Coulomb Autoencoders

Emanuele Sansone, Hafiz Tiomoko Ali, Sun Jiacheng
Huawei Noah's Ark Lab (London)

HUAWEI

Contents

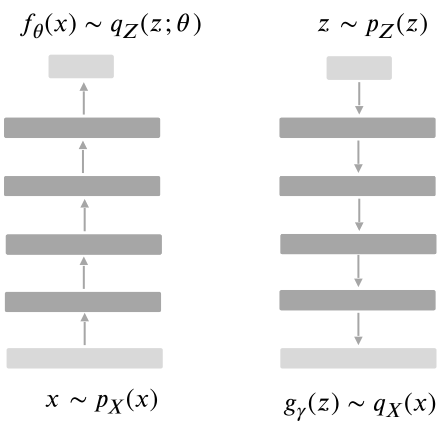
Motivation

BigGANs [Brock et al. 2019]

Improvement of deep generative models (GANs, Flow, Autoregressive, VAEs) in recent years
Lack of theoretical understanding:
1. Training (i.e. convergence guarantees to optimal solutions)
2. Generalization (i.e. quality of solutions with finite number of samples)

- Motivation
- **Background on Autoencoders**
- The Problem of Local Minima
- Generalization Analysis
- Conclusions

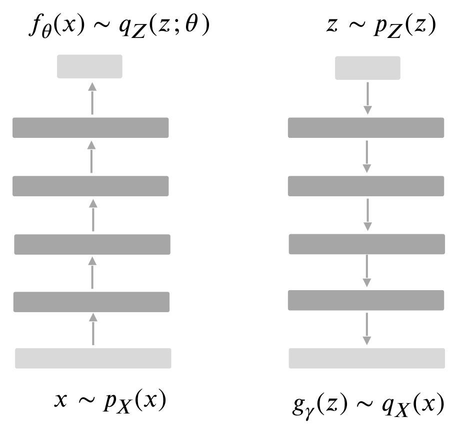
Background on Autoencoders - I



Goal

Implicitly learning the unknown density $p_X(x)$

Background on Autoencoders - I

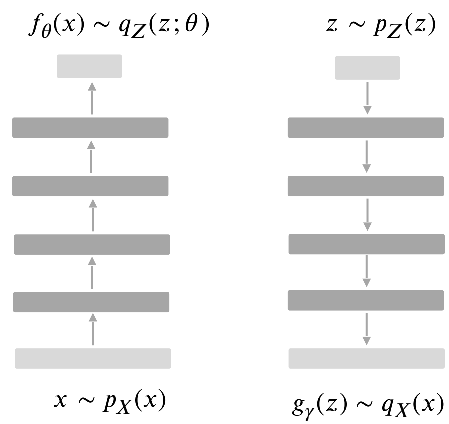


Goal
Implicitly learning the unknown density $p_X(x)$

Problem formulation
In order to ensure that $p_X(x) = q_X(x)$, we need:

1. Left-invertibility $x = g_\gamma(f_\theta(x))$ on the support of $p_X(x)$
2. Density matching $q_Z(z;\theta) = p_Z(z)$

Background on Autoencoders - I



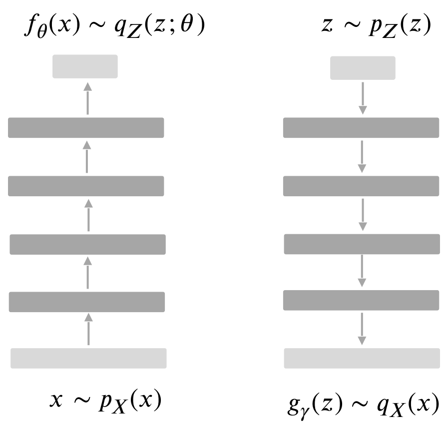
Goal
Implicitly learning the unknown density $p_X(x)$

Problem formulation
In order to ensure that $p_X(x) = q_X(x)$, we need:

1. Left-invertibility $x = g_\gamma(f_\theta(x))$ on the support of $p_X(x)$
2. Density matching $q_Z(z;\theta) = p_Z(z)$

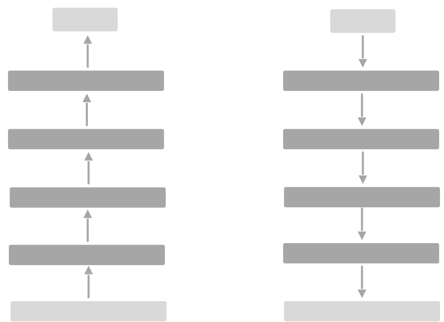
Objective: $\mathscr{L}(\theta, \gamma) = REC(g_\gamma \circ f_\theta) + \lambda D(q_Z, p_Z)$

Background on Autoencoders - II



$$\mathscr{L}(\theta,\gamma) = REC(g_\gamma \circ f_\theta) + \lambda D(q_Z, p_Z)$$

Background on Autoencoders - II



$$\mathcal{L}(\theta, \gamma) = REC(g_\gamma \circ f_\theta) + \lambda D(q_Z, p_Z)$$

Properties

1. $REC(g_\gamma \circ f_\theta)$ is typically the L2 loss, which is convex
2. $D(q_Z, p_Z)$ has many forms, all of them are non-convex
   - Kullback-Leibler Divergence (KL) in Variational Autoencoders
     [Kingma and Welling 2014]
     [Rezende et al. 2014]
   - Maximum-Mean Discrepancy (MMD) in Generative Moment Matching Networks
     [Li et al. 2015]
     Wasserstein (WAE)
     [Tolstikhin et al. 2018]
     Coulomb Autoencoders (CouAEs)

Background on Autoencoders - III

Why MMD should be preferred over KL?

1. KL term is not a proper metric, while MMD is an integral probability metric

Background on Autoencoders - III

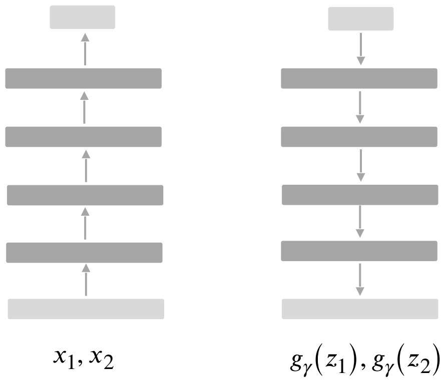
Why MMD should be preferred over KL?

1. KL term is not a proper metric, while MMD is an integral probability metric
2. KL is not always defined (e.g. empirical distributions, namely superposition of Dirac impulses)

Background on Autoencoders - III

Why MMD should be preferred over KL?

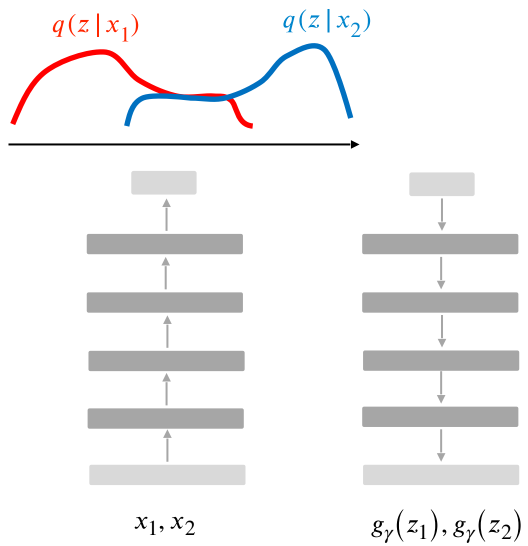
1. KL term is not a proper metric, while MMD is an integral probability metric
2. KL is not always defined (e.g. empirical distributions, namely superposition of Dirac impulses)
3. Local minima (problem of reconstruction/posterior collapse, due to stochastic encoder)



Background on Autoencoders - III

Why MMD should be preferred over KL?

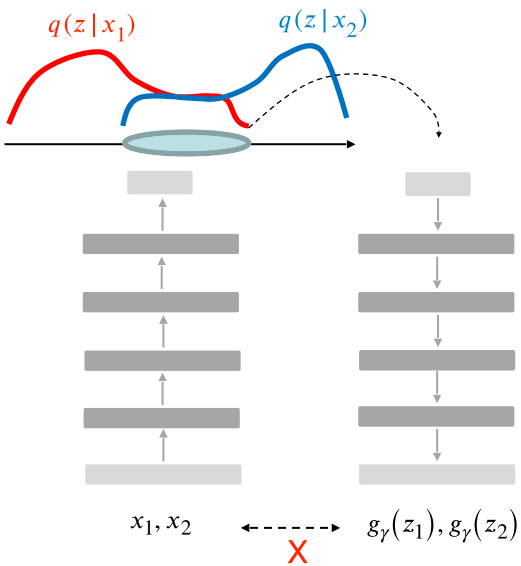
1. KL term is not a proper metric, while MMD is an integral probability metric
2. KL is not always defined (e.g. empirical distributions, namely superposition of Dirac impulses)
3. Local minima (problem of reconstruction/posterior collapse, due to stochastic encoder)



Background on Autoencoders - III

Why MMD should be preferred over KL?

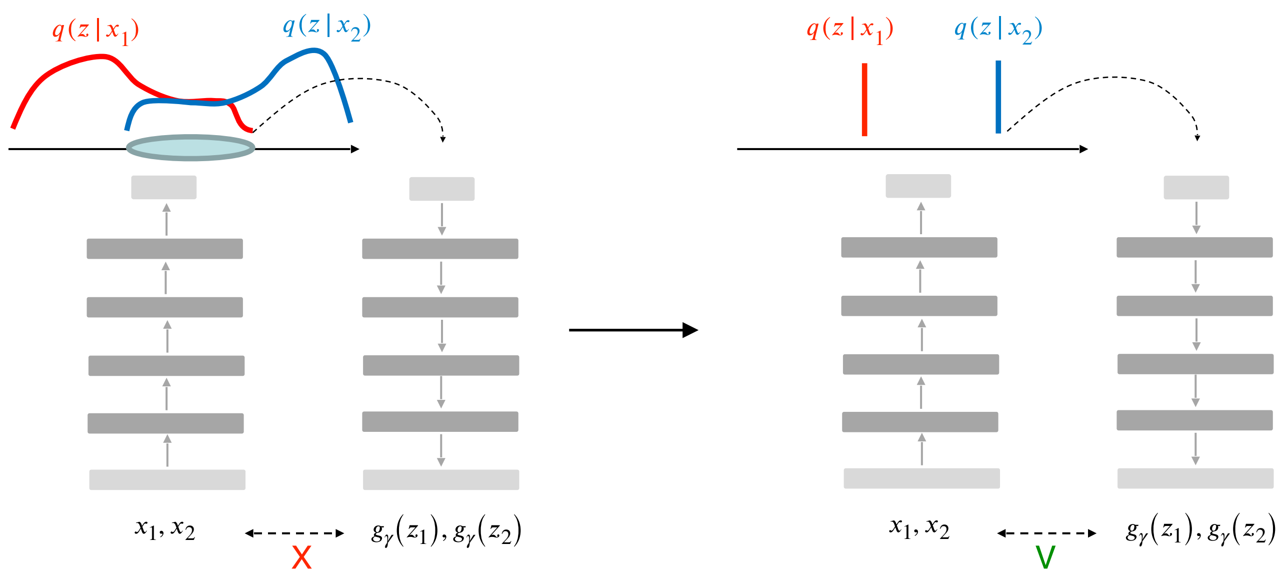
1. KL term is not a proper metric, while MMD is an integral probability metric
2. KL is not always defined (e.g. empirical distributions, namely superposition of Dirac impulses)
3. Local minima (problem of reconstruction/posterior collapse, due to stochastic encoder)



Background on Autoencoders - III

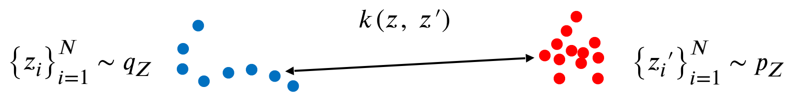
Why MMD should be preferred over KL?

1. KL term is not a proper metric, while MMD is an integral probability metric
2. KL is not always defined (e.g. empirical distributions, namely superposition of Dirac impulses)
3. Local minima (problem of reconstruction/posterior collapse, due to stochastic encoder)



- Motivation
- Background on Autoencoders
- **The Problem of Local Minima**
- Generalization Analysis
- Conclusions

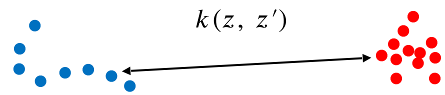
Properties of MMD



$$\mathrm{M}MD\left(\{z_i\}_{i=1}^{N}, \{z_i'\}_{i=1}^{N}\right) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} k(z_i', z_j') + \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} k(z_i, z_j) - \frac{2}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} k(z_i', z_j)$$

Properties of MMD

$\{z_i\}_{i=1}^N \sim q_Z$



$\{z_i'\}_{i=1}^N \sim p_Z$

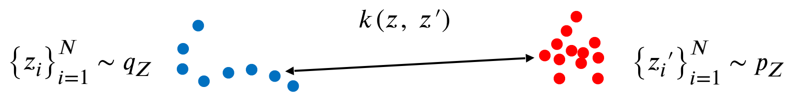
$$MMD\left(\{z_i\}_{i=1}^N, \{z_i'\}_{i=1}^N\right) = \boxed{\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} k(z_i', z_j')} + \boxed{\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} k(z_i, z_j)} - \boxed{\frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k(z_i', z_j)}$$

Intra-similarity | Intra-similarity | Inter-similarity

Minimization of MMD wrt $\{z_i\}_{i=1}^N \approx$ maximization of inter-similarity and minimization of intra-similarities

Used for density matching or two sample test [Gretton et al. 2012], recently used in autoencoders [Tolstikhin et al. 2018]

Properties of MMD



$$MMD\left(\{z_i\}_{i=1}^N, \{z_i'\}_{i=1}^N\right) = \underbrace{\frac{1}{N(N-1)}\sum_{i=1}^N\sum_{j\neq i} k(z_i', z_j')}_{\text{Intra-similarity}} + \underbrace{\frac{1}{N(N-1)}\sum_{i=1}^N\sum_{j\neq i} k(z_i, z_j)}_{\text{Intra-similarity}} - \underbrace{\frac{2}{N^2}\sum_{i=1}^N\sum_{j=1}^N k(z_i', z_j)}_{\text{Inter-similarity}}$$

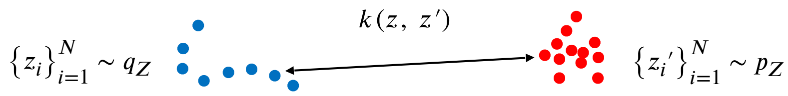
Minimization of MMD wrt $\{z_i\}_{i=1}^N \approx$ maximization of inter-similarity and minimization of intra-similarities

Used for density matching or two sample test [Gretton et al. 2012], recently used in autoencoders [Tolstikhin et al. 2018]

⬇

The choice of kernel function is related with the problem of local minima

Properties of MMD and Coulomb kernel

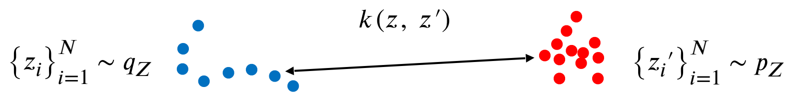


$$MMD\left(\{z_i\}_{i=1}^{N}, \{z_i'\}_{i=1}^{N}\right) = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} k(z_i', z_j') + \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} k(z_i, z_j) - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k(z_i', z_j)$$

Intra-similarity | Intra-similarity | Inter-similarity

Coulomb kernel $k(z, z') = \dfrac{1}{\left\|z - z'\right\|^{h-2}}$ $\quad N > h > 2$

Properties of MMD and Coulomb kernel



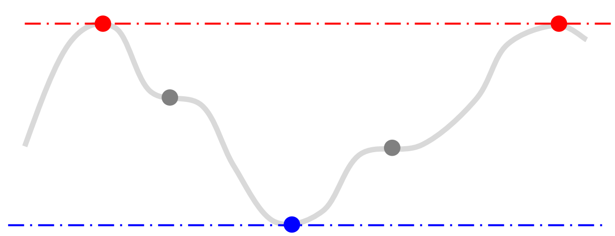
$$MMD\left(\{z_i\}_{i=1}^N, \{z_i'\}_{i=1}^N\right) = \underbrace{\frac{1}{N(N-1)}\sum_{i=1}^N \sum_{j\neq i} k(z_i', z_j')}_{\text{Intra-similarity}} + \underbrace{\frac{1}{N(N-1)}\sum_{i=1}^N \sum_{j\neq i} k(z_i, z_j)}_{\text{Intra-similarity}} - \underbrace{\frac{2}{N^2}\sum_{i=1}^N \sum_{j=1}^N k(z_i', z_j)}_{\text{Inter-similarity}}$$

Coulomb kernel $k(z, z') = \dfrac{1}{\|z - z'\|^{h-2}}$ $\quad N > h > 2$

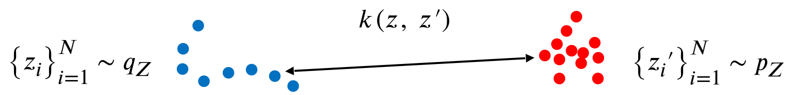
Theorem

Minimization of MMD wrt $\{z_i\}_{i=1}^N$

1. All local extrema are global
2. The set of saddle points has measure zero



Properties of MMD and Coulomb kernel



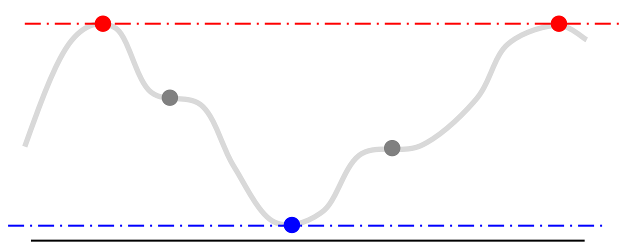
$$MMD\left(\{z_i\}_{i=1}^{N}, \{z_i'\}_{i=1}^{N}\right) = \underbrace{\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} k(z_i', z_j')}_{\text{Intra-similarity}} + \underbrace{\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} k(z_i, z_j)}_{\text{Intra-similarity}} - \underbrace{\frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k(z_i', z_j)}_{\text{Inter-similarity}}$$

Coulomb kernel $k(z, z') = \dfrac{1}{\|z - z'\|^{h-2}}$ $\quad N > h > 2$

Theorem

Minimization of MMD wrt $\{z_i\}_{i=1}^{N}$

1. All local extrema are global
2. The set of saddle points has measure zero

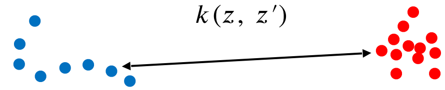


Remark:

Convergence to global minimum when optimized through local-search methods!

Properties of MMD and Coulomb kernel

$\{z_i\}_{i=1}^{N} \sim q_Z$ $\quad k(z, z')$ $\quad \{z_i'\}_{i=1}^{N} \sim p_Z$



$$MMD\left(\{z_i\}_{i=1}^{N}, \{z_i'\}_{i=1}^{N}\right) = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} k(z_i', z_j') + \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} k(z_i, z_j) - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k(z_i', z_j)$$

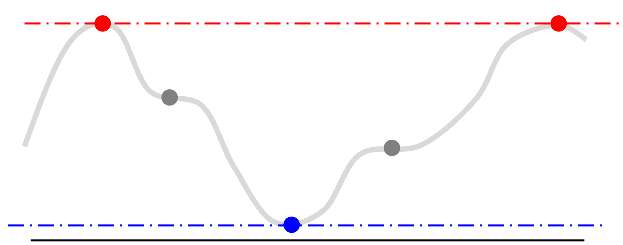
Intra-similarity | Intra-similarity | Inter-similarity

Coulomb kernel $k(z, z') = \frac{1}{\|z - z'\|^{h-2}}$ $\quad N > h > 2$

Theorem

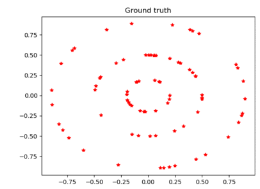
Minimization of MMD wrt $\{z_i\}_{i=1}^{N}$

1. All local extrema are global
2. The set of saddle points has measure zero

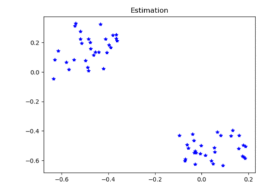


Remark:

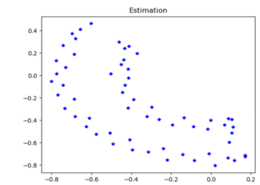
Convergence to global minimum when optimized through local-search methods!



Ground truth



Gaussian kernel



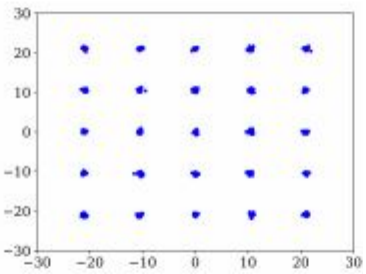
Coulomb kernel

Experiments

Eval. Metric	Data/Method	VAE	WAE	CouAE
Test Log-likel.	Grid			

Experiments

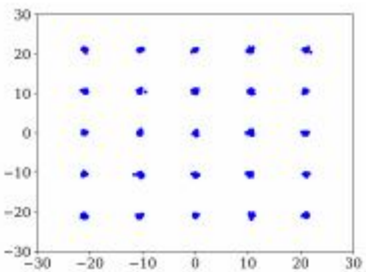
Eval. Metric	Data/Method	VAE	WAE	CouAE
Test Log-likel.	Grid			



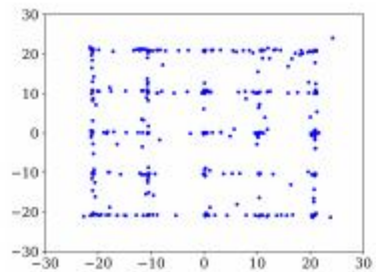
Ground truth

Experiments

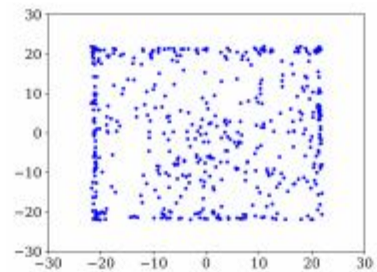
Eval. Metric	Data/Method	VAE	WAE	CouAE
Test Log-likel.	Grid	-4.4±0.2	-6.4±1.1	**-4.3±0.1**



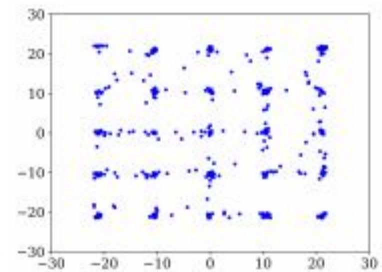
Ground truth



VAE (KL)



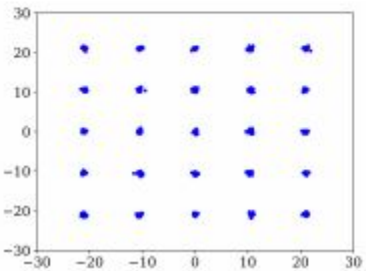
WAE (MMD)



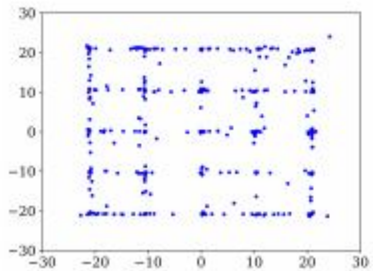
CouAE (MMD + Coulomb)

Experiments

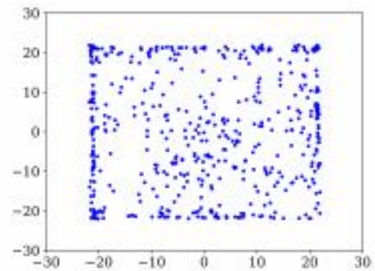
Eval. Metric	Data/Method	VAE	WAE	CouAE
Test Log-likel.	Grid	-4.4±0.2	-6.4±1.1	**-4.3±0.1**
FID	CelebA			



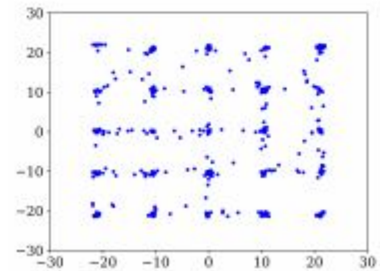
Ground truth



VAE (KL)



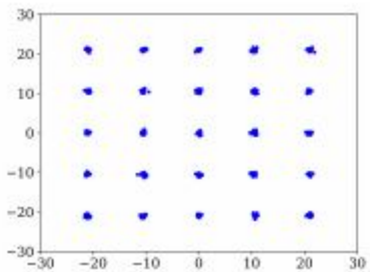
WAE (MMD)



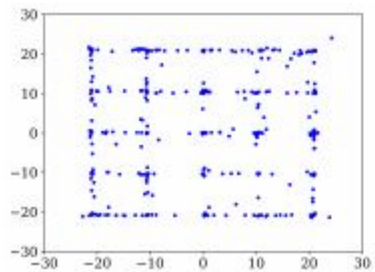
CouAE (MMD + Coulomb)

Experiments

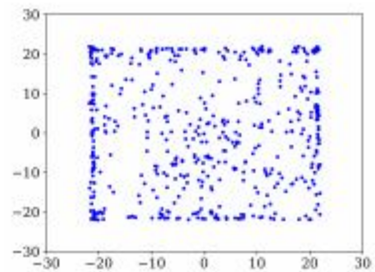
Eval. Metric	Data/Method	VAE	WAE	CouAE
Test Log-likel.	Grid	-4.4±0.2	-6.4±1.1	**-4.3±0.1**
FID	CelebA	63	55	**47**



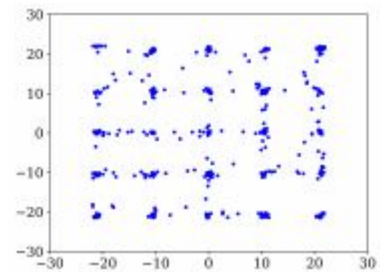
Ground truth



VAE (KL)

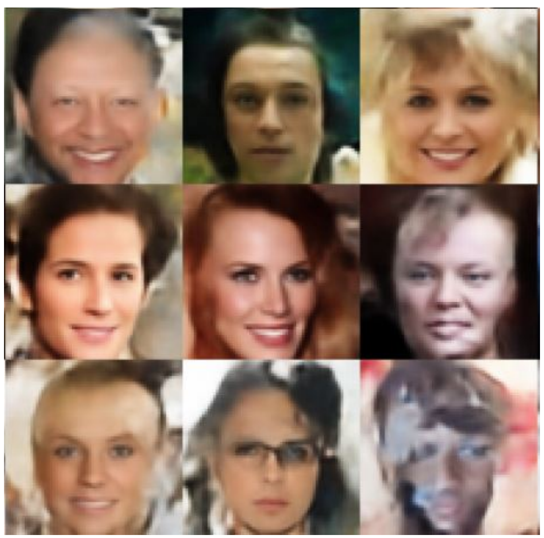


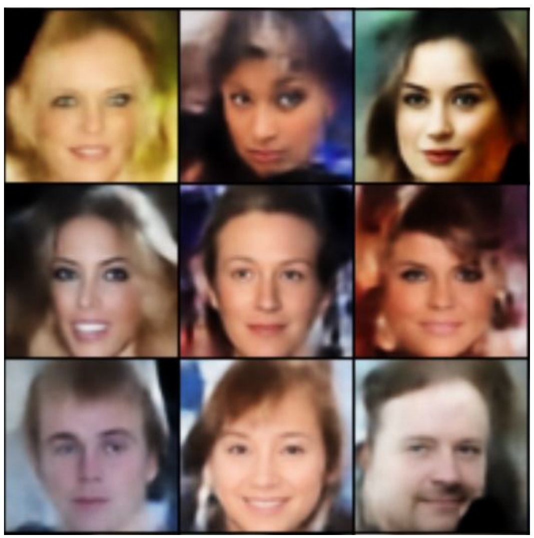
WAE (MMD)



CouAE (MMD + Coulomb)


VAE (KL)


WAE (MMD)

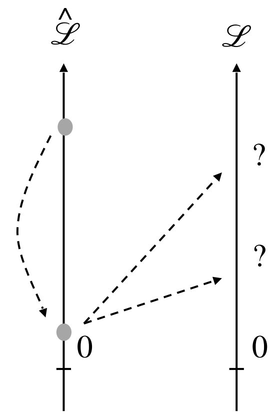

CouAE (MMD + Coulomb)

- Motivation
- Background on Autoencoders
- The Problem of Local Minima
- **Generalization Analysis**
- Conclusions

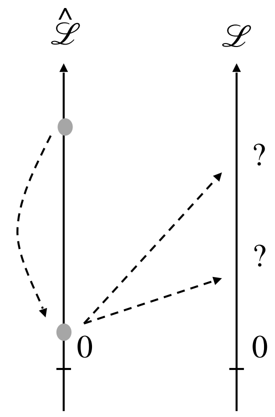
Generalization Analysis

$\hat{\mathscr{L}} = R\hat{E}C + \lambda M\hat{M}D$ (finite number of samples)

$\mathscr{L} = REC + \lambda MMD$ (infinite number of samples)



Generalization Analysis



$\hat{\mathscr{L}} = R\hat{E}C + \lambda M\hat{M}D$ (finite number of samples)
$\mathscr{L} = REC + \lambda MMD$ (infinite number of samples)

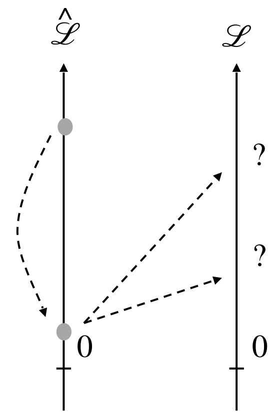
Theorem

$0 \le k(z, z') = 1/(\|z - z'\|^{h-2} + \epsilon) \le K$
$0 \le REC \le \xi$

For any $s,\ t > 0$

$$\Pr\left\{ \left| \hat{\mathscr{L}} - \mathscr{L} \right| > t + \lambda s \right\} \le 2exp\left\{ -\frac{2Nt^2}{\xi^2} \right\} + 6exp\left\{ -\frac{2\lfloor N/2 \rfloor s^2}{9K^2} \right\}$$

Generalization Analysis



$\hat{\mathscr{L}} = R\hat{E}C + \lambda M\hat{M}D$ (finite number of samples)
$\mathscr{L} = REC + \lambda MMD$ (infinite number of samples)

Theorem

$0 \leq k(z, z') = 1/(\|z - z'\|^{h-2} + \epsilon) \leq K$
$0 \leq REC \leq \xi$

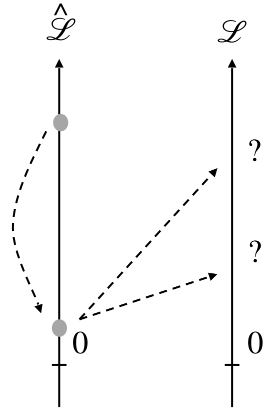
For any $s,\ t > 0$

Contribution of recon. error

Contribution of MMD

$$\Pr\left\{ \left| \hat{\mathscr{L}} - \mathscr{L} \right| > t + \lambda s \right\} \leq 2exp\left\{ -\frac{2Nt^2}{\xi^2} \right\} + 6exp\left\{ -\frac{2\lfloor N/2 \rfloor s^2}{9K^2} \right\}$$

Generalization Analysis



$\hat{\mathscr{L}} = R\hat{E}C + \lambda M\hat{M}D$ (finite number of samples)
$\mathscr{L} = REC + \lambda MMD$ (infinite number of samples)

Theorem

$0 \leq k(z, z') = 1/(\left\| z - z' \right\|^{h-2} + \epsilon) \leq K$
$0 \leq REC \leq \xi$

For any $s,\ t > 0$

Contribution of recon. error

Contribution of MMD

$$\Pr\left\{ \left| \hat{\mathscr{L}} - \mathscr{L} \right| > t + \lambda s \right\} \leq 2exp\left\{ -\frac{2Nt^2}{\xi^2} \right\} + 6exp\left\{ -\frac{2\lfloor N/2 \rfloor s^2}{9K^2} \right\}$$

How can we make ξ small?

1. Estimation of ξ -> maximum reconstruction error on both training and validation data
2. Minimization of ξ -> Finding proper network architecture (e.g. layer width, networks' depth, residual connections)

Experiments

Controlling ξ by changing total number of hidden neurons (capacity)

Experiments

Controlling ξ by changing total number of hidden neurons (capacity)

Eval. Metric	**Data/Width factor**	$\times 0.25$	$\times 0.5$	$\times 1$
Test Log-likel.	Grid	-5.8±0.4	-4.8±0.4	-4.3±0.1
FID	CelebA	53	51	47

Experiments

Controlling ξ by changing total number of hidden neurons (capacity)

Eval. Metric	Data/Width factor	×0.25	×0.5	×1
Test Log-likel.	Grid	-5.8±0.4	-4.8±0.4	-4.3±0.1
FID	CelebA	53	51	47

x0.25

x0.5

x1

Experiments

Controlling ξ by changing total number of hidden neurons (capacity)

Eval. Metric	**Data/Width factor**	×0.25	×0.5	×1
Test Log-likel.	Grid	-5.8±0.4	-4.8±0.4	-4.3±0.1
FID	CelebA	53	51	47

x0.25

x0.5

x1

Remarks

1. Network architecture is fundamental to control generalization
2. Increasing capacity (the number of hidden neurons) leads to better generalization (as long as ξ is decreased)
3. Other architectural choices (e.g. depth, residual connections) may further decrease ξ

Experiments

Controlling ξ by changing total number of hidden neurons (capacity)

Eval. Metric	Data/Width factor	×0.25	×0.5	×1
Test Log-likel.	Grid	-5.8±0.4	-4.8±0.4	-4.3±0.1
FID	CelebA	53	51	47

x0.25

x0.5

x1

Remarks

1. Network architecture is fundamental to control generalization
2. Increasing capacity (the number of hidden neurons) leads to better generalization (as long as ξ is decreased)
3. Other architectural choices (e.g. depth, residual connections) may further decrease ξ

Open Question

What is/are the optimal network architecture/s minimizing ξ?

- Motivation
- Background on Autoencoders
- The Problem of Local Minima
- Generalization Analysis
- **Conclusions**

Conclusions

1. Problem of local minima, MMD + Coulomb kernel behaves similarly to a convex functional
2. Generalization analysis, probabilistic bound giving insights on possible directions to improve autoencoder in principled manner

Thank You