Wasserstein-2 Generative Networks

Alexander Korotin 1  Vage Egiazarian 1  Arip Asadulaev 2  Evgeny Burnaev 1

Abstract
Modern generative learning is mainly associated with Generative Adversarial Networks (GANs). Training such networks is always hard due to the minimax nature of the optimization objective. In this paper we propose a novel algorithm for training generative models, which gets rid of minimax GAN objective, thus significantly simplified model training. The proposed algorithm uses the variational approximation of Wasserstein-2 distances by Input Convex Neural Networks. We also provide the results of computational experiments, which confirms the efficiency of our algorithm in application to latent spaces optimal transport and image-to-image style transfer.

1. Introduction
Generative learning framework have become wide spread upon last couple of years, tentatively starting with the introduction of generative adversarial networks (GANs) by (Goodfellow et al., 2014). The framework aims is to define a stochastic procedure to sample from a given complex probability distribution $Q$ on a space $Y \subset \mathbb{R}^D$, e.g. latent vector or images of faces, etc. The usual generative pipeline includes sampling from simple distribution $P$ (typically Gaussian or uniform) on space $X$ and applying a generative mapping $g: X \to Y$, that transforms noise distribution $P$ into the desired complicated distribution $Q$ on $Y$.

In many cases for a pair of probability distributions $P, Q$ there may exist several different generative mappings. For example, in Figure 1 two possible generative mappings for a pair of 1-dimensional distributions are shown. Intuitively, the mapping in Figure 1b is better than the one in Figure 1a and should be preferred. The mapping in Figure 1b is straightforward, well-structured and invertible.

Existing generative learning approaches mainly do not focus on the structural properties of the obtained generative mapping. For example, GAN-based approaches ($f$-GAN by (Nowozin et al., 2016; Yadav et al., 2017), Wasserstein-1 metric is W-GAN by (Arjovsky et al., 2017) among others (Li et al., 2017; Mroueh & Sercu, 2017; Mroueh et al., 2017)) approximate generative mapping by a neural network, architecture of which is based on the specific properties of the problem (e.g. convolutional for image-related tasks).

The reasonable question is how to find a generative mapping $g \circ P = Q$ that is well-structured. Typically, the better the structure of the mapping is, the easier is to find such a mapping. There are many ways to define what the well structured is, but usually such a mapping is expected to be continuous at least and, if possible, invertible. One may note that when $P$ and $Q$ are both one-dimensional ($X, Y \subset \mathbb{R}^1$), the only class of mappings $g : X \to Y$ satisfying these properties, are monotone mappings, i.e. $\forall x, x' \in X$

$$\langle g(x) - g(x'), x - x' \rangle > 0. \quad (1)$$

It turns out, that the intuition of 1-dimensional spaces can be easily extended to $X = Y = \mathbb{R}^D$. We can require the similar to 1 condition to hold true: $\forall x, x' \in X$

$$\langle g(x) - g(x'), x - x' \rangle > 0. \quad (2)$$

The condition is called cycle monotonicity, and every function satisfying this condition is invertible, see (Villani, 2008). What is important, is that for almost every two continuous
distributions $P, Q$ on $X = Y = \mathbb{R}^D$ there exists a unique cycle monotone mapping $g : X \rightarrow Y$ satisfying $g \circ P = Q$, see e.g. (Brenier, 1991). Thus, instead of searching for arbitrary generative mapping, one may try to significantly reduce the considered approximating class of functions by considering only cycle monotone ones.

According to (Rockafellar, 1966), every cycle monotone mapping $g$ is contained in the sub-gradient of some convex function $\psi : X \rightarrow \mathbb{R}$. Thus, every convex class of functions may produce cycle monotone mappings (by considering the gradients of such functions). In practice, deep input convex neural networks (ICNN, see (Amos et al., 2017)) can be used as convex approximations for a convex class of functions.

Formally, to fit a cycle monotone generative mapping one may apply any existing approach, such as GANs (Goodfellow et al., 2014) with the set of generators restricted to gradients of ICNN. However, these approaches require solving a minimax optimization problem, which is typically a Herculean task.

It turns out that the cycle monotone generators are highly related to Wasserstein-2 distances. Recently developed approach of (Taghvaei & Jalali, 2019) uses dual form of the Wasserstein-2 distance to find the optimal generative mapping, which is cycle monotone. It uses the gradient-descent-based algorithm by (Chartrand et al., 2009) for computing Wasserstein-2 distances. The key drawback of the algorithm is similar to the one of GANs – its optimization objective minimax.

In general, the obvious drawback of the approach with cycle monotone generators is that it is limited to spaces of the same dimension, i.e. $X, Y \subset \mathbb{R}^D$. However, in practice this issue can be easily solved by combining a generative model with a decoding part of a pre-trained autoencoder, i.e. training a generative map into a latent space (Figure 2).

In this paper, we investigate cyclically monotone generative models. The main contributions of this paper are as follows:

1. Developing an algorithm for training cyclically monotone generative models. The proposed algorithm uses non-minimax optimization objective, which improves the speed and stability of the training process in comparison to (Taghvaei & Jalali, 2019).
2. Developing a class of Input Convex Neural Networks (in particular, convolutional), gradients of which are used to approximate cyclically monotone mappings.
3. Conducting computational experiments with the proposed algorithm for problems of
   - Optimal Transport in latent spaces
   - Image-to-image Translation.

In contrast to classical GAN objectives, the proposed algorithm gets rid of minimax optimization problem and solves simpler non-convex minimization problem. The key idea of the proposed approach is to use the cycle regularization, which penalizes gradients of discriminating functions for being non-inverse.

2. Related Work

Generative learning is mainly associated with Generative Adversarial Netwoks, a paradigm invented by (Goodfellow et al., 2014). The suggested model consists of two networks: the first one, generator, is used as a generative mapping, the second one, discriminator, is used to estimate the quality of the generated samples. Several methods to estimate the proximity of the generated samples exist. They use difference measures of distance between the distributions, e.g. $f$-divergences are use in $f$-GAN by (Nowozin et al., 2016; Roth et al., 2017), Wasserstein-1 distances are used in W-GAN by (Arjovsky et al., 2017; Gulrajani et al., 2017), Maximum Mean Discrepancy in MMD GAN by (Li et al., 2017), Fisher and Sobolev distances in Fisher and Sobolev GANs respectively by (Mroueh & Sercu, 2017; Mroueh et al., 2017) among some others. Although such models sometimes report superior performance (Karras et al., 2017; Ledig et al., 2017; Zhu et al., 2017), they are extremely hard to train in practice, because they require solving non

---

**Figure 2.** Generative mapping into the latent space of an autoencoder.

**Figure 3.** Image-to-image style translation.
convex-concave minimax optimization problem (w.r.t. generator and discriminator parameters). No general efficient algorithms exists, the primary algorithm that is used is the Simultaneous Gradient Descent/Ascent (Liang & Stokes, 2018), initially suggested by (Goodfellow et al., 2014), but still it does not have any guarantees of performance.

The second paradigm, which is closely related to generative learning is the optimal transport (OT). The main problem studied in OT is how to find an optimal (generative) mapping from one distribution $\mathbb{P}$ to an other $\mathbb{Q}$.\textsuperscript{1} The formal goal is to find the mapping $g : \mathcal{X} \to \mathcal{Y}$, such that it minimized the transport cost (i.e. optimal), given by a function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$:

$$
\min_{g\in\Theta} \int_{\mathcal{X}} c(x, g(x))d\mathbb{P}(x). \quad (3)
$$

It is worth noting that if $c(x, y) = \|x - y\|$ is a norm, then the optimal transport cost (3) is equal to Wasserstein-1 distances, used in W-GAN (Arjovsky et al., 2017; Gulrajani et al., 2017). For the quadratic cost $c(x, y) = \|x - y\|^2$, the optimal transport distance is known as Wasserstein-2 distance. Several algorithms exist for finding optimal transport generative mappings (Lei et al., 2019; Seguy et al., 2017). Mainly, such algorithms solve minimization problem (instead of minimax), which is easier than minimax problem considered in GANs. However, the OT methods typically require specific restrictions during optimization, such as 1-Lipschitz discriminators (Gulrajani et al., 2017), conjugate discriminators (Seguy et al., 2017). In practice, this is often achieved by some form of regularization, which typically hepls but adds additional bias, which in some cases (Taghvaei & Jalali, 2019) makes it impossible to fit the entire distribution with required small error.

(Taghvaei & Jalali, 2019) combine the advantages of GANs with several ideas of the optimal transport. More precisely, they use the awesome properties of Wasserstein-2 distances. First of all, such distance can be computed using simple gradient-descent based algorithm (Chartrand et al., 2009), which exploits the dual form of Wasserstein Distances (Villani, 2008). Secondly, the found optimal discriminator that appears during this process, immediately leads to optimal generative mapping, which transforms on distribution into an other. Although theoretically the algorithm sounds like a powerful tool for generative learning, in practice it suffers from computational complexity: during the main optimization cycle it solves additional optimization problem. This was noted in the original paper (Taghvaei & Jalali, 2019), but is de-facto confirmed by the lack of experiments with complicated distributions.

\textsuperscript{1}Mainly, it is assumed that the underlying spaces $\mathcal{X}, \mathcal{Y}$ are of equal dimension $D$.

\section{Wasserstein-2 Optimal Transport}

We recall the main properties of Wasserstein-2 distances and their relation to cycle monotone mappings. According to (Villani, 2008), the dual form of Wasserstein-2 distance can be expressed as follows:

$$
W_2(\mathbb{P}, \mathbb{Q}) = \min_{\psi \in \text{Conv}} \left[ \int_{\mathcal{X}} \psi(x)d\mathbb{P}(x) + \int_{\mathcal{Y}} \psi^*(y)d\mathbb{Q}(y) \right] + \text{Const}(\mathbb{P}, \mathbb{Q}),
$$

where the minimum is taken over all the input convex functions (discriminators) $\psi : \mathcal{X} \to \mathbb{R}$ and

$$
\psi^*(y) = \max_x \left( \langle x, y \rangle - \psi(x) \right)
$$

is the Frenchet conjugate to $\psi$ (Fenchel, 1949) discriminator.\textsuperscript{2}

For Wasserstein-2 distances, it turns out that the optimal discriminator $\psi^*$ immediately (Chartrand et al., 2009) leads to generative mapping from $g : \mathcal{X} \to \mathcal{Y}$, such that it attains the minimum of (3). More precisely, the mapping is $g = \nabla_x \psi(x)$, i.e. gradient of a convex function. Thus, it is cycle monotone (Rockafellar, 1966). In particular, the inverse mapping can be obtain by taking the gradient w.r.t. input of the conjugate optimal discriminator $(\psi^*)^c(y)$ (McCann et al., 1995). Thus, the following properties hold true:

\begin{equation}
\begin{aligned}
\psi^{-1}(x) &= (\nabla_x \psi^*(x))^{-1} = \nabla_{\psi} (\psi^*)^c(y).
\end{aligned}
\end{equation}

In fact, one may approximate the the primal discriminator $\psi$ by input convex functions $\psi_\theta$ (e.g. input convex neural networks), and solve the optimization problem

$$
\min_{\theta \in \Theta} \left[ \int_{\mathcal{X}} \psi_\theta(x)d\mathbb{P}(x) + \int_{\mathcal{Y}} \psi_\theta^*(y)d\mathbb{Q}(y) \right].
$$

For the first integral it is straightforward to compute the derivative w.r.t. $\theta$. To compute the derivative w.r.t the second part, one need to use the properties of the conjugate function (Taghvaei & Jalali, 2019):

\begin{equation}
\begin{aligned}
\psi_\theta^*(y) &= \max_x \left( \langle x, y \rangle - \psi_\theta(x) \right) = \\
&\langle (\nabla_x \psi_\theta)^{-1}(y), y \rangle - \psi((\nabla_x \psi_\theta)^{-1}(y))
\end{aligned}
\end{equation}

Thus, to compute the gradients one needs to compute the inverse values of the current mapping $\nabla_x \psi_\theta$. To do this, the optimization problem has to be solved

$$
x = (\nabla_x \psi_\theta)^{-1}(y) \iff x = \arg \max_x \left( \langle x, y \rangle - \psi_\theta(x) \right).
$$

Although the optimization problem is convex, it is very complex, because it requires taking the gradient of $\psi$.

\textsuperscript{2}For Wasserstein-1 distances $\psi = -\psi^*$. However, the discriminators must be 1-Lipshtitz functions (Gulrajani et al., 2017), rather than convex.
We propose an approach for minimizing Wasserstein-2 distances and finding the optimal cyclically monotone generator. Note that

\[
\min_{\psi \in \mathcal{C}} \left[ \int_{\mathcal{X}} \psi(x) d\mathcal{P}(x) + \int_{\mathcal{Y}} \psi^\circ(y) d\mathcal{Q}(y) \right]
\]

where \( T : \mathcal{Y} \rightarrow \mathcal{X} \) is an arbitrary function. Formally, by considering arbitrary functions \( T \) we obtain an upper bound, which matches the entire value for \( T = (\nabla_x \psi)^{-1}(y) = \nabla_y \psi^\circ(y) \), i.e. the gradient of a convex function. Thus, the reasonable approach is to approximate both primal and dual discriminators by two different networks \( \psi_\theta \) and \( \psi_\omega^\circ \). The optimization objective is for this problem minimax and has the form

\[
\min_{\psi \in \mathcal{C}} \left[ \int_{\mathcal{X}} \psi_\theta(x) d\mathcal{P}(x) + \int_{\mathcal{Y}} \max_{\psi_\omega^\circ \in \mathcal{C}} \left[ (\nabla_y \psi_\omega^\circ(y), y) - \psi_\omega^\circ(y) \right] d\mathcal{Q}(y) \right]
\]

We optimize objective via stochastic gradient descent/ascent \( (\nabla_\psi \psi_\theta, \psi_\omega^\circ) \), which matches the entire value for \( T = (\nabla_x \psi)^{-1}(y) = \nabla_y \psi^\circ(y) \), i.e. the gradient of a convex function. Thus, the reasonable approach is to approximate both primal and dual discriminators by two different networks \( \psi_\theta \) and \( \psi_\omega^\circ \). The optimization objective is for this problem minimax and has the form

\[
\min_{\psi \in \mathcal{C}} \left[ \int_{\mathcal{X}} \psi_\theta(x) d\mathcal{P}(x) + \int_{\mathcal{Y}} \max_{\psi_\omega^\circ \in \mathcal{C}} \left[ (\nabla_y \psi_\omega^\circ(y), y) - \psi_\omega^\circ(y) \right] d\mathcal{Q}(y) \right]
\]

and can be optimized via the gradient descent over parameters \( \theta, \omega \) of primal and dual discriminators respectively.

### 3.3. Neural Network Architectures

We approximate convex functions by Input Convex Neural Networks (Amos et al., 2017). Based on the mentioned paper we construct concept of our neural networks. The overall architecture is schematically presented on Figure 4. An input convex network consists of two blocks:

1. **First block** simply equals usual linear (or convolutional) layer;
2. **Second block** is convexity preserving. It can consist almost every existing layer, but all the weight in such layer must be not negative (excluding biases), activation functions – convex and monotone.

Such a network will always be input convex.

![Schematic architecture of Input Convex Neural Network](image)

**Figure 4.** Schematically presented architecture of Input Convex Neural Network.

In all our experiments Conjugate network and Discriminator had the same architecture. We used the following networks:

1. **Multi Layer Perceptron ICNN (MLP-ICNN)** with several layers is used to get optimal mapping when input and output are latent vectors. We used the CELU activation function, which is convex and Monotone.
2. **Convolutional ICNN (Conv-ICNN)** is used for image to image style transfer task. We implemented typical architecture which consists of several convolutional layers followed by fully connected. Here we also used CELU activation’s between convolution layers.

The training was done via gradient descent with weight clipping for some weights to the positive quadrant. We also tried instead of model weight cropping, using softplus, exponent on weights and regularization, but none of these worked well.
4. Experiments

In this section, we experimentally evaluate the proposed model on several practical tasks. Each subsection below includes a particular problem description, the corresponding training procedure, and experimental results. In section 4.1 we test the model on simple toy example of fitting Gaussian mixture. In section 4.2 we apply our method for latent space optimal transport. In section 4.3 we test our method for the problem of image-to-image style translation.

To begin with, we implemented algorithm proposed by (Taghvaee & Jalali, 2019), but the model convergence even on toy examples was very slow and often it didn’t converge at all. Next, we tested our modification from section 3.1, which uses minimax objective and trains two neural networks (discriminator and conjugate discriminator). In all the convergence was much more faster, the method worked not only on toy examples. It overall got better results. Despite this, the we experienced severe difficulties (inherent in minimax problems, e.g. GANs) during training, such as hardcore balancing of the learning speed for two neural networks (by choosing learning rates, inner cycle steps, etc.). To soften these issues, we used the regularization from section 3.2, which gets rid of the minimax objective of 3.1.

4.1. Toy experiment

To begin with, we test the performance of our method in application to toy task of mapping Normal Distribution to Gaussian Mixture.

We tried our network on toy examples on various configurations such as uniform to normal, normal to Gaussian mixture and e.t.c. The example of generation from Normal Distribution to Gaussian Mixture shown on Figure 5.

4.2. Latent Space Optimal Transport

In the next series of experiments, we investigate the usage of this method on real datasets by using latent space in autoencoder. We use publicly available MNIST10k dataset. We encode $N \approx 6000$ images from the train set of the MNIST dataset into 16-dimensional vectors (by using AE); After that, we use MLP-ICNN to fit a cyclically monotone generative mapping $\mathbb{R}^D \rightarrow \mathbb{R}^D$ to map normal distribution into latent space.

The generated images by our model are shown on top in Figure 6. The bottom part of the Figure corresponds to simple latent space gaussian fit (by Gaussian mixture Model) combined with decoding model.

![Figure 6. Results on Celeba dataset. On top MLP-ICNN(Ours) and bottom single Gaussian distribution](image)

4.3. Image-to-Image Style Translation

As another application, we consider an image to image style transfer problem. In the original paper for Cycle GAN (Zhu et al., 2017) authors used 4 different neural networks - two generators and two discriminators - and the minimax objective to train the model. Our method used only 2 networks - discriminator and its conjugate neural network. Generators for forward and inverse making are gradients from those two networks.

Experimental setup. We experiment with two publicly available datasets:

- **PhototoCezanne** $N \approx 1200$ unpaired picture of size 256x256 photo and Cezanne pictures. For training we cropped images into many $128 \times 128$ parts and feed them to our neural networks.

- **WintertoSummer** This dataset consist $N \approx 1000$ unpaired picture of size $256 \times 256$ Winter and Summer photos. For training, as in previous dataset, we cropped images into many $64 \times 64$ parts and feed them to our neural networks.

We trained our model for 2 hours for each task on single NVIDIA GeForce GTX 1080 Ti. Results for Photo-toCezanne are presented on Figure 7 and on WinterSummer on Figure 8.
5. Conclusion

In this paper we developed algorithm for training generative models. The proposed algorithm imposes regularization for approximate Wasserstein-2 distances. To do this, two neural networks (discriminators) are used. The key advantage of the algorithm w.r.t. existing approaches such as GANs is that the objective is not adversarial, i.e. does not require solving minimax problem.

The results of computational experiments confirm the efficiency of the algorithm in application to latent spaces optimal transport and image-to-image style transfer.

References

Amos, B., Xu, L., and Kolter, J. Z. Input convex neural networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 146–155. JMLR. org, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. arXiv preprint arXiv:1701.07875, 2017.

Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. Communications on pure and applied mathematics, 44(4):375–417, 1991.

Chartrand, R., Wohlberg, B., Vixie, K., and Bollt, E. A gradient descent solution to the monge-kantorovich problem. Applied Mathematical Sciences, 3(22):1071–1080, 2009.

Fenchel, W. On conjugate convex functions. Canadian Journal of Mathematics, 1(1):73–77, 1949.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Advances in neural information processing systems, pp. 2672–2680, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In Advances in Neural Information Processing Systems, pp. 5767–5777, 2017.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681–4690, 2017.

Lei, N., Su, K., Cui, L., Yau, S.-T., and Gu, X. D. A geometric view of optimal transportation and generative
model. *Computer Aided Geometric Design*, 68:1–21, 2019.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pp. 2203–2213, 2017.

Liang, T. and Stokes, J. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *arXiv preprint arXiv:1802.06132*, 2018.

McCann, R. J. et al. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309–324, 1995.

Mroueh, Y. and Sercu, T. Fisher gan. In *Advances in Neural Information Processing Systems*, pp. 2513–2523, 2017.

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., and Cheng, Y. Sobolev gan. *arXiv preprint arXiv:1711.04894*, 2017.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pp. 271–279, 2016.

Rockafellar, R. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 17(3): 497–510, 1966.

Roth, K., Lucchi, A., Nowozin, S., and Hofmann, T. Stabilizing training of generative adversarial networks through regularization. In *Advances in neural information processing systems*, pp. 2018–2028, 2017.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

Taghvaei, A. and Jalali, A. 2-wasserstein approximation via restricted convex potentials with application to improved training for gans. *arXiv preprint arXiv:1902.07197*, 2019.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Yadav, A., Shah, S., Xu, Z., Jacobs, D., and Goldstein, T. Stabilizing adversarial nets with prediction methods. *arXiv preprint arXiv:1705.07364*, 2017.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.