PROBGAN: Towards probabilistic GAN with theoretical guarantees (ICLR 2019)

Bayesian Modelling and Monte Carlo Inference for GAN (ICML 2018)

Hao He, Hao Wang, Guang-He Lee, Yonglong Tian

Presenter: Gyuseung Baek

May 17, 2020

Introduction

- Slightly improve Bayesian GAN(BGAN, Saatchi and Wilson).
  - Change likelihood(operation switch) and prior(informative)
- Theoretically and empirically prove PROBGAN is better than BGAN.
  - PROBGAN converges to the true data generation(if true is in our model).
  - BGAN is not suitable for any minimax-style GAN objective.
  - There are toy example BGAN fails in converging.
  - Experiments
- Algorithm: BGAN algorithm + alpha

GAN Framework

- Data space $\mathcal{X}$, Latent space $\mathcal{Z}$. True data generator: $p_{data} : \mathcal{Z} \to \mathcal{X}$.
- Generator with para. θ_g: $p_{gen}(\cdot|\theta_g) : \mathcal{Z} \to \mathcal{X}$
- Discriminator with para. θ_d: $D(\cdot|\theta_d) : \mathcal{X} \to [0, 1]$

- GAN(Goodfellow, 2016): Find θ_g and θ_d s.t.

$$\max_{\theta_d} \min_{\theta_g} \mathbb{E}_{x\sim p_{data}} [\log D(x|\theta_d)] + \mathbb{E}_{x\sim p_{gen}(\cdot|\theta_g)} [\log (1 - D(x|\theta_d))]$$

- General GAN Framework: Find θ_g and θ_d s.t.

$$\max_{\theta_d} \mathcal{J}_d(\theta_d|\theta_g) = \mathbb{E}_{x\sim p_{data}} [\phi_1 (D(x|\theta_d))] + \mathbb{E}_{x\sim p_{gen}(\cdot|\theta_g)} [\phi_2 (D(x|\theta_d))]$$

$$\max_{\theta_g} \mathcal{J}_g(\theta_g|\theta_d) = \mathbb{E}_{x\sim p_{gen}(\cdot|\theta_g)} [\phi_3 (D(x|\theta_d))]$$

- minimax-style: $\phi_2 = -\phi_3$

- Mode Collapse (Control ϕ, Multiple generator, Bayesian GAN)

BGAN

- Give a (underying) distn. for $\theta_d(q_d)$ and $\theta_g(q_g)$.
- Generator: $p_{model}(x|q_g) = \mathbb{E}_{\theta_g \sim q_g(\theta_g)} [p_{gen}(x|\theta_g)]$ for $x \in \mathcal{X}$.
- Goal: estimate posterior of q
- Information

$$p(\theta_g|\theta_d) \propto \exp\{\mathcal{J}_g(\theta_d|\theta_g)\}\, prior(\theta_g|\alpha_g)$$
$$p(\theta_d|\theta_g) \propto \exp\{\mathcal{J}_d(\theta_g|\theta_d)\}\, prior(\theta_d|\alpha_d)$$

BGAN

- Posterior estimation for q: for given $q_d^{(t)}$ and $q_g^{(t)}$, update q_d and q_g as:

$$q_g^{(t+1)}(\theta_g)|q_d^{(t)} \propto \exp\left\{\mathbb{E}_{\theta_d \sim q_d^{(t)}} \mathcal{J}_g(\theta_g|\theta_d)\right\} prior(\theta_g|\alpha_g)$$
$$q_d^{(t+1)}(\theta_d)|q_g^{(t)} \propto \exp\left\{\mathbb{E}_{\theta_g \sim q_g^{(t)}} \mathcal{J}_d(\theta_d|\theta_g)\right\} prior(\theta_d|\alpha_d)$$

- prior: weak informative prior (indep. to θ_g, θ_d respectively).

PROBGAN

- prior at time t: informative prior for generator. (Compatibility)

$$prior^{(t)}(\theta_g|\alpha_g) = q_g^{(t)}(\theta_g)$$

$\left(q_g^{(t+1)}(\theta_g)|q_d^{(t)} \rightarrow q_g^{(t+1)}(\theta_g)|q_d^{(t)}, q_g^{(t)}\right)$

- likelihood: switch $\mathbb{E}$ and $\mathcal{J}$ (intuitive & empirical)
- Posterior estimation for q: for given $q_d^{(t)}$ and $q_g^{(t)}$, update q_d and q_g as:

$$q_g^{(t+1)}(\theta_g) \propto \exp\left\{\mathcal{J}_g(\theta_g|\mathbb{E}_{\theta_d \sim q_d^{(t)}}\theta_d)\right\} q_g^{(t)}(\theta_g)$$

$$q_d^{(t+1)}(\theta_d) \propto \exp\left\{\mathcal{J}_d(\theta_d|\mathbb{E}_{\theta_g \sim q_g^{(t)}}\theta_g)\right\}$$

(Thm 1) Guarantee of Convergence

Assume the GAN objective(ϕ_1, ϕ_2, ϕ_3) and the discriminator space are symmetry. If there exist a distn. q_g^* for θ_g s.t. $p_{model}(x|q_g^*) = \mathbb{E}_{\theta_g \sim q_g^*} [p_{gen}(x|\theta_g)] = p_{data}(x)$ for all $x \in \mathcal{X}$, there exists a ideal discriminator distn. q_d^* s.t. $D(x|q_g^*) = \mathbb{E}_{\theta_d \sim q_d^*} D(\cdot|\theta_d) = Const.$. Moreover, q_g^* and q_d^* is an equilibrium of the dynamic in previous page.

(Lemma 1) Compatibility Issue

Consider a joint distribution $p(x, y)$ of variable X and Y. Its conditional distributions can be represented in the forms of $p(x|y) \propto \exp\{L(x, y)\}q_x(x)$ and $p(y|x) \propto \exp\{-L(x, y)\}q_y(y)$ only if X and Y are independent and $L(x, y)$ is decomposable, i.e. $\exists L_x$ and L_y, $L(x, y) = L_x(x) + L_y(y)$.

(Lemma 2) Convergence Issue

- Set Data space $\mathcal{X} = \{0, 1\}$, para. space for generator $\Theta_g = \{\theta_g^0, \theta_g^1\}$, para. space for discriminator $\Theta_d = \{\theta_d^0, \theta_d^1\}$.
- Generator: $p_{gen}(x|\theta_g^0) = Bern(0)$, $p_{gen}(x|\theta_g^1) = Bern(1)$
  - Distn. of generators: $q_g(\theta_g; |\gamma) = \gamma\mathbb{I}(\theta_g = \theta_g^1) + (1-\gamma)\mathbb{I}(\theta_g = \theta_g^0)$
- Discriminator: $D(x|\theta_d^0) = \epsilon\mathbb{I}(x = 1) + (1-\epsilon)\mathbb{I}(x = 0)$, $D(x|\theta_d^1) = \epsilon\mathbb{I}(x = 0) + (1-\epsilon)\mathbb{I}(x = 1)$

Lemma 2. For every $\lambda \in (0, 1)$ s.t. the desired generator distribution $q_g^*(\theta_g) \triangleq q_g(\theta_g|\gamma = \lambda)$ is not a fixed point of the iterative dynamics of PROBGAN.

SGHMC based - same with BGAN

Algorithm 1 Our Meta Inference Algorithm

Input: Initial Monte Carlo samples of $\{\theta_{d,m}^{(0)}\}_{m=1}^{M_d}$ and $\{\theta_{g,m}^{(0)}\}_{m=1}^{M_g}$, learning rate η, SGHMC noise factor α, number of updates in SGHMC procedure L.

for $t = 1, \cdots$ **do**
  for $m = 1$ **to** M_d **do**
    $\theta_{d,m} \leftarrow \theta_{d,m}^{(t)}$
    for $l = 1$ **to** L **do**
      $\mathbf{n} \sim \mathcal{N}(0, 2\alpha\eta I)$
      $\mathbf{v} \leftarrow (1-\alpha)\mathbf{v} + \eta\nabla_{\theta_d} \log q_d^{(t+1)}(\theta_{d,m}) + \mathbf{n}$
      $\theta_{d,m} \leftarrow \theta_{d,m} + \mathbf{v}$
    end for
    $\theta_{d,m}^{(t+1)} \leftarrow \theta_{d,m}$
  end for
  for $m = 1$ **to** M_g **do**
    $\theta_{g,m} \leftarrow \theta_{g,m}^{(t)}$
    for $l = 1$ **to** L **do**
      $\mathbf{n} \sim \mathcal{N}(0, 2\alpha\eta I)$
      $\mathbf{v} \leftarrow (1-\alpha)\mathbf{v} + \eta\nabla_{\theta_g} \log q_g^{(t+1)}(\theta_{g,m}) + \mathbf{n}$
      $\theta_{g,m} \leftarrow \theta_{g,m} + \mathbf{v}$
    end for
    $\theta_{g,m}^{(t+1)} \leftarrow \theta_{g,m}$
  end for
end for

Difference from BGAN

- The only difference is prior term of generator.

$$\nabla_{\theta_g} \log q_g^{(t+1)}(\theta_g) = \nabla_{\theta_g} \mathcal{J}(\theta_g | \mathbb{E}_{\theta_d \sim q_d^{(t)}} \theta_d) + \nabla_{\theta_g} \log q_g^{(t)}(\theta_g)$$

- Solution
  - Gaussian Mixutre Approximation (GMA): From Monte Carlo samples of θ_g at time t $\left\{\theta_{g,m}^{(t)}\right\}_{m=1}^{M_g}$, approax $q_g^{(t+1)}(\theta_g)$ as:

$$q_g^{(t+1)}(\theta_g) \simeq C \exp\left\{\sum_{m=1}^{M_g} \frac{\|\theta_g - \theta_{g,m}^{(t)}\|_2^2}{2\sigma^2}\right\}$$

  - Partial Summation Approximation (PSA) Above equation can be expressed inductively:

$$\nabla_{\theta_g} \log q_g^{(t+1)}(\theta_g) = \sum_{i=0}^{t} \nabla_{\theta_g} \mathcal{J}(\theta_g | \mathbb{E}_{\theta_d \sim q_d^{(i)}} \theta_d)$$

    Therefore, if we store all historical discriminator samples, it can be calculated.
    Practically, we store subset of discriminators.

High-Dimensional Multi-modal synthetic dataset

- Dataset: latent dim(d) = 2, Data dim(D) = 100, number of modes(n) = 10

  $z \sim \mathcal{U}[-1,1]^d, x = A_i(z + b_i), A_i \sim N(0, \sigma_A^2 I_{D\times d}), b_i \sim N(0, \sigma_b^2 I_d)(i = 1, .., n)$

  $\sigma_A = \sigma_B = 5$. Generate K samples $\{x_k\}_{k=1}^K \sim p_{model}$
- Metric: projection distance
  $\epsilon_p(x) = \min_{1\leq i\leq n} \epsilon_i(x) \triangleq \|x - A_i(A_i^\top A_i)^{-1}A_i^\top x\|_2$.
  Hit set $\mathcal{H}_i \triangleq \{x_k | \epsilon_i(x_k) < \eta\}$ (η: threshold. makes $\mathcal{H}_i$s are indep.)
  Projected hit set $\mathcal{PH}_i \triangleq \{(A_i^\top A_i)^{-1}A_i^\top x - b_i | x \in \mathcal{H}_i\}$
  - Hit ratio $\mathcal{H}_r \triangleq \sum_{i=1}^n |\mathcal{H}_i|/K$
  - Hit distance $\mathcal{H}_d \triangleq \sum_{i=1}^n \sum_{x\in\mathcal{H}_i} \epsilon_i(x) / \sum_{i=1}^n |\mathcal{H}_i|$
  - Cover error $\mathcal{C}_\epsilon \triangleq \frac{1}{n}\sum_{i=1}^n KL\left(\hat{p}(\cdot|\mathcal{PH}_i)||\mathcal{U}[-1,1]^d\right)$

High-Dimensional Multi-modal synthetic dataset

Table 2: Hit ratios ($\mathcal{H}_r$), hit distances ($\mathcal{H}_d$), cover errors ($\mathcal{C}_e$) results. Note, if the model failed to capture all the modes of real data, by definition its cover error is ∞. In that case, we report the averaged KL-divergence on modes captured by the model in brackets.

	$\mathcal{H}_r$ (HIGHER IS BETTER), $\mathcal{H}_d$ (LOWER IS BETTER)				$\mathcal{C}_e$ (LOWER IS BETTER)			
	GAN-MM	GAN-NS	WGAN	LSGAN	GAN-MM	GAN-NS	WGAN	LSGAN
GAN	0.86, 22.6	0.85, 23.1	0.78, 26.7	0.74, 23.1	12.11	8.86	7.20	∞ (12.07)
MGAN	0.82, 24.2	0.84, 25.5	0.67, 31.7	0.81, 23.6	5.46	6.31	5.00	∞ (4.25)
BGAN	1.0, **5.5**	1.0, **6.4**	1.0, **12.1**	1.0, 6.3	∞ (1.73)	1.76	4.32	1.80
ProbGAN-GMA	1.0, 7.4	1.0, 7.7	1.0, 15.5	1.0, **5.3**	1.84	**1.73**	3.01	1.79
ProbGAN-PSA	1.0, 5.8	1.0, **6.4**	1.0, 12.5	1.0, 6.4	**1.75**	1.75	**2.28**	**1.74**

Table 3: Inception score and FID results on CIFAR-10. Results of each model trained with 4 different GAN objectives are all reported.

	Inception scores (higher is better)				FIDs (lower is better)			
	GAN-MM	GAN-NS	WGAN	LSGAN	GAN-MM	GAN-NS	WGAN	LSGAN
DCGAN	6.53	7.21	7.19	7.36	35.57	27.68	28.31	29.11
MGAN	7.19	7.25	7.18	7.34	30.01	27.55	28.37	30.72
BGAN	7.21	7.37	7.26	**7.46**	29.87	24.32	29.87	29.19
ProbGAN-PSA	**7.75**	**7.53**	**7.28**	7.36	**24.60**	**23.55**	**27.46**	**26.90**

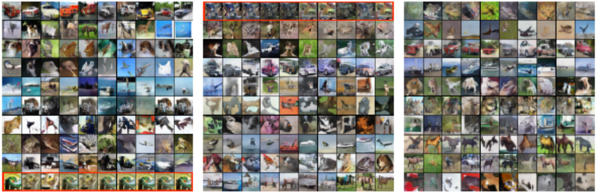
Natural Image Dataset

- Dataset: CIFAR10, STL-10, ImageNet
- Metric:
  - Inception Score: $\exp\left(\mathbb{E}_x\left[KL(p(y|x)\|p(y))\right]\right)$ where $p(y|x)$: pre-trained inception model(googlenet) and $p(y)$ is average of $p(y|x)$ over all images in dataset.

  Frechet Inception Distance(FID): measure the similarity between the real and synthetic data.

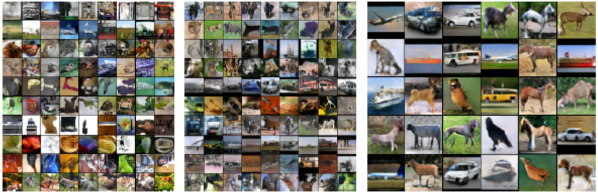
DATASET	STL-10		IMAGENET	
	INCEPTION SCORES	FIDS	INCEPTION SCORES	FIDS
DCGAN	8.05 ± 0.101	51.01	7.66 ± 0.113	48.99
MGAN	8.72 ± 0.096	51.56	7.77 ± 0.108	45.75
BGAN	8.84 ± 0.100	47.35	8.52 ± 0.075	29.68
PROBGAN-PSA	**8.87** ± 0.095	**46.74**	**8.57** ± 0.073	**27.69**

Natural Image Dataset



(a) MGAN (Epoch 250) (b) BGAN (Epoch 250) (c) ProbGAN-PSA (Epoch 250)

Figure 3: Images generated by MGAN, BGAN and our model trained on CIFAR 10 with GAN-NS objective. The tenth generator of MGAN (Figure 3(a)) and the first of BGAN (Figure 3(b)) collapse while generators of our method all work well. DCGAN (Figure 10 in the appendix) also presents 'single generator collapse' issue. Note that, mode collapse also happens when baseline models trained with other GAN objectives.



(a) ImageNet (randomly picked) (b) STL-10 (randomly picked) (c) STL-10 (cherry-picked)

Figure 4: Images generated by ProbGAN trained on ImageNet (left) and STL-10 (middle, right). Figure 4(c) are cherry-picked synthetic images on STL-10.