GANs with Conditional Independence Graphs: On Subadditivity of Probability Divergences

Mucong Ding
University of Maryland, College Park
mcding@umd.edu

Constantinos Daskalakis
Massachusetts Institute of Technology

Soheil Feizi
University of Maryland, College Park

Abstract

Generative Adversarial Networks (GANs) are modern methods to learn the underlying distribution of a data set. GANs have been widely used in sample synthesis, de-noising, domain transfer, etc. GANs, however, are designed in a *model-free* fashion where *no* additional information about the underlying distribution is available. In many applications, however, practitioners have access to the underlying independence graph of the variables, either as a Bayesian network or a Markov Random Field (MRF). We ask: how can one use this additional information in designing *model-based* GANs? In this paper, we provide theoretical foundations to answer this question by studying subadditivity properties of probability divergences, which establish upper bounds on the distance between two high-dimensional distributions by the sum of distances between their marginals over (local) neighborhoods of the graphical structure of the Bayes-net or the MRF. We prove that several popular probability divergences satisfy some notion of subadditivity under mild conditions. These results lead to a principled design of a model-based GAN that uses a set of simple discriminators on the neighborhoods of the Bayes-net/MRF, rather than a giant discriminator on the entire network, providing significant statistical and computational benefits. Our experiments on synthetic and real-world datasets demonstrate the benefits of our principled design of model-based GANs.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 Introduction

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have been successfully used to model complex distributions such as image data. GANs model the learning problem as a *min-max* game between generator and discriminator functions. Depending on the specific cost function and constraints on the discriminator network, the associated optimization problem aims at estimating a Wasserstein distance (Arjovsky et al., 2017), an Integral Probability Measure (IPM) (Müller, 1997), an f-divergence (Nowozin et al., 2016), etc., between the target and generated distributions.

GANs are often designed in a *model-free* fashion where *no* additional information about the underlying distribution is available[1]. In some applications, however, one may have some side information about the data distribution. For example, one may know that there is a Markov chain governing the underlying independence graph of the variables. In general, the underlying independence graph of variables may be available as a Bayesian network (i.e. a directed graph) or a Markov Random Field (i.e. an undirected graph). In this paper, we ask: how can we use this additional information in a principled *model-based* design of GANs?

In this paper, we provide theoretical foundations to answer the aforementioned question for high-dimensional distributions with conditional independence structure captured by either a Bayesian network or a Markov Random Field (MRF). We mainly focus on the application to GANs, while the theory developed can be used by any other type of adversarial learning that exploits discriminator networks. The pertinent question

[1]Some works have studied GANs under some strict assumptions on the input data distribution. For example, Feizi et al. (2017) has designed GANs for multivariate Gaussians while Balaji et al. (2019) and Farnia et al. (2020) have studied GANs for mixtures of Gaussians. In contrast, our method is applicable to any Bayesian network or Markov Random Field, which are significantly richer families of distributions.

is whether a known Bayes-net or MRF structure can be exploited to design a GAN with multiple discriminators that are localized and simple. In particular, we are interested in whether we can replace the large discriminator of the vanilla GAN implementation with several simple discriminators that are used to enforce constraints on local neighborhoods of the Bayes-net or the MRF (i.e. local discriminators). Ignoring the underlying conditional independence structure we might know about the target distribution and letting the GAN "learn it on its own" requires a very large discriminator network, especially in applications where data is gathered across many time steps. Large discriminators face computational and statistical challenges, given that min-max training is computationally challenging, and statistical hypothesis testing in large dimensions requires sample complexity exponential in the dimension; see e.g. discussions by Daskalakis and Pan (2017); Daskalakis et al. (2019); Canonne et al. (2020).

Our proposed framework is based on *subadditivity* properties of probability divergences over a Bayes-net or a MRF, which establish upper bounds on the distance between two high-dimensional distributions with the same Bayes-net or MRF structure by the sum of distances between their marginals over (local) neighborhoods of the graphical structure of the Bayes-net or the MRF (Daskalakis and Pan, 2017). For a Bayes-Net, each local neighborhood is defined as the union of a node i and its parents Π_i, as it is the smallest set that encodes conditional dependence. For a MRF, the set of local neighborhoods can be defined as the set of maximal cliques $\mathcal{C}$ of the underlying graph.

Let δ be some divergence or probability metric, such as some Wasserstein distance or f-divergence, that is estimated by each of the local discriminators in their dedicated neighborhood. If we train a generator with the set of local discriminators, it samples a distribution Q that minimizes the sum of divergences δ between marginals of P and Q over the local neighborhoods, where P is the target distribution. As per our description of what the local neighborhoods are in each case, the optimization objective becomes $\sum_{i=1}^{n} \delta(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}})$ on a Bayes-net, and $\sum_{C \in \mathcal{C}} \delta(P_{X_C}, Q_{X_C})$ on a MRF. However, our real goal is to minimize some divergence $\delta'(P, Q)$ of interest measured on the joint (high-dimensional) distributions. We say that $\delta(.,.)$ satisfies *generalized subadditivity* if the sum $\sum_{i=1}^{n} \delta(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}})$ or $\sum_{C \in \mathcal{C}} \delta(P_{X_C}, Q_{X_C})$ upper-bounds the divergence $\delta'(P, Q)$ of interest up to some constant factor $\alpha > 0$ and additive error $\epsilon \geq 0$, i.e. $\delta'(P, Q) - \epsilon \leq \alpha \cdot \sum_{i=1}^{n} \delta(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}})$ (on Bayes-nets), or $\delta'(P, Q) - \epsilon \leq \alpha \cdot \sum_{C \in \mathcal{C}} \delta(P_{X_C}, Q_{X_C})$ (on MRFs), where δ' can be the same or different from δ. In this sense, the generator effectively minimizes $\delta'(P, Q)$ by minimizing its upper-bound. Since, in many applications, local neighborhoods can be significantly smaller than the entire graph, local discriminators targeting each of these neighborhoods will enjoy improved computational and statistical properties in comparison to a global discriminator targeting the entire graph.

The key question is which divergences or metrics exhibit subadditivity to be used in our proposed framework. For testing the identity of Bayes-nets, Daskalakis and Pan (2017) shows that squared Hellinger distance, Kullback-Leibler divergence, and Total Variation distance satisfy some notion of generalized subadditivity. Since our goal in this paper is to exploit subadditivity in the design of GANs, we are interested in establishing generalized subadditivity bounds for distances and divergences that are commonly used in GAN formulations. In this work, we prove that

- Jensen-Shannon divergence used in the original GAN model (Goodfellow et al., 2014),
- Wasserstein distance used in Wasserstein GANs (Arjovsky et al., 2017), and Integral Probability Metric (IPM) (Müller, 1997) used in Wasserstein, MMD and Energy-based GANs (Li et al., 2015; Zhao et al., 2017),
- and nearly all f-divergences used in f-GANs (Nowozin et al., 2016),

satisfy some notion of generalized subadditivity over Bayes-nets under some mild conditions.[2] Moreover, we prove that under some mild conditions

- Wasserstein distance and IPM satisfy generalized subadditivity on MRFs.

These results establish theoretical foundations for using underlying conditional independence graphs in GAN's designs. We demonstrate benefits of our design over several synthetic and real datasets such as the synthetic "ball throwing trajectory" dataset and two real Bayes-net datasets: *the EARTHQUAKE dataset* (Korb and Nicholson, 2010) and *the CHILD dataset* (Spiegelhalter, 1992).

2 Related Works

In many applications, adversarial learning has been used in a broader sense where *multiple local discriminators* have been employed in the learning framework. For example, in image-to-image translation methods (Isola et al., 2017; Zhu et al., 2017; Yi et al., 2017; Choi et al., 2018; Yu et al., 2019; Demir and Unal, 2018), local

[2] We discuss the notion of "local subadditivity" in Section 6 and Appendix F.

discriminators are applied to different patches of images (Li and Wand, 2016). In the analysis of time-series data as well as natural language processing (NLP) tasks, local discriminators based on sliding windows (Li et al., 2019), self-attention (Clark et al., 2019), recurrent neural networks (RNNs) (Esteban et al., 2017; Mogren, 2016), convolution neural networks (CNNs) (Nie et al., 2018), and dilated causal convolutions (Oord et al., 2016; Donahue et al., 2019) have been applied on different subsequences of the data. These models have been applied to a wide range of tasks including image style transfer (Isola et al., 2017; Zhu et al., 2017; Yi et al., 2017; Choi et al., 2018), inpainting (Yu et al., 2019; Demir and Unal, 2018), and texture synthesis (Li and Wand, 2016), as well as time-series generation (Esteban et al., 2017; Mogren, 2016), imputation (Liu et al., 2019), anomaly detection (Li et al., 2019), and even video generation (Clark et al., 2019) and inpainting (Chang et al., 2019).

Intuitively, these methods aim at structuring the generation process and/or narrowing down the purview of the discriminator to capture known dependencies leading to improved computational and statistical properties. These methods, however, are mostly not accompanied by theoretical foundations. In particular, it is not clear what subset of features each local discriminator should be applied to, how many local discriminators should be used in the learning process, and what the effect of the discriminator localization is on estimating the distance between the generated and target distributions.

3 Notation

Consider a Directed Acyclic Graph (DAG) G with nodes $\{1,\dots,n\}$. Let Π_i be the set of parents of node i in G. Assume that $(1,\dots,n)$ is a topological ordering of G, i.e. $\Pi_i \subseteq \{1,\dots,i-1\}$ for all i. A probability distribution $P(x)$ defined over space $\Omega = \{(x_1,\dots,x_n)\}$ is a *Bayes-net with respect to graph G* if it can be factorized as $P(x) = \prod_{i=1}^n P_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i})$.

Given an undirected graph G with nodes $\{1,\dots,n\}$, a probability distribution $P(x)$ defined over space $\Omega = \{(x_1,\dots,x_n)\}$ is a *MRF with respect to graph G* if any two disjoint subsets of variables $A, B \subseteq \{1,\dots,n\}$ are conditionally independent conditioning on a separating subset S of variables (i.e. S such that all paths in G from nodes in A to nodes in B pass through S). This conditional independence property is denoted $X_A \perp\!\!\!\perp X_B \mid X_S$. Such $P(x)$ can be factorized as $P(x) = \prod_{C\in\mathcal{C}} \psi_C(X_C)$, where $\mathcal{C}$ is the set of maximal cliques in G. In this paper, unless otherwise noted, we always assume $X_i \in \mathbb{R}^d$, thus $\Omega \subseteq \mathbb{R}^{nd}$, and use the Euclidean metric. We always assume the density exists.

4 Generalized Subadditivity on Bayes-nets

In this section, we define the notion of *generalized subadditivity* of a statistical divergence δ on Bayes-nets. We discuss subadditivity on MRFs in Section 5.

Definition 1 (Generalized Subadditivity of Divergences on Bayes-nets)**.** *Consider two Bayes-nets P, Q over the same sample space $\Omega = \{(x_1,\dots,x_n)\}$ and defined with respect to the same DAG, G, i.e. factorizable as $P(x) = \prod_{i=1}^n P_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i})$, $Q(x) = \prod_{i=1}^n Q_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i})$, where Π_i is the set of parents of node i in G. For a pair of statistical divergences δ and δ', and constants $\alpha > 0$ and $\epsilon \geq 0$, if the following holds for all Bayes-nets P, Q as above:*

$$\delta'(P,Q) - \epsilon \leq \alpha \cdot \sum_{i=1}^n \delta(P_{X_i\cup X_{\Pi_i}}, Q_{X_i\cup X_{\Pi_i}}),$$

then we say that δ satisfies α-linear subadditivity with error ϵ with respect to δ' on Bayes-nets. *For the common case $\epsilon = 0$ and $\delta' = \delta$, we say that δ satisfies* α-linear subadditivity on Bayes-nets. *When additionally $\alpha = 1$, we say that δ satisfies subadditivity on Bayes-nets.*

We refer to the right-hand side of the subadditivity inequality as the subadditivity upper bound. If a statistical divergence δ satisfies linear subadditivity with respect to δ', minimizing the subadditivity upper bound serves as a proxy to minimizing $\delta'(P,Q)$. The subadditivity upper bound is often used as the objective function in adversarial learning when local discriminators are employed.

We argue that subadditivity of δ on (1) product measures, and (2) length-3 Markov Chains suffices to imply subadditivity on all Bayes-nets. The claim is implicit in the proof of Theorem 2.1 by Daskalakis and Pan (2017); we state it explicitly here and provide its proof in Appendix A.1 for completeness. Roughly speaking, the proof follows because we can always combine nodes of a Bayes-net into super-nodes to obtain a 3-node Markov Chain or a 2-node product measure, and apply the Markov Chain/Product Measure subadditivity property recursively.

Theorem 1. *If a divergence δ satisfies the following:*

(1) For any two Bayes-nets P and Q on DAG $X \to Y \to Z$, the following subadditivity holds: $\delta(P_{XYZ}, Q_{XYZ}) \leq \delta(P_{XY}, Q_{XY}) + \delta(P_{YZ}, Q_{YZ})$.

(2) For any two product measures P and Q over variables X and Y, the following subadditivity holds: $\delta(P_{XY}, Q_{XY}) \leq \delta(P_X, Q_X) + \delta(P_Y, Q_Y)$.

then δ satisfies subadditivity on Bayes-nets.

Using Theorem 1, it is not hard to prove that squared Hellinger distance has subadditivity on Bayes-nets, as shown by Daskalakis and Pan (2017). For completeness, we provide proof of the following in Appendix A.2

Theorem 2 (Theorem 2.1 by Daskalakis and Pan (2017)). *The squared Hellinger distance defined as* $\mathrm{H}^2(P,Q) \coloneqq 1 - \int \sqrt{PQ}\ \mathrm{d}x$ *satisfies subadditivity on Bayes-nets.*

4.1 Subadditivity of f-Divergences

For two probability distributions P and Q on Ω, the f-divergence of P from Q, denoted $D_f(P,Q)$, is defined as $D_f(P,Q) = \int_\Omega f\left(P(x)/Q(x)\right) Q(x)\mathrm{d}x$. We assume P is absolutely continuous with respect to Q, written as $P \ll Q$. Common f-divergences are Kullback-Leibler divergence (KL), Symmetric KL divergence (SKL), Jensen-Shannon divergence (JS), and Total Variation distance (TV); see Appendix B. The subadditivity of KL-divergence on Bayes-nets is claimed by Daskalakis and Pan (2017) without a proof. We provide a proof in Appendix A.3 for completeness.

Theorem 3 (Claimed by Daskalakis and Pan (2017)). *The KL-divergence defined as* $\mathrm{KL}(P,Q) \coloneqq \int P \log\left(P/Q\right) \mathrm{d}x$ *satisfies subadditivity on Bayes-nets.*

It follows from the proof of Theorem 3 that the following conditions suffice for the KL subadditivity to become additivity: $\forall i,\ P_{X_{\Pi_i}} = Q_{X_{\Pi_i}}$ (almost everywhere). From the investigation of local subadditivity of f-divergences (Theorem 21 in Appendix F), we will see that this is the minimum set of requirements possible. The subadditivity of KL divergence easily implies the subadditivity of the Symmetric KL divergence.

Corollary 4. *The Symmetric KL divergence defined as* $\mathrm{SKL}(P,Q) \coloneqq \mathrm{KL}(P,Q) + \mathrm{KL}(Q,P)$ *satisfies subadditivity on Bayes-nets.*

Moreover, the linear subadditivity of Jensen-Shannon divergence (JS) follows from the subadditivity property of squared Hellinger distance; see Appendix A.4.

Corollary 5. *The Jensen-Shannon divergence defined as* $\mathrm{JS}(P,Q) \coloneqq \frac{1}{2}\mathrm{KL}\left(P,(P+Q)/2\right) + \frac{1}{2}\mathrm{KL}\left(Q,(P+Q)/2\right)$ *satisfies* $(1/\ln 2)$*-linear subadditivity on Bayes-nets.*

Using a slightly modified version of Theorem 1, it is not hard to derive the linear subadditivity of Total Variation distance, which is stated without proof by Daskalakis and Pan (2017). We provide a proof in Appendix A.5 for completeness.

Theorem 6 (Claimed by Daskalakis and Pan (2017)). *The Total Variation distance defined as* $\mathrm{TV}(P,Q) \coloneqq \frac{1}{2}\int |P-Q|\,\mathrm{d}x$ *satisfies* 2*-linear subadditivity on Bayes-nets.*

4.2 Subadditivity of Wasserstein Distance and IPMs

Suppose Ω is a metric space with distance $d(\cdot,\cdot)$. The p-Wasserstein distance W_p is defined as $\mathrm{W}_p(P,Q) \coloneqq \left(\inf_{\gamma\in\Gamma(P,Q)} \int_{\Omega\times\Omega} d(x,y)^p \mathrm{d}\gamma(x,y)\right)^{1/p}$, where $\gamma \in \Gamma(P,Q)$ denotes the set of all possible couplings of P and Q; see Appendix C.

In general, Wasserstein distance does not satisfy subadditivity on Bayes-nets and MRFs shown by a counterexample using Gaussian distributions (Appendix E). However, based on the linear subadditivity of TV on Bayes-nets, one can prove that all p-Wasserstein distances with $p \geq 1$ satisfy α-linear subadditivity when space Ω is discrete and finite (Appendix A.6).

Corollary 7. *If* Ω *is a finite metric space,* p*-Wasserstein distance for* $p \geq 1$ *satisfies* $(2^{1/p}\mathrm{diam}(\Omega)/d_{\min})$*-linear subadditivity on Bayes-nets, where* $\mathrm{diam}(\Omega)$ *is the diameter and* $d_{\min}$ *is the smallest distance between pairs of distinct points in* Ω*.*

Integral Probability Metrics (IPMs) are a class of probability distances defined as $d_{\mathcal{F}}(P,Q) \coloneqq \sup_{\phi\in\mathcal{F}} \{\mathbb{E}_{x\sim P}[\phi(x)] - \mathbb{E}_{x\sim Q}[\phi(x)]\}$, which include the Wasserstein distance, Maximum Mean Discrepancy, and Total Variation distance. The IPM with $\mathcal{F}$ being all 1-Lipschitz functions is the 1-Wasserstein distance (Villani, 2008). Practical GANs take $\mathcal{F}$ as a parametric function class, $\mathcal{F} = \{\phi_\theta(x) | \theta \in \Theta\}$, where $\phi_\theta(x)$ is a neural network. The resulting IPMs are called neural distances (Arora et al., 2017).

Next, we prove that neural distances (even those expressible by a single ReLU neuron) satisfy generalized subadditivity with respect to the Symmetric KL divergence. This property establishes substantive theoretical justification for the local discriminators used in GANs based on IPMs.

Theorem 8. *Consider two Bayes-nets* P, Q *on* $\Omega = \{(X_1,\ldots,X_n)\} \subseteq \mathbb{R}^{nd}$ *with a common DAG* G*, and any set of function classes* $\{\mathcal{F}_1,\ldots,\mathcal{F}_n\}$*. Suppose the following conditions are fulfilled:*

(1) the space Ω *is bounded, i.e.* $\mathrm{diam}(\Omega) < \infty$*;*
(2) each discriminator class $(\mathcal{F}_i)$ *is larger than the set of single neuron networks with ReLU activations, i.e.* $\{\max\{w^T x + b, 0\} \big| \|[w,b]\|_2 = 1\}$*; and*
(3) $\log(P_{X_i\cup X_{\Pi_i}}/Q_{X_i\cup X_{\Pi_i}})$ *are bounded and Lipschitz continuous for all* i*.*

Then the neural distances defined by $\mathcal{F}_1,\ldots,\mathcal{F}_n$ *satisfy the following* α*-linear subadditivity with error* ϵ *with respect to the Symmetric KL divergence on Bayes-nets:*

$$\mathrm{SKL}(P,Q) - \epsilon \leq \alpha \cdot \sum_{i=1}^{n} d_{\mathcal{F}_i}(P_{X_i\cup X_{\Pi_i}}, Q_{X_i\cup X_{\Pi_i}}),$$

where α and ϵ are constants independent of P, Q and $\{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$, satisfying

$$\alpha > R\Big((k_{\max}+1)d\Big) \text{ and } \epsilon = \mathcal{O}\left(n\alpha^{-\frac{2}{(k_{\max}+1)d+1}} \log \alpha\right),$$

where $R((k_{\max}+1)d)$ is a function that only depends on $k_{\max}$ (the maximum in-degree of G) and d (the dimensionality of each variable of the Bayes-net).

Regarding condition (1), bounded space Ω still allows many real-world data-types, including images and videos. Regarding condition (2), all practical neural networks using ReLU activations satisfy this requirement. Thus, the only non-trivial requirement is condition (3). In practical GAN training, Q is the output distribution of a generative model, which can be regarded as a transformation of a Gaussian distribution. Thus, in general, Q is bounded and Lipschitz. If we have $P \ll Q$, for bounded and Lipschitz real distribution P, the condition (3) is satisfied. If the subadditivity upper bound is minimized, we can minimize $\mathrm{SKL}(P, Q)$ up to $\mathcal{O}(n)$. For the detailed proof, see Appendix A.7.

5 Generalized Subadditivity on MRFs

The definition of *generalized subadditivity* of a statistical divergence with respect to another one over MRFs is the same as in Definition 1, except that the local neighborhoods are defined as maximal cliques $C \in \mathcal{C}$ of the MRF. For an alternative definition of subadditivity on MRFs, see Appendix D.

The clique factorization of MRFs (i.e. $P(x) = \prod_{C\in\mathcal{C}} \psi_C^P(X_C)$) offers a special method to prove the subadditivity of IPMs on MRFs. Consider the Symmetric KL divergence $\mathrm{SKL}(P, Q) := \mathrm{KL}(P, Q) + \mathrm{KL}(Q, P) = \mathbb{E}_{x\sim P}[\log(P/Q)] - \mathbb{E}_{x\sim Q}[\log(P/Q)]$. Clique factorization of P and Q decomposes $\mathrm{SKL}(P, Q)$ into $\mathrm{SKL}(P, Q) = \sum_{C\in\mathcal{C}}(\mathbb{E}_{x_C\sim P_{X_C}}[\log(\psi_C^P/\psi_C^Q)] - \mathbb{E}_{x_C\sim Q_{X_C}}[\log(\psi_C^P/\psi_C^Q)])$, where each term in the summation is upper-bounded by an IPM $d_{\mathcal{F}_C}(P_{X_C}, Q_{X_C})$ on the clique C, as long as $\log(\psi_C^P/\psi_C^Q) \in \mathcal{F}_C$. This implies the subadditivity of 1-Wasserstein distance with respect to the Symmetric KL divergence, whenever each $\log(\psi_C^P/\psi_C^Q)$ is Lipschitz continuous; see Appendix A.8 for the proof.

Theorem 9. *Consider two MRFs P, Q with the same factorization. If any of the following is fulfilled:*

(1) The space Ω is discrete and finite.
(2) $\log(\psi_C^P/\psi_C^Q)$ are Lipschitz continuous for all $C \in \mathcal{C}$.

Then, the 1-Wasserstein distance satisfies α-linear subadditivity with respect to the Symmetric KL Divergence on MRFs, for some constant $\alpha > 0$ independent of P and Q.

Using the aforementioned property of Symmetric KL divergence, the subadditivity of neural distances (Theorem 8) can be generalized to MRFs; see Appendix A.9.

Corollary 10. *For two MRFs P, Q on a common graph G and a set of function classes $\{\mathcal{F}_C | C \in \mathcal{C}\}$, if all of the three conditions in Theorem 8 are fulfilled (with condition (3) replaced by: $\log(\psi_C^P/\psi_C^Q)$ are bounded and Lipschitz continuous for all $C \in \mathcal{C}$), the neural distances induced by $\{\mathcal{F}_C | C \in \mathcal{C}\}$ satisfy α-linear subadditivity with error ϵ with respect to the Symmetric KL divergence on MRFs, i.e. $\mathrm{SKL}(P, Q) - \epsilon \le \alpha \cdot \sum_{C\in\mathcal{C}} d_{\mathcal{F}_C}(P_{X_C}, Q_{X_C})$, where α and ϵ are constants independent of P, Q and $\{\mathcal{F}_C | C \in \mathcal{C}\}$, satisfying $\alpha > R(c_{\max}d)$ and $\epsilon = \mathcal{O}\left(|\mathcal{C}|\alpha^{-\frac{2}{c_{\max}d+1}} \log \alpha\right)$. $|\mathcal{C}|$ is the number of maximal cliques in G and $R(c_{\max}d)$ is a function that only depends on $c_{\max} = \max\{|C| \big| C \in \mathcal{C}\}$ (the maximum size of the cliques in G) and d.*

6 Local Subadditivity

So far, we have stated and proved the subadditivity or generalized subadditivity of some f-divergences on Bayes-nets or MRFs. However, many divergences may not enjoy subadditivity property (see such a counterexample of 2-Wasserstein distance in Appendix E). It is difficult to formulate a general framework for determining which divergence is subadditive.

In this section, we consider a particular scenario when two distributions P, Q are *close* to each other, which can happen after some initial training steps in a GAN. In this case, we are able to determine if an arbitrary f-divergence satisfies generalized subadditivity on Bayes-nets. We only report our main results here. See Appendix F and Appendix G for more details and proofs. We consider two notions of "closeness" for distributions.

Definition 2. *Distributions P, Q are one-sided ϵ-close for some $0 < \epsilon < 1$, if $\forall x \in \Omega \subseteq \mathbb{R}^{nd}$, $P(x)/Q(x) < 1 + \epsilon$. Moreover, P, Q are two-sided ϵ-close, if $\forall x$, $1-\epsilon < P(x)/Q(x) < 1+\epsilon$. Note this requires $P \lll\gg Q$.*

We find that most f-divergences satisfy generalized linear subadditivity when the distributions are one- or two-sided ϵ-close.

Theorem 11. *An f-divergence whose $f(\cdot)$ is continuous on $(0, \infty)$ and twice differentiable at 1 with $f''(1) > 0$ satisfies α-linear subadditivity, when P, Q are two-sided $\epsilon(\alpha)$-close with $\epsilon > 0$, where $\epsilon(\alpha)$ is a non-increasing function and $\lim_{\epsilon\downarrow 0} \alpha = 1$.*

Theorem 12. *An f-divergence whose $f(\cdot)$ is continuous and strictly convex on $(0, \infty)$, twice differentiable at $t = 1$, and has finite $f(0) = \lim_{t\downarrow 0} f(t)$, satisfies α-linear subadditivity, when P, Q are one-sided $\epsilon(\alpha)$-close with $\epsilon > 0$, where $\epsilon(\alpha)$ is a non-increasing function and $\lim_{\epsilon\downarrow 0} \alpha > 0$.*

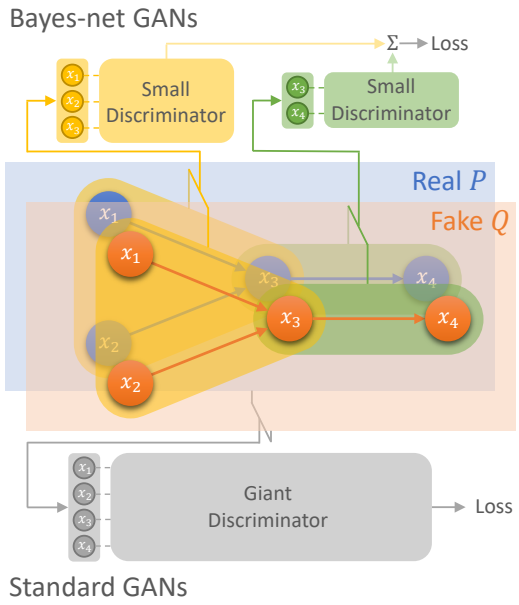


Figure 1: Conceptual diagram of the Bayes-net GANs with local discriminators compared with the standard GANs.

7 GANs with Bayes-Nets/MRFs

Our proposed model-based GAN minimizes the generalized subadditivity upper bound of a divergence measure δ. For example, a Bayes-net GAN[3] is formulated as the following optimization problem:

$$\min_{Q} \quad \sum_{i=1}^{n} \delta(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}}).$$

Similar to a standard GAN (Goodfellow et al., 2014; Arjovsky et al., 2017), the generated distribution Q is characterized as $G(Z)$ where $G(.)$ is the generator function and Z is a normal distribution. Note that the discriminator is implicit in the definition of the δ (Figure 1). Since local neighborhoods are often significantly smaller than the entire graph, our proposed model-based GAN enjoys improved computational and statistical properties compared to a model-free GAN that uses a global discriminator targeting the entire graph.

8 Experiments

In this section, we provide experimental results demonstrating the benefits of exploiting the underlying Bayes-net or MRF structure of the data in the design of model-based GANs. In our experiments, we consider a synthetic *ball throwing trajectory dataset* as well as two real Bayes-net datasets: *the EARTHQUAKE dataset* (Korb and Nicholson, 2010) and *the CHILD dataset* (Spiegelhalter, 1992). Unless otherwise stated, the Wasserstein GAN (Arjovsky et al., 2017) with gradient penalty (Gulrajani et al., 2017) is used in the experiments. Detailed experimental setups (including network architectures and hyper-parameters) can be found in Appendix K. The experiments on MRF datasets and more experimental findings on Bayes-nets including the sensitivity analysis of Bayes-net GANs are reported in Appendix J.

8.1 Synthetic Ball throwing trajectories

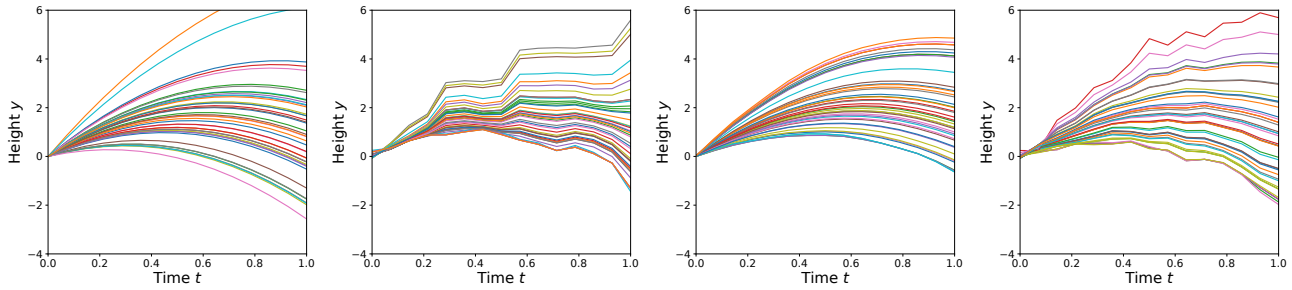
In this section, we consider a simple synthetic dataset that consists of single-variate time-series data $(y_1, \ldots, y_{15})$ representing the y-coordinates of ball throwing trajectories lasting 1 second, where $y_t = v_0 * (t/15) - g(t/15)^2/2$. v_0 is a Gaussian random variable and $g = 9.8$ is the gravitational acceleration. These trajectories are Bayes-nets, where the underlying DAG has the following structure: each node $t \in \{1, \ldots, 15\}$ has two parents, $(t-1)$ and $(t-2)$ (if they exist). This is because, given g and without known v_0, one can determine y_t from y_{t-1} and y_{t-2}.

We train two types of GANs to generate "ball throwing trajectories": (1) Bayes-net GANs with local discriminators where each discriminator has a certain *time localization width* and (2) a standard GAN with one global discriminator. From the underlying physics of this dataset, we know that a proper discriminator design should have at least a localization width of 3 since one needs at least three consecutive coordinates y_{t-2}, y_{t-1}, y_t to estimate the gravitational acceleration g. Thus, from the theory, a GAN trained using local discriminators with a localization width of 2 should not be able to generate high-quality samples. This is in fact verified by our experiments. In Fig. 2, we see samples generated by the local-width 3 GAN (Fig. 2(c)) are visually very similar to the ground truth trajectories (Fig. 2(a)), while samples generated by the local-width 2 GAN demonstrate poor quality.

Note that increasing the localization width of the discriminators enhances their discrimination power, but at the same time, it increases the model complexity, which can cause statistical and computational issues during the training. To understand this trade-off, we progressively increase the localization width from 3 to 15, obtaining one giant discriminator at the end. The quality of generated trajectories from the standard GAN (corresponding to the giant discriminator) is, in fact, worse (Fig. 2(d)).

In Fig. 3, we compare the estimation errors of the gravi-

[3] A model-based GAN on MRFs can be formulated similarly.



(a) Ground truth (b) Local-width 2 (c) Local-width 3 (d) Standard GAN

Figure 2: GAN-generated ball throwing trajectories by (b) the Bayes-net GAN (ours) with *localization width* 2 (the width of the local neighborhoods that the discriminators test on), (c) the Bayes-net GAN with local-width 3, and (d) the standard GAN.

Dataset	GAN used	**Energy Stats.** ($\times 10^{-2}$) (smaller is better)	**Detection AUC** (smaller is better)	**Rel. BIC** ($\times 10^{2}$) (larger is better)	**Rel. GED** (smaller is better)
EARTHQUAKE	Bayes-net (ours)	0.24 ± 0.04	0.523 ± 0.005	$+1.68 \pm 0.17$	0.4 ± 0.7
	Standard	1.72 ± 0.08	0.564 ± 0.012	-4.30 ± 0.21	5.6 ± 0.7
CHILD	Bayes-net (ours)	2.37 ± 0.10	0.644 ± 0.008	$+0.6 \pm 1.5$	9 ± 4
	Standard	4.40 ± 0.22	0.689 ± 0.019	-7.1 ± 2.0	24 ± 8

Table 1: Quality metrics of samples generated by the standard and Bayes-net GANs trained on the Bayes-nets.

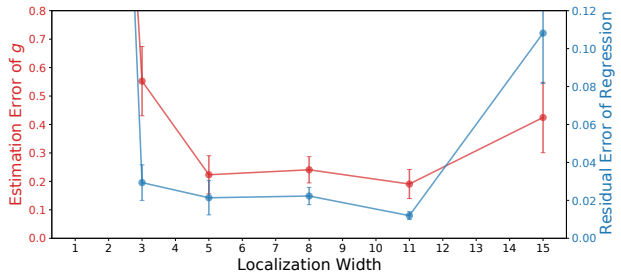


Figure 3: Estimation errors of gravitational acceleration g and residual errors of degree-2 polynomial regression on the generated trajectories with varying localization width.

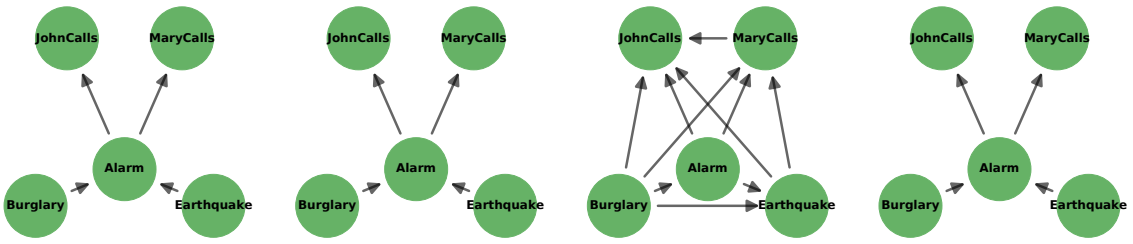
tational acceleration g and the residual errors of degree-2 polynomial regression (which evaluate the "smoothness" of generated trajectories) among GANs with different localization widths. Interestingly, the curves of both metrics demonstrate a U-shaped behavior indicating that there is an optimal localization width balancing between the discrimination power and the model complexity and its resulting statistical/computational burden.

8.2 Real Bayes-nets

Next, we consider two real Bayes-net datasets: (1) *the EARTHQUAKE dataset* which is a small Bayes-net with 5 nodes and 4 edges characterizing the alarm system against burglary which can get occasionally set off by an earthquake (Korb and Nicholson, 2010), and (2) *the CHILD dataset* which is a Bayes-net for diagnosing congenital heart disease in a newborn "blue baby" (Spiegelhalter, 1992), with 20 nodes and 25 edges. The underlying Bayes-nets of both datasets are known. We first generate samples from the Bayes-nets, then train both standard GANs and Bayes-net GANs (using the subadditivity upper-bound as objectives) on them (Since all the features are categorical, we use *Gumbel-Softmax* (Jang et al., 2016) as a differentiable approximation to the *Softmax* function in the generator; see Appendix K.)

If a GAN learns the Bayes-net well, it should learn both the joint distribution and the conditional dependencies. We evaluate the quality of the generated samples by four scores:

- **Energy Statistics** measuring how close the real and fake empirical distributions based on a statistical potential energy (a function of distances between observables) (Székely and Rizzo, 2013),
- **Detection AUC**: AUC scores of binary classifiers trained to distinguish fake samples from real ones,
- **Relative BIC**: the Bayesian information criterion of fake samples (a log-likelihood score with an additional penalty for the network complexity) (Koller and Friedman, 2009) subtracted by the BIC of real ones, and



(a) Ground truth (b) Direct Prediction (c) Standard GAN (d) Bayes-net GAN (ours)

Figure 4: Causal structures predicted from (b) the observed data, (c) the data generated by the standard GAN, and (d) the data generated by the Bayes-net GAN (ours).

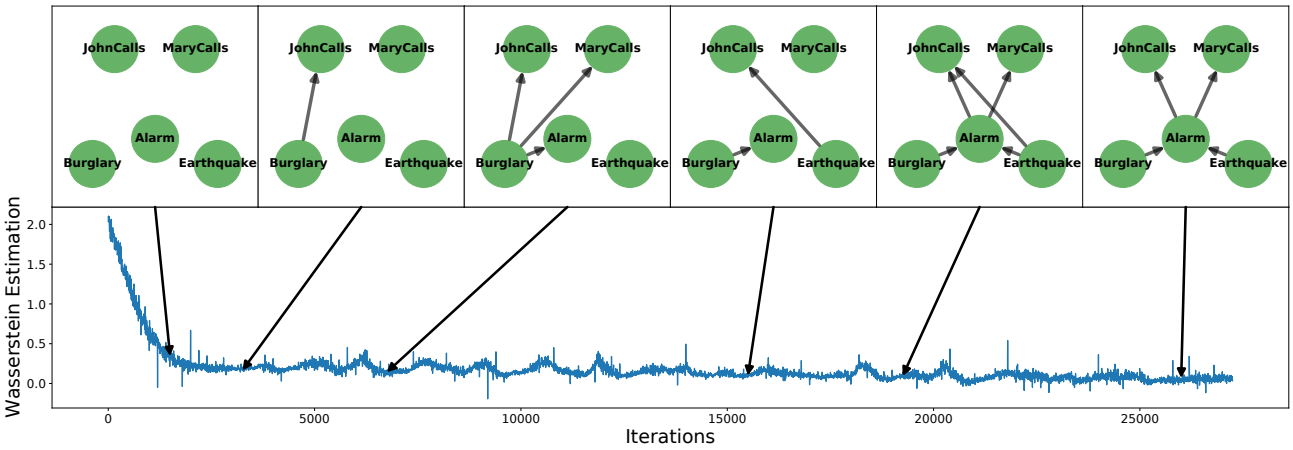


Figure 5: Causal structures predicted from the data generated by the Bayes-net GAN at different stages of training and the Wasserstein loss curve.

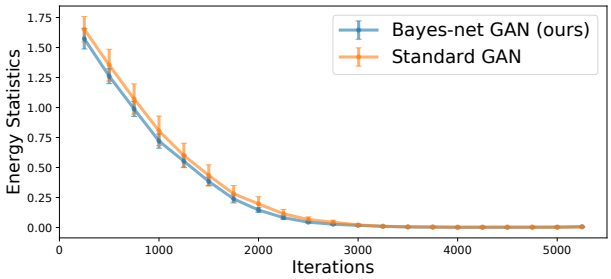


Figure 6: Energy statistics between generated and observed samples at different stages of training.

- **Relative GED**: the graph editing distance between the DAGs predicted from the fake and real samples by a greedy search starting from the ground truth DAG.

The first two metrics characterize the similarity between the joint distributions, while the last two evaluate how accurately the causal structure is learned.

We find that the Bayes-net GAN using the ground truth causal graph consistently outperforms the model-free standard GAN on all four quality metrics (Table 1). For Bayes-net GANs, the relative BIC scores (the second last column) are positive, i.e., the BIC of samples generated by the Bayes-net GANs is even higher than the BIC of observed data. Because the Bayes-net GANs are designed to conveniently capture the ground truth causal dependencies (compared to the other correlations), the likelihood of the ground truth causal structure can further increase. On *the EARTHQUAKE dataset*, we can usually recover the true causal graph from the data generated by the Bayes-net GAN (Fig. 4(d)). This is not the case if we use standard GANs (Fig. 4(c)), where any pair of nodes are directly dependent on each other. In this regard, we conclude the standard GANs cannot efficiently capture the conditional independence relationships among variables.

Next, we study how a Bayes-net GAN learns the causal structure during the training (Fig. 5). In general, discrete Bayes-nets are multi-modals. The Bayes-net GAN learns some strong conditional dependencies at first,

e.g. "Burglary" leads to "JohnCalls" in the second snapshot, although it is not a direct dependence (in fact, "Burglary" triggers "Alarm", then "JohnCalls"). After some training, the dependence relation is further specified, and the edge ("Burglary"→"JohnCalls") is replaced by a pair of new edges, ("Burglary"→"Alarm") and ("Alarm"→"JohnCalls") in the second last snapshot. During training, we rarely observe that the Bayes-net GAN captures any non-existing dependencies (e.g. "Earthquake" and "Burglary"). However, this happens often for standard GANs; see Fig. 4(c) for an example.

The success of learning causal independence structures also simplifies the task of learning joint distribution. Without changing any setup or hyper-parameters, replacing the discriminator with a set of local discriminators brings a performance gain on the first two scores as well (Table 1). Moreover, Bayes-net GANs are computationally efficient when the Bayes-nets are not very large. On average, they converge faster than the standard GAN on Bayes-nets; see Fig. 6 for the averaged curves of energy statistics on *the EARTHQUAKE dataset*. These results highlight the statistical and computational benefits of our principled design of Bayes-net GANs.

Acknowledgements

We thank the Simons Institute for the Theory of Computing, where this collaboration started, during the "Foundations of Deep Learning" program. This project was supported in part by NSF CAREER AWARD 1942230, Simons Fellowship, Qualcomm Faculty Award, IBM Faculty Award, DOE Award 302629-00001 and a sponsorship from Capital One. Constantinos Daskalakis was supported by NSF Awards IIS-1741137, CCF-1617730 and CCF-1901292, by a Simons Investigator Award, by the DOE PhILMs project (No. DE-AC05-76RL01830), by the DARPA award HR00111990021, by a Google Faculty award, and by the MIT Frank Quick Faculty Research and Innovation Fellowship.

References

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

Soheil Feizi, Farzan Farnia, Tony Ginart, and David Tse. Understanding gans: the lqg setting. *arXiv preprint arXiv:1710.10793*, 2017.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Normalized wasserstein for mixture distributions with applications in adversarial learning and domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6500–6508, 2019.

Farzan Farnia, William Wang, Subhro Das, and Ali Jadbabaie. Gat-gmm: Generative adversarial training for gaussian mixture models. *arXiv preprint arXiv:2006.10293*, 2020.

Constantinos Daskalakis and Qinxuan Pan. Square hellinger subadditivity for bayesian networks and its applications to identity testing. In *Conference on Learning Theory*, pages 697–703, 2017.

Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing ising models. *IEEE Transactions on Information Theory*, 65(11):6829–6852, 2019.

Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing bayesian networks. *IEEE Transactions on Information Theory*, 66(5): 3132–3170, 2020.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. In *International Conference on Learning Representations*, 2017.

Kevin B Korb and Ann E Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.

DJ Spiegelhalter. Learning in probabilistic expert systems. *Bayesian statistics*, 4:447–465, 1992.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using

cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019.

Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.

Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*, pages 702–716. Springer, 2016.

Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In *International Conference on Artificial Neural Networks*, pages 703–716. Springer, 2019.

Aidan Clark, Jeff Donahue, and Karen Simonyan. Efficient video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019.

Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

Olof Mogren. C-rnn-gan: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904*, 2016.

Weili Nie, Nina Narodytska, and Ankit Patel. Relgan: Relational generative adversarial networks for text generation. In *International Conference on Learning Representations*, 2018.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. In *International Conference on Learning Representations*, 2019.

Yukai Liu, Rose Yu, Stephan Zheng, Eric Zhan, and Yisong Yue. Naomi: Non-autoregressive multiresolution sequence imputation. In *Advances in Neural Information Processing Systems*, pages 11236–11246, 2019.

Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9066–9075, 2019.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 224–232. JMLR. org, 2017.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Igal Sason and Sergio Verdu. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11): 5973–6006, 2016.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10): 4394–4412, 2006.

Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

Adam M Oberman and Yuanlong Ruan. An efficient linear programming method for optimal transportation. *arXiv preprint arXiv:1509.03668*, 2015.

Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48: 257–263, 1982.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

Anuran Makur. *A study of local approximations in information theory*. PhD thesis, Massachusetts Institute of Technology, 2015.

Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in GANs. In *International Conference on Learning Representations*, 2018.

Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.