Objectives Towards Stable Adversarial Training Without Gradient Penalties

Christos Tsirigotis*
Mila, Université de Montréal,
Aristotle University of Thessaloniki

R. Devon Hjelm
Mila, Université de Montréal,
Microsoft Research

Aaron Courville†
Mila, Université de Montréal

Pericles A. Mitkas
Aristotle University of Thessaloniki

Abstract

Recent advances in adversarial learning observe that stabilization with gradient penalties trades off generated sample quality. We try to address this limitation by proposing novel objectives inspired by the logical XOR operation, which should not depend on gradient penalty regularization in order to be locally stable. In the sections that follow, we will present a theoretical study on this type of objective functions. We will prove global optimality conditions with similar assumptions as those made in the original GAN paper by Goodfellow et al., we will notice connections between the XOR-type objectives and the original GAN, and we will define non zero-sum parametric objectives based on that connection. Finally, we will attempt to study the local stability of the continuous-time training dynamical system around desirable equilibria.

1 Introduction

Generative Adversarial Networks (GAN) [1] offer a training methodology which has produced state-of-the-art generative models in terms of sample quality, enabling solutions that scale to vast datasets [2] as well as to large resolution images [3]. Though successful, stabilizing the training procedure still requires considerable effort from a practitioner's point of view, and as a consequence, their training dynamics have attracted the research interest of the communities of optimization and game theory. Two complementary research directions are considered in this endeavour. The first tries to solve the stabilization problem by proposing, alternative to stochastic gradient descent, algorithms which are more suitable for saddle-point optimization [4, 5]. The second one tries instead to propose or modify the training objectives to be optimized, which in turn hopefully leads to better behaved training procedures [6, 7, 8]. This work proceeds in the second direction by proposing and studying novel adversarial objectives, that solve the same generative problem, utilizing the GAN framework, while trying to establish stable training and to avoid limitations introduced by existing methods.

A common limitation among existing methods is the usage of objectives which correspond to a zero-sum game between the generator and the critic network [9]. This is a known issue regarding the zero-sum game setting: An algorithm or regularized dynamics trying to reach a Nash equilibrium will eventually lead to "limit" cycles around desired equilibria. Modifications to the original objectives which depart from a zero-sum game, like the gradient penalty [10, 6, 7] or the non-saturating

*Work done for diploma thesis at Aristotle University of Thessaloniki. Correspondence to: christos.tsirigotis@umontreal.ca

†Canadian Institute for Advanced Research (CIFAR) Fellow

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NeurIPS 2019), Vancouver, Canada.

standard GAN objectives [1, 11], offer in certain cases stable training procedures in the cost of approximating solutions of a slightly different problem. The gradient penalty, in particular, tries to regularize the Lipschitz constant of the critic function and it is considered as standard technology for the robust training of GANs.

1.1 Discussion on constraining critic function's Lipschitz constant

While originally devised [10] in order to satisfy the function class constraint of the Kantorovich metric between probability measures, which was introduced to the GAN framework by Arjovsky et al. (WGAN) [12], variations of the gradient penalty were soon proposed to stabilize adversarial training [6, 7]. These methods regularize the GAN or WGAN objectives so that certain theoretical properties about local convergence can be shown and they have been deployed in practice successfully. However, there have been experiments which suggest that utilizing a gradient penalty does not yield state-of-the-art performance, even though the training procedure is effectively stabilized [2]. Similar concerns have been attempted to be put forward by theoretical arguments [13]. Looking closely at proposition 1.1, we postulate that this behaviour is due to the final critic having zero gradient with respect to its input everywhere on the support of the real distribution. Mescheder et al. [7] made a seemingly necessary assumption for studying the equilibria of GAN and WGAN training; that the critic will have a constant value in a local neighbourhood of the real distribution's support at equilibrium. Thus, the final critic is locally constant in all directions on every point in the real distribution's support. So by combining this, with the hypothesis that the real distribution lies on a lower dimensional manifold [11], we hypothesize in particular that this condition predicts the limitation of the critic's capacity to locally discriminate between real and fake samples, which are nearby, but not on, the support of the real distribution.

Proposition 1.1 (***Stationary points of GAN and WGAN (Mescheder et al. [7])***)

Let C_ψ be a parametric model of the critic function and G_θ be the parametric model of the generator function. Also, $Q_\theta := G_\theta \# Z$ the induced measure in sample space by pushing random variable Z through model G_θ. P is the target probability measure. Points (θ^*, ψ^*) of the joint parameter space consist equilibria of a system optimizing with GAN (1a) or WGAN (1b) parametric objectives.

$$\text{(GAN)} \quad Q_\theta = P \quad \text{and} \quad C_\psi(x) = 0 \quad \text{and} \quad \nabla_x C_\psi(x) = 0 \quad \forall x \in \text{supp}\{P\} \tag{1a}$$

$$\text{(WGAN)} \quad Q_\theta = P \quad \text{and} \quad \nabla_x C_\psi(x) = 0 \quad \forall x \in \text{supp}\{P\} \tag{1b}$$

A parallel successful attempt to satisfy a Lipschitz constraint in the critic's class is spectral normalization [8]. It modifies critic's architecture in such as a way so that its Lipschitz constant is approximately equal to 1. This has been deployed successfully in practice [2] for the original GAN objectives, leading to superior sample quality and easier optimization, however it does not seem to suffice for guaranteeing training's stability. This could be because such normalization does not guarantee that the final condition for the critic's gradients (proposition 1.1) will be met in the case of GAN's or WGAN's desired training equilibria. So, we believe that this method acts complimentary to those guaranteeing local stability around training's equilibria.

2 Methodology

We motivate our search for alternative adversarial objective functions in methods which utilize more than one sample as inputs to the discriminator function. Some of them [14, 15, 16, 17, 18, 19, 13] derive the expressions for the objectives from the Maximum Mean Discrepancy [20] metric between probability measures, while others motivate their objective functions in an attempt to mitigate the mode dropping problem [17, 21, 22]. A recent method in this line of thought [23] suggests experimentally that defining objectives which act on the relative discrimination between real and fake samples may have actual benefits on the stability of a GAN's training.

In this work we will use a differentiable anagolue of the XOR logical operation, in order to define relative discrimination objective functions for adversarial generation. In particular, we think of a discriminator function $D: \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ which accepts two sample inputs and tries to discriminate whether these two samples have been drawn from the same or from a different distribution. We are going to have D learn an XOR-type relation between samples from the two distribution. In this sense, D is trained to assign the "true" value when two samples are drawn from different distributions, and the "false" value when they are drawn from the same. Consequently, the adversarial game, we will initially consider, is formulated as such:

$$\min_Q \max_D \; \mathop{E}_{\substack{x \; P \\ y \; P}} \log\left(1 - D(x,y)\right) + \mathop{E}_{\substack{x \; P \\ y \; Q}} \log D(x,y) + \mathop{E}_{\substack{x \; Q \\ y \; Q}} \log\left(1 - D(x,y)\right) + \mathop{E}_{\substack{x \; Q \\ y \; P}} \log D(x,y) \tag{2}$$

3 Theoretical Results

3.1 Global Optimality Analysis

For solving analytically the game described in eq. (2), we will make the following assumptions: First, analysis concerns the realizable case. Second, measures P and Q are absolutely continuous between themselves. Third, they both admit probability density functions, p and q respectively, under a common measure of reference, .

Lemma 3.1 (Optimal discriminator of (2))

$$D(x,y) = \frac{a(x,y)}{a(x,y) + b(x,y)} \quad \forall x,y \in \operatorname{supp}\{\quad\} \tag{3a}$$

$$a(x,y) = \frac{1}{2}\left(p(x)\,q(y) + q(x)\,p(y)\right) \tag{3b}$$

$$b(x,y) = \frac{1}{2}\left(p(x)\,p(y) + q(x)\,q(y)\right) \tag{3c}$$

Outside $\operatorname{supp}\{\quad\}$, D can take any real value.

Proof: Upon expanding the expectations in (2) and after algebraic manipulations, we get:

$$D = \arg\max_D \; 2 \int \left\{ \log D(x,y)\; a(x,y) + \log\left(1 - D(x,y)\right) b(x,y) \right\} d(x)\,d(y) \tag{4}$$

Expressions (3b) and (3c) have an integral equal to 1 and consequently a and b can be considered as "mixture" densities of probability measures A and B on $\operatorname{supp}\{\quad\} \times \operatorname{supp}\{\quad\}$, absolutely continuous with respect to . Likewise to the analysis in Goodfellow et al. [1], the integral takes maximum value iff each integrated term is maximal for each $(x,y) \in \operatorname{supp}\{\quad\}$. So:

$$D(x,y) = \arg\max_d \; \log(d)\, a(x,y) + \log(1 - d)\, b(x,y) \qquad \forall (x,y) \in \operatorname{supp}\{\quad\} \tag{5}$$

The function of d to be maximized has unique extremal at $\frac{a}{a+b}$ and is strictly concave in $(0,1)$. □

Proposition 3.2 (Optimal generator of (2))

$$Q = P \tag{6}$$

Proof: Substituting (3) in (2), we get the following equivalent optimization problem for Q:

$$Q = \arg\min_Q \left\{ 4\,\mathrm{JSD}\left(A(Q)\,;B(Q)\right) - 4\log 2 \right\} \iff A_Q = B_Q \iff \tag{7a}$$

$$p(x)q(y) + q(x)p(y) = p(x)p(y) + q(x)q(y) \iff \tag{7b}$$

$$\left(p(x) - q(x)\right)\left(p(y) - q(y)\right) = 0 \iff q = p \iff Q = P \tag{7c}$$

□

3.2 XORGAN Parametric Objectives

We notice that we can factor the expression of optimal discriminator in lemma 3.1 with terms of the original GAN's optimal discriminator expression [1], $D_1(x)$, as such:

$$D_{XOR}(x,y) = \frac{p(x)q(y) + q(x)p(y)}{\left(p(x) + q(x)\right)\left(p(y) + q(y)\right)} = D_1(x)\left(1 - D_1(y)\right) + \left(1 - D_1(x)\right) D_1(y) \tag{8}$$

This motivates us to de ne the XORGAN's objectives using a critic function of a single sample input. Furthermore, this allows us to state the following equivalent non zero-sum game between the critic and the generator function:

Definition 3.3 (Parametric XORGAN objective functions)

$$\arg\max \; L_{GAN}(;) := \underset{x \; P}{E} \log \; C \; (x) \; + \underset{x \; Q}{E} \log \; C \; (x) \qquad (9)$$

$$\arg\min \; L_{XOR}(;) := S \; P; P; C \; + S \; Q \; ; Q \; ; C \; + 2 \, D \; P; Q \; ; C \qquad (10)$$

$$D(P; Q; C) := \underset{\substack{x \; P \\ y \; Q}}{E} \log \; C(x) \; C(y) \; + \; C(x) \; C(y) \qquad (11)$$

$$S(P; Q; C) := \underset{\substack{x \; P \\ y \; Q}}{E} \log \; C(x) \; C(y) \; + \; C(x) \; C(y) \qquad (12)$$

The expressions above can be derived by modeling $D_1(x; \;)$ by $C \; (x)$ as in [1] and substituting appropriately into a model for $D_{XOR}(x; y; \;)$, and then in eq. (2).

3.3 Stability Analysis in Continuous-Time Training Dynamics

We now present the training of a XORGAN as a continuous-time dynamical system, as in [24, 7], by considering gradient descent/ascent as its optimization algorithm:

$$v(;) := \begin{pmatrix} - \\ - \end{pmatrix} = \begin{pmatrix} r \; L_{GAN}(;) \\ r \; L_{XOR}(;) \end{pmatrix} \qquad (13)$$

Then, the following statements can be proven regarding its stability analysis:

Proposition 3.4 (Stationary points of XORGAN)

Points(;) of the joint parameter space consist equilibria of the system which occurs from the optimization of XORGAN parametric objectives, as described in de nition 3.3.

$$\text{(XORGAN)} \qquad Q \; = P \;\; \text{and} \;\; C \; (x) = 0 \quad 8x \; 2 \; \text{supp} f \, Pg \qquad (14)$$

Also, the system's Jacobian at these points is negative semi-de nite and it has real eigenvalues only.

Proof of proposition 3.4 can be found in supplementary material.

4 Conclusions and Future Work

We hope that these primary ndings can serve as motivation for further investigation towards this direction. First of all, however, we are still seeking to nalize an argument towards asymptotic stability to a forward invariant set of the training dynamics, using Lyapunov arguments. That would reassure us about XORGAN's local stability properties compared to existing analyses of other objectives. Nevertheless, the nding of desirable and provable equilibria with negative semi-de nite Jacobian indicates that we are searching in a good direction. This view can be reinforced by the simulation experiments on toy problems, presented in g. 1 and in supplementary material. Second, in future work, we seek to compare experimentally in image generation tasks against benchmark adversarial objectives, like GAN [1] or WGAN [12], as well as with more recent approaches like the RGAN [23]. Finally, we hope that current work will motivate simple experiments which would investigate the limitations of gradient penalty regularization in adversarial learning, if any.

Figure 1: Simulation with $C \; (x) = \; _0 x + \; _1$ and $Q \; =$

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, 2019.

[3] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[4] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with Optimism. In *International Conference on Learning Representations*, 2018.

[5] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A Variational Inequality Perspective on Generative Adversarial Networks. In *International Conference on Learning Representations*, 2019.

[6] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing Training of Generative Adversarial Networks through Regularization. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2018–2028. Curran Associates, Inc., 2017.

[7] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which Training Methods for GANs do actually Converge? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3481–3490. PMLR, Stockholmsmässan, Stockholm Sweden, July 2018.

[8] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. In *International Conference on Learning Representations*, 2018.

[9] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in Adversarial Regularized Learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pages 2703–2717, Philadelphia, PA, USA, 2018. Society for Industrial and Applied Mathematics.

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved Training of Wasserstein GANs. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5767–5777. Curran Associates, Inc., 2017.

[11] Martin Arjovsky and Léon Bottou. Towards Principled Methods for Training Generative Adversarial Networks. In *International Conference on Learning Representations*, 2017.

[12] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, International Convention Centre, Sydney, Australia, 2017.

[13] Wei Wang, Yuan Sun, and Saman Halgamuge. Improving MMD-GAN Training with Repulsive Loss Function. In *International Conference on Learning Representations*, 2019.

[14] Yujia Li, Kevin Swersky, and Rich Zemel. Generative Moment Matching Networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1718–1727. PMLR, Lille, France, 2015.

[15] Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

[16] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabas Poczos. MMD GAN: Towards Deeper Understanding of Moment Matching Network. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2203–2213. Curran Associates, Inc., 2017.

[17] Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, Martin Heusel, Hubert Ramsauer, and Sepp Hochreiter. Coulomb GANs: Provably Optimal Nash Equilibria via Potential Fields. In *International Conference on Learning Representations*, 2018.

[18] Miko aj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

[19] Michael Arbel, Dougal Sutherland, Miko aj Bińkowski, and Arthur Gretton. On gradient regularizers for MMD GANs. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6701–6711. Curran Associates, Inc., 2018.

[20] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schölkopf. Injective Hilbert space embeddings of probability measures. In *Proceedings of the 21st Annual Conference on Learning Theory*, Helsinki, Finland, 2008.

[21] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. PacGAN: The power of two samples in generative adversarial networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1505–1514. Curran Associates, Inc., 2018.

[22] Thomas Lucas, Corentin Tallec, Yann Ollivier, and Jakob Verbeek. Mixed batches and symmetric discriminators for GAN training. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2844–2853. PMLR, Stockholmsmässan, Stockholm Sweden, July 2018.

[23] Alexia Jolicoeur-Martineau. The relativistic discriminator: A key element missing from standard GAN. In *International Conference on Learning Representations*, 2019.

[24] Vaishnavh Nagarajan and J. Zico Kolter. Gradient descent GAN optimization is locally stable. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5585–5595. Curran Associates, Inc., 2017.

[25] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A Semantic Loss Function for Deep Learning with Symbolic Knowledge. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5502–5511. PMLR, Stockholmsmässan, Stockholm Sweden, July 2018.

[26] Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

5 Supplementary Material

5.1 Remarks on XORGAN Optimal Discriminator

5.1.1 In relation to the original GAN objectives

Let us look into eq. (3) of the optimal discriminator. We could try to represent this with a model of two samples inputs. However, this would be more computationally inefficient than representing it with a model of one sample input.

Remark 5.1 (optimal GAN discriminator appears in optimal XORGAN discriminator). Substitute expressions for A (3b) and B (3c) and after algebraic manipulations we get:

$$D_{XOR}(x;y) = \frac{p(x)q(y) + q(x)p(y)}{\left(p(x)+q(x)\right)\left(p(y)+q(y)\right)} \tag{15a}$$

$$= \frac{p(x)}{p(x)+q(x)} \cdot \frac{q(y)}{p(y)+q(y)} + \frac{q(x)}{p(x)+q(x)} \cdot \frac{p(y)}{p(y)+q(y)} \tag{15b}$$

$$= \frac{p(x)}{p(x)+q(x)} \left(1 - \frac{p(y)}{p(y)+q(y)}\right) + \left(1 - \frac{p(x)}{p(x)+q(x)}\right) \frac{p(y)}{p(y)+q(y)} \tag{15c}$$

$$= D_{GAN}(x)\left(1 - D_{GAN}(y)\right) + \left(1 - D_{GAN}(x)\right) D_{GAN}(y) \tag{15d}$$

This way we can isolate the same term multiple times within the same expression, $D_{GAN} = \frac{p}{p+q}$, hence simplifying the representation into an operation between functions of a single sample input.

We could have isolated the term $\frac{q}{p+q}$ instead. This would mean that an original GAN targetted 1 for the fake data and 0 for the real. However, such a choice does not affect the final expression for XORGAN's optimal discriminator as it is symmetric with respect to p and q.

Furthermore, it was possible to derive our relativistic objectives in the same way for the NXOR logical operation. The optimal discriminator, we would find in that case, would be:

$$D_{NXOR}(x;y) = \frac{p(x)p(y) + q(x)q(y)}{\left(p(x)+q(x)\right)\left(p(y)+q(y)\right)} \tag{16}$$

where again we could have isolated the same terms of a single sample input without affecting the final expression due to symmetry. It is true, however, that, independently from such choices, the objective functional expression remains invariant for every adversarial objectives studied (GAN, WGAN, XORGAN, and others). In any case, the final objective expression is irrelevant to the choice of predetermined targets for the discriminator. In particular, this seems to be the case for XORGAN as well because the following relation holds:

$$D_{XOR}(x;y) + D_{NXOR}(x;y) = 1 \tag{17}$$

5.1.2 In relation to the logical events: same vs different distribution

We chose to represent bijectively the interval $[0,1]$, which is the image of probability measures, by using the sigmoid function. By substituting the optimal discriminator for the GAN objectives with the composite function $\sigma \circ C$, where as $C: \mathbb{R}^d \rightarrow \mathbb{R}$ we define the critic model:

$$D_{XOR}(x;y) = \sigma\left(C(x)\right)\left(1 - \sigma\left(C(y)\right)\right) + \left(1 - \sigma\left(C(x)\right)\right)\sigma\left(C(y)\right) \tag{18}$$

$$= \sigma\left(C(x)\right)\sigma\left(-C(y)\right) + \sigma\left(-C(x)\right)\sigma\left(C(y)\right)$$

Similarly, for the NXOR alternative, we can show that:

$$D_{NXOR}(x;y) = \sigma\left(C(x)\right)\sigma\left(C(y)\right) + \sigma\left(-C(x)\right)\sigma\left(-C(y)\right) \tag{19}$$

We will attempt to give semantics to the derived discriminator expressions. We warn the reader, however, that the following attempt is rushed and immature, more intuitive than well-defined mathematically. More mature and general connection of objective functions to probabilistic semantics can be found in the work of Xu et al. [25]. We find however the following perspective to be useful to our intuition, so we are going to develop it in the current text.

We like to think that the expression for $D_{XOR}(x; y)$ (18) corresponds to the logical proposition:

$$Dxy :\Leftrightarrow (Px \wedge Qy) \vee (Qx \wedge Py) \tag{20}$$

Proposition D (20) is evaluated as true if and only if its inputs are samples of different distributions. Similarly, the expression for $D_{NXOR}(x; y)$ (19) corresponds to the logical proposition:

$$Sxy :\Leftrightarrow (Px \wedge Py) \vee (Qx \wedge Qy) \tag{21}$$

Proposition S (21) is evaluated as true if and only if its inputs are samples from the same distribution. We further see that the disjunction of these two proposition is a tautology (22), which is aligned with the observation made in eq. (17). Due to symmetry, we would derive the same remarks even if we considered $C(x)$ to represent $\frac{q}{p+q}$.

$$\models (Dxy \vee Sxy) \Leftrightarrow (Px \wedge Qy) \vee (Qx \wedge Py) \vee (Px \wedge Py) \vee (Qx \wedge Qy) \tag{22}$$

Having remarked these intuitive relations, we define two objective functionals with respect to two probability measure inputs and a critic function, which the discriminator is consisted of, C. We name them by the first letters of the english words "Different" and "Same", reminding their intuitive utility:

$$D(P; Q; C) := \mathop{\mathbb{E}}_{\substack{x \sim P \\ y \sim Q}} \log \left(C(x) \quad C(y) + \quad C(x) \quad C(y) \right) \tag{23}$$

$$S(P; Q; C) := \mathop{\mathbb{E}}_{\substack{x \sim P \\ y \sim Q}} \log \left(C(x) \quad C(y) + \quad C(x) \quad C(y) \right) \tag{24}$$

The initial problem (2) is written as such, when expressed by the functionals (23) and (24):

$$\arg\min_{Q} \arg\max_{C} \quad S(P; P; C) + D(P; Q; C) + D(Q; P; C) + S(Q; Q; C) \tag{25}$$

By observing that the functional D (23) is symmetric with respect to its probability measure inputs, we can simplify the last expression:

$$\arg\min_{Q} \arg\max_{C} \quad S(P; P; C) + S(Q; Q; C) + 2D(P; Q; C) \tag{26}$$

5.2 Proofs about Stability of Non Zero-sum XORGAN Training

As it is described in definition 3.3, we chose to have the original GAN objective for training the discriminator in the non zero-sum game formulation. We remind that this choice does not alter the solutions of the initial problem, due to the symmetry of the optimal XOR discriminator as we have seen in section 5.1.1. We are restating the non zero-sum objectives (definition 3.3) that we are going to study, as well as the definitions for the functionals D (23) and S (24).

Definition 5.2 (Non zero-sum XORGAN objective functions)

$$\arg\max \quad L_{GAN}(\ ;\ ) := \mathop{\mathbb{E}}_{x \sim P} \log \quad C(x) + \mathop{\mathbb{E}}_{x \sim Q} \log \quad C(x) \tag{27a}$$

$$\arg\min \quad L_{XOR}(\ ;\ ) := S\left(P; P; C\right) + S\left(Q; Q; C\right) + 2D\left(P; Q; C\right) \tag{27b}$$

Training with gradient descent/ascent, stated as a continuous-time dynamical system, becomes:

$$v(\ ;\ ) := \begin{pmatrix} - \\ - \end{pmatrix} = \begin{pmatrix} \nabla \quad L_{GAN}(\ ;\ ) \\ \nabla \quad L_{XOR}(\ ;\ ) \end{pmatrix} \tag{28}$$

$$= \begin{pmatrix} \mathop{\mathbb{E}}_{x \sim P}\left[C(x) \nabla \quad C(x) \right] \quad \mathop{\mathbb{E}}_{x \sim Q}\left[C(x) \nabla \quad C(x) \right] \\ \nabla_{1} S\left(Q_{1}; Q; C\right)_{1=} \quad \nabla_{2} S\left(Q; Q_{2}; C\right)_{2=} \quad 2\nabla \quad D\left(P; Q; C\right) \end{pmatrix}^{1}$$

Proposition 5.3 (Non zero-sum XORGAN equilibria)

Points(;) of the joint parameter space consist stationary points of the system which occurs from the optimization of non zero-sum XORGAN parametric objectives, as described in eq. (28).

$$Q = P \quad \text{and} \quad C \quad (x) = 0 \quad 8x\ 2\ \text{suppf}\ Pg \tag{29}$$

Proof: We can easily verify that points (29) are indeed stationary points, as the time-derivative of the states equals to zero:

$$r \quad L_{GAN}(;) \quad = \quad = \tag{30}$$

$$= \mathop{E}_{x \quad P} \quad C \quad (x) \;\; r \quad C \quad (x) \quad = \quad \mathop{E}_{x \quad Q} \quad C \quad (x) \;\; r \quad C \quad (x) \quad =$$

$$= \mathop{E}_{x \quad P} \quad (0) \quad (0) \;\; r \quad C \quad (x) \quad = \;\; = 0$$

For nding the derivative with respect to, let us remember that the functionals S (24) and D (23) are symmetric with respect to their measure inputs, as also that $D_{NXOR} = 1 \quad D_{XOR}$, which is apparent from the following de nitions:

$$D_{NXOR}(x; y; \;) := \quad C \;(x) \quad C \;(y) \;+ \quad C \;(x) \quad C \;(y) \tag{31a}$$

$$D_{XOR}(x; y; \;) := \quad C \;(x) \quad C \;(y) \;+ \quad C \;(x) \quad C \;(y) \tag{31b}$$

Calculating the derivatives of the constituent expressions:

$$r \quad S \quad P; Q \; ; C \quad = r \begin{Bmatrix} \mathop{E}_{\substack{x \quad P \\ z \quad Z}} \log \quad D_{NXOR} \quad x; G \;(z) \; ; \end{Bmatrix} \tag{32a}$$

$$= \mathop{E}_{\substack{x \quad P \\ z \quad Z}} \left[\frac{1}{D_{NXOR} \quad x; G \;(z) \; ;} r \quad G \;(z) \, r_y D_{NXOR}(x; y; \;) \;\; _{y = G \;(z)} \right]$$

$$r \quad D \quad P; Q \; ; C \quad = \mathop{E}_{\substack{x \quad P \\ z \quad Z}} \left[\frac{1}{D_{XOR} \quad x; G \;(z) \; ;} r \quad G \;(z) \, r_y D_{XOR}(x; y; \;) \;\; _{y = G \;(z)} \right] \tag{32b}$$

Also for D_{NXOR} and D_{XOR}:

$$r_y D_{NXOR}(x; y; \;) = \quad C \;(y) \quad C \;(y) \quad C \;(x) \quad \Big[\quad C \;(x) \Big] \; r_y C \;(y) \tag{33a}$$

$$r_y D_{XOR}(x; y; \;) = \; r_y D_{XOR}(x; y; \;) \tag{33b}$$

As $8x \;; y \;\; 2\ \text{suppf}\ Pg \quad C \quad (x \;) = 0$, according to eq. (29), we nd the following values for eqs. (31) and (33):

$$D_{NXOR}(x \;; y \;; \;\;) = \quad (0) \quad (0) + \quad (0) \quad 0 = \frac{1}{2} \tag{34a}$$

$$D_{XOR}(x \;; y \;; \;\;) = \frac{1}{2} \tag{34b}$$

$$r_y D_{NXOR}(x \;; y; \;\;) \;\; _{y = y} \; = r_y D_{NXOR}(x \;; y; \;\;) \;\; _{y = y} \; = 0 \tag{34c}$$

Now to nd the values for (32), we make the following observation, beginning from eq. (29):

$$Q \quad = P \;=) \quad \text{suppf}\ Q \quad g = \text{suppf}\ Pg \tag{35a}$$

$$=) \quad G \quad (z)\ 2\ \text{suppf}\ Pg \quad 8z\ 2\ \text{suppf}\ Zg \tag{35b}$$

$$=) \quad C \quad G \quad (z) \quad = 0 \quad 8z\ 2\ \text{suppf}\ Zg \tag{35c}$$

By using eqs. (29), (34) and (35c) in (32), we observe that those are zeroed out, and consequently by substituting into (13), we verify that indeed (;) are stationary points. □

We will examine the behaviour of the system (13) locally around the stationary points (29). For this reason we will linearize the system by expanding its rst order. So we are extracting the Jacobian of the system J (;):

$$! \quad - \; ' \; J(\;\; ; \;\;) \quad = \tag{36}$$
$$-$$

$$\begin{pmatrix} r^2 \; L_{GAN}(\; ; \;) \; {}_{=}^{=} & r^2 \; L_{GAN}(\; ; \;) \; {}_{=}^{=} \\ r^2 \; L_{XOR}(\; ; \;) \; {}_{=}^{=} & r^2 \; L_{XOR}(\; ; \;) \; {}_{=}^{=} \end{pmatrix}$$

Lemma 5.4 (Jacobian at SP (29) of non zero-sum XORGAN)

The Jacobian of the system (28), J , at the stationary points (29) is:

$$J(\;\; ; \;\;) = \begin{matrix} J & J \\ J & J \end{matrix} \tag{37a}$$

$$J \quad = \quad \frac{1}{2} \, {}_{x} E_{P} \;\; r \quad C \;\; (x) \;\; {}_{=} \quad r \quad C \;\; (x) \;\; {}^{T}_{=} \tag{37b}$$

$$J \quad = \quad \frac{1}{4} r \quad {}_{x} E_{Q} \quad \overset{h}{C} \quad (x) + 2 \;\; r \quad C \;\; (x) \;\; {}_{=} \quad \overset{i}{}_{=} \tag{37c}$$

$$J \quad = 0 \tag{37d}$$

$$J \quad = \quad \frac{1}{2} r \quad {}_{x} E_{Q} \quad C \quad (x) \quad {}_{=} \quad r \quad {}_{x} E_{Q} \quad C \quad (x) \quad {}^{T}_{=} \tag{37e}$$

Proof: We will rst nd the values for J and J terms of J , which correspond to the second derivatives of the original GAN critic objective function (27a). We are going to reuse observations that we used in the proof of proposition 5.3.

$$r^2 \quad L_{GAN}(\; ; \;) \quad {}_{=}^{=} \quad = \tag{38}$$

$$= {}_{x} E_{P} \quad C \quad (x) \quad C \quad (x) \;\; r \quad C \;\; (x) \quad {}_{=} \quad r \quad C \;\; (x) \quad {}^{T}_{=} \quad + \quad C \quad (x) \;\; r^2 \;\; C \;\; (x) \quad {}_{=}$$

$$+ {}_{x} E_{Q} \quad C \quad (x) \quad C \quad (x) \;\; r \quad C \;\; (x) \quad {}_{=} \quad r \quad C \;\; (x) \quad {}^{T}_{=} \quad C \quad (x) \;\; r^2 \;\; C \;\; (x) \quad {}_{=}$$

$$= \quad \frac{1}{2} \, {}_{x} E_{P} \;\; r \quad C \;\; (x) \quad {}_{=} \quad r \quad C \;\; (x) \quad {}^{T}_{=}$$

$$r^2 \quad L_{GAN}(\; ; \;) \quad {}_{=}^{=} \quad = \tag{39}$$

$$= {}_{z} E_{Z} \quad " \quad C \quad (x) \quad C \quad (x) \;\; r \quad C \;\; (x) \quad {}_{=} \quad r \;_{x} C \quad (x)^{T}$$

$$C \quad (x) \;\; r^2_{x} \; C \;\; (x) \quad {}_{=} \quad {}_{x = G \quad (z)} \; r \quad G \;\; (z) \quad {}_{=} \quad \#$$

$$= \quad \frac{1}{4} r \quad {}_{x} E_{Q} \quad \overset{h}{C} \quad (x) + 2 \;\; r \quad C \;\; (x) \quad {}_{=} \quad \overset{i}{}_{=}$$

Now we will proceed to calculate the values for J and J . To this end, we will make extensive use of the property $D_{NXOR} + D_{XOR} = 1$, which implies that all derivatives of the summation are equal to 0: $\nabla \{ D_{\mathrm{NXOR}} + D_{\mathrm{XOR}} \} = \nabla \{1\} = 0$. We also observe that $\forall x^*; y^* \in \operatorname{supp}\{\mathrm{P}\}$:

$$\nabla^2_{xy} D_{\mathrm{NXOR}}(x; y; \ ^*) \Big|_{\substack{x=x \\ y=y}} = \tag{40}$$

$$= 2 \quad C \quad (x^*) \qquad C \quad (x^*) \qquad C \quad (y^*) \qquad C \quad (y^*) \ \nabla_x C \quad (x) \Big|_{x=x} \ \nabla_y C \quad (y) \Big|_{y=y}^T$$

$$= \frac{1}{8} \nabla_x C \quad (x) \Big|_{x=x} \ \nabla_y C \quad (y) \Big|_{y=y}^T$$

So:

$$\nabla^2 \ L_{\mathrm{XOR}}(\ ; \ ) \Big|_{\substack{= \\ =}} = 2\nabla^2 \ S \ \mathrm{Q} \ ; \mathrm{Q} \ ; C \Big|_{\substack{= \\ =}} + 2\nabla^2 \ D \ \mathrm{P}; \mathrm{Q} \ ; C \Big|_{\substack{= \\ =}} \tag{41a}$$

$$= 2 \left\{ 2 \mathop{\mathrm{E}}_{\substack{x\sim \mathrm{P} \\ z\sim Z}} \left[\left(\nabla^2_{\ y} D_{\mathrm{NXOR}}(x; y; \ ) \Big|_{\substack{= \\ y=G \ (z)}} + \nabla^2_{\ y} D_{\mathrm{XOR}}(x; y; \ ) \Big|_{\substack{= \\ y=G \ (z)}} \right) \nabla^T G \ (z) \Big|_{=} \right] \right\}$$

$$= 2 \left\{ 2 \mathop{\mathrm{E}}_{\substack{x\sim \mathrm{P} \\ z\sim Z}} \left[\nabla^2 \ 1 \quad \nabla^T G \ (z) \Big|_{=} \right] \right\} = 0$$

$$\nabla^2 \ L_{\mathrm{XOR}}(\ ; \ ) \Big|_{\substack{= \\ =}} = \nabla^2 \ L_{\mathrm{XOR}}(\ ; \ ) \Big|_{\substack{= \\ =}}^T = 0 \tag{41b}$$

$$\nabla^2 \ L_{\mathrm{XOR}}(\ ; \ ) \Big|_{\substack{= \\ =}} = \tag{41c}$$

$$= 2\nabla^2_{\ _1 \ _2} S \ \mathrm{Q} \ _1; \mathrm{Q} \ _2; C \Big|_{\substack{_1 = \\ _2 =}} + 2\nabla^2 \ S \ \mathrm{Q} \ ; \mathrm{Q} \ ; C \Big|_{=} + 2\nabla^2 \ D \ \mathrm{P}; \mathrm{Q} \ ; C \Big|_{=}$$

$$= 4 \mathop{\mathrm{E}}_{\substack{z_1\sim Z \\ z_2\sim Z}} \left[\nabla \ G \ (z_1) \Big|_{=} \ \nabla^2_{xy} D_{\mathrm{NXOR}}(x; y; \ ^*) \Big|_{\substack{x=G \ (z_1) \\ y=G \ (z_2)}} \nabla \ G \ (z_2) \Big|_{=}^T \right]$$

$$= \frac{4}{8} \mathop{\mathrm{E}}_{\substack{z_1\sim Z \\ z_2\sim Z}} \left[\nabla \ G \ (z_1) \Big|_{=} \ \nabla_x C \quad (x) \Big|_{x=G \ (z_1)} \ \nabla_x C \quad (x) \Big|_{x=G \ (z_2)}^T \ \nabla \ G \ (z_2) \Big|_{=}^T \right]$$

$$= \frac{1}{2} \mathop{\mathrm{E}}_{z_1\sim Z} \ \nabla \ C \quad G \ (z_1) \Big|_{=} \ \mathop{\mathrm{E}}_{z_2\sim Z} \ \nabla \ C \quad G \ (z_2) \Big|_{=}^T$$

$$= \frac{1}{2} \nabla \mathop{\mathrm{E}}_{x\sim \mathrm{Q}} \ C \quad (x) \Big|_{=} \ \nabla \mathop{\mathrm{E}}_{x\sim \mathrm{Q}} \ C \quad (x) \Big|_{=}^T$$

□

In order to proceed with our analysis at this point, we will state some assumptions. Specifically, we are going to define the following *reparameterization* manifolds:

$$M_D := \left(\ | C \quad (x) = 0 \quad \forall x \in \operatorname{supp}\{\mathrm{P}\} \ \right)_2 \tag{42}$$

$$M_G := \ | \arg\min \ \mathop{\mathrm{E}}_{x\sim \mathrm{Q}} \ C \quad (x) \tag{43}$$

We observe that M_D (42) simply describes $\ ^*$ part out of the SP of non zero-sum XORGAN, according to proposition 5.3. In addition, we assume that for each $(\ ^*; \ ^*)$ there exist -balls, $B (\ ^*)$ and $B (\ ^*)$, around $\ ^*$ and $\ ^*$ at their respective subspaces, such that $M_D \cap B (\ ^*)$ and $M_G \cap B (\ ^*)$ define C^1-manifolds. Finally, we can express M_D equivalently as:

$$M_D := \ | \arg\min \ \left(\mathop{\mathrm{E}}_{x\sim \mathrm{P}} C \quad (x) \right)^2 \tag{44}$$

To understand the equivalence, we notice that eq. (44) describes a condition which is minimized if and only if it is equal to 0, as it is non-negative. This condition is equal to 0, when all (non-negative) square terms in the integral are also equal to 0. The integral is evaluated on $\mathrm{supp}\{P\}$, thus, for all points in the support, the square terms, and consequently the critic C_ψ, must be equal to 0.

Lemma 5.5 (***Condition for negative definite*** $J_{\psi\psi}$)

If vector $u \neq 0$ does not lie in the tangent space of $\mathcal{M}_D$ (44) at ψ^*, $T_{\psi^*}\mathcal{M}_D$, then $u^T J_{\psi\psi} u < 0$.

Proof: From lemma 5.4 we have

$$u^T J_{\psi\psi} u = -\frac{1}{2} \mathop{\mathbb{E}}_{x\sim P}\left[\left(u^T \nabla_\psi C_\psi(x) \Big|_{\psi=\psi^*} \right)^2\right] \tag{45}$$

which implies $u^T J_{\psi\psi} u \leq 0$. We get equality if and only if:

$$u^T \nabla_\psi C_\psi(x) \Big|_{\psi=\psi^*} = 0 \quad \forall x \in \mathrm{supp}\{P\}. \tag{46}$$

Let:

$$h(\psi) := \mathop{\mathbb{E}}_{x\sim P}\left[C_\psi(x)^2 \right] \tag{47}$$

$$\implies \nabla_\psi h(\psi) = 2 \mathop{\mathbb{E}}_{x\sim P} C_\psi(x) \nabla_\psi C_\psi(x) \tag{48}$$

$$\implies \nabla^2_\psi h(\psi) = 2 \mathop{\mathbb{E}}_{x\sim P}\left[\nabla_\psi C_\psi(x) \nabla_\psi C_\psi(x)^T \right] + 2 \mathop{\mathbb{E}}_{x\sim P}\left[C_\psi(x) \nabla^2_\psi C_\psi(x) \right] \tag{49}$$

By using the expression of ψ^* from proposition 5.3, we observe that:

$$h(\psi^*) = 0 \tag{50}$$

$$\nabla_\psi h(\psi) \Big|_{\psi=\psi^*} = 0 \tag{51}$$

$$\nabla^2_\psi h(\psi) \Big|_{\psi=\psi^*} = 2 \mathop{\mathbb{E}}_{x\sim P} \nabla_\psi C_\psi(x) \Big|_{\psi=\psi^*} \left(\nabla_\psi C_\psi(x) \right)^T \Big|_{\psi=\psi^*} \succeq 0 \tag{52}$$

so it achieves the minimum of the expression h and thus $\psi^* \in \mathcal{M}_D$. The vector $u \in T_{\psi^*}\mathcal{M}_D$ if and only if the second directional derivative at $\psi = \psi^*$ is equal to 0, which is iff eq. (46) holds. □

Lemma 5.6 (***Condition for negative definite*** $J_{\theta\theta}$)

If vector $w \neq 0$ does not lie in the tangent space of $\mathcal{M}_G$ (43) at θ^*, $T_{\theta^*}\mathcal{M}_G$, then $w^T J_{\theta\theta} w < 0$.

Proof: From lemma 5.4 we have

$$w^T J_{\theta\theta} w = -\frac{1}{2} \left(w^T \nabla_\theta \mathop{\mathbb{E}}_{x\sim Q_\theta} C_\psi(x) \Big|_{\theta=\theta^*} \right)^2 \tag{53}$$

which implies $w^T J_{\theta\theta} w \leq 0$. We get equality if and only if:

$$w^T \nabla_\theta \mathop{\mathbb{E}}_{x\sim Q_\theta} C_\psi(x) \Big|_{\theta=\theta^*} = 0 \tag{54}$$

Let:

$$g(\theta) := \left(\mathop{\mathbb{E}}_{x\sim Q_\theta} C_\psi(x) \right)^2 \tag{55}$$

$$\implies \nabla_\theta g(\theta) = 2 \left(\mathop{\mathbb{E}}_{x\sim Q_\theta} C_\psi(x) \right) \nabla_\theta \mathop{\mathbb{E}}_{x\sim Q_\theta} C_\psi(x) \tag{56}$$

$$\implies \nabla^2_\theta g(\theta) = 2 \nabla_\theta \left(\mathop{\mathbb{E}}_{x\sim Q_\theta} C_\psi(x) \right) \nabla_\theta \left(\mathop{\mathbb{E}}_{x\sim Q_\theta} C_\psi(x) \right)^T + 2 \left(\mathop{\mathbb{E}}_{x\sim Q_\theta} C_\psi(x) \right) \nabla^2_\theta \mathop{\mathbb{E}}_{x\sim Q_\theta} C_\psi(x) \tag{57}$$

By using proposition 5.3, we observe that:

$$g(\theta^*) = 0 \tag{58}$$

$$\nabla_\theta g(\theta)\Big|_{\theta=\theta^*} = 0 \tag{59}$$

$$\nabla^2_\theta g(\theta)\Big|_{\theta=\theta^*} = 2\nabla_\theta \mathop{\mathbb{E}}_{x\sim\mathbb{Q}_\theta} C_\phi(x)\Big|_{\theta=\theta^*} \left(\nabla_\theta \mathop{\mathbb{E}}_{x\sim\mathbb{Q}_\theta} C_\phi(x)\Big|_{\theta=\theta^*}\right)^T \succeq 0 \tag{60}$$

so it achieves the minimum of the expression g and thus $\theta^* \in M_G$. The vector $w \in T_{\theta^*} M_G$ if and only if the second directional derivative at $\theta = \theta^*$ is equal to 0, which is iff eq. (54) holds. □

We will attempt to draw some conclusion about the eigenvalues of Jacobian $J(\theta^*, \phi^*)$. For this purpose we prove the following lemma.

Lemma 5.7 (*Eigenvalues of block upper triangular matrix*)

Let matrix

$$J := \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} \tag{61}$$

where $A \in \mathbb{R}^{n\times n}$, $D \in \mathbb{R}^{m\times m}$. Then, $\sigma\{J\} = \sigma\{A\} \cup \sigma\{D\}$.

Proof: Initially we will show that $\sigma\{J\} = \sigma\{A\} \cup \sigma\{D\}$. Let $u^T = \left[x^T; y^T\right]$ be an eigenvector of matrix J, with $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, and the corresponding eigenvalue λ:

$$Ju = \lambda u \iff \begin{pmatrix} Ax + By \\ Dy \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix} \tag{62}$$

If $y \neq 0$, then from $Dy = \lambda y$ we understand that y consists an eigenvector of D with eigenvalue λ. Thus, $\lambda \in \sigma\{D\}$. Otherwise $y = 0$, and then $Ax = \lambda x$ will hold, from which we understand that x consists an eigenvector of A with eigenvalue λ. In total, by combining the two possibilities, it ought to hold that $\sigma\{J\} \subseteq \sigma\{A\} \cup \sigma\{D\}$.

In inverse, let λ be an eigenvalue of A with corresponding eigenvector $x \neq 0$. Then $Ax = \lambda x$ and we observe that the vector $\left[x^T; 0\right]^T$ is an eigenvector of J, because

$$J\begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} Ax \\ 0 \end{pmatrix} = \lambda \begin{pmatrix} x \\ 0 \end{pmatrix} \tag{63}$$

and with λ as its eigenvalue. Thus $\sigma\{A\} \subseteq \sigma\{J\}$. Finally, we consider λ to be an eigenvalue of D, which is not also an eigenvalue of A, so $\lambda \in \sigma\{D\} \setminus \sigma\{A\}$, and its corresponding eigenvector $y \neq 0$. Then $Dy = \lambda y$, but also the matrix $A - \lambda I$ is not singular, and hence invertible. Then we observe that the vector $\left[x^T; y^T\right]^T$, with $x := (-A + \lambda I)^{-1}By$, consists an eigenvector of J with eigenvalue λ.

$$J\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} A(-A + \lambda I)^{-1}By + By \\ Dy \end{pmatrix} = \begin{pmatrix} \left(A + (-A + \lambda I)\right)(-A + \lambda I)^{-1}By \\ \lambda Dy \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix} \tag{64}$$

Consequently, from all possibilities of the inverse, we have $\sigma\{A\} \cup \sigma\{D\} \setminus \sigma\{A\} \subseteq \sigma\{J\}$.

By combining the direct and the inverse, we have: $\sigma\{J\} = \sigma\{A\} \cup \sigma\{D\}$. □

Corollary 5.7.1 (*Eigenvalues of Jacobian at SP (29) of non zero-sum XORGAN*)

Matrix J, found at lemma 5.4, is negative semi-definite in the parameter space $\left[\theta^T; \phi^T\right]^T$. However, in the subspace $V(\theta^*, \phi^*) := \mathrm{span}\left\{\left[u^T; w^T\right]^T \mid u \notin T_{\phi^*} M_D, w \notin T_{\theta^*} M_G\right\}$, J is negative definite. In addition, all of its eigenvalues are real numbers.

Proof: From lemma 5.4, we see that the symmetric matrices $J_{\theta\theta}$ and $J_{\phi\phi}$ are negative semi-definite. In consequence, lemma 5.7 implies that J is negative semi-definite in the joint parameter space. In subspace V though, $J_{\theta\theta}$ and $J_{\phi\phi}$ are negative definite, according to lemmata 5.5 and 5.6. Thus, J has some eigenvalues with negative real part. Furthermore, all eigenvalues are real numbers because $J_{\theta\theta}$ and $J_{\phi\phi}$ are symmetric matrices □