Lecture 17: Spectral Analysis and Stability

Start Recording!

Reminders

- Office Hours tomorrow with Adrien (12-1PM)
- Talks this Friday.

References for this lecture:

1. Mescheder, Lars, Sebastian Nowozin, and Andreas Geiger. "The numerics of gans." *Neurips* (2017).
2. Gidel, Gauthier, et al. "Negative momentum for improved game dynamics." *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019.
3. Azizian, Waïss, et al. "A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games." International Conference on Artificial Intelligence and Statistics. PMLR, 2020.

Today: General tools to analyse convergence AND stability of gradient based methods

Variational Inequality Perspective

We only 'care' about the gradient-based updates, i.e., the vector field:

$$F(\theta_t, \phi_t) := \begin{pmatrix} \nabla_\theta \mathcal{L}(\theta_t, \varphi_t) \\ -\nabla_\phi \mathcal{L}(\theta_t, \varphi_t) \end{pmatrix}$$

Previous plots. We represented the joint space (θ_t, ϕ_t)
More compact formalism:

$$\omega_t := (\theta_t, \phi_t)$$

Variational Inequality Perspective

Goal: Find a stationary point of the vector field:

$$F(\omega^*) = 0$$

In zero sum game: Equivalent to find a point with 0 gradient for each player

If the game is convex concave: equivalent to find a Nash!

Gradient ~~Descent~~ Method

Update rule:

$$\omega_{t+1} = \omega_t - \eta F(\omega_t)$$

What we will look at:

$$\underbrace{\|\omega_t - \omega^*\|^2}$$

Arbitrary norm !!!

Spectral Analysis

$$\begin{aligned}\|\omega_{t+1} - \omega^*\| &= \|\omega_t - \omega^* - \eta(F(\omega_t) - F(\omega^*))\| \\ &\approx \|\omega_t - \omega^* - \eta \nabla F(\omega^*)(\omega_t - \omega^*)\| \\ &\lesssim \underbrace{\|I_d - \eta \nabla F(\omega^*)\|}_{\text{Matrix norm induced by } \|\,\|} \|\omega_t - \omega^*\|\end{aligned}$$

$$\lesssim \|I_d - \eta \nabla F(\omega^*)\| \|\omega_t - \omega^*\|$$

Classical Result on Matrix Norm and Spectral Radius

We have:

$$\|\omega_{t+1} - \omega^*\| \lesssim \|I_d - \eta \nabla F(\omega^*)\| \|\omega_t - \omega^*\|$$

For any matrix A, there exists a norm such that:

$$\|A\| \approx \rho(A) := \sup\{|\lambda| \,:\, \lambda \in Sp(A)\}$$

Thus:

$$\|\omega_{t+1} - \omega^*\| \lesssim \boxed{\rho(I_d - \eta \nabla F(\omega^*))} \|\omega_t - \omega^*\|$$

Spectral Radius and Norm

Questions (Bo Wen Peng and Martin Dalaire): How do we prove this

$$\|A\| \approx \rho(A) := \sup\{|\lambda| \ : \ \lambda \in Sp(A)\}$$

Idea:

$$\|A^k\|_2^{1/k} \to \rho(A)$$

Theorem

$$\rho := \rho(I_d - \eta \nabla F(\omega^*))$$

Theorem:

1. If $\rho < 1$ then for any **ε**>0, there exists a constant C such that for:

Note

2. If ρ … the

opti…

Olivier Ethier: What is the criterion to know if we initialize close enough to w*?

Carl Perreault-Lafleur:
How are we assured that the norm for which the convergence theorem (slide 10) holds is relevant? ie. we would be happy to converge w.r.t. l2 norm, but what if it converges w.r.t. a weird norm?

Conclusion

Connection between :

- **convergence** (numerical analysis) and
- **eigenvalues** (spectral analysis)

Quantity of interest:

$$\rho(I_d - \eta \nabla F(\omega^*)) = \max\{\underbrace{|1 - \eta\lambda|}_{\text{This has to be smaller than 1}} \;:\; \lambda \in Sp(\underbrace{\nabla F(\omega^*)}_{\text{Jacobian of the Vector Field at the optimum}})\}$$

Spectral radius

This has to be smaller than 1

Jacobian of the Vector Field at the optimum

First idea

$$|1 - \eta\lambda|^2 = 1 - 2\eta\Re(\lambda) + \eta^2|\lambda|^2$$
$$\approx 1 - 2\eta\Re(\lambda)$$

For a small step size

Reminder: We want this quantity to be < 1!!!!

Strongly Convex function: Positive Hessian.
This: Generalization of (strong) convexity for games

Conclusion: We need $\Re(\lambda) > 0 \, , \forall\lambda \in Sp(\nabla F(\omega^*))$

Visualization

$\Re(\lambda) > 0 \implies \lambda$ is there

Changing η moves • to • on the dotted lines

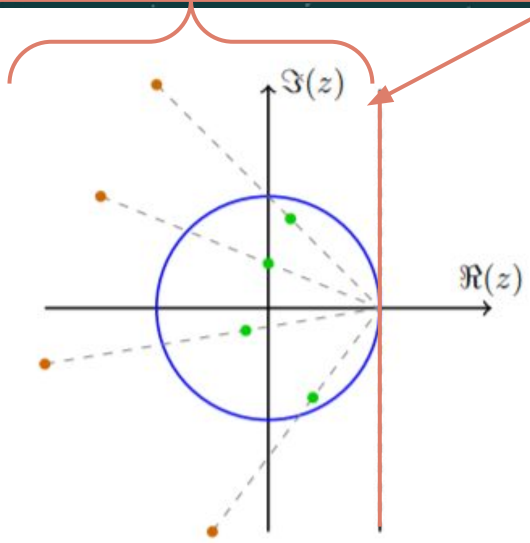


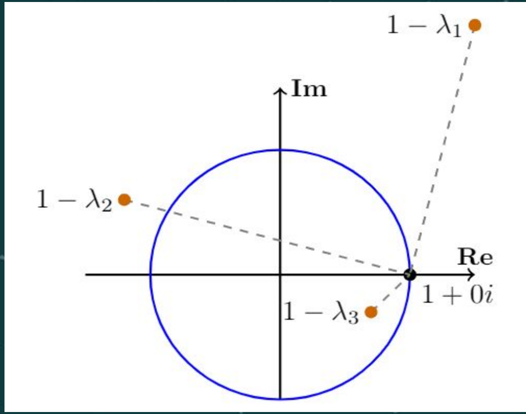
Figure from Mescheder, et al 2018

Goal: Move **all** the • inside the **blue circle (norm =1)**

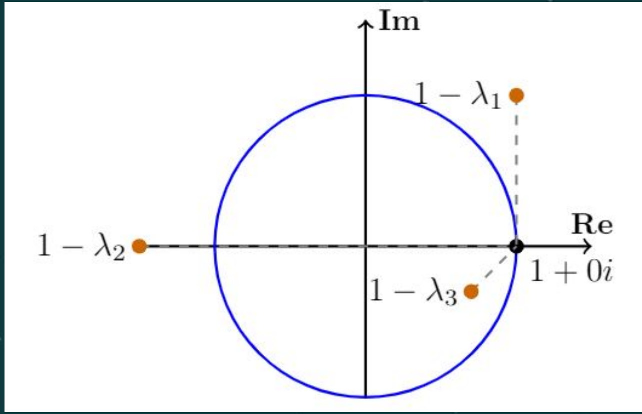
Fact: with η small enough it will happen.

Three sets of eigenvalues:

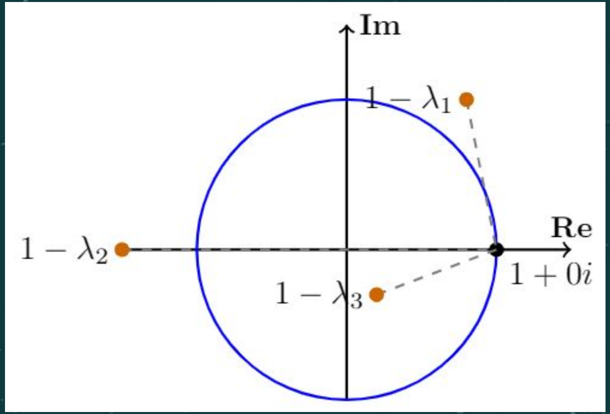
Example 1



Example 2



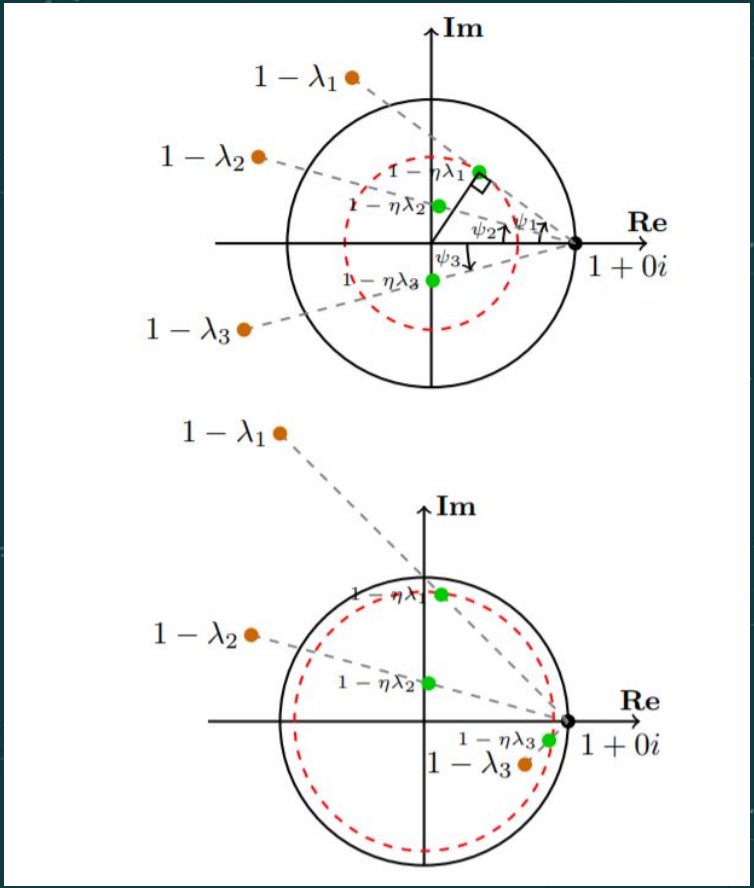
Example 3



Interpretation: The convergence rate is given by the radius of the red circle.

Problem: for each matrix,

we need to find the **best** step size η

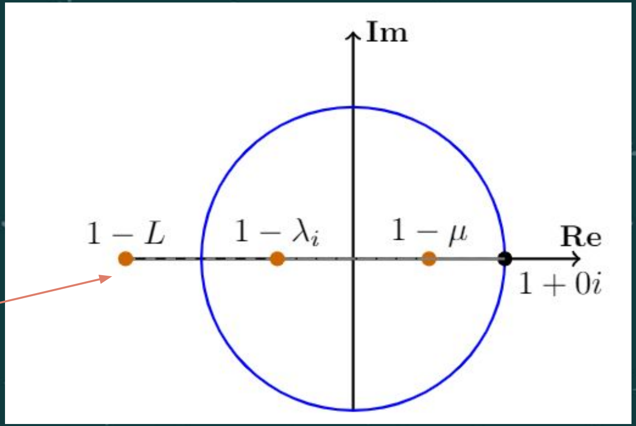


Special case: Gradient Descent

Gradient Descent:

$$\nabla F(\omega) = \nabla^2 g(\omega)$$

Symmetric matrix => **real eigenvalues** !!



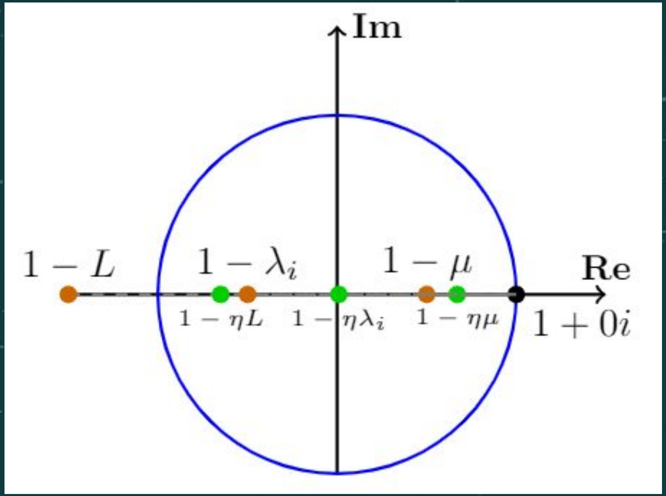
Question (several students) : why does the Hessian has to have real eigenvalues?

$$\|Av\|_2^2 = \langle Av, Av \rangle = (Av)^* Av = \lambda \|v\|^2 \in \mathbb{R}$$

Theorem for Gradient Descent

Problem: find that optimal η

$$\min_{\eta} \max_{1 \le i \le n} |1 - \eta\lambda_i|^2$$



$$\eta^* = \frac{2}{L + \mu}$$

⇒

$$\rho(I_d - \eta^* \nabla g(\omega^*)) = 1 - \frac{2\mu}{L + \mu}$$

Lecture on Optimization

Strongly convex and Lipschitz function: (numerical proof)

Condition number: The quantity of interest for convergence speed.

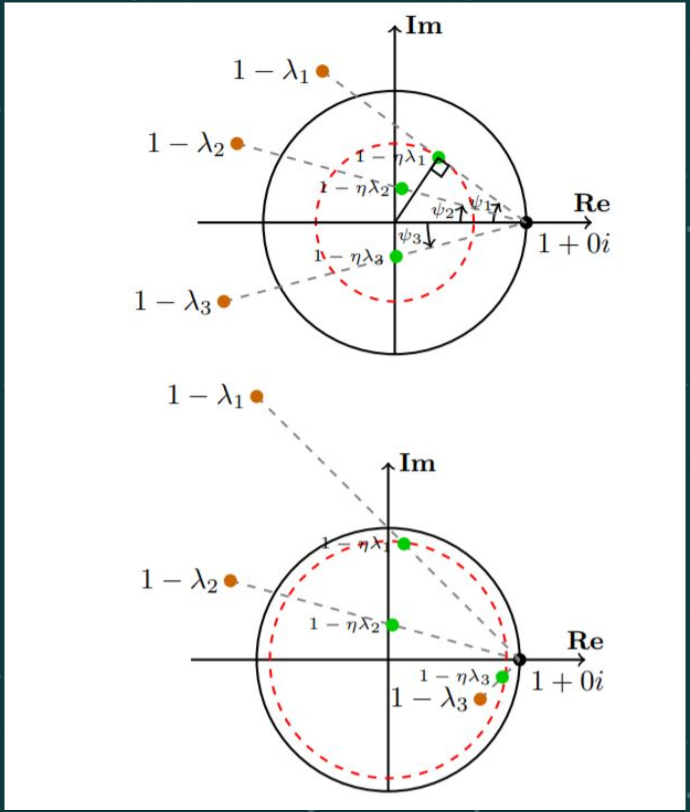
$$g(\theta_t) - g^* \leq (1 - \frac{\mu}{L})^t (g(\theta_0) - g^*)$$

Question (Elio) : What about the non-smooth case? Does the bounds of the spectral radius still hold in the case of non-smooth convex game ?

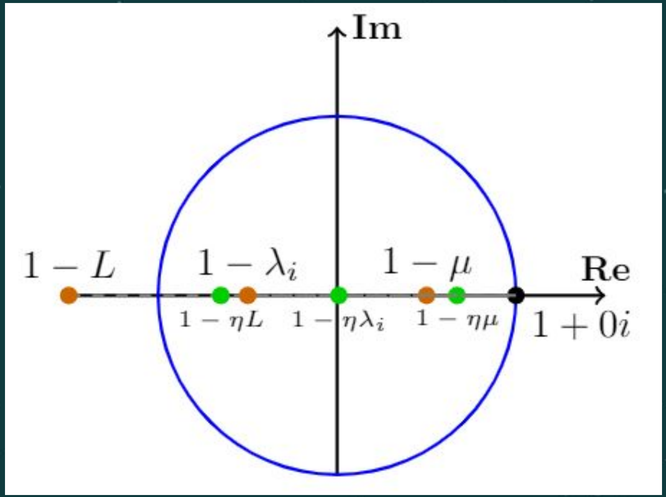
Strongly convex and Lipschitz function: (spectral proof)

$$\|\omega_t - \omega^*\|^2 \leq (1 - \frac{2\mu}{L + \mu})^t \|\omega_0 - \omega^*\|^2$$

Why are games more challenging to optimize (and analyze)?



Imaginary (games) vs Real eigenvalues (minimization)



Theorem

Theorem 3. *Let ω^* be a stationary point of v and denote by σ^* the spectrum of $\nabla v(\omega^*)$. If the eigenvalues of $\nabla v(\omega^*)$ all have positive real parts, then*

(i). (Gidel et al., 2019b) For $\eta = \min_{\lambda \in \sigma^*} \Re(1/\lambda)$, the spectral radius of F_η can be upper-bounded as

$$\varrho^2 \leq 1 - \min_{\lambda \in \sigma^*} \Re(1/\lambda) \min_{\lambda \in \sigma^*} \Re(\lambda) \, .$$

(ii). For all $\eta > 0$, the spectral radius of the gradient operator F_η at ω^* is lower bounded by

$$\varrho^2 \geq 1 - 4 \min_{\lambda \in \sigma^*} \Re(1/\lambda) \min_{\lambda \in \sigma^*} \Re(\lambda) \, .$$

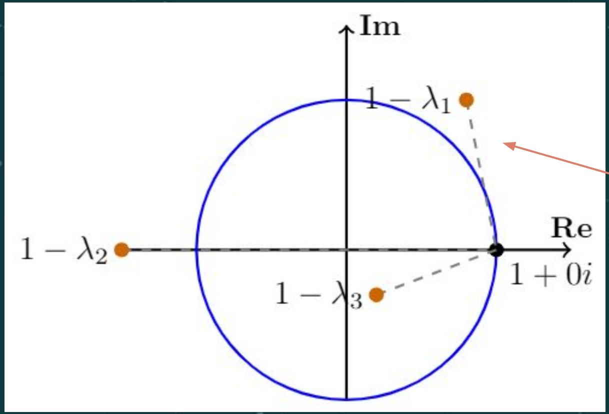
$$v(\omega) = F(\omega)$$

Intuition:

$$\rho(I_d - \eta^* F(\omega^*)) \approx 1 - \underbrace{\min_i \Re(1/\lambda_i)}_{\text{Equivalent of } 1/L} \underbrace{\min_i \Re(\lambda_i)}_{\text{Equivalent of } \backslash mu}$$

Equivalent of 1/L

Equivalent of \mu



$$\Re(1/\lambda) = \frac{\Re(\lambda)}{|\lambda|^2}$$

Conclusion

We have powerful tool to analyze **local convergence** of games using spectral analysis.

- The bian of the

Charlie Gauthier:
How can this be used in practice?

- For **minimization**, the Jacobian is a Hessian (thus only has **real** eigenvalues).
- The (sufficient) condition for local convergence was to have only eigenvalues with **positive real part.**