Convergence, generalisation and privacy

in generative adversarial networks

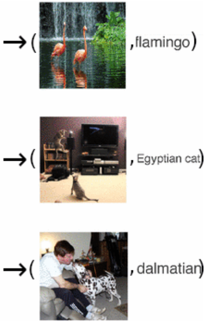
Olof Mogren, PhD. RISE Research institutes of Sweden

Discriminative modelling

- Model conditional distribution $P(Y|X)$

Generative modelling

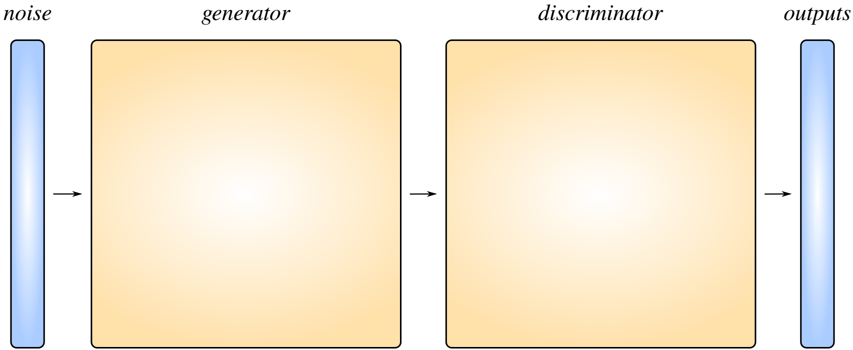
- Model joint distribution $P(X, Y)$



Generative modelling of fashion segmentation

*M. Korneliusson, J. Martinsson, **O. Mogren***

Generative adversarial networks (GANs)



Goodfellow, et.al., 2014

GAN properties

- Generate realistic images
- Discriminate between generated and real images
- Training: min-max game
- $\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{real}} [\log D_{\theta_D}(\boldsymbol{x})] + \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{G_{\theta_G}}} [\log(1 - D_{\theta_D}(\boldsymbol{x}))]$
- No expensive normalizing constant

GAN properties

- Generate realistic images
- Discriminate between generated and real images
- Training: min-max game
- $\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{real}} [\log D_{\theta_D}(\boldsymbol{x})] + \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{G_{\theta_G}}} [\log(1 - D_{\theta_D}(\boldsymbol{x}))]$
- No expensive normalizing constant

GAN properties

- Generate realistic images
- Discriminate between generated and real images
- Training: min-max game
- $\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{real}} \left[\log D_{\theta_D}(\boldsymbol{x})\right] + \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{G_{\theta_G}}} \left[\log(1 - D_{\theta_D}(\boldsymbol{x}))\right]$
- No expensive normalizing constant

GAN properties

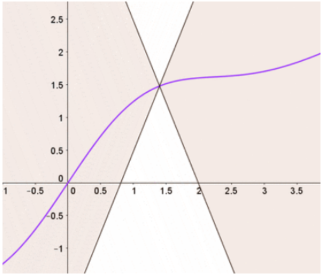
- Generate realistic images
- Discriminate between generated and real images
- Training: min-max game
- $\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{real}} [\log D_{\theta_D}(\boldsymbol{x})] + \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{G_{\theta_G}}} [\log(1 - D_{\theta_D}(\boldsymbol{x}))]$
- No expensive normalizing constant

GAN properties

- Generate realistic images
- Discriminate between generated and real images
- Training: min-max game
- $\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{real}} [\log D_{\theta_D}(\boldsymbol{x})] + \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{G_{\theta_G}}} [\log(1 - D_{\theta_D}(\boldsymbol{x}))]$
- No expensive normalizing constant

RI. SE

Regularization

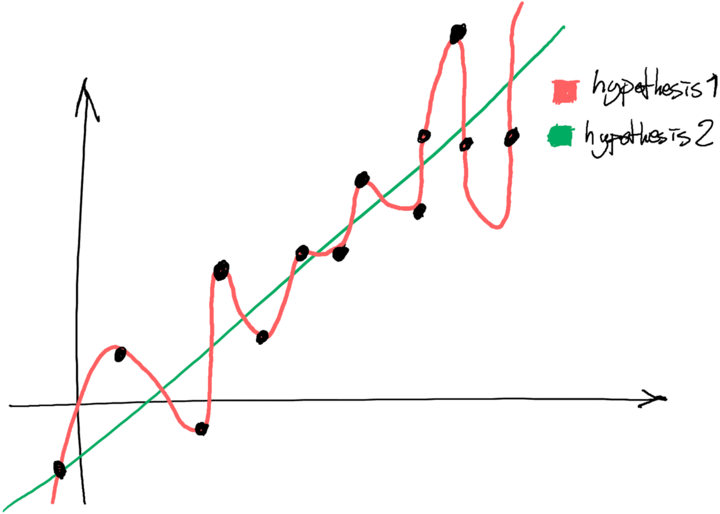


- Lipschitz continuity: gradient bound
- Loss-sensitive GAN: loss that restricts D to satisfy Lipschitz condition (Qi, 2017)
- Spectral normalization: regularization on the weight parameters (Miyato, ICLR 2018)
- Wasserstein GAN: penalty constrains the magnitude of the gradient (Arjovsky, 2017

Sufficiently large discriminator

- Capacity of D, and data: large enough
- If G "wins", then the generated distribution D is close to D_{real}
- But "large enough" could mean $\exp(d)$!

Generalization; intuition



Learning objectives

- Supervised learning: minimize loss
- GAN: find Nash equilibrium

Definition of generalization

- $\hat{\mathcal{D}}_{\text{real}}$ - empirical version, m samples
- $\mathcal{D}_G$ *generalizes* if with high probability:

$$|d(\mathcal{D}_{\text{real}}, \mathcal{D}_G) - d(\hat{\mathcal{D}}_{\text{real}}, \hat{\mathcal{D}}_G)| \leq \epsilon$$

- $\hat{\mathcal{D}}_G$ - empirical version of $\hat{\mathcal{D}}_G$, polynomial number of samples
- $d(\cdot, \cdot)$ - divergence or distance
- ϵ - generalization error.

Arora, et.al., ICML 2017

Neural net distance

- Jensen-Shannon divergence and Wasserstein distance **don’t generalize**
- A weaker distance, the Neural net distance **does**

(Details)

MIX+GAN

- A mixture of generators achieves provable approximate pure equilibria
- Experiments show that this can also help in practice

Arora, et.al., ICML 2017

RI.
SE

MIX+DCGAN

DCGAN

Method	Score
SteinGAN [Wang and Liu, 2016]	6.35
Improved GAN [Salimans et al., 2016]	8.09±0.07
AC-GAN [Odena et al., 2016]	8.25 ± 0.07
S-GAN (best variant in [Huang et al., 2017])	8.59± 0.12
DCGAN (as reported in Wang and Liu [2016])	6.58
DCGAN (best variant in Huang et al. [2017])	7.16±0.10
DCGAN (5x size)	7.34±0.07
MIX+DCGAN (Ours, with 5 components)	7.72±0.09
Wasserstein GAN	3.82±0.06
MIX+WassersteinGAN (Ours, with 5 components)	4.04±0.07
Real data	11.24±0.12

Arora, et.al., ICML 2017

Differential privacy

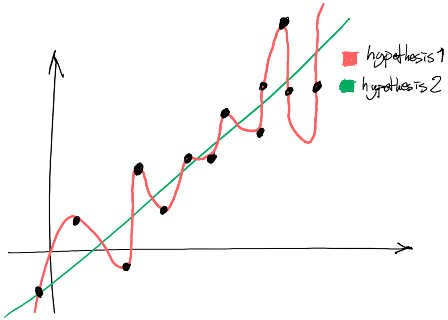
A randomized algorithm $\mathcal{A} : D \rightarrow R$ satisfies ϵ-differential privacy if for any two adjacent datasets $\mathcal{S}, \mathcal{S}' \subseteq D$ and for any subset of outputs $O \subseteq R$ it holds:

$$P[\mathcal{A}(\mathcal{S} \in O)] \leq e^{\epsilon} P[\mathcal{A}(\mathcal{S}') \in O]$$

Wu, et.al., NeurIPS 2019

Generalization/privacy

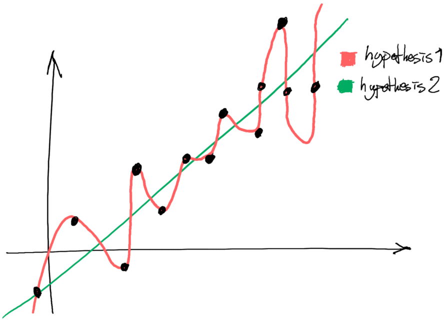
- Common goal: learn the population features
- Membership attacks



Wu, et.al., NeurIPS 2019

Generalization/privacy

- Common goal: learn the population features
- Membership attacks



Wu, et.al., NeurIPS 2019

Generalization/privacy

- Differential privacy $\rightarrow$ RO-stability*
- RO-stability $\rightarrow$ Generalization

Theorem 1 (Generalization gap) *If an algorithm $\mathcal{A}$ satisfies ϵ-differential privacy, then the generalization gap can be bounded by a data-independent constant.*

* ***R**eplace **O**ne element in the inputl; Wu, et.al., NeurIPS 2019*

Regularization and privacy

- Lipschitz condition crucial for privacy

Wu, et.al., NeurIPS 2019

Experimental validation

- Membership attack
- GAN information leakage

Attacker, α

- Given the discriminator D_{θ_D} and an image from the attack testing dataset
- α sets a threshold $t \in (0, 1)$
- α outputs 1 if $D_{\theta_D}/b \geq t$, otherwise, it outputs 0.

Wu, et.al., NeurIPS 2019

Experimental validation

Table 1: Evaluation results of DCGAN trained with different strategies. IS denotes the Inception score. N/A indicates that the strategy leads to failure/collapse of the training. The last row presents the Inception scores of the real data (training images of these two datasets).

Strategy	LFW				IDC			
	F1	AUC	Gap	IS	F1	AUC	Gap	IS
-JS divergence-								
Original	0.565	0.729	0.581	3.067	0.445	0.531	0.138	2.148
Weight Clipping	0.486	0.501	0.113	3.112	0.378	0.502	0.053	2.083
Spectral Normalization	0.482	0.506	0.106	3.104	0.416	0.508	0.124	2.207
Gradient Penalty	N/A				N/A			
-Wasserstein-								
W/o clipping	N/A				N/A			
Weight Clipping	0.484	0.512	0.042	3.013	0.388	0.513	0.045	1.912
Spectral Normalization	0.515	0.505	0.017	3.156	0.415	0.507	0.013	2.196
Gradient Penalty	0.492	0.503	0.031	2.994	0.426	0.504	0.017	1.974
IS (Real data)	4.272				3.061			

Wu, et.al., NeurIPS 2019

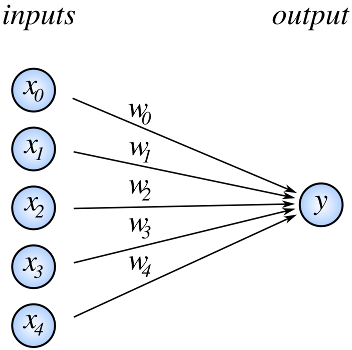
Thank you.

Appendix

RI.
SE

Byggstenarna i deep learning

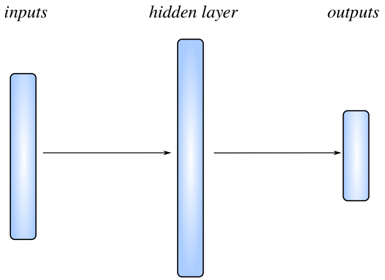
- Varje lager innehåller ett antal enheter/neuroner
- *Löst* inspirerade av biologiska neuroner
- Ett djupt nät kan innehålla miljontals enheter
- $w_1, \ldots, w_n$ inlärda parametrar



(Tillbaka)

Lager i djupa neuronnät

- I praktiken arrangeras neuronerna i lager
- Varje lager:
  - linjär transformation av input-vektorn
  - icke-linjär aktiveringsfunktion



(Tillbaka)

Neural net distance

- Jensen-Shannon divergence and Wasserstein distance **don't generalize**
- A weaker distance, the Neural net distance **does**

$$d_{\mathcal{F},\phi}(\mu,\nu) = \sup_{D\in\mathcal{F}} \mathbb{E}_{\mathbf{x}\sim\mu}\left[\phi D_{\theta_D}(\boldsymbol{x})\right] + \mathbb{E}_{\mathbf{x}\sim\nu}\left[\phi(1 - D_{\theta_D}(\boldsymbol{x}))\right] - 2\phi(1/2)$$

(Back)

RO-stability

Define 2 ***(Uniform RO-stability)*** *The randomized algorithm $\mathcal{A}$ is uniform RO-stable with respect to the discriminator loss function (Equation 2) in our case, if for all adjacent datasets S, S', it holds that:*

$$\sup_{x \in S} |\mathbb{E}_{\theta_d \sim \mathcal{A}(S)}[\phi(\mathbf{d}(x;\theta_d))] - \mathbb{E}_{\theta_d \sim \mathcal{A}(S')}[\phi(\mathbf{d}(x;\theta_d))]| \leq \epsilon_{stable}(m) \tag{6}$$

A well-known heuristic observation is that differential privacy implies uniform stability. The prior work [35] has formlized this observation into the following lemma:

Lemma 1 ***(Differential privacy $\Rightarrow$ uniform RO-stability)*** *If a randomized algorithm $\mathcal{A}$ is ϵ-differentially private, then the algorithm $\mathcal{A}$ satisfies $(e^\epsilon - 1)$-RO-stability.*

The stability of the algorithm is also related to the generalization gap. Numerous studies [30, 23] focus on exploring the relationship in various settings. Formally, we have the following lemma:

Lemma 2 *If an algorithm $\mathcal{A}$ is uniform RO-stable with rate $\epsilon_{stable}(m)$, then $|F_U(\mathcal{A})|$ (Equation 4) can be bounded: $|F_U(\mathcal{A})| \leq \epsilon_{stable}(m)$.*

*Replace **O**ne element in the input; (Back)*

Generalization gap (Wu, et.al., NeurIPS 2019)

$$F_U(\mathcal{A}_d) = \mathbb{E}_{\theta_d \sim \mathcal{A}_d(S)} \mathbb{E}_{S \sim p_{data}^m} [\hat{U}(\theta_d, \theta_g^*) - U(\theta_d, \theta_g^*)]$$

(Back)