IFT 6756 - Lecture 17
Spectral Analysis and Stability

This version of the notes has not yet been thoroughly checked. Please report any bugs to the scribes or instructor.

Scribes **Instructor:** Gauthier Gidel
Winter 2021: Uros Petricevic and Olivier Ethier

1 Summary

In the previous lecture, we introduced a new method referred to as extragradient. This method serves as an explicit approximation of the proximal point method, which is implicit and thus hard to implement. Then we defined montonicity and strong monotonicity for vector fields and found the convergence rates of extragradient in each case. Moreover, we introduced a variant of extragradient, the optimistic method, and found its convergence rates for monotone and strongly monotone vector fields as well.

In this lecture we see how it is possible to look at the spectral radius of a particular matrix to assess the convergence of a given vector field. Needless to say, this particular matrix is constructed from the derivative of the vector field itself.

2 Variational Inequality Perspective (reminder)

One way to formulate the minmax optimization problem is to concatenate the individual gradients. More precisely, we concatenate the gradient of the first player and minus the gradient of the second player as we can see in definition 1.

Definition 1. $F(\theta_t, \phi_t)$

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

Note that we will refer to the joint space (θ_t, ϕ_t) as:

$$w_t = (\theta_t, \phi_t)$$

2.1 Goal

Using this formulation, the minmax problem is solved at a stationary point of the vector field $F(\theta_t, \phi_t)$. We are looking for w^* such that :

$$F(w^*) = 0$$

In zero sum game, this is equivalent to finding a point where the gradient is 0 for each player. Furthermore, if the game is convex concave, this is equivalent to find a Nash equilibrium. Even in more challenging games where we are not sure to have a Nash equilibrium, we can still hope for a local Nash equilibrium.

3 Gradient Method

For standard Gradient method, the update rule is

$$w_{t+1} = w_t - \eta F(w_t) \tag{1}$$

We can be tempted to compare this to a gradient descent update, but doing so might cause some confusion due to our intuitive understanding of the word ”descent”. This update rule is indeed a descent of the loss surface w.r.t. parameters θ, but it is also an ascent of the same surface w.r.t. parameters ϕ. Remember, this is a minmax problem.
In order to assess the convergence, we look at the distance to the optimum:

$$||w_t - w^*||^2$$

Notice that we did not make any assumption regarding the norm. Hence, it is an arbitrary norm.

4 Spectral Analysis

We look at $||w_{t+1} - w^*||$ in order to see if the distance is increasing or decreasing to the optimum. By replacing with the update rule, we get:

$$||w_t - w^*|| = ||w_t - w^* - \eta(F(w_t) - F(w*))|| \tag{2}$$

Let's remember that by definition $F(w*) = \mathbf{0}$. Also, if w_t is close to the optimum w^*, we can approximate $F(w_t)$ with its first order Taylor expansion around w^*:

$$\approx ||w_t - w^* - \eta\nabla F(w*)(w_t - w*)|| = ||(I_d - \eta\nabla F(w*))(w_t - w*)||$$

From here we simply use the submultiplicative property of the norm to find an upper bound. We finally have:

$$||w_{t+1} - w^*|| \lesssim ||I_d - \eta\nabla F(w*)||\ ||(w_t - w*)|| \tag{3}$$

This result shows that if $||I_d - \eta\nabla F(w*)|| \leq 1$, the distance to the optimum will decrease at each step, leading to convergence. Therefore, we are interested in finding a simple way to verify if $||I_d - \eta\nabla F(w*)|| \leq 1$.

4.1 Introducing the Spectral Radius

For any matrix A, there exists a norm such that the norm of the matrix A is approximately equal to the spectral radius of A:

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

Remark on spectral radius

The spectral radius is not a matrix norm.

Proof. Let's consider the counterexample

$$M = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

The greatest eigenvalue of M is $\rho(M) = 0$. By definition of any matrix norm, $||M|| = 0 \iff M = 0$, but in this case $M \neq 0$. Therefore, the spectral radius of a matrix is not a norm. □

Note that this proof is valid for any nilpotent matrix $M : M^k = 0$ for all integer k.

We can use this property to rewrite equation 3:

$$||w_{t+1} - w^*|| \lesssim \rho(I_d - \eta\nabla F(w*))\ ||(w_t - w*)||$$

Note that this is equivalent to :

$$\forall\, \epsilon,\ \exists\, ||\cdot|| : ||w_{t+1} - w^*|| \leq [\rho(I_d - \eta \nabla F(w*)) + \epsilon]\ \ ||(w_t - w*)|| \tag{4}$$

Now, we need to do a spectral analysis of $\rho(I_d - \eta \nabla F(w*))$ in order to assess convergence. In other words, we have convergence if all the eigenvalues of this matrix are strictly smaller than 1 because it would imply that the distance to the optimum decreases at each step.

4.2 Convergence

Definition 2 (ρ).

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

Theorem 1. *Let $\epsilon \in \mathbb{R} : \epsilon > 0$ and the quantity ρ be defined for a vector field F (as in definition 1) and a standard update rule (as in equation 1).*

1. *If $\rho < 1$, there exists a constant C such that:*

$$||w_t - w^*|| \leq C(\rho + \epsilon)^t$$

This gives the convergence rate if initialized close enough to the optimum.

2. *If $\rho > 1$ then for almost all initialization, the gradient method does not converge to the optimum.*

Proof. The proof actually started at the beginning of section 4. We can pick up from equation 4 and finish it via the following simple step. Let the distance to the optimum be $||w_t - w^*|| \leq (\rho + \epsilon)||w_{t-1} - w^*||$ after t steps with the spectral radius ρ and $\epsilon > 0$. We can recursively replace the distance to the optimum and get:

$$||w_t - w^*|| \leq (\rho + \epsilon)||w_{t-1} - w^*|| \leq (\rho + \epsilon)^2||w_{t-2} - w^*|| \leq ... \leq (\rho + \epsilon)^t||w_0 - w^*|| = C(\rho + \epsilon)^t$$

1. Keeping in mind that $\epsilon > 0$, if $\rho > 1$, then $(\rho + \epsilon) > 1$. Thus the distance to the optimum explodes as $t \to \infty$, so there is no convergence.
2. If $\rho < 1$ and $(\rho + \epsilon) < 1$, then the distance to the optimum tends to zero when $t \to \infty$. So there is convergence with rate $(\rho + \epsilon)$.

□

Remark on the choice of the norm

Remember that to reach equation 4, we made no assumption with regards to the norm. What if we care about the L2 norm in particular? In general for a finite dimension d, we have:

$$\exists\, c_1, c_2\ : \forall w \in \mathbb{R}^d\ \ c_1||w||_2 \leq ||w|| \leq c_2||w||_2$$

In our case, this yields:

$$c_1||w_t - w^*||_2 \leq ||w_t - w^*|| \leq C(\rho + \epsilon)^t$$

Therefore:

$$\forall \epsilon, \exists c_1 : ||w_t - w^*||_2 \leq \frac{C}{c_1}(\rho + \epsilon)^t$$

We see that fixing a particular norm only scales the constant C and does not affect the convergence rate $(\rho + \epsilon)$ in any way. In other words, we might have to do more update steps, but fixing the norm won't prevent the method from converging.

Remark on initialization

Remember that to reach equation 4, we approximated $F(w_t)$ as its first order Taylor expansion around w^*. This implies that we need to initialize close enough to w^* for the approximation to be true. The exact Taylor expansion is:

$$F(w_t) = F(w*) + \nabla F(w^*)(w_t - w^*) + O(||w_t - w^*||^2)$$

In order for the first order expansion to be a good approximation, we need $O(||w_t - w^*||^2)$ to be small. By definition of the big-O notation, $O(||w_t - w^*||^2)$ = $M||w_t - w^*||^2$ where M is an upper-bound on $||\nabla^2 F(w)||$. Therefore, we initialize close enough to the optimum if

$$M||w_t - w^*||^2 \ll ||\nabla F(w^*)(w_t - w^*)||$$

Example :

$$F(w) = Aw$$
$$\nabla F(w) = A$$
$$\nabla^2 F(w) = 0$$

In this case, M=0 and we have the exact relation $F(w_t) = F(w^*) + \nabla F(w_t)(w_t - w*)$. Consequently, we can initialize anywhere.

4.3 Partial conclusion

We now have a connection between the **convergence** (numerical analysis) and the **eigenvalues** (spectral analysis). In fact, as long as we initialize close enough to the optimum, the convergence rate for a given vector field $F(w)$ is:

$$\rho(I_d - \eta\nabla F(w*)) = \max\{|1 - \eta\lambda| : \lambda \in Sp(\nabla F(w*))\}$$

Where

- $|1 - \eta\lambda|$ has to be smaller than 1 for all eigenvalues in order to achieve convergence
- $Sp(\nabla F(w*))$ is the set of all eigenvalues of the vector field's Jacobian at the optimum

4.4 A practical approach

In our machine learning context, $\nabla F(w*)$ is actually the Hessian of the loss (by definition of F) around the minima and so it will most likely always be real. Even though the matrix $\nabla F(w*)$ is composed of real elements, the eigenvalues may be complex numbers. To address this, we look at the square of $|1 - \lambda|$:

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

Where $\Re(\lambda)$ is the real part of λ and η is the step size. This approach is convenient because for a small step size, we can focus only on the real part of the eigenvalues:

$$|1 - \eta\lambda|^2 \approx 1 - 2\eta\Re(\lambda) \ \text{, if } \eta^2 \ll 1$$

Then, for $|1 - \eta\lambda|$ to be smaller than one, we need:

$$\Re(\lambda) > 0, \forall\lambda \in Sp(\nabla F(w^*))$$

Notice that this condition is similar to a generalization of strong convexity for games. As a matter of fact, we need the Hessian around w^* to be definite positive to have convergence. Equivalently, we can say that we want the curvature around the minima to look like a small bowl in order to converge to the minima.

Remark on the nature of λ

Sometimes (like for standard gradient descent with a single player and a single loss), the Hessian of the loss is symmetrical. In these cases, the eigenvalues of the Hessian are real because the eigenvalues of a symmetric matrix are always real[1].

Proof. Let S be a real and symmetric matrix $S \in \mathbb{R}^{n\times n} : S^T = S$. Since S is symmetric, it has an eigendecomposition $Su = \lambda u$ with $u \in \mathbb{C}^n$ and $\lambda \in \mathbb{C}$. Since S is real, we have $S^* = S$ and

$$\begin{aligned} Su = \lambda u &\iff Su^* = \lambda^* u^* \\ \implies u^{*\top} Su &= u^{*\top}(Su) = u^{*\top}\lambda u = \lambda\langle u, u\rangle \\ \implies u^{*\top} Su &= (Su^*)^\top u = (\lambda^* u^*)^\top u = \lambda^*\langle u, u\rangle \end{aligned}$$

Since we consider a general case where $u \neq 0$, we have $\lambda^* = \lambda$ and thus $\lambda \in \mathbb{R}$.

□

4.5 Visualisation

In figure 1, the orange points represent $(1-\lambda)\ \forall\lambda \in Sp(\nabla F(w*))$. The grey lines show how these points move when adding the step size η and varying it from 1 to 0. Indeed, every point is at (1,0) if $\eta = 0$. The green dots represent $(1-\eta\lambda)$ when η is somewhere between 0 and 1. The orange dots are actually the particular case $\eta = 1$.
We showed that the method will converge only if all eigenvalues respect $\Re(\lambda) > 0$. This condition is easily represented by the red line in figure 1. More precisely, the condition is fulfilled when all dots are on the left side of the red line. But this condition is not sufficient to achieve convergence, we also need to have $(1-\eta\lambda) < 1\ \forall\lambda$. This second condition is satisfied when all dots are inside the blue circle. If both conditions are satisfied, the method converges.
Figure 2 shows three examples with different sets of values $(1-\eta\lambda)$. In example 1, one of the value is to the right of the red line, so it cannot converge. In example 2, a value is exactly on the red line, but since we want the real part of λ to be strictly greater than zero, there is no convergence here either. In example 3, all dots are to the left of the red line. Thus there exist a step size η for which all dots are inside the blue circle, which in turn indicates convergence.

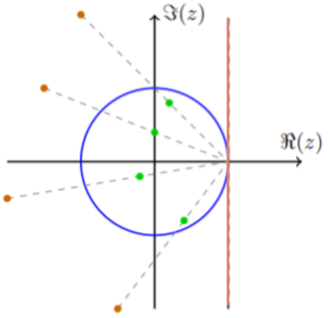

Figure from Mescheder, et al 2018

Figure 1: Nice titre/caption

Once we have been able to put all the values $(1-\eta\lambda)$ in the blue circle by adjusting η, we now try to put them in the smallest circle possible. The interpretation is that the convergence rate is given by the largest radius $|1-\eta\lambda|$ of the eigenvalues λ (represented as the red circle in figure 3). Coincidentally, the smaller the radius the better the convergence rate. In other words, the best step size η is the one that puts the eigenvalues in the smallest circle.

4.5.1 Special Case: Gradient Descent

As mentioned earlier, the Hessian $\nabla F(w) = \nabla^2 g(w)$ for a regular Gradient Descent is symmetric and it implies that the eigenvalues are real. This means that the values $(1-\eta\lambda)$ are always on the real axis as shows figure 4. We note

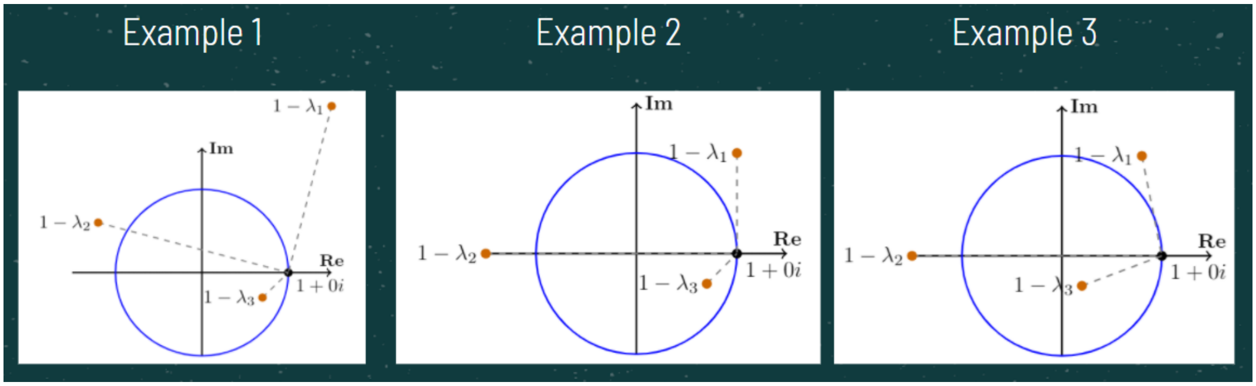

Figure 2: Nice titre/caption

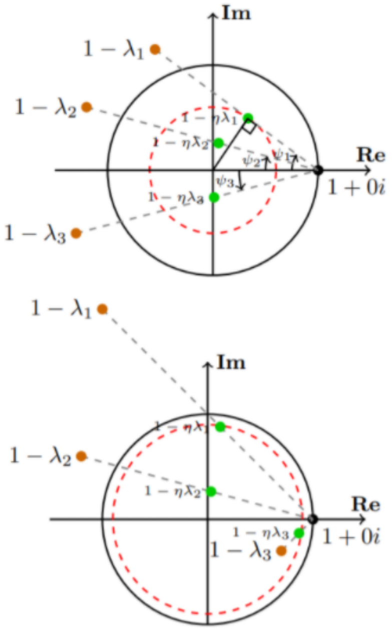

Figure 3: Nice titre/caption

the largest eigenvalue L and the smallest μ.
The problem of finding the optimal η can be written as:

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

By symmetry of the problem, we know that the smallest radius circumscribing all values is achieved when the distance from the origin to $(1-L)$ equals the distance from the origin to $(1-\mu)$:

$$(1-\eta^* L)^2 = (1-\eta^*\mu)^2$$

This equation yields two solutions for η^*. The first one is not interesting as it is $\eta^* = 0$. The second one is:

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

By replacing with the optimal step size found:

$$\rho(I_d - \eta^*\nabla^2 g(w^*)) = \max\{|1-\eta^*\lambda| : \lambda \in Sp(\nabla^2 g(w^*))\} = \left|1 - \frac{2\mu}{L+\mu}\right| = \frac{L-\mu}{L+\mu}$$

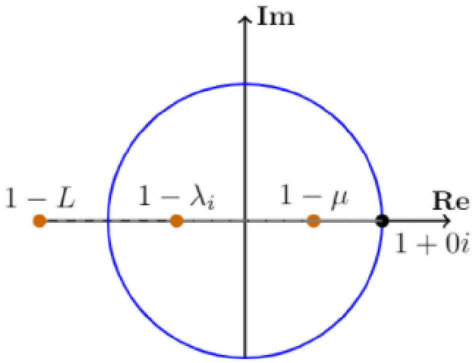

Figure 4: Nice titre/caption

4.5.2 Reminder on optimization

In the last lectures, we have seen by numerical proof that the convergence for strongly convex and Lipschitz functions have the following **global** convergence rate:

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

Where the larger $\frac{\mu}{L}$ is, the better the convergence rate. But we have just seen using spectral proof that the convergence for strongly convex and Lipschitz functions is:

$$||w_t - w^*||^2 \leq \left(1 - \frac{2\mu}{L + \mu}\right)^t ||w_0 - w^*||^2$$

Since $\frac{2\mu}{L+\mu}$ is always larger than $\frac{\mu}{L}$, the spectral proof gives a better convergence rate than the numerical proof. Why is that? simply because the spectral proof provides a **local** convergence rate (with stronger assumptions).

Remark on non-smooth cases

The spectral analysis doesn't work for a non-smooth case because the Jacobian is not continuous. The first order Taylor approximation made at the start of the lecture does not hold oin this context.

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

In minimization, we get all the eigenvalues on the real ($\Re$) axis, while in games the Jacobian can have imaginary ($\Im$) eigenvalues. Thus it can be way more challenging to solve :

$$\min_\eta \max_i |1 - \eta\lambda_i|^2, 1 \leq i \leq n$$

In other words it could be more challenging to fit all the value inside the red circle of figure 3 because of the angle ψ.

Remark on interpretation of ψ

The presense of real and imaginary parts indicates a trade-off between cooperation and adversity in games. It can be shown that purely cooperative games (also known as potential games) have their eigenvalues on the real axis, while purely adversarial games (also known as Hamiltonian games) have their eigenvalues on the imaginary axis.
Moreover, the gradient of the vector field is symmetric for cooperative games and anti-symmetric for adversarial games. Therefore, we can express the vector field's gradient of any game which is not purely cooperative or purely adversarial as:

$$\nabla F(w^*) = J = S + A$$

Where $S^T = S$ and $A^T = -A$. Since S is symmetric and A is anti-symmetric, they respectively have pure real and pure imaginary eigenvalues. For a mixed game (both cooperative and adversarial), we thus could be tempted to interpret the real part x of the eigenvalues $\lambda = x + iy$ as the cooperative contribution (from S) and the imaginary part iy as the adversarial contribution (from A). But this intuition is actually erroneous. It is only valid for non-mixed games:

$$\text{if } A = 0 \implies J = S \implies eig(J) \in \mathbb{R} = eig(S)$$

and

$$\text{if } S = 0 \implies J = A \implies eig(J) \in \mathbb{R} = eig(A)$$

This leaves the mapping from the real and imaginary parts of the eigenvalues to cooperation and adversity ambiguous for games which are not purely cooperative or adversarial.

Remark about GANs

In the paper [2], they compute eigenvalues for GANs to see where they are. The result is that the values are between there is some eigenvalues that have an important imaginary part and some that are pure real values. One interesting observation is that the gradient penality seems to make GANs more comperative since there is no more pure real values.

6 Theorem

$$\rho(I_d - \eta^* \nabla F(w^*)) \approx 1 - min_i \Re(1/\lambda_i) min_i \Re(\lambda_i)$$

Where $min_i \Re(1/\lambda_i)$ is equivalent to 1/L
and $min_i \Re(\lambda_i)$ to μ.

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

If we have a eigenvalue that is very large but with a real part small, $\Re(1/\lambda)$ would be large. That could be an explanation why gradient descent is very slow, as shown in the paper [3]

7 Conclusion

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

1. The analysis of games is more challenging because the Jacobian of the vector field ∇F may have imaginary eigenvalues.
2. For minimization, the Jacobian is a Hessian (thus only has real eigenvalues).
3. The (sufficient) condition for local convergence was to have only eigenvalues with positive real part.

References

[1] Math 2940: Symmetric matrices have real eigenvalues. `http://pi.math.cornell.edu/~jerison/math2940/real-eigenvalues.pdf`.

[2] H. Berard, G. Gidel, A. Almahairi, P. Vincent, and S. Lacoste-Julien. A closer look at the optimization landscapes of generative adversarial networks, 2020.

[3] L. Mescheder, S. Nowozin, and A. Geiger. The numerics of gans, 2018.