Learning Generative Models Across Incomparable Spaces

Charlotte Bunne[*1,2] **David Alvarez-Melis**[1] **Andreas Krause**[2] **Stefanie Jegelka**[1]
[1]CSAIL, Massachusetts Institute of Technology (MIT)
[2]Department of Computer Science, Eidgenössische Technische Hochschule (ETH)

Abstract

Adversarial training has become the *de facto* standard for generative modeling. While adversarial approaches have shown remarkable success in learning a distribution that faithfully recovers a reference distribution *in its entirety*, they are not applicable when one wishes the generated distribution to recover some —but not all— aspects of it. For example, one might be interested in modeling purely relational or topological aspects (such as cluster or manifold structure) while ignoring or constraining absolute characteristics (e.g., global orientation in Euclidean spaces). Furthermore, such absolute aspects are not available if the data is provided in an intrinsically relational form, such as a weighted graph. In this work, we propose an approach to learn generative models across such *incomparable* spaces that relies on the Gromov-Wasserstein distance, a notion of discrepancy that compares distributions *relationally* rather than absolutely. We show how the resulting framework can be used to learn distributions across spaces of different dimensionality or even different data types.

1 Introduction

Generative Adversarial Networks (GANs) [10] and its variations [17, 1, 12] are powerful models for learning complex distributions. In broad terms, these methods rely on an *adversary* that compares (either directly or indirectly) samples from the *true* and *generated* distributions, giving rise to a notion of divergence between them. Current methods require the two distributions to be supported in sets that are *identical* or at the very least *comparable*, so that a coherent ground metric across them can be defined and lifted to a distance between distributions, e.g., via optimal transport distances [19, 9] or Integral Probability Metrics (IPM) [15, 22, 14]. In all of these cases, the spaces over which the distributions are defined must have the same dimensionality (e.g., the space of 28×28-pixel vectors for MNIST), and the generated distribution that minimizes the objective is one with the same support as the reference one. This is of course desirable when the goal is to learn to generate samples that are indeed *indistinguishable* from those of the reference distribution.

However, one might be interested in modeling only topological or relational aspects of the reference distribution, either because the absolute location of the data manifold is irrelevant (e.g., distributions over spaces of learned representations, such as word embeddings, are defined only up to rotations) or not available (e.g., if the data is accessible only as a weighted graph among sample points). Owing to their reliance on direct comparison of samples from the two distributions, traditional generative adversarial approaches are not applicable in these settings.

In this work, we propose a general method to learn generative models across *incomparable* spaces, e.g., spaces of different dimensionality or data type. Here, the topology (e.g., the relational information between samples) of the reference data manifold is preserved, but its surface-level form characteristics can vary. A key component of our method is the Gromov-Wasserstein (GW) distance [13] to compare distributions, a generalization of classic optimal transport (OT) distances to the case where the ground

[*]bunnec@ethz.ch

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

spaces are incomparable. Similar to existing OT-based generative models [19, 9], we leverage the differentiability of this distance to provide gradients for the generator, and, for efficiency, further parametrize it via a learnable adversary. The added flexibility of the GW distance necessitates additional constraining of the adversary. We achieve this by a novel Procrustes-based orthogonality regularization principle, which might be of independent interest.

While this novel framework subsumes the traditional GAN formulation (i.e., learning an identical distribution) as a particular case (Fig. 1a and e), it allows to learn cluster-preserving distributions across spaces of different dimensionality (Fig. 1b and c) and even across different data types, e.g., from graphs to Euclidean space (Fig. 1d). Thus, our method can also be understood as performing dimensionality reduction or manifold learning, but, departing from classical approaches to these problems, it recovers, in addition to the manifold structure of the data, the probability distribution defined over it. The main contribution of this work is therefore to provide a framework that substantially expands the potential applicability of generative adversarial learning.

2 Model

We consider a dataset of n observations $(x_1, ..., x_n)$ drawn from a reference distribution $p \in \mathcal{P}(\mathcal{X})$. Our goal is to learn a generative model g_θ parametrized by θ which resembles the data distribution purely based on relational and intra-structural characteristics of the dataset. The generative model, defined as $g_\theta : \mathcal{Z} \to \mathcal{Y}$ (typically a neural network), maps random noise $z \in \mathcal{Z}$ to a generator space $\mathcal{Y}$ that is independent of data space $\mathcal{X}$.

2.1 Gromov-Wasserstein Discrepancy

Classical statistical divergences are only applicable when comparing measures whose supports lie in the same metric space, or when a meaningful distance between them can be computed. Instead of relying on a metric *across* the spaces, the Gromov-Wasserstein distance [13] compares distributions by computing a discrepancy between the metrics defined *within* each of the spaces. As a consequence, it is oblivious to the specific characteristics or dimensionality of the spaces.

When the distributions $p \in \mathcal{P}(\mathcal{X})$ and $q \in \mathcal{P}(\mathcal{Y})$ being compared are accessible only through finite samples, the discrete formulation of the problem requires a similarity (or distance) matrix between the samples and a probability vector for each space, say $(D, \mathbf{p})$ and $(\bar{D}, \mathbf{q})$, with $(D, \mathbf{p}) \in \mathbb{R}^{n \times n} \times \Sigma_n$, where n is the sample size. Then, the GW discrepancy is defined as

$$GW(D, \bar{D}, \mathbf{p}, \mathbf{q}) := \min_{T \in \mathcal{U}_{\mathbf{p},\mathbf{q}}} \sum_{ijkl} L(D_{ik}, \bar{D}_{jl}) T_{ij} T_{kl}, \tag{1}$$

with coupling T, and $\mathcal{U}_{\mathbf{p},\mathbf{q}}$ being the set of all couplings between $\mathbf{p}$ and $\mathbf{q}$. If $L = L_2$, then $GW^{1/2}$ defines a (true) distance [13]. Problem (1) is a quadratic programming problem, and solving directly is prohibitive for large n. Instead, adding an entropy regularization term $\epsilon H(T)$ leads to a smoothed problem that can be solved much more efficiently through projected gradient descent methods [16], where the projection steps rely on the Sinkhorn-Knopp scaling algorithm [8]. We refer to the resulting divergence as GW_ϵ.

With entropy regularization, GW_ϵ is not a distance any more, as the discrepancy of identical metric measure spaces is then no longer zero. Similar to the Wasserstein metric [2], the estimation of $GW_\epsilon(\cdot)$ from samples yields biased gradients. We thus propose a normalized entropy regularized Gromov-Wasserstein discrepancy defined as

$$\overline{GW}_\epsilon(D, \bar{D}, \mathbf{p}, \mathbf{q}) := 2 \times GW_\epsilon(D, \bar{D}, \mathbf{p}, \mathbf{q}) - GW_\epsilon(D, D, \mathbf{p}, \mathbf{p}) - GW_\epsilon(\bar{D}, \bar{D}, \mathbf{q}, \mathbf{q}). \tag{2}$$

2.2 Gromov-Wasserstein Generative Model

As in traditional adversarial approaches, we parametrize the generator $g_\theta : \mathcal{Z} \to \mathcal{Y}$ as a neural network that maps noise samples z to features y. We train g_θ by using $\overline{GW}_\epsilon$ as a loss, i.e., for mini-batches X and Y of reference and generated samples, respectively, we compute pairwise distance matrices D and $\bar{D}$ and solve the $\overline{GW}_\epsilon$ problem, taking $\mathbf{p}$ and $\mathbf{q}$ as uniform distributions. Motivated by Salimans et al. [19] and justified by the envelope theorem [5], we do not backpropagate the gradient through the iterative computation of the GW_ϵ coupling T (Problem (1)).

While this procedure alone is often sufficient for simple problems, in high dimensions the statistical efficiency of the classical metrics is generally poor and a large number of input samples is needed to achieve a good discrimination between generator and data distribution [19]. To improve the discriminability of generator and data samples, the intra-space distance computation is learned adversarially. Data and generator samples are mapped into feature spaces in which intra-space distances are measured using the Euclidean metric, denoted by

$$D^{\omega}_{ij} := \|f_\omega(x_i) - f_\omega(x_j)\|_2\,, \text{ where } f_\omega \colon \mathcal{X} \to \mathbb{R}^s \tag{3}$$

with f_ω modeled by a neural network mapping to s-dimensional output. The feature space might not only reduce the dimensionality of $\mathcal{X}$ through a mapping onto $\mathbb{R}^s$ but also extract the important features. The original minimization problem of the generator θ thus becomes a min-max problem

$$\min_{\theta} \max_{\omega=(\tilde{\omega},\hat{\omega})} \quad \overline{GW}_\epsilon(D^{\tilde{\omega}}, D^{\hat{\omega}}, \mathbf{p}, \mathbf{q}), \tag{4}$$

where $D^{\tilde{\omega}}$ and $D^{\hat{\omega}}$ denote pairwise distance matrices of samples originating from the generator and reference domain, respectively, mapped into the feature space via f_ω (Eq. 3). Adversary f_ω and generator g_θ are optimized in an alternating training scheme.

Note that Problem (4) makes very few assumptions on the spaces $\mathcal{X}$ and $\mathcal{Y}$, requiring only that a metric be defined on them. This remarkable flexibility can be exploited to enforce various characteristics on the generated distribution. We briefly discuss some of these in Section 2.3. On the other hand, this flexibility, combined with the added degrees of freedom brought by the learned adversarial metric, requires us to regularize the adversary to ensure stable training and prevent it from overpowering the generator. We propose an effective method to do so in Section 2.4.

2.3 Constraining the Generator

Training the generator using the $\overline{GW}_\epsilon$ loss encourages it to recover the relational and geometrical properties of the reference dataset, but leaves other global aspects undetermined. We can thus *shape* the generated distribution by enforcing desired properties through constraints. For example, while any translation of a distribution would achieve the same $\overline{GW}_\epsilon$ loss, we can enforce centering around the origin by penalizing the norm of the generated samples (we use ℓ_1 regularization for the examples in Figure 1a). For computer vision tasks, filters such as total variation denoising [18] assist the training process and improve the quality of the result (see Fig. 1e).

2.4 Regularizing the Adversary

To avoid arbitrary distortion of the space by the adversary, we propose to regularize f_ω by (approximately) enforcing it to define a unitary transformation, thus restricting the magnitude of stretching it can do. Note that *directly* parametrizing f_ω as an orthogonal matrix would defeat its purpose, as the Frobenius norm is unitarily invariant. Instead, we allow f_ω a more general form, but limit its expansivity and contractivity through approximate orthogonality [23]. Orthogonal regularization has been explored as a means to prevent exploding gradients and stabilize training of neural networks, and various approaches exist to enforce it in neural networks. Saxe et al. [20] introduced a new class of random orthogonal initial conditions on the weights of neural networks stabilizing the initial training phase. By enforcing the weight matrices to be Parseval tight frames, layerwise orthogonality constraints are introduced [7, 3, 4] and deviations of the weights from orthogonality is penalized via $R_\beta(W_k) := \beta\|W_k^\top W_k - I\|_F^2$, where W_k are weights of layer k and $\|\cdot\|_F$ is the Frobenius norm.

However, these approaches enforce orthogonality on the weights of each layer rather than constraining the network f_ω in its entirety to function as an orthogonal operator. To achieve this, we introduce a new orthogonal regularization approach, which ensures orthogonality of a network by minimizing the distance to its closest orthogonal matrix P^*. The regularization term is defined as

$$R_\beta(f_\omega(X), X) := \beta\|f_\omega(X) - XP^{*\top}\|_F^2, \tag{5}$$

where P^* is an orthogonal matrix that most closely maps X to $f_\omega(X)$, and β is a hyperparameter. The matrix $P^* = \arg\min_{P\in O(s)} \|f_\omega(X) - XP^\top\|_F$, where $O(s) = \left\{P \in \mathbb{R}^{s\times s} \mid P^\top P = I\right\}$ and s the dimensionality of the feature space, can be obtained by solving the orthogonal Procrustes problem, with closed-form solution $P^* = UV^\top$ [21]. Here, U and V are the left and right singular vectors of $f_\omega(X)^\top X$, i.e. $U\Sigma V^\top = \text{SVD}(f_\omega(X)^\top X)$. This regularization approach efficiently enforces orthogonality; we use it for training the adversary of the GW generative model.

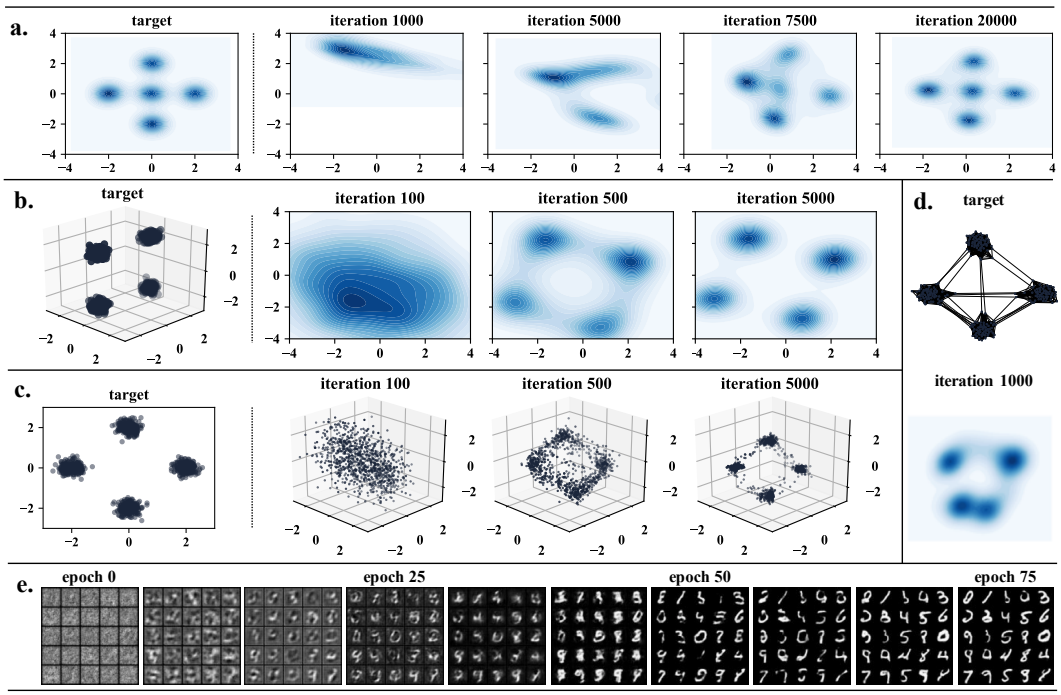

Figure 1: **Results of the GW generative model.** Learning a mixture of Gaussian distributions **a.** with adversary f_ω and ℓ_1-regularization ($\beta = 1$). The GW generative model can be applied to generate samples of **c.** reduced and **d.** increased dimensionality compared to the target distribution or **e.** map graph data into $\mathbb{R}^2$. **e.** Learning to generate MNIST digits ($\beta = 10$).

3 Results

As a proof of concept for the formulation and regularization, we illustrate the versatility of our method on popular problems for generative modeling. As a sanity check, we test the model's ability to recover 2D mixtures of Gaussians. The GW generative model recovers mixtures of Gaussians both with a static Euclidean as intra-distance representation and with a learned adversary f_ω (Fig. 1a) that stabilizes the learning. We also observe that ℓ_1-regularization indeed helps position the learned distributions around the origin. We further test whether the GW generative model can translate across different dimensionalities: as Figures 1b and c demonstrate, it learns distributions in lower and higher dimensions. In addition, we let our model learn from geodesic distances between the nodes of a graph; it indeed captures the graph structure in Euclidean space (Fig. 1d). For the experiments on synthetic datasets, generator and adversary architectures are multilayer perceptrons (MLPs) with ReLU activation functions. To illustrate the ability of the GW generative model to generate images, we train the model on MNIST [11]. Figure 1e displays generated MNIST digits throughout the training process. Both generator and adversary follow the deep convolutional architecture introduced by Chen et al. [6]. For both datasets, the adversary was constrained to approximate an orthogonal operator. The results highlight the effectiveness of the orthogonal Procrustes regularization, which allows successful learning of complex distributions without further adjustment.

4 Discussion

We presented a new generative model that learns across *incomparable* spaces by comparing intra-distances of each space. Disentangling data and generator space enables a wide range of new applications, and our preliminary experimental results suggest the effectiveness of this approach. Our novel Procrustes-based regularization principle enforces networks to function as orthogonal operators. Future work includes exploring the applicability of this regularizer in a wider range of contexts, and designing constraints to achieve specific characteristics of the generated samples.

Acknowledgments. We thank Suvrit Sra for a question initiating this research. This research was partially supported by NSF CAREER award 1553284.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, volume 70. PMLR, 2017.

[2] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The Cramer Distance as a Solution to Biased Wasserstein Gradients. *arXiv preprint arXiv:1705.10743*, 2017.

[3] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural Photo Editing with Introspective Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[4] A. Brock, J. Donahue, and K. Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[5] M. Carter. *Foundations of Mathematical Economics*. MIT Press, 2001.

[6] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[7] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval Networks: Improving Robustness to Adversarial Examples. In *International Conference on Machine Learning (ICML)*, volume 70. PMLR, 2017.

[8] M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[9] A. Genevay, G. Peyré, and M. Cuturi. Learning Generative Models with Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84. PLMR, 2018.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.

[12] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. MMD GAN: Towards Deeper Understanding of Moment Matching Network. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[13] F. Mémoli. Gromov-Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11(4), 2011.

[14] Y. Mroueh, T. Sercu, and V. Goel. McGan: Mean and Covariance Feature Matching GAN. In D. Precup and Y. W. Teh, editors, *International Conference on Machine Learning (ICML)*, volume 70. PMLR, 2017.

[15] A. Müller. Integral Probability Metrics and Their Generating Classes of Functions. *Advances in Applied Probability*, 29(2), 1997.

[16] G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In *International Conference on Machine Learning (ICML)*, volume 48, 2016.

[17] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2016.

[18] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4), 1992.

[19] T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving GANs Using Optimal Transport. In *International Conference on Learning Representations (ICLR)*, 2018.

[20] A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[21] P. H. Schönemann. A generalized solution of the Orthogonal Procrustes problem. *Psychometrika*, 31(1), 1966.

[22] B. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. Lanckriet. On the Empirical Estimation of Integral Probability Metrics. *Electronic Journal of Statistics*, 6, 2012.

[23] E. Vorontsov, C. Trabelsi, S. Kadoury, and C. Pal. On orthogonality and learning recurrent networks with long term dependencies. In *International Conference on Machine Learning (ICML)*, volume 70. PMLR, 2017.