Mila

Some results on GAN dynamics

Ioannis Mitliagkas

Game dynamics are ~~weird~~
fascinating

Start with optimization dynamics

Optimization

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

Smooth, differentiable cost function, L

→ Looking for stationary (fixed) points (gradient is 0)

→ Gradient descent

Optimization

$$v(\boldsymbol{\theta}) = \nabla \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta})$$

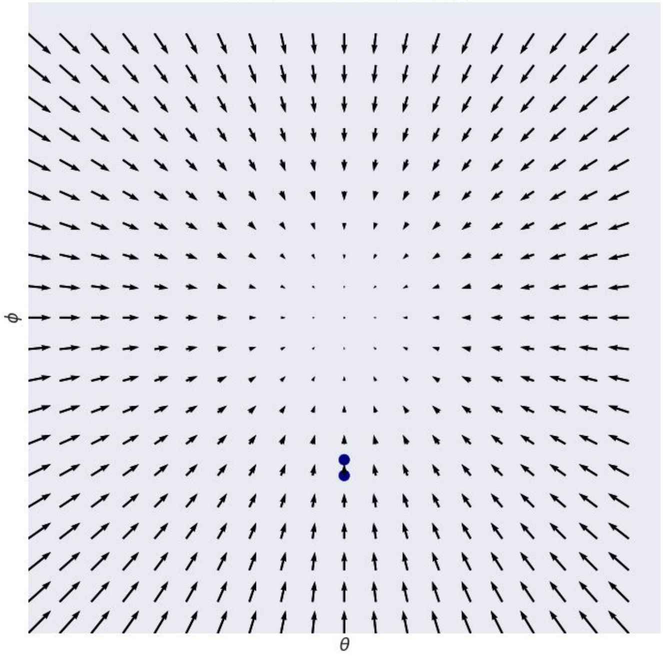
Conservative vector field

→

Straightforward dynamics

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{v}(\boldsymbol{\theta}_t)$$

Ferenc Huszar



Gradient descent

$$\boldsymbol{v}(\boldsymbol{\theta}) = \nabla \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta})$$

Conservative vector field

→

Straightforward dynamics

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{v}(\boldsymbol{\theta}_t)$$

Fixed-point analysis

$$F_\eta(\boldsymbol{\theta}) = \boldsymbol{\theta} - \eta \boldsymbol{v}(\theta)$$

Jacobian of operator

$$\nabla F_\eta(\boldsymbol{\theta}) = I - \eta \underline{\nabla \boldsymbol{v}(\theta)}$$

Hessian of objective, L

Mila

Université de Montréal

Local convergence

Theorem 1 (Prop. 4.4.1 Bertsekas [1999]). *If the spectral radius* $\rho_{\max} \stackrel{def}{=} \rho(\nabla F_\eta(\boldsymbol{\omega}^*)) < 1$, *then, for* $\boldsymbol{\omega}_0$ *in a neighborhood of* $\boldsymbol{\omega}^*$, *the distance of* $\boldsymbol{\omega}_t$ *to the stationary point* $\boldsymbol{\omega}^*$ *converges at a linear rate of* $\mathcal{O}\left((\rho_{\max} + \epsilon)^t\right)$, $\forall \epsilon > 0$.

Eigenvalues of op. Jacobian

$$\lambda(\nabla F_\eta(\boldsymbol{\theta})) = 1 - \eta\lambda(\nabla \boldsymbol{v}(\theta))$$

If ρ(θ*)=max |λ(θ*)|<1, then fast local convergence

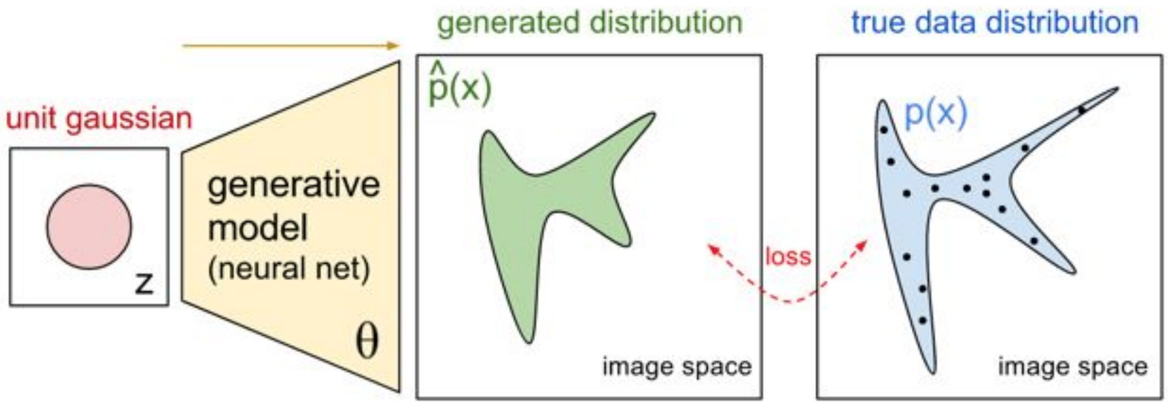
Jacobian of operator

$$\nabla F_\eta(\boldsymbol{\theta}) = I - \eta \underline{\nabla \boldsymbol{v}(\theta)}$$

Hessian of objective, L
Symmetric, real-eigenvalues

Games

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

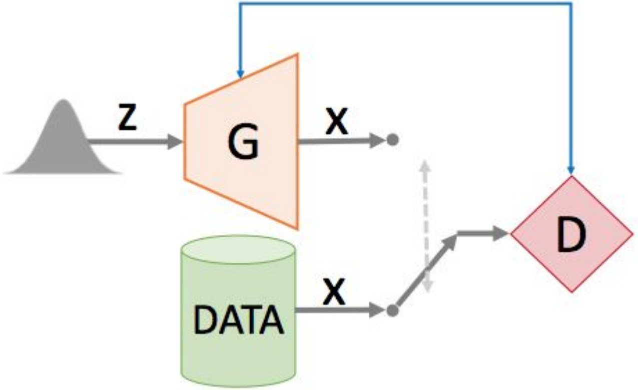
Both differentiable

Generator network, G

Given latent code, z, produces sample G(z)

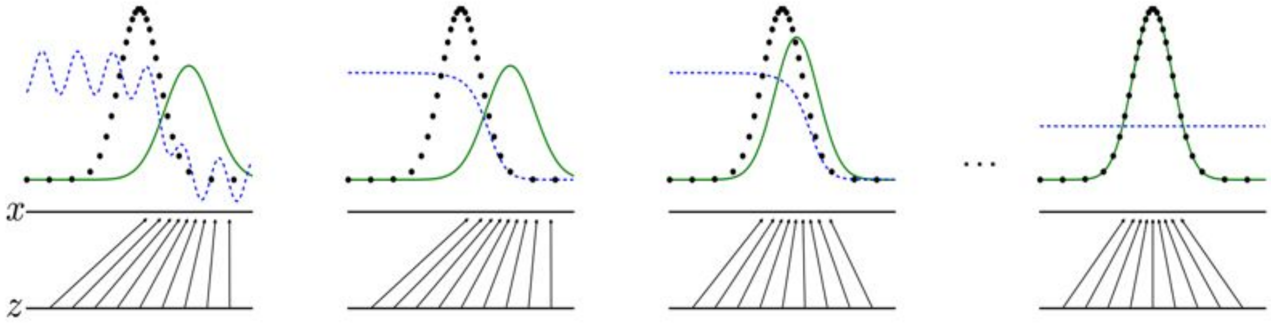
Discriminator network, D

Given sample x or G(z), estimates probability it is real



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

Generative Adversarial Networks



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

Games

Nash Equilibrium

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta}, \boldsymbol{\varphi}^*)$$

$$\boldsymbol{\varphi}^* \in \arg\min_{\boldsymbol{\varphi}\in\boldsymbol{\varphi}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\theta}^*, \boldsymbol{\varphi})$$

Smooth, differentiable L
→ Looking for local Nash eq

→ Gradient descent
- → **Simultaneous**
- → **Alternating**

Game dynamics

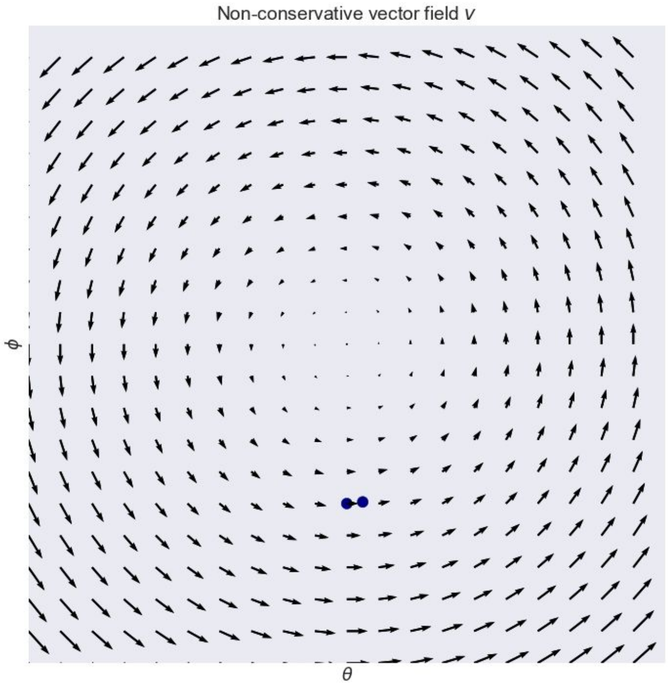
$$\boldsymbol{v}(\boldsymbol{\varphi}, \boldsymbol{\theta}) := \begin{bmatrix} \nabla_{\boldsymbol{\varphi}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\varphi}, \boldsymbol{\theta}) \\ \nabla_{\boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\varphi}, \boldsymbol{\theta}) \end{bmatrix}$$

Non-conservative vector field

→

Rotational dynamics

$$F_\eta(\boldsymbol{\varphi}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \begin{bmatrix} \boldsymbol{\varphi} & \boldsymbol{\theta} \end{bmatrix}^\top - \eta\, \boldsymbol{v}(\boldsymbol{\varphi}, \boldsymbol{\theta})$$

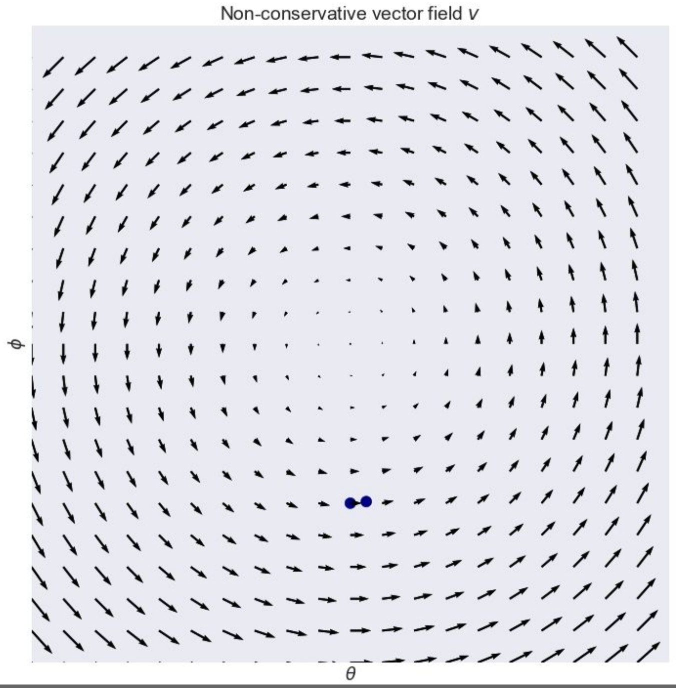


Game dynamics under gradient descent

$$F_\eta(\varphi, \theta) \stackrel{\text{def}}{=} [\varphi \quad \theta]^\top - \eta\, v(\varphi, \theta)$$

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

$$\mathcal{L}^{(\boldsymbol{\phi})}(\boldsymbol{\phi}, \boldsymbol{\theta}) = -f(\boldsymbol{\phi}, \boldsymbol{\theta})$$
$$\mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\phi}, \boldsymbol{\theta}) = -g(\boldsymbol{\phi}, \boldsymbol{\theta})$$

Warning: $\mathcal{L} \neq L$

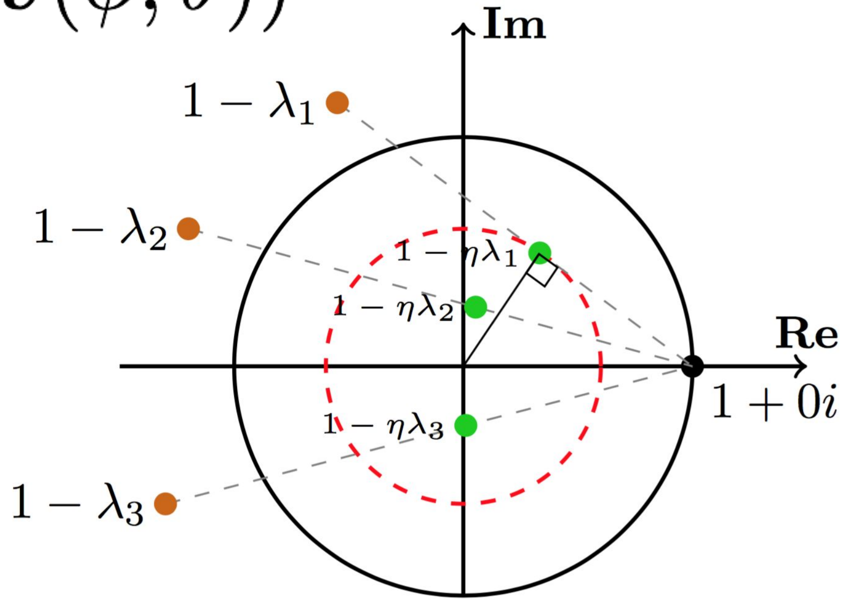
Step size

$$\eta = h$$

Eigen-analysis, gradient descent

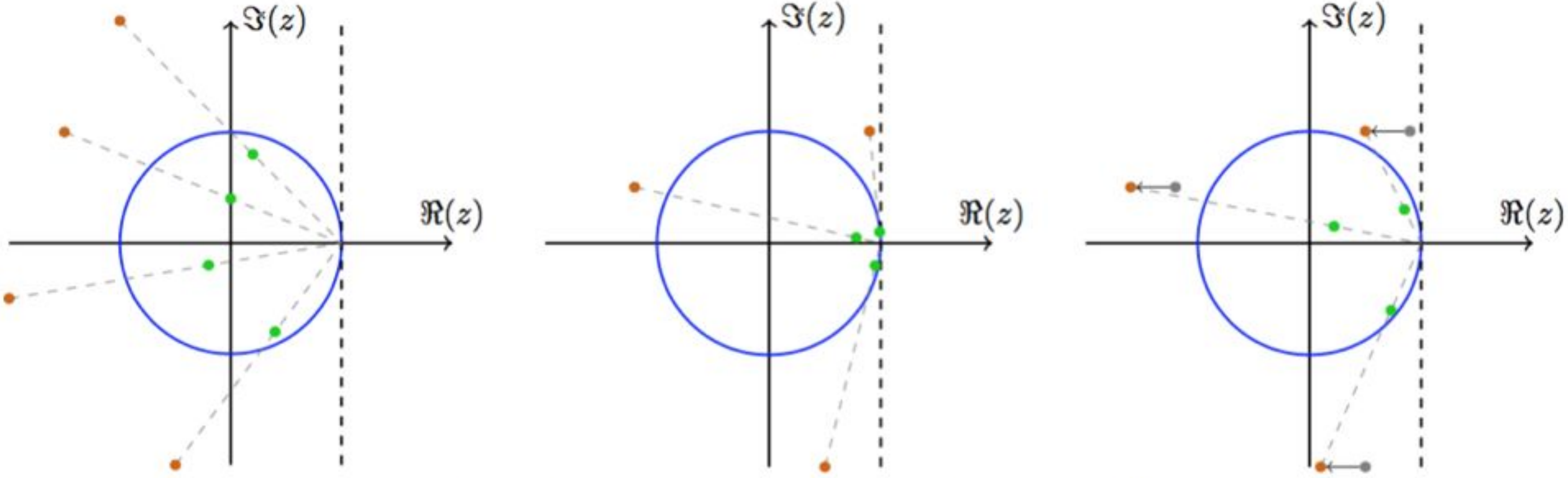
Theorem 1 (Prop. 4.4.1 Bertsekas [1999]). *If the spectral radius* $\rho_{\max} \stackrel{def}{=} \rho(\nabla F_\eta(\boldsymbol{\omega}^*)) < 1$*, then, for* $\boldsymbol{\omega}_0$ *in a neighborhood of* $\boldsymbol{\omega}^*$*, the distance of* $\boldsymbol{\omega}_t$ *to the stationary point* $\boldsymbol{\omega}^*$ *converges at a linear rate of* $\mathcal{O}\left((\rho_{\max} + \epsilon)^t\right)$, $\forall \epsilon > 0$.

$$\lambda(\nabla F_\eta(\boldsymbol{\phi}, \boldsymbol{\theta})) = 1 - \eta\lambda(\nabla \boldsymbol{v}(\phi, \theta))$$



Mila

The Numerics of GANs



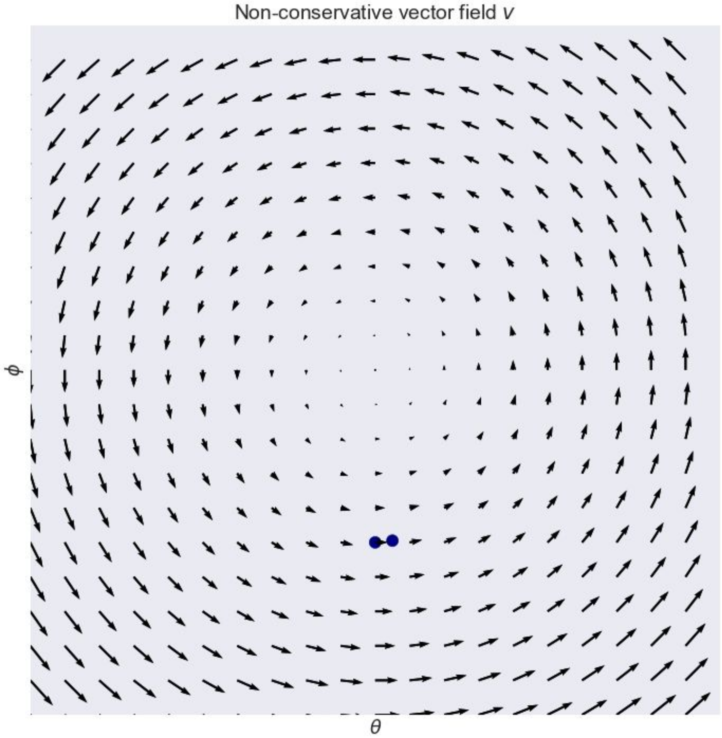
(a) Illustration how the eigenvalues are projected into unit ball.

(b) Example where h has to be chosen extremely small.

(c) Illustration how our method alleviates the problem.

$$\lambda(\nabla F_\eta(\boldsymbol{\phi}, \boldsymbol{\theta})) = 1 - \eta\lambda(\nabla \boldsymbol{v}(\phi, \theta))$$

Make vector field “more conservative”

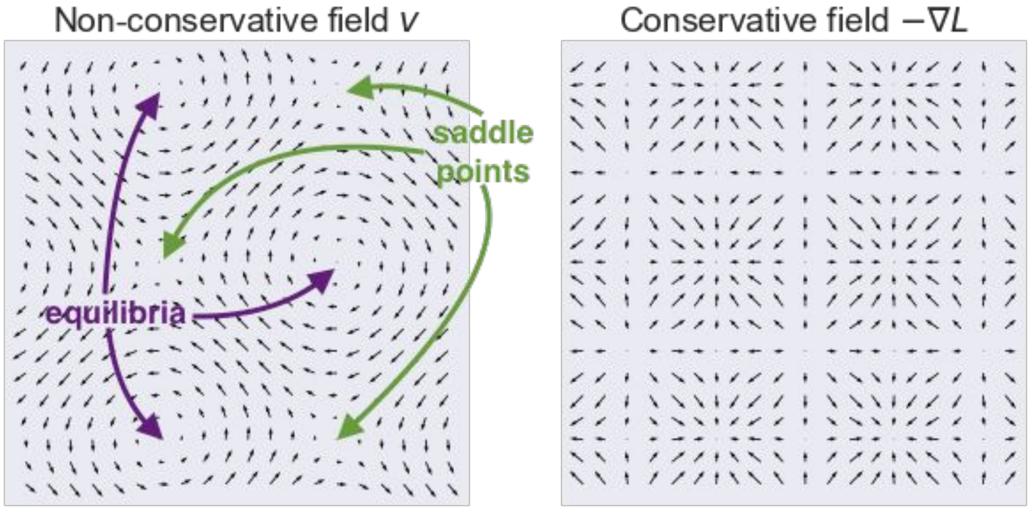


Idea 1: Minimize the norm of the gradient

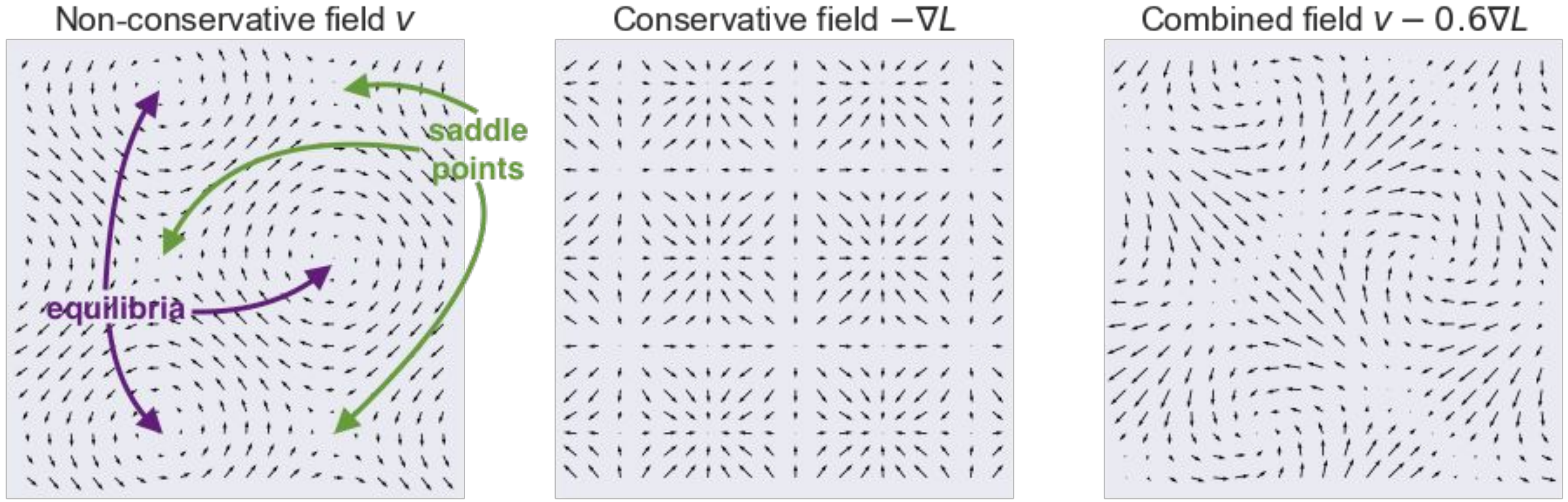
$$L(\boldsymbol{\phi}, \boldsymbol{\theta}) = \frac{1}{2}\|v(\boldsymbol{\phi}, \boldsymbol{\theta})\|^2$$

Idea 1: Minimize vector field norm

$$L(\boldsymbol{\phi}, \boldsymbol{\theta}) = \frac{1}{2} \|v(\boldsymbol{\phi}, \boldsymbol{\theta})\|^2$$



Idea 2: use L as regularizer

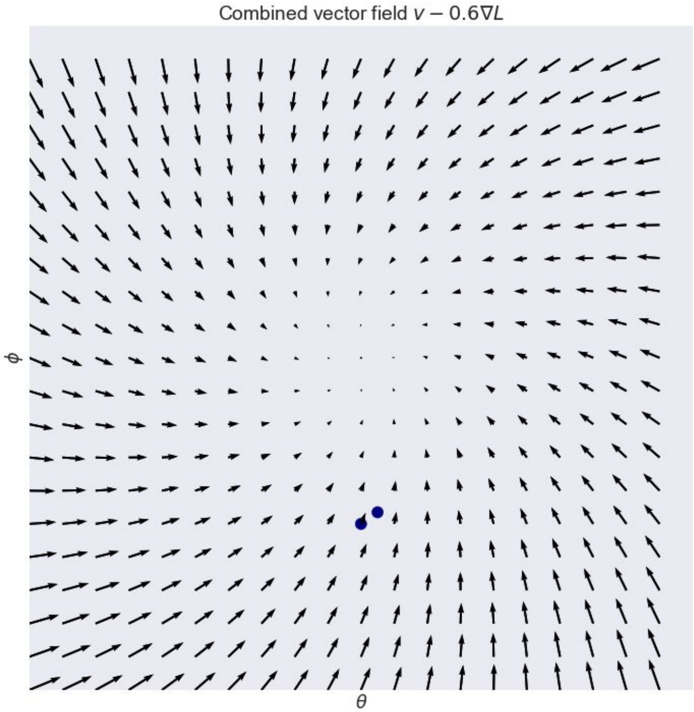


Idea 2: use L as regularizer

Algorithm 2 Consensus optimization

1: **while** not converged **do**
2: $\quad v_\phi \leftarrow \nabla_\phi(f(\theta, \phi) - \gamma L(\theta, \phi))$
3: $\quad v_\theta \leftarrow \nabla_\theta(g(\theta, \phi) - \gamma L(\theta, \phi))$
4: $\quad \phi \leftarrow \phi + h v_\phi$
5: $\quad \theta \leftarrow \theta + h v_\theta$
6: **end while**

Idea 2: use L as regularizer



Other ways to control these rotations?

Momentum (heavy ball, Polyak 1964)

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{v}(\boldsymbol{\theta}_t) + \beta(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})$$

Jacobian of momentum operator

$$\nabla F_{\eta,\beta}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}) = \begin{bmatrix} \boldsymbol{I}_n & \mathbf{0}_n \\ \boldsymbol{I}_n & \mathbf{0}_n \end{bmatrix} - \eta \begin{bmatrix} \nabla \boldsymbol{v}(\boldsymbol{\theta}_t) & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{0}_n \end{bmatrix} + \beta \begin{bmatrix} \boldsymbol{I}_n & -\boldsymbol{I}_n \\ \mathbf{0}_n & \mathbf{0}_n \end{bmatrix}$$

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.):

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \boldsymbol{v}(\boldsymbol{x}_t) + \beta(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}), \quad \boldsymbol{x}_t = (\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)$$

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

$$F_{\eta,\beta}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) := \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0}_n \\ \boldsymbol{I}_n & \boldsymbol{0}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix} - \eta \begin{bmatrix} \boldsymbol{v}(\boldsymbol{x}_t) \\ \boldsymbol{0}_n \end{bmatrix} + \beta \begin{bmatrix} \boldsymbol{I}_n & -\boldsymbol{I}_n \\ \boldsymbol{0}_n & \boldsymbol{0}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix}$$

Bilinear game

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\varphi}} \ \boldsymbol{\theta}^\top \boldsymbol{A} \boldsymbol{\varphi}$$

Method	β	Bounded	Converges
Simultaneous	$\beta \in \mathbb{R}$	✗	✗
Alternated	>0	✗	✗
	0	✓	✗
	<0	✓	✓

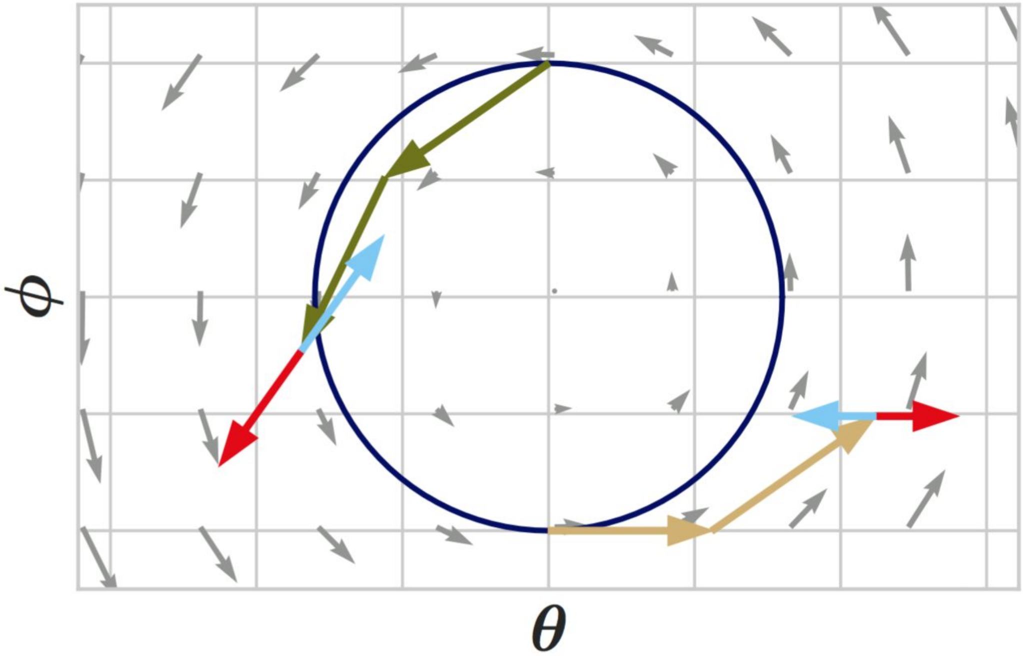
“Proof by picture”

Gradient descent

- → Simultaneous
- → Alternating

Momentum

- → Positive
- → Negative

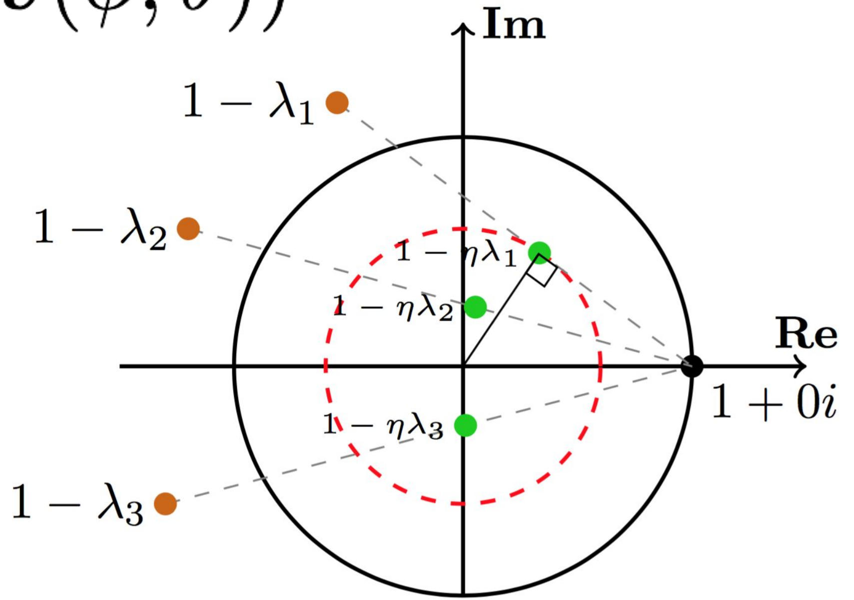


General games

Eigen-analysis, 0 momentum

Theorem 1 (Prop. 4.4.1 Bertsekas [1999]). *If the spectral radius $\rho_{\max} \stackrel{def}{=} \rho(\nabla F_\eta(\boldsymbol{\omega}^*)) < 1$, then, for $\boldsymbol{\omega}_0$ in a neighborhood of $\boldsymbol{\omega}^*$, the distance of $\boldsymbol{\omega}_t$ to the stationary point $\boldsymbol{\omega}^*$ converges at a linear rate of* $\mathcal{O}\big((\rho_{\max} + \epsilon)^t\big)$, $\forall \epsilon > 0$.

$$\lambda(\nabla F_\eta(\boldsymbol{\phi}, \boldsymbol{\theta})) = 1 - \eta\lambda(\nabla \boldsymbol{v}(\phi, \theta))$$

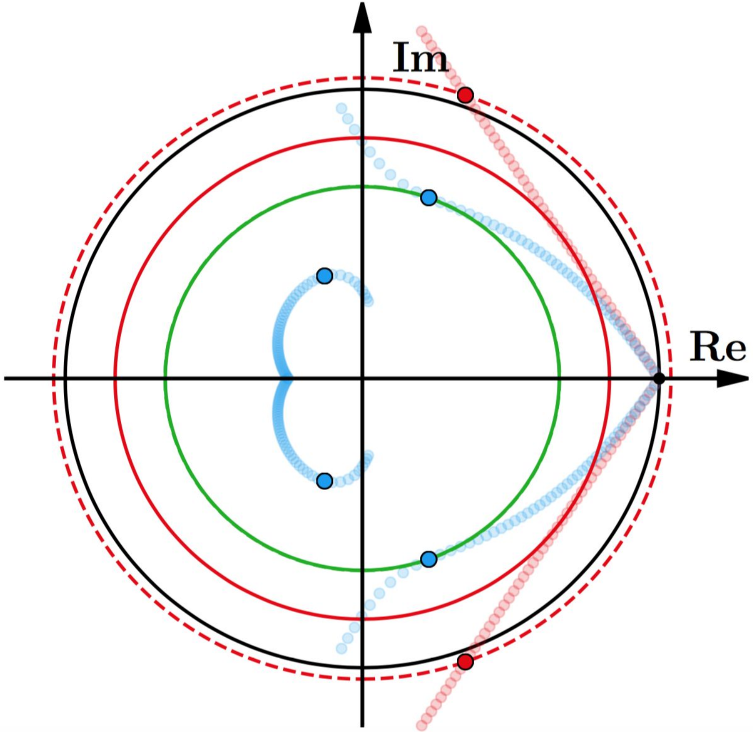


Mila

Zero vs negative momentum

Momentum

- **Zero**
- **Negative**



Negative Momentum

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

$$\mu_\pm(\beta, \eta, \lambda) := (1 - \eta\lambda + \beta)\frac{1 \pm \Delta^{\frac{1}{2}}}{2}, \tag{9}$$

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

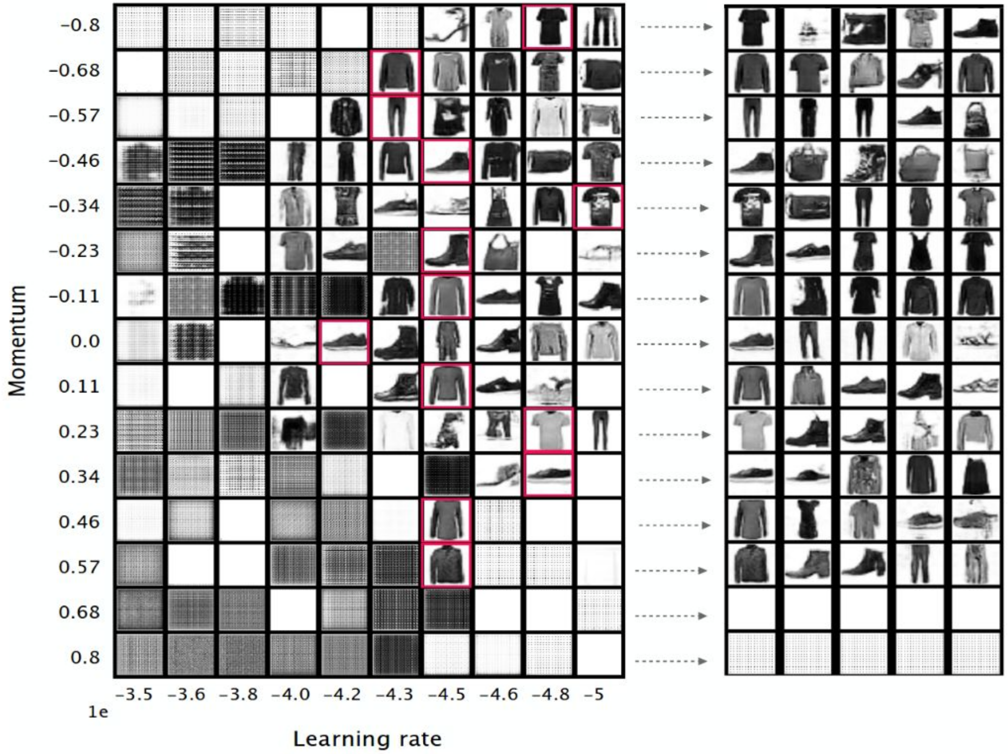
$$\mu_+(\beta, \eta, \lambda) = 1 - \eta\lambda - \beta\frac{\eta\lambda}{1 - \eta\lambda} + O(\beta^2) \quad \textit{and} \quad \mu_-(\beta, \eta, \lambda) = \frac{\beta}{1 - \eta\lambda} + O(\beta^2) \tag{10}$$

[3] If Δ 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