A Variational Inequality Perspective on GANs

Gauthier Gidel*¹, Hugo Berard*¹², Gaëtan Vignoud¹, Pascal Vincent¹², Simon Lacoste-Julien¹

*equal contribution
¹ MILA, Université de Montréal
² Facebook AI Research (FAIR), Montréal
Generative Adversarial Networks (GANs)

[Goodfelow et al. NIPS 2014]

Noise → Generator → Fake Data

Generator

Discriminator

Fake or Real

True Data

Hugo Berard,
Montreal AI Symposium, August 28, 2018
Generative Adversarial Networks (GANs)

[Goodfelow et al. NIPS 2014]

If $D$ is non-parametric: $L(\theta) = \text{JSD}(p_D || p_\theta)$

Non-saturating GAN:

Loss of Generator:
\[
\min_{\theta} -\mathbb{E}_{z \sim p_Z} \left[ \log(D_\phi(G_\theta(z))) \right]
\]

Loss of Discriminator:
\[
\max_{\phi} \mathbb{E}_{x \sim p_D} \left[ \log(D_\phi(x)) \right] + \mathbb{E}_{z \sim p_Z} \left[ \log(1 - D_\phi(G_\theta(z))) \right]
\]
Two-player Games

\[
\begin{align*}
\min_{\theta} L_\theta(\theta, \phi) & \quad \text{and} \quad \min_{\phi} L_\phi(\theta, \phi)
\end{align*}
\]

Zero-sum game if: \( L_\theta(\theta, \phi) = -L_\phi(\theta, \phi) \) also called Saddle Point (SP).

Example: WGAN formulation [Arjovsky et al. 2017]

\[
\min_{\theta} \max_{\phi, \|f_\phi\|_L \leq 1} \left[ \mathbb{E}_{x \sim p_D}[f_\phi(x)] - \mathbb{E}_{z \sim p_Z}[f_\phi(g_\theta(z))] \right] = L_\theta(\theta, \phi) = -L_\phi(\theta, \phi)
\]

Hugo Berard,
Montreal AI Symposium, August 28, 2018
“Saddle Points are Hard to Optimize ...”

Gradient vector field: \( F(\theta, \phi) = \begin{pmatrix} \nabla_\theta L_\theta(\theta, \phi) \\ \nabla_\phi L_\phi(\theta, \phi) \end{pmatrix} \)

**Bilinear** saddle point = Linear in \( \theta \) and \( \phi \)

\( \Rightarrow \) “Cycling behavior” (see right).

**Example**: WGAN with linear discriminator and generator

\[
\min_\theta \max_{\phi, \|f_\phi\|_{L \leq 1}} \phi^T \mathbb{E}_{x \sim p_D}[x] - \phi^T \theta \mathbb{E}_{z \sim p_Z}[z]
\]

(https://www.inference.vc/my-notes-on-the-numerics-of-gans/)
… but saddle points can be optimized!

**Non-convergent**
- **Blue**: Simultaneous gradient method.
- **Orange**: Alternating gradient method.

**Convergent**
- **Green**: Gradient method with **averaging**.
- **Purple**: **Extragradient** method.

from *Variational Inequality* literature
GANs as a Variational Inequality

New perspective for GANs:
- Based on stationary conditions.
- Relates to vast literature with standard algorithms.

Nash-Equilibrium:
\[
\begin{align*}
\theta^* &= \arg \min_{\theta} L_{\theta}(\theta, \phi^*) \\
\phi^* &= \arg \min_{\phi} L_{\phi}(\theta^*, \phi)
\end{align*}
\]

No player can improve its cost

Stationary Conditions:
\[
\begin{align*}
\nabla_{\theta} L_{\theta}(\theta^*, \phi^*)^T (\theta - \theta^*) &\geq 0 \\
\nabla_{\phi} L_{\phi}(\theta^*, \phi^*)^T (\phi - \phi^*) &\geq 0
\end{align*}
\]

\forall (\theta, \phi) \in \Theta \times \Phi

can be constraint sets.
GANs as a Variational Inequality

Stationary Conditions:
\[
\begin{align*}
\nabla_\theta L_\theta(\theta^*, \phi^*)^T(\theta - \theta^*) & \geq 0 \\
\nabla_\phi L_\phi(\theta^*, \phi^*)^T(\phi - \phi^*) & \geq 0
\end{align*}
\quad \forall (\theta, \phi) \in \Theta \times \Phi
\]

Can be written as:
\[
F(\omega) = \begin{pmatrix}
\nabla_\theta L_\theta(\omega) \\
\nabla_\phi L_\phi(\omega)
\end{pmatrix}
\quad \omega = (\theta, \phi)
\]

\[
F(\omega^*)^T(\omega - \omega^*) \geq 0 
\quad \forall \omega \in \Omega
\]

\omega^* solves the Variational Inequality
GANs as a Variational Inequality

Takeaways:
- GAN can be formulated as a Variational Inequality.
- Encompass most of GANs formulations.
- Standard algorithms from Variational Inequality can be used for GANs.
- Theoretical Guarantees (for convex and stochastic cost functions).

\[
\begin{align*}
\theta^* &= \arg\min_{\theta} L_\theta(\theta, \phi^*) \\
\phi^* &= \arg\min_{\phi} L_\phi(\theta^*, \phi)
\end{align*}
\]

\[F(\omega^*)^T(\omega - \omega^*) \geq 0 \quad \forall \omega \in \Omega\]
Standard Algorithms from Variational Inequality

Method 1: **Averaging**

\[ \bar{\omega}_T = \frac{\sum_t \rho_t \omega_t}{\sum_t \rho_t} \]

- Converge even for "cycling behavior".
- Easy to implement.
- Can be combined with any method.

Method 2: **Extragradient**

- **Step 1:** \[ \omega_{t+\frac{1}{2}} = \omega_t - \gamma_t F(\omega_t) \]
- **Step 2:** \[ \omega_{t+1} = \omega_t - \gamma_t F(\omega_{t+\frac{1}{2}}) \]

**Intuition:** Look 1 step in the future and anticipate next move of adversary.

- Standard in the literature.
- Does not require averaging.
- Theoretically and empirically faster.

Hugo Berard, Montreal AI Symposium, August 28, 2018
OneSEM: Re-using the gradients

**Problem**: Extragradient requires to compute two gradients at each step.

**Solution**: OneSEM \( \text{Re-use} \) gradient.

- **Step 1**: \( \omega_{t+\frac{1}{2}} = \omega_t - \gamma_t F(\omega_{t-\frac{1}{2}}) \) \( \text{re-use from previous iteration.} \)
- **Step 2**: \( \omega_{t+1} = \omega_t - \gamma_t F(\omega_{t+\frac{1}{2}}) \) \( \text{(same as extragradient).} \)
Experimental Results

Bilinear Stochastic Objective:

$$\frac{1}{n} \sum_{i=1}^{n} \left( x^\top M^{(i)} y + x^\top a^{(i)} + y^\top b^{(i)} \right).$$
Experimental Results: WGAN on CIFAR10

Inception Score vs nb of generator updates

- AltSGD $\gamma = 1 \cdot 10^{-4}$
- SimSGD $\gamma = 1 \cdot 10^{-4}$
- OneSEM $\gamma = 1 \cdot 10^{-4}$
- SEM $\gamma = 1 \cdot 10^{-4}$

Inception Score on CIFAR10

<table>
<thead>
<tr>
<th>Method</th>
<th>no averaging</th>
<th>with averaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimSGD</td>
<td>5.01 ± 0.06</td>
<td>5.15 ± 0.15</td>
</tr>
<tr>
<td>AltSGD</td>
<td>5.13 ± 0.03</td>
<td>5.28 ± 0.08</td>
</tr>
<tr>
<td>SEM</td>
<td>5.52 ± 0.08</td>
<td>5.62 ± 0.08</td>
</tr>
<tr>
<td>OneSEM</td>
<td>5.45 ± 0.10</td>
<td>5.63 ± 0.12</td>
</tr>
</tbody>
</table>

Extragradient Methods
Experimental Results: WGAN-GP on CIFAR10

Inception Score vs Time

Inception Score on CIFAR10

<table>
<thead>
<tr>
<th>Method</th>
<th>no averaging</th>
<th>with averaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimSGD</td>
<td>6.00 ± 0.07</td>
<td>6.01 ± 0.08</td>
</tr>
<tr>
<td>AltSGD</td>
<td>6.25 ± 0.05</td>
<td>6.51 ± 0.04</td>
</tr>
<tr>
<td>SEM</td>
<td>6.22 ± 0.04</td>
<td>6.35 ± 0.09</td>
</tr>
<tr>
<td>OneSEM</td>
<td>6.27 ± 0.06</td>
<td>6.23 ± 0.13</td>
</tr>
</tbody>
</table>

Extragradient Methods

Mila
Hugo Berard,
Montreal AI Symposium, August 28, 2018
Conclusion

- GAN can be formulated as a **Variational Inequality**.

- Bring *standard methods* from *optimization literature* to the GAN community.

- **Averaging** helps improve the inception score (further evidence by [Yazici et al. 2018]).

- **Extragradient** is faster and achieve better convergence.

- Introduce **OneSEM** a cheaper version of extragradient.

- We can design better algorithm for GANs inspired from Variational Inequality.
Thank you!