Pourquoi les matrices aléatoires expliquent l'apprentissage ? Un argument d'universalité offert par les GANs

Mohamed El Amine SEDDIK[1,2], Mohamed TAMAAZOUSTI[1], Romain COUILLET[2,3]

[1]CEA, LIST, 8 Avenue de la Vauve, 91120 Palaiseau, France

[2]Laboratoire de Signaux et Systèmes, Centralesupélec, 3 Rue Joliot Curie, 91190 Gif-sur-Yvette, France

[3]GIPSA-lab, Université Grenoble-Alpes, 11 Rue des Mathématiques, 38400 Saint-Martin-d'Hères, France

mohamedelamine.seddik@cea.fr, mohamed.tamaazousti@cea.fr
romain.couillet@gipsa-lab.grenoble-inp.fr

Résumé – Afin de mettre en évidence les aspects d'universalité des grandes matrices aléatoires sur les données réelles, nous étudions dans cet article le comportement spectral de la matrice de Gram pour une large classe de vecteurs aléatoires, les dits *vecteurs concentrés*, qui sont plus riches que des vecteurs gaussiens. L'hypothèse de concentration est particulièrement motivée par le fait que l'on peut générer, grâce aux GANs, des données réalistes par transformations Lipschitziennes de vecteurs gaussiens. En particulier, en générant des images à travers un GAN, nous montrons que le comportement spectral de la matrice de Gram est le même sur ces données que sur un modèle de mélange gaussien, rendant ainsi l'estimation des performances de classification prédictible à travers la théorie des matrices aléatoires pour des données réelles.

Abstract – In order to highlight the universality aspects of large random matrices on real data, we study in this paper the spectral behaviour of the Gram matrix for a large class of random vectors, the so-called *concentrated vectors*, which are richer than Gaussian vectors. The concentration hypothesis is particularly motivated by the fact that realistic data can be generated using GANs by Lipschitzian transformations of Gaussian vectors. Particularly, by generating images across a GAN, we show that the spectral behaviour of the Gram matrix is the same on these data as on a Gaussian mixture model, thereby making the estimation of classification performance predictable through random matrix theory for real data.

1 Introduction

À l'ère des données massives, le traitement de données et l'apprentissage statistique s'orientent vers le développement de méthodes de classification ou d'inférence basées sur des grands jeux de données de grandes tailles. On comprend aujourd'hui que ces grandes dimensions induisent de nombreux phénomènes contre-intuitifs, entraînant une compréhension erronée du comportement des algorithmes souvent conçus pour des données de petites dimensions. En profitant de (au lieu de subir) ce caractère multidimensionnel, la théorie des matrices aléatoires (TMA) est capable de prédire les performances d'algorithmes aussi complexes que certains réseaux de neurones aléatoires [1] ainsi que de nombreuses méthodes à noyau telles que les SVM [2], la classification semi-supervisée [3], l'analyse en composantes principales [4] ou encore le regroupement spectral [5]. Fréquemment, pour caractériser les performances de ces algorithmes théoriquement, le modèle sous-jacent considéré est un mélange gaussien (GMM). Un résultat clé, démontré dans [5], est que la matrice à noyau d'un ensemble de données de grandes dimensions issues d'un modèle GMM multi-classes se "développe aux deux premiers ordres" en un modèle analysable ; notamment, les liens entre classes et fonction à noyau deviennent très facilement lisible à travers une étude des vecteurs propres dominants du modèle. Plus étonnamment, les auteurs de [5] montrent empiriquement que la matrice à noyau présente le même comportement sur des données réelles MNIST que sur des vecteurs gaussiens, suggérant ainsi un aspect d'*universalité* des grandes matrices aléatoires vis-à-vis de la distribution des données.

Le but de cet article est de relaxer l'hypothèse de gaussianité à une classe plus large de distributions, afin de mettre tout d'abord en évidence les propriétés d'universalité que présente la TMA. En effet, la plupart des données réelles (notamment les images ou leurs représentations, souvent utilisées en vision par ordinateur [6]) sont de nature et de structures complexes et paraissent donc a priori peu susceptibles d'être bien représentées par des gaussiennes. Néanmoins, grâce à de nouvelles avancées sur les modèles génératifs depuis l'avènement des réseaux de neurones génératifs adverses (les dits GANs), il est maintenant possible de générer des données complexes et structurées (mimant à s'y méprendre des données réelles) en appliquant des opérations *lipschitziennes* successives à des vecteurs aléatoires gaussiens standards [7]. Parallèlement, la théorie de la *concentration de la mesure* nous apprend que les transformations lipschitziennes de vecteurs gaussiens vérifient des propriétés dites de concentrations [1] et fournit un cadre mathématique puissant pour traiter élégamment ces objets. Ceci suggère que modéliser des données réalistes par le biais de *vecteurs concentrés* est un moyen plus riche mais tout aussi mathématiquement accessible

que via de simples GMM.

Dans ce papier, nous précisons, à travers les GANs, l'intérêt de considérer le cadre de *mélanges de vecteurs concentrés* [1] comme modèle pour les données réelles. En particulier, nous rappellerons, comme nous l'avons montré dans [6], que les matrices à noyaux non-linéaires générées par des données concentrées de grandes tailles ont le *même comportement asymptotique* que pour des données GMM (à savoir approximables par un développement au deuxième ordre). En particulier, un corollaire immédiat est que pour la plupart des méthodes de classification ou régression à noyaux, seuls les statistiques d'ordres un et deux des données concentrées sont informatives. Cette observation démontre ainsi un puissant résultat d'universalité vis-à-vis de la distribution des données qui tend à démontrer que le traitement statistique et d'apprentissage sur des données réelles (si on assimile de vraies images à des images qui pourraient être générées par un GAN) peuvent être analysés théoriquement par le biais de la théorie des matrices aléatoires. Ceci valide par conséquent la pertinence de cette théorie pour l'apprentissage statistique moderne.

2 Mélanges de vecteurs concentrés

2.1 Notions de concentration de la mesure [1]

Nous définissons tout d'abord la notion de *vecteur concentré*. Même si plusieurs notions de concentration ont récemment été développées en vue de l'apprentissage en grandes dimensions [1], pour simplicité, nous nous concentrons ici sur la concentration q-exponentielle.

Définition 1 (Concentration q-exponentielle)**.** *Étant donné un espace normé $(E, \|\cdot\|_E)$ et un réel q, un vecteur $X \in E$ est dit q-exponentiellement concentré si pour toute fonction $\mathcal{F} : E \to \mathbb{R}$ 1-Lipschitz, il existe des constantes $C, c > 0$ telles que*

$$\forall t > 0,\ \mathbb{P}\left\{|\mathcal{F}(X) - \mathbb{E}\mathcal{F}(X)| \geq t\right\} \leq Ce^{-ct^q}.$$

On note alors $X \in \mathcal{O}(e^{-\cdot^q})$ dans $(E, \|\cdot\|_E)$.

L'archétype du vecteur concentré est le vecteur gaussien standard.

Remarque 1 (Concentration d'un vecteur Gaussien [1])**.** *Le vecteur $X \sim \mathcal{N}(0, \mathbf{I}_p)$ est 2-exponentiellement concentré.*

Une propriété intéressante de la concentration est sa stabilité par application de fonctions lipschitziennes. Plus précisément, nous avons la remarque suivante :

Remarque 2 (Stabilité lipschitzienne)**.** *Soit $X \in \mathcal{O}(e^{-\cdot^q})$ dans $(E, \|\cdot\|_E)$ et $\mathcal{G} : E \to F$ une application ℓ-Lipschitz. Alors, la propriété de concentration sur X se transfère à $\mathcal{G}(X)$, à savoir :*

$$\mathcal{G}(X) \in \mathcal{O}(e^{-(\cdot/\ell)^q}) \text{ dans } (F, \|\cdot\|_F).$$

2.2 Des vecteurs concentrés au mélange

Soient $n, p \gg 1$, et une suite de vecteurs aléatoires indépendants $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, distribués en k classes de distributions $\mu_1, \ldots, \mu_k$ de moyennes distinctes mais, pour simplicité ici, de même covariance $\mathbf{C}$.[1] Nous considérons l'hypothèse de concentration q-exponentielle sur la matrice de données

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n},$$

c'est-à-dire qu'il existe $q \geq 2$ tel que pour toute famille de vecteurs indépendants $\mathbf{y}_1, \ldots, \mathbf{y}_m$ de loi μ_ℓ, on a la concentration :

$$[\mathbf{y}_1, \ldots, \mathbf{y}_m] \in \mathcal{O}(e^{-\cdot^q}) \text{ dans } (\mathbb{R}^{p \times m}, \|\cdot\|_F).$$

On nommera $\mathbf{m}_\ell$ la moyenne de μ_ℓ et n_ℓ le nombre de vecteurs $\mathbf{x}_i$ qui suivent la distribution μ_ℓ.

Sans perte de généralité, nous ordonnons les vecteurs $\mathbf{x}_i$ dans la matrice de données $\mathbf{X}$ de sorte que, pour chaque ℓ :

$$\mathbf{x}_{1+\sum_{j=1}^{\ell-1} n_j}, \ldots, \mathbf{x}_{\sum_{j=1}^{\ell} n_j} \sim \mu_\ell.$$

Pour usage ultérieur, nous introduisons les matrices suivantes :

$$\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_k] \in \mathbb{R}^{p \times k},$$
$$\mathbf{J} = \frac{1}{\sqrt{p}}[\mathbf{j}_1, \ldots, \mathbf{j}_k] \in \mathbb{R}^{n \times k},$$
$$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p \times n}$$

où $\mathbf{j}_\ell \in \mathbb{R}^n$ est le vecteur canonique de la distribution μ_ℓ, défini par $(\mathbf{j}_\ell)_i = \delta_{x_i \sim \mu_\ell}$. Les vecteurs $\mathbf{z}_i$ sont les versions centrées des $\mathbf{x}_i$, à savoir $\mathbf{z}_i = \mathbf{x}_i - \mathbf{m}_\ell$ pour $x_i \sim \mu_\ell$.

3 Comportement de la matrice $\mathbf{X}^\intercal\mathbf{X}$

Ces définitions et notations étant posées, nous nous penchons maintenant sur le comportement en grande dimension (pour n, p larges) de la matrice de Gram

$$\mathbf{G} = \frac{1}{p}\mathbf{X}^\intercal\mathbf{X} \tag{1}$$

qui est au cœur de nombreuses méthodes statistiques d'inférence, régression et classification. Notamment, une description fine du comportement de $\mathbf{G}$ donne accès au fonctionnement interne et aux performances de nombreuses méthodes d'apprentissage et de traitement des données, telles que le regroupement spectral, les SVMs, l'apprentissage semi-supervisé ou par transfert, etc.

Le résultat principal est le suivant : avec les notations introduites dans la Section 2.2, la matrice $\mathbf{G}$ peut s'écrire :

$$\mathbf{G} = \underbrace{\mathbf{J}\mathbf{M}^\intercal\mathbf{M}\mathbf{J}^\intercal}_{\text{information}} + \underbrace{\frac{1}{p}\mathbf{Z}^\intercal\mathbf{Z}}_{\text{bruit}} + * + o_p(1) \tag{2}$$

où $*$ contient un ensemble de termes peu utiles pour notre propos et $o_p(1)$ est compris dans le sens où la norme opérateur

1. Cette hypothèse peut évidemment être levée.

entre membres gauche et droite de l'équation tend vers zéro en probabilité lorsque $p, n \to \infty$ proportionnellement. Ainsi, $\mathbf{G}$ se décompose en une partie informative contenant les vecteurs canoniques $\mathbf{J}$ des classes et un bruit porté par la matrice $\mathbf{Z}^\intercal\mathbf{Z}$. Dans une analyse de rapport signal-à-bruit, si les moyennes des différentes classes sont suffisamment distinctes (à savoir, si $\|\mathbf{M}^\intercal\mathbf{M}\|$ est élevé), les vecteurs propres dominants de $\mathbf{G}$ seront fortement alignés aux vecteurs $\mathbf{J}$). Du fait de la forme du modèle (2), la TMA garantit même l'existence d'un seuil ζ fonction du rapport p/n tel que la détection des classes au moyen des vecteurs propres dominant de $\mathbf{G}$ (comme dans la méthode du regroupement spectral) n'est asymptotiquement possible que si $\|\mathbf{M}^\intercal\mathbf{M}\| > \zeta$.

Une étude fine de l'expansion (2) donne même accès à un *équivalent déterministe* de la résolvante $\mathbf{Q} \equiv (\frac{1}{p}\mathbf{X}^\intercal\mathbf{X} - z\mathbf{I}_p)^{-1}$ de $\frac{1}{p}\mathbf{X}^\intercal\mathbf{X}$, outil fondamental pour l'analyse d'un large spectre de méthode en apprentissage. On a précisément le résultat :

Proposition 1 (Équivalent déterministe pour $\frac{1}{p}\mathbf{X}^\intercal\mathbf{X}$, extension de [1])**.** *Sous les hypothèses de la Section 2.2, pour tout $z \in \mathbb{C}$ avec $\Im[z] > 0$ et $p/n \to c > 0$, on note $\alpha(z) = \frac{cz^{-1}}{1+\delta(z)}$*

$$\|\mathbb{E}[\mathbf{Q}] - \mathbf{Q}_\delta\| \to 0,$$
$$\mathbf{Q}_\delta = -\alpha(z)\left(\mathbf{I}_n + \alpha(z)\mathbf{U}(\mathbf{I}_k + \mathbf{U}^\intercal\mathbf{U})^{-1}\mathbf{U}^\intercal\right) + *',$$

où $\delta(z)$ est l'unique solution de $\delta(z) = \frac{1}{p}\operatorname{tr}(\mathbf{C}\bar{\mathbf{Q}}_\delta)$, $\bar{\mathbf{Q}}_\delta$ est l'équivalent déterministe développé dans [1] ($*'$ est une matrice déterministe non utile pour notre propos), et $\mathbf{U} = \mathbf{J}\mathbf{M}^\intercal$.

Ce résultat permet en outre de (i) localiser les valeurs propres limites de $\frac{1}{p}\mathbf{X}^\intercal\mathbf{X}$, (ii) de déterminer le seuil de détectabilité spectrale mentionné plus haut, (iii) d'évaluer le "contenu" asymptotique des vecteurs propres isolés de $\frac{1}{p}\mathbf{X}^\intercal\mathbf{X}$ et, de manière bien plus fondamentale, (iv) d'inférer les performances asymptotiques d'algorithmes d'apprentissage et inférence reposant sur des fonctions simples de $\frac{1}{p}\mathbf{X}^\intercal\mathbf{X}$ (LS-SVM, regroupement spectral, PCA, etc.).

En lisant la Proposition 1 dans le détail, on s'aperçoit que seuls les moments d'ordres un et deux des μ_ℓ suffisent à décrire le comportement spectral de $\frac{1}{p}\mathbf{X}^\intercal\mathbf{X}$. En outre, le résultat est strictement identique à celui relevé dans le cas d'un modèle GMM. Le comportement spectral asymptotique de $\frac{1}{p}\mathbf{X}^\intercal\mathbf{X}$ est donc en ce sens *universel*. Nous mettons en évidence cet aspect d'universalité dans la section suivante et soulevant l'importance de cette propriété dans le cadre de la modélisation des données réelles.

4 Images générées par un GAN : un exemple de données concentrées

Il est peu concevable de modéliser des données réelles par de simples vecteurs gaussiens ; cela impose notamment qu'un blanchiment linéaire rende les entrées du vecteur indépendantes. Les vecteurs concentrés analysés précédemment ne souffrent pas de cette contrainte.

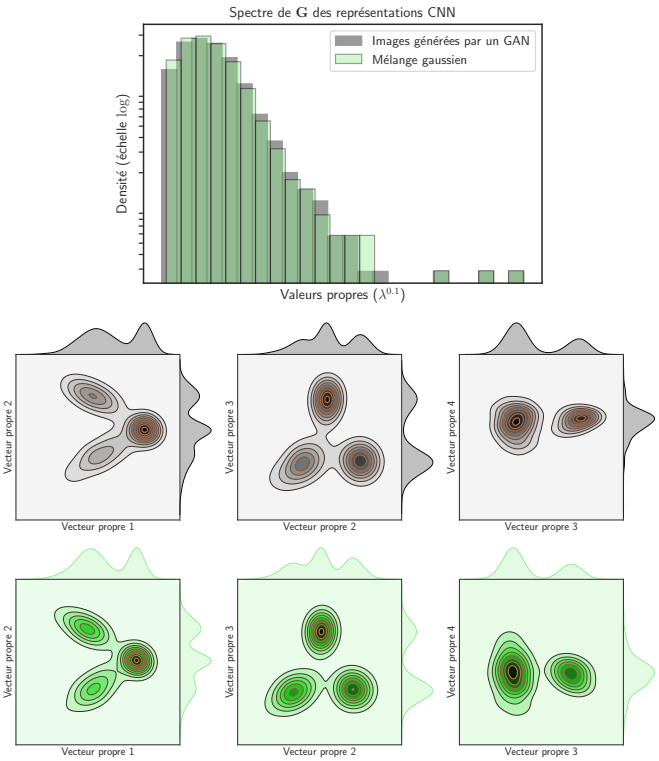

FIGURE 1 – (En haut) Histogramme des valeurs propres de la matrice $\mathbf{G}$ pour les représentations CNN des images MNIST générées par un GAN (en gris) et un modèle GMM (en vert) de mêmes moyennes et covariances. (En bas) Les vecteurs propres dominants correspondants aux valeurs propres isolées de la matrices $\mathbf{G}$ pour les images GAN (en gris) et pour le modèle GMM (en vert). Les classes considérées sont des images des chiffres $\{0, 1, 2\}$, $p = 784$ et $n = 500$.

Le second intérêt des vecteurs concentrés que nous discutons dans cette section est qu'ils sont déjà implicitement utilisés en pratique comme substituts efficaces à des vraies données. En effet, les dits réseaux de neurones génératifs adverses (GANs) produisent des données *réalistes* (notamment des images fidèles à la réalité) et consistent ni plus ni moins qu'en l'application successive d'opérations lipschitziennes sur des vecteurs gaussiens [7]. Schématiquement,

$$\text{Donnée Réelle} \approx \text{Donnée GAN} = \underbrace{\mathcal{F}_1 \circ \cdots \circ \mathcal{F}_N}_{\text{opérations Lipschitz}}(\mathcal{N}(0, \mathbf{I})).$$

Comme les vecteurs gaussiens standards sont concentrés (Remarque 1) et que toute opération lipschitzienne maintient la concentration (Remarque 2), les vecteurs de données issus des GANs sont des vecteurs concentrés. La proximité entre données réelles et données générées par des GANs donne ainsi à penser que les mélanges de vecteurs concentrés forment un cadre très réaliste de modélisation de données réelles.

Dans ce qui suit, nous illustrons l'équivalence vis-à-vis des propriétés spectrales de la matrice $\mathbf{G}$ entre le modèle GMM et les mélanges de vecteurs concentrés sur des données générées par un GAN. À cette fin, nous considérons (i) des vecteurs

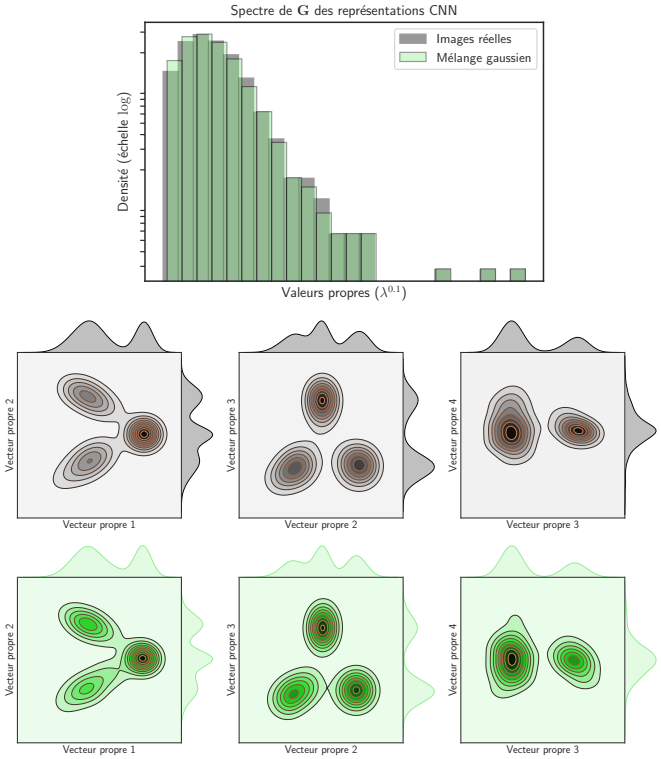

FIGURE 2 – (En haut) Histogramme des valeurs propres de la matrice $\mathbf{G}$ pour les représentations CNN des images MNIST réelles de test (en gris) et un modèle GMM (en vert) de mêmes moyennes et covariances. (En bas) Les vecteurs propres dominants correspondants aux valeurs propres isolées de la matrices $\mathbf{G}$ pour les images MNIST réelles (en gris) et pour le modèle GMM (en vert). Les classes considérées sont des images des chiffres $\{0, 1, 2\}$, $p = 784$ et $n = 500$.

$\mathbf{x}_i$ de représentations CNN (à savoir obtenus à la sortie d'un réseau de neurones convolutionnel entraîné à la classification d'images) d'images générées par un GAN dont nous comparons le comportement à (ii) des images réelles.

Pour ce faire, nous avons généré (à travers un GAN appris sur MNIST) des images des classes $\{0, 1, 2\}$ et utilisé leurs représentations CNN [2] comme exemple de données concentrées. L'équivalent GMM de ces données est donc obtenu en estimant empiriquement les moyennes et covariances au sein de chaque classe. La Figure 1 représente l'histogramme des valeurs propres ainsi que l'espace propre dominant (qui contient les informations sur les classes) de la matrice $\mathbf{G}$, pour les données générées et leur mélange gaussien équivalent (à savoir de même moyennes et covariances). Nous constatons que la matrice $\mathbf{G}$ présente le même comportement en terme de spectre et d'espaces propres dominants, validant ainsi l'aspect universel de cette matrice par rapport à la distribution des données et rendant ainsi l'estimation des performances de classification prédictible à travers la théorie des matrices aléatoires, comme avancé en Section 3. De plus, du fait de la capacité des GANs à générer des données visuellement très similaires aux données réelles, on observe le même comportement de la matrice $\mathbf{G}$ des données réelles, comme illustré sur la Figure 2. Cette observation valide ainsi la pertinence du modèle de *mélanges de vecteurs concentrés* [1] pour le traitement des données réelles.

2. Avant dernière couche d'un réseau CNN appris à classifier les 10 classes du dataset MNIST.

5 Conclusion

Cet article a mis en avant la pertinence d'une modélisation de données réalistes, notamment en traitement des images, par des vecteurs satisfaisant le phénomène de concentration de la mesure. Une propriété fondamentale d'universalité, seulement valable en grandes dimensions, révèle qu'il suffit par ailleurs de traiter mathématiquement ces vecteurs comme de simples mélanges gaussiens, sans perte de généralité. Ce constat est à la fois surprenant mais pratiquement rassurant : il constitue un premier pas notable vers l'espoir futur d'une compréhension et d'une maîtrise accrues d'algorithmes avancés d'apprentissage pour des données structurées complexes. En outre, l'article a démontré explicitement notre capacité, à travers cette théorie, à anticiper les performances de classifieurs, régresseurs et estimateurs pour des données aussi réalistes que les surprenantes images générées par des GANs.

Références

[1] Cosme Louart and Romain Couillet. Concentration of measure and large random matrices with an application to sample covariance matrices. *submitted*, 2019.

[2] Zhenyu Liao and Romain Couillet. Random matrices meet machine learning : A large dimensional analysis of ls-svm. In *ICASSP*, pages 2397–2401. IEEE, 2017.

[3] Xiaoyi Mai and Romain Couillet. A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *arXiv preprint arXiv :1711.03404*, 2017.

[4] Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet. A kernel random matrix-based approach for sparse PCA. In *ICLR*, 2019.

[5] Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1) :1393–1454, 2016.

[6] Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet. Kernel random matrices of large concentrated data : The example of gan-generated images. In *ICASSP*, 2019.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.