that the sample $x$ comes from the data distribution $P$ as in the GAN case; see Figure 1. It is also instructive to look at the second term $-f^*(g_f(v))$ in the saddle objective (7). This term is typically (except for the Pearson $\chi^2$ divergence) a decreasing function of the output $V_\omega(x)$ favoring variational functions that output negative numbers for samples from the generator.

We can see the GAN objective,

$$
F(\theta, \omega) = E_{x \sim P} [\log D_\omega(x)] + E_{z \sim Q_\theta} [\log(1 - D_\omega(z))],
$$

as a special instance of (7) by identifying each terms in the expectations of (7) and (8). In particular, choosing the last nonlinearity in the discriminator as the sigmoid $D_\omega(x) = 1/(1 + e^{-V_\omega(x)})$, corresponds to output activation function is $g_f(v) = -\log(1 + e^{-v})$; see Table 2.

#### 2.5 Example: Univariate Mixture of Gaussians

To demonstrate the properties of the different $f$-divergences and to validate the variational divergence estimation framework we perform an experiment similar to the one of [24].

**Setup.** We approximate a mixture of Gaussians by learning a Gaussian distribution. We represent our model $Q_\theta$ using a linear function which receives a random $z \sim \mathcal{N}(0, 1)$ and outputs $G_\theta(z) = \mu + \sigma z$, where $\theta = (\mu, \sigma)$ are the two scalar parameters to be learned. For the variational function $T_\omega$ we use a neural network with two hidden layers having 64 units each and tanh activations. We optimize the objective $F(\omega, \theta)$ by using the single-step gradient method presented in Section 3. In each step we sample batches of size 1024 from $p(x)$ and $p(z)$ and we use a step-size of $\eta = 0.01$ for updating both $\omega$ and $\theta$. We compare the results to the best fit provided by the exact optimization of $D_f(P||Q_\theta)$ w.r.t. $\theta$, which is feasible in this case by solving the required integrals in [2] numerically. We use $(\hat{\omega}, \hat{\theta})$ (learned) and $\theta^*$ (best fit) to distinguish the parameters sets used in these two approaches.

**Results.** The left side of Table 3 shows the optimal divergence and objective values $D_f(P||Q_{\theta^*})$ and $F(\hat{\omega}, \hat{\theta})$ as well as the corresponding (optimal) means and standard deviations. Note that the results are in line with the lower bound property, having $D_f(P||Q_{\theta^*}) \geq F(\hat{\omega}, \hat{\theta})$. There is a good correspondence between the gap in objectives and the difference between the fitted means and standard deviations. The right side of Table 3 shows the results of the following experiment: (1) we train $T_\omega$ and $Q_\theta$ using a particular divergence, then (2) we estimate the divergence and re-train $T_\omega$ while keeping $Q_\theta$ fixed. As expected, $Q_\theta$ performs best on the divergence it was trained with. We present further details and plots of the fitted Gaussians and variational functions in the supplementary materials.
We now discuss numerical methods to find saddle points of the objective (6).

Algorithm 1
Suppose that there is a saddle point is sufficiently smooth so that there is a constant conditions follows (the proof is given in the supplementary material):

For convenience, let’s define framework under these assumptions.

Motivated by the success of the alternating gradient method with a single inner step, we propose an even simpler algorithm shown in Algorithm 1. The algorithm differs from the original one in that there is no inner loop and the gradients with respect to \( \omega \) even simpler algorithm shown in Algorithm 1. The algorithm differs from the original one in that there is no inner loop and the gradients with respect to \( \omega \) and \( \theta \) are computed in a single back-propagation.

**Algorithm 1 Single-Step Gradient Method**

1: function SINGLESTEPGRADIENTITERATION(\( P, \theta, \omega, B, \eta \))
2: Sample \( X_P = \{x_1, \ldots, x_B\} \) and \( X_Q = \{x'_1, \ldots, x'_B\} \), from \( P \) and \( Q_\omega \), respectively.
3: Update: \( \omega^{t+1} = \omega + \eta \nabla_\omega F(\theta^t, \omega^t) \).
4: Update: \( \theta^{t+1} = \theta^t - \eta \nabla_\theta F(\theta^t, \omega^t) \).
5: end function

**Analysis.** Here we show that Algorithm 1 geometrically converges to a saddle point \( (\theta^*, \omega^*) \) if there is a neighborhood around the saddle point in which \( F \) is strongly convex in \( \theta \) and strongly concave in \( \omega \). These assumptions are similar to those made in [10]. Formally, we assume:

\[
\nabla_\theta F(\theta^*, \omega^*) = 0, \quad \nabla_\omega F(\theta^*, \omega^*) = 0, \quad \nabla^2_\theta F(\theta, \omega) \geq \delta I, \quad \nabla^2_\omega F(\theta, \omega) \leq -\delta I, \quad (9)
\]

for \((\theta, \omega)\) in the neighborhood of \((\theta^*, \omega^*)\). Note that although there could be many saddle points that arise from the structure of deep networks [6], they would not qualify as the solution of our variational framework under these assumptions.

For convenience, let’s define \( \pi^t = (\theta^t, \omega^t) \). Now the convergence of Algorithm 1 can be stated as follows (the proof is given in the supplementary material):

**Theorem 1.** Suppose that there is a saddle point \( \pi^* = (\theta^*, \omega^*) \) with a neighborhood that satisfies conditions (9). Moreover, we define \( J(\pi) = \frac{1}{2} \| \nabla F(\pi) \|_2^2 \) and assume that in the above neighborhood, \( F \) is sufficiently smooth so that there is a constant \( L > 0 \) such that \( \| \nabla J(\pi') - \nabla J(\pi) \|_2 \leq L \| \pi' - \pi \|_2 \) for any \( \pi, \pi' \) in the neighborhood of \( \pi^* \). Then using the step-size \( \eta = \delta / L \) in Algorithm 1 we have

\[
J(\pi^t) \leq \left(1 - \frac{\delta^2}{L} \right)^t J(\pi^0).
\]
3.2 Practical Considerations

Here we discuss principled extensions of the heuristic proposed in \[10\] and real/fake statistics discussed by Larsen and Sønderby\[33\]. Furthermore we discuss practical advice that slightly deviate from the principled viewpoint.

Goodfellow et al. \[10\] noticed that training GAN can be significantly sped up by maximizing $\mathbb{E}_{x \sim Q_0} \left[ \log D_\theta(x) \right]$ instead of minimizing $\mathbb{E}_{x \sim Q_0} \left[ \log \left( 1 - D_\theta(x) \right) \right]$ for updating the generator. In the more general f-GAN Algorithm \[1\], this means that we replace line 4 with the update

$$ \theta^{t+1} = \theta^t + \eta \nabla_{\theta} \mathbb{E}_{x \sim Q_\theta} \left[ g_f(V_{\omega^t}(x)) \right], \quad (10) $$

thereby maximizing the variational function output on the generated samples. We can show that this transformation preserves the stationary point as follows (which is a generalization of the argument in \[10\]): note that the only difference between the original direction (line 4) and (10) is the scalar factor $f^\omega(T_\theta(x))$, which is the derivative of the conjugate function $f^\ast$. Since $f^\ast$ is the inverse of $f^\omega$ (see Cor. 1.4.4, Chapter E, \[15\]), if $T = T^\ast$, using \[5\], we can see that this factor would be the density ratio $p(x)/q(x)$, which would be one at the stationary point. We found this transformation useful also for other divergences. We found Adam \[17\] and gradient clipping to be useful especially in the large scale experiment on the LSUN dataset.

The original implementation \[10\] of GAN’s and also Larsen and Sønderby monitor certain real and fake statistics, which are defined as the true positive and true negative rates of the variational function viewing it as a binary classifier. Since our output activation $g_f$ are all monotone, we can derive similar statistics for any $f$-divergence by only changing the decision threshold. Due to the link between the density ratio and the variational function \[5\], the threshold lies at $f^\omega(1)$ (see Table 2). That is, we can interpret the output of the variational function as classifying the input $x$ as a true sample if the variational function $T_\omega(x)$ is larger than $f^\omega(1)$, and classifying it as a generator sample otherwise.

4 Experiments

We now train generative neural samplers based on VDM on the MNIST and LSUN datasets.

**MNIST Digits.** We use the MNIST training data set (60,000 samples, 28-by-28 pixel images) to train the generator and variational function model proposed in \[10\] for various $f$-divergences. With $z \sim \text{Uniform}_{100}(−1, 1)$ as input, the generator model has two linear layers each followed by batch normalization and ReLU activation and a final linear layer followed by the sigmoid function. The variational function $V_{\omega}(x)$ has three linear layers with exponential linear unit \[4\] in between. The final activation is specific to each divergence and listed in Table 2. As in \[27\] we use Adam with a learning rate of $\alpha = 0.0002$ and update weight $\beta = 0.5$. We use a batchsize of 4096, sampled from the training set without replacement, and train each model for one hour. We also compare against variational autoencoders \[18\] with 20 latent dimensions.

**Results and Discussion.** We evaluate the performance using the kernel density estimation (Parzen window) approach used in \[10\]. To this end, we sample 16k images from the model and estimate a Parzen window estimator using an isotropic Gaussian kernel bandwidth using three fold cross validation. The final density model is used to evaluate the average log-likelihood on the MNIST test set (10k samples). We show the results in Table 4 and some samples from our models in Figure 2.

The use of the KDE approach to log-likelihood estimation has known deficiencies \[33\]. In particular, for the dimensionality used in MNIST ($d = 784$) the number of model samples required to obtain accurate log-likelihood estimates is infeasibly large. We found a large variability (up to 50 nats) between multiple repetitions. As such the results are not entirely conclusive. We also trained the same KDE estimator on the MNIST training set, achieving a significantly higher holdout likelihood. However, it is reassuring to see that the model trained for the Kullback-Leibler divergence indeed achieves a high holdout likelihood compared to the GAN model.

\[5\] Available at https://github.com/goodfeli/algorithm

\[10\] Available at http://torch.ch/blog/2015/11/13/gan.html
Table 4: Kernel Density Estimation evaluation on the MNIST test data set. Each KDE model is build from 16,384 samples from the learned generative model. We report the mean log-likelihood on the MNIST test set (n = 10,000) and the standard error of the mean. The KDE MNIST result is using 60,000 MNIST training images to fit a single KDE model.

<table>
<thead>
<tr>
<th>Training divergence</th>
<th>KDE $\langle LL \rangle$ (nats)</th>
<th>± SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kullback-Leibler</td>
<td>416</td>
<td>5.62</td>
</tr>
<tr>
<td>Reverse Kullback-Leibler</td>
<td>319</td>
<td>8.36</td>
</tr>
<tr>
<td>Pearson $\chi^2$</td>
<td>429</td>
<td>5.53</td>
</tr>
<tr>
<td>Neyman $\chi^2$</td>
<td>300</td>
<td>8.33</td>
</tr>
<tr>
<td>Squared Hellinger</td>
<td>-708</td>
<td>18.1</td>
</tr>
<tr>
<td>Jeffrey</td>
<td>-2101</td>
<td>29.9</td>
</tr>
<tr>
<td>Jensen-Shannon</td>
<td>367</td>
<td>8.19</td>
</tr>
<tr>
<td>GAN</td>
<td>305</td>
<td>8.97</td>
</tr>
<tr>
<td>Variational Autoencoder</td>
<td>445</td>
<td>5.36</td>
</tr>
<tr>
<td>KDE MNIST train (60k)</td>
<td>502</td>
<td>5.99</td>
</tr>
</tbody>
</table>

LSUN Natural Images. Through the DCGAN work [27] the generative-adversarial approach has shown real promise in generating natural looking images. Here we use the same architecture as in [27] and replace the GAN objective with our more general f-GAN objective.

We use the large scale LSUN database [35] of natural images of different categories. To illustrate the different behaviors of different divergences we train the same model on the classroom category of images, containing 168,103 images of classroom environments, rescaled and center-cropped to 96-by-96 pixels.

Setup. We use the generator architecture and training settings proposed in DCGAN [27]. The model receives $z \in \text{Uniform}_{d_{rand}}(-1, 1)$ and feeds it through one linear layer and three deconvolution layers with batch normalization and ReLU activation in between. The variational function is the same as the discriminator architecture in [27] and follows the structure of a convolutional neural network with batch normalization, exponential linear units [4] and one final linear layer.

Results. Figure 3 shows 16 random samples from neural samplers trained using GAN, KL, and squared Hellinger divergences. All three divergences produce equally realistic samples. Note that the difference in the learned distribution $Q_{\theta}$ arise only when the generator model is not rich enough.

Figure 3: Samples from three different divergences.

5 Related Work

We now discuss how our approach relates to existing work. Building generative models of real world distributions is a fundamental goal of machine learning and much related work exists. We only discuss work that applies to neural network models.
Mixture density networks \cite{bishop1994mixture} are neural networks which directly regress the parameters of a finite parametric mixture model. When combined with a recurrent neural network this yields impressive generative models of handwritten text \cite{graves2013speech}.

NADE \cite{mnih2014neural} and RNADE \cite{zheng2015neural} perform a factorization of the output using a predefined and somewhat arbitrary ordering of output dimensions. The resulting model samples one variable at a time conditioning on the entire history of past variables. These models provide tractable likelihood evaluations and compelling results but it is unclear how to select the factorization order in many applications.

Diffusion probabilistic models \cite{grosse2016probabilistic} define a target distribution as a result of a learned diffusion process which starts at a trivial known distribution. The learned model provides exact samples and approximate log-likelihood evaluations.

Noise contrastive estimation (NCE) \cite{gutmann2010noise} is a method that estimates the parameters of unnormalized probabilistic models by performing non-linear logistic regression to discriminate the data from artificially generated noise. NCE can be viewed as a special case of GAN where the discriminator is constrained to a specific form that depends on the model (logistic regression classifier) and the generator (kept fixed) is providing the artificially generated noise (see supplementary material).

The generative neural sampler models of \cite{malhotra2015generative} and \cite{chan2017variational} did not provide satisfactory learning methods; \cite{malhotra2015generative} used importance sampling and \cite{chan2017variational} expectation maximization. The main difference to GAN and to our work really is in the learning objective, which is effective and computationally inexpensive.

Variational auto-encoders (VAE) \cite{kingma2013auto,rezende2014stochastic} are pairs of probabilistic encoder and decoder models which map a sample to a latent representation and back, trained using a variational Bayesian learning objective. The advantage of VAEs is in the encoder model which allows efficient inference from observation to latent representation and overall they are a compelling alternative to $f$-GANs and recent work has studied combinations of the two approaches \cite{choi2016generative}.

As an alternative to the GAN training objective the work \cite{gretton2012kernel} and independently \cite{audebert2015kernel} considered the use of the kernel maximum mean discrepancy (MMD) \cite{gretton2006kernel, gretton2012kernel} as a training objective for probabilistic models. This objective is simpler to train compared to GAN models because there is no explicitly represented variational function. However, it requires the choice of a kernel function and the reported results so far seem slightly inferior compared to GAN. MMD is a particular instance of a larger class of probability metrics \cite{peyre2019computational} which all take the form $D(P, Q) = \sup_{T \in \mathcal{T}} |\mathbb{E}_{x \sim P}[T(x)] - \mathbb{E}_{x \sim Q}[T(x)]|$, where the function class $\mathcal{T}$ is chosen in a manner specific to the divergence. Beyond MMD other popular metrics of this form are the total variation metric (also an $f$-divergence), the Wasserstein distance, and the Kolmogorov distance.

A previous attempt to enable minimization of the KL-divergence in deep generative models is due to Goodfellow et al. \cite{goodfellow2014generative}, where an approximation to the gradient of the KL divergence is derived.

In \cite{kim2018generative} another generalization of the GAN objective is proposed by using an alternative Jensen-Shannon divergence that interpolates between the KL and the reverse KL divergence and has Jensen-Shannon as its mid-point. We discuss this work in more detail in the supplementary materials.

6 Discussion

Generative neural samplers offer a powerful way to represent complex distributions without limiting factorizing assumptions. However, while the purely generative neural samplers as used in this paper are interesting their use is limited because after training they cannot be conditioned on observed data and thus are unable to provide inferences.

We believe that in the future the true benefits of neural samplers for representing uncertainty will be found in discriminative models and our presented methods extend readily to this case by providing additional inputs to both the generator and variational function as in the conditional GAN model \cite{mirza2014conditional}.

We hope that the practical difficulties of training with saddle point objectives are not an underlying feature of the model but instead can be overcome with novel optimization algorithms. Further investigations, such as \cite{salimans2016improved}, are needed to investigate and hopefully overcome these difficulties.

Acknowledgements. We thank Ferenc Huszár for discussions on the generative-adversarial approach.
References