# ATIAM 2018 - ML Project
# Regularized auto-encoders (VAE/WAEs) applied to latent audio synthesis

**Adrien Bitton, Philippe Esling***
Institut de Recherche et Coordination Acoustique Musique (IRCAM)
UPMC - CNRS UMR 9912 - 1, Place Igor Stravinsky, F-75004 Paris
{adrien.bitton, esling}@ircam.fr

## Abstract

Auto-Encoders are a major class of unsupervised representation learning models that mirror a data distribution with a latent space that supports continuous generation by sampling and decoding latent codes to data domain. Unregularized auto-encoders do not have a training objective that structures the latent encoding, hence continuous generation is not satisfying as there is no model of the distribution that lies in-between the encoded coordinates seen during training. To address this limitation, the encoding distribution can be regularized against a latent prior and is jointly optimized with the reconstruction objective of the auto-encoder.

This learning can be implemented through stochastic Variational Inference in the form of the Variational Auto-Encoder which trains on the Evidence Lower Bound: the Negative Log-Likelihood (reconstruction) and the Kullback-Leibler Divergence that assesses the distance of each encoding to the unit Gaussian prior.

Alternatively, the Wasserstein Auto-Encoder minimizes the Wasserstein distance and leads to a different latent regularizer based on the Optimal Transport theory. The Maximum Mean Discrepancy is used to assess the distance of the mini-batch encoding to any latent prior sampling.

Their potential is to be investigated in this Machine Learning project, applied first to traditional image datasets such as MNIST and then tailored to audio synthesis.

**check the footnote ATIAM 2017 ...**

## 1 Introduction

Unsupervised learning models data and shapes representations without explicit targets nor direct assumptions over the generative processes involved. Amongst them, a key model is the Variational Auto-Encoder **VAE** (Kingma and Welling [2013]). In this, an encoding space is learned under certain constraints and regularizations. Encoder and decoder networks are jointly trained in order to fit such objectives while being able to retrieve the original data. As an example, we may refer to dimensionality reduction or data compression: a reduced code that explains the data and naturally highlights its salient properties. The corresponding training objective is called Evidence Lower Bound (ELBO) in which reconstruction is assessed with Negative Log-Likelihood (*NLL*) and regularization is performed by the point-wise Kullback-Leibler Divergence (KLD).

Alternatively, Wasserstein Auto-Encoders **WAEs** (Tolstikhin et al. [2018]) have been later introduced to address some limitations of VAEs. Two kind of latent regularizations are proposed in this

*https://esling.github.io/

implementation. The WAE-GAN uses adversarial training in the latent space where an additional discriminator is trained at distinguishing encoded latent coordinates and samples from the prior while the encoder is pushed at fooling this adversarial, hence matching the prior. Adversarial training has been broadly developed in the form of Generative Adversarial Networks (*GANs*), enabling some the most impressive artificial image generations but being prone to instability (along with other limitations). This project will focus on the second proposal, the WAE-MMD that uses the Maximum Mean Discrepancy **MMD** (Gretton et al. [2012]) to assess the distance of the mini-batch encoding against any latent prior sampling.

Hence it provides a more general implementation than the VAE:

- it does not require an analytic divergence corresponding to a chosen prior
- it computes the latent regularization over the full encoding compared to the point-wise KLD regularization in the VAE
- it is not restricted to the NLL reconstruction objective (but authors only report training on Mean Squared Error (*MSE*))

The Gaussian prior and the point-wise regularization are said to induce the blurriness reported in VAE generations. For that reasons WAEs are expected to improve generation in auto-encoders.

Both VAE and WAE training objectives comprise a reconstruction loss and a weighted regularization. As in Sønderby et al. [2016], gradually introducing the regularization by linearly increasing its weight from zero to the desired strength throughout the first training epochs has been reported to improve the generative performances of so called *beta-VAEs* regularized with KLD. Such warm-up procedure may also be applied to WAE-MMD in this project (which was not tested in the original implementation).

In the first place, evaluation of the generative qualities of each will be done on traditional image datasets such as MNIST (hand-written digits) or Fashion-MNIST (a more complex image dataset of clothing items). The resulting models can be tested on image reconstructions but also on the generative factors encoded in the latent dimensions. To visualize the later, each encoding dimension can be traversed while keeping fixed the others in order to display the data variations enabled per latent variable. Random samples and interpolations in the latent space are also of interest in order to visualize data variations jointly produced by all latent dimensions.

In the second place, evaluation will be carried on audio data in order to investigate sound synthesis with respect to latent space regularizations. Indeed, for a given set of data, the latent structures learned by the models will depend on the training regularization and the underlying probabilistic model considered, later impacting their topology and synthesis qualities. In addition to tailoring input data representation and model architecture to sound processing, generative processes specific to audio and their evaluation remain an open question to be discussed throughout the project.

## 2 Datasets

Standard image datasets will be used for validating the first model implementations and evaluations. The most widespread available is MNIST, a collection of hand-written digits (in 10 classes from #0 to #9) which has extensively been covered in the literature. Supplementary evaluation can be done on Fashion-MNIST, a dataset sharing the same base properties as MNIST (28x28 grayscale images across 10 classes) but featuring clothing items instead of digits, hence providing a possibly more complex data domain to be fitted. Both datasets can be directly loaded from torchvision.datasets (https://pytorch.org/docs/stable/torchvision/datasets.html).

To investigate audio domains, the Studio-On-Line (*SOL*) dataset may be used. It features labeled single note recordings in different dynamics, playing styles and spanning the tessitura of 12 orchestral instruments (winds, strings, brass, keyboard). To the extent of this project, we propose using spectro-temporal input data representations invertible from magnitude to audio using Griffin-Lim (eg. STFT or NSGT if scaling on log-frequency eg. Mel). This pre-processing facilitates the subsequent unsupervised learning and simplifies the network architecture in comparison with waveform models such as WaveNet.

## 3 Expected works

### 3.1 Preliminary study

Image datasets are to be used in order to validate model implementations and draw the first comparisons of generative qualities across the different auto-encoders. WAE models are to be tested against a working "reference" VAE architecture and training procedure. Amongst the possible WAE variations are:

- different reconstruction objectives (eg. NLL, MSE, cross-entropy ...)
- different MMD kernels (eg. RBF, inverse multiquadratic ...)
- different priors than unit Gaussian

A code base is provided on MNIST and Fashion-MNIST, tutorials may also be found online eg.
https://github.com/pytorch/examples/tree/master/vae
https://github.com/schelotto/Wasserstein_Autoencoders
This should also help learning the Pytorch framework, which detailed documentation can be found at https://pytorch.org/docs/stable/index.html along with diverse tutorials (https://pytorch.org/tutorials/).

Based on this, model variations are to be tested on a benchmark in test reconstructions. As reported in (Tolstikhin et al. [2018]), the closer the encoding remains to the prior, the better sample quality is to be expected. The resulting latent dissimilarities may also be reported and compared across configurations. Moreover, learned latent structures are to be visualized with respect to class labels. Finally, a first set of generative evaluations are expected, which may include latent traversals, interpolations and random latent sampling.

**Exercise 1 - Benchmark of VAEs and WAEs on image reconstructions**

1. Identifying the possible model variations to be tested and their expected impacts on the training
2. Defining a common evaluation setup that accounts for test reconstruction quality and latent matching to the prior(s)
3. Implementing, training and benchmark of the different models
4. Drawing conclusions on the possible model variations: what are their observed effects (on the training curves, on the results ...) ? are they beneficial to the purpose of this project ? are they specific to VAEs, WAEs or effective on both ?
5. Validating the code and a set of reference models to be later trained on audio data

**Exercise 2 - Exploring image generation from latent space**

1. Visualizing and comparing structures in the latent space encoding of train and test sets with respect to image classes
2. Defining and implementing a set of possible image generation techniques from the latent space (eg. sampling, traversal, interpolation ...)
3. Visualizing the generative power of the models validated in test reconstructions
4. Qualitative comparison of the results across these models
5. Quantitative comparison of the quality of generated samples (classification, MMD in the data domain ...)

### 3.2 Application to audio

Pre-processed audio data (SOL dataset) and inversion codes from magnitude can be provided for that experiment. The preliminary study is to be transferred to sound processing. Most concepts are still relevant but assessing audio generation requires further developments. At least two questions are opened, against which the different model variations may be compared.

How to synthesize audio from latent space ? As for the previous image implementations, the generative potentials of the models may be explored through latent traversals, interpolations and random sampling. The concept of generating data from a latent path (series of latent coordinates to

be decoded) becomes even more central as spectrogram data is often sliced into successive windows (that may be overlapped). The encoding of a single note recording will then result into a latent path (the series of coordinates corresponding to all its short-term spectrogram windows).

Such note to path mapping does not imply a recurrent principle (eg. using RNNs to model temporal correlations) but as successive spectrogram windows display "smooth" timbre variations, they populate the latent space into short-term features that can be consistently merged and interpolated into new timbre paths. For the purpose of generating new synthesis paths, geodesic curves may be computed on the latent manifold as proposed in Shao et al. [2017]. As the learned topology heavily depends on the prior and regularizer used for training, the resulting geodesic shootings may differ and highlight different qualities.

**Exercise 3 - Benchmark of VAEs and WAEs on spectrogram reconstructions**

1. Defining the pre/post-processing pipelines from audio signal to input training data to inversion back to signal
2. Preparing the dataset (data, metadata, subsets ...) and validate the pre/post-processing on it
3. Benchmark of the models according to the previous performance measures and with additional scores specific to spectrograms: do the models efficiently adapt to this data ? do they rank similarly or do some features crucially affect performances on the considered audio dataset ?
4. Selecting a set of efficient models for the later generative evaluations

**Exercise 4 - Exploring sound synthesis from latent space**

1. Visualizing and comparing latent structures (here might be more metadata and possible classes such as instrument, note, velocity, style ...)
2. Applying and visualizing previous generation schemes on the selected models for audio domains
3. Implementing some synthesis paths generators: traversals, random paths, circles, spirals, geodesic curves ...
4. Enjoy your latent synthesizer ... any qualitative comparison ?

How to evaluate sound generation ? Visual inspections are often more straight-forward than listening tests. Indeed visual concepts are clearer than the sound vocabulary available to define noises and musical contents. In addition to reconstruction and discrepancy scores which are 'data agnostic' (eg. images, sounds ...), evaluation and visualization of the sound synthesis may rely on audio descriptor distributions and variations.
These can be extracted using Python package LibROSA http://librosa.github.io/librosa/index.html (eg. librosa.feature.spectral_centroid/bandwidth/contrast/flatness/rolloff).

**Exercise 5 - Evaluating sound synthesis from latent space** The question is open for proposals and trials. The idea is to use the synthesis procedures implemented previously and compare them. Possibly in-between them for a given model or for a given synthesis technique to be tested across models. Audio descriptors might be used, to estimate variations, mappings, audio correlations.

## References


Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Scholkopf. Wasserstein auto-encoders. 2018. URL `https://openreview.net/pdf?id=HkL7n1-0b`.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, March 2012.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. How to train deep variational autoencoders and probabilistic ladder networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, 2016.

Hang Shao, Abhishek Kumar, and P. Thomas Fletcher. The riemannian geometry of deep generative models. *CoRR*, abs/1711.08014, 2017.