Some results on GAN dynamics

Ioannis Mitliagkas
Game dynamics are weird fascinating
Start with optimization dynamics
Optimization

$\theta^* \in \arg \min_{\theta \in \Theta} L^{(\theta)}(\theta)$

Smooth, differentiable cost function, L
→ Looking for stationary (fixed) points (gradient is 0)
→ Gradient descent
Optimization

\[ \mathbf{v}(\theta) = \nabla \mathcal{L}^{(\theta)}(\theta) \]

Conservative vector field

\[ \theta_{t+1} = \theta_t - \eta \mathbf{v}(\theta_t) \]
Gradient descent

\[ \nu(\theta) = \nabla \mathcal{L}^{(\theta)}(\theta) \]

Conservative vector field

\[ \theta_{t+1} = \theta_t - \eta \nu(\theta_t) \]

Straightforward dynamics

Fixed-point analysis

\[ F_\eta(\theta) = \theta - \eta \nu(\theta) \]

Jacobian of operator

\[ \nabla F_\eta(\theta) = I - \eta \nabla \nu(\theta) \]

Hessian of objective, L
Local convergence

**Theorem 1** (Prop. 4.4.1 Bertsekas [1999]). If the spectral radius $\rho_{\text{max}} \overset{\text{def}}{=} \rho(\nabla F_\eta(\omega^*)) < 1$, then, for $\omega_0$ in a neighborhood of $\omega^*$, the distance of $\omega_t$ to the stationary point $\omega^*$ converges at a linear rate of $O((\rho_{\text{max}} + \epsilon)^t)$, $\forall \epsilon > 0$.

Eigenvalues of op. Jacobian

$$\lambda(\nabla F_\eta(\theta)) = 1 - \eta \lambda(\nabla \nu(\theta))$$

If $\rho(\theta^*)=\max |\lambda(\theta^*)| < 1$, then fast local convergence

Jacobian of operator

$$\nabla F_\eta(\theta) = I - \eta \nabla \nu(\theta)$$

Hessian of objective, $L$

Symmetric, real-eigenvalues
Implicit generative models

- Generative moment matching networks [Li et al. 2017]
- Other, domain-specific losses can be used
- Variational AutoEncoders [Kingma, Welling, 2014]
- Autoregressive models (PixelRNN [van den Oord, 2016])
Generative Adversarial Networks

Both differentiable

Generator network, $G$
- Given latent code, $z$, produces sample $G(z)$

Discriminator network, $D$
- Given sample $x$ or $G(z)$, estimates probability it is real

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_x}[\log D(x)] + \mathbb{E}_{z \sim P_z}[\log(1 - D(G(z)))]$$
Generative Adversarial Networks

\[
\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_x}[\log D(x)] + \mathbb{E}_{z \sim P_z}[\log(1 - D(G(z)))]
\]
Games

Nash Equilibrium

\[ \theta^* \in \arg \min_{\theta \in \Theta} \mathcal{L}^{(\theta)}(\theta, \varphi^*) \]

\[ \varphi^* \in \arg \min_{\varphi \in \Phi} \mathcal{L}^{(\varphi)}(\theta^*, \varphi) \]

Smooth, differentiable $L$ → Looking for local Nash eq

→ Gradient descent
  → Simultaneous
  → Alternating
Game dynamics

\[ \mathbf{v}(\varphi, \theta) := \begin{bmatrix} \nabla_\varphi \mathcal{L}(\varphi)(\varphi, \theta) \\ \nabla_\theta \mathcal{L}(\theta)(\varphi, \theta) \end{bmatrix} \]

Non-conservative vector field

\[ \Rightarrow \]

Rotational dynamics

\[ F_\eta(\varphi, \theta) \overset{\text{def}}{=} [\varphi \quad \theta]^T - \eta \mathbf{v}(\varphi, \theta) \]
Game dynamics under gradient descent

\[ F_\eta(\varphi, \theta) \overset{\text{def}}{=} \left[ \begin{array}{c} \varphi \\ \theta \end{array} \right]^\top - \eta \, v(\varphi, \theta) \]

Jacobian is non-symmetric, with complex eigenvalues → Rotations in decision space

Games demonstrate rotational dynamics.
The Numerics of GANs
by Mescheder, Nowozin, Geiger
A word on notation and formulation

Maximization vs minimization

\[ \mathcal{L}^{(\phi)}(\phi, \theta) = -f(\phi, \theta) \]
\[ \mathcal{L}^{(\theta)}(\phi, \theta) = -g(\phi, \theta) \]

Warning: \( \mathcal{L} \neq L \)

Step size

\[ \eta = h \]
Eigen-analysis, gradient descent

**Theorem 1** (Prop. 4.4.1 Bertsekas [1999]). If the spectral radius $\rho_{\text{max}} \overset{\text{def}}{=} \rho(\nabla F_\eta(\omega^*)) < 1$, then, for $\omega_0$ in a neighborhood of $\omega^*$, the distance of $\omega_t$ to the stationary point $\omega^*$ converges at a linear rate of $O((\rho_{\text{max}} + \epsilon)^t)$, $\forall \epsilon > 0$.

$$\lambda(\nabla F_\eta(\phi, \theta)) = 1 - \eta \lambda(\nabla v(\phi, \theta))$$
The Numerics of GANs

\[ \lambda(\nabla F_\eta(\phi, \theta)) = 1 - \eta \lambda(\nabla v(\phi, \theta)) \]

(a) Illustration how the eigenvalues are projected into unit ball.
(b) Example where \( h \) has to be chosen extremely small.
(c) Illustration how our method alleviates the problem.
Make vector field “more conservative”

Idea 1: Minimize the norm of the gradient

\[ L(\phi, \theta) = \frac{1}{2} \| v(\phi, \theta) \|^2 \]
Idea 1: Minimize vector field norm

\[ L(\phi, \theta) = \frac{1}{2} \| v(\phi, \theta) \|^2 \]
Idea 2: use $L$ as regularizer

Non-conservative field $v$

Conservative field $-\nabla L$

Combined field $v - 0.6\nabla L$
Idea 2: use L as regularizer

Algorithm 2 Consensus optimization

1: \textbf{while} not converged \textbf{do}
2: \hspace{1em} v_\phi \leftarrow \nabla_\phi (f(\theta, \phi) - \gamma L(\theta, \phi))
3: \hspace{1em} v_\theta \leftarrow \nabla_\theta (g(\theta, \phi) - \gamma L(\theta, \phi))
4: \hspace{1em} \phi \leftarrow \phi + h v_\phi
5: \hspace{1em} \theta \leftarrow \theta + h v_\theta
6: \textbf{end while}
Idea 2: use $L$ as regularizer
Other ways to control these rotations?
Momentum (heavy ball, Polyak 1964)

\[ \theta_{t+1} = \theta_t - \eta \nabla \nu(\theta_t) + \beta (\theta_t - \theta_{t-1}) \]

Jacobian of momentum operator

\[ \nabla F_{\eta, \beta}(\theta_t, \theta_{t-1}) = \begin{bmatrix} I_n & 0_n \\ I_n & 0_n \end{bmatrix} - \eta \begin{bmatrix} \nabla \nu(\theta_t) & 0_n \\ 0_n & 0_n \end{bmatrix} + \beta \begin{bmatrix} I_n & -I_n \\ 0_n & 0_n \end{bmatrix} \]

Non-symmetric, with complex eigenvalues
→ Rotations in augmented state-space
Summary

Positive momentum can be bad for adversarial games

Practice that was very common when GANs were first invented.
   → Recent work reduced the momentum parameter.
   → Not an accident
Negative Momentum for Improved Game Dynamics

Gidel, Askari Hemmat, Pezeshki, Huang, Lepriol, Lacoste-Julien, Mitliagkas
AISTATS 2019
Our results

Negative momentum is optimal on simple bilinear game

Negative momentum values are locally preferrable near 0 on a more general class of games

Negative momentum is empirically best for certain zero sum games like “saturating GANs”
Momentum on games

Recall Polyak’s momentum (on top of simultaneous grad. desc.):

\[ x_{t+1} = x_t - \eta v(x_t) + \beta (x_t - x_{t-1}), \quad x_t = (\theta_t, \phi_t) \]

Fixed point operator requires a **state augmentation**:
(because we need previous iterate)

\[
F_{\eta, \beta}(x_t, x_{t-1}) := \begin{bmatrix}
I_n & 0_n \\
I_n & 0_n \\
\end{bmatrix}
\begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix} - \eta \begin{bmatrix} v(x_t) \\ 0_n \end{bmatrix} + \beta \begin{bmatrix} I_n & -I_n \\
0_n & 0_n \end{bmatrix}
\begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix}
\]
### Bilinear game

\[ \min_\theta \max_\phi \theta^\top A\phi \]

<table>
<thead>
<tr>
<th>Method</th>
<th>( \beta )</th>
<th>Bounded</th>
<th>Converges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simultaneous</td>
<td>( \beta \in \mathbb{R} )</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Alternated</td>
<td>&gt;0</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>&lt;0</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
“Proof by picture”

Gradient descent
→ Simultaneous
→ Alternating

Momentum
→ Positive
→ Negative
General games
Eigen-analysis, 0 momentum

**Theorem 1** (Prop. 4.4.1 Bertsekas [1999]). If the spectral radius $\rho_{\text{max}} \overset{\text{def}}{=} \rho(\nabla F_\eta(\omega^*)) < 1$, then, for $\omega_0$ in a neighborhood of $\omega^*$, the distance of $\omega_t$ to the stationary point $\omega^*$ converges at a linear rate of $O((\rho_{\text{max}} + \epsilon)^t)$, $\forall \epsilon > 0$.

$$\lambda(\nabla F_\eta(\phi, \theta)) = 1 - \eta \lambda(\nabla v(\phi, \theta))$$
Zero vs negative momentum

Momentum
→ Zero
→ Negative
Theorem 3. The eigenvalues of $\nabla F_{\eta, \beta}(\phi^*, \theta^*)$ are

$$
\mu_{\pm}(\beta, \eta, \lambda) := (1 - \eta \lambda + \beta) \frac{1 \pm \Delta^{\frac{1}{2}}}{2},
$$

where $\Delta := 1 - \frac{4 \beta}{(1 - \eta \lambda + \beta)^2}$, $\lambda \in \text{Sp}(\nabla v(\phi^*, \theta^*))$ and $\Delta^{\frac{1}{2}}$ is the complex square root of $\Delta$ with positive real part\(^3\). Moreover we have the following Taylor approximation,

$$
\mu_+(\beta, \eta, \lambda) = 1 - \eta \lambda - \beta \frac{\eta \lambda}{1 - \eta \lambda} + O(\beta^2) \quad \text{and} \quad \mu_-(\beta, \eta, \lambda) = \frac{\beta}{1 - \eta \lambda} + O(\beta^2)
$$

\(^3\) If $\Delta$ is a negative real number we set $\Delta^{\frac{1}{2}} := i \sqrt{-\Delta}$
Empirical results
What happens in practice?

Fashion MNIST:
What happens in practice?

CIFAR-10:
Negative Momentum

To sum up:

- Negative momentum seems to improve the behaviour due to “bad” eigenvalues.
- Optimal for a class of games
- Empirically optimal on “saturating” GANs