ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200218

Dual Rejection Sampling for Wasserstein Auto-Encoders

Liang Hou[1, 2] and **Huawei Shen**[1, 2] and **Xueqi Cheng**[1, 2]

Abstract. Deep generative models enhanced by Wasserstein distance have achieved remarkable success in recent years. Wasserstein Auto-Encoders (WAEs) are auto-encoder based generative models that aim to minimize the Wasserstein distance between the data distribution and the generated distribution. The quality of generated samples of WAE depends on the distance between the data distribution and the generated distribution. However, WAE actually minimizes a Wasserstein distance between the data distribution and the reconstructed distribution in data space plus a penalty divergence between the aggregated posterior and the prior in latent space, leading a gap between theory and practice. Consequently, the quality of generated samples of WAE is not satisfactory. In this paper, we propose a novel *dual rejection sampling* method to improve the performance of WAE on the generated samples in the sampling phase. The proposed method first corrects the generative prior by a discriminator based rejection sampling scheme in latent space and then rectifies the generated distribution by another discriminator based rejection sampling method in data space. Our method is validated, both qualitatively and quantitatively, by extensive experiments on three real-world datasets.

1 INTRODUCTION

Generative models, one of the most promising research fields in machine learning, have recently achieved impressive success for many real-world applications, e.g., image synthesis [5, 17, 28], video generation [29, 33] and image classification [6, 7]. Generative models aim to learn the distribution of high-dimensional complex data and imply a generated distribution. With deep learning dominating generative models, Variational Auto-Encoders (VAEs) [19] and Generative Adversarial Networks (GANs) [11] are two well-known deep latent variable generative models, also known as implicit models. VAE is a theoretically elegant approach that maximizes a lower bound of the log-likelihood of data, leading to minimizing the Kullback–Leibler (KL) divergence between the data distribution and the generated distribution. GAN is designed as a minimax game between a generator and a discriminator. The valina GAN approximately minimizes the Jensen-Shannon (JS) divergence between the data distribution and the generated distribution. Many other divergences included by f-divergences can also be applied to GAN's framework by modifying its objective function [25]. However, those divergences are often suffered from gradient vanishing problem when two probability distributions have no intersecting supporting domain [1], which is common in high-dimensional data space.

[1] CAS Key Laboratory of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, email: {houliang17z,shenhuawei,cxq}@ict.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China

Wasserstein distance, introduced by Optimal Transport (OT) problem [35], provides a much weaker topology than many other divergences, including KL divergence, JS divergence, and other f-divergences. The weaker topology makes the Wasserstein distance have meaningful gradients wherever two probability distributions lie on. Wasserstein GAN (WGAN) [2] and WGAN-gp [13] are two GAN based generative models that minimize the Wasserstein-1 distance, also known as Earth-Mover (EM) distance, between the data distribution and the generated distribution from the dual form of OT cost. Nevertheless, they lack the inference mechanism and are notorious to train in literature [9, 30]. Wasserstein Auto-Encoders (WAEs) [32] aim to learn auto-encoder based generative models by trying to minimize the Wasserstein distance between the data distribution and the generated distribution from the modified primal form of OT cost. The primal form allows WAE to optimize many Wasserstein distances defined by different cost function, but the dual form cannot own this flexibility, which makes WGANs limited to Wasserstein-1 distance. WAE also has an inference mechanism due to the existence of the encoders and stable training processes by avoiding adversarial training in high-dimensional data space. The theory of WAE requires that the aggregated posterior induced by encoders and the generative prior must be precisely the same, which is a hard constraint that is not feasible to optimize using stochastic gradient descent. So WAE relaxed the constraint to a penalty divergence between the aggregated posterior and the prior. However, the relaxed objective function does not minimize the exact Wasserstein distance between the data distribution and the generated distribution anymore. The practical compromise deviates from the theoretical goal of WAE, resulting in unsatisfactory quality of the generated samples.

In this paper, we first analyze that the objective optimized by WAE actually consists of a Wasserstein distance between the data distribution and the reconstructed distribution plus an arbitrary divergence between the aggregated posterior and the prior. We then present a novel Dual Rejection Sampling method for Wasserstein Auto-Encoders (DRSWAE) to improve the quality of generated samples of WAE by encouraging the generated distribution closer to the data distribution in the sampling phase. DRSWAE contains two rejection sampling schemes located in latent space and in data space, respectively. The first rejection sampling in latent space aims to make the prior matching the aggregated posterior better by rejecting some latent codes with a probability according to a discriminator, which discriminates latent codes from the aggregated posterior rather than the prior. We argue that the corresponding generated distribution generated from the resampled prior through the decoder becomes closer to the reconstructed distribution, which is also closer to the data distribution than the originally generated distribution intuitively, leading a better quality of generated samples. The second rejection sampling in data space seeks to push the resampled generated distribution matching the data distribution further by rejecting unreal samples in

agreement with the probability given by another discriminator, which discriminates data samples from the data distribution rather than the resampled generated distribution. We conduct extensive experiments on three real-world datasets: MNIST, CelebA, and SVHN. The qualitative and quantitative results demonstrate the effectiveness and superiority of our proposed DRSWAE over baselines: WAE and two single rejection sampling methods for WAE in latent or data space and their oppositions.

The main contributions of this paper are summarized as follows:

- We show that the practical goal of WAE is to minimize a Wasserstein distance between the data distribution and the reconstructed distribution plus a divergence between the aggregated posterior and the prior.
- We present a novel dual rejection sampling method to improve the quality of generated samples of a trained WAE.
- We conduct extensive experiments on three real-world datasets. The qualitative and quantitative results demonstrate the effectiveness of our method.
- We perform ablation study to verify the superiority of dual rejection sampling compared with single rejection sampling in latent or data space.

In the rest of this paper, we first introduce the necessary concepts in Section 2. We then present our novel dual rejection sampling method for WAE in Section 3. We conduct extensive experiments and analyze the qualitative and quantitative results in Section 4. We introduce the most related work and discuss the relationship between us in Section 5. Finally, we conclude this work and point out future work directions in Section 6.

2 BACKGROUND

2.1 Wasserstein Auto-Encoders

Wasserstein Auto-Encoders (WAEs) are auto-encoder based generative models by trying to minimize the Wasserstein distance between data distribution P_X and generated distribution P_G theoretically [32]. Wasserstein distance is induced by OT problem. The Kantorovich's formulation of the problem is given by [35]

$$W_c(P_X, P_G) = \inf_{\tau \in \mathcal{P}(X \sim P_X, Y \sim P_G)} \mathbb{E}_{(X,Y)\sim\tau}[c(X,Y)], \quad (1)$$

where $c(X,Y) : \mathcal{X}, \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$ represents any cost function and $\mathcal{P}(X \sim P_X, Y \sim P_G)$ is a set of all joint distributions of (X,Y) with marginals P_X and P_G respectively.

Given the generated probability $p_G(x) = \int_{\mathcal{Z}} p_G(x|z)p(z)dz$ induced by a deterministic generative model $p_G(x|z) = \delta(x - G(z))$, where G represents the generator, the Kantorovich's formulation of OT problem can be converted into [24]

$$W_c(P_X, P_G) = \inf_{Q: Q_Z = P_Z} \mathbb{E}_{P_X} \mathbb{E}_{Q_{Z|X}}[c(X, G(Z))], \quad (2)$$

by introducing an inferenced distribution $Q_{Z|X}$ with the encoder Q, and Q_Z is its marginal distribution (the aggregated posterior) with probability $q(z) = \int_{\mathcal{X}} q(z|x)p(x)dx$. P_Z is the prior distribution.

Unfortunately, it is difficult to optimize objectives with hard constraints. WAE relaxed the constraint by adding a penalty into its objective function

$$D_{\text{WAE}}(P_X, P_G) = \inf_{Q_{Z|X} \in \mathcal{Q}} \mathbb{E}_{P_X} \mathbb{E}_{Q_{Z|X}}[c(X, G(Z))] + \lambda \cdot D_Z(Q_Z, P_Z), \quad (3)$$

where $\mathcal{Q}$ is a set of non-parametric encoder function, D_Z is an arbitrary divergence function between two distributions in latent space, and $\lambda > 0$ is a hyper-parameter.

2.2 Discriminator Rejection Sampling

Discriminator Rejection Sampling (DRS) performs a rejection sampling using the GAN discriminator to approximately correct errors in the GAN generator distribution [3]. Rejection sampling is a sampling method that can recover a target distribution Q_X, which cannot be sampled directly but can access the probability $q(x)$, by sampling from an easily sampled proposal distribution P_X. First, one can find a finite number M such that $Mp(x) \geq q(x)$ for $\forall x \in \mathcal{X}$, where $\mathcal{X}$ is the domain of Q_X and P_X (assume that Q_X and P_X have the same domain). Then a proposal sample x drawn from P_X will be accepted for a finally generated sample with acceptance probability $q(x)/Mp(x)$, and be rejected otherwise.

As one of the prerequisites for rejection sampling, the ratio of two probabilities $q(x)/p(x)$ is obtained by utilizing an optimal discriminator in DRS. A discriminator is defined with a sigmoid function to discriminate two classes and can be written as

$$D(x) = \sigma(d(x)) = \frac{1}{1 + e^{-d(x)}}, \quad (4)$$

where $D(x)$ is the final discriminator output, and $d(x)$ is the logit. From the Proposition 1 in the original GAN paper [11], for any fixed generator, the ratio can be derived from the optimal discriminator that discriminate a sample $x \sim Q_X$ from the prior P_X:

$$D^*(x) = \frac{q(x)}{q(x) + p(x)} \Rightarrow \frac{q(x)}{p(x)} = e^{d^*(x)}, \quad (5)$$

where d^* is the logit function of the optimal discriminator D^*.

Another remaining thing is to finding M such that scale $q(x)/Mp(x) \in [0,1], \forall x \in \mathcal{X}$ to make it as a valid acceptance probability. In practice, one can approximately estimate

$$M = \max_{x \in \mathcal{X}} \frac{q(x)}{p(x)} = \max_{x \in \mathcal{X}} e^{d^*(x)} \approx \max_{x \in \bar{\mathcal{X}}} e^{d^*(x)} = e^{d^*(x^*)}, \quad (6)$$

where $\bar{\mathcal{X}}$ is an empirical distribution on $\mathcal{X}$ and $x^* = \arg\max_{x \in \bar{\mathcal{X}}} e^{d^*(x)}$ is the sample that maximizes it.

Then the acceptance probability of a proposal sample x is given by $q(x)/Mp(x) = e^{d^*(x) - d^*(x^*)}$.

To alleviate the practical issue that rejection sampling has low acceptance probability when the target distribution is in high-dimensional space [23], DRS first considers a probability that is implicitly followed by a sigmoid function

$$\frac{1}{1 + e^{-F(x)}} = e^{d^*(x) - d^*(x^*)}. \quad (7)$$

And rewrite $F(x)$ as follows:

$$F(x) = d^*(x) - d^*(x^*) - \log(1 - e^{d^*(x) - d^*(x^*)}). \quad (8)$$

To improve the efficiency of sampling, DRS instead computes

$$\hat{F}(x) = d^*(x) - d^*(x^*) - \log(1 - e^{d^*(x) - d^*(x^*) - \epsilon}) - \gamma, \quad (9)$$

where ϵ is a small constant added for numerical stability, and γ is a hyper-parameter modulating overall acceptance probability. Negative γ will promote the acceptance of most samples, and encourage to reject for positive γ.

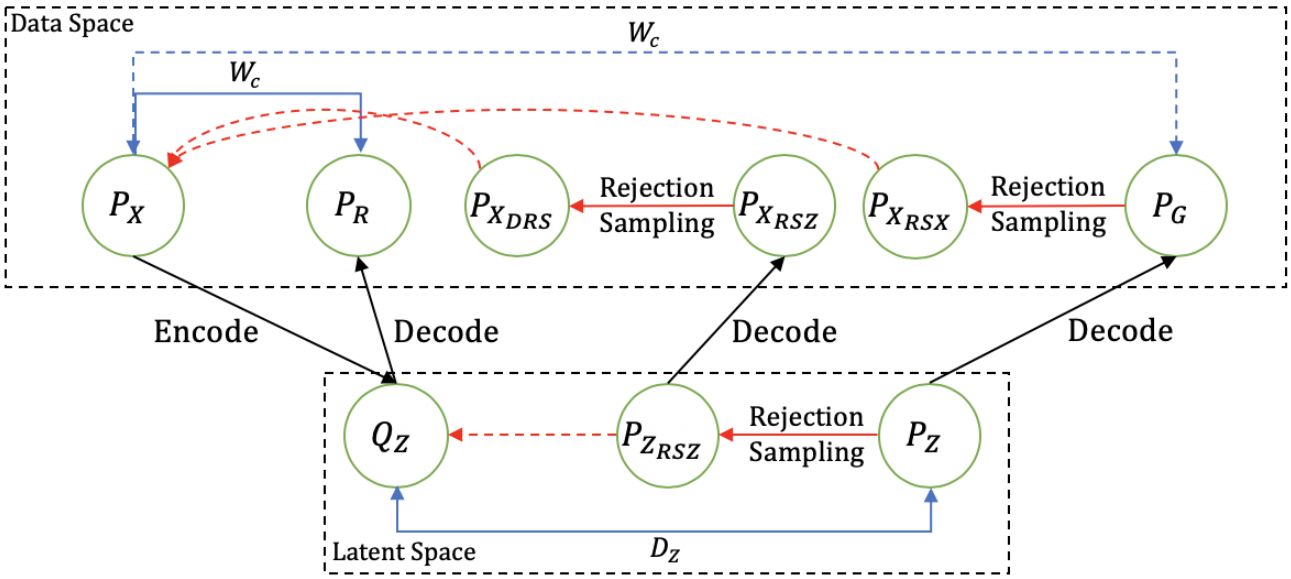


Figure 1. Illustration of our analysis on WAE and our proposed method DRSWAE. (a) The dotted blue line indicates the theoretical objective of WAE. Two solid blue lines represent the practical objectives of WAE. (b) For DRSWAE, solid red lines indicate the rejection sampling processes, and dotted red lines imply their corresponding targets. DRSWAE produces a distribution $P_{X_{DRS}}$, which is constructed through two rejection sampling processes in latent and data spaces. $P_{X_{RSZ}}$ and $P_{X_{RSX}}$ are obtained from two simplified versions, which conduct single rejection sampling in latent or data space, respectively.

3 DUAL REJECTION SAMPLING FOR WAE

3.1 Analysis on Wasserstein Auto-Encoders

We realize that the objective used in WAE is not truly minimizing the exact Wasserstein distance between data distribution P_X and generated distribution P_G anymore, because the added penalty regularization cannot guarantee that the aggregated posterior Q_Z exactly matches the prior P_Z, which should be the same in theoretical. Now the question is what the objective function of WAE optimizes? To answer this question, we first define the reconstructed distribution through the auto-encoder pair

$$\begin{aligned} p_R(y) &= \int_{\mathcal{Z}} p_G(y|z)[\int_{\mathcal{X}} q(z|x)p(x)dx]dz \\ &= \int_{\mathcal{Z}} p_G(y|z)q(z)dz. \end{aligned} \tag{10}$$

Then the Wasserstein distance between the data distribution P_X and the reconstructed distribution P_R is given by

$$W_c(P_X, P_R) = \inf_{Q} \mathbb{E}_{P_X} \mathbb{E}_{Q_{Z|X}}[c(X, G(Z)]. \tag{11}$$

The proof uses the same theory of Theorem 1 in WAE paper [32], but different from the prior used in the reconstructed distribution and the generated distribution, which is Q_Z and P_Z, respectively. This equation reveals that the reconstruction error is a Wasserstein distance between the data distribution and the reconstructed distribution. After substituting Equation 11 back into Equation 3, the objective function in WAE can be re-written as

$$D_{\text{WAE}}(P_X, P_G) = W_c(P_X, P_R) + \lambda \cdot D_Z(Q_Z, P_Z). \tag{12}$$

Now we find that WAE actually minimizes a Wasserstein distance between the data distribution P_X and the reconstructed distribution P_R in data space plus a divergence between the aggregated posterior Q_Z and the prior P_Z in latent space weighted by a hyper-parameter λ (see Figure 1).

3.2 Dual Rejection Sampling

We hypothesize that the poor performance of WAE on generation lies in the difference between practice and theory. To remedy this issue, we propose a novel Dual Rejection Sampling method for Wasserstein Auto-Encoders (DRSWAE) to minimize the distance between the generated distribution and the data distribution in the sampling phase. Therefore, we can improve the quality of generated samples of WAE. DRSWAE performs a rejection sampling in latent space (see Section 3.2.1) and a rejection sampling in data space (see Section 3.2.2) sequentially. The rejection sampling scheme is accomplished by utilizing an optimal discriminator. The details are described in Algorithm 1, which executes rejection sampling (Algorithm 2) twice.

Algorithm 1 Dual Rejection Sampling for Wasserstein Auto-Encoders (DRSWAE)

Require: dataset $\mathcal{D} = \{x_1, x_2, ..., x_n\}$, encoder E_ω, decoder/generator G_θ, discriminator in latent space D_{ϕ_Z}, discriminator in data space D_{ϕ_X}, prior P_Z

Ensure: generated samples $\mathcal{D}_G = \{y_1, y_2, ..., y_m\}$

1: $\mathcal{Z} \leftarrow \text{Sample}(P_Z)$
2: $\mathcal{Z}_E \leftarrow \text{Encode}(E_\omega, \mathcal{D})$
3: $\mathcal{D}_P \leftarrow \text{RejectionSampling}(\mathcal{Z}_E, \mathcal{Z}, \text{null}, D_{\phi_Z}, \text{null}, P_Z)$
4: $\mathcal{D}_G \leftarrow \text{RejectionSampling}(\mathcal{D}, \mathcal{D}_P, G_\theta, D_{\phi_X}, D_{\phi_Z}, P_Z)$
5: **return** $\mathcal{D}_G$

3.2.1 *Rejection sampling in latent space*

As mentioned above, WAE minimizes a Wasserstein distance $W_c(P_X, P_R)$ plus an arbitrary divergence $D_Z(Q_Z, P_Z)$. In practice, the prior P_Z is almost always impossible to exactly match the posterior Q_Z, because we optimize it in parametric space rather than function space. We also experimentally find that there exists a mismatch between the aggregated posterior and the prior (see Table 2).

Algorithm 2 Rejection Sampling in Latent/Data Space

Require: target dataset $\mathcal{D}_T = \{x_1, x_2, ..., x_n\}$, proposal dataset $\mathcal{D}_P = \{z_1, z_2, ..., z_n\}$, generator G_θ, discriminator D_ϕ, filter F_η, prior P_Z
Ensure: generated samples $\mathcal{D}_{RS} = \{y_1, y_2, ..., y_m\}$
$D_{\phi^*} \leftarrow \text{Train}(D_\phi, \mathcal{D}_T, \mathcal{D}_P)$ // Train discriminator till optimal
$M \leftarrow \text{EstimateM}(D_{\phi^*}, \{\mathcal{D}_T, \mathcal{D}_P\})$ // Equation 6
$\mathcal{D}_{RS} \leftarrow \phi$
while $|\mathcal{D}_{RS}| < m$ **do**
  $z \leftarrow \text{Sample}(P_Z)$
  if $F_\eta = \text{null}$ **then**
    $x \leftarrow z$ // Rejection sampling in latent space
  else
    while $\text{Reject}(F_\eta, z)$ **do**
      $z \leftarrow \text{Sample}(P_Z)$
    end while
    $x \leftarrow G_\theta(z)$ // Rejection sampling in data space
  end if
  $a \leftarrow \sigma(\hat{F}(x, M, \epsilon, \gamma))$ // Equation 9 and Equation 6
  $p \leftarrow \text{Uniform}(0, 1)$
  if $p \leq a$ **then**
    $\mathcal{D}_{RS} \leftarrow \text{Append}(\mathcal{D}_{RS}, x)$
  end if
end while
return $\mathcal{D}_{RS}$

Consequently, the reconstructed distribution P_R is closer to the data distribution P_X than the generated distribution P_G (see Table 1). In other words, the reconstructed samples have better quality than directly generated samples. Thus, a natural idea to improve the quality of generated samples is to sample from the reconstructed distribution P_R. Notice that we do not mean to take the reconstructed data, which do not have generalization since it is trying to memorize the empirical data, as the generated samples. To this end, we need to generalize the aggregated posterior Q_Z as the generative prior. However, it is not feasible to achieve this goal because we cannot sample from the generalized Q_Z directly. Fortunately, the assumed simple prior P_Z in deep latent variable generative models, which is usually isotropic Gaussian distribution or Uniform distribution, is easy to sample. We propose to recover the aggregated posterior Q_Z by leveraging a discriminator based rejection sampling method and regard the prior P_Z as the proposal distribution.

Then, the new resampled prior is given by

$$p_{Z_{RSZ}}(z) = \frac{1}{Z_Z} \int_{\mathcal{Z}} p(z) \frac{q(z)}{M_Z p(z)} dz, \tag{13}$$

where $Z_Z = \int_{\mathcal{Z}} p(z) \frac{q(z)}{M_Z p(z)} dz$ is the normalization term.

3.2.2 *Rejection sampling in data space*

After finishing the rejection sampling in latent space, the obtained resampled prior $P_{Z_{RSZ}}$ is closer to the aggregated posterior Q_Z than the original prior P_Z (see Table 2). The corresponding generated distribution is

$$p_{X_{RSZ}}(x) = \int_{\mathcal{Z}} p_G(x|z) p_{Z_{RSZ}}(z) dz. \tag{14}$$

We continue to conduct another discriminator based rejection sampling in data space to push the new generated distribution $P_{X_{RSZ}}$ closer to the data distribution P_X, as there is still a gap between the data distribution and the reconstructed distribution. Therefore, we can further improve the overall quality of the generated samples. One can do this by considering the target distribution as the data distribution P_X and the proposal distribution as $P_{X_{RSZ}}$.

Thus, the final generated distribution is given by

$$p_{X_{DRS}}(x) = \frac{1}{Z_X} \int_{\mathcal{X}} p_{X_{RSZ}}(x) \frac{p(x)}{M_X p_{X_{RSZ}}(x)} dx, \tag{15}$$

where $Z_X = \int_{\mathcal{X}} p_{X_{RSZ}}(x) \frac{p(x)}{M_X p_{X_{RSZ}}(x)} dx$ is the normalization term.

Notice that we do not have to calculate this expression, and we only need to able to sample from it. We share Algorithm 2 to describe the rejection sampling in latent and data spaces because they are only different from their target and proposal distributions but common with many other operations.

We also introduce two simplified versions of DRSWAE: RSZWAE and RSXWAE, which performs one rejection sampling in latent space or data space, respectively. RSXWAE can be seen as a direct application of DRS to WAE. As illustrated in Figure 1, the generated distribution of RSZWAE and RSXWAE are represented by $P_{X_{RSZ}}$ and $P_{X_{RSX}}$, respectively. Intuitively, it is promising that DRSWAE can outperform RSZWAE as it further minimizes the distance directly in data space. Compared with RSXWAE, we argue that DRSWAE achieves better results because the proposal distribution $P_{X_{RSZ}}$ is often closer to P_X than P_G, offering better conditions for rejection sampling in data space in practice.

4 EXPERIMENTS

4.1 Datasets

We conduct experiments on three real-world datasets:

- MNIST[3]: consisting of 60k hand-written digital images.
- CelebA[4]: containing roughly 203k face images.
- SVHN[5]: including about 73k street view house number images.

4.2 Model architecture

We specify $D_Z(Q_Z, P_Z) = D_{\text{JS}}(Q_Z, P_Z)$ by employing a discriminator to estimate the JS divergence. Hence, the WAE-GAN [32] model is selected as the implementation of WAE. We reuse the discriminator to perform rejection sampling in latent space. The additional network of our DRSWAE compared with WAE is a discriminator that discriminates samples from the data distribution than the generated distribution for conducting rejection sampling in data space. For encoder-decoder pairs on MNIST and CelebA datasets, We follow the exact model architecture given in the WAE paper [32]. For all datasets, the architecture of the discriminator in latent space is the same as the description in the WAE paper, and the architecture of the discriminator in data space is the same as the encoder except for its final output layer because it must output a scalar. For the new SVHN dataset, the encoder and decoder are similar to the DCGAN ones reported by [27], which both use fully convolutional architectures with

[3] http://yann.lecun.com/exdb/mnist/
[4] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[5] http://ufldl.stanford.edu/housenumbers/

4×4 convolutional filters and designed as follows:

$$\begin{aligned} z \in \mathbb{R}^{28\times 28} &\rightarrow \text{Conv}_{128} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\rightarrow \text{Conv}_{256} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\rightarrow \text{Conv}_{512} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\rightarrow \text{Conv}_{1024} \rightarrow \text{BN} \rightarrow \text{ReLU} \rightarrow \text{FC}_{32} \end{aligned}$$

$$\begin{aligned} z \in \mathbb{R}^{32} &\rightarrow \text{FC}_{2\times 2\times 1024} \rightarrow \text{ReLU} \\ &\rightarrow \text{FSConv}_{512} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\rightarrow \text{FSConv}_{256} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\rightarrow \text{FSConv}_{128} \rightarrow \text{BN} \rightarrow \text{ReLU} \rightarrow \text{FSConv}_{3} \\ &\rightarrow \text{Sigmoid} \end{aligned}$$

4.3 Experimental settings

For training basic WAE models, the experimental settings on MNIST and CelebA datasets are consistent with the original one described in the WAE paper [32]. For SVHN dataset, the parameters are the same as MNIST dataset except for the hyper-parameter $\lambda = 1$.

Because the optimal discriminators in both latent and data spaces are needed by DRSWAE, we train two discriminators for each dataset as follows:

- **MNIST** The discriminator in latent space is trained for 5 epochs with learning rate $1e^{-4}$, and the discriminator in data space is trained for 1 epoch with learning rate $1e^{-4}$. The hyper-parameter γ_z that modulating overall acceptance probability for rejection sampling in latent space is set to 80^{th} percentile of each batch (the sample at 80^{th} percentile of each batch has 50% probability of being accepted), and γ_x is set to 30^{th} percentile as of each batch in data space.
- **CelebA** The discriminator in latent space is trained for 5 epochs with learning rate $1e^{-5}$, and the discriminator in data space is trained for 1 epoch with learning rate $1e^{-5}$. We set 80^{th} percentile as γ_z and 90^{th} percentile as γ_x.
- **SVHN** The discriminator in latent space is trained for 3 epochs with learning rate $1e^{-4}$, and the discriminator in data space is trained for 1 epoch with learning rate $1e^{-4}$. We set 80^{th} percentile as γ_z and 80^{th} percentile as γ_x.

The optimizer is Adam [18] with $\beta_1 = 0.5, \beta_2 = 0.999$. In order to avoid over-fitting, we add weight decay $1e^{-8}$ for all discriminators except for the discriminator in data space on CelebA dataset, which is $1e^{-5}$. The small constant is set to $\epsilon = 1e^{-8}$.

4.4 Evaluation metrics

For evaluating the quality of the generated samples, we adopt two widely used evaluation metrics: Inception Score (IS) [30] and Fréchet Inception Distance (FID) [14]. IS computes the KL divergence between the conditional class distribution and marginal class distribution to measure the diversity and quality of the generated samples. We replace the Inception Network used in IS with a pre-trained classifier on MNIST dataset with accuracy 99.10% when calculating IS on MNIST dataset. FID is the Fréchet distance [8] between two sets of features obtained by the Inception Network. The Fréchet distance is the Wasserstein-2 distance between two Gaussian distributions: We calculate IS[6] and FID[7] using their official implementations by generating 50,000 samples. We use the pre-calculated statistics of CelebA[8] and SVHN[9] as the data distribution on computing FID.

For evaluating the distance between the aggregated posterior and the generative prior, which verifies the effectiveness of the first rejection sampling in latent space, two evaluation metrics are adopted: Fréchet Distance (FD) [8] and Sliced Wasserstein Distance (SWD) [21]. SWD is an approximation of Wasserstein Distance (WD) by randomly projecting a distribution to many one-dimensional distributions. Furthermore, the Wasserstein distance between two one-dimensional distributions has closed-form solution [20]. We calculate SWD by projecting distractions to 1,000 one-dimensional distributions.

4.5 Baselines

- WAE: the WAE-GAN model that we aim to improving.
- RSZWAE+: the accepted samples of RSZWAE.
- RSZWAE−: the rejected samples of RSZWAE.
- RSXWAE+: the accepted samples of RSXWAE.
- RSXWAE−: the rejected samples of RSXWAE.

4.6 Results

4.6.1 Results on Data Space

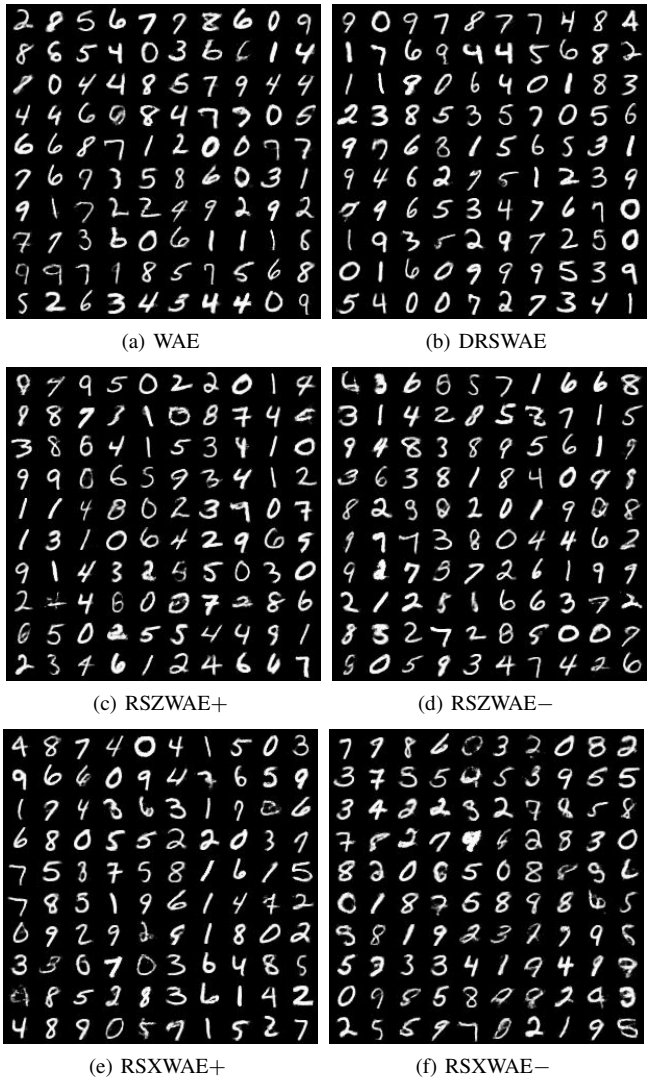
Table 1. IS and FID for samples on MNIST, CelebA and SVHN. WAE-reconstruction means the reconstructed samples of WAE.

Methods	IS (↑) MNIST	FID (↓) CelebA	FID (↓) SVHN
WAE-reconstruction	$\underline{9.62} \pm 0.02$	$\underline{33.27}$	$\underline{17.41}$
WAE	9.20 ± 0.03	39.05	24.86
RSZWAE+	9.23 ± 0.03	38.95	24.51
RSZWAE−	9.18 ± 0.03	39.20	24.61
RSXWAE+	9.29 ± 0.04	38.97	23.97
RSXWAE−	8.38 ± 0.04	39.23	25.54
DRSWAE	$\mathbf{9.51} \pm 0.02$	$\mathbf{37.28}$	$\mathbf{18.94}$

From the reported IS on MNIST dataset in Table 1, WAE performs poorly compared with WAE-reconstruction, which represents the reconstructed samples of WAE, offering potential improvements. RSZWAE+ slightly improves the quantitative result of WAE, and RSXWAE+ also increases the value. Moreover, our proposed method DRSWAE outperforms all baselines heavily. Through visual comparison on Figure 2, WAE has bad performance on visual effects. Although RSZWAE+ achieves numerical superiority on IS, it still generates some poor visual quality samples. This is probably due to its lack of correcting distribution directly in data space. RSXWAE+ moderately improves the visual quality but still synthesises a few unrecognizable digits, e.g., the sample in row 9, column 1 in Figure 2(e). The samples produced by DRSWAE have the best visual quality. It is convincing that DRSWAE outperforms RSZWAE because RSZWAE lacks a direct correction on the generated distribution. We conjecture that the reason for this poorer result of RSXWAE+, compared with DRSWAE, is that the originally

[6] https://github.com/openai/improved-gan
[7] https://github.com/bioinf-jku/TTUR
[8] http://bioinf.jku.at/research/ttur/ttur_stats/fid_stats_celeba.npz
[9] http://bioinf.jku.at/research/ttur/ttur_stats/fid_stats_svhn_train.npz

generated distribution P_G is far away from the data distribution P_X than the resampled generated distribution $P_{X_{RSZ}}$, increasing difficulty for rejection sampling in data space in practice. Both qualitative and quantitative results of DRSWAE verify our conjecture.



(a) WAE (b) DRSWAE

(c) RSZWAE+ (d) RSZWAE−

(e) RSXWAE+ (f) RSXWAE−

Figure 2. Qualitative results of MNIST.

From the FID on CelebA and SVHN datasets reported in Table 1. The results of WAE-reconstruction and WAE show that the reconstructed distribution P_R is closer to the data distribution P_X than the originally generated distribution P_G, which verifies the motivation of performing rejection sampling in latent space. Moreover, our proposed method DRSWAE achieves the best results compared with any baseline with single rejection sampling in latent or data space and their oppositions.

We also give the visual results on CelebA dataset in Figure 3, which shows that the subjective visual quality of samples generated by DRSWAE with high acceptance probability is considerably better than that rejected by DRSWAE with high rejection probability, because Figure 3(a) have clearer textures and richer colors than Figure 3(b).

(a) Samples accepted by DRSWAE with high probability in data space.

(b) Samples rejected by DRSWAE with high probability in data space.

Figure 3. Qualitative results on CelebA.

4.6.2 *Results on Latent Space*

In order to validate the first rejection sampling in latent space, we report the results that indicate the distance between the aggregated posterior and the generative prior used in each method in Table 2. First, we provide the lower limit of these metrics caused by sampling empirical distributions, named by SAME, which means the minimum value calculated by sampling two sets of independent samples drawn from the same distribution. Since the value of WAE are greater than SAME, we can conclude that WAE cannot fully optimize the penalty divergence, resulting in a mismatch between the aggregate posterior

Table 2. FD and SWD for samples on MNIST, CelebA and SVHN in latent space. The numbers in parentheses after datasets is the dimension of the latent space. We calculate FD and SWD based on the latent codes drawn from the aggregated posterior and the latent codes that finally generate samples.

Methods	MNIST(8)		CelebA(64)		SVHN(32)	
	FD($\times 10^{-2}$)	SWD($\times 10^{-4}$)	FD($\times 10^{-2}$)	SWD($\times 10^{-4}$)	FD($\times 10^{-2}$)	SWD($\times 10^{-4}$)
SAME	0.12	1.43	8.72	2.92	1.94	1.28
WAE	0.45	4.82	18.54	6.27	5.70	8.56
RSZWAE+	**0.30**	**4.07**	**15.53**	**4.42**	**1.84**	**2.77**
RSZWAE−	0.99	8.40	27.15	18.42	11.99	19.12
RSXWAE+	2.93	30.11	96.68	91.72	11.99	22.27
RSXWAE−	10.80	115.97	20.03	7.49	7.82	11.76
DRSWAE	1.78	17.90	101.60	99.20	27.64	70.91

and the prior. Those bold numbers demonstrate that the first rejection sampling in latent space really makes the resampled prior match the aggregated posterior better than the original prior, which provides more suitable conditions for rejection sampling in data space implicitly. Although our proposed DRSWAE shifts its generative prior far away from the aggregated posterior, we believe that it is in line with our expectations. The reason is that the decoder is not the inverse of the encoder. The prior, perfectly matched the aggregated posterior, only generating the reconstructed distribution, but not the data distribution. That is, the best prior fitted with the data distribution is not the aggregated posterior given the decoder.

5 RELATED WORK

5.1 Wasserstein Auto-Encoders

The Wasserstein distance inherited in WAE gives a theoretical interpretation of Adversarial Auto-Encoders (AAEs) [24]. Independent on our work, f-Wasserstein Autoencoder [16] also points out that the objective function of WAE is composed of a Wasserstein distance adding a divergence, but they do not improve the performance of WAE. To improving WAE, some works attempted to use different divergence or distance on regularizing the penalty in the relaxed objective of WAE, e.g., sliced Wasserstein distance [20] and Wasserstein distance [37]. Sinkhorn Auto-Encoders [26] utilized the Sinkhorn algorithm to minimizes the p-Wasserstein distance in latent space. A mixture of Gaussian prior to enhancing the modeling capability of WAE was proposed in [10]. However, our proposed DRSWAE method is a generic framework for improving the quality of generated samples in the sampling phase so that it can be applied to existing WAE models.

5.2 Resampled Distribution on Latent Space

The simple generative prior used in implicit models may limit their ability to create. There are some works that enrich the prior or posterior in literature. A rejection sampling for VAE to push the variational posterior to match the true posterior in training phase was proposed in [12]. As pointed in [15], the Evidence Lower Bound Objective (ELBO) in VAE can be divided into three parts: average reconstruction, index-code mutual information, and marginal KL divergence to prior. It reveals that WAE removed the mutual information term of VAE and alternated the KL divergence to any divergence. Resampled priors for VAE were presented in [4] to construct a richer prior for enhancing the representation ability by introducing acceptance networks. As a special case of WAE, AAE was improved by proposing learned priors to enrich its expression ability in [36]. However, these works only focused on the training stage but not the sampling stage, which is different from us. Recently, Discriminator Optimal Transport (DOT) [31] proposed to transport the generative codes in latent space in the generation phase to rectify the generated samples directed by an optimal discriminator. However, like the above methods, they all lacked a mechanism that polishes the final generated distribution in data space.

5.3 Resampled Distribution on Data Space

Discriminator Rejection Sampling (DRS) [3] leveraged the most relevant idea to our work. It employed a rejection sampling scheme using the GAN discriminator to approximately correct errors in the GAN generator distribution. Nevertheless, we focus our attention on WAE. Another main difference is that there is only one rejection sampling process in DRS but two rejection sampling processes in DRSWAE. We show that our proposed DRSWAE outperforms single rejection sampling in latent or data space in experiments. There are few attempts that also worked on remedying the generated distribution in literature. MH-GAN replaced the rejection sampling scheme with the Metropolis-Hastings sampling method to accelerate the speed of sampling in [34]. Moreover, in [22], the authors employed the discriminator to refine the hidden representation of generated samples to obtain better visual quality. Although DRSWAE utilizes rejection sampling methods for correcting distributions, we believe that any other sampling method, such as Metropolis-Hastings sampling, can also be adopted for a choice.

6 CONCLUSION

In this paper, we aim to improve the quality of generated samples of WAE in the sampling phase. As quality of samples is related to the distance between the generated distribution and the data distribution. We shift our goal to push the generated distribution closer to the data distribution. We find that the objective of WAE actually minimizes a Wasserstein distance between the data distribution and the reconstructed distribution plus an arbitrary divergence between the aggregated posterior and the prior. Inspired by this finding, we present a novel dual rejection sampling method to first push the prior closer to the aggregated posterior using a rejection sampling in latent space, implying a improved generated distribution. Then we further push the resampled generated distribution closer to the data distribution via another rejection sampling in data space. Extensive experiments on three real-world datasets demonstrate the effectiveness of our method. In the future, we will explore more efficient sampling methods such as MCMC.

ACKNOWLEDGEMENTS

This work is funded by the National Natural Science Foundation of China under grant numbers 61425016, 61433014, and 91746301. Huawei Shen is also funded by K.C. Wong Education Foundation and Beijing Academy of Artificial Intelligence (BAAI).

REFERENCES

[1] Martín Arjovsky and Léon Bottou, 'Towards principled methods for training generative adversarial networks', in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, (2017).

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou, 'Wasserstein generative adversarial networks', in *Proceedings of the 34th International Conference on Machine Learning*, eds., Doina Precup and Yee Whye Teh, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223, International Convention Centre, Sydney, Australia, (06–11 Aug 2017). PMLR.

[3] Samaneh Azadi, Catherine Olsson, Trevor Darrell, Ian Goodfellow, and Augustus Odena, 'Discriminator rejection sampling', in *International Conference on Learning Representations*, (2019).

[4] Matthias Bauer and Andriy Mnih, 'Resampled priors for variational autoencoders', in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 66–75, (2019).

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan, 'Large scale GAN training for high fidelity natural image synthesis', in *International Conference on Learning Representations*, (2019).

[6] Justin Cosentino and Jun Zhu, 'Generative well-intentioned networks', in *Advances in Neural Information Processing Systems 32*, eds., H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, 13077–13088, Curran Associates, Inc., (2019).

[7] Jinhao Dong and Tong Lin, 'Margingan: Adversarial training in semi-supervised learning', in *Advances in Neural Information Processing Systems 32*, eds., H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, 10440–10449, Curran Associates, Inc., (2019).

[8] DC Dowson and BV Landau, 'The fréchet distance between multivariate normal distributions', *Journal of multivariate analysis*, **12**(3), 450–455, (1982).

[9] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville, 'Adversarially learned inference', *arXiv preprint arXiv:1606.00704*, (2016).

[10] Benoit Gaujac, Ilya Feige, and David Barber, 'Improving gaussian mixture latent variable model convergence with optimal transport', (2018).

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

[12] Aditya Grover, Ramki Gummadi, Miguel Lazaro-Gredilla, Dale Schuurmans, and Stefano Ermon, 'Variational rejection sampling', in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, eds., Amos Storkey and Fernando Perez-Cruz, volume 84 of *Proceedings of Machine Learning Research*, pp. 823–832, Playa Blanca, Lanzarote, Canary Islands, (09–11 Apr 2018). PMLR.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, 'Improved training of wasserstein gans', in *Advances in Neural Information Processing Systems 30*, eds., I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5767–5777, Curran Associates, Inc., (2017).

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, 'Gans trained by a two time-scale update rule converge to a local nash equilibrium', in *Advances in Neural Information Processing Systems 30*, eds., I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 6626–6637, Curran Associates, Inc., (2017).

[15] Matthew D Hoffman and Matthew J Johnson, 'Elbo surgery: yet another way to carve up the variational evidence lower bound', in *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, (2016).

[16] Hisham Husain, Richard Nock, and Robert C Williamson, 'A primal-dual link between gans and autoencoders', in *Advances in Neural Information Processing Systems 32*, eds., H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, 413–422, Curran Associates, Inc., (2019).

[17] Tero Karras, Samuli Laine, and Timo Aila, 'A style-based generator architecture for generative adversarial networks', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (June 2019).

[18] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*, (2014).

[19] Diederik P Kingma and Max Welling, 'Auto-encoding variational bayes', *arXiv preprint arXiv:1312.6114*, (2013).

[20] Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde, 'Sliced wasserstein auto-encoders', in *International Conference on Learning Representations*, (2019).

[21] Soheil Kolouri, Yang Zou, and Gustavo K. Rohde, 'Sliced wasserstein kernels for probability distributions', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (June 2016).

[22] Yuejiang Liu, Parth Ashit Kothari, and Alexandre Alahi, 'Collaborative sampling in generative adversarial networks', Technical report, (2019).

[23] David JC MacKay and David JC Mac Kay, *Information theory, inference and learning algorithms*, Cambridge university press, 2003.

[24] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey, 'Adversarial autoencoders', *arXiv preprint arXiv:1511.05644*, (2015).

[25] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka, 'f-gan: Training generative neural samplers using variational divergence minimization', in *Advances in Neural Information Processing Systems 29*, eds., D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 271–279, Curran Associates, Inc., (2016).

[26] Giorgio Patrini, Rianne van den Berg, Patrick Forré, Marcello Carioni, Samarth Bhargav, Max Welling, Tim Genewein, and Frank Nielsen, 'Sinkhorn autoencoders', *arXiv preprint arXiv:1810.01118*, (2018).

[27] Alec Radford, Luke Metz, and Soumith Chintala, 'Unsupervised representation learning with deep convolutional generative adversarial networks', *arXiv preprint arXiv:1511.06434*, (2015).

[28] Ali Razavi, Aaron van den Oord, and Oriol Vinyals, 'Generating diverse high-fidelity images with vq-vae-2', in *Advances in Neural Information Processing Systems 32*, eds., H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, 14837–14847, Curran Associates, Inc., (2019).

[29] Masaki Saito, Eiichi Matsumoto, and Shunta Saito, 'Temporal generative adversarial nets with singular value clipping', in *The IEEE International Conference on Computer Vision (ICCV)*, (Oct 2017).

[30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen, 'Improved techniques for training gans', in *Advances in Neural Information Processing Systems 29*, eds., D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 2234–2242, Curran Associates, Inc., (2016).

[31] Akinori Tanaka, 'Discriminator optimal transport', in *Advances in Neural Information Processing Systems 32*, eds., H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, 6813–6823, Curran Associates, Inc., (2019).

[32] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf, 'Wasserstein auto-encoders', in *International Conference on Learning Representations*, (2018).

[33] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz, 'Mocogan: Decomposing motion and content for video generation', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (June 2018).

[34] Ryan Turner, Jane Hung, Eric Frank, Yunus Saatchi, and Jason Yosinski, 'Metropolis-Hastings generative adversarial networks', in *Proceedings of the 36th International Conference on Machine Learning*, eds., Kamalika Chaudhuri and Ruslan Salakhutdinov, volume 97 of *Proceedings of Machine Learning Research*, pp. 6345–6353, Long Beach, California, USA, (09–15 Jun 2019). PMLR.

[35] Cédric Villani, *Optimal transport: old and new*, volume 338, Springer Science & Business Media, 2008.

[36] Hui-Po Wang, Wei-Jan Ko, and Wen-Hsiao Peng, 'Learning priors for adversarial autoencoders', in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1388–1396. IEEE, (2018).

[37] Shunkang Zhang, Yuan Gao, Yuling Jiao, Jin Liu, Yang Wang, and Can Yang. Wasserstein-wasserstein auto-encoders, 2019.