

---

# Game-Theoretic Models for Generative Learning

## PhD Thesis Proposal

---

### Abstract

Machine learning has achieved great success in classification tasks with the booming of powerful deep neural networks. However, there's still a big gap between computer and human intelligence. I identify three levels of artificial intelligence: memorization, recognition and creativity. Discriminative learning attempts to infer knowledge from data by modeling the conditional distribution of outputs given inputs, while generative learning tries to estimate the entire distribution of data and produces new samples. My thesis will focus on the latter problem and investigate generative learning from a game-theoretic perspective. We first formulate the problem as a distributionally robust game with payoff uncertainty, and then develop a robust optimization algorithm to solve the Nash equilibrium. Meanwhile, we will study the distance metric to measure similarity between distributions, which is vital to build the objective functions. Last, we plan to test our approach in simulations and apply it to train several generative models for images and audio.

## 1 Introduction

In the last decade, machine learning approaches have achieved great success with the boom of deep neural networks. There are three levels of human intelligence: memorization, recognition and generation. Computers are approaching humans on the first two levels, and now going to the third.

Apart from recognition and classification tasks, people hope to learn the entire distribution of data and generate new samples. In mathematics, it means to learn a function describing the structure of unlabeled data, for example, density estimation is a direct application. It requires inferring the real data distribution given a set of observations. There's no straightforward way to evaluate the accuracy of generative models, because the goal is to produce lifelike artificial examples and the evaluation is subjective in most cases.

In generative learning, new samples produced by the learned model should be indistinguishable from the original data and have enough diversity. Generative models are powerful tools for many tasks such as image and audio generation, video prediction, style transfer, voice conversion, and semi-supervised learning.

If we work on low-dimension data, or just want some statistics instead of producing new samples, there's no need to model the data distribution. A bunch of unsupervised learning algorithms work well, e.g., k-means, Gaussian mixture model, EM, PCA, etc. But for complex tasks such as generating images, audios and videos, neural networks demonstrate more power. There are three popular mainstream methods in deep generative learning, generative adversarial network (GAN), variational auto-encoders (VAE) and autoregressive models (e.g., WaveNet, PixelRNN).

In general, training deep generative models is hard and time consuming. For example, it requires 8 GPUs training 6 days to learn the WaveNet autoencoder. The high dimensional training data and complex objective structures lead to many problems in optimization, such as algorithm instability, saturation, and mode collapse. Moreover, the model should have strong generalization power to produce diverse artificial examples instead of just memorizing the training set.

In this research work, we plan to develop a new game-theoretic framework for generative learning. We formulate the problem as a distributionally robust game with payoff uncertainty, and develop

41 a learning algorithm to solve the robust Nash equilibrium. In this game there are several groups  
42 of players that are competitive, noncooperative and have different objectives. Each player works  
43 in a continuous action space to optimize its expected worst-case performance. The players are  
44 implemented with neural network models to perform prediction, classification or data generation  
45 tasks. Agents with similar objectives form a group and work together against the others in order to  
46 optimize their expected payoffs. The distributionally robust Nash equilibrium is achieved by solving  
47 a minimax optimization problem. We consider using stochastic optimization techniques to deal with  
48 large-scale learning tasks.

49 Iterative optimization algorithms travel to the equilibria by minimizing a loss function, which is in our  
50 case a distance metric defined to measure the similarity between two distributions. As an important  
51 part of this research, We will study the pros and cons of several kinds of distance metrics, and compare  
52 Wasserstein distance with the most prevalent information-based metrics such as Kullback-Leibler and  
53 Jensen-Shannon divergence. We plan to develop a practical method to approximately calculate the  
54 distance between distributions and give theoretical analysis and numerical evaluations.

55 We will first verify the theoretical results through simulations, and then apply our approach on real  
56 datasets for image and audio generation. We plan to work on several tasks such as object detection,  
57 vehicle tracking, image generation and audio synthesis. This research contributes to the areas of  
58 distributionally robust game, deep generative learning, stochastic optimization and time-series data  
59 analysis.

## 60 **2 Distributionally Robust Games**

### 61 **2.1 Introduction**

#### 62 **2.1.1 Distribution Uncertainty Set**

#### 63 **2.1.2 Related Work**

### 64 **2.2 Problem Formulation**

#### 65 **2.2.1 From unsupervised learning to Generative Model**

#### 66 **2.2.2 Game Theoretic Framework for Learning**

#### 67 **2.2.3 Definition**

#### 68 **2.2.4 The Existence of Distributionally Robust Nash Equilibria**

### 69 **2.3 Minimax Robust Game**

#### 70 **2.3.1 From Duality to Triality Theory**

#### 71 **2.3.2 Dimension Reduction**

#### 72 **2.3.3 Evaluation**

### 73 **2.4 Case Study: Learning a Generative Adversarial Model**

## 74 **3 Wasserstein Metric**

75 All headings should be lower case (except for first word and proper nouns), flush left, and bold.

76 First-level headings should be in 12-point type.

### 77 **3.1 Introduction**

#### 78 **3.1.1 Optimal Transportation Problem**

#### 79 **3.1.2 Definition**

80 Second-level headings should be in 10-point type.

81 **3.2 From KL divergence to Wasserstein Metric**

82 Third-level headings should be in 10-point type.

83 **3.3 Other Metrics: L1, L2, Maximum Mean Discrepancy**

84 **3.4 Dynamic Optimal Transport**

85 There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline  
86 with the text, with the heading followed by 1 em of space.

87 **3.5 Case Study: A Toy Example**

88 **4 Learning Algorithms**

89 These instructions apply to everyone.

90 **4.1 Bregman Learning under f-divergence**

91 The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as  
92 long as you maintain internal consistency. As to the format of the references themselves, any style is  
93 acceptable as long as it is used consistently.

94 The documentation for `natbib` may be found at

95 `http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

96 Of note is the command `\citet`, which produces citations appropriate for use in inline text. For  
97 example,

98 `\citet{hasselmo}` investigated\dotso

99 produces

100 Hasselmo, et al. (1995) investigated...

101 If you wish to load the `natbib` package with options, you may add the following before loading the  
102 `nips_2018` package:

103 `\PassOptionsToPackage{options}{natbib}`

104 If `natbib` clashes with another package you load, you can add the optional argument `nonatbib`  
105 when loading the style file:

106 `\usepackage[nonatbib]{nips_2018}`

107 **4.2 Distributionally Robust Optimization**

108 Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number<sup>1</sup>  
109 in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote  
110 with a horizontal rule of 2 inches (12 picas).

111 Note that footnotes are properly typeset *after* punctuation marks.<sup>2</sup>

112 **4.3 Train a Deep Generative Model**

113 All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction.  
114 The figure number and caption always appear after the figure. Place one line space before the figure

---

<sup>1</sup>Sample of the first footnote.

<sup>2</sup>As in this example.



Figure 1: Sample figure caption.

Table 1: Sample table title

Part		
Name	Description	Size ( $\mu\text{m}$ )
Dendrite	Input terminal	$\sim 100$
Axon	Output terminal	$\sim 10$
Soma	Cell body	up to $10^6$

caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

#### 4.4 Case Study: Unsupervised Learning for Clustering

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules*. We strongly suggest the use of the booktabs package, which allows for typesetting high-quality, professional tables:

<https://www.ctan.org/pkg/booktabs>

This package was used to typeset Table 1.

#### 4.5 Case Study: Generative Modeling for Image Synthesis

### 5 Domain Transfer as a Minimax Game

#### 5.1 Introduction

Many problems in machine learning involve translating data from one domain to another. For example, transfer photograph to artistic painting, convert one person’s voice to another, translate music to imitate different musical instruments, etc. In supervised learning, it assumes that the test samples have the same distribution as the training set. However, it is not valid in many practical cases, so the knowledge needs to be transferred across domains to capture the domain shift. The general problem is to learn a mapping from one domain to another, on condition that it changes the appearance style while keeps the underlying content.

Let  $x^s \in \mathbb{X}^S$  be samples in the source domain and  $x^t \in \mathbb{X}^T$  be samples in the target domain. The goal of domain transfer is to learn a mapping:  $G : \mathbb{X}^S \rightarrow \mathbb{X}^T$  such that the generated output  $\hat{x}^t = G(x^s)$

141 is indistinguishable from the real samples drawn from the target domain. The optimal mapping  $G$   
 142 transports  $\mathbb{X}^S$  to  $\hat{\mathbb{X}}^T$ , which should have the same distribution as  $\mathbb{X}^T$ .

143 We propose a game-theoretic approach to learn the mapping. The approach is based on neural  
 144 network representations to capture the high-level and low-level features, and distributionally robust  
 145 optimization to find the Nash equilibrium. In the game, there are several groups of players with  
 146 different objectives. The intergroup competition and intragroup collaboration enable the players to  
 147 learn from others and optimize their worst-case performance.

148 This work has a wide range of use cases. In classification, since labelled data is scarce and expensive,  
 149 people want to learn from unlabelled target domain by exploring the knowledge learnt from a  
 150 well-labeled source domain. Domain transfer algorithms help to translate source data samples to  
 151 target domain as well as the corresponding labels. In visual and performing arts, it’s inspiring to  
 152 automatically generate artificial paintings with user-specified style or play synthetic music with  
 153 desired timbre and musical instrument. In informatics, it’s useful to transform speaker identity by  
 154 modifying his voice to sound like another person. It is also possible to learn and mimic animal’s  
 155 vocalization and study the feedback on the artificially generated sound. For case study, we apply our  
 156 approach in two typical domain transfer tasks: image style transfer and emotional voice conversion.

## 157 5.2 Motivation

158 In machine learning, discriminative models predict labels from data by learning a conditional  
 159 distribution  $p(y|x)$ , while generative models produce new data with desired labels by drawing  
 160 samples from distribution  $p(x|y)$ . From a probabilistic modeling perspective, domain transfer models  
 161 traslate samples from source domain to target domain by learning a joint distribution  $p(x^s, x^t)$ . If  
 162 we do not have paired training data  $\{x_i^s, x_i^t\}_{i=1}^N$ , the problem is estimating joint distribution by  
 163 its marginal distributions  $p(x^s), p(x^t)$ . According to coupling theory [1], the choice of valid joint  
 164 distribution is generally not unique. Therefore, we need to make a choice so that the marginal  
 165 distributions are related in a desirable way.

166 There are bunch of assumptions and constrains proposed to deal with this ill-posed problem. Some  
 167 suggest to partially preserve the content of source domain data, such as pixel intensity, gradient and  
 168 object boundary [2][3]; others propose to keep certain properties unchanged during the transfer, such  
 169 as semantic features and class labels [4].

170 Gatys et al. [7] separate and recombine the content and style of images under the Convolutional  
 171 Neural Network (CNN) representation. In the hierarchical network, high-level content information  
 172 is stored in higher layers of the network, and low-level information such as texture and artistic  
 173 style is captured by the correlations between filter responses in different CNN layers. They assume  
 174 the representations of content and style in the Convolutional Neural Network are well separable.  
 175 However, the style comes from only one image, not the entire training set. This limination prevents it  
 176 from capturing the general theme of the target domain.

177 Zhu et al. [10] proposed a very straightforward constraint called cycle-consistency. It assumes if we  
 178 transfer a sample from the source domain to the target domain and then translate back, we should get  
 179 exactly the same sample. Choi et al. [11] generalized it to perform multiple-domain translation using  
 180 a single generative model. However, domain transfer is not a one-to-one mapping, but many-to-many.  
 181 That’s why the cycle-consistency constraint is too strong and leads to the lack of diversity in the  
 182 translated outputs.

183 Based on the similar idea, Liu et al. [8] developed the UNIT framework by making a fully shared  
 184 latent space assumption: corresponding images across domains can be mapped to the same latent  
 185 code in a shared-latent space. This assumption implies the cycle-consistency constraint. Xun et  
 186 al. [9] extended this idea to a partially shared latent space assumption, where each data sample is  
 187 generated from a shared latent code for content and a domain-specific latent code for style. Images  
 188 are translated across domains by reconstructing through encoder and decoder networks.

189 Another idea is to make assumptions over data distributions. Covariate shift [5] assumes unchanged  
 190 conditional distributions  $p(y^s|x^s), p(y^t|x^t)$  and the only difference across domains exist in the input  
 191 distributions. If the distributions share a common support, then importance weighting [6] can  
 192 help to estimate the target density  $\hat{p}(x^t) = \hat{w}(x)p(x^s)$ , where  $w(x) = p(x^s)/p(x^t)$  is estimated  
 193 by minimizing the Kullback–Leibler divergence  $KL(p(x^t), \hat{p}(x^t))$ . The weighting parameters

correspond to the last layer of the decoder network and the first layer of the encoder network, which are low-level features representing the style.

Other approaches [11][12][13] assume there exist a transformation  $\mathcal{T}$  so that the source and target distributions can be matched with the new representations,  $p(\mathcal{T}(x^s)) = p(\mathcal{T}(x^t))$ . It requires to minimize the distance between distributions, which has been discussed in chapter 3. This assumption is equivalent to finding a transposition plan  $\mathcal{T}$  such that  $p(\mathcal{T}(x^s)) = p(x^t)$ . A particular solution is the optimal plan with the minimum transportation cost, i.e., the Wasserstein distance between the source and target distributions. As a by-product of this optimization problem, minimizing the transportation cost is equivalent to matching the samples with common representations and labels, yielding better knowledge transfer across domains. Under this idea, Damodaran et al. [14] propose to minimize the Wasserstein distance between joint distributions of data-label pair  $p(x^s, y^s), p(x^t, f(x^t))$ , which aligns data samples from source and target domains as well as transfers the discriminative information to the classifier  $f$  in the target domain.

### 5.3 Related Work

Learning generative models for domain transfer is an open problem. There are several topics closely related to our work, but with different definitions.

**Deep Generative Modeling** The original objective of this topic is to generate new samples from scratch by learning complicated data distributions in an unsupervised way. At test time, it takes random noise as input and outputs realistic samples. In some cases, it can also take in conditional information to produce user-specified output. There are three main frameworks of deep generative modeling: generative adversarial networks (GANs) [15], variational auto-encoders (VAEs) [16], and auto-regressive models [17].

GANs build the generative model on the top of a discriminative network to force the output to be indistinguishable from the real samples. This model works pretty well for generating images with impressive visual quality [18] and high resolution [19]. Variations under this framework include conditional GAN [20] that generate samples conditioned on class labels, LAPGAN [21] that generates images in a coarse-to-fine fashion, WGAN-GP [22] that enables stable training of GANs without hyperparameter tuning.

VAEs use an encoder-decoder framework to model data in a latent space and optimize the reconstruction loss plus a regularizer. The generative process has two steps of sampling: first draw latent variables from  $p(z)$  and then draw datapoints from the conditional distribution  $p(x|z)$ . At test time, the encoder part is discarded and the decoder takes random noise as input to generate new samples. However, the reconstructed samples are blurry. This is because the VAE decoder assumes  $p(x|z)$  to be an isotropic Gaussian, which leads to the use of L2 loss. To remedy this, VAE-GAN [23] suggests learning the loss through a GAN discriminator.

Auto-regressive model is quite different from the above two. It aims at modeling time-varying processes by assuming that the value of a time series depends on its previous values and a stochastic term. For a sequential data sample  $x = (x_1, x_2, \dots, x_T)$ , the joint distribution  $p(x)$  is factorised as a product of conditional distributions

$$p(x) = \prod_{t=1}^T p(x_t | x_1, \dots, x_{t-1}) \quad (1)$$

This idea is quite straightforward for modeling audio sequence [24], but it also works for images. In PixelRNN [25], each image is written as a sequence, in which pixels are taken row by row from the image. The two-dimensional spatial autocorrelation of pixels is modeled by one-dimensional temporal correlations. Since the generation process is sequential, it requires a lot of GPU memory and computation time (200K updates over 32 GPUs) even after some modifications [26].

**Image Style Transfer** There are two types of style transfer problems: example-based style transfer where the style comes from one image, and domain-based style transfer where the style is learnt from a collection of images in a specific domain. The former problem originates from nonphoto-realistic rendering (NPR) [27] in computer graphic, and has the similar meaning of realistic image manipulation. The goal is to edit image in a user-specified way and keep it as realistic as possible.

243 Practical issues include texture synthesis and transfer [28], photo manipulation of shape and color  
 244 [29], photorealistic image stylization [30], etc. In general, the output should be similar to the input in  
 245 high-level structures and varies in low-level details such as color and texture.

246 Recently, Gatys et al. [31] claimed the image content and style information are separable in Convolutional  
 247 Neural Network representations. They introduced a method [32] to separate and recombine  
 248 content and style of natural images by matching feature correlations (Gram matrix) in different convolutional  
 249 layers. However, their synthesis process is slow (an hour for a 512\*512 image). Moreover,  
 250 the style from a single image is ambiguous and may not capture the general theme of an entire domain  
 251 of images.

252 The second problem, also named as image-to-image translation, learns a mapping to transfer images  
 253 from one domain to another. For example, super-resolution [39] maps low-dimensional images to high-  
 254 dimension, colorization [40] maps gray images to color; other cases include day to night, dog to cat,  
 255 young to old, summer to winter, photographs to paintings, aerial photos to maps [30,34,35,36,37,38,41].  
 256 The mapping can be learnt in a supervised or unsupervised manner. In supervised settings [33,42,43],  
 257 corresponding image pairs across domains are available for training. In unsupervised settings  
 258 [2,4,8,9,10], there’s no paired data and the training set only contains independent set of images for  
 259 each domain. Our work is under the unsupervised settings because it is more applicable, and the  
 260 training data is almost free and unlimited.

261 **Domain Adaptation** Most recognition algorithms are developed and evaluated on the same data  
 262 distributions, e.g., the public datasets ImageNet, MS-COCO, CIFAR-10, MNIST. In real applications,  
 263 people often confront performance degradation when apply a classifier trained on a source domain to  
 264 a target domain.

265 In unsupervised domain adaptation, source domain has labeled data  $\mathcal{D}_s = \{x_i^s, y_i^s\}_{i=1}^{n_s}$  while target  
 266 domain contains data without labels  $\mathcal{D}_t = \{x_i^t\}_{i=1}^{n_t}$ . The goal is to learn a classifier  $f : x_i^t \mapsto y_i^t$   
 267 for the unseen target samples by exploring the knowledge learnt from the source domain. Domain  
 268 adaptation algorithms attempt to transfer knowledge across domains by solving the domain shift  
 269 problem, i.e., the data-label distributions  $p(x^s, y^s)$  and  $p(x^t, y^t)$  are different.

270 There are many approaches to address this issue. One is to extract transferable features that are  
 271 invariant across domains [45,46], or learn representative hash codes [47] to find a common latent  
 272 space where the classifier can be used without considering the data’s origin. Another trend is to learn  
 273 the transformation between domains [48] to align the source and target datapoints through barycentric  
 274 mapping, and train a classifier on the transferred source data. Courty [49] and Damodaran[14]  
 275 proposed to look for a transformation that matches the data-label joint distributions  $p(x^s, y^s)$  in  
 276 source domain to its equivalent version  $p(x^t, y^t)$  in target domain. The predictive function  $f$  is learnt  
 277 by minimizing the optimal transport loss between the distributions  $p(x^s, y^s)$  and  $p(x^t, f(x^t))$ . As a  
 278 by-product, minimizing the optimal transport cost is equivalent to mapping a source domain sample  
 279 to a target domain sample with similar semantic content, and this is the domain transfer problem.

280 **Voice Conversion** Voice conversion (VC) aims to change a speaker’s voice to make it sounds  
 281 like spoken by another person. It is a special case of voice transformation (VT), whose goal is to  
 282 modify human speech without changing its content. VC transforms speaker identity by replacing  
 283 speaker-dependent components of the signal while maintaining the linguistic information. Speech  
 284 quality and speaker similarity are two important factors to evaluate a VC system. There are a bunch  
 285 of VC applications, such as movie dubbing, personalized TTS (Text To Speech) systems, speaker  
 286 accent or emotion transformation, speaking-aid devices, call quality enhancement, etc.

287 Most VC frameworks involve three steps: feature extraction, feature conversion, waveform generation.  
 288 In speech analysis, waveform signals are encoded into feature representations that are easy to control  
 289 and modify. Spectral envelope, mel-cepstrum, fundamental frequency ( $f_0$ ), formant frequencies and  
 290 bandwidths are the most widely used features to represent speech in short-time segments. To capture  
 291 contextual information across frames, implicit methods such as hidden Markov models (HMMs),  
 292 Long Short-Term Memory (LSTM) and recurrent neural networks (RNNs) [63] were developed.

293 The main work in VC is to transform the source feature sequences to target feature sequences that  
 294 capture the speaker identity. Most traditional VC systems perform frame-by-frame mapping under the  
 295 assumption that speech segments are independent from each other. Some recent models such as HMM  
 296 and RNN incorporate speech dynamics implicitly. There are four typical approaches to learn the

mapping function: codebook mapping (e.g., Vector quantization (VQ) [56]), mixed linear mappings (e.g., Gaussian mixture model (GMM) [57]), neural network mapping (e.g., RBM, DNN, RNN [55]), and exemplar-based mapping (e.g., non-negative matrix factorization (NMF) [58]). Beyond these, an autoregressive neural network model called WaveNet [24] was proposed. It can directly learn the mapping based on raw audio and generate speech waveforms conditioning on the speaker identity.

There are various assumptions in speech analysis and waveform generation. Source-filter models assume speech to be generated by excitation signals passing through a vocal tract, and encode speech waveforms as acoustic features that represent sound source and vocal tract independently. However, the original phase information will lose under this assumption. At conversion time, the converted target features are passed through a vocoder based on the source filter model to reconstruct the waveform. Quality degradation may happen due to the inaccurate assumption. Iterative phase reconstruction algorithm Griffin-Lim [59] was adopted to alleviate this issue. Harmonic plus noise models (HNM) [54] assume speech to be a combination of a noise component and a harmonic component, i.e., sinusoidal waves with frequencies relevant to pitch. Speech is parameterized by the fundamental frequency  $f_0$  and a spectrum which consists of a lower band of harmonic and a higher band of noise. Other assumptions include stationary speech signal, frame-by-frame mapping, time-invariant linear filter, etc. Recently, Tamamori et. al [53] proposed a speaker-dependent WaveNet vocoder that does not require explicit modeling of excitation signals and those assumptions.

[Software: Merlin baseline system, sprocket baseline system, STRAIGHT, WORLD, Griffin-Lim, Speech Signal Processing Toolkit \(SPTK\)](#)

In terms of conversion conditions, VC can be categorized into parallel and non-parallel, text-dependent and text-independent systems [50]. In parallel systems, the training corpus consists of paired recordings from the source and target speakers with same linguistic contents. The shared acoustic features can be used to train the mapping model. To get parallel feature sequences of equal length, a time-alignment step must be included to remove the temporal differences in the recordings, for example, the dynamic time warping (DTW) [56] algorithm. Phoneme transcriptions are also useful for time alignment. Non-parallel system does not require sentences with the same linguistic contents. It is much more useful and practical because non-parallel speech data is easier to collect and therefore can get larger training sets. There are several ways to learn the mapping without paired data: (1) use unit selection method [60] to choose matched linguistic feature pairs; (2) build pseudo parallel sentences using an extra automatic speech recognition (ASR) module [61]; (3) extract speaker-independent features in a latent space [62]; (4) use unpaired image-to-image translation approaches such as CycleGAN [10].

Parallel, text-dependent systems are supposed to have better performance. However, parallel utterance pairs are difficult to get. Most parallel VC systems require time alignment to extract parallel source-target features. The misalignment in automatic time alignment algorithms often leads to degradation in speech quality, while manual correction is arduous. Recently, the winner of VC Challenge 2018 [51] showed their algorithm can achieve similar results in both parallel and non-parallel settings. It first uses a lot of external speech data with phonetic transcriptions to train a speaker-independent content-posterior-feature extractor, followed by a speaker-dependent LSTM-RNN to predict fundamental frequency  $f_0$  and STRAIGHT spectral features [52], and then reconstruct the waveforms with a speaker-dependent WaveNet vocoder [53]. Moreover, Kaneko et.al [64] and Fang et. al [54] claimed their nonparallel, text-independent VC algorithms based on CycleGAN [10] perform comparable to or better than the state-of-the-art parallel approaches.

## Music Style Transfer

### 5.4 Method

### 5.5 Case Study: Image Style Transfer

### 5.6 Case Study: Emotional Voice Conversion

## 6 Dissertation Outline

The research will be split into the following four stages:



## 346 6.1 Introduction

## 347 6.2 Related Work

## 348 6.3 Distributionally Robust Games

349 In this part we introduce distributionally robust games and develop new filtering and learning  
350 architectures under this framework. The system may contain several competing neural networks: the  
351 attackers learn to generate synthetic samples that are supposed to have the same distribution as the  
352 original ones, while the defenders try to find counter-examples and create difficulties for the other  
353 side. Each player tries to perform better and beat the others, which forms a multi-agent zero-sum  
354 game with uncertain payoffs. The players use a robust optimization approach to contend with the  
355 worst-case scenario payoff. The attacker network is constructed based on the outcome of defender  
356 networks, and vice versa. The competing networks are trained together iteratively until achieving the  
357 distributional robust Nash equilibrium.

## 358 6.4 Wasserstein Metric

359 The loss function is designed to measure the similarity of two probability distributions. Unsupervised  
360 learning is conducted by minimizing the loss. We plan to study the properties of several widely used  
361 loss metrics:

- 362 • Compare L1, L2-loss, KL-divergence, f-divergence, and Wasserstein distance
- 363 • Study the time-dependent formulation of the optimal transportation cost
- 364 • Test the effect of translation and perturbation for a certain loss metric

## 365 6.5 Learning Algorithms for Robust Optimization

- 366 • Develop a specific learning algorithm to find robust Nash equilibria, which should be stable  
367 and efficient
- 368 • Compare with existing numerical optimization approaches in large-scale machine learning:  
369 SGD, Adam, Momentum, Ishikawa-Nesterov, Newton's method, conjugate gradient, natural  
370 gradient, etc.
- 371 • Compare with existing deep generative models: RBM, VAE, GAN, WGAN, etc.

## 372 6.6 Generative Modeling for Vehicle Tracking

## 373 6.7 Generative Modeling for Image Synthesis

## 374 6.8 Generative Modeling for Domain Transfer

## 375 6.9 Experiments on Large-scale Machine Learning datasets

- 376 • Maryland Traffic Surveillance Dataset (vehicle tracking, done)
- 377 • Large-scale CelebFaces Attributes Dataset (CelebA, image synthesis, done)
- 378 • Large-scale Scene Understanding Challenge (LSUN, image synthesis)
- 379 • Interactive Emotional Dyadic Motion Capture (IEMOCAP, speech synthesis)

## 380 6.10 Discussion

## 381 6.11 Conclusion and Future Work

# 382 7 Research Plan

## 383 7.1 Research Progress

384 Literature review, planning

385 Theory part on distributional robust games, Bregman learning and convex optimization

386           Theoretical analysis and comparison for L2 distance, f-divergence and Wasserstein metric  
 387           Algorithm design, overall integration, simulations, specific implementations on real prob-  
 388           lems  
 389           Application part on large-scale machine learning: experiments, evaluation and revision  
 390           Documentation and Defence

## 391   **7.2   Application on Image and Audio Synthesis**

- 392           • Test on large-scale image dataset MNIST, CelebA and LSUN
- 393           • Literature review on emotional speech classification and audio synthesis
- 394           • Compare two sound representations in generative learning: waveform and spectrogram
- 395           • Design deep generative models for emotional speech generation
- 396           • Test on voice conversion or music style transfer if possible

## 397   **7.3   Timeline**

Fall 2018	Study on Generative Models for Image Style Transfer
Spring 2019	Study on Generative Models for Audio Style Transfer
Summer 2019	Writing of PhD Thesis
August 2019	Defense of PhD Thesis

## 398   **8   Conclusion**

399   Tackling the aforementioned problems would take us much closer to real intelligent systems, and  
 400   defines three core pillars of Artificial Intelligence. However, there are many other problems which  
 401   need to be solved and integrated to achieve a fully intelligent system, e.g. navigation, learning by  
 402   imitation, cooperation, and many others.

## 403   **9   List of Publications**

### 404   **9.1   Thesis Related Publications**

- 405           Jian Gao and Hamidou Tembine, Distributionally Robust Games for Deep Generative  
 406           Learning, July 2018. DOI: 10.13140/RG.2.2.15305.44644
- 407           Jian Gao, Yida Xu, Julian Barreiro-Gomez, Massa Ndong, Michalis Smyrnis and Hamidou  
 408           Tembine (September 5th 2018) Distributionally Robust Optimization. In Jan Valdman,  
 409           Optimization Algorithms, IntechOpen. DOI: 10.5772/intechopen.76686. ISBN: 978-1-  
 410           78923-677-4
- 411           Jian Gao and Hamidou Tembine, Distributionally Robust Games: Wasserstein Metric,  
 412           International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, July  
 413           2018
- 414           Jian Gao and Hamidou Tembine, Bregman Learning for Generative Adversarial Networks,  
 415           Chinese Control and Decision Conference (CCDC), Shenyang, China, June 2018 (*Best*  
 416           *Paper Finalist Award*)
- 417           Jian Gao and Hamidou Tembine, Distributed Mean-Field-Type Filter for Vehicle Tracking,  
 418           in American Control Conference (ACC), Seattle, USA, May 2017 (*Student Travel Award*)
- 419           Dario Bauso, Jian Gao and Hamidou Tembine, Distributionally Robust Games: f-Divergence  
 420           and Learning, 11th EAI International Conference on Performance Evaluation Methodologies  
 421           and Tools (VALUETOOLS), Venice, Italy, Dec 2017

## 9.2 Other Publications

J. Gao and H. Tembine, "Distributed Mean-Field-Type Filters for Traffic Networks," in IEEE Transactions on Intelligent Transportation Systems. doi: 10.1109/TITS.2018.2816811

J. Gao and H. Tembine, "Empathy and berge equilibria in the forwarding dilemma in relay-enabled networks," 2017 International Conference on Wireless Networks and Mobile Communications (WINCOM), Rabat, 2017, pp. 1-8. doi: 10.1109/WINCOM.2017.8238199 (*Best paper Award*)

J. Gao and H. Tembine, "Correlative mean-field filter for sequential and spatial data processing," IEEE EUROCON 2017 -17th International Conference on Smart Technologies, Ohrid, 2017, pp. 243-248. doi: 10.1109/EUROCON.2017.8011113

Fanhui Shi, Jian Gao, Xixia Huang, An affine invariant approach for dense wide baseline image matching. International Journal of Distributed Sensor Networks (IJDSN) 12(12) (2016)

J. Gao and H. Tembine, "Distributed Mean-Field-Type Filters for Big Data Assimilation," 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Sydney, NSW, 2016, pp. 1446-1453. doi: 10.1109/HPCC-SmartCity-DSS.2016.0206

[1]

## References

- [1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems* 27, pages 2672–2680. Curran Associates, Inc., 2014.