Limits of Deepfake Detection: A Robust Estimation Viewpoint

Sakshi Agarwal and Lav R. Varshney, *Senior Member, IEEE*

Abstract

Deepfake detection is formulated as a hypothesis testing problem to classify an image as genuine or GAN-generated. A robust statistics view of GANs is considered to bound the error probability for various GAN implementations in terms of their performance. The bounds are further simplified using a Euclidean approximation for the low error regime. Lastly, relationships between error probability and epidemic thresholds for spreading processes in networks are established.

Index Terms

authentication, detection, estimation, generative adversarial networks, hypothesis testing, misinformation

I. INTRODUCTION

Deepfake, a portmanteau of "deep learning" and "fake" refers to realistic audiovisual content artificially generated using advanced generative algorithms like generative adversarial networks (GANs), with an implied use for unethical purposes. Concerns have been raised over the dissemination of misinformation via online channels [1], and the mistrust that deepfakes engender [2].

The gravity of the deepfake problem has been outlined in [3]. Technologies capable of generating hyperrealistic fake images and videos are used for dating scams, "astroturfing," "catfishing," and to gain the victim's trust for blackmail, harassment, or sabotage [4]. With an active community of developers creating free and easy tools to commoditize such technology, there is eroded trust in visual content [5]. Such "democratization of fraud," can have particularly grave consequences for politics and international affairs where it could be used to incite violence, discredit leaders and institutions, or even tip elections [6] and exacerbate disinformation wars that disrupt domestic politics [7], [8], [9].

Such alarming consequences of deepfakes present an urgent need to computationally discern fake content from genuine as they are almost indistinguishable to humans. While several studies use features extracted based on visual artifacts, image quality, lipsync, blinking, or warping for classification (and might soon be obsolete) [10], [11], [12], [13], this work gives a generalizable statistical framework with guarantees on its reliability. In particular, we build on the information-theoretic study of authentication [14] to cast deepfake detection as a hypothesis testing problem specifically for outputs of GANs, themselves viewed through a generalized robust statistics framework [15], [16].

II. PROBLEM FORMULATION

A. GAN Formulation

The GAN trains on infinite sample images with distribution $\mathbb{P}_X$, to select an optimal generator function $\hat{g}$ from a family of functions $\mathcal{G}$, which takes an input of Gaussian noise $Z \sim \mathcal{N}(0, I_r)$ to generate an output distribution $\mathbb{P}_{\hat{g}(Z)}$ closest to the input distribution as follows:

$$\hat{g} = \arg \min_{g \in \mathcal{G}} L(\mathbb{P}_X, \mathbb{P}_{g(Z)}). \tag{1}$$

Here $L : \mathcal{M}(\mathbb{R}^d) \times \mathcal{M}(\mathbb{R}^d) \to \mathbb{R}_{\geq 0}$ measures the distance between the two distributions. We assume that the minimum is attained. The alternate case can be dealt with by a standard limiting argument.

1) Perturbation View of GAN : Intuitively, the generator family $\mathcal{G}$ of the GAN should be designed such that $\inf_{g \in \mathcal{G}} L(\mathbb{P}_X, \mathbb{P}_{g(Z)})$ is small. We view GANs as a generalization of the robust statistics framework [16] with the true distribution $\mathbb{P}_X$ as a slightly perturbed version of a generated distribution $\mathbb{P}_{g(Z)}$ under the distance measure $L(\cdot, \cdot)$. Define:

$$OPT := \inf_{g \in \mathcal{G}} L(\mathbb{P}_X, \mathbb{P}_{g(Z)}) \tag{2}$$

as the oracle error, which is the minimum distance between the generated and input distributions, for a particular GAN and input distribution. The oracle error is fixed for a choice of L and $\mathcal{G}$. Approximately solving the minimization problem by training on finite samples from X, we still get a generated distribution within $O(OPT)$ distance of the true input distribution [16]. For error analysis, we choose the following L functions — Kullback-Leibler (KL) divergence, total variational distance, Jensen-Shannon divergence [17], and Wasserstein metric [18].

B. Hypothesis Test

The distribution of legitimate images is $\mathbb{P}_X$. The distribution generated by the GAN is $\mathbb{P}_{g(Z)}$. The n i.i.d. pixels of the input image are $Y_1, Y_2, Y_3, \ldots, Y_n$. Accordingly, the test is

$$H_0 := Y \sim \mathbb{P}_X \qquad\qquad \text{—Legitimate}$$
$$H_1 := Y \sim \mathbb{P}_{\hat{g}(Z)} \qquad\qquad \text{—GAN generated}$$

where $\hat{g}$ is close to the optimal result defined in (1). The assumption of i.i.d. pixels is to make the hypothesis test easier than the non i.i.d. case seen in realistic images.

III. ERROR BOUNDS

A. General Hypothesis Testing Bounds

We have the following error bounds [19, ch. 11] for hypothesis testing:

$$\text{Neyman-Pearson error: } \beta_n^\epsilon \doteq \exp(-nD(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)})) \tag{3}$$
$$\text{Bayesian error: } P_e^{(n)} \leq \exp\{-nC(\mathbb{P}_X, \mathbb{P}_{\hat{g}(Z)})\}. \tag{4}$$

where C is the Chernoff information, which can further be bounded in terms of total variational distance [20].

$$C(\mathbb{P}_X, \mathbb{P}_{\hat{g}(Z)}) \geq -\frac{1}{2}\log(1 - TV(\mathbb{P}_X, \mathbb{P}_{\hat{g}(Z)})^2) \tag{5}$$

We assume n is sufficiently large for asymptotic bounds to be valid, which mostly holds as the images we generally deal with are fairly high resolution (approximately 10^6).

B. Deriving Bounds for Specific L-functions

We derive bounds for the Neyman-Pearson and Bayesian error probabilities for the chosen L-function implementations of the GAN.

1) KL Divergence: We have

$$L(\mathbb{P}_X, \mathbb{P}_{g(Z)}) = D(\mathbb{P}_X||\mathbb{P}_{g(Z)}) \geq OPT \qquad \forall g \in \mathcal{G}.$$

Neyman-Pearson:

$$\beta_n^\epsilon \doteq \exp(-nD(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)})) \leq \exp(-nOPT). \tag{6}$$

Bayesian: From the bound on Chernoff information in (5)

$$P_e^{(n)} \leq \exp\{\frac{n}{2}\log(1 - TV(\mathbb{P}_X, \mathbb{P}_{\hat{g}(Z)})^2)\} \tag{7}$$

and from reverse Pinsker's inequality (A.54), assuming $P_g^* = \min_x \mathbb{P}_{\hat{g}(Z)}(x) > 0$,

$$P_e^{(n)} \leq \exp\{\frac{n}{2}\log(1 - \frac{P_g^*}{4}D(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)}))\} \tag{8}$$
$$\leq \exp\{\frac{n}{2}\log(1 - \frac{P_g^*}{4}OPT)\} \tag{9}$$
$$= \left(1 - \frac{P_g^*}{4}OPT\right)^{n/2}. \tag{10}$$

2) Total Variational Distance: We have

$$L(\mathbb{P}_X, \mathbb{P}_{g(Z)}) = TV(\mathbb{P}_X, \mathbb{P}_{g(Z)}) \geq OPT \qquad \forall g \in \mathcal{G}.$$

Neyman-Pearson: Using Pinsker's inequality (A.53)

$$\beta_n^\epsilon \leq \exp(-2nTV(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)})^2) \tag{11}$$
$$\leq \exp(-2nOPT^2). \tag{12}$$

Bayesian: From the bound on Chernoff information in (5)

$$P_e^{(n)} \leq \exp\{\frac{n}{2}\log(1 - TV(\mathbb{P}_X, \mathbb{P}_{\hat{g}(Z)})^2)\} \tag{13}$$
$$\leq \exp\{\frac{n}{2}\log(1 - OPT^2) \tag{14}$$
$$= (1 - OPT^2)^{n/2}. \tag{15}$$

TABLE I
SUMMARY OF THE ERROR BOUNDS

L-function	Neyman-Pearson Bound $\beta_n^\epsilon \leq$	Bayesian Bound $P_e^{(n)} \leq$
KL divergence	$\exp(-nOPT)$	$(1 - \frac{P_g^*}{4}OPT)^{n/2}$
TV distance	$\exp(-2nOPT^2)$	$(1 - OPT^2)^{n/2}$
JS divergence	$\exp(-\frac{n}{2}OPT^2)$	$(1 - \frac{1}{4}OPT^2)^{n/2}$
Wasserstein	$\exp\left(-\frac{2n}{\text{diam}(\mathcal{X})^2}OPT^2\right)$	$\left(1 - \left(\frac{OPT}{\text{diam}(\mathcal{X})}\right)^2\right)^{n/2}$

3) Jensen-Shannon Divergence: From the bound on Jensen-Shannon divergence (A.55):

$$2TV(\mathbb{P}_X, \mathbb{P}_{g(Z)}) \geq JS(\mathbb{P}_X, \mathbb{P}_{g(Z)}). \tag{16}$$

Thus, we have $\forall g \in \mathcal{G}$

$$2TV(\mathbb{P}_X, \mathbb{P}_{g(Z)}) \geq JS(\mathbb{P}_X, \mathbb{P}_{g(Z)}) \geq OPT. \tag{17}$$

Neyman-Pearson: From Pinsker's inequality and (17)

$$\beta_n^\epsilon \leq \exp(-2nTV(\mathbb{P}_X || \mathbb{P}_{\hat{g}(Z)})^2) \tag{18}$$

$$\leq \exp(-\frac{n}{2}OPT^2) \tag{19}$$

Bayesian: From the bound on Chernoff information in (5)

$$P_e^{(n)} \leq \exp\{\frac{n}{2}\log(1 - TV(\mathbb{P}_X, \mathbb{P}_{\hat{g}(Z)})^2)\} \tag{20}$$

$$\leq \exp\left\{\frac{n}{2}\log\left(1 - \frac{OPT^2}{4}\right)\right\} \tag{21}$$

$$= \left(1 - \frac{OPT^2}{4}\right)^{n/2}. \tag{22}$$

4) Wasserstein Metric: From the bound on Wasserstein metric (A.56)

$$\text{diam}(\mathcal{X})TV(\mathbb{P}_X, \mathbb{P}_{g(Z)}) \geq W(\mathbb{P}_X, \mathbb{P}_{g(Z)}) \geq OPT. \tag{23}$$

where $\text{diam}(\mathcal{X})$ is the diameter of the space.

Neyman-Pearson: From Pinsker's inequality and (23)

$$\beta_n^\epsilon \leq \exp(-2n\ TV(\mathbb{P}_X || \mathbb{P}_{\hat{g}(Z)})^2) \tag{24}$$

$$\leq \exp\left(-\frac{2n}{\text{diam}(\mathcal{X})^2}OPT^2\right) \tag{25}$$

Bayesian: From the bound on Chernoff information in (5)

$$P_e^{(n)} \leq \exp\{\frac{n}{2}\log(1 - TV(\mathbb{P}_X, \mathbb{P}_{\hat{g}(Z)})^2)\} \tag{26}$$

$$\leq \exp\left\{\frac{n}{2}\log\left(1 - \left(\frac{OPT}{\text{diam}(\mathcal{X})}\right)^2\right)\right\} \tag{27}$$

$$= \left(1 - \left(\frac{OPT}{\text{diam}(\mathcal{X})}\right)^2\right)^{n/2}. \tag{28}$$

C. Discussion

The bounds are summarized in Table I. As OPT increases, i.e., the GAN used is less accurate, it is exponentially easier to detect deepfakes in the Neyman-Pearson case, and polynomially so in the Bayesian case. The bound decays exponentially with the resolution n. Thus, if we require very high resolution in order to trust images, an extremely accurate GAN would be required to go undetected.

D. Generalizing the Problem

Images typically have regions depicting features like eyes, mouth, and hair or objects like vehicles and trees that follow a common distribution. Consider the image to consist of k patches with m pixels each, following the distribution $\mathbb{P}_X^{(i)}$, such that m is still large for the hypothesis testing bounds to be valid. We now similarly define individual oracle errors for each patch as $OPT_i := \inf_{g \in \mathcal{G}} L(\mathbb{P}_X^{(i)}, \mathbb{P}_{g(Z)})$, considering that the GAN generates each patch independently. The generalized bounds can be obtained *mutatis mutandis* by multiplying together the corresponding bounds for each of the k patches obtained by using m for n, OPT_i for OPT, $P_{g_i}^*$ for P_g^*, and $\mathrm{diam}(\mathcal{X}_i)$ for $\mathrm{diam}(\mathcal{X})$ to represent the values defined for each patch of the image.

IV. EUCLIDEAN INFORMATION THEORY

With advances in deep learning and cheap availability of data, GANs have grown ever more complex and can generate a very close version of the true distribution. This development suggests further analysis in the low error regime using the concept of Euclidean information theory, which gives the following approximation for KL divergence [21].

Lemma IV.1 (Euclidean approximation). *When two distributions are close to each other, i.e. $P \approx Q$, the KL divergence between them can be approximated as*

$$D(P||Q) \approx \frac{1}{2}||Q - P||_{\hat{P}}^2 = \frac{1}{2}||[\hat{P}^{-1/2}](Q - P)||^2 \tag{29}$$

where the weight $\hat{P}$ is any distribution in the neighborhood of P and Q.

A. Simplified Bounds Using Euclidean Approximation

On applying the approximation in the hypothesis testing error bounds, we get the following:

$$\beta_n^\epsilon \doteq \exp(-nD(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)})) \approx \exp(-\frac{n}{2}||\mathbb{P}_X - \mathbb{P}_{\hat{g}(Z)}||_P^2) \tag{30}$$

where the weight (subscript P) is any distribution in the neighborhood of the two distributions. Also, from [19, ch. 11]

$$P_e^{(n)} \leq \exp(-nD^*(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)})), \text{ where} \tag{31}$$

$$D^*(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)}) = \max_\lambda \min(D(\mathbb{P}_X||\mathbb{P}_\lambda), D(\mathbb{P}_{g(\hat{Z})}||\mathbb{P}_\lambda)) \tag{32}$$

$$\mathbb{P}_\lambda = \frac{\mathbb{P}_X^\lambda \mathbb{P}_{\hat{g}(Z)}^{1-\lambda}}{\sum_{a \in \mathcal{X}} \mathbb{P}_X^\lambda(a) \mathbb{P}_{\hat{g}(Z)}^{1-\lambda}(a)}. \tag{33}$$

Assuming $\mathbb{P}_X \approx \mathbb{P}_{\hat{g}(Z)}, \mathbb{P}_\lambda$ is also close to the two distributions. So using the approximation in Lemma IV.1, the maximizing λ is such that the two quantities are equal.

$$D^*(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)}) \approx \frac{1}{2}||\frac{\mathbb{P}_X + \mathbb{P}_{\hat{g}(Z)}}{2} - \mathbb{P}_X||_{\mathbb{P}_X}^2 \tag{34}$$

$$= \frac{1}{8}||\mathbb{P}_X - \mathbb{P}_{\hat{g}(Z)}||_{\mathbb{P}_X}^2 \tag{35}$$

$$\approx \frac{1}{4}D(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)}). \tag{36}$$

1) KL Divergence: **Neyman-Pearson:** The bound stays the same.
Bayesian:

$$P_e^{(n)} \leq \exp\big(-\frac{n}{4}D(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)})\big) \tag{37}$$

$$\leq \exp\big(-\frac{n}{4}OPT\big). \tag{38}$$

2) Total Variational Distance: **Neyman-Pearson:** The bound stays the same.
Bayesian: Using Pinsker's inequality

$$P_e^{(n)} \leq \exp\big(-\frac{n}{4}D(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)})\big) \tag{39}$$

$$\leq \exp\big(-\frac{n}{4}2TV(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)})^2\big) \tag{40}$$

$$\leq \exp\big(-\frac{n}{2}OPT^2\big) \tag{41}$$

TABLE II
SIMPLIFIED BOUNDS USING EUCLIDEAN APPROXIMATION

L-function	Neyman-Pearson Bound $\beta_n^\epsilon \leq$	Bayesian Bound $P_e^{(n)} \leq$
KL divergence	$\exp(-nOPT)$	$\exp(-\frac{n}{4}OPT)$
TV distance	$\exp(-2nOPT^2)$	$\exp(-\frac{n}{2}OPT^2)$
JS divergence	$\exp(-2nOPT)$	$\exp(-\frac{n}{2}OPT)$
Wasserstein	$\exp\left(-\frac{2n}{\text{diam}(\mathcal{X})^2}OPT^2\right)$	$\exp\left(-\frac{n}{2\,\text{diam}(\mathcal{X})^2}OPT^2\right)$

3) Jensen-Shannon Divergence: When $P_1 \approx P_2$, Jensen-Shannon divergence can also be approximated in the Euclidean regime as:

$$JS(P_1, P_2) = \frac{1}{2}[D(P_1||\tfrac{P_1+P_2}{2}) + D(P_2||\tfrac{P_1+P_2}{2})] \tag{42}$$

$$\approx \frac{\left||P_1 - P_2|\right|_{P_1}^2}{4} \approx \frac{D(P_1||P_2)}{2}. \tag{43}$$

Thus, we have

$$OPT \leq JS(\mathbb{P}_X, \mathbb{P}_{g(Z)}) \approx \frac{D(\mathbb{P}_X||\mathbb{P}_{g(Z)})}{2}. \tag{44}$$

Neyman-Pearson:

$$\beta_n^\epsilon \doteq \exp(-nD(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)})) \leq \exp(-2n\, OPT) \tag{45}$$

Bayesian:

$$P_e^{(n)} \leq \exp\left(-\frac{n}{4}D(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)})\right) \leq \exp\left(-\frac{n}{2}\, OPT\right) \tag{46}$$

4) Wasserstein Metric: Combining the bound for Wasserstein metric in (A.56) and Pinsker's inequality, we have

$$OPT \leq W(\mathbb{P}_X, \mathbb{P}_{g(Z)}) \leq \text{diam}(\mathcal{X})\sqrt{\frac{D(\mathbb{P}_X, \mathbb{P}_{g(Z)})}{2}}. \tag{47}$$

Neyman-Pearson: The bound remains the same.
Bayesian:

$$P_e^{(n)} \leq \exp\left(-\frac{n}{4}D(\mathbb{P}_X||\mathbb{P}_{\hat{g}(Z)})\right) \tag{48}$$

$$\leq \exp\left(-\frac{n}{2\,\text{diam}(\mathcal{X})^2}OPT^2\right) \tag{49}$$

B. Discussion

Table II summarizes the simplified bounds using Euclidean approximation. For the same value of OPT, we see that GANs with Jensen-Shannon divergence as the L-function are the easiest to detect with lowest bound on the Bayesian error probability. Also, while the relationship between the error probability and image resolution remains the same, it is exponentially easier to detect deepfakes from GANs with higher OPT for both tests unlike the results in Table I. The exponent of Neyman-Pearson bound is simply that of the Bayesian bound multiplied by a factor of four for all L-functions.

V. EPIDEMIC THRESHOLD THEORY AND DEEPFAKES

Deepfakes may disperse rapidly in social networks [22] with spreading dynamics similar to diseases. The SIR (susceptible-infected-recovered) model for epidemics can be useful to assess the risk posed by a deepfake. Epidemic threshold characterizes the critical level λ_c for effective spreading rate λ above which a global epidemic occurs and the spreading cannot be contained. The threshold can be predicted based purely on the network structure [23]. The spreading rate is expressed as

$$\lambda = \frac{\beta}{\gamma} \tag{50}$$

where β is the probability of transmission from an infected to susceptible node, and γ is the probability of recovery, i.e., probability of correctly detecting the deepfake $1 - P_e$. Let the bound for $P_e^{(n)}$ be $\exp(-nf(OPT))$. Thus, we have

$$\lambda \leq \frac{\beta}{1 - \exp(-nf(OPT))}. \tag{51}$$

Along with the condition for containment $\lambda \leq \lambda_c$, we get the following condition on OPT for the deepfake to be locally containable.

$$f(OPT) \geq -\frac{1}{n} \ln \left(1 - \frac{\beta}{\lambda_c}\right). \tag{52}$$

When $f(OPT)$ is greater than the expression on the right, which depends only on n and the network structure, the deepfake can be locally confined. From Tables I and II, f is always increasing. A higher OPT, i.e., worse GAN accuracy, guarantees higher robustness to global spread of misinformation. Thus the network will be less easily fooled by a weaker deepfake generation system.

VI. CONCLUSION

This work provides a statistical framework to detect deepfakes and error guarantees for these tests. Avenues to extend this study include— accounting for a possible gender, race, etc. based bias in the training data to prevent incorrect classification, developing a similar framework for conditional GANs [24] for detecting face swapped content, and deriving error bounds in terms of more commonly used GAN evaluation metrics as in [25]. Further, by incorporating these results with studies on spreading dynamics of infectious diseases [26], [27], we can obtain bounds on source detection probability for deepfakes.

REFERENCES

[1] C. Day, "The future of misinformation," *Computing in Science Engineering*, vol. 21, no. 1, pp. 108–108, Jan. 2019.
[2] O. Schwartz, "You thought fake news was bad? Deep fakes are where truth goes to die," *The Guardian*, Nov. 2018. [Online]. Available: https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth
[3] G. Patrini and F. Cavalli, "The state of deepfakes: Reality under attack," Annual Report v.2, Deeptrace B.V., Amsterdam, the Netherlands, 2018.
[4] G. Fleishman, "How to spot the realistic faces creeping into your timelines," *Fast Company*, Apr. 2019.
[5] M. H. Maras and A. Alexandrou, "Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos," *The International Journal of Evidence & Proof*, Oct. 2018.
[6] R. Chesney and D. Citron, "Deepfakes and the new disinformation war: The coming age of post-truth geopolitics author," *Foreign Affairs*, vol. 98, no. 1, pp. 147–155, Jan./Feb. 2019.
[7] A. Breland, "The bizarre and terrifying case of the "deepfake" video that helped bring an African nation to the brink," *Mother Jones*, Mar. 2019. [Online]. Available: https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo
[8] S. Poonam and S. Bansal, "Misinformation is endangering Indias election," *The Atlantic*, Apr. 2019. [Online]. Available: https://www.theatlantic.com/international/archive/2019/04/india-misinformation-election-fake-news/586123
[9] J. F. Boylan, "Will deep-fake technology destroy democracy?" *The New York Times*, Oct. 2018. [Online]. Available: https://www.nytimes.com/2018/10/17/opinion/deep-fake-technology-democracy.html
[10] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? Assessment and detection," IDIAP Research Institute, Martigny, Switzerland, Tech. Rep. Idiap-RR-18-2018, Dec. 2018.
[11] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, Jan. 2019, pp. 83–92.
[12] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI generated fake face videos by detecting eye blinking," arXiv:1806.02877v2 [cs.CV], Jun. 2018.
[13] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," arXiv:1811.00656v2 [cs:CV], Mar. 2019.
[14] U. M. Maurer, "Authentication theory and hypothesis testing," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1350–1356, Jul. 2000.
[15] C. Gao, J. Liu, Y. Yao, and W. Zhu, "Robust estimation and generative adversarial nets," arXiv:1810.02030v3 [stat.ML], Feb. 2019.
[16] B. Zhu, J. Jiao, and D. Tse, "Deconstructing generative adversarial networks," arXiv:1901.09465v2 [cs.LG], Feb. 2019.
[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 2672–2680.
[18] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," arXiv:1701.07875v3 [stat.ML], Dec. 2017.
[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley-Interscience, 2006.
[20] I. Sason, "Tight bounds for symmetric divergence measures and a refined bound for lossless source coding," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 701–707, Feb. 2015.
[21] S. Borade and L. Zheng, "Euclidean information theory," in *2008 IEEE International Zurich Seminar on Communications*, Mar. 2008, pp. 14–17.
[22] M. Ingram, "Facebook slammed by UN for its role in Myanmar genocide," *Columbia Journalism Review*, Nov. 2018. [Online]. Available: https://www.cjr.org/the_media_today/facebook-un-myanmar-genocide.php
[23] W. Wang, Q.-H. Liu, L.-F. Zhong, M. Tang, H. Gao, and H. E. Stanley, "Predicting the epidemic threshold of the susceptible-infected recovered model," *Scientific Reports*, vol. 6, Apr. 2016.
[24] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv:1411.1784v1 [cs.LG], Nov. 2014.
[25] D. J. Im, H. Ma, G. W. Taylor, and K. Branson, "Quantitatively evaluating GANs with divergences proposed for training," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, 2018.
[26] R. Dutta, A. Mira, and J.-P. Onnela, "Bayesian inference of spreading processes on networks," *Proceedings of the Royal Society A*, vol. 474, no. 2215, Jul. 2018.
[27] D. Shah and T. Zaman, "Detecting sources of computer viruses in networks: Theory and experiment," *SIGMETRICS Perform. Eval. Rev.*, vol. 38, no. 1, pp. 203–214, Jun. 2010.
[28] I. Sason and S. Verdu, "f-divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.
[29] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1007–1016, Mar. 2006.
[30] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.
[31] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review / Revue Internationale de Statistique*, vol. 70, no. 2, pp. 419–435, Dec. 2002.

APPENDIX

IMPORTANT INEQUALITIES

A. Pinsker's inequality:

$$TV(P_1, P_2)^2 \leq \frac{1}{2} D(P_1 || P_2). \tag{A.53}$$

Proof. The proof is given in [28]. □

B. Reverse Pinsker's inequality:

For the case of finite alphabet $\mathcal{X}$, when $P_2^* = \min_{x \in \mathcal{X}} P_2(x) > 0$,

$$\frac{4}{P_2^*} TV(P_1, P_2)^2 \geq D(P_1 || P_2). \tag{A.54}$$

Proof. This reverse bound on total variational distance has been proved by Csiszár and Talata [29, p. 1012]. □

C. Bound on Jensen-Shannon divergence:

$$JS(P_1, P_2) \leq 2TV(P_1, P_2). \tag{A.55}$$

Proof. The proof is given in [30]. □

D. Bound on Wasserstein metric:

$$W(P_1, P_2) \leq \text{diam}(\mathcal{X}) TV(P_1, P_2) \tag{A.56}$$

where $\text{diam}(\mathcal{X}) = \sup\{d(x, y) : (x, y) \in \mathcal{X}\}$ is the diameter of the space.

Proof. The proof is given in [31]. □