Approximating the Wassersten Metric in GANs

Anton Mallasto — University of Copenhagen
Guido Montúfar — UCLA & MPI
Augusto Gerolin — Vrije Universiteit Amsterdam

Summary

- Generative models often aim at minimizing a similarity measure between a model and a given data distribution, in order to learn to sample from the latter.
- *Generative adversarial networks* (GANs) have been especially popular in generative modelling. Initially GANs suffered from unstable training, resulting from the Jensen-Shannon divergence they were minimizing.
- The 1-Wasserstein metric, originating from optimal transport, was proposed to be minimized instead, resulting in *Wasserstein GANs* (WGANs). In this work, we study how well do different WGAN implementations actually model the 1-Wasserstein metric.

Optimal Transport

- Let $(\mathcal{X}, d)$ be a polish space, and $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a cost function, then the optimal transport problem between two probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is

$$\mathrm{OT}_c(\mu,\nu) := \min_{\gamma} \mathbb{E}_{\gamma}[c], \tag{1}$$

where γ is a joint distribution of μ and ν, and $\mathbb{E}_\mu[f] = \int_{\mathcal{X}} f(x) d\mu(x)$.

- When $c = d^p$ and $p \geq 1$, we get the *p-Wasserstein metric*

$$W_p(\mu,\nu) := \mathrm{OT}_{d^p_{\mathcal{X}}}(\mu,\nu)^{\frac{1}{p}}, \tag{2}$$

which defines a metric distance function between probability measures with finite p^{th} moments.

- Call (φ, ψ) *admissable*, if $\varphi \oplus \psi \leq c$. Then, the optimal transport problem admits the dual

$$\mathrm{OT}_c(\mu,\nu) = \sup_{(\varphi,\psi) \text{ admissable}} \left\{ \mathbb{E}_\mu[\varphi] + \mathbb{E}_\nu[\psi] \right\}, \tag{3}$$

where the optimal (φ,ψ) are called *Kantorovich potentials*. It can be shown that the optimal ψ satisfies

$$\psi = \varphi^c, \ \varphi^c = \inf_{x \in \mathcal{X}} \{c(x,y) - \varphi(x)\}, \tag{4}$$

where φ^c is called the *c-transform* of φ.

Proposition 1 *The c-transform enforces the constraints in (3), that is, $(\varphi, \varphi^c) \in \mathrm{ADM}(c)$ for any $\varphi \in L^1(\mu)$.*

Proposition 2 *Let φ be a 1-Lipschitz function and $c = d$, the metric on $\mathcal{X}$. Then, $\varphi^c = -\varphi$.*

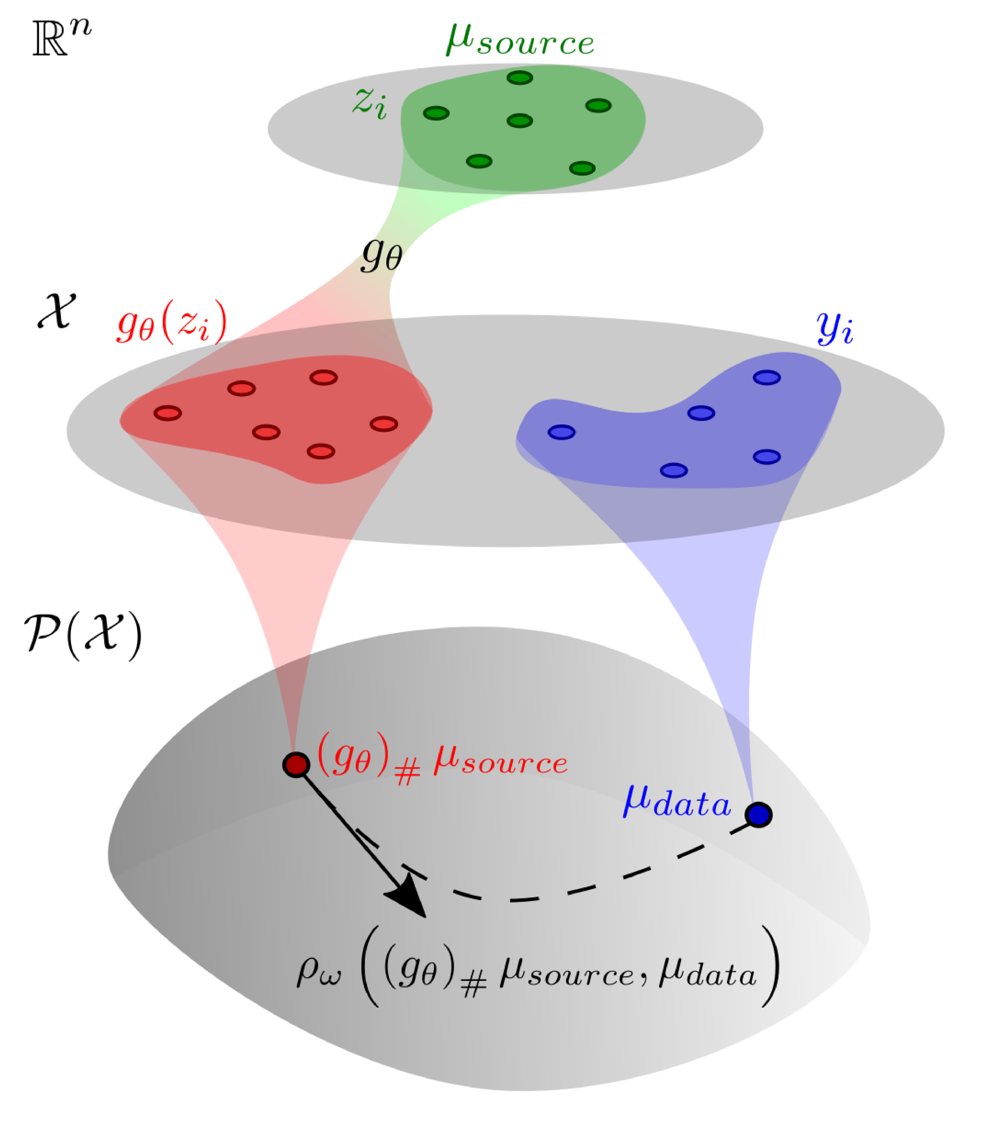

Figure 1: Visualization of the GAN setting. A source distribution μ_{source} is used to generate low dimensional elements in $\mathbb{R}^n$. These are then mapped to the data space $\mathcal{X}$ with the generator g_θ. The generated distribution and the data distribution can then be viewed as elements in the space of probabilities $\mathcal{P}(\mathcal{X})$ over $\mathcal{X}$. This space is then equipped with a similarity measure ρ_ω, which typically is computed by maximizing some quantity with respect to a discriminator φ_ω and its parameters ω. The aim is then to minimize this similarity with respect to the generator parameter θ.

Wasserstein GANs

- GANs, introduced by Goodfellow et al., aim at learning to sample from a given target data distribution $y_i \sim \mu_{data}$ by minimizing a similarity measure ρ (the Jensen-Shannon divergence in the original paper) between a model, defined by sampling $z_i \sim \mu_{source}$ lying in some low-dimensional space and mapping the points with a generator $x_i = g_\theta(z_i)$, resulting in the measure $(g_\theta)_\# \mu_{source}$. The minimization is carried out with respect to the generator parameter θ. See Fig. 1.
- The source distribution lives in some low dimensional space, which is then *pushed-forward* to the data space by the generator g_θ. The low dimensionality is justified by the *manifold hypothesis*.
- Introduced by Arjovsky et al. in 2017, the WGANs minimize the 1-Wasserstein distance, that is, $\rho = W_1$, as follows. Model the Kantorovich potential as a neural network φ_ω with parameters ω, then by Proposition 2, we can write (3) batch-wise as

$$\begin{aligned} &\min_\theta W_1\left((g_\theta)_\# \mu_{source}, \mu_{data}\right) \\ &\approx \min_\theta \max_\omega \left\{ \frac{1}{N}\sum_{i=1}^N \varphi_\omega(g_\theta(z_i)) - \frac{1}{N}\sum_{i=1}^N \varphi_\omega(y_i) \right\}, \end{aligned} \tag{5}$$

where $y_i \sim \mu_{data}$ and $z_i \sim \mu_{source}$ for $i = 1, 2, ..., N$. This defines the objective for WGANs.

Remark 1 *The main implementational difficulty is ensuring, that φ_ω is indeed 1-Lipschitz, so that $\varphi^c = -\varphi$. We could also directly ensure, that (φ,ψ) is admissable.*

Computing the Wasserstein metric

- **Weight clipping** ensures Lipschitzness for φ_ω, by clipping the weights of the neural network to lie inside some box $[-c, c]$ with $c > 0$ small. This was the strategy used by Arjovsky et al (2017).
- **Gradient penalty** was introduced by Gulrajani et al. (2017). They start by noticing, that 1-Lipschitzness over a the support of the joint distribution of μ and ν implies $\|\nabla_x \varphi_\omega(x)\| \leq 1$. This is then enforced by adding a penalty term to the objective in (5).
- **The *c*-transform**, given in (4), can be computed batch-wise. This does not yield the exact *c*-transform, as this would require a minimization over the entire support of μ. However, it does provide an admissable pair. This gives the batch-wise objective

$$\begin{aligned} &\max_\omega \left\{ \frac{1}{N}\sum_{i=1}^N \varphi_\omega(x_i) + \frac{1}{N}\sum_{i=1}^N \widehat{\varphi^c_\omega}(y_i) \right\}, \\ &\widehat{\varphi^c_\omega}(y_i) = \min_j \left\{ c(x_j, y_i) - \varphi_\omega(x_j) \right\}. \end{aligned} \tag{6}$$

Experiments

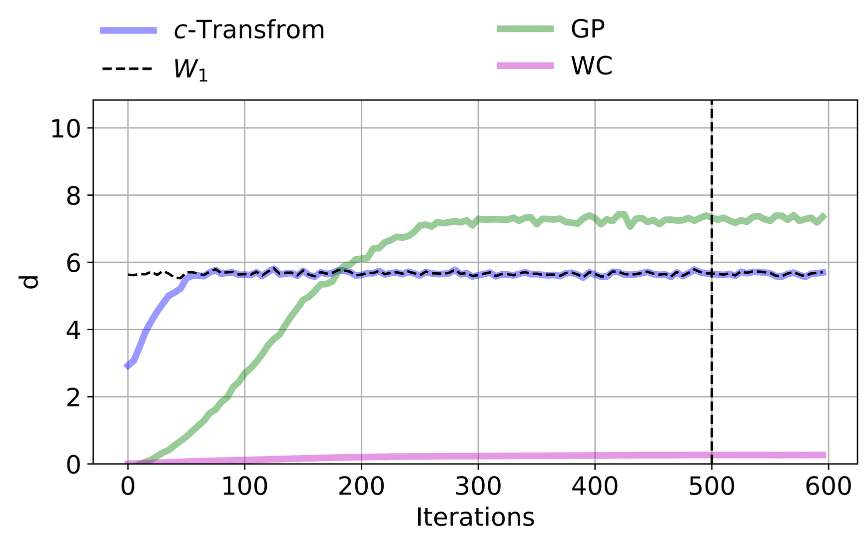

Figure 2: Estimating the distance between two standard 2-dimensional Gaussian distributions that have been shifted by $\pm[1,1]$. The discriminators are trained for the first 500 mini-batches, after which we assess how the discriminators are able to estimate the batch-wise distance.

Error	MNIST	CIFAR10	CelebA
WC	14.98 ± 0.32	27.26 ± 0.61	48.65 ± 1.29
GP	14.89 ± 0.38	27.14 ± 0.87	48.00 ± 2.88
c-transform	0.82 ± 0.16	1.53 ± 0.29	2.84 ± 0.49

Table 1: For each method, the discriminators are trained 20 times for 500 iterations on mini-batches of size 64, after which training is stopped and the error between the ground truth and the estimate are computed.

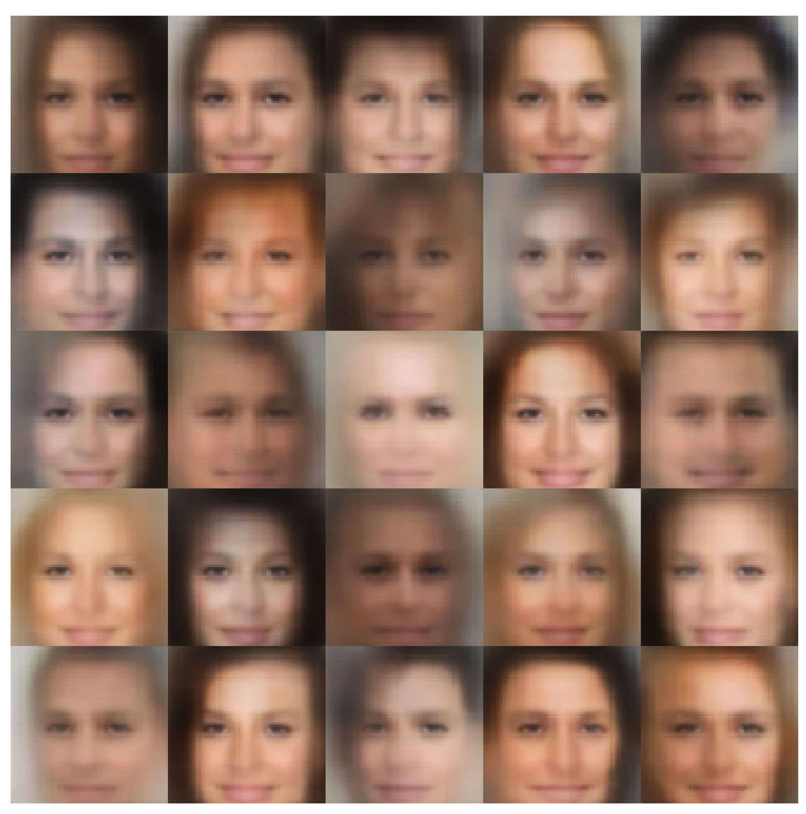

Figure 3: Images generated by a GAN with the *c*-transform. Although *c*-transform is far superior at estimating the batch-wise distance, it does not seem to be favorable for the GAN setting.

Acknowledgements

This research was supported by Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation. Part of this research was performed while the author was visiting the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation (Grant No. DMS-1440415).