

Continual Learning with Fully Probabilistic Models

Anonymous CVPR 2021 submission

Paper ID ****

Abstract

We present an approach for continual learning (CL) that is based on fully probabilistic (or: generative) models of machine learning. In contrast to, e.g., GANs that are “generative” in the sense that they can generate samples, fully probabilistic models aim at modeling the data distribution directly. Consequently, they provide functionalities that are highly relevant for continual learning, such as density estimation (outlier detection) and sample generation. As a concrete realization of generative continual learning, we propose Gaussian Mixture Replay (GMR). GMR is a pseudo-rehearsal approach using a Gaussian Mixture Model (GMM) instance for both generator and classifier functionalities. Relying on the MNIST, FashionMNIST and Devanagari benchmarks, we first demonstrate unsupervised task boundary detection by GMM density estimation, which we also use to reject untypical generated samples. In addition, we show that GMR is capable of class-conditional sampling in the fashion of a cGAN. Lastly, we verify that GMR, despite its simple structure, achieves state-of-the-art performance on common class-incremental learning problems at very competitive time and memory complexity.

1. Introduction

Context This conceptual work is in the context of continual learning (CL). In its most general formulation, CL assumes that the distribution of training data changes over the training time of a machine learning model (concept drift). Often, this is restricted to a succession of sub-tasks having a stable data distribution, with abrupt changes in data distribution occurring at *sub-task boundaries* only. This is what we term a *sequential learning task* (SLT), see Sec. 2.

Although the CL paradigm is completely agnostic w.r.t. the type of learning that is involved. Most current work on CL is about supervised learning, often in the context of classification which usually requires discriminative machine learning methods. Since such methods are not well-suited for *outlier detection*, the *recognition of sub-task boundaries*

is problematic. The problem is usually circumvented by simply assuming that sub-task boundaries are known.

What renders CL different from conventional machine learning is the fact learning occur continuously over long times. This implies a number of constraints. First, access to data is limited, typically to samples from the current sub-task, for memory reasons. Of course, a small subset of samples from previous sub-tasks may be retained. More useful still is the *generation* of such samples. Second, training times for new sub-tasks should scale sub-linearly (ideally: $\mathcal{O}(1)$) with the total number of samples seen by the model. Otherwise, CL could not be scaled to learning tasks with an infinite number of sub-tasks.

Motivation The presented work is motivated by the fact that many functionalities evoked in the previous paragraphs are in fact typical of generative, unsupervised learning methods. Mixture models, for example, are very commonly used for outlier detection and sample generation and have a very benign forgetting behavior when faced with changes in data distribution. In this article, we aim at integrating mixture models into a hybrid approach for supervised CL, which we term Gaussian Mixture Replay (GMR), and to show the various benefits for CL on standard benchmarks.

1.1. Related Work on CL

The field of CL is expanding rapidly, see [1, 2, 3, 4] for reviews. Systematic comparisons between different approaches to avoid CF are performed in, e.g., [5, 6]. As discussed in [6], many recently proposed methods demand specific experimental setups, which deviate significantly from application scenarios. For example, some methods require access to samples from *all* sub-tasks for tuning hyper-parameters, whereas others need access to all samples from past tasks. Many proposed methods have a time and/or memory complexity that scales at least linearly with the number of sub-tasks and thus may fail if this number is large. Among the proposed remedies to CF, three major directions may be distinguished according to [4]: parameter isolation, regularization and rehearsal.

Parameter Isolation Isolation methods aim at determining (or creating) a group of DNN parameters that are mainly “responsible” for a certain sub-task. CL is then avoided by *protecting* these parameters when training on successive sub-tasks. Representative works are [7, 8, 9, 10, 11, 12].

Regularization Regularization methods mostly propose to modify the loss function, including additional terms that protect knowledge acquired in previous sub-tasks. Actual approaches are very diverse: SSL [13] focuses on enhancing sparsity of neural activities, whereas approaches such as LwF [14] rely on knowledge distillation mechanisms. A method that has attracted significant attention is Elastic Weight Consolidation (EWC) [15]. EWC inhibits changes to weights that are important to previous sub-tasks, measuring this importance based on the Fisher information matrix (FIM). Synaptic intelligence [16] is pursuing a similar goal. Even an online variant of EWC is published [17]. Incremental Moment Matching (IMM) [18] makes use of the FIM to merge the parameters obtained for different sub-tasks. The Matrix of Squares (MasQ) method [19] is similar in spirit to EWC, but relies on the calculus of derivatives to assess the importance of parameters for a sub-task. It is more simple w.r.t. its concepts and much more memory-efficient.

Rehearsal Rehearsal methods come mainly in two forms: rehearsal and pseudo-rehearsal. *Rehearsal methods* store a subset of samples from past sub-tasks preventing CF, either by putting constraints on current sub-task training or by adding retained samples to the current sub-task training set. Typical representatives of rehearsal methods are iCaRL [20], (A-)GEM [21, 22], GBSS [23] and TEM [24]. *Pseudo-rehearsal* or *generative replay* methods, in contrast, do not store samples but generate them using a dedicated *generator* that is trained along with the learner, see Fig. 1. Typical models used as generators are generative adversarial networks (GANs), variational autoencoders (VAEs) and their variants, see [25] and [26]. The GMR model that we propose here belongs to this type as well.

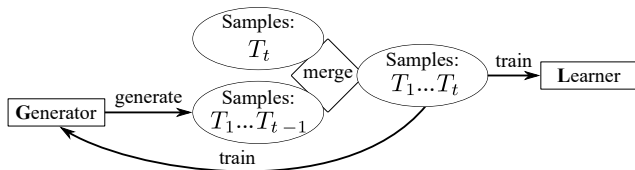


Figure 1. The replay approach to continual learning: a Learner, e.g., a DNN, is trained on several sub-tasks sequentially. To avoid forgetting, a Generator is trained to generate samples from past sub-tasks. For training L, G generates samples from past sub-tasks, which are merged with current sub-task samples.

Training and Evaluation Paradigms for CL In the context of CL, a wide range of training and evaluation paradigms are proposed, see [27, 28, 29, 30, 5, 31, 32].

1.2. Gaussian Mixture Replay

Gaussian Mixture Replay (GMR) is a CL approach based on pseudo-rehearsal, with a Gaussian Mixture Model (GMM) serving as generator. Mixture models describe the probability density of data \mathbf{X} as a weighted superposition of parametric distributions $p(\cdot; \beta_j)$:

$$p(\mathbf{X}) = \prod_i p(\mathbf{x}_i) = \prod_i \sum_{j=1}^K \pi_j p(\mathbf{x}_i; \beta_j).$$

For GMR, we use Gaussian parametric distributions defined by centroids μ_j and covariance matrices Σ_j : $p(\mathbf{x}; \beta_j) \equiv \mathcal{N}(\mathbf{x}; \mu_j, \Sigma_j) \equiv \mathcal{N}_j(\mathbf{x})$. For simplicity, we describe GMR including a single GMM “layer” only, but a generalization to deep convolutional GMMs is straightforward, see [33]. Data vectors entering the trained GMM are transformed into the GMM’s a posteriori distribution (or *responsibility*) γ as $\gamma_i(\mathbf{x}) = \frac{\exp(\mathcal{N}_i(\mathbf{x}))}{\sum_z \exp(\mathcal{N}_z(\mathbf{x}))}$. Responsibilities are bounded in the interval $[0, 1]$ and normalized to have unit sum: $\sum_z \gamma_z(\mathbf{x}) = 1$. This makes them well suited as inputs for a linear classifier which transforms responsibilities into class membership probabilities. The data flow through a GMR instance is shown in Fig. 2.



Figure 2. Principal structure of the GMR model, composed of a GMM modeling the distribution of training samples (left). A linear classifier operating on the posterior probabilities (also termed *responsibilities*) produced by the GMM. The coupled GMM/classifier implements the Learner, whereas the GMM implements the Generator from Fig. 1. The GMM sampling process is informed by feedback from the classifier.

A major point about GMR is that the generator and learner are not separate entities. The GMM performs generative tasks (sampling and outlier detection), and, at the same time, provides the learner (i.e., the linear classifier) with a high-level data representation.

1.3. Differences to Related Work

Gaussian Mixture Replay (GMR) aims to improve the following aspects of recent work on continual learning:

Outlier Detection Discriminative machine learning models such as DNNs or CNNs, which are at the heart of most current CL approaches, allow *supervised* outlier detection only. Here, outliers are simply samples with high loss, and concept drift is assumed to occur if the loss changes significantly. However, loss computation requires targets for supervised learning, which are not always available. More problematic still, in such an approach it is impossible to determine whether concept drift is occurring in the data, or it is

just the targets that are drifting. And lastly, outlier detection for individual samples cannot be trusted: high-loss inliers cannot be distinguished from outliers, unless classification is near-perfect.

Sample Generation In pseudo-rehearsal methods such as [25], GANs (cGANs, WGANs) are employed as generators. While these can generate impressive samples, it is not clear whether these samples represent the full probability distribution that they are supposed to sample from. In fact, there is the problem of *mode collapse*, where GANs focus on a small part of the data distribution only. Mode collapse is difficult to detect automatically since GANs do not possess a (differentiable) loss function that expresses the models’ current ability to sample.

Resource Efficiency Pseudo-rehearsal approaches contain generator and learner components. For GANs, the generator is further composed of a generator and a discriminator. All of these components are usually implemented as DNNs or CNNs requiring a considerable amount of resources, in particular memory.

Scalability Since the generators are implemented as CNNs or DNNs, they are very sensitive to class balance. For each new sub-task, the generator must therefore produce the precise number of samples that ensures that classes from previous and current sub-tasks are balanced. As a consequence, the number of generated samples grows linearly with the number of sub-tasks, which may be prohibitive for problems with many sub-tasks.

1.4. Novel Contributions

GMR offers several novel contributions to the field of CL:

- Unsupervised outlier detection: consistency ensured by relying on a fully probabilistic GMM
- Resource-efficiency: pseudo-rehearsal integrating learner and generator in a single structure
- Robustness: model collapse excluded by theoretical guarantees for GMM training
- Competitiveness: state-of-the-art CL performance on standard problems

To validate our approach, we perform a comparison to Elastic Weight Consolidation (EWC) model what is assumed to be a “standard model” for CL in many recent publications. A simple generative-replay approach as presented in [25] is used as baseline. Furthermore, we provide a public TensorFlow 2 implementation¹.

2. Data

Image Benchmarks In order to measure the impact of forgetting during continual learning, three public image classi-

fication benchmarks are used to construct sequential learning tasks, see Tab. 2. All datasets consist of grayscale images with dimensions of $28 \times 28 \times 1$ or $32 \times 32 \times 3$, whose entries are normalized to the $[0, 1]$ interval. We merge the provided train and test sets for each benchmark, and split the merged data in a proportion of 90% to 10% into training and test data. All datasets exhibit an almost equal distribution of samples within classes.

MNIST contains images of handwritten digits (0-9) with a resolution of 28×28 pixels. It is probably the most commonly used benchmark for classification problems. FashionMNIST contains pictures of different types of clothes. This data set is supposed to be harder to classify compared to MNIST (same resolution) and thus leads to lower accuracies. Similar to MNIST, the Devanagari data set contains written Devanagari letters. It is available in a resolution of 32×32 pixels per image. Since there are more classes included than needed, we randomly select 10 classes.

Sequential Learning Tasks Sequential Learning Tasks (SLTs) simulate a continuous learning scenario by dividing data sets given in Sec. 2. The resulting sub data sets are enumerated and contain only samples of non-overlapping classes. For example, a D_{5-5} task consists of two sub-data sets consisting of 5 classes each. Each sub-task is identified by its order, e.g., T_1, T_2, \dots, T_x . Baseline experiments (D_{10}) contain all available classes to investigate the effect of incremental task-by-task training.

With SLTs basic experiments can be carried out to determine the effect of forgetting under the above conditions. To measure the impact of the number of classes contained in a task, different combinations and subdivisions are evaluated. Tab. 1 displays all evaluated SLTs and their definition of sub-tasks can be taken.

3. Gaussian Mixture Replay in Detail

As stated in Sec. 1.2, GMR is comprised of a generator realized by a GMM, and a learner realized by a linear classifier. Both can indeed be replaced by more complex, “deeper” methods, but we limit us here to simplest case.































The generator consists of K Gaussian mixture components, each maintaining a separate μ_k centroid and covari-

Table 1. Definition of Sequential Learning Tasks (SLTs) and the class divisions of their sub-tasks.

SLT	Sub-Tasks
D_{10}	$T_1(0, 1, 2, 3, 4, 5, 6, 7, 8, 9)$
D_{9-1a}	$T_1(0, 1, 2, 3, 4, 5, 6, 7, 8) \quad T_2(9)$
D_{9-1b}	$T_1(0, 1, 2, 4, 5, 6, 7, 8, 9) \quad T_2(3)$
D_{5-5a}	$T_1(0, 1, 2, 3, 4) \quad T_2(5, 6, 7, 8, 9)$
D_{5-5b}	$T_1(0, 1, 2, 6, 7) \quad T_2(3, 4, 5, 8, 9)$
$D_{2-2-2-2-2a}$	$T_1(0, 1) \quad T_2(2, 3) \quad T_3(4, 5) \quad T_4(6, 7) \quad T_5(8, 9)$
$D_{2-2-2-2-2b}$	$T_1(1, 7) \quad T_2(0, 2) \quad T_3(6, 8) \quad T_4(4, 5) \quad T_5(3, 9)$

¹<https://github.com/cvpr2021-anonymous/CLwFPM>

Table 2. Detailed information to the used data sets (including examples of each class).

Dataset	Ref.	Resolution	Number of Training Samples	Number of Test Samples	Random Examples (from classes)									
					0	1	2	3	4	5	6	7	8	9
MNIST	[34]	28×28	50 000	10 000										
FashionMNIST	[35]	28×28	60 000	10 000										
Devanagari	[36]	32×32	18 000	2 000										

ance matrix Σ_j . Covariance matrices are always taken to be diagonal (a justification for this is given in the discussion).

As the basic data flow in GMR has been outlined in Sec. 1.2, we will describe the procedure for training, sampling and outlier detection, as well as outline the principal GMR hyper-parameters.

3.1. Outlier Detection

Outlier detection is performed by the generator according to standard GMM procedures. Essentially, it is based on the value of the loss function for a given sample, and anything too far below the “normal” loss value is considered an outlier. To achieve this, we compute of the mean and the variance of the GMM loss during training:

$$\begin{aligned}\hat{\mu}(\mathcal{L}) &= \mathbb{E}_i \mathcal{L}(x_i) \\ \hat{\Sigma}^2(\mathcal{L}) &= \mathbb{E}_i (\mathcal{L}(x_i) - \hat{\mu}(\mathcal{L}))^2.\end{aligned}$$

A sample x is considered an outlier if, and only if, $\mathcal{L}(x) < \hat{\mu}(\mathcal{L}) - c\sqrt{\hat{\Sigma}^2(\mathcal{L})}$, where c is a free parameter. Smaller values of c will detect more outliers and vice versa.

3.2. Unconditional Sampling

Sampling is again conducted according to GMM standard procedures. It consists of first drawing a GMM component from a multinomial distribution parameterized by the GMM weights π : $k \sim \mathcal{M}(\pi)$. Then, a random vector $z \in \mathbb{R}^d$, $z \sim \mathcal{N}(0, \mathbf{I})$ of the same dimensions d as the data is drawn. The vector is transformed into a sample x as $x = \Sigma_k z + \mu_k$, which ensures that $x \sim \mathcal{N}_k(\cdot; \mu_k, \Sigma_k)$. In Sec. 7.3, we will prove that the GMM log-likelihood on training data provides a lower bound for the log-likelihood of samples generated in this way. Thus, if we have higher training log-likelihoods, we can expect to generate better samples. To show this, we shall prove the following **Proposition**: The training loss of a GMM is a lower bound on the expected loss of generated samples.

Proof: To prove the proposition, it is sufficient to prove the proposition for the case of a single Gaussian component density, which shall be denoted $\mathcal{N}(x; \mu, \Sigma) \equiv \mathcal{N}(x)$. After decomposing the covariance matrix Σ as $\Sigma = \mathbf{A}\mathbf{A}^\top$, a set of samples $G \supset g$ can be obtained (see Sec. 3.2). This is achieved by transforming a random normal variable $z \sim \mathcal{N}(0, \mathbf{I})$ as $g = \mathbf{A}z + \mu$. The loss on the generated

samples is expressed as

$$\begin{aligned}\mathcal{L}(g) &= \ln \mathcal{N}(g) = \ln \mathcal{N}(\mathbf{A}z + \mu) \\ &\sim f(\Sigma) - \frac{1}{2}(\mathbf{A}z)^\top \Sigma^{-1}(\mathbf{A}z) = f(\Sigma) - \frac{1}{2}\|z\|^2 \\ \mathcal{L}(G) &= \sum_i \mathcal{L}(g) = Nf(\Sigma) - \frac{dN}{2}.\end{aligned}\quad (1)$$

If the training samples follow a Gaussian distribution, their mean and variance coincide with the parameters μ, Σ of the Gaussian component density. Thus, their loss is identical to Eq. (1) by the same reasoning. If training samples deviate from Gaussianity, as may be expected in practice, their loss will be lower. This is trivial to show by expanding their distribution around a Gaussian one into an Edgeworth series (see [37]), and plugging this expansion into Eq. (1). Thus, we know that the loss that is actually obtained on test data represents a lower bound for the loss of generated samples.

3.3. Class-Conditional Sampling

This form of sampling has the goal of generating samples belonging to a given class c . To provide the GMM with this information, we fix a certain output vector o of the linear classifier and try to infer what inputs i would produce it:

$$o = s(\mathbf{W}i + b) \Rightarrow i \approx \mathbf{W}^\top (s^{-1}(o) + k - b).$$

Since the softmax function is shift-invariant, the inverse is defined only up to a constant k which we set to 0. To first approximation, we assume that the weight matrix \mathbf{W} of the linear classifier has orthogonal columns. Entries of o must be bounded in the $[0, 1]$ interval, have a unit sum, and express a confident decision for a given class C . We choose $o_C = 0.95$ and normalize accordingly to obtain a *control signal* i for the GMM. This control signal represents the expected posterior probabilities of the GMM for a given class C . It is therefore consistent to use it for unconditional sampling (previous paragraph) instead of the GMM weights π .

3.4. Replay

Prior to training generator and learner at sub-task $T \geq 1$, samples from previous sub-tasks $1 \leq t < T$ must be produced by the generator. If we let $\nu(t)$ denote the number of data samples for any sub-task t , and $\xi(t)$ the number of samples to generate for sub-task t , then two strategies may

be discerned for choosing $\xi(t)$. The *proportional* strategy which chooses $\xi(t) = \sum_{t'}^{t-1} \nu(t')$ and the *constant* strategy with $\xi(t) = \kappa \nu(t)$.

Training Once samples have been generated, generator and learner are trained concurrently, each with its own loss function. For the GMM, we use plain stochastic gradient descent (SGD) to maximize the *log-likelihood* of the training data under the model, expressed in the notation of Sec. 1.2 as:

$$\mathcal{L}(X) = \ln p(\mathbf{X}) = \sum_i \log \sum_k \pi_j \mathcal{N}_j(\mathbf{x}),$$

using the efficient training procedure for high-dimensional streaming data described in [38]. The linear classifier receives the GMM responsibilities γ as input and is trained by minimizing the usual cross-entropy loss

$$\mathbf{y}_i = s(\mathbf{W}\gamma_i(\mathbf{x}) + \mathbf{b})$$

$$\mathcal{L}^{CE} = \frac{1}{N} \sum_i \log y_{ij} t_{ij}$$

by SGD, with $s(\cdot)$ denoting the softmax function. SGD learning rates for GMM and linear classifier are denoted by ϵ^G and ϵ^C .

Hyper-Parameters The principal hyper-parameters of GMR are, first of all, the number K of GMM components, and the GMM learning rate ϵ^{GMM} . All GMM hyper-parameters are selected according to [38]. In particular, the crucial parameter K follows a “the more the better” logic so it is easy to select. For the linear classifier, the learning rate ϵ^C plays a role as well. Since inputs to the linear classifier are normalized and bounded in the $[0, 1]$ interval, the optimal learning rate is rather task-independent can be selected as a function of the GMM parameter K .

4. Elastic Weight Consolidation

The approach from [15] is a typical regularization-based model for DNNs, see Sec. 1.1. EWC stores DNN parameters $\theta_i^{T_t}$ after training on sub-task T_t . In addition, EWC computes the “importance” of each parameter after training on sub-task T_t . This is done by approximating the diagonal \vec{F}^{T_t} of the Fisher Information Matrix (see [19] for a discussion of this approximation). The EWC loss function contains additional terms, see Eq. (2) besides the cross-entropy loss computed on the current sub-task T_c . These additional terms punish deviations from “important” DNN parameter values obtained after training on past sub-tasks:

$$\mathcal{L}^{EWC} = \mathcal{L}_{T_c}(\theta) + \frac{\lambda}{2} \sum_{i=1}^{c-1} \sum_i F_i^{T_t} (\theta_i - \theta_i^{T_t})^2 \quad (2)$$

EWC is optimized using the Adam optimizer. EWC hyper-parameters are the SGD step size ϵ^{EWC} , the regularization

constant λ and of course the number and size of layers in the DNN. In [15], it is proposed to set $\lambda = 1/\epsilon^{EWC}$, thus eliminating one hyper-parameter.

5. Generative Replay

We implement generative replay (GR) as described in [25] with a GAN-based generator. The precise configurations of generator and learner is given in App. A. Batch sizes are \mathcal{B} for the first sub-task and $2\mathcal{B}$ for sub-tasks $t > 1$. Important hyper-parameters are the SGD step size ϵ^G , and the number of epochs for training. At each sub-task, the generator produces as many samples as contained in all previous sub-tasks to maintain balance. Alternatively, a fixed number of generated samples is possible as well.

6. Experiments

For validating the goals as outlined in Sec. 1, we conduct the following experiments on sequential learning tasks (SLTs) constructed as described in Sec. 2. In Sec. 6.2, we demonstrate unsupervised outlier detection to identify sub-task boundaries without reference to class labels. A demonstration of sampling quality as measured by the GMM-loglikelihood is given in Sec. 6.3. As a by-product of the GMR architecture, we present, results on class-conditional sampling on all three datasets in Sec. 6.4. Sec. 6.5 shows that GMR achieves state-of-the-art classification performance on the SLTs when compared to generative replay and EWC.

6.1. Hyper-Parameters

GMR In the terms of Sec. 3, we chose $K=100$, $\epsilon^G=0.01$, $\mathcal{B}=100$, $\epsilon^C=0.01$. For the constant replay strategy (see Sec. 3.4), a proportionality constant of $S=2$ is used. Training epochs are empirically set to 20 for each task. The other hyper-parameters are set to default values as defined in [38].

GR In terms of Sec. 5 and Fig. 1, generators are always trained for 50 epochs (\mathcal{E}) and solvers for 25 epochs. The Adam optimizer is used for effecting gradient descent, using a step size $\epsilon^G=0.001$ for solver and generators. Samples are generated such as to maintain balance between previous and current classes.

EWC For each SLT, we perform a grid search for the parameter ϵ . We vary the learn rate for EWC ϵ^{EWC} as $\epsilon^{EWC} \in \{0.001, 0.0001, 0.00001, 0.000001, 0.0000001\}$. Depending on this λ is always set to $\frac{1}{\epsilon^{EWC}}$. We fix the model architecture to a three-layer DNN, each of size 800. Training epochs \mathcal{E} are empirically set to 10 for each training task. The best hyper-parameters and experiments are selected based on the highest average accuracy (over 10 repetitions) on all classes measured after the last sub-task.

6.2. Task Boundary Detection

We train GMR on the SLT $D_{2-2-2-2-2a}$, while updating the sliding average and variance for the log-likelihood as indicated in Sec. 3.1. For each sample x in a mini-batch \mathcal{B} , we test whether they are inliers as discussed in Sec. 3.1, using a value of $c = 1$. We then compute the empirical probability of inliers in the mini-batch. Each time this probability drops by more than 20%, we assume a task boundary has occurred. The results are shown in Fig. 3.

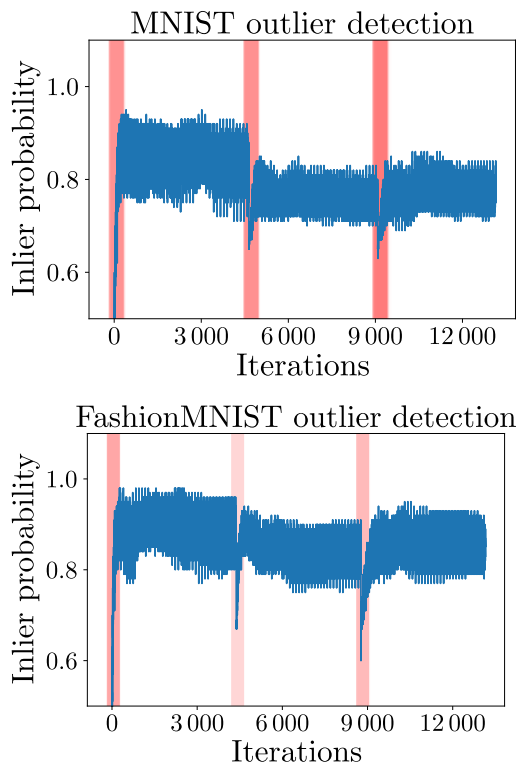


Figure 3. Detection of task boundaries on the first three sub-tasks of $D_{2-2-2-2-2a}$. Areas highlighted in red signal automatically detected task boundaries.

6.3. Sampling

In this experiment, we verify that the GMM loss of generated samples (sampling loss) is always higher than the GMM training loss. A proof for this was given in Sec. 3.2: here, we give an empirical validation. This experiment is independent of continual learning, which is why we use the baseline SLT D_{10} for all three datasets. Fig. 4 shows results for all three datasets, and we observe that the sampling loss is indeed higher than the asymptotic training loss, often by quite a margin.

6.4. Class-Conditional Sampling with GMR

For this experiment, we train a GMR instance on SLT D_{10} for each dataset, i.e., on all classes at once. Subsequently,

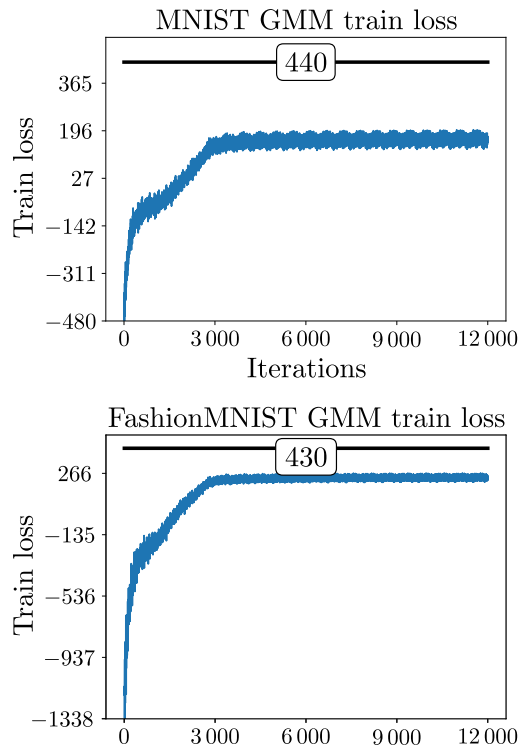


Figure 4. Training and sampling loss for SLT D_{10} . The sampling loss is superimposed on each graph as a black horizontal line. Its value is given in the box.

we use each of the three trained models to conditionally generate 50 samples: 25 from classes $\mathcal{C} = \{1, 2\}$, and 25 samples from from classes $\mathcal{C} = \{5, 7\}$. For each generated sample, the class c is drawn from \mathcal{C} with equal probability. Control signals to the GMM for generating a sample from class c are obtained and applied according to Sec. 3.3. The results can be viewed in Fig. 5. We observe that samples are very reliably selected from the given set \mathcal{C} . In some cases, errors occur for samples that are visually very similar to elements of \mathcal{C} : this reflects simply the fact that the classification accuracy is not perfect. For perfect classification, we expect no such sampling inaccuracies.

Additionally, we perform class-conditional sampling in the same way as just described, but using a deep convolutional GMM (DCGMM) as described in [33]. Model details are given in App. B. The generated samples for MNIST are shown in Fig. 6.

6.5. Comparison of GR, GMR and EWC

We train EWC, GMR and GR on all SLTs listed in Sec. 2, according to the hyper-parameter settings described earlier in this section, see Sec. 6.1. Classification accuracy is read off after completing training on the last sub-task. Baseline accuracy on a non-continual learning task (D_{10}) is recorded for all datasets and methods. For GR and GMR, we use the

Table 3. Results of the conducted GMR, EWC and GR experiments. The accuracy in % is stated as baseline for all experiments based on the available classes for each dataset. For each best SLT experiment (defined in Tab. 1) the difference to the baseline is given. Therefore, the maximum measured accuracy value is used, averaged over 10 experiment repetitions. To measure the accuracy, the joint test dataset consisting of all tasks (D_{10}) is used.

model dataset	GMR						EWC						GR					
	MNIST		FashionMNIST		Devanagari		MNIST		FashionMNIST		Devanagari		MNIST		FashionMNIST		Devanagari	
	acc. %	std	acc. %	std	acc. %	std	acc. %	std	acc. %	std	acc. %	std	acc. %	std	acc. %	std	acc. %	std
D_{10} baseline	87.4	0.59	73.9	0.26	74.1	0.73	97.57	0.26	87.55	0.38	95.58	0.56	99.3	0	99.3	0	99.1	0
	diff.	std	diff.	std	diff.	std	diff.	std	diff.	std	diff.	std	diff.	std	diff.	std	diff.	std
D_{9-1a}	-1.3	0.59	-2.7	0.26	-3.2	0.73	-41.8	0.26	-9.6	0.38	-56.6	0.56	-15.1	0.7	-25.6	0.5	-13.5	0.54
D_{9-1b}	-3.5	2.19	-1.5	0.87	-1.4	0.88	-50.7	7.77	-20.1	2.52	-29.7	13.34	-21.8	0.9	-16.5	1.1	-10.9	0.41
D_{5-5a}	-0.6	1.53	-1.2	1.53	-6.8	1.38	-35.3	6.65	-32.7	4.22	-46.0	15.38	-10.0	1.4	-19.8	3.2	-6.7	0.3
D_{5-5b}	-1.3	1.92	-1.9	0.49	-4.7	1.59	-35.0	1.83	-36.0	2.72	-47.1	0.11	-11.9	1.0	-17.7	4.0	-7.6	0.37
$D_{2-2-2-2-2a}$	-9.5	3.83	-8.5	0.91	-22.5	2.71	-72.2	7.43	-55.6	4.05	-72.1	2.75	-41.4	3.8	-25.0	5.9	-40.0	3.3
$D_{2-2-2-2-2b}$	-10.4	5.28	-5.7	2.37	-14.7	2.94	-72.6	3.22	-57.3	4.99	-73.2	2.31	-34.8	4.1	-29.4	7.3	-34.7	7.6

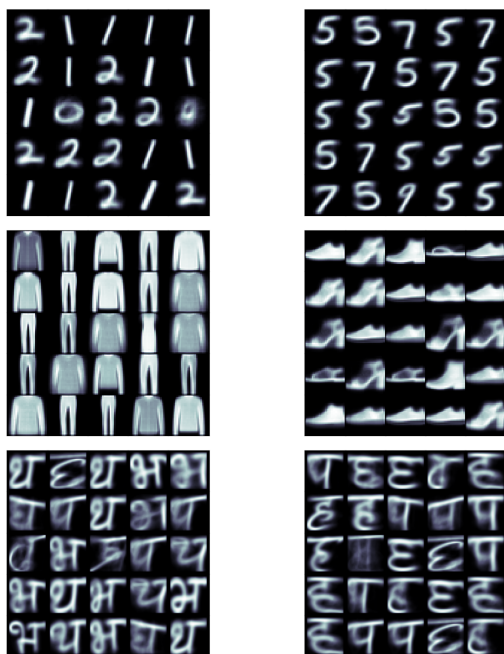


Figure 5. Conditional sampling results for GMR models trained on MNIST (top), FashionMNIST (middle) and Devanagari (bottom). In each row, 25 samples for classes 1,2 (left) and 25 samples for classes 5,7 (right) are generated.



Figure 6. Conditional sampling results for GMR models trained on MNIST using a deep convolutional GMM (DCGMM). To be read as Fig. 5.

proportional sample generation strategy, see Sec. 3.4. We present the results as deviations from baseline performance

in Tab. 3. The comparison is not entirely fair since GMR has a much lower baseline performance. On the contrary, we observe that the drop in classification accuracy due to CL is generally much smaller. We take the view that is really this drop that characterizes continual learning performance.

7. Principal Conclusions and Discussion

7.1. State-Of-The-Art GMR Performance

From the experiments of Sec. 6, we can conclude that GMR can equalize the performance of GR, and that both GMR and GR outperform EWC by a large margin. Here, we are talking about *continual learning* performance as defined in Sec. 6.5. GMR performance on the non-continual baseline D_{10} is markedly inferior to that of a standard DNN. This makes it even more remarkable that continual learning performance is similar to GR, which after all included a fully-fledged CNN classifier.

7.2. Memory Requirements

GMR has a low memory footprint because the generator (the GMM) is re-used for classification, see Fig. 2. Since the GMM itself is “flat”, its memory requirements are modest. For an input dimensionality of $d=1000$, with $K=100$ GMM components and 10 classes, the total memory required to store the complete GMR model is $2Kd + 10K + 10 = 201\,010$. The corresponding GR model consists of a three-layer DNN (the learner) and two CNNs for the generator and a discriminator. It contains 3 770 204 parameters, which is more than two orders of magnitude larger than the corresponding GMR model. This is again remarkable since continual learning performance is quite comparable.

7.3. Quality of Generated Samples

In contrast to models like, e.g., GANs, GMMs provide strong guarantees concerning the quality of generated samples via their loss function, see Sec. 3.2. A direct implication is that sample generation capacity can be monitored *at*

training time by monitoring the loss. In particular, if the loss should decrease significantly during training, this would be a strong sign for mode collapse. This is virtually excluded due to SGD that aims to maximize the loss, and such behavior was never observed in all the experiments conducted in Sec. 6.3.

7.4. Simple Conditional Sampling

As shown in Sec. 6.4, conditional sampling is a reliable way to obtain samples from certain given classes only. The simplicity of the process is an appealing feature of GMR, realized through the GMMs ability for (unconditional) sampling.

7.5. Respecting Real-World Constraints

GMR is attractive for real-world applications because it fully respects several important constraints (see [6] for a more comprehensive discussion of real-world constraints).

No Looking Back GMR does not require access to data from past sub-tasks to determine when to stop re-training, see Sec. 1. This property is shared with GR but not with EWC, whose performance drops after a certain time, see Sec. 6.5. To determine the optimal point for stopping training, EWC requires access to data from past sub-tasks, which is in direct contradiction to the continual learning paradigm, see Sec. 1.

No Looking Ahead The hyper-parameters for EWC, notably the balancing parameter λ and the DNN parameters, need to be determined by grid-search, since performance depends upon these parameters in a complex way. This requires repeating the whole experiment many times with different parameters, and thus determining hyper-parameters for given sub-task based on sub-tasks that come later. That violates the CL paradigm, which states that only one sub-task at a time can be accessed. See [6] for a longer discussion on this point. In contrast, GMR has only one really free parameter, the number of GMM components K , which follows a “the more the better”. Therefore, it is possible to determine a good value for K on sub-task T_1 only, and thus to respect the CL paradigm. A large number of training epochs does not affect learning adversely, and can thus be liberally selected on T_1 , just as the learning rate.

7.6. Model Limitations

As far as *continual* learning performance is concerned, the presented GMR method has state-of-the-art performance on SLTs derived from simple benchmarks such as MNIST or FashionMNIST. It is however strongly inferior w.r.t. *non-continual* (baseline) performance as shown in Sec. 6.5. Neither can it be expected to perform well on more difficult SLTs constructed, e.g., from the SVHN benchmark. Mainly, such a complex benchmark would require an ex-

remely high number K of GMM components for high-quality sampling. In addition, the representation provided by an “flat” GMM may not be expressive enough to allow accurate classification. A solution to both problems may well lie in using deep GMM variants such as presented in [39] or [33].

8. Outlook and Next Steps

Applying GMR to more challenging problems requires principally to improve the GMM’s sample generation capacity without excessive resource requirements. We will investigate two main directions:

Using a Deep Convolutional Generator Just as DNNs and CNNs can represent more complex functions than single-layer perceptrons, deep convolutional GMMs can model more complex distributions. We plan to investigate the models proposed in [39] or [33] for replacing the current “flat” GMM.

Different GMR Design Choices A major design choice in GMR is to restrict GMMs to diagonal covariance matrices, see Sec. 3. Full covariance matrices are out of the question due to memory reasons: for $K = 100$ and data dimensionality $d \approx 1000$. This would involve 10^8 parameters for storage alone, not talking about memory requirements on a GPU due to parallel processing. A compromise might be the use of a MFA (mixture of factor analyzers) instead of a GMM model. This may, at reasonable memory overhead, significantly enhance the GMM’s sample generation capacity as demonstrated in, e.g., [40].

References

- [1] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, may 2019. 1
- [2] Tyler L. Hayes, Ronald Kemker, Nathan D. Cahill, and Christopher Kanan. New Metrics and Experimental Paradigms for Continual Learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2144–2147, jun 2018. 1
- [3] Andrea Soltoggio, Kenneth O. Stanley, and Sebastian Risi. Born to learn: The inspiration, progress, and future of evolved plastic artificial neural networks. *Neural Networks*, 108:48–67, 2018. 1
- [4] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1

- [5] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L. Hayes, and Christopher Kanan. Measuring Catastrophic Forgetting in Neural Networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 3390–3398. AAAI Press, 2018. 1, 2
- [6] Benedikt Pfülb and Alexander Gepperth. A comprehensive, application-oriented study of catastrophic forgetting in DNNs. *International Conference on Learning Representations (ICLR)*, 2019. 1, 8
- [7] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A. Rusu, Alexander Pritzel, and Daan Wierstra. PathNet: Evolution Channels Gradient Descent in Super Neural Networks. *CoRR*, abs/1701.08734, 2017. 2
- [8] A. Mallya and S. Lazebnik. PackNet: Adding Multiple Tasks to a Single Network by Iterative Pruning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7765–7773, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society. 2
- [9] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights. In Vittorio Ferrari, Cristian Sminchisescu, Yair Weiss, and Martial Hebert, editors, *Computer Vision ECCV 2018 - 15th European Conference, 2018, Proceedings*, Lecture Notes in Computer Science, pages 72–88. Springer-Verlag Berlin Heidelberg, 2018. 2
- [10] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming Catastrophic Forgetting with Hard Attention to the Task. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4548–4557, Stockholm, Sweden, 10–15 Jul 2018. PMLR. 2
- [11] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive Neural Networks. *CoRR*, abs/1606.04671, 2016. 2
- [12] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert Gate: Lifelong Learning With a Network of Experts. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2
- [13] Rahaf Aljundi, Marcus Rohrbach, and Tinne Tuytelaars. Selfless Sequential Learning. In *International Conference on Learning Representations*, 2019. 2
- [14] Zhizhong Li and Derek Hoiem. Learning Without Forgetting. In *European Conference on Computer Vision*, pages 614–629. Springer, 2016. 2
- [15] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 2, 5
- [16] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual Learning Through Synaptic Intelligence. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 2
- [17] Jonathan Schwarz, Jelena Luketina, Wojciech M. Czarnecki, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. *35th International Conference on Machine Learning, ICML 2018*, 10:7199–7208, 2018. 2
- [18] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming Catastrophic Forgetting by Incremental Moment Matching. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 46554665, Red Hook, NY, USA, 2017. Curran Associates Inc. 2
- [19] Alexander Gepperth and Florian Wiech. Simplified computation and interpretation of fisher matrices in incremental learning with deep neural networks. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Deep Learning*, pages 481–494, Cham, 2019. Springer International Publishing. 2, 5
- [20] S. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. iCaRL: Incremental Classifier and Representation Learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542, 2017. 2
- [21] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient Episodic Memory for Continual Learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 64706479, Red Hook, NY, USA, 2017. Curran Associates Inc. 2

- 972 [22] Arslan Chaudhry, MarcAurelio Ranzato, Marcus 1026
973 Rohrbach, and Mohamed Elhoseiny. Efficient Life- 1027
974 long Learning with A-GEM. In *ICLR*, 2019. 2 1028
975 1029
976 [23] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and 1030
977 Yoshua Bengio. Gradient based sample selection for 1031
978 online continual learning. In *Advances in Neural In-* 1032
979 *formation Processing Systems 32: Annual Confer-* 1033
980 *ence on Neural Information Processing Systems 2019,* 1034
981 *NeurIPS 2019, December 8-14, 2019, Vancouver, BC,* 1035
982 *Canada*, pages 11816–11825, 2019. 2 1036
983 1037
984 [24] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elho- 1038
985 seiny, Thalaisyasingam Ajanthan, Puneet K. Dokania, 1039
986 Philip H. S. Torr, and Marc’Aurelio Ranzato. On Tiny 1040
987 Episodic Memories in Continual Learning. 2019. 2 1041
988 1042
989 [25] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon 1043
990 Kim. Continual Learning with Deep Generative Re- 1044
991 play. In *Proceedings of the 31st International Con-* 1045
992 *ference on Neural Information Processing Systems,* 1046
993 *NIPS’17*, page 29943003, Red Hook, NY, USA, 2017. 1047
994 Curran Associates Inc. 2, 3, 5 1048
995 1049
996 [26] Nitin Kamra, Umang Gupta, and Yan Liu. Deep Gen- 1050
997 erative Dual Memory Network for Continual Learn- 1051
998 ing. *CoRR*, abs/1710.10368, 2017. 2 1052
999 1053
1000 [27] Pietro Buzzega, Matteo Boschini, Angelo Porrello, 1054
1001 Davide Abati, and Simone Calderara. Dark Experi- 1055
1002 ence for General Continual Learning: a Strong, Sim- 1056
1003 ple Baseline. In *33. Annual Conference on Neural In-* 1057
1004 *formation Processing Systems (NIPS)*, 2020. 2 1058
1005 1059
1006 [28] Joseph K J and Vineeth N Balasubramanian. Meta- 1060
1007 Consolidation for Continual Learning. In *Advances in* 1061
1008 *Neural Information Processing Systems*, volume 33, 1062
1009 pages 14374–14386. Curran Associates, Inc., 2020. 2 1063
1010 1064
1011 [29] Adel Tameem, Nguyen Cuong V., Turner Richard E., 1065
1012 Ghahramani Zoubin, and Weller Adrian. Interpretable 1066
1013 Continual Learning. pages 1–11, 2019. 2 1067
1014 1068
1015 [30] Sebastian Farquhar and Yarín Gal. Towards Ro- 1069
1016 bust Evaluations of Continual Learning. *CoRR*, 1070
1017 abs/1805.09733, 2018. 2 1071
1018 1072
1019 [31] Timothe Lesort, Andrei Stoian, Jean-François Goudou, 1073
1020 and David Filliat. Training Discriminative Models to 1074
1021 Evaluate Generative Ones. 2 1075
1022 1076
1023 [32] Martin Mundt, Yong Won Hong, Iuliia Pliushch, and 1077
1024 Visvanathan Ramesh. A wholistic view of contin- 1078
1025 ual learning with deep neural networks: Forgotten 1079
lessons and the bridge to active and open world learning. *arXiv*, pages 1–32, 2020. 2
- [33] Alexander Gepperth and Benedikt Pfülb. Image Modeling with Deep Convolutional Gaussian Mixture Models. In *International Joint Conference on Neural Networks*, 2021. Submitted. 2, 6, 8, 11
- [34] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 4
- [35] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. pages 1–6, 2017. 4
- [36] Shailesh Acharya, Ashok Kumar Pant, and Prashna Kumar Gyawali. Deep learning based large scale handwritten Devanagari character recognition. *SKIMA 2015 - 9th International Conference on Software, Knowledge, Information Management and Applications*, 2016. 4
- [37] Sergei Blinnikov and Richhild Moessner. Expansions for nearly Gaussian distributions. *Astronomy and Astrophysics Supplement Series*, 130(1):193–205, 1998. 4
- [38] Alexander Gepperth and Benedikt Pfülb. Gradient-Based Training of Gaussian Mixture Models in High-Dimensional Spaces. *arXiv*, 2019. 5
- [39] Aäron van den Oord and Benjamin Schrauwen. Factoring Variations in Natural Images with Deep Gaussian Mixture Models. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3518–3526, 2014. 8
- [40] Eitan Richardson and Yair Weiss. On GANs and GMMs. *Advances in Neural Information Processing Systems*, 2018-December(NeurIPS):5847–5858, 2018. 8