Research on the stability of generative adversarial network

Zubkov Maxim, Filatov Andrey

Moscow Institute of Physics and Technology (MIPT), Moscow, Russia

20 August 2020

Introduction

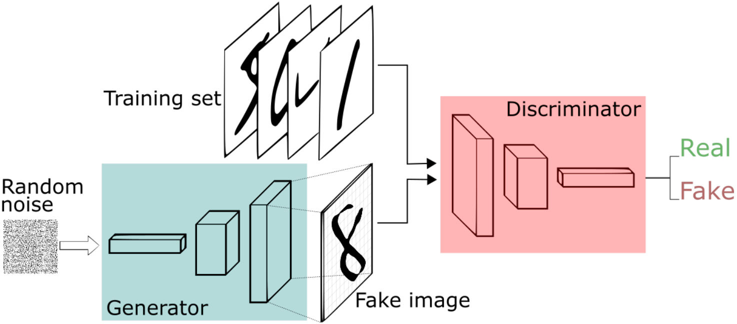


Figure 1: GAN structure

Image credit: jrmerwin.github.io

Training issues

- Vanishing gradinets
- Discriminator's overfitting
- Hyperparameter sensitivity
- Mode collapse

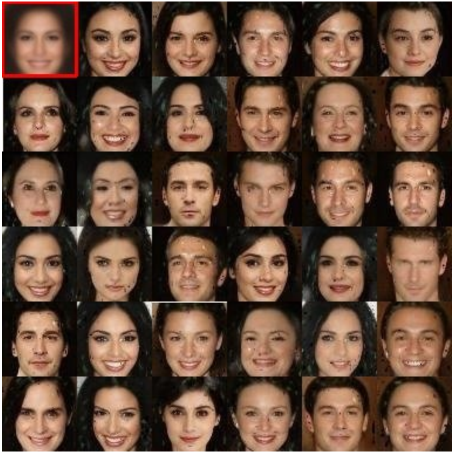


Image credit: `https://arxiv.org/pdf/1805.12462`

Problem statement

Let Φ_w be a discriminator and μ_θ be a distribution produced by generator
In common case min-max problem is to solved :

$$\min_{\mu_\theta} J(\mu_0, \mu_\theta) = \min_{\mu_\theta} \max_{\Phi_w} \Psi(\mu_\theta, \Phi_w)$$

Problem statement

Let Φ_w be a discriminator and μ_θ be a distribution produced by generator

In common case min-max problem is to solved :

$$\min_{\mu_\theta} J(\mu_0, \mu_\theta) = \min_{\mu_\theta} \max_{\Phi_w} \Psi(\mu_\theta, \Phi_w)$$

Wasserstein-GAN metric:

$$J(\mu_0, \mu_\theta) = \max_{\Phi_w} \Psi(\mu_\theta, \Phi_w) = \max_{\|\Phi_w\|_{Lip} \leq 1} \mathbb{E}_{x \sim \mu_0}[\Phi_w(x)] - \mathbb{E}_{x \sim \mu_\theta}[\Phi_w(x)]$$

Theoretical insights

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ is a L-smooth and has a lower bound. Let $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$. Then $\|\nabla f(x_k)\| \to 0$ at $k \to \infty$.

Theorem (Chu et al.)

Let $J : \mu_\theta \to \mathbb{R}$ be a convex function. Fix $\mu := \mu_\theta$ and consider the optimal discriminator: $\Phi_\mu : Y \to [0, 1]$. Let it satisfies regularity conditions. Then $\theta \mapsto J(\mu_\theta)$ is L-smooth.

Regularity conditions

(D1) $x \mapsto \Phi_\mu(x)$ – α-Lipschitz,

(D2) $x \mapsto \nabla_x \Phi_\mu(x)$ – β_1-Lipschitz,

(D3) $\mu \mapsto \nabla_x \Phi_\mu(x)$ – β_2-Lipschitz w.r.t 1-Wasserstein distance.

Also, let $f_\theta(\omega)$ be a family of generators that satisfies:

(G1) $\theta \mapsto f_\theta(z)$ A-Lipschitz in expectation for $z \sim \omega$, i.e,
$\mathbb{E}_{z\sim\omega}[\|f_{\theta_1}(z) - f_{\theta_2}(z)\|_2] \leq A\|\theta_1 - \theta_2\|_2$,

(G2) $\theta \mapsto D_\theta f_\theta(z)$ B-Lipschitz in expectation for $z \sim \omega$, i.e.,
$\mathbb{E}_{z\sim\omega}[\|D_{\theta_1} f_{\theta_1}(z) - D_{\theta_2} f_{\theta_2}(z)\|_2] \leq B\|\theta_1 - \theta_2\|_2$.

And smooth constant $L = \alpha B + A^2(\beta_1 + \beta_2)$.

Methods

Aim	Methods
Discriminator optimality	Vary number of dicriminator iterations
Lipschitz discriminator	Spectral Norm Adversarial Lipschitz Regularization
Lipschitz gradient of discriminator	Smooth activation function, Spectral Norm

Figure 2: Methods

Algorithm 1

while θ has not converged **do**
  for $t = 0, ..., n_{cr}$ **do**
    Sample $\{\mathbf{x}^{(i)}\}_{i=1}^{m} \sim \mu_0$ a batch from the real data.
    Sample $\{\mathbf{z}^{(i)}\}_{i=1}^{m} \sim \mu_\theta$ a batch of prior samples.
    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} \Phi_w(\mathbf{x}^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} \Phi_w(f_\theta(\mathbf{z}^{(i)}))\right]$
    $w \leftarrow w + \alpha \cdot \mathrm{Adam}(w, g_w)$
  end for
  Sample $\{\mathbf{z}^{(i)}\}_{i=1}^{m} \sim \mu_\theta$ a batch of prior samples.
  $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} \Phi_w(f_\theta(\mathbf{z}^{(i)}))$
  $\theta \leftarrow \theta - \alpha \cdot \mathrm{Adam}(\theta, g_\theta)$
end while

Training set

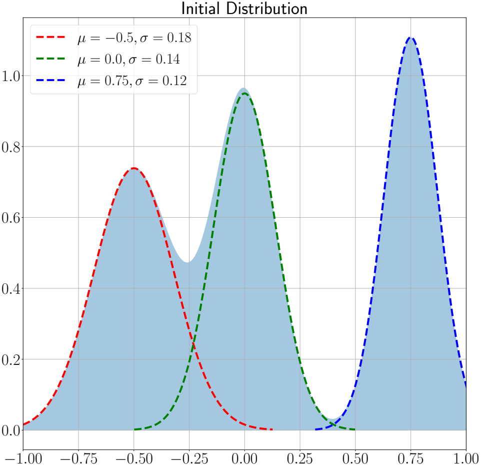


Figure 3: Linear combination of normal distributions

Results

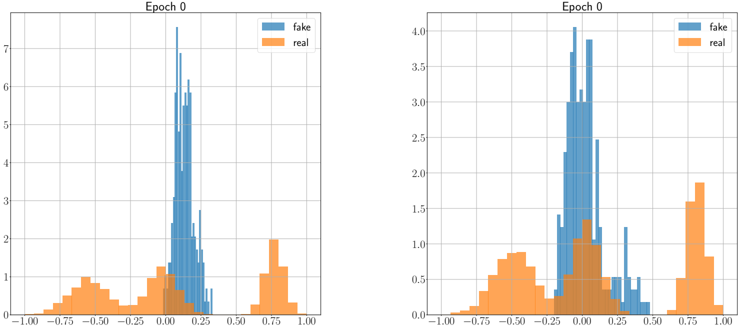
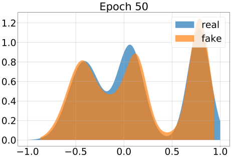
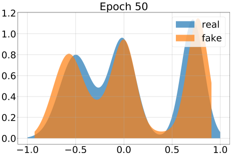


Figure 4: GAN training process

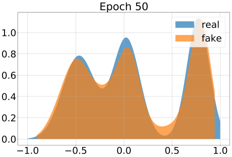
Results



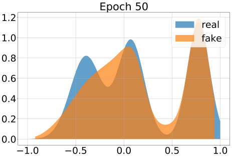
(a) Improved, $n_{cr} = 2$



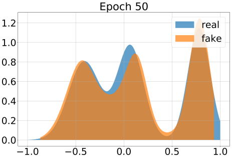
(b) Default, $n_{cr} = 2$



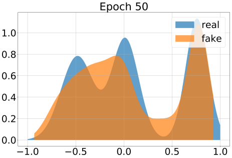
(c) Improved, $n_{cr} = 4$



(d) Default, $n_{cr} = 4$

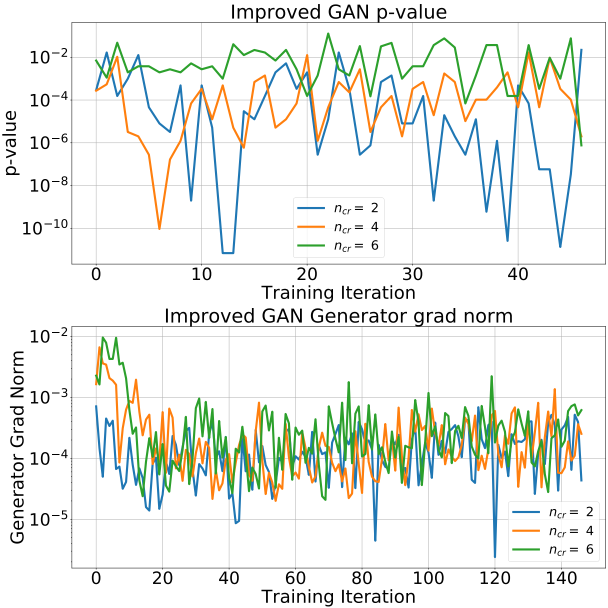


(e) Improved, $n_{cr} = 6$



(f) Default, $n_{cr} = 6$

Comparison of Improved GAN with different n_{cr}



Conclusion

- Gradient step size, proposed in the theorem guarantees convergence
- n_{cr} barely affect Improved GAN, but causes learning process destabilize it at Default GAN

Improved	n_{cr}	Epoch time	Convergence
-	2	2.92 ± 0.04 c	+
-	4	2.77 ± 0.04 c	-
-	6	2.62 ± 0.03 c	-
+	2	4.59 ± 0.04 c	+
+	4	4.26 ± 0.05	+
+	6	4.12 ± 0.02	$\pm$

Future work

- Ability to generate complex data such as picture
- Examine more complex neural architectures
- Apply this approach to other gradient methods

Thank you for attention!

Telegram: @mv2357, @aerowind