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

Alex Lin   Melissa Yu

Harvard University

April 5, 2018

Jensen-Shannon divergence

Minimize Jensen-Shannon divergence between true data distribution p_d and generative model p_g:

$$\min_G \mathcal{D}_{JS}(p_d(x)||p_g(x))$$

where

$$\mathcal{D}_{JS}(p_d(x)||p_g(x)) = \frac{1}{2}\mathcal{D}_{KL}(p_d||\frac{p_d + p_g}{2}) + \frac{1}{2}\mathcal{D}_{KL}(p_g||\frac{p_d + p_g}{2})$$

Generative Adversarial Networks (GANs)

Minimizing JSD corresponds to finding best G when D is optimal

$$\min_G \max_D V(D, G) = \mathbb{E}_{p_d(x)}[\log D(x)] + \mathbb{E}_{p_z(z)}[\log(1 - D(G(z)))]$$

Min-max game between 2 neural networks

- generator $G(z)$: prior $p_z(z)$ and likelihood $p(x \mid z)$
- discriminator $D(x)$: predicts probability x comes from p_d, not p_g

VAEs

Encoder $q(z \mid x)$, decoder $p(x \mid z)$. Prior on latent codes $p(z)$.

$$\begin{aligned} &\min_q \mathbb{E}_{q(z \mid x)}[-\log p(x \mid z)] + \mathcal{D}_{KL}(q(z \mid x) \mid\mid p(z)) \\ &= \text{Reconstruction} + \text{Regularization} \end{aligned}$$

Adversarial Auto-encoders (AAEs)

- Aggregated posterior

$$q(z) = \int_x q(z \mid x) p_d(x) dx$$

- Replaces VAE's $\mathcal{D}_{KL}(q(z \mid x) \,||\, p(z))$ regularizer with $\mathcal{D}_{JS}(q(z) \,||\, p(z))$

AAE: Adversarial Networks

- Directly minimizing $\mathcal{D}_{JS}(p(z) \mid\mid q(z))$ is intractable
- $\Rightarrow$ Attach adversarial network on top of z
  - Encoder of autoencoder is also generator of adversarial network

$$G(z) = q(z \mid x)$$

  - Discriminator D distinguishes $p(z)$ (positive samples) from $q(z)$ (negative samples)

AAE

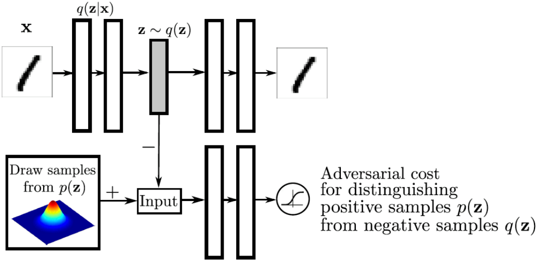


Figure 1: Architecture of an adversarial autoencoder. The top row is a standard autoencoder that reconstructs an image $\mathbf{x}$ from a latent code $\mathbf{z}$. The bottom row diagrams a second network trained to discriminatively predict whether a sample arises from the hidden code of the autoencoder or from a sampled distribution specified by the user.

AAE

Trained by alternately minimizing reconstruction error and training adversarial network

1. *Reconstruction phase.* Encoder q and decoder p are trained to minimize reconstruction error:

$$\min_{q,p} \; \mathbb{E}_{q(z \mid x)}[-\log p(x \mid z)] \tag{1}$$

2. *Regularization phase.* Adversarial network is trained on GAN objective:

$$\min_{q} \max_{D} \; \mathbb{E}_{p(z)}[\log D(z)] + \mathbb{E}_{q(z \mid x)}[\log(1 - D(z))]$$

Semi-supervised AAE

- *Semi-supervised phase.* The encoder for y, $q(y \mid x)$, is trained to minimize the cross entropy cost on labeled examples.

Choice of Encoder q

- Deterministic: q is deterministic function of x
  - may not produce smooth mapping; empirical distribution of data is fixed by training set
- Gaussian posterior: q is Gaussian distribution whose parameters are predicted by encoder network

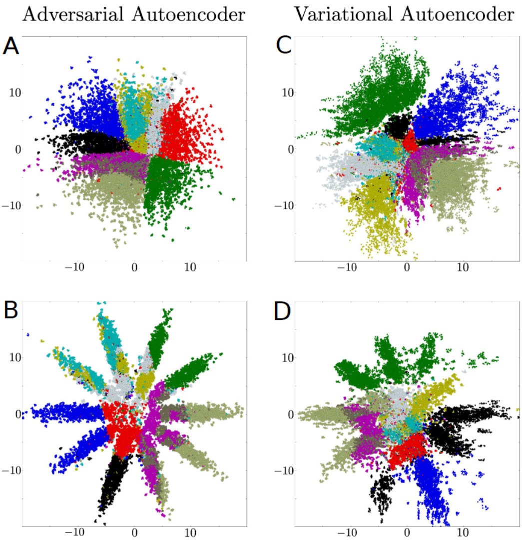
$$z \sim \mathcal{N}(\mu(x), \sigma(x))$$

- Universal approximator posterior

Encoder (cont.)

- Gaussian posterior + universal approximator posterior give network additional sources of stochasticity that could help it in the adversarial regularization stage by smoothing out $q(z)$.
- In practice, authors obtain similar test likelihoods for all 3 choices.

AAE vs. VAE



Adversarial Autoencoder
Variational Autoencoder
A
C
B
D
10
0
−10
−10
0
10

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 $x_1, \dots, x_n \in \mathcal{X}$ 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

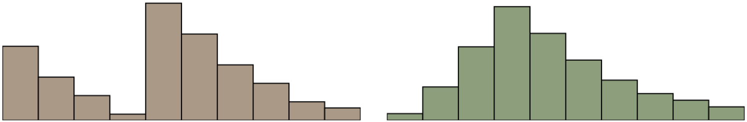
- Alternative way to measure distance between probability distributions
- Aka Earth Mover's Distance, Kantorovich-Rubinstein Metric, Optimal Transfer Plan

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

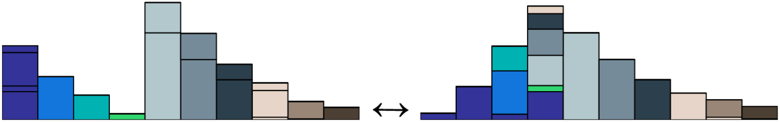
- The objective is

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

Using Wasserstein instead of KL

- The objective is

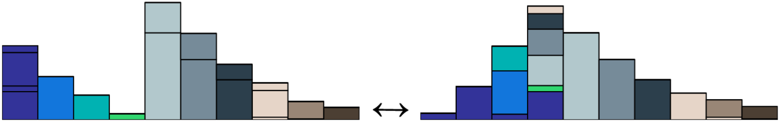
$$W_c(P_X, P_Y) = \inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_Y)} \mathbb{E}_{(X,Y)\sim\Gamma}[c(X, Y)]$$



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.

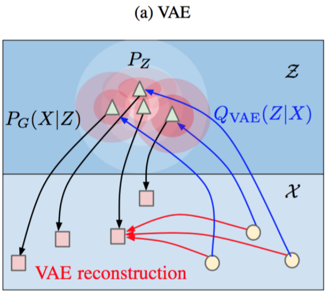
Wasserstein Auto-Encoder

- 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.

- 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 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—

  - 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.*