Adversarial Auto-Encoders (AAEs) and Wasserstein Auto-Encoders (WAEs)

Alex Lin Melissa Yu

Harvard University

April 5, 2018

WAE Paper

Published as a conference paper at ICLR 2018

Wasserstein Auto-Encoders

Ilya Tolstikhin
MPI for Intelligent Systems
Tübingen, Germany
ilya@tue.mpg.de

Olivier Bousquet
Google Brain
Zürich, Switzerland
obousquet@google.com

Sylvain Gelly
Google Brain
Zürich, Switzerland
sylvaingelly@google.com

Bernhard Schölkopf
MPI for Intelligent Systems
Tübingen, Germany
bs@tue.mpg.de

A Quick Review of VAEs

- Let our data $x_1, \ldots, x_n \in \mathcal{X}$ be distributed according to P_X.

A Quick Review of VAEs

- Let our data $x_1, \ldots, x_n \in \mathcal{X}$ be distributed according to P_X.
- Goal of generative modeling is to find a model G with $P_G : \mathcal{X} \to [0, 1]$ that minimizes a specified distance $D(P_X, P_G)$.

A Quick Review of VAEs

- Let our data $x_1, \ldots, x_n \in \mathcal{X}$ be distributed according to P_X.
- Goal of generative modeling is to find a model G with $P_G : \mathcal{X} \to [0,1]$ that minimizes a specified distance $D(P_X, P_G)$.
- Variational Auto-Encoders (VAEs) seek to minimize

$$D_{KL}(P_X, P_G) = \underbrace{-\mathbb{E}_{P_X}[\log P_G(X)]}_{\text{negative log-likelihood}} + \underbrace{\mathbb{E}_{P_X}[\log P_X(X)]}_{\text{entropy of data}}$$

A Quick Review of VAEs

- Let our data $x_1, \ldots, x_n \in \mathcal{X}$ be distributed according to P_X.
- Goal of generative modeling is to find a model G with $P_G : \mathcal{X} \to [0, 1]$ that minimizes a specified distance $D(P_X, P_G)$.
- Variational Auto-Encoders (VAEs) seek to minimize

$$D_{KL}(P_X, P_G) = \underbrace{-\mathbb{E}_{P_X}[\log P_G(X)]}_{\text{negative log-likelihood}} + \underbrace{\mathbb{E}_{P_X}[\log P_X(X)]}_{\text{entropy of data}}$$

- The NLL cannot be optimized directly, so VAEs use an upper bound.

$$NLL \leq \underbrace{\inf_{Q(Z|X) \in \mathcal{Q}}}_{\text{min over all encoders}} -\mathbb{E}_{P_X}[\underbrace{\mathbb{E}_{Q(Z|X)}[\log P_G(X|Z)]}_{\text{reconstruction loss}} - \underbrace{D_{KL}(Q(Z|X), P_Z)}_{\text{regularization loss}}]$$

A Quick Review of VAEs

- Let our data $x_1, \ldots, x_n \in \mathcal{X}$ be distributed according to P_X.
- Goal of generative modeling is to find a model G with $P_G : \mathcal{X} \to [0,1]$ that minimizes a specified distance $D(P_X, P_G)$.
- Variational Auto-Encoders (VAEs) seek to minimize

$$D_{KL}(P_X, P_G) = \underbrace{-\mathbb{E}_{P_X}[\log P_G(X)]}_{\text{negative log-likelihood}} + \underbrace{\mathbb{E}_{P_X}[\log P_X(X)]}_{\text{entropy of data}}$$

- The NLL cannot be optimized directly, so VAEs use an upper bound.

$$NLL \leq \underbrace{\inf_{Q(Z|X) \in \mathcal{Q}}}_{\text{min over all encoders}} -\mathbb{E}_{P_X}[\underbrace{\mathbb{E}_{Q(Z|X)}[\log P_G(X|Z)]}_{\text{reconstruction loss}} - \underbrace{D_{KL}(Q(Z|X), P_Z)}_{\text{regularization loss}}]$$

- Take-Away: VAEs minimize an upper bound on KL divergence between the data and the model.

Using Wasserstein instead of KL

- Alternative way to measure distance between probability distributions

Using Wasserstein instead of KL

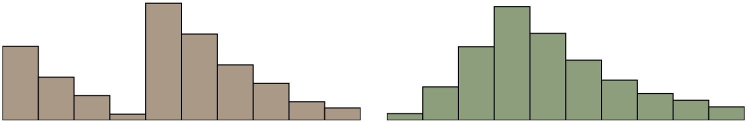
- Alternative way to measure distance between probability distributions
- Aka Earth Mover's Distance, Kantorovich-Rubinstein Metric, Optimal Transfer Plan

Using Wasserstein instead of KL

- Alternative way to measure distance between probability distributions
- Aka Earth Mover's Distance, Kantorovich-Rubinstein Metric, Optimal Transfer Plan
- Defined with respect to a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ as

$$W_c(P_X, P_Y) = \inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_Y)} \mathbb{E}_{(X,Y)\sim\Gamma}[c(X, Y)]$$

where $\Gamma(x, y)$ is the "transport plan".



Using Wasserstein instead of KL

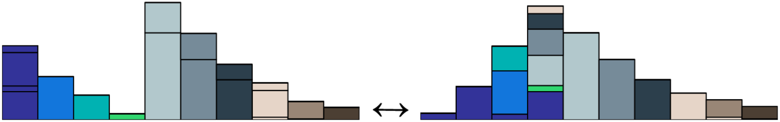
- The objective is

$$W_c(P_X, P_Y) = \inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_Y)} \mathbb{E}_{(X,Y)\sim\Gamma}[c(X, Y)]$$

Using Wasserstein instead of KL

- The objective is

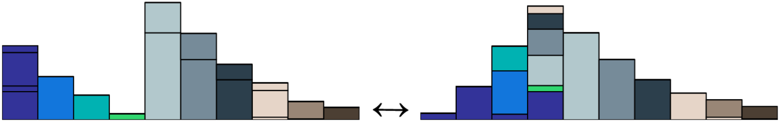
$$W_c(P_X, P_Y) = \inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_Y)} \mathbb{E}_{(X,Y)\sim\Gamma}[c(X, Y)]$$



Using Wasserstein instead of KL

- The objective is

$$W_c(P_X, P_Y) = \inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_Y)} \mathbb{E}_{(X,Y)\sim\Gamma}[c(X, Y)]$$



- Γ must be some joint distribution of X, Y because
  - $\int_x \Gamma(x, y)dx = P_X(x) \Rightarrow$ Total earth leaving x = total earth at x
  - $\int_y \Gamma(x, y)dy = P_Y(y) \Rightarrow$ Total earth entering y = total earth at y

Wasserstein Auto-Encoder

- Consider a latent space $\mathcal{Z}$ with prior P_Z. We want to find a model (i.e. decoder) $G : \mathcal{Z} \to \mathcal{X}$ that minimizes the Wasserstein distance

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

Wasserstein Auto-Encoder

- Consider a latent space $\mathcal{Z}$ with prior P_Z. We want to find a model (i.e. decoder) $G : \mathcal{Z} \to \mathcal{X}$ that minimizes the Wasserstein distance

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

- **[Bousquet et al. (2017)]** This is equivalent to minimizing

$$W_c(P_X, P_G) = \inf_{\underbrace{Q\,:\,Q_Z = P_Z}_{\text{min over encoders with marginal } P_Z}} \mathbb{E}_{P_X} \underbrace{\mathbb{E}_{Q(Z|X)}[c(X, G(Z))]}_{\text{reconstruction cost}}$$

where $Q_Z = \int Q(Z|X)P_X(X)dX$.

Wasserstein Auto-Encoder

- The WAE objective relaxes the constraint $Q_Z = P_Z$ by adding a penalty. It minimizes

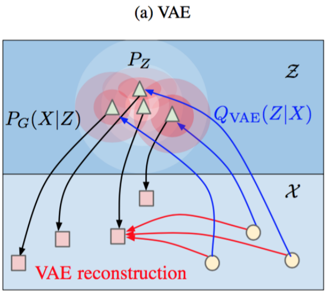
$$D_{WAE}(P_X, P_G) = \inf_{Q(Z|X) \in \mathcal{Q}} \mathbb{E}_{P_X} \mathbb{E}_{Q(Z|X)}[c(X, G(Z)] + \lambda \cdot \mathcal{D}_Z(Q_Z, P_Z)$$

where $\mathcal{D}$ is some divergence and λ is some regularization parameter.

- The WAE objective relaxes the constraint $Q_Z = P_Z$ by adding a penalty. It minimizes

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

  where $\mathcal{D}$ is some divergence and λ is some regularization parameter.
- Claims to fix blurriness issue of VAEs



Wasserstein Auto-Encoder

- The WAE objective minimizes

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

Wasserstein Auto-Encoder

- The WAE objective minimizes

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

- There are two sources of customization.

Wasserstein Auto-Encoder

- The WAE objective minimizes

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

- There are two sources of customization.
  1. Choose divergence between Q_Z and P_Z.

Wasserstein Auto-Encoder

- The WAE objective minimizes

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

- There are two sources of customization.
  1. Choose divergence between Q_Z and P_Z.

WAE-GAN	WAE-MMD
$\mathcal{D}_Z = D_{JS}$	$\mathcal{D}_Z = MMD_k$
Introduce adversarial discriminator in $\mathcal{Z}$	Use unbiased U-statistic estimator

Wasserstein Auto-Encoder

- The WAE objective minimizes

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

- There are two sources of customization.
  1. Choose divergence between Q_Z and P_Z.

WAE-GAN	WAE-MMD
$\mathcal{D}_Z = D_{JS}$	$\mathcal{D}_Z = MMD_k$
Introduce adversarial discriminator in $\mathcal{Z}$	Use unbiased U-statistic estimator

  2. Choose reconstruction cost function c.

Wasserstein Auto-Encoder

- The WAE objective minimizes

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

- There are two sources of customization.
  1. Choose divergence between Q_Z and P_Z.

WAE-GAN	WAE-MMD
$\mathcal{D}_Z = D_{JS}$	$\mathcal{D}_Z = MMD_k$
Introduce adversarial discriminator in $\mathcal{Z}$	Use unbiased U-statistic estimator

  2. Choose reconstruction cost function c.
     - If $c(x, x') = \|x - x'\|_2^2$, then WAE-GAN is equivalent to AAE

Wasserstein Auto-Encoder

- The WAE objective minimizes

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

- There are two sources of customization.
  1. Choose divergence between Q_Z and P_Z.

WAE-GAN	WAE-MMD
$\mathcal{D}_Z = D_{JS}$	$\mathcal{D}_Z = MMD_k$
Introduce adversarial discriminator in $\mathcal{Z}$	Use unbiased U-statistic estimator

  2. Choose reconstruction cost function c.
     - If $c(x, x') = \|x - x'\|_2^2$, then WAE-GAN is equivalent to AAE
     - Theoretical justification of AAE as minimizing 2-Wasserstein distance

Experiments

- Test VAE, WAE-GAN, WAE-MMD on MNIST and CelebA
- Record Frechet Inception Distance (FID) to assess quality of images

Algorithm	**FID**
VAE	82
WAE-MMD	55
WAE-GAN	42

Table 1: FID scores for samples on CelebA (smaller is better).

Extensions

ADVERSARIALLY REGULARIZED AUTOENCODERS

Junbo (Jake) Zhao[1], Yoon Kim[2], Kelly Zhang[1], Alexander M. Rush[2], Yann LeCun[1,3]
[1] Department of Computer Science, New York University
[2] School of Engineering and Applied Sciences, Harvard University
[3] Facebook AI Research
{jakezhao,kz918,yann}@cs.nyu.edu, {yoonkim,srush}@seas.harvard.edu

WASSERSTEIN AUTO-ENCODERS:
LATENT DIMENSIONALITY AND RANDOM ENCODERS

Paul Rubenstein,* Bernhard Schölkopf, Ilya Tolstikhin
Empirical Inference
Max Planck Institute for Intelligent Systems, Tübingen
{paul.rubenstein,bs,ilya}@tuebingen.mpg.de

LEARNING DISENTANGLED REPRESENTATIONS WITH
WASSERSTEIN AUTO-ENCODERS

Paul Rubenstein,* Bernhard Schölkopf, Ilya Tolstikhin
Empirical Inference
Max Planck Institute for Intelligent Systems, Tübingen
{paul.rubenstein,bs,ilya}@tuebingen.mpg.de

Our Project

- Build on work of Zhao et. al (2018) in applying WAEs to text

Our Project

- Build on work of Zhao et. al (2018) in applying WAEs to text
- Semi-Supervised Learning for transfering between different styles of English

Our Project

- Build on work of Zhao et. al (2018) in applying WAEs to text
- Semi-Supervised Learning for transfering between different styles of English
  - Shakespearean English vs. Modern English

ORIGINAL TEXT	MODERN TEXT
JULIET By and by, I come.— To cease thy strife and leave me to my grief. Tomorrow will I send.	**JULIET** Alright, I'm coming!—I beg you to stop trying for me and leave me to my sadness. Tomorrow I'll send the messenger.
ROMEO So thrive my soul—	**ROMEO** My soul depends on it—

Our Project

- Build on work of Zhao et. al (2018) in applying WAEs to text
- Semi-Supervised Learning for transfering between different styles of English
  - Shakespearean English vs. Modern English

ORIGINAL TEXT	MODERN TEXT
JULIET	**JULIET**
By and by, I come.— To cease thy strife and leave me to my grief. 155 Tomorrow will I send.	Alright, I'm coming!—I beg you to stop trying for me and leave me to my sadness. Tomorrow I'll send the messenger.
ROMEO	**ROMEO**
So thrive my soul—	My soul depends on it—

  - Formal English vs. Informal English

Informal:	*I'd say it is punk though.*
Formal:	*However, I do believe it to be punk.*
Informal:	*Gotta see both sides of the story.*
Formal:	*You have to consider both sides of the story.*