2022 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, AUG. 22–25, 2022, XI'AN, CHINA

DECENTRALIZED GAN TRAINING THROUGH DIFFUSION LEARNING

Zhuoyue Wang, Flávio R. M. Pavan, and Ali H. Sayed

Adaptive Systems Laboratory, EPFL

ABSTRACT

Most available studies on distributed GAN architectures focus on implementations with a fusion center. In this work, we propose a fully decentralized scheme by employing a diffusion strategy to train a network of GANs. We interpret the training procedure as a team competition problem and use the paradigm of competing adaptive networks to solve it. We explain that the local discriminators and generators will cluster around their respective centroids. We present simulation results to illustrate that the proposed strategy allows local agents to match the performance of the centralized GAN. More importantly, we also illustrate that local GANs are able to generate different types of images from a dataset, even when they are locally trained with a subset that does not contain all image types.

Index Terms— Generative adversarial networks, competing diffusion, decentralized training, distributed optimization

1. INTRODUCTION

Generative adversarial networks (GANs) are a powerful type of generative models whose training can be understood as an adversarial game [1]. Widely used in fake image generation tasks, GANs consist of two distinct network components known as the generator and discriminator. These components compete against each other, with the generator trying to drive the discriminator away from its objective. In practice, GANs can be employed in several different ways. Distributed approaches, in which a local GAN is assigned to every agent in a network, are of particular interest. Since only training parameters are shared—either between neighbors or with a central aggregator—the training of local generators can be improved while enhancing the privacy of local datasets.

Most existing studies proposing distributed GAN architectures focus on centralized settings with fusion centers. Recent literature has dealt with the centralized scenario where every agent is only equipped with a local discriminator, except for a central agent that is equipped with a generator [2, 3]. In this case, since discriminator agents do not have generators, they are incapable of generating fake samples by themselves. In a different, but still centralized approach, the work [4] extends GAN applications to federated learning. So far, however, only a few studies have investigated the training of a distributed GAN architecture in a decentralized manner. Although gossiping GANs [5] make use of a decentralized training scheme, they require each agent to be able to communicate with any other agent. Due to such requirement, this training approach is actually equivalent to centralized training.

In this paper, we propose a fully decentralized training algorithm for a network of GANs. The idea of this training approach is to allow each local GAN in the network to be able to generate all types of images from a dataset, even when these local GANs are only trained with a subset of the data that does not contain all image types. For example, consider the MNIST dataset [6] and assume each GAN in the network is trained with images of only two types of digits, e.g., one agent is trained with images of the digits 7 and 8, while a second agent is trained with images of the digits 1 and 3, and so forth. Can we devise a decentralized algorithm that allows the agents to generate fake images of digits for which the local GAN has not seen any samples during training? The procedure we devise here, as we proceed to explain, allows for this possibility. We will employ a diffusion strategy [7, 8] to solve an aggregated min-max problem considering that each agent is equipped with both a local discriminator and a local generator.

The paper is organized as follows. In Section 2, we formulate the network of GANs problem starting from a single agent case. In Section 3, we derive the diffusion training algorithm and, in Section 4, we present a brief convergence analysis based on network centroid representations. In Section 5, we provide simulation results for both homogeneous and non-homogeneous datasets. In Section 6, we present the conclusions of this work.

2. PROBLEM FORMULATION

2.1. Single agent case

The training of a single GAN involves solving a min-max problem of the form [1]:

$$\min_{w^{\mathrm{G}}} \max_{w^{\mathrm{D}}} J(w^{\mathrm{D}}, w^{\mathrm{G}}) \tag{1}$$

where w^{D} and w^{G} are the parameters of the discriminator and generator, respectively. The objective function is given by:

$$J(w^{\mathrm{D}}, w^{\mathrm{G}}) \triangleq \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(z)}\, \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}(x)}\, Q(w^{\mathrm{D}}, w^{\mathrm{G}}; \boldsymbol{x}, \boldsymbol{z}) \tag{2}$$

E-mail addresses: {zhuoyue.wang, flavio.pavan, ali.sayed}@epfl.ch

978-1-6654-8547-0/22/$31.00

where $\boldsymbol{x}$ and $\boldsymbol{z}$ denote the training data for the discriminator and the input noise for the generator, respectively, with distributions $p_{\boldsymbol{x}}(x)$ and $p_{\boldsymbol{z}}(z)$. Traditionally, the gain/loss function $Q(w^{\mathrm{D}}, w^{\mathrm{G}}; \boldsymbol{x}, \boldsymbol{z})$ takes the form [1]:

$$Q(w^{\mathrm{D}}, w^{\mathrm{G}}; \boldsymbol{x}, \boldsymbol{z}) = \log D(w^{\mathrm{D}}; \boldsymbol{x}) + \log(1 - D(w^{\mathrm{D}}; G(w^{\mathrm{G}}; \boldsymbol{z}))) \tag{3}$$

where $D(\cdot;\cdot)$ and $G(\cdot;\cdot)$ represent the outputs of the discriminator and generator, respectively.

2.2. Network of GANs

We consider the decentralized network setting shown in Fig. 1, in which K agents, each equipped with an individual GAN and distinct datasets, are allowed to share information and coordinate on a common task. Each agent k of the network seeks to solve a local min-max problem of the form:

$$\min_{w_k^{\mathrm{G}}} \max_{w_k^{\mathrm{D}}} J_k(w_k^{\mathrm{D}}, w_k^{\mathrm{G}}) \tag{4}$$

where w_k^{D} and w_k^{G} are the discriminator and generator parameters for agent k, whose local objective is defined as:

$$J_k(w_k^{\mathrm{D}}, w_k^{\mathrm{G}}) \triangleq \mathbb{E}_{\boldsymbol{z}_k \sim p_{\boldsymbol{z}}(z)} \, \mathbb{E}_{\boldsymbol{x}_k \sim p_{\boldsymbol{x}_k}(x_k)} \, Q(w_k^{\mathrm{D}}, w_k^{\mathrm{G}}; \boldsymbol{x}_k, \boldsymbol{z}_k). \tag{5}$$

Here, $\boldsymbol{x}_k$ represents the local training data at agent k, drawn from a sample space $\mathcal{X}_k$ and with distribution $p_{\boldsymbol{x}_k}(x_k)$. Assuming that all GANs in the network have the same architecture, the gain/loss function on the right-hand side of (5) becomes the same for all agents.

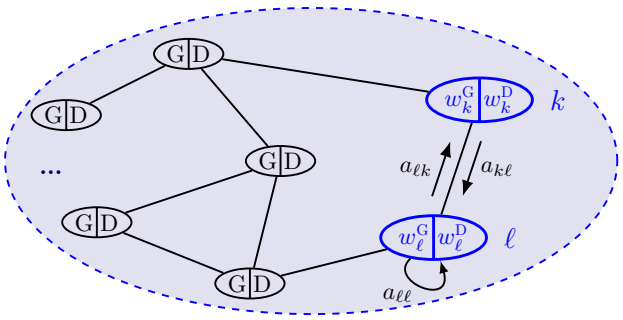


Fig. 1. A network of GANs where each node represents a local GAN. Agents interact over a graph topology.

Within the network, agents are connected by a graph with combination matrix $A = [a_{\ell k}] \in \mathbb{R}^{K \times K}$. The objective of the network is to solve:

$$\min_{w^{\mathrm{G}}} \max_{w^{\mathrm{D}}} J(w^{\mathrm{D}}, w^{\mathrm{G}}) \tag{6}$$

where the global objective is a weighted aggregate of the local costs:

$$J(w^{\mathrm{D}}, w^{\mathrm{G}}) = \sum_{k=1}^{K} p_k J_k(w^{\mathrm{D}}, w^{\mathrm{G}}), \tag{7}$$

and p_k are positive weights normalized to add up to one. These weights are specified further ahead.

For convenience, we define separate local objective functions for the discriminators and generators as follows:

$$J_k^{\mathrm{D}}(w^{\mathrm{D}}, w^{\mathrm{G}}) \triangleq -J_k(w^{\mathrm{D}}, w^{\mathrm{G}}), \tag{8a}$$

$$J_k^{\mathrm{G}}(w^{\mathrm{D}}, w^{\mathrm{G}}) \triangleq J_k(w^{\mathrm{D}}, w^{\mathrm{G}}). \tag{8b}$$

Definitions (8a)–(8b) allow us to interpret the local min-max problem (4) at the kth agent as equivalent to a minimization of both $J_k^{\mathrm{D}}(w_k^{\mathrm{D}}, w_k^{\mathrm{G}})$ over w_k^{D} and $J_k^{\mathrm{G}}(w_k^{\mathrm{D}}, w_k^{\mathrm{G}})$ over w_k^{G}.

3. DIFFUSION TRAINING

The training of the network of GANs can be understood as a particular case of the decentralized competing networks framework proposed in [9]. Based on this framework, we will first derive the training algorithm from the perspective of the discriminators, initially assuming that the generator parameter w^{G} is fixed and known. Considering (7) and (8a), the global problem (6) reduces to:

$$\min_{w^{\mathrm{D}}} \sum_{k=1}^{K} p_k J_k^{\mathrm{D}}(w^{\mathrm{D}}, w^{\mathrm{G}}). \tag{9}$$

This corresponds to a traditional decentralized optimization problem for the discriminators, whose solutions can be pursued by a number of algorithms for decentralized stochastic optimization [7]. We adopt the adapt-then-combine (ATC) diffusion strategy given by:

$$\boldsymbol{\phi}_{k,i}^{\mathrm{D}} = \boldsymbol{w}_{k,i-1}^{\mathrm{D}} - \mu \widehat{\nabla J_k^{\mathrm{D}}}(\boldsymbol{w}_{k,i-1}^{\mathrm{D}}, w^{\mathrm{G}}) \tag{10a}$$

$$\boldsymbol{w}_{k,i}^{\mathrm{D}} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\phi}_{\ell,i}^{\mathrm{D}} \tag{10b}$$

where μ is a small step size, $\widehat{\nabla J_k^{\mathrm{D}}}(\cdot,\cdot)$ denotes an estimate for the gradient of the local objective at the kth agent with respect to the discriminator parameters, and $\mathcal{N}_k$ is the set including the kth agent and agents connected to it with nonzero weights. After diffusion is run over the discriminators at all agents, we tackle the optimization problem for generators using a similar strategy. We additionally consider that, at the kth agent, the generator and discriminator have complete access to each other's parameters. In order to balance out the competition, at a given iteration we will allow for the repetition of diffusion n_{d} times over the discriminators and n_{g} times over the generators. Then, in a manner similar to [9], we obtain the construction of Algorithm 1.

4. CONVERGENCE ANALYSIS

Next, we comment briefly on the convergence of the Diffusion GAN algorithm. Our goal is twofold. First, to show that, after

Algorithm 1 Diffusion GAN

Initialize for each agent k: $w_k^{\mathrm{D}} \leftarrow w_{k,0}^{\mathrm{D}}$ and $w_k^{\mathrm{G}} \leftarrow w_{k,0}^{\mathrm{G}}$

while not done **do**
  for $n = 1, \ldots, n_{\mathrm{d}}$ **do**
    for each agent k **do**
      $\boldsymbol{\phi}_k^{\mathrm{D}} \leftarrow \boldsymbol{w}_k^{\mathrm{D}} - \mu\widehat{\nabla J_k^{\mathrm{D}}}(\boldsymbol{w}_k^{\mathrm{D}}, \boldsymbol{w}_k^{\mathrm{G}})$
      $\boldsymbol{w}_k^{\mathrm{D}} \leftarrow \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\phi}_\ell^{\mathrm{D}}$
    end for
  end for
  for $n = 1, \ldots, n_{\mathrm{g}}$ **do**
    for each agent k **do**
      $\boldsymbol{\phi}_k^{\mathrm{G}} \leftarrow \boldsymbol{w}_k^{\mathrm{G}} - \mu\widehat{\nabla J_k^{\mathrm{G}}}(\boldsymbol{w}_k^{\mathrm{D}}, \boldsymbol{w}_k^{\mathrm{G}})$
      $\boldsymbol{w}_k^{\mathrm{G}} \leftarrow \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\phi}_\ell^{\mathrm{G}}$
    end for
  end for
end while

sufficient iterations, all local discriminators and generators will cluster around their respective average parameters, i.e., centroids. This part of the analysis is based on results from [8, 10, 11]. Our second goal is to relate the min-max problem over the centroids with the problem of training a single GAN with samples from all local datasets. This will allow us to conclude that, when an agent does not have access to a particular type of data, it may still be able to generate it by learning from its neighbors.

4.1. Analysis Assumptions

Assumption 1 (Connectivity). The combination matrix A is primitive and doubly-stochastic. Thus, from the Perron-Frobenius theorem [10], matrix A has a single eigenvalue at one associated with a Perron eigenvector $p = \mathbb{1}/K$, whose entries determine the scalars in (7). Also, we have:

$$\lambda_2 \triangleq \rho\left(A - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}}\right) < 1 \tag{11}$$

where $\rho(\cdot)$ denotes the spectral radius of its matrix argument. □

We choose the Metropolis rule [7, 12] as our combination policy, whose combination matrix satisfies Assumption 1.

Assumption 2 (Smoothness). For each $k = 1, \ldots, K$, the gradients of discriminator and generator objectives are Lipschitz, i.e., for all $\alpha_1, \alpha_2, \beta_1, \beta_2$ they satisfy:

$$\|\nabla J_k^{\mathrm{D}}(\alpha_1, \beta) - \nabla J_k^{\mathrm{D}}(\alpha_2, \beta)\| \le L_{\mathrm{D}}\|\alpha_1 - \alpha_2\| \tag{12a}$$
$$\|\nabla J_k^{\mathrm{G}}(\alpha, \beta_1) - \nabla J_k^{\mathrm{G}}(\alpha, \beta_2)\| \le L_{\mathrm{G}}\|\beta_1 - \beta_2\| \tag{12b}$$

for some $L_{\mathrm{D}}, L_{\mathrm{G}} \ge 0$. □

Assumption 3 (Bounded gradients). For each agent $k = 1, \ldots, K$, the gradients of the discriminator and generator objectives are bounded, i.e., for all α, β:

$$\|\nabla J_k^{\mathrm{D}}(\alpha, \beta)\| \le B_{\mathrm{D}} \quad \text{and} \quad \|\nabla J_k^{\mathrm{G}}(\alpha, \beta)\| \le B_{\mathrm{G}} \tag{13}$$

for some $B_{\mathrm{D}}, B_{\mathrm{G}} \ge 0$. □

Assumption 4 (Gradient noise processes). We define the gradient noise for the discriminator and generator objectives as the difference between the true and approximate gradient vectors:

$$\boldsymbol{s}_{k,i}^{\mathrm{D}}(\boldsymbol{w}_{k,i-1}^{\mathrm{D}}, \boldsymbol{w}_{k,i-1}^{\mathrm{G}}) \triangleq \widehat{\nabla J_k^{\mathrm{D}}}(\boldsymbol{w}_{k,i-1}^{\mathrm{D}}, \boldsymbol{w}_{k,i-1}^{\mathrm{G}}) - \nabla J_k^{\mathrm{D}}(\boldsymbol{w}_{k,i-1}^{\mathrm{D}}, \boldsymbol{w}_{k,i-1}^{\mathrm{G}}), \tag{14a}$$
$$\boldsymbol{s}_{k,i}^{\mathrm{G}}(\boldsymbol{w}_{k,i-1}^{\mathrm{D}}, \boldsymbol{w}_{k,i-1}^{\mathrm{G}}) \triangleq \widehat{\nabla J_k^{\mathrm{G}}}(\boldsymbol{w}_{k,i-1}^{\mathrm{D}}, \boldsymbol{w}_{k,i-1}^{\mathrm{G}}) - \nabla J_k^{\mathrm{G}}(\boldsymbol{w}_{k,i-1}^{\mathrm{D}}, \boldsymbol{w}_{k,i-1}^{\mathrm{G}}). \tag{14b}$$

We assume these noises are unbiased and have bounded second-order moments:

$$\mathbb{E}\{\boldsymbol{s}_{k,i}^{\mathrm{D}}(\boldsymbol{w}_{k,i-1}^{\mathrm{D}}, \boldsymbol{w}_{k,i-1}^{\mathrm{G}})|\boldsymbol{\mathcal{F}}_{i-1}\} = 0 \tag{15a}$$
$$\mathbb{E}\{\boldsymbol{s}_{k,i}^{\mathrm{G}}(\boldsymbol{w}_{k,i-1}^{\mathrm{D}}, \boldsymbol{w}_{k,i-1}^{\mathrm{G}})|\boldsymbol{\mathcal{F}}_{i-1}\} = 0 \tag{15b}$$
$$\mathbb{E}\{\|\boldsymbol{s}_{k,i}^{\mathrm{D}}(\boldsymbol{w}_{k,i-1}^{\mathrm{D}}, \boldsymbol{w}_{k,i-1}^{\mathrm{G}})\|^2|\boldsymbol{\mathcal{F}}_{i-1}\} \le \sigma_{\mathrm{D}}^2 \tag{15c}$$
$$\mathbb{E}\{\|\boldsymbol{s}_{k,i}^{\mathrm{G}}(\boldsymbol{w}_{k,i-1}^{\mathrm{D}}, \boldsymbol{w}_{k,i-1}^{\mathrm{G}})\|^2|\boldsymbol{\mathcal{F}}_{i-1}\} \le \sigma_{\mathrm{G}}^2 \tag{15d}$$

where $\boldsymbol{\mathcal{F}}_i$ denotes the filtration generated by the random processes $\boldsymbol{w}_{k,j}^{\mathrm{D}}$ and $\boldsymbol{w}_{k,j}^{\mathrm{G}}$ for all $k = 1, \ldots, K$ and $j \le i$. □

4.2. Single Training Round for Discriminators

Holding the generators fixed, we apply diffusion over the discriminators according to (10a)–(10b). To analyze the diffusion recursions, we introduce the discriminator network centroid, which averages the discriminator parameter vectors from across the network at a given time i:

$$\boldsymbol{w}_{\mathsf{c},i}^{\mathrm{D}} \triangleq \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{w}_{k,i}^{\mathrm{D}}. \tag{16}$$

Theorem 1 (Network disagreement). *Under Assumptions 1–4, the network disagreement between the discriminator centroid and local discriminators is bounded after sufficient iterations:*

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\,\|\boldsymbol{w}_{k,i}^{\mathrm{D}} - \boldsymbol{w}_{\mathsf{c},i}^{\mathrm{D}}\|^2 \le \frac{\mu^2\lambda_2^2(B_{\mathrm{D}}^2 + \sigma_{\mathrm{D}}^2)}{(1-\lambda_2)^2} + O(\mu^3) \tag{17}$$

for $i \ge (3\log\mu/\log\lambda_2) + O(1)$.

Proof. Omitted due to space limitations but similar to arguments used in [8]. □

After sufficient iterations, we can replace the local models $\boldsymbol{w}_{k,i}^{\mathrm{D}}$, for $k = 1, \ldots, K$, by the discriminator network centroid $\boldsymbol{w}_{\mathsf{c},i}^{\mathrm{D}}$ since, due to Theorem 1, all the local agents would have approximately approached this same discriminator.

Theorem 2 (Stationary point). *Suppose that the global objective of the discriminators is bounded from below, i.e.,* $J^{\mathrm{D}}(w^{\mathrm{D}}, w^{\mathrm{G}}) \triangleq -J(w^{\mathrm{D}}, w^{\mathrm{G}}) \geq J'_{\mathrm{D}}$. *Then, the centroid* $\boldsymbol{w}_{\mathsf{c},i}^{\mathrm{D}}$ *will reach an* $O(\mu)$*-mean-square-stationary point in at most* $O(1/\mu^2)$ *iterations. Specifically for some time* $i^\star$*, we have*

$$\mathbb{E}\,\|\nabla J^{\mathrm{D}}(\boldsymbol{w}_{\mathsf{c},i^\star}^{\mathrm{D}}, w^{\mathrm{G}})\|^2 \leq 2\mu\frac{c_2}{c_1} \tag{18}$$

where

$$c_1 \triangleq \frac{1 - 2\mu L_{\mathrm{D}}}{2} = O(1), \quad c_2 \triangleq \frac{L_{\mathrm{D}}\sigma_{\mathrm{D}}^2}{2} + O(\mu) = O(1), \tag{19}$$

and

$$i^\star \leq \frac{J^{\mathrm{D}}(w_{\mathsf{c},0}^{\mathrm{D}}, w^{\mathrm{G}}) - J'_{\mathrm{D}}}{\mu^2 c_2}. \tag{20}$$

Proof. Omitted due to space limitations but similar to arguments used in [8]. □

4.3. Single Training Round for Generators

After the discriminators converge to $w_{\mathsf{c}}^{\mathrm{D}}$, we can consider substituting the discriminator centroid into (6). Taking (7) and (8b) into account, we obtain the following problem for the generators:

$$\min_{w^{\mathrm{G}}} \sum_{k=1}^{K} p_k J_k^{\mathrm{G}}(w_{\mathsf{c}}^{\mathrm{D}}, w^{\mathrm{G}}). \tag{21}$$

Since the assumptions that the generator objectives satisfy are similar to the assumptions for the discriminator objectives, we can again verify that all generators will approach a centroid when running enough iterations, in an analogous manner to Theorem 1. In order to analyze the training process in this situation, it is also reasonable to replace the local generator models $\boldsymbol{w}_{k,i}^{\mathrm{G}}$, for $k = 1, \ldots, K$, by $\boldsymbol{w}_{\mathsf{c},i}^{\mathrm{G}}$.

4.4. Convergence of the Generator Centroid

Considering that both discriminator and generator parameters have converged to their corresponding centroid vectors, we can express the global min-max problem (6) as:

$$\min_{w_{\mathsf{c}}^{\mathrm{G}}} \max_{w_{\mathsf{c}}^{\mathrm{D}}} J(w_{\mathsf{c}}^{\mathrm{D}}, w_{\mathsf{c}}^{\mathrm{G}}). \tag{22}$$

4.4.1. Homogeneous Datasets

If all agents sample training data independently and from the same sample space with equal distributions, i.e., $\mathcal{X}_k = \mathcal{X}$ and $p_{\boldsymbol{x}_k}(\cdot) = p_{\boldsymbol{x}}(\cdot)$ for all $k = 1, \ldots, K$, then the local data samples $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K\}$ can be regarded as independent and identically distributed (iid) random vectors. From (5), we obtain for the kth agent the objective function:

$$J_k(w_{\mathsf{c}}^{\mathrm{D}}, w_{\mathsf{c}}^{\mathrm{G}}) = \mathbb{E}_{\boldsymbol{z}_k \sim p_{\boldsymbol{z}}(z)}\, \mathbb{E}_{\boldsymbol{x}_k \sim p_{\boldsymbol{x}}(x)}\, Q(w_{\mathsf{c}}^{\mathrm{D}}, w_{\mathsf{c}}^{\mathrm{G}}; \boldsymbol{x}_k, \boldsymbol{z}_k) \tag{23}$$

which does not depend on k. Then, the global objective (7) reduces to:

$$J(w_{\mathsf{c}}^{\mathrm{D}}, w_{\mathsf{c}}^{\mathrm{G}}) = J_k(w_{\mathsf{c}}^{\mathrm{D}}, w_{\mathsf{c}}^{\mathrm{G}}) \sum_{k=1}^{K} p_k \overset{\text{(a)}}{=} J_k(w_{\mathsf{c}}^{\mathrm{D}}, w_{\mathsf{c}}^{\mathrm{G}}) \tag{24}$$

where (a) is due to Assumption 1. We note that this objective is equal to the objective (2) of a single GAN evaluated on the discriminator and generator centroids. Therefore, problems (22) and (1) are equivalent and the centroids are expected to converge to the solution for a single GAN trained with data from all local datasets.

4.4.2. Non-homogeneous Datasets

Next, we consider a more challenging situation that is closer to real applications. We suppose a worst-case scenario in which the agents have non-homogeneous and disjoint datasets. Specifically, we assume that the sample spaces $\mathcal{X}_k$, for $k = 1, \ldots, K$, form a finite partition of a sample space $\mathcal{X}$, i.e.,

$$\bigcup_{k=1}^{K} \mathcal{X}_k = \mathcal{X} \tag{25}$$

and $\mathcal{X}_k \cap \mathcal{X}_j = \varnothing$ for all $k, j = 1, \ldots, K$ with $k \neq j$. The training data distributions can be expressed as $p_{\boldsymbol{x}_k}(\cdot) = p_{\boldsymbol{x}|\boldsymbol{y}_k}(\cdot|y_k)$, where $\boldsymbol{y}_k$ represents the labels contained in the dataset of the kth local agent and their respective probabilities. From (5), the objective function for the kth agent results in:

$$J_k(w_{\mathsf{c}}^{\mathrm{D}}, w_{\mathsf{c}}^{\mathrm{G}}) = \mathbb{E}_{\boldsymbol{z}_k \sim p_{\boldsymbol{z}}(z)}\, \mathbb{E}_{\boldsymbol{x}_k \sim p_{\boldsymbol{x}|\boldsymbol{y}_k}(x|y_k)}\, Q(w_{\mathsf{c}}^{\mathrm{D}}, w_{\mathsf{c}}^{\mathrm{G}}; \boldsymbol{x}_k, \boldsymbol{z}_k). \tag{26}$$

For simplicity, we assume that each agent sees the same number of distinct classes of equiprobable data. Then, the prior probability for all $k = 1, \ldots, N$ is given by $p_{\boldsymbol{y}_k}(y_k) = 1/K$. From Assumption 1, we note that this probability is equal to the Perron eigenvector entry p_k. Considering this after substituting (26) into (7), it follows from the law of total expectation that the global objective evaluated at the centroids becomes:

$$J(w_{\mathsf{c}}^{\mathrm{D}}, w_{\mathsf{c}}^{\mathrm{G}}) = \mathbb{E}_{\boldsymbol{z}_k \sim p_{\boldsymbol{z}}(z)}\, \mathbb{E}_{\boldsymbol{x}_k \sim p_{\boldsymbol{x}}(x)}\, Q(w_{\mathsf{c}}^{\mathrm{D}}, w_{\mathsf{c}}^{\mathrm{G}}; \boldsymbol{x}_k, \boldsymbol{z}_k). \tag{27}$$

Once again, the min-max problem of the centroids takes the same form as the training problem of a single GAN with objective (2).

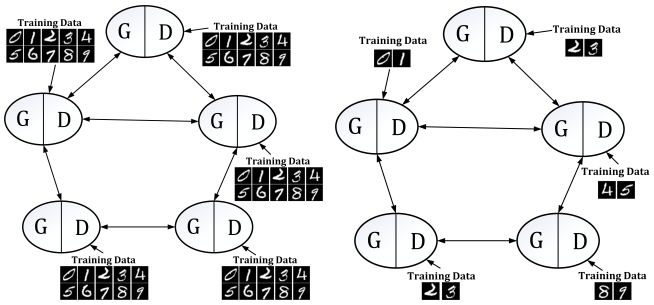
As a final remark, notice that the discriminator centroid is optimized to approximate the Jensen-Shannon divergence between the data distribution $p_{\boldsymbol{x}}(\cdot)$ and the generated distribution $p_{\mathrm{G}}(\cdot)$ resulting from the centroid parameters [1]. Due to the convexity of the Jensen-Shannon divergence [13] on $p_{\mathrm{G}}(\cdot)$, this distribution can converge to $p_{\boldsymbol{x}}(\cdot)$ through careful optimization.

5. SIMULATION RESULTS

The experiments are based on a network of GANs with $K = 5$ agents, using MNIST [6] and Fashion-MNIST [14] as training datasets. Each of these datasets has different instances of 10 classes of images.

In the case of homogeneous datasets, we divide the data randomly, and each agent has access to 12 000 training images comprising all classes of samples. We compare the performance of Algorithm 1 (Diffusion GAN) with non-cooperative (i.e., agents do not communicate) and centralized training (equivalent to single GAN with all the training data). In the case of non-homogeneous datasets, each agent has access to 12 000 training images composed of only two classes of samples.

The topology of the simulated network of GANs is shown in Fig. 2. For each agent, we use a deep convolutional GAN structure similar to [15]. We consider stochastic gradient descent as the optimizer with $\mu = 0.05$, $n_{\mathrm{d}} = 5$ and $n_{\mathrm{g}} = 1$. We use the Fréchet inception distance (FID) [16] to compare, in a quantitative manner, the similarity between the generated data distribution $p_{\mathrm{G}}(\cdot)$ and real data distribution $p_{\boldsymbol{x}}(\cdot)$. Lower values of the FID score should indicate greater similarity between distributions.

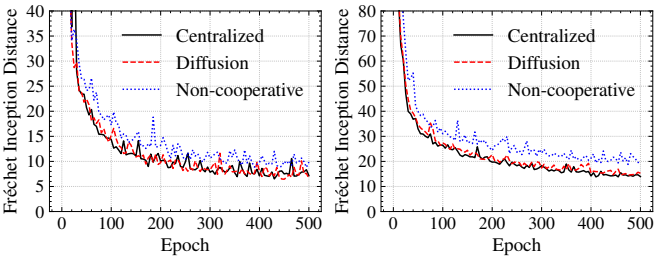


(a) Homogeneous datasets. (b) Non-homogeneous datasets.

Fig. 2. Topology and training data example (MNIST) for the simulated network of GANs.

5.1. Homogeneous Datasets

Figure 3 shows the FID score for different training approaches on MNIST and Fashion-MNIST datasets. We note that training through diffusion outperforms non-cooperative GANs. Also, the FID score for diffusion and centralized training are very close, which indicates that the Diffusion GAN algorithm allows agents with limited training data to achieve performance similar to a centralized agent with access to all training data. Samples generated through different approaches with homogeneous datasets are shown in Figs. 4 and 5. In these figures, each square corresponds to 100 fake images generated by a given agent.



(a) MNIST. (b) Fashion-MNIST.

Fig. 3. FID scores in the case of homogeneous datasets.

(a) Non-cooperative. (b) Centralized. (c) Diffusion.

Fig. 4. Generated fake images: homogeneous MNIST datasets.

(a) Non-cooperative. (b) Centralized. (c) Diffusion.

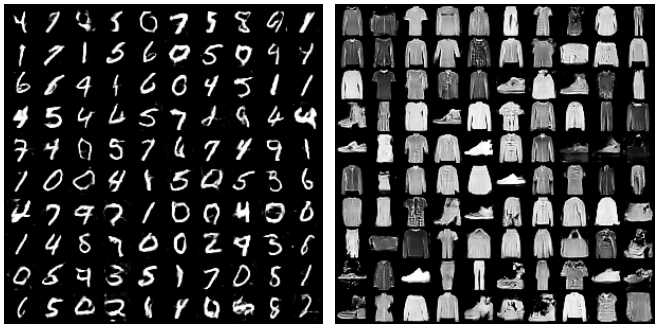
Fig. 5. Generated fake images: homogeneous Fashion-MNIST datasets.

5.2. Non-Homogeneous Datasets

In this scenario, each agent has access to only two classes of samples. Without cooperation, agents would only be able to generate digits of the same class as the training samples contained in their own datasets. However, by using the Diffusion GAN algorithm, agents are now able to generate all types of digits. For the MNIST dataset, fake samples generated by an agent with access to images of only 0s and 1s are shown in Fig. 6(a). For the Fashion-MNIST dataset, fake samples generated by an agent with access to images of only T-shirts and trousers are shown in Fig. 6(b).

6. CONCLUSION

In this work, we proposed a diffusion-based, fully decentralized algorithm for training a network of GANs. We explained that all local discriminators and generators will cluster around



(a) Agent sees only 0s and 1s.

(b) Agent sees only T-shirts and trousers.

Fig. 6. Generated fake images: same agent, non-homogeneous datasets.

their respective centroids after sufficient iterations. We also showed that the discriminator centroid will reach a stationary point, which will consist of an approximation for the Jensen-Shannon divergence between the distributions of training and generated data. Then, by writing the min-max problem over the network centroids, we are able to reduce it to the problem of training a single GAN with samples from all local datasets. Notably, this allows us to show that, although an agent might not have access to a particular type of data, it may still be able to generate it due to the sharing of information with its neighbors.

Through numerical simulations, we validated the proposed algorithm for different training settings and obtained encouraging results. For homogeneous datasets, the algorithm allowed local agents to match the performance of the centralized GAN. In the non-homogeneous case, the diffusion training was able to successfully allow agents with limited types of training data to generate all classes of fake samples.

7. REFERENCES

[1] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 1–9.

[2] C. Hardy, E. Le Merrer, and B. Sericola, "MD-GAN: Multi-discriminator generative adversarial networks for distributed datasets," in *Proc. IEEE Int. Parallel and Distributed Processing Symp. (IPDPS)*, 2019, pp. 866–877.

[3] S. Augenstein *et al.*, "Generative models for effective ML on private, decentralized datasets," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2020, pp. 1–26.

[4] M. Rasouli, T. Sun, and R. Rajagopal, "FedGAN: Federated generative adversarial networks for distributed data," *arXiv preprint*, 2020, Accessed on: Jan. 17, 2022. [Online]. Available: https://arxiv.org/abs/2006.07228v2.

[5] C. Hardy, E. Le Merrer, and B. Sericola, "Gossiping GANs: Position paper," in *Proc. Workshop on Distributed Infrastructures for Deep Learning (DIDL)*, 2018, pp. 25–28.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[7] A. H. Sayed, "Adaptation, Learning, and Optimization over Networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

[8] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments—part I: Agreement at a linear rate," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1242–1256, Jan. 2021.

[9] S. Vlaski and A. H. Sayed, "Competing adaptive networks," in *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, 2021, pp. 71–75.

[10] S. Vlaski and A. H. Sayed, "Diffusion learning in non-convex environments," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5262–5266.

[11] M. Kayaalp, S. Vlaski, and A. H. Sayed, "Dif-MAML: Decentralized multi-agent meta-learning," *IEEE Open Journal of Signal Processing*, vol. 3, pp. 71–93, Jan. 2022.

[12] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, Jun. 1953.

[13] J. Burbea and C. Rao, "On the convexity of some divergence measures based on entropy functions," *IEEE Transactions on Information Theory*, vol. 28, no. 3, pp. 489–495, May 1982.

[14] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint*, 2017, Accessed on: Feb. 17, 2022. [Online]. Available: https://arxiv.org/abs/1708.07747.

[15] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2016, pp. 1–16.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 1–12.