EECS 6322 Week 10, Paper 2
f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper proposes a variational divergence estimation approach, known as f-divergence, as a more general formulation of the generative-adversarial approach, which is used to train GANs. The f-divergence method can be used to train generative neural samplers with different choices of divergence functions.

Generative neural samplers are feedforward neural network models that generate samples from a given probability distribution. These include models such as Generative Adversarial Networks (GANs) and Variational Autoencoders. For GANs, the neural samplers are estimated by approximating the minimization of the symmetric Jensen-Shannon divergence. Since this Jensen-Shannon divergence is a divergence measure between two given distributions P and Q, it implies that the true distribution P can also be approximated if there are sufficient training samples and Q is rich enough to represent P. Hence, the authors propose a generalized approach for the GAN training objective using arbitrary f-divergences.

The f-divergence for any two distributions P and Q having continuous density functions p and q with respect to a base measure dx defined on the domain X can be given by: $D_f(P||Q) = \int_x q(x)\ f(\frac{p(x)}{q(x)})dx$ *where the generator function is a lower-semicontinuous convex function which satisfies f(1) = 0. The different choices of the* f *function lead to different popular divergences, commonly used in machine learning and statistics. The authors further propose variational divergence minimization (VDM) that estimates the model parameters for a given model by extending the existing methods to estimate the f-divergences given only samples from P and Q. For a given lower-semicontinuous convex function* f*, its convex conjugate function* $f*$ *can be given by:* $f*(t) = \sup_{u\ \epsilon\ dom_f} \{ut - f(u)\}$ *where* $f*$ *is also a lower-semicontinuous convex function and* $f** = f$*. Thus, the f-divergence can be represented in terms of this* $f*$ *function given by:* $D_f(P||Q) = \sup_{T\ \epsilon\ \mathcal{T}} (E_{x\sim P}[T(x)] - E_{x\sim Q}[f*(T(x))])$ *where* $\mathcal{T}$ *is an arbitrary class of functions* $T: X \rightarrow R$*. Under mild conditions on the function* f*, a lower bound can also be given by:* $T*(x) = f'(\frac{p(x)}{q(x)})$ *where* f' *denotes the first order derivative of* f*.*

Now, to obtain the Variational Divergence Minimization (VDM), a generative model Q_θ *is trained with parameters* θ *by finding a saddle-point of the f-GAN objective function given by:* $F(\theta,\omega) = E_{x\sim P}[T_\omega(x)] - Ex \sim Q_\theta[f*(T_\omega(x))]$ *where the model is minimized with respect to* θ *and maximized with respect to* ω*. To approximate this function F, both the expectations in the formula are approximated using minibatch samples from the training set and the generative model* Q_θ*. This variational objective function can be applied to different f-divergences by representing* $T_\omega(x)$ *by* $T_\omega(x) = g_f(V_\omega(x))$*, which modified the objective function as:* $F(\theta,\omega) = E_{x\sim P}[g_f(V_\omega(x))] - Ex \sim Q_\theta[f*(g_f(V_\omega(x)))]$ *where* g_f *is an output activation function which is specific to the f-divergence used, and* $V_\omega : X \rightarrow R$ *without any range constraints on the output. Thus, the GAN objective function given by:* $F(\theta,\omega) = E_{x\sim P}[log\ D_\omega(x)] - Ex \sim Q_\theta[log(1 - D_\omega(x))]$ *can be seen as a special case of this modified objective function.*

To demonstrate the properties of the different f-divergences, the model performs an experiment to approximate the Gaussian parameters of a mixture of Gaussian models. The results show that when the generative model is misspecified, the divergence function used for estimation has a strong influence on which model is learned. The authors also discuss a numeric single-step gradient descent method to find the saddle points of the objective function through a direct single-step optimization procedure. The algorithm performs gradient ascent on the objective function to optimize the parameter ω *and performs gradient descent to optimize the parameters* θ*. Thus, the authors prove that the algorithm geometrically converges to a saddle point* $(\theta*, \omega*)$ *if there is a neighborhood around the saddle point in which F is strongly convex in* θ *and strongly concave in* ω*. For practical implementation of updating the generator in GANs, the objective can be changed from maximizing* $Ex \sim Q_\theta[log(D_\omega(x))]$ *instead of minimizing* $Ex \sim Q_\theta[log(1 - D_\omega(x))]$*, thus speeding up the algorithm.*

The model with the Variational Divergence Minimization (VDM) was tested on several datasets such as MNIST and LSUN. For MNIST, a training dataset of 60k samples and 28-by-28 pixel images were used to train the generator and variational function model for different f-divergences. The performance of the model was evaluated using the kernel density estimation approach. The results from the experiment varied significantly between each repetition due to which they were not entirely conclusive. However, the model trained for the Kullback-Leibler divergence achieved a high holdout likelihood compared to the GAN model objective. For the

LSUN dataset, the model was trained on the classroom category of images, containing around 168k images of classroom environments, rescaled and center-cropped to 96-by-96 pixels. The results from different divergence functions such as GAN, KL, and squared Hellinger divergences produced equally realistic samples, having no major differences between them. Thus, although the modified objective function proposed in the paper is efficient and computationally inexpensive, the use of the generative neural samplers applied in the paper are limited since they cannot be conditioned on observed data after training and thus are unable to provide inferences.

II. Strengths

The f-divergence method gives a generalized objective function for training generative models. Though the model is similar to GANs, the objective function for GANs is a specific form of the objective function as proposed in the paper, which is also effective and computationally inexpensive as compared to GANs. Thus, the proposed work enables the application of different divergence functions to generative models and enables the comparison between different models and their respective divergence functions.

III. Weaknesses

The use of purely generative neural samplers used in the work proposed by the authors is limited since they cannot be conditioned on observed data after training and thus are unable to provide inferences. The results for the MNIST dataset were also not entirely conclusive by using the kernel density estimation (KDE) approach to log-likelihood estimation.

IV. Correctness

The claims and empirical methodology as proposed in the paper are correct and are supported by sufficient theoretical grounding and the experiment results as given in the paper. Some of the divergence functions have also been applied and have proved to be a specific form of the general objective function as proposed by the authors. The modification in the GANs algorithm as proposed has also proved to be computationally efficient, inexpensive, and converges to a local saddle point, thus proving the groundwork to be correct.

V. Clarity

The paper is well written in general with clear formatting and visuals. The tables, graphs and illustrations used for comparison between the results supplement the understanding of the reader. The theoretical proof, along with the analysis of the modification in the GANs algorithms have also been included to increase the clarity of the work proposed. Additionally, the resemblance and the differences from the related work makes it easier for the reader to understand the novelty of the work proposed.

VI. Relation to prior work

Prior work in building generative models include neural network models such as mixture density networks, NADE and RNADE, diffusion probabilistic models, noise contrastive estimation, variational auto-encoders, and GANs. These methods use different sampling and optimization methods such as importance sampling and expectation maximization. The main resemblance of the proposed work is to the generative adversarial networks (GANs) which uses a more specific version of the general objective function as proposed in the paper. The objective function as proposed by the authors is also effective and computationally inexpensive as compared to GANs.

VII. Reproducibility

Enough details are given in terms of the theoretical grounding, implementation details of the model in specific experiments, and the code implementation on GitHub to reproduce the major details of the work proposed by the authors.

VIII. Additional comments

The algorithm proposed for the Variational Divergence Minimization (VDM) requires a neighborhood around the saddle point in which F is strongly convex in θ and strongly concave in ω. However, such a scenario is not very common in some of the practical applications and hence further discussion on such scenarios would supplement the proposed work.