Multi-objective training of Generative Adversarial Networks with multiple discriminators

Isabela Albuquerque*, João Monteiro*, Thang Doan, Breandan Considine, Tiago Falk, and Ioannis Mitliagkas

*Equal contribution
Recent literature proposed to tackle GANs training instability* issues with multiple discriminators (Ds)

2. Stabilizing GANs training with multiple random projections, Neyshabur et al. (2017)
3. Online Adaptative Curriculum Learning for GANs, Doan et al. (2018)

*Mode-collapse or vanishing gradients
The multiple discriminators GAN setting
Our work

Multiple discriminators GANs

Multi-objective optimization

Our work
Our work

\[ \min \mathcal{L}_G(z) = [l_1(z), l_2(z), \ldots, l_K(z)]^T \]

- Each \( l_k = -\mathbb{E}_{z \sim p_z} \log D_k(G(z)) \) is the loss provided by the \( k \)-th discriminator
Our work

\[
\min \mathcal{L}_G(z) = [l_1(z), l_2(z), ..., l_K(z)]^T
\]

- Multiple gradient descent (MGD) is a natural choice to solve this problem
  - But it might be too costly
- Alternative: maximize the hypervolume (HV) of a single solution
Multiple gradient descent

- Seeks a Pareto-stationary solution
- Two steps:
  1. Find a common descent direction $\forall l_k$
   1.1 Minimum norm element within the convex hull of all $\nabla l_k(x)$
  2. Update the parameters with $x_{t+1} = x_t - \lambda \frac{w_t^*}{\|w_t^*\|}$, where

\[
    w_t^* = \arg\min ||w||^2, \quad w = \sum_{k=1}^{K} \alpha_k \nabla l_k(x_t),
\]

\[
    \text{s.t.} \quad \sum_{k=1}^{K} \alpha_k = 1, \quad \alpha_k \geq 0 \quad \forall k
\]
Hypervolume maximization for training GANs
Hypervolume maximization for training GANs

\[ \mathcal{L}_G = -\log \left( \prod_{k=1}^{K} (\eta - l_k) \right) \]

\[ \mathcal{L}_G = - \sum_{k=1}^{K} \log(\eta - l_k) \]

\[ \frac{\partial \mathcal{L}_G}{\partial \theta} = \sum_{k=1}^{K} \frac{1}{\eta - l_k} \frac{\partial l_k}{\partial \theta} \]
Hypervolume maximization for training GANs

\[ \mathcal{L}_G = -\log \left( \prod_{k=1}^{K} (\eta - l_k) \right) \]

\[ \mathcal{L}_G = -\sum_{k=1}^{K} \log(\eta - l_k) \]

\[ \frac{\partial \mathcal{L}_G}{\partial \theta} = \sum_{k=1}^{K} \frac{1}{\eta - l_k} \frac{\partial l_k}{\partial \theta} \]

\[ \eta^t = \delta \max_k \{ l_k^t \}, \quad \delta > 1 \]
MGD vs. HV maximization vs. Average loss minimization

- MGD seeks a Pareto-stationary solution
  - $x_{t+1} \prec x_t$
- HV maximization seeks Pareto-optimal solutions
  - $HV(x_{t+1}) > HV(x_t)$
  - For the single-solution case, central regions of the Pareto-front are preferred
- Average loss minimization does not enforce equally good individual losses
  - Might be problematic in case there is a trade-off between discriminators
MNIST

- Same architecture, hyperparameters, and initialization for all methods
- 8 Ds, 100 epochs
- FID was calculated using a LeNet trained on MNIST until 98% test accuracy
Upscaled CIFAR-10 - Computational cost

- Different GANs with both 1 and 24 Ds + HV
- Same architecture and initialization for all methods
- Comparison of minimum FID obtained during training, along with computation cost in terms of time and space

<table>
<thead>
<tr>
<th></th>
<th># Disc.</th>
<th>FID-ResNet</th>
<th>FLOPS*</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCGAN</td>
<td>1</td>
<td>4.22</td>
<td>8e10</td>
<td>1292</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>1.89</td>
<td>5e11</td>
<td>5671</td>
</tr>
<tr>
<td>LSGAN</td>
<td>1</td>
<td>4.55</td>
<td>8e10</td>
<td>1303</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>1.91</td>
<td>5e11</td>
<td>5682</td>
</tr>
<tr>
<td>HingeGAN</td>
<td>1</td>
<td>6.17</td>
<td>8e10</td>
<td>1303</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>2.25</td>
<td>5e11</td>
<td>5682</td>
</tr>
</tbody>
</table>

* Floating point operations per second

- Additional cost → performance improvement
Cats $256 \times 256$
Thank you!

Questions? Come to our poster! #4

Code: https://github.com/joaomonteirof/hGAN