Which Training Methods for GANs do actually Converge?

Reproduced by: Zijie Pan, Kaiming Cheng

Dec 06. 2019
Motivation

Generative Adversarial Networks (GANs), proposed by Goodfellow in 2014, is a powerful latent variable model, showing dominant abilities to generate realistic image samples after training on sample data.

**Problem**: GANs are hard to train and gradient descent optimization results in no convergence.

**Main Question in this paper**: How will GAN training become locally asymptotically stable in the general case?
Tasks in this paper
1. Proposed Dirac-GAN configurations: Prove the necessity of absolute continuity.
2. Analyze unregularized and common regularized GAN training algorithm stability on Dirac-GAN
3. Proposed simplified gradient penalties leads to convergence
GANs are defined by a min-max two-player game between a discriminative network $D_{\psi}(x)$ and generative network $G_{\theta}(z)$.

Objective function:

$$\min_{G} \max_{D} \left( \mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p_{latent}} [\log(1 - D(G(z)))] \right).$$

Our goal when training GANs is to find a Nash-equilibrium.

(Mescheder et al. (2017)) shown that local convergence of GAN training near an equilibrium point $(\theta^*, \psi^*)$ can be analyzed by looking at the spectrum of the Jacobian $F_{\theta^* h}(\theta^*, \psi^*)$ at the equilibrium:

Eigenvalues with absolute value **bigger** than 1: Not Converge

Eigenvalues with absolute value **smaller** than 1: Converge with linear rate $O(|\lambda_{max}|^k)$

Eigenvalues are all on the **unit circle**: Converge (sublinear rate), Diverge or Neither
Mescheder et al. (2017)

When $\lambda^2$ is very close to zero, it is very likely to get imagery number for $\lambda$. Thus, we can require intractably small learning rates to achieve convergence.

Sønderby et al., 2016; Arjovsky & Bottou, 2017:

Show that for common use cases of GANs, we don’t have the property of absolute continuity for the data distributions like natural images.

Techniques that lead to local convergence:

- Arjovsky et al. (2017): Propose WGAN training
- Sønderby et al., 2016; Arjovsky & Bottou, 2017: Propose instance noise
- Roth et al., 2017: Propose zero-centered gradient penalties
...
Proposed Counter-example:
- Dirac-GAN
  - Not absolute continuity → Nonconvergence
  - No optimal discriminator parameter (except 0)
  - No incentive for the discriminator to move to the equilibrium when generator is the target distribution.
Vector Field of Dirac-GAN for Different Training Algorithm

<table>
<thead>
<tr>
<th>Method</th>
<th>Local convergence (a.c. case)</th>
<th>Local convergence (general case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>unregularized (Goodfellow et al., 2014)</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>WGAN (Arjovsky et al., 2017)</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>WGAN-GP (Gulrajani et al., 2017)</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DRAGAN (Kodali et al., 2017)</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Instance noise (Sønderby et al., 2016)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ConOpt (Mescheder et al., 2017)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Gradient penalties (Roth et al., 2017)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Gradient penalty on real data only</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Gradient penalty on fake data only</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

(a) Standard GAN
(b) Non-saturating GAN
(c) WGAN ($n_d = 5$)
(d) WGAN-GP ($n_d = 5$)
(e) Consensus optimization
(f) Instance noise
(g) Gradient penalty
(h) Gradient penalty (CR)
Animated Convergence Results for Unregularized GAN vs Gradient Penalty

Unregularized → Not always stable

WGAN and WGAN GP → Not always stable

Instance noise & zero-centered & gradient penalties -> stable
Proposed Solution (1)

Inspired from zero-center gradient penalties (Roth et al., 2017)

Simplified regularization term:

\[ R_1(\psi) := \frac{\gamma}{2} \mathbb{E}_{p_D(x)} \left[ \| \nabla D_\psi(x) \|^2 \right] \]

\[ R_2(\theta, \psi) := \frac{\gamma}{2} \mathbb{E}_{p_\theta(x)} \left[ \| \nabla D_\psi(x) \|^2 \right] \]
Data Summary

There are in total three different datasets:

- 2-D Example (2D Gaussian, Line, Circle, Four Lines)
  - (not implemented)
- CIFAR-10 dataset
Experimental Results

```
(resnet_5_0): ResnetBlock(
  (conv_0): Conv2d(1024, 1024, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (conv_1): Conv2d(1024, 2048, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (conv_s): Conv2d(1024, 2048, kernel_size=(1, 1), stride=(1, 1), bias=False)
)
  (fc): Linear(in_features=32768, out_features=1, bias=True)
)
=> Loading checkpoint from url...
Computing inception score...
/home/ec2-user/anaconda3/envs/pytorch_p36/lib/python3.6/site-packages/torch.nn.functional.py:2639: UserWarning: Default upsampling behavior when mode=bilinear is changed to align_corners=False since 0.4.0. Please specify align_corners=True if the old behavior is desired. See the documentation of nn.Upsample for details.
  "See the documentation of nn.Upsample for details.".format(mode))
Inception score: 2.4715 ± 0.0223
Creating samples...
100% 1/1 [00:02<00:00, 2.12s/it]
```
Experimental Results

Good behavior of WGAN-GP is surprising

Explanation in the next slide
We see that the R1- and R2-regularizers and WGAN-GP perform similarly and they achieve good results.

While we find that unregularized GAN training quickly leads to mode-collapse for these problems, our simple R1-regularizer enables stable training.

Reason:
WGAN-GP oscillates in narrow circles around the equilibrium which might be enough to produce images of sufficiently high quality.
Reproduced Results

• Sample Output (Regularized) on different datasets, showing with regularizer we can have stable training:
  • - Celebrate
  • - Bedroom
  • - Tower
  • - Bridge
Reproduced Results

4.1. Simplified gradient penalties

Our analysis suggests that the main effect of the zero-centered gradient penalties proposed by Roth et al. (2017) on local stability is to penalize the discriminator for deviating from the Nash-equilibrium. The simplest way to achieve this is to penalize the gradient on real data alone: when the generator distribution produces the true data distribution and the discriminator is equal to 0 on the data manifold, the gradient penalty ensures that the discriminator cannot create a non-zero gradient orthogonal to the data manifold without suffering a loss in the GAN game.

This leads to the following regularization term:

$$R_1(\psi) := \frac{\gamma}{2} E_{D(x)} \left[ \| \nabla D_\psi(x) \|^2 \right].$$  (9)
Reproduced Results

Loss with unregularized → Not Converge
Reproduced Results

Unregularized → not converging!
Conclusion and Future Work

Results we have so far:

- Negative hyperparameter: No convergence
- For eqn above: the second term has magic properties:
  - Near Nash Equilibrium, No rotation
  - Away from the Nash Equilibrium, transition from rotational convergence to non-convergence
- Convex combination of R1 and R2 have same convergence results.

\[
\lambda_{1/2} = -\frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} - f'(0)^2}
\]
Conclusion and Future Work (Cont.)

Conclusion Cont.

- Unregularized gradient based GAN optimization is not always locally convergent.
- WGANs and WGAN-GP do not always lead to local convergence whereas instance noise and zero-centered gradient penalties do.
- Local convergence achieved for simplified zero-centered gradient penalties under suitable assumptions.

Future Work

- Extend the theory to the non-realizable case (Not well understood or well-behaved to be modelled accurately)
References

- .... (http://proceedings.mlr.press/v80/mescheder18a/mescheder18a.pdf)
Working Split

• Kaiming Cheng: Coding, Model Training, Presentation

• Zijie Pan: Concept Research, Coding, Presentation

Thank you!