On some theoretical limitations of Generative Adversarial Networks

Benoit Oriol, Alexandre Miot

On some theoretical limitations of Generative Adversarial Networks

Oriol, Benoît
benoit.oriol@polytechnique.edu
École Polytechnique
Palaiseau, 91120, France

Miot, Alexandre
alexandre.miot@sgcib.com
Société Générale
La Défense, 92800, France

October 22, 2021

Abstract

Generative Adversarial Networks have become a core technique in Machine Learning to generate unknown distributions from data samples. They have been used in a wide range of context without paying much attention to the possible theoretical limitations of those models. Indeed, because of the universal approximation properties of Neural Networks, it is a general assumption that GANs can generate any probability distribution. Recently, people began to question this assumption and this article is in line with this thinking. We provide a new result based on extreme value theory showing that GANs can't generate heavy tailed distributions. The full proof of this result is given.

1 Introduction

The universal approximation property of neural networks (see [8] and [4]) might make us assume that GANs can simulate any distribution from a Gaussian prior. However, neural networks, as functions are by design almost everywhere differentiable functions with bounded derivatives to limit exploding gradients phenomenons (see [10]). By Rademacher (see [7] for a proof) and mean value theorems, this is nearly equivalent to say that neural networks functions are Lipschitz continuous. This fact basically sets the limitations of GANs to express any probability distribution given a Gaussian prior. The are numerous definitions of the concept of "fat", "longed" or "heavy" tailed distributions. They are usually not equivalent but all convey a sense of having a larger probability of being "big" compared to a Gaussian or Exponential distribution. Here we focus on two possible ways to define the concept. One, similarly to [12], is focussing on finite samples and relies on classical concentration inequalities. The other is asymptotic and uses Extreme Value Theory to prove a new theorem in the continuity of the theoretical work of Huster et al. in [9] and the experimental approach of [6].

Notations, in the following, we make use of the following notations:

- f is a Lipschitz function $\mathbb{R}^n \mapsto \mathbb{R}$, $||f||_l = \sup_{x,y\in\mathbb{R}^n, x\neq y} \frac{||f(x)-f(y)||}{||x-y||}$ its semi-norm,

- γ_n the Gaussian measure on $\mathbb{R}^n$,
- d the Euclidean distance in $\mathbb{R}^n$ i.e. $d(x,y) = ||y-x||$,
- for any set $S \subset \mathbb{R}^n$ the ϵ neighbourhood $S_\epsilon = \{x \in \mathbb{R}^n \text{ such that } d(x,S) < \epsilon\}$, where $\epsilon > 0$,
- $\bar{A}$ the complement of a subset $A \subset \mathbb{R}^n$,
- M the median of a mapping $\mathbb{R}^n \mapsto \mathbb{R}$ for the γ_n measure i.e. $\gamma_n(\{f \geq M\}) \geq \frac{1}{2}$ and $\gamma_n(\{f \leq M\}) \geq \frac{1}{2}$,
- X a standard Gaussian random variable in $\mathbb{R}^n$,
- $\bar{\Psi} = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} e^{-u^2/2} du$ the Gaussian tail function.

2 Limitations through sub-gaussianity

In this section we prove that given a Lipschitz function and a Gaussian prior X, $f(X)$ is sub-gaussian: a GAN with a Gaussian prior can only generate sub-gaussian distributions.

Definition 1. *A real valued random variable Y is said to be sub-gaussian if it satisfies one of the two following equivalent properties :*

$$\Updownarrow \begin{array}{l} \exists K \in \mathbb{R}, \forall t \geq 0, P(|Y| \geq t) \leq 2e^{-\frac{t^2}{K^2}} \\ \exists K' \in \mathbb{R}, \forall p \geq 1, ||X||_{L^p} \leq K'\sqrt{p} \ . \end{array}$$

For a proof see [15].

If $G = f(X)$, where f is Lipschitz, by Lipschitz continuity and $\mathbb{E}(X) = 0$

$$\forall x \geq 0 \quad \mathbb{P}(|G - f(0)| \geq x) \leq \mathbb{P}\left(|X| \geq \frac{x}{||f||_l}\right) \ .$$

In particular if X is one dimensional then using a standard upper bound of the gaussian tail function G will be sub-gaussian as a sum of two independant sub-gaussian functions, considering $f(0)$ as a constant random variable.

If X was n dimensional then,

$$\mathbb{P}\left(|X| \geq \frac{x}{||f||_l}\right) = \mathbb{P}\left(|X|^2 \geq \frac{x^2}{||f||_l^2}\right) \ ,$$

and $|X|^2$ following a χ^2 distribution with n degrees of freedom the sub-gaussianity of G would seem to be dependent on the dimension of X. Yet, this is not the case as stated by the following remarkable result:

Theorem 1 (Gaussian concentration theorem [14], [13] and [3])**.** *Let X be a standard gaussian random variable on $\mathbb{R}^n$ and f a Lipschitz function then $f(X) - \mathbb{E}(f(X))$ is sub-gaussian. More precisely,*

$$\forall \epsilon > 0, \mathbb{P}(|f(X) - \mathbb{E}(f(X))| \geq \epsilon) \leq 2e^{-2\epsilon^2/||f||_l^2} \ .$$

So in particular, a GAN with a Gaussian prior will not be able to generate any realistic samples even if trained on a "fat tailed" distribution. This is not the first time that concentration of measure gives some strong theoretical limits to machine learning methods see [5] or [11] for a more recent paper. Limitations of GANs has also been explored from a different perspective in [16].

This theorem is also true when we replace the mean $\mathbb{E}(f(X))$ by a median of $f(X)$. The original proof of the theorem can be found in [14]. The proof is quite technical, a more accessible one can be found in [2] . To get a sense of the concentration of measure phenomenon we

provide here a simple proof with the median.

The gaussian concentration theorem is a consequence of:

Theorem 2 (Gaussian Isoperimetric Theorem [3])**.** *Let's A be a Borel set in $\mathbb{R}^n$ and $H = \{x \in \mathbb{R}^n$ such that $x_1 < a\}$ with $a \in \mathbb{R}$ such that $\gamma_n(A) = \gamma_n(H)$ then*

$$\forall \epsilon \geq 0 \; \gamma_n(A_\epsilon) \geq \gamma_n(H_\epsilon) \; .$$

It is easily seen that $\gamma_n(H_r) = \Psi(a+r)$ where Ψ is the cumulative distributive function of the one dimensional standard Gaussian distribution. It is not obvious at first sight what is the link between this theorem and the Gaussian concentration theorem for Lipschitz functions. The link is made defining the following 'isoperimetric function' for $a \in [0,1]$ and $\epsilon > 0$

$$\begin{aligned} \eta_a(\epsilon) &= \sup_{A \text{ borel set } \subset \mathbb{R}^n} \{\gamma_n(\bar{A_\epsilon}) \mid \gamma_n(A) \geq a\} \\ &= 1 - \inf_{A \text{ borel set } \subset \mathbb{R}^n} \{\gamma_n(A_\epsilon) \mid \gamma_n(A) \geq a\} \; . \end{aligned}$$

Lemma 1. *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a Lipschitz function and M a median for the Gaussian measure, then*

$$\forall \epsilon > 0 \quad \gamma_n(|f - M| > \epsilon) \leq \eta_{\frac{1}{2}} \left(\frac{\epsilon}{||f||_l} \right) \; .$$

Proof. Let $A = \{f \leq M\}$, $\epsilon > 0$ and $x \in A_\epsilon$ then

$$\exists y \in A \text{ such that } d(x, A) \leq d(x, y) < \epsilon$$

so, f being Lipschitz and $y \in A$

$$|f(x) - f(y)| \leq ||f||_l \; \epsilon \; .$$

So, $f(x) \leq M + ||f||_l \; \epsilon$ i.e. $\{f \geq M + ||f||_l \; \epsilon\} \subset \bar{A_\epsilon}$. Changing $\epsilon \to \frac{\epsilon}{||f||_l}$ we have proved that for any Lipschitz function f of median M

$$\forall \epsilon > 0 \quad \gamma_n(\{f > M + \epsilon\}) \leq \eta_{\frac{1}{2}} \left(\frac{\epsilon}{||f||_l} \right) \; .$$

Noticing that f if Lipschitz iif $-f$ is, $||f||_l = ||-f||_l$, if M is a median of f then $-M$ is a median of $-f$ and applying what we just proved to $-f$ the result follows. □

We can now prove the Gaussian concentration theorem.

Proof. Let A be a Borel set such that $\gamma_n(A) \geq \frac{1}{2}$, there exists a half-space $H = \mathbb{R}^{n-1} \times]-\infty, a[$ such that $\gamma_n(A) = \gamma_n(H) = \Psi(a) \geq \frac{1}{2}$. From the Isoperimetric Gaussian Theorem

$$\begin{aligned} \forall \epsilon \geq 0, \quad & \gamma_n(A_\epsilon) \geq \gamma_n(H_\epsilon) \\ & \gamma_n(A_\epsilon) \geq \Psi(a + \epsilon) \geq \Psi(\epsilon) \; . \end{aligned}$$

Ψ being non decreasing and $a \geq 0$ as $\Psi(a) \geq \frac{1}{2}$. Taking the infinum on the left side and noticing that the infinimum is reached for H

$$\forall \epsilon > 0,$$

$$\inf_{A \text{ borel set } \subset \mathbb{R}^n} \{\gamma_n(A_\epsilon) \mid \gamma_n(A) \geq \frac{1}{2}\} = \Psi(\epsilon) \; .$$

That is to say,

$$\eta_{\frac{1}{2}} = 1 - \Psi = \bar{\Psi}.$$

□

3 Limitations through Extreme Value Theory

In this section, we prove the main result of this paper: given a Lipschitz function and a Gaussian

prior X, if $f(X)$ is in a domain of attraction of an extreme value distribution of parameter ξ then $\xi \leq 0$. In particular, $f(X)$ can't be "heavy tailed". In fact, we prove the theorem for a wider range of distributions.
In the following, we use the notations of Extreme Value Theory that are introduced in [9].

Theorem 3. *Let $n \in \mathbb{R}^*$, $f : \mathbb{R}^d \to \mathbb{R}$ a $\mathcal{C}^1$ a.e Lipschitz function with semi norm $L = ||f||_l$. Let G_k^d be a real random variable of probability distribution function $g_{G_k^d}(x) \propto \|x\|_2^k e^{-\frac{\|x\|_2^2}{2}}$, where $k \in \mathbb{N}$. If $f(G_k^d)$ is in the domain of attraction of the extreme value distribution of parameter $\xi \in \mathbb{R}$, i.e $f(G_k^d) \in \mathcal{D}(\mathcal{H}_\xi)$, then $\xi \leq 0$.*

Proof. **Case $d = 1$.** We prove by contradiction that $\xi \leq 0$. Supposing that $\xi > 0$, by theorem 8.a [1], $\forall\, \gamma \in]0, \xi[$, $\mathbb{E}[f(G_k^1)^\gamma]$ is finite and

$$\begin{aligned} c_\gamma &= \lim_{t\to\infty} \mathbb{E}\left[\left(\frac{f(G_k^1)}{t}\right)^\gamma | f(G_k^1) > t\right] \\ &= \left(1 - \frac{\gamma}{\xi}\right)^{-1} . \end{aligned}$$

Let $\gamma \in]0, \xi[$. We are only interested in the behaviour of the previous integral when t goes to $+\infty$ so we can suppose that $t > f(0) + 1$ and $t > \sqrt{|k-1| + \gamma} L + |f(0)|$.
$f^{-1}(]t, \infty[)$ is an open set of $\mathbb{R}$ by continuity. Open intervals are a countable base of $\mathbb{R}$, so we can write $f^{-1}(]t, \infty[) = \bigcup_{i\in\mathcal{I}}]a_i, b_i[$ where $\mathcal{I}$ is finite or countable. We can also suppose that any of those intervals are disjoints. So we can suppose that :

- $f^{-1}(]t, \infty[) = \bigcup_{i\in\mathcal{I}}]a_i, b_i[$ where $\mathcal{I}$ is finite or countable
- $0 \notin]a_i, b_i[$, $a_i \neq 0$, $b_i \neq 0$ and f is strictly positive on each interval
- $-\infty \leq a_0 < b_0 \leq a_1 < b_2 \leq \ldots < b_{m^*} \leq +\infty$, where m^* is equal to m or ∞ according to $\mathcal{I}$ cardinality
- $\forall i \in \mathcal{I}$ $f(a_i) = t$ if $a_i > -\infty$ by continuity of f
- $\forall i \in \mathcal{I}$ $f(b_i) = t$ if $b_i < \infty$ by continuity of f
- $\forall x \in]a_i, b_i[$ $|x| > \frac{t - f(0)}{L}$ as f is L-Lipschitz and $]a_i, b_i[\subset f^{-1}(]t, +\infty[$. In particular, noting $t^* = \min(t, t - f(0))$, $|x| > \frac{t^*}{L}$

Also, we are only interested in $t \to +\infty$ so we can suppose ${t^*}^2 > |k-1|L^2$. If $\mathcal{I} = \emptyset$, the case is trivial: $\mathbb{E}\left[\left(\frac{f(G_k^1)}{t}\right)^\gamma | f(G_k^1) > t\right]$ is not defined, which is a contradiction.
Otherwise, the conditional expectation is well defined and finite and we have:

$$\mathbb{E}\left[\left(\frac{f(G_k^1)}{t}\right)^\gamma | f(G_k^1) > t\right] = \frac{\sum_{i\in\mathcal{I}} \int_{a_i}^{b_i} \left(\frac{f(x)}{t}\right)^\gamma |x|^n e^{-\frac{x^2}{2}} dx}{\sum_{i\in\mathcal{I}} \int_{a_i}^{b_i} |x|^n e^{-\frac{x^2}{2}} dx} . \quad (1)$$

Let $i \in \mathcal{I}$. For the numerator, integrating by part:

$$\int_{a_i}^{b_i} \left(\frac{f(x)}{t}\right)^\gamma |x|^k e^{-\frac{x^2}{2}} dx = \left[-\frac{|x|^k}{x}\left(\frac{f(x)}{t}\right)^\gamma e^{-\frac{x^2}{2}}\right]_{a_i}^{b_i} + \int_{a_i}^{b_i} \underbrace{\left(\frac{k-1}{x^2} + \gamma\frac{f'(x)}{x f(x)}\right)}_{M} \left(\frac{f(x)}{t}\right)^\gamma |x|^k e^{-\frac{x^2}{2}} dx$$

The first integrated part is equal to $\left[-\frac{|x|^k}{x} e^{-\frac{x^2}{2}}\right]_{a_i}^{b_i}$, as we have seen on the interval

bounds either f is equal to t or $f = O_{\pm\infty}(x)$. We can bound the M term as $|x| \geq \frac{t^*}{L}$, $f(x) > t > 0$ and $|f'| < L$,

$$|M| \leq |k-1|\frac{L^2}{t^{*2}} + \gamma\frac{L^2}{t^{*2}} \ .$$

We deduce the following inequalities:

$$\int_{a_i}^{b_i} \left(\frac{f(x)}{t}\right)^\gamma |x|^k e^{-\frac{x^2}{2}} dx \leq \left[-|x|^{k-1}e^{-\frac{x^2}{2}}\right]_{a_i}^{b_i} + \frac{(|k-1|+\gamma)L^2}{t^{*2}} \int_{a_i}^{b_i} \left(\frac{f(x)}{t}\right)^\gamma |x|^k e^{-\frac{x^2}{2}} dx \ , \quad (2)$$

$$\int_{a_i}^{b_i} \left(\frac{f(x)}{t}\right)^\gamma |x|^k e^{-\frac{x^2}{2}} dx \geq \left[-|x|^{k-1}e^{-\frac{x^2}{2}}\right]_{a_i}^{b_i} - \frac{(|k-1|+\gamma)L^2}{t^{*2}} \int_{a_i}^{b_i} \left(\frac{f(x)}{t}\right)^\gamma |x|^k e^{-\frac{x^2}{2}} dx \ . \quad (3)$$

The denominator has still a dependance on f through the domain of integration, so $|x| > \frac{t^*}{L}$ is still valid and similarly:

$$\int_{a_i}^{b_i} |x|^k e^{-\frac{x^2}{2}} dx \leq \left[-|x|^{k-1}e^{-\frac{x^2}{2}}\right]_{a_i}^{b_i} + \frac{|k-1|L^2}{t^{*2}} \int_{a_i}^{b_i} |x|^k e^{-\frac{x^2}{2}} dx \ , \quad (4)$$

$$\int_{a_i}^{b_i} |x|^k e^{-\frac{x^2}{2}} dx \geq \left[-|x|^{k-1}e^{-\frac{x^2}{2}}\right]_{a_i}^{b_i} - \frac{|k-1|L^2}{t^{*2}} \int_{a_i}^{b_i} |x|^k e^{-\frac{x^2}{2}} dx \ . \quad (5)$$

Combining equations (3) and (4) in (1), we have:

$$\mathbb{E}\left[\left(\frac{f(G_k^1)}{t}\right)^\gamma | f(G_k^1) > t\right] \geq \frac{1 - \frac{|k-1|L^2}{t^{*2}}}{1 + \frac{(|k-1|+\gamma)L^2}{t^{*2}}} \ .$$

And combining (2) and (5) in (1), as we chose $t > \sqrt{|k-1|+\gamma}L + |f(0)|$, we have:

$$\mathbb{E}\left[\left(\frac{f(G_k^1)}{t}\right)^\gamma | f(G_k^1) > t\right] \leq \frac{1 + \frac{|k-1|L^2}{t^{*2}}}{1 - \frac{(|k-1|+\gamma)L^2}{t^{*2}}}$$

So $c_\gamma = \lim_{t\to\infty} \mathbb{E}\left[\left(\frac{f(G_k^1)}{t}\right)^\gamma | f(G_k^1) > t\right] = 1$. Assuming $f(G_k^1) \in \mathcal{D}(\mathcal{H}_\xi), \xi > 0$, entails $c_\gamma = (1 - \frac{\gamma}{\xi})^{-1}$ and $\gamma = 0$. We conclude that $\xi \leq 0$.

Case $d \in \mathbb{N}^*$**.** We prove that $\xi \leq 0$ by contradiction. If $\xi > 0$ using theorem 8.a [1], $\forall 0 < \gamma < \xi$, $\mathbb{E}[f(G_k^d)^\gamma]$ is finite and

$$c_\gamma = \lim_{t\to\infty} \mathbb{E}\left[\left(\frac{f(G_k^d)}{t}\right)^\gamma | f(G_k^d) > t\right] = \left(1 - \frac{\gamma}{\xi}\right)^{-1} \ .$$

Let $\gamma \in]0, \xi[$ and $t \in \mathbb{R}_+^*$ such that $t > f(0_d)$ and $t > L\sqrt{|k+d-2|+\gamma}L + |f(0)|$. Using the hyperspherical coordinates, we introduce the operator $H : \mathcal{L}([0,\pi]^{d-2} \times [0, 2\pi]) \mapsto \mathbb{R}$:

$$H(f) = \alpha \int_0^\pi ... \int_0^\pi \int_0^{2\pi} \prod_{i=1}^{d-2} \sin(\theta_i)^{d-i-1} f(\theta) d\theta,$$

with α the normalising term for the G_k^d distribution and $d\theta = d\theta_1 ... d\theta_{d-1}$.

Then, we have:

$$\mathbb{E}\left[\left(\frac{f(G_k^d)}{t}\right)^\gamma 1_{f(G_k^d)>t}\right] =$$
$$H\Big(\theta \mapsto \underbrace{\int_0^{+\infty} 1_{f_\theta(r)>t}\left(\frac{f_\theta(r)}{t}\right)^\gamma r^{k+d-1}e^{-\frac{r^2}{2}}dr}_{\mathbb{E}\left[\left(\frac{f_\theta(G_{k+d-1}^1)}{t}\right)^\gamma 1_{f_\theta(G_{k+d-1}^1)>t}\right]}\Big)$$

with $x = r(x_1, ..., x_d)$ and

$$\begin{aligned} x_1 &= \sin(\theta_1) \\ x_2 &= \sin(\theta_1)\cos(\theta_2) \\ &\vdots \\ x_{d-1} &= \sin(\theta_1)\sin(\theta_2)\dots\cos(\theta_{d-1}) \\ x_d &= \sin(\theta_1)\sin(\theta_2)\dots\sin(\theta_{d-1}) \ . \end{aligned}$$

For $\theta = (\theta_1, ...\theta_{d-1})$, $f_\theta : r \in \mathbb{R}_+ \mapsto f(rx_1, ..., rx_d)$ is L-Lipschitz as $(x_1, \dots, x_d)$ is on the unit sphere. Also, $f(0) = f_\theta(0)$. We can use the bounds from the 1-dimensional proof. We note:

$$M_+ = \frac{1 + \frac{|k+d-2|L^2}{t^{*2}}}{1 - \frac{(|k+d-2|+\gamma)L^2}{t^{*2}}}$$
$$M_- = \frac{1 - \frac{|k+d-2|L^2}{t^{*2}}}{1 + \frac{(|k+d-2|+\gamma)L^2}{t^{*2}}} \ .$$

We obtain:

$$\mathbb{E}\left[\left(\frac{f(G_k^d)}{t}\right)^\gamma 1_{f(G_k^d)>t}\right] \leq$$
$$M_+ H\left(\theta \mapsto \mathbb{P}(1_{f_\theta(G_{k+d-1}^1)>t})\right) \ .$$

That is to say,

$$\mathbb{E}\left[\left(\frac{f(G_k^d)}{t}\right)^\gamma 1_{f(G_k^d)>t}\right] \leq M_+\mathbb{P}(f(G_k^d) > t) \ .$$

Similarly, we have:

$$\mathbb{E}\left[\left(\frac{f(G_k^d)}{t}\right)^\gamma 1_{f(G_k^d)>t}\right] \geq M_-\mathbb{P}(f(G_k^d) > t) \ .$$

Thus, we conclude similarly that c_γ is well-defined, finite, and $c_\gamma = 1$ that is $\gamma = 0$, which is absurd as $\gamma > 0$. □

4 Conclusion and future work

Because of the intrinsic Lipschitz characteristics of Neural Networks, GANs expressivity is limited. In particular, a Gaussian prior cannot be used to simulate heavy tailed distributions. In the EVT framework, the question of the existence of a tail index for the generated distribution, or the conditions for its existence, remains. A theoretical partial answer is given in [9] for GANs with ReLU or Leaky-ReLU activation functions and a finite number of neurons. The general case is still an open question. Likewise, determining the thinnest tail prior being able to simulate samples exhibiting heavy tails is an important question needing further investigations.
Moreover, experimentally, the problem of training GANs with a heavy-tailed prior remains too. Indeed, with such priors GANs are hard to train and exhibit numerical instabilities.

References

[1] A. A. Balkema and Laurens de Haan. Residual life time at great age. *The Annals of Probability*, 1974.

[2] Sergei Bobkov. An isoperimetric inequality on the discrete cube, and an elementary proof of the isoperimetric inequality in gauss space. *The Annals of Probability*, 1997.

[3] Christer Borell. The brunn-minkowski inequality in gauss space. *Inventiones mathematicae*, 1975.

[4] George V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 1989.

[5] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. *ArXiv*, 2018.

[6] Richard Feder, Philippe Berger, and George Stein. Nonlinear 3d cosmic web simulation with heavy-tailed generative adversarial networks. *Phys. Rev. D*, 2020.

[7] Juha Heinonen. Lectures on lipschitz analysis. *Rep. Dept. Math. Stat*, 2005.

[8] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 1989.

[9] Todd Huster, Jeremy Cohen, Zinan Lin, Kevin Chan, Charles Kamhoua, Nandi O. Leslie, Cho-Yu Jason Chiang, and Vyas Sekar. Pareto gan: Extending the representational power of gans to heavy-tailed distributions. In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2021.

[10] Razvan Pascanu, Tomás Mikolov, and Yoshua Bengio. Understanding the exploding gradient problem. *CoRR*, 2012.

[11] Jack Prescott, Xiao Zhang, and David Evans. Improved estimation of concentration under L_p-norm distance metrics using half spaces. *International Conference on Learning Representations*, 2021.

[12] Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning representations of GAN-data behave as Gaussian mixtures. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2020.

[13] Vladimir Sudakov and Boris Tsirelson. Extremal properties of half-spaces for spherically invariant measures. *Journal of Soviet Mathematics*, 1978.

[14] Boris Tsirelson, Ildar Ibragimov, and Vladimir Sudakov. Norms of gaussian sample functions. *Proceedings of the Third Japan — USSR Symposium on Probability Theory*, 1976.

[15] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[16] Magnus Wiese, Robert Knobloch, and Ralf Korn. Copula & marginal flows: Disentangling the marginal from its joint. *CoRR*, 2019.