McGan: Mean and Covariance Feature Matching GAN

arXiv:1702.08398, Feb. 27th, 2017

Youssef Mroueh Tom Sercu Vaibhava Goel

IBM Watson Research Center, New York

May 12th, 2017

Presented by: Dan Salo

Vanilla GAN Limitations

Summary:

- Extends theory presented in WGAN paper to other first order and second order feature matching.
- Results do not make a strong case for using second order.

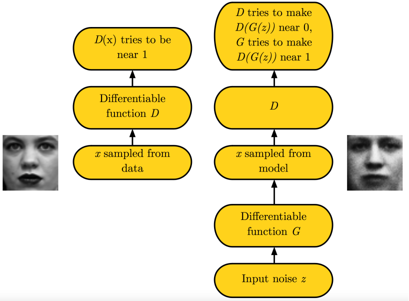
Outline:

- GAN, Limitations, and Wasserstein metric (3 slides)
- McGan Math and Implementation (5 slides)
- Experiments (2 slides)

GAN (Goodfellow NIPS 2014)

Idea: Find Nash Equilibrium of two-player (D, G) minimax game.

- $g_\theta : \mathcal{Z} \subset \mathbb{R}^{n_z} \to \mathcal{X}$, function (DNN) with parameters θ.
- $\mathbb{P}_\theta$ is the distrib. of $g_\theta(z)$, with p_z a fixed distrib. on $\mathcal{Z}$.
- $\mathbb{P}_r$ is the distrib. of real data.
- Discriminator (D) critiques Generator (G)



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

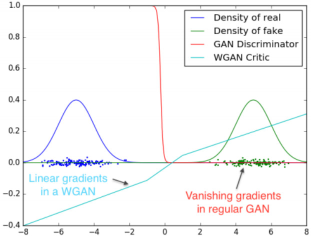
Vanilla GAN Limitations

Limitations:

- "Mode dropping" in Generator and "vanishing gradients" from Discriminator
- Instability and unmeaningful loss functions during training
- Requires a specific alternating learning schedule
- Gradient descent decreases one loss but changes the other loss; not ideal for Nash Equilibrium
- Theory assumes search over function space, but algorithm searches over parameter space

WGAN (Arjovsky 2017a)

Idea: Replace JS divergence with a metric that induces a weaker topology, i.e. the Wasserstein-1 or Earth-Mover distance over a family of functions $\{f_w\}_{w\in\mathcal{W}}$ that are all K-Lipschitz.



$$W_{\mathbb{P}_r||\mathbb{P}_\theta} = \inf_{\gamma\in\prod(P,Q)} \mathop{\mathbb{E}}_{(x,y)\sim\gamma} ||x-y|| \tag{2}$$

$$= \sup_{||f||_L\leq 1} \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r} f(x) - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_\theta} f(x) \tag{3}$$

$$= \sup_{||f||_L\leq 1} \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r} f(x) - \mathop{\mathbb{E}}_{z\sim p_z} f(g_\theta(z)) \tag{4}$$

IPM for Learning Generative Models

Integral Probability Measure: Find the function f from space $\mathscr{F}$ that maximizes the discrepancy between the means of two distributions, $\mathbb{P}$ and $\mathbb{Q}$.

$$d_{\mathscr{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathscr{F}} \left\{ \mathop{\mathbb{E}}_{x\sim\mathbb{P}} f(x) - \mathop{\mathbb{E}}_{x\sim\mathbb{Q}} f(x) \right\} \tag{5}$$

General GAN Objective with IPM:

$$\begin{aligned}
\mathcal{L}_{GAN} &= \min_{g_\theta} d_{\mathscr{F}}(\mathbb{P}_r, \mathbb{P}_\theta) && (6)\\
&= \min_{g_\theta} \sup_{f\in\mathscr{F}} \left\{ \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r} f(x) - \mathop{\mathbb{E}}_{z\sim p_z} f(g_\theta(z)) \right\} && (7)\\
&= \min_{g_\theta} \sup_{f\in\mathscr{F}} \frac{1}{N}\sum_{i=1}^{N} f(x_i) - \frac{1}{M}\sum_{j=1}^{M} f(g_\theta(z_j)) && (8)
\end{aligned}$$

where $\{x_i, 1\ldots N\} \sim \mathbb{P}_r$ and $\{z_i, 1\ldots M\} \sim p_z$

IPM$_{\mu,q}$: Mean Feature Matching GAN

Idea: Define function space $\mathscr{F}$ as a finite dimensional Hilbert space with bounded parameter space Ω.

$$\mathscr{F}_{v,\omega,p} = \{f(x) = \langle v, \Phi(x)\rangle | v \in \mathbb{R}^m, ||v||_p \leq 1, \Phi_\omega : \mathcal{X} \rightarrow \mathbb{R}^m, \omega \in \Omega\}$$

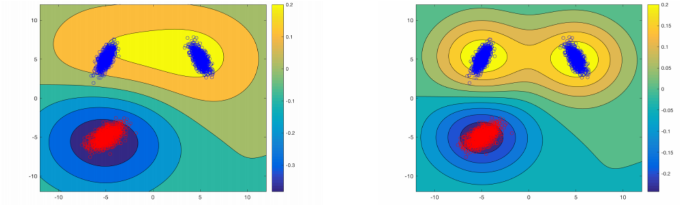
$$d_{\mathscr{F}}(\mathbb{P}_r, \mathbb{P}_\theta) = \max_{\omega\in\Omega, v, ||v||_p\leq 1} \left\langle v, \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r} \Phi_\omega(x) - \mathop{\mathbb{E}}_{z\sim p_z} \Phi_\omega(g_\theta(z)) \right\rangle \quad (9)$$

$$= \max_{\omega\in\Omega} \left[\max_{v, ||v||_p\leq 1} \left\langle v, \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r} \Phi_\omega(x) - \mathop{\mathbb{E}}_{z\sim p_z} \Phi_\omega(g_\theta(z)) \right\rangle \right] \quad (10)$$

$$= \max_{\omega\in\Omega} ||\mu_\omega(\mathbb{P}_r) - \mu_\omega(\mathbb{P}_\theta)||_q \quad (11)$$

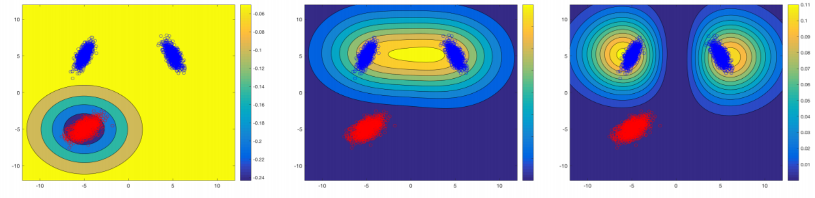
where $\mu_\omega(\mathbb{P}_\theta) = \mathbb{E}\left[\Phi_\omega(g_\theta(z))\right]$ is the mean of the feature vector and $||\cdot||_q$ is the q-norm.

Motivation for Second Order



a) IPM $_{\mu,2}$: Level sets of $f(x) = \langle v^*, \Phi_\omega(x)\rangle$ $v^* = \frac{\mu_w(\mathbb{P})-\mu_w(\mathbb{Q})}{\|\mu_w(\mathbb{P})-\mu_w(\mathbb{Q})\|_2}$.

b) IPM $_\Sigma$: Level sets of $f(x) = \sum_{j=1}^{k} \langle u_j, \Phi_\omega(x)\rangle \langle v_j, \Phi_\omega(x)\rangle$ $k = 3, u_j, v_j$ left and right singular vectors of $\Sigma_w(\mathbb{P}) - \Sigma_w(\mathbb{Q})$.



Level Sets of c) $\langle u_1, \Phi_\omega(x)\rangle \langle v_1, \Phi_\omega(x)\rangle$ d) $\langle u_2, \Phi_\omega(x)\rangle \langle v_2, \Phi_\omega(x)\rangle$ e) $\langle u_3, \Phi_\omega(x)\rangle \langle v_3, \Phi_\omega(x)\rangle$

Figure: IPM_Σ characterizes real data (blue) better than $\text{IPM}_{\mu,2}$

IPM_Σ: Covariance Feature Matching GAN

Idea: Motivated by PCA, define function space $\mathscr{F}$ of bilinear functions in Φ_ω with bounded parameter space Ω.

$$\mathscr{F}_{U,V,\omega} = \{f(x) = \langle U^T \Phi_\omega(x) \rangle \langle V^T \Phi_\omega(x) \rangle \mid U, V \in \mathbb{R}^{m \times k}, U^T U = I_k, V^T V = I_k, \omega \in \Omega\}$$

$$\begin{aligned} d_{\mathscr{F}_{U,V,\omega}}(\mathbb{P}_r, \mathbb{P}_\theta) &= \max_{\omega \in \Omega} \Big[\max_{U,V \in \mathcal{O}_{m,k}} \mathop{\mathbb{E}}_{x \sim \mathbb{P}_r} \left\langle U^T \Phi_\omega(x), V^T \Phi_\omega(x) \right\rangle - \\ &\qquad \mathop{\mathbb{E}}_{x \sim p_z} \left\langle U^T \Phi_\omega(g_\theta(z)), V^T \Phi_\omega(g_\theta(z)) \right\rangle \Big] &(12) \\ &= \max_{\omega \in \Omega} \max_{U,V \in \mathcal{O}_{m,k}} \text{tr}[U^T (\Sigma_\omega(\mathbb{P}_r) - \Sigma_\omega(\mathbb{P}_\theta)) V] &(13) \\ &= \max_{\omega \in \Omega} ||[\Sigma_\omega(\mathbb{P}_r) - \Sigma_\omega(\mathbb{P}_\theta)]_k||_* &(14) \end{aligned}$$

where $\Sigma_\omega(\mathbb{P}) = \mathop{\mathbb{E}}_{x \sim \mathbb{P}} \Phi_\omega(x) \Phi_\omega(x)^T$ is the uncentered feature covariance embedding of $\mathbb{P}$, and $|| \cdot ||_*$ is the nuclear norm.

TensorFlow Loss Functions

Vanilla GAN Losses:

```
D_loss = -tf.reduce_mean(tf.log(D_real) + tf.log(1 - D_fake))
G_loss = -tf.reduce_mean(tf.log(D_fake))
```

$\mathrm{IPM}_{\mu,\infty}$ or WGAN Losses:

```
D_loss = tf.reduce_mean(D_real) - tf.reduce_mean(D_fake)
G_loss = -tf.reduce_mean(D_fake)
```

Bounded Ω with weight clipping:

```
clip_D = [p.assign(tf.clip_by_value(p, -0.01, 0.01)) for p in theta_D]
```

Bounded Ω with gradient penalty (Arjovsky 2017b):

```
ddx = tf.gradients(d_hat, x_hat)
ddx = tf.sqrt(tf.reduce_sum(tf.square(ddx), axis=1))
D_loss = D_loss_WGAN + ddx
```

IPM$_\mu$: LSUN Generation

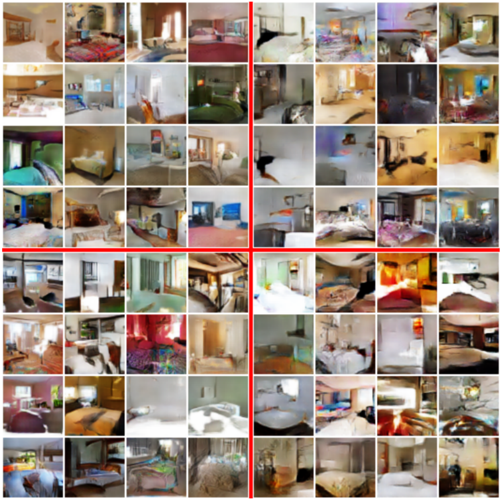


Figure: Primal (left), Dual (right), l_1 (top), l_2 (bottom)

IPM_Σ: Conditional Cifar-10 Generation

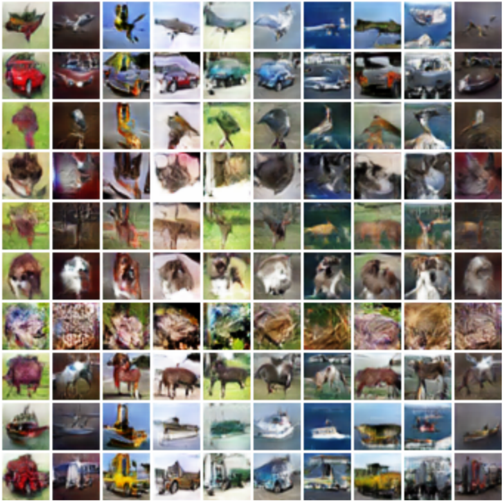


Figure: Rows: same class conditioning, Columns: same $z \sim p_z$ sample

McGan Algorithms

Algorithm 1 Mean Matching GAN - Primal (P_μ)

Input: p to define the ball of v ,η Learning rate, n_c number of iterations for training the critic, c clipping or weight decay parameter, N batch size
Initialize v, ω, θ
repeat
  for $j = 1$ **to** n_c **do**
    Sample a minibatch $x_i, i = 1 \dots N, x_i \sim \mathbb{P}_r$
    Sample a minibatch $z_i, i = 1 \dots N, z_i \sim p_z$
    $(g_v, g_\omega) \leftarrow (\nabla_v \hat{\mathscr{L}}_\mu(v, \omega, \theta), \nabla_\omega \hat{\mathscr{L}}_\mu(v, \omega, \theta))$
    $(v, \omega) \leftarrow (v, \omega) + \eta$ RMSProp $((v, \omega), (g_v, g_\omega))$
    {Project v on ℓ_p ball, $B_{\ell_p} = \{x, \|x\|_p \leq 1\}$}
    $v \leftarrow \text{proj}_{B_{\ell_p}}(v)$
    $\omega \leftarrow \text{clip}(\omega, -c, c)$ {Ensure Φ_ω is bounded}
  end for
  Sample $z_i, i = 1 \dots N, z_i \sim p_z$
  $d_\theta \leftarrow -\nabla_\theta \left\langle v, \frac{1}{N} \sum_{i=1}^{N} \Phi_\omega(g_\theta(z_i)) \right\rangle$
  $\theta \leftarrow \theta - \eta$ RMSProp (θ, d_θ)
until θ converges

Algorithm 3 Covariance Matching GAN - Primal (P_Σ)

Input: k the number of components ,η Learning rate, n_c number of iterations for training the critic, c clipping or weight decay parameter, N batch size
Initialize U, V, ω, θ
repeat
  for $j = 1$ **to** n_c **do**
    Sample a minibatch $x_i, i = 1 \dots N, x_i \sim \mathbb{P}_r$
    Sample a minibatch $z_i, i = 1 \dots N, z_i \sim p_z$
    $G \leftarrow (\nabla_U, \nabla_V, \nabla_\omega)\hat{\mathscr{L}}_\sigma(U, V, \omega, \theta)$
    $(U, V, \omega) \leftarrow (U, V, \omega) + \eta$ RMSProp $((U, V, \omega), G)$
    { Project U and V on the Stiefel manifold $\mathcal{O}_{m,k}$}
    $Q_u, R_u \leftarrow QR(U) \quad s_u \leftarrow \text{sign}(\text{diag}(R_u))$
    $Q_v, R_v \leftarrow QR(V) \quad s_v \leftarrow \text{sign}(\text{diag}(R_v))$
    $U \leftarrow Q_u \text{Diag}(s_u)$
    $V \leftarrow Q_v \text{Diag}(s_v)$
    $\omega \leftarrow \text{clip}(\omega, -c, c)$ {Ensure Φ_ω is bounded}
  end for
  Sample $z_i, i = 1 \dots N, z_i \sim p_z$
  $d_\theta \leftarrow -\nabla_\theta \frac{1}{N} \sum_{j=1}^{N} \langle U\Phi_\omega(g_\theta(z_j)), V\Phi_\omega(g_\theta(z_j)) \rangle$
  $\theta \leftarrow \theta - \eta$ RMSProp (θ, d_θ)
until θ converges