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 counter-example 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 $P \ll\gg Q$.*

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 $\lim_{\epsilon\downarrow 0} \alpha > 0$.*

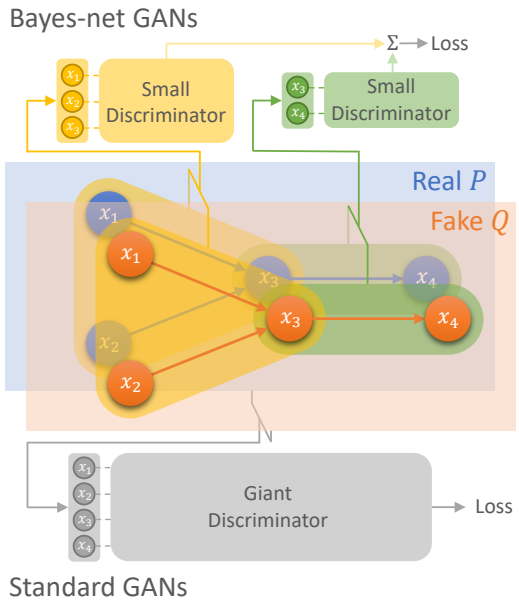


Figure 1: Conceptual diagram of the Bayes-net GANs with local discriminators compared with the standard GANs.

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

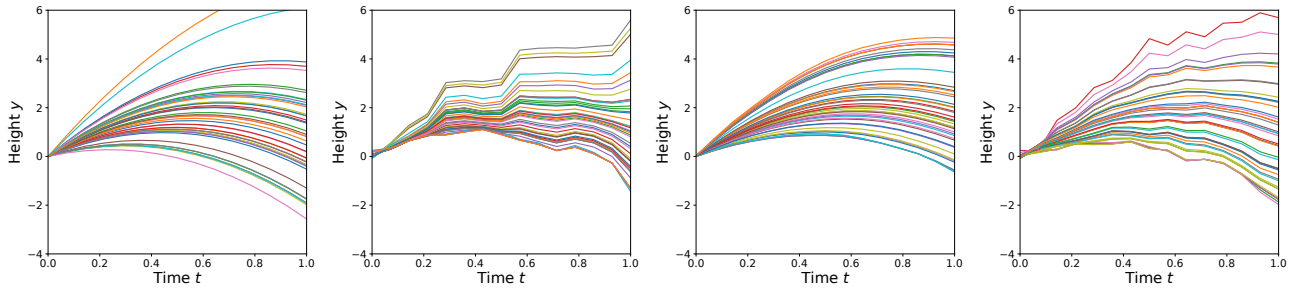
In this section, we consider a simple synthetic dataset that consists of single-variate time-series data $(y_1, \ldots, y_{15})$ representing the y-coordinates of ball throwing trajectories lasting 1 second, where $y_t = v_0 * (t/15) - g(t/15)^2/2$. v_0 is a Gaussian random variable and $g = 9.8$ is 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.

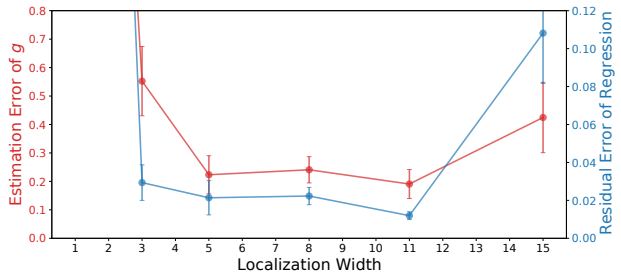


Figure 3: Estimation errors of gravitational acceleration g and residual errors of degree-2 polynomial regression on the generated trajectories with varying localization width.

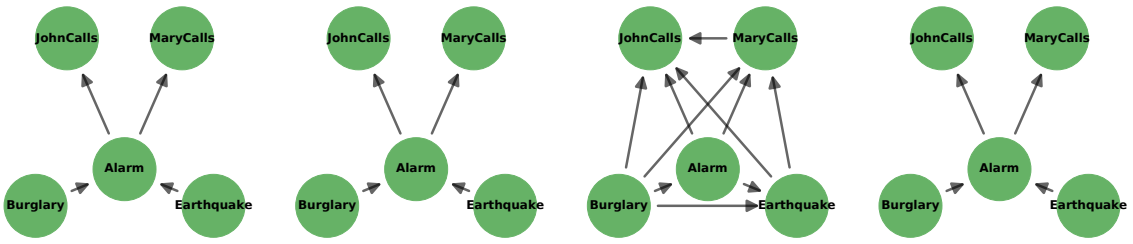
tational acceleration g and the residual errors of degree-2 polynomial regression (which evaluate the "smoothness" of generated trajectories) among GANs with different localization widths. Interestingly, the curves of both metrics demonstrate a U-shaped behavior indicating that there is an optimal localization 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).

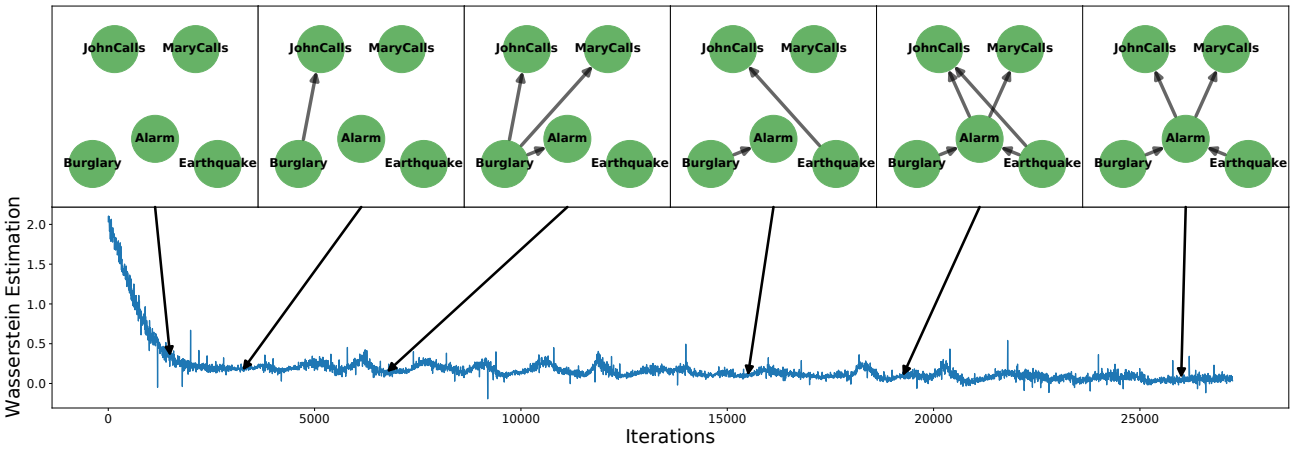


Figure 5: Causal structures predicted from the data generated by the Bayes-net GAN at different stages of training and the Wasserstein loss curve.

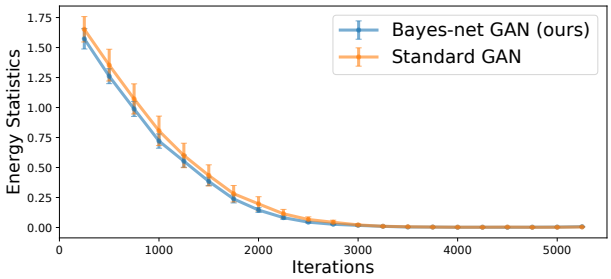


Figure 6: Energy statistics between generated and observed samples at different stages of training.

- **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.

Appendix to GANs with Conditional Independence Graphs: On Subadditivity of Probability Divergences

A Proofs

A.1 Proof of Theorem 1

Proof. The theorem is implicit in (Daskalakis and Pan, 2017). For completeness, we provide a full argument here.

For a pair of Bayes-nets P and Q with respect to a Directed Acyclic Graph (DAG) G, consider the topological ordering $(1, \cdots, n)$ of the nodes of G. Consistent with the topological ordering, consider the following Markov Chain on super-nodes: $X_{\{1,\cdots,n-1\}\setminus\Pi_n} \to X_{\Pi_n} \to X_n$, where Π_n is the set of parents of node n and $\Pi_n \subseteq \{1, \cdots, n-1\}$. We distinguish three cases:

1. $\Pi_n \neq \varnothing$ and $\Pi_n \subsetneqq \{1, \cdots, n-1\}$: In this case, we apply the subadditivity property of δ with respect to Markov Chains to obtain $\delta(P, Q) \leq \delta(P_{\cup_{i=1}^{n-1} X_i}, Q_{\cup_{i=1}^{n-1} X_i}) + \delta(P_{X_{\Pi_n} \cup X_n}, Q_{X_{\Pi_n} \cup X_n})$.
2. $\Pi_n = \{1, \cdots, n-1\}$: In this case, it is trivial that $\delta(P, Q) \equiv \delta(P_{X_{\Pi_n} \cup X_n}, Q_{X_{\Pi_n} \cup X_n}) \leq \delta(P_{\cup_{i=1}^{n-1} X_i}, Q_{\cup_{i=1}^{n-1} X_i}) + \delta(P_{X_{\Pi_n} \cup X_n}, Q_{X_{\Pi_n} \cup X_n})$.
3. $\Pi_n = \varnothing$: In this case, X_n is independent from $(X_1, \ldots, X_{n-1})$ in both Bayes-nets. Thus we apply the subadditivity of δ with respect to product measures to obtain $\delta(P, Q) \leq \delta(P_{\cup_{i=1}^{n-1} X_i}, Q_{\cup_{i=1}^{n-1} X_i}) + \delta(P_{X_n}, Q_{X_n}) \equiv \delta(P_{\cup_{i=1}^{n-1} X_i}, Q_{\cup_{i=1}^{n-1} X_i}) + \delta(P_{X_{\Pi_n} \cup X_n}, Q_{X_{\Pi_n} \cup X_n})$.

We proceed by induction. For each inductive step $k = 1, \cdots, n-2$, we consider the following Markov Chain on super-nodes: $X_{\{1,\cdots,n-k-1\}\setminus\Pi_{n-k}} \to X_{\Pi_{n-k}} \to X_{n-k}$. No matter what Π_{n-k} is, we always have: $\delta(P_{\cup_{i=1}^{n-k} X_i}, Q_{\cup_{i=1}^{n-k} X_i}) \leq \delta(P_{\cup_{i=1}^{n-k-1} X_i}, Q_{\cup_{i=1}^{n-k-1} X_i}) + \delta(P_{X_{\Pi_{n-k}} \cup X_{n-k}}, Q_{X_{\Pi_{n-k}} \cup X_{n-k}})$. In the end of the induction, we obtain: $\delta(P, Q) \leq \delta(P_{X_1}, Q_{X_1}) + \sum_{i=2}^{n} \delta(P_{\Pi_i \cup X_i}, Q_{\Pi_i \cup X_i}) \equiv \sum_{i=1}^{n} \delta(P_{\Pi_i \cup X_i}, Q_{\Pi_i \cup X_i})$, since $\Pi_1 \equiv \varnothing$. The subadditivity of δ on Bayes-nets is proved. □

A.2 Proof of Theorem 2

Proof. The subadditivity of squared Hellinger distance is proved in Theorem 2.1 of (Daskalakis and Pan, 2017). Here, we repeat the proof for completeness.

Given Theorem 1, we only need to show the following:

1. For two Markov Chains P, Q on variables $X \to Y \to Z$, it holds that $\mathrm{H}^2(P_{XYZ}, Q_{XYZ}) \leq \mathrm{H}^2(P_{XY}, Q_{XY}) + \mathrm{H}^2(P_{YZ}, Q_{YZ})$.
2. For two product measures P, Q on variables X, Y, it holds that $\mathrm{H}^2(P_{XY}, Q_{XY}) \leq \mathrm{H}^2(P_X, Q_X) + \mathrm{H}^2(P_Y, Q_Y)$.

We first show the subadditivity with respect to Markov Chains. Using the Markov property, we know $P_{XYZ} = P_{Z|XY} P_{XY} = P_{Z|Y} P_{XY}$ (and the same holds for Q), thus,

$$
\begin{aligned}
&\mathrm{H}^2(P_{XYZ}, Q_{XYZ}) \\
&= 1 - \int \sqrt{P_{XYZ} Q_{XYZ}} \mathrm{d}x \mathrm{d}y \mathrm{d}z \\
&= 1 - \int \sqrt{P_{XY} Q_{XY}} \left(\int \sqrt{P_{Z|Y} Q_{Z|Y}} \mathrm{d}z \right) \mathrm{d}x \mathrm{d}y \\
&= 1 - \int \frac{1}{2} (P_Y + Q_Y) \left(\int \sqrt{P_{Z|Y} Q_{Z|Y}} \mathrm{d}z \right) \mathrm{d}y + \int \frac{1}{2} \left(\sqrt{P_{XY}} - \sqrt{Q_{XY}} \right)^2 \left(\int \sqrt{P_{Z|Y} Q_{Z|Y}} \mathrm{d}z \right) \mathrm{d}x \mathrm{d}y
\end{aligned}
$$

Since all densities are non-negative, we have $\sqrt{P_Y Q_Y} \leq \frac{1}{2}(P_Y + Q_Y)$ and $\sqrt{P_{Z|Y} Q_{Z|Y}} \leq \frac{1}{2}\left(P_{Z|Y} + Q_{Z|Y}\right)$ point-wisely. Thus,

$$
\begin{aligned}
&\mathrm{H}^2(P_{XYZ}, Q_{XYZ}) \\
&\leq 1 - \int \sqrt{P_Y Q_Y}\left(\int \sqrt{P_{Z|Y} Q_{Z|Y}}\mathrm{d}z\right)\mathrm{d}y + \int \frac{1}{2}\left(\sqrt{P_{XY}} - \sqrt{Q_{XY}}\right)^2 \left(\int \frac{1}{2}\left(P_{Z|Y} + Q_{Z|Y}\right)\mathrm{d}z\right)\mathrm{d}x\mathrm{d}y \\
&= \left(1 - \int \sqrt{P_{YZ} Q_{YZ}}\mathrm{d}y\mathrm{d}z\right) + \frac{1}{2}\int \left(\sqrt{P_{XY}} - \sqrt{Q_{XY}}\right)^2 \mathrm{d}x\mathrm{d}y \\
&= \mathrm{H}^2(P_{XY}, Q_{XY}) + \mathrm{H}^2(P_{YZ}, Q_{YZ})
\end{aligned}
$$

It remains to show the subadditivity with respect to product measures. If P, Q are product measures over X, Y, then $P_{XY} = P_X P_Y$ and $Q_{XY} = Q_X Q_Y$. Since all densities are non-negative, we have $\sqrt{P_Y Q_Y} \leq \frac{1}{2}(P_Y + Q_Y)$ point-wise. Hence,

$$
\begin{aligned}
&\mathrm{H}^2(P_{XY}, Q_{XY}) \\
&= 1 - \int \sqrt{P_{XY} Q_{XY}}\mathrm{d}x\mathrm{d}y \\
&= 1 - \int \sqrt{P_X Q_X}\left(\int \sqrt{P_Y Q_Y}\mathrm{d}y\right)\mathrm{d}x \\
&= 1 - \int \frac{1}{2}(P_X + Q_X)\left(\int \sqrt{P_Y Q_Y}\mathrm{d}y\right)\mathrm{d}x + \int \frac{1}{2}\left(\sqrt{P_X} - \sqrt{Q_X}\right)^2 \left(\int \sqrt{P_Y Q_Y}\mathrm{d}y\right)\mathrm{d}x \\
&\leq 1 - \left(\int \frac{1}{2}(P_X + Q_X)\mathrm{d}x\right)\left(\int \sqrt{P_Y Q_Y}\mathrm{d}y\right) + \int \frac{1}{2}\left(\sqrt{P_X} - \sqrt{Q_X}\right)^2 \left(\int \frac{1}{2}(P_Y + Q_Y)\mathrm{d}y\right)\mathrm{d}x \\
&= 1 - \int \sqrt{P_Y Q_Y}\mathrm{d}y + \int \frac{1}{2}\left(\sqrt{P_X} - \sqrt{Q_X}\right)^2 \mathrm{d}x \\
&= \mathrm{H}^2(P_X, Q_X) + \mathrm{H}^2(P_Y, Q_Y)
\end{aligned}
$$

□

A.3 Proof of Theorem 3

Proof. The subadditivity of KL-divergence is claimed in (Daskalakis and Pan, 2017) without proof. Here, we provide a proof for completeness.

Given Theorem 1, we only need to show the following:

1. For two Markov Chains P, Q on variables $X \to Y \to Z$, it holds that $\mathrm{KL}(P_{XYZ}, Q_{XYZ}) \leq \mathrm{KL}(P_{XY}, Q_{XY}) + \mathrm{KL}(P_{YZ}, Q_{YZ})$.
2. For two product measures P, Q on variables X, Y, it holds that $\mathrm{KL}(P_{XY}, Q_{XY}) \leq \mathrm{KL}(P_X, Q_X) + \mathrm{KL}(P_Y, Q_Y)$.

We first show the subadditivity with respect to Markov Chains. The Markov property implies $P_{XYZ} = P_{XY} P_{YZ} / P_Y$ (and the same holds for Q). Thus,

$$
\begin{aligned}
\mathrm{KL}(P_{XYZ}, Q_{XYZ}) &= \int P_{XYZ} \log\left(\frac{P_{XY}}{Q_{XY}}\frac{P_{YZ}}{Q_{YZ}}\Big/\frac{P_Y}{Q_Y}\right)\mathrm{d}x\mathrm{d}y\mathrm{d}z \\
&= \int P_{XY} \log\left(\frac{P_{XY}}{Q_{XY}}\right)\mathrm{d}x\mathrm{d}y + \int P_{YZ} \log\left(\frac{P_{YZ}}{Q_{YZ}}\right)\mathrm{d}y\mathrm{d}z - \int P_Y \log\left(\frac{P_Y}{Q_Y}\right)\mathrm{d}y \\
&= \mathrm{KL}(P_{XY}, Q_{XY}) + \mathrm{KL}(P_{YZ}, Q_{YZ}) - \mathrm{KL}(P_Y, Q_Y)
\end{aligned}
$$

The subadditivity follows from the non-negativity of KL-divergence. Additivity holds when $\mathrm{KL}(P_Y, Q_Y) = 0$.

It remains to show the subadditivity with respect to product measures. We will, in fact, show additivity rather

than subadditivity. If P, Q are product measures over X, Y, then $P_{XY} = P_X P_Y$ and $Q_{XY} = Q_X Q_Y$, hence,

$$\begin{aligned}\mathrm{KL}(P_{XY}, Q_{XY}) &= \int P_{XY} \log\left(\frac{P_X}{Q_X}\frac{P_Y}{Q_Y}\right) \mathrm{d}x\mathrm{d}y \\ &= \int P_X \log\left(\frac{P_X}{Q_X}\right) \mathrm{d}x + \int P_Y \log\left(\frac{P_Y}{Q_Y}\right) \mathrm{d}y \\ &= \mathrm{KL}(P_X, Q_X) + \mathrm{KL}(P_Y, Q_Y).\end{aligned}$$

□

A.4 Proof of Corollary 5

Proof. The subadditivity of Jensen-Shannon divergence follows from:

1. The subadditivity of squared Hellinger distance (Theorem 2).
2. f-Divergence inequalities (Theorem 11 of (Sason and Verdu, 2016), repeated as Theorem 16 in Appendix B.2): for any two densities P and Q,

$$(\ln 2)\mathrm{H}^2(P, Q) \leq \mathrm{JS}(P, Q) \leq \mathrm{H}^2(P, Q)$$

Combining the inequalities implies that, for any pair of Bayes-nets P, Q with respect to a DAG G, we have,

$$\mathrm{JS}(P, Q) \leq \mathrm{H}^2(P, Q) \leq \sum_{i=1}^{n} \mathrm{H}^2(P_{\Pi_i \cup X_i}, Q_{\Pi_i \cup X_i}) \leq \frac{1}{\ln 2} \sum_{i=1}^{n} \mathrm{JS}(P_{\Pi_i \cup X_i}, Q_{\Pi_i \cup X_i})$$

This proves that Jensen-Shannon divergence satisfies $(1/\ln 2)$-linear subadditivity on Bayes-nets.

Note that we assume natural logarithm is used in the definition of Jensen-Shannon divergence when deriving the inequalities between $\mathrm{JS}(P, Q)$ and $\mathrm{H}^2(P, Q)$ (see Theorem 16 for details). However, the choice of the base of the logarithm does not affect the $(1/\ln 2)$-linear subadditivity of Jensen-Shannon divergence. □

A.5 Proof of Theorem 6

In the following proofs, we extensively use the Integral Probability Metric (IPM) formula of Total Variation distance (Müller, 1997). If $\mathcal{F}$ is the set of measurable functions on Ω taking values in $[0, 1]$, then,

$$\mathrm{TV}(P, Q) = \sup_{\phi \in \mathcal{F}} \left| \mathbb{E}_{x \sim P}[\phi(x)] - \mathbb{E}_{x \sim Q}[\phi(x)] \right|$$

Lemma 13. *Let P and Q be two Bayes-nets with respect to DAG $X \to Y \to Z$. Then,*

$$\mathrm{TV}(P_{XYZ}, Q_{XYZ}) \leq \mathrm{TV}(P_{XY}, Q_{XY}) + \mathrm{TV}(P_Y, Q_Y) + \mathrm{TV}(P_{YZ}, Q_{YZ})$$

Proof. We do a hybrid argument. By the triangle inequality, we have:

$$\mathrm{TV}(P_{XYZ}, Q_{XYZ}) \leq \mathrm{TV}(P_{XYZ}, P_{XY} Q_{Z|Y}) + \mathrm{TV}(P_{XY} Q_{Z|Y}, Q_{XYZ})$$

We bound each term on the right-hand side separately.

Let us start with the second term. Let $\mathcal{F}_{xy}$ be the set of measurable functions on variables x and y taking values in $[0, 1]$, and $\mathcal{F}_{xyz}$ be the set of measurable functions on variables x, y, z taking values in $[0, 1]$, etc. Using the Markov property, we know $P_{XYZ} = P_{XY} P_{Z|Y} = P_Y P_{X|Y} P_{Z|Y}$ (and the same holds for Q). Then,

$$\begin{aligned}\mathrm{TV}(P_{XY} Q_{Z|Y}, Q_{XYZ}) &= \sup_{\phi \in \mathcal{F}_{xyz}} \left| \mathbb{E}_{P_{XY} Q_{Z|Y}}[\phi(x, y, z)] - \mathbb{E}_{Q_{XYZ}}[\phi(x, y, z)] \right| \\ &= \sup_{\phi \in \mathcal{F}_{xyz}} \left| \mathbb{E}_{P_{XY}}\left[\mathbb{E}_{Q_{Z|Y}}[\phi(x, y, z)]\right] - \mathbb{E}_{Q_{XY}}\left[\mathbb{E}_{Q_{Z|Y}}[\phi(x, y, z)]\right] \right| \\ &\leq \sup_{\phi \in \mathcal{F}_{xy}} \left| \mathbb{E}_{P_{XY}}\left[\phi(x, y)\right] - \mathbb{E}_{Q_{XY}}\left[\phi(x, y)\right] \right| \\ &\equiv \mathrm{TV}(P_{XY}, Q_{XY})\end{aligned}$$

Let us now bound the first term,

$$\begin{aligned}
\mathrm{TV}(P_{XYZ}, P_{XY}Q_{Z|Y}) &= \sup_{\phi\in\mathcal{F}_{xyz}} \left|\mathbb{E}_{P_{XYZ}}[\phi(x,y,z)] - \mathbb{E}_{P_{XY}Q_{Z|Y}}[\phi(x,y,z)]\right| \\
&= \sup_{\phi\in\mathcal{F}_{xyz}} \left|\mathbb{E}_{P_YP_{Z|Y}}\left[\mathbb{E}_{P_{X|Y}}[\phi(x,y,z)]\right] - \mathbb{E}_{P_YQ_{Z|Y}}\left[\mathbb{E}_{P_{X|Y}}[\phi(x,y,z)]\right]\right| \\
&\le \sup_{\phi\in\mathcal{F}_{yz}} \left|\mathbb{E}_{P_YP_{Z|Y}}\left[\phi(y,z)\right] - \mathbb{E}_{P_YQ_{Z|Y}}\left[\phi(y,z)\right]\right| \\
&\le \sup_{\phi\in\mathcal{F}_{yz}} \left|\mathbb{E}_{P_YP_{Z|Y}}\left[\phi(y,z)\right] - \mathbb{E}_{Q_YQ_{Z|Y}}\left[\phi(y,z)\right]\right| \\
&\quad + \sup_{\phi\in\mathcal{F}_{yz}} \left|\mathbb{E}_{Q_YQ_{Z|Y}}\left[\phi(y,z)\right] - \mathbb{E}_{P_YQ_{Z|Y}}\left[\phi(y,z)\right]\right| \\
&= \mathrm{TV}(P_{YZ}, Q_{YZ}) + \sup_{\phi\in\mathcal{F}_{yz}} \left|\mathbb{E}_{Q_Y}\left[\mathbb{E}_{Q_{Z|Y}}\left[\phi(y,z)\right]\right] - \mathbb{E}_{P_Y}\left[\mathbb{E}_{Q_{Z|Y}}\left[\phi(y,z)\right]\right]\right| \\
&\le \mathrm{TV}(P_{YZ}, Q_{YZ}) + \sup_{\phi\in\mathcal{F}_{y}} \left|\mathbb{E}_{Q_Y}\left[\phi(y)\right] - \mathbb{E}_{P_Y}\left[\phi(y)\right]\right| \\
&\le \mathrm{TV}(P_{YZ}, Q_{YZ}) + \mathrm{TV}(P_Y, Q_Y)
\end{aligned}$$

Combining the two inequalities concludes the proof. □

Lemma 14. *Let P and Q be two product measures over variables X and Y. Then,*

$$\mathrm{TV}(P_{XY}, Q_{XY}) \le \mathrm{TV}(P_X, Q_X) + \mathrm{TV}(P_Y, Q_Y)$$

Proof. By the triangle inequality, we have:

$$\mathrm{TV}(P_{XY}, Q_{XY}) \le \mathrm{TV}(P_{XY}, P_XQ_Y) + \mathrm{TV}(P_XQ_Y, Q_{XY})$$

We bound each term on the right hand side separately. Let $\mathcal{F}_{xy}$ be the set of measurable functions on variables x and y taking values in $[0,1]$, and $\mathcal{F}_y$ be the set of measurable functions on variable y taking values in $[0,1]$, etc. Then,

$$\begin{aligned}
\mathrm{TV}(P_{XY}, P_XQ_Y) &= \sup_{\phi\in\mathcal{F}_{xy}} \left|\mathbb{E}_{P_{XY}}[\phi(x,y)] - \mathbb{E}_{P_XQ_Y}[\phi(x,y)]\right| \\
&= \sup_{\phi\in\mathcal{F}_{xy}} \left|\mathbb{E}_{P_Y}\left[\mathbb{E}_{P_X}[\phi(x,y)]\right] - \mathbb{E}_{Q_Y}\left[\mathbb{E}_{P_X}[\phi(x,y)]\right]\right| \\
&\le \sup_{\phi\in\mathcal{F}_{y}} \left|\mathbb{E}_{P_Y}\left[\phi(y)\right] - \mathbb{E}_{Q_Y}\left[\phi(y)\right]\right| \\
&\equiv \mathrm{TV}(P_Y, Q_Y)
\end{aligned}$$

Similarly, we get $\mathrm{TV}(P_XQ_Y, Q_{XY}) \le \mathrm{TV}(P_X, Q_X)$. Combining the two inequalities concludes the proof. □

Proof of Theorem 6: Similar to the proof of Theorem 1, for a pair of Bayes-nets P and Q with respect to a DAG G, we perform induction on each nodes of G. Consider the topological ordering $(1,\cdots,n)$ of the nodes of G. Consistent with the topological ordering, consider the following Markov Chain on super-nodes: $X_{\{1,\cdots,n-1\}\setminus\Pi_n} \to X_{\Pi_n} \to X_n$, where Π_n is the set of parents of node n and $\Pi_n \subseteq \{1,\cdots,n-1\}$. We distinguish three cases:

1. $\Pi_n \neq \varnothing$ and $\Pi_n \subsetneqq \{1,\cdots,n-1\}$: In this case, we apply Lemma 13 to get $\mathrm{TV}(P,Q) \le \mathrm{TV}(P_{\cup_{i=1}^{n-1}X_i}, Q_{\cup_{i=1}^{n-1}X_i}) + \mathrm{TV}(P_{X_{\Pi_n}}, Q_{X_{\Pi_n}}) + \mathrm{TV}(P_{X_{\Pi_n}\cup X_n}, Q_{X_{\Pi_n}\cup X_n})$.

2. $\Pi_n = \{1,\cdots,n-1\}$: In this case, it is trivial that $\mathrm{TV}(P,Q) \equiv \mathrm{TV}(P_{X_{\Pi_n}\cup X_n}, Q_{X_{\Pi_n}\cup X_n}) \le \mathrm{TV}(P_{\cup_{i=1}^{n-1}X_i}, Q_{\cup_{i=1}^{n-1}X_i}) + \mathrm{TV}(P_{X_{\Pi_n}}, Q_{X_{\Pi_n}}) + \mathrm{TV}(P_{X_{\Pi_n}\cup X_n}, Q_{X_{\Pi_n}\cup X_n})$.

3. $\Pi_n = \varnothing$: In this case, X_n is independent from $(X_1, \ldots, X_{n-1})$ in both Bayes-nets. Thus we apply Lemma 14 to get $\mathrm{TV}(P, Q) \leq \mathrm{TV}(P_{\cup_{i=1}^{n-1} X_i}, Q_{\cup_{i=1}^{n-1} X_i}) + \mathrm{TV}(P_{X_n}, Q_{X_n}) \equiv \mathrm{TV}(P_{\cup_{i=1}^{n-1} X_i}, Q_{\cup_{i=1}^{n-1} X_i}) + \mathrm{TV}(P_{X_{\Pi_n}}, Q_{X_{\Pi_n}}) + \mathrm{TV}(P_{X_{\Pi_n} \cup X_n}, Q_{X_{\Pi_n} \cup X_n})$, where $\mathrm{TV}(P_{X_{\Pi_n}}, Q_{X_{\Pi_n}}) = 0$ and $\mathrm{TV}(P_{X_{\Pi_n} \cup X_n}, Q_{X_{\Pi_n} \cup X_n}) = \mathrm{TV}(P_{X_1}, Q_{X_1})$ as $\Pi_n = \varnothing$.

We proceed by induction. For each inductive step $k = 1, \cdots, n-2$, we consider the following Markov Chain on super-nodes: $X_{\{1,\cdots,n-k-1\} \setminus \Pi_{n-k}} \to X_{\Pi_{n-k}} \to X_{n-k}$. No matter what Π_{n-k} is, we always have: $\mathrm{TV}(P_{\cup_{i=1}^{n-k} X_i}, Q_{\cup_{i=1}^{n-k} X_i}) \leq \mathrm{TV}(P_{\cup_{i=1}^{n-k-1} X_i}, Q_{\cup_{i=1}^{n-k-1} X_i}) + \mathrm{TV}(P_{X_{\Pi_{n-k}}}, Q_{X_{\Pi_{n-k}}}) + \mathrm{TV}(P_{X_{\Pi_{n-k}} \cup X_{n-k}}, Q_{X_{\Pi_{n-k}} \cup X_{n-k}})$. In the end of the induction, we obtain: $\mathrm{TV}(P, Q) \leq \mathrm{TV}(P_{X_1}, Q_{X_1}) + \sum_{i=2}^{n} \big(\mathrm{TV}(P_{\Pi_i \cup X_i}, Q_{\Pi_i \cup X_i}) + \mathrm{TV}(P_{\Pi_i}, Q_{\Pi_i})\big)$. Since $\Pi_1 \equiv \varnothing$, we know $\mathrm{TV}(P_{X_{\Pi_1}}, Q_{X_{\Pi_1}}) = 0$ and $\mathrm{TV}(P_{X_{\Pi_1} \cup X_1}, Q_{X_{\Pi_1} \cup X_1}) = \mathrm{TV}(P_{X_1}, Q_{X_1})$. Hence, we conclude that,

$$\mathrm{TV}(P, Q) \leq \sum_{i=1}^{n} \Big(\mathrm{TV}(P_{\Pi_i \cup X_i}, Q_{\Pi_i \cup X_i}) + \mathrm{TV}(P_{\Pi_i}, Q_{\Pi_i})\Big)$$

Now we relate this inequality to the notion of linear subadditivity. For two densities P and Q on variables X, Y, it holds that,

$$\begin{aligned}
\mathrm{TV}(P_X, Q_X) &\equiv \frac{1}{2} \int \Big| P_X - Q_X \Big| \mathrm{d}x \\
&= \frac{1}{2} \int \Big| \int P_{XY} \mathrm{d}y - \int Q_{XY} \mathrm{d}y \Big| \mathrm{d}x \\
&\leq \frac{1}{2} \int \left(\int \Big| P_{XY} - Q_{XY} \Big| \mathrm{d}y \right) \mathrm{d}x \\
&\equiv \mathrm{TV}(P_{XY}, Q_{XY})
\end{aligned}$$

Applying this inequality to X_{Π_i} and X_i, for any $i \in \{1, \cdots, n\}$, we obtain, $\mathrm{TV}(P_{\Pi_i}, Q_{\Pi_i}) \leq \mathrm{TV}(P_{\Pi_i \cup X_i}, Q_{\Pi_i \cup X_i})$. Thus,

$$\mathrm{TV}(P, Q) \leq 2 \sum_{i=1}^{n} \mathrm{TV}(P_{\Pi_i \cup X_i}, Q_{\Pi_i \cup X_i})$$

This concludes that Total Variation distance satisfies 2-linear subadditivity on Bayes-nets. □

A.6 Proof of Corollary 7

Proof. If Ω is a finite (and therefore bounded) metric space, there exist two-way bounds between p-Wasserstein distance and Total Variation distance (see Theorem 20 in Appendix C.1 for details), namely,

$$\mathrm{W}_p(P, Q)^p / \mathrm{diam}(\Omega)^p \leq \mathrm{TV}(P, Q) \leq \mathrm{W}_p(P, Q)^p / d_{\min}^p$$

where $\mathrm{diam}(\Omega) = \max\{d(x, y) | x, y \in \Omega\}$ is the diameter of the space Ω and $d_{\min} = \min_{x \neq y} d(x, y)$ is the smallest distance between pairs of distance points in Ω. For $p \geq 1$, this directly implies the $(2^{1/p} \mathrm{diam}(\Omega)/d_{\min})$-linear subadditivity of p-Wasserstein distance on Bayes-nets on finite Ω,

$$\mathrm{W}_p(P, Q) \leq \frac{2^{1/p} \mathrm{diam}(\Omega)}{d_{\min}} \sum_{i=1}^{n} \mathrm{W}_p(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}})$$

via the 2-linear subadditivity of Total Variation distance (Theorem 6). □

A.7 Proof of Theorem 8

Proof. For reference, we repeat the three conditions of the subadditivity of neural distances here:

(1) The space Ω is bounded, i.e. $\mathrm{diam}(\Omega) < \infty$.

(2) For any $i \in \{1, \cdots, n\}$, discriminator class $\mathcal{F}_i$ is larger than the set of neural networks with a single neuron, which have ReLU activation and bounded parameters, i.e. $\mathcal{F}_i \supseteq \{\max\{w^T x + b, 0\} \big| w \in \mathbb{R}^{D_i}, b \in \mathbb{R}, \|[w, b]\|_2 = 1\}$, where D_i is the number of dimensions of variables $X_i \cup X_{\Pi_i}$.

(3) For any $i \in \{1, \cdots, n\}$, $\log(P_{X_i \cup X_{\Pi_i}} / Q_{X_i \cup X_{\Pi_i}})$ exists, and is bounded and Lipschitz continuous.

For two distributions P, Q and a set of discriminators $\mathcal{F}$ satisfying all the three conditions, by Theorem 28 we know that for any $i \in \{1, \cdots, n\}$, $\log(P_{X_i \cup X_{\Pi_i}} / Q_{X_i \cup X_{\Pi_i}})$ is inside the closure of the linear span of $\mathcal{F}_i$, i.e. $\log(P_{X_i \cup X_{\Pi_i}} / Q_{X_i \cup X_{\Pi_i}}) \in \text{cl}(\text{span}\mathcal{F}_i)$. Moreover, each $\log(P_{X_i \cup X_{\Pi_i}} / Q_{X_i \cup X_{\Pi_i}})$ is approximated by the corresponding $\mathcal{F}_i$ with an error decay function, denoted by $\varepsilon_i(r)$. Using Theorem 27, we upper-bound each Symmetric KL divergence between local marginals, $\text{SKL}(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}})$, by a linear function of the corresponding neural distance $d_{\mathcal{F}}(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}})$,

$$\text{SKL}(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}}) \leq 2\varepsilon_i(r) + r d_{\mathcal{F}_i}(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}}) \qquad \forall r \geq 0, \forall i \in \{1, \cdots, n\}$$

Because of the condition (3): each $\log(P_{X_i \cup X_{\Pi_i}} / Q_{X_i \cup X_{\Pi_i}})$ is bounded and Lipschitz continuous, there exists a constant $\eta_i > 0$, such that,

$$\left|\log(P_{X_i \cup X_{\Pi_i}} / Q_{X_i \cup X_{\Pi_i}})\right| < \eta_i$$

and for any $x, y \in \Omega_i$ (which is the space of variables $X_i \cup X_{\Pi_i}$), it holds that,

$$\left|\log(P_{X_i \cup X_{\Pi_i}}(x) / Q_{X_i \cup X_{\Pi_i}}(x)) - \log(P_{X_i \cup X_{\Pi_i}}(y) / Q_{X_i \cup X_{\Pi_i}}(y))\right| \leq \frac{\eta_i}{\text{diam}(\Omega_i)} \|x - y\|$$

Again, by Theorem 28, we get an efficient upper-bound on $\varepsilon_i(r)$,

$$\varepsilon_i(r) \leq C(D_i)\eta_i \left(\frac{r}{\eta_i}\right)^{-\frac{2}{D_i+1}} \log\left(\frac{r}{\eta_i}\right) \qquad \forall r \geq R(D_i) > \mathrm{e}^{\frac{D_i+1}{2}}\eta_i, \forall i \in \{1, \cdots, n\}$$

where $C(D_i)$ and $R(D_i)$ are constants that only depend on the dimensionality, D_i, of variables $X_i \cup X_{\Pi_i}$. More specifically, $D_i = (k_i + 1)d \leq (k_{\max} + 1)d$, where k_i is the in-degree of node i, d is the dimensionality of each variable of the Bayes-nets, and $k_{\max}$ is the maximum in-degree of G.

Because $C(D_i)$ and $R(D_i)$ are increasing functions of the dimensionality D_i, and for $r \geq R(D_i) > \mathrm{e}^{\frac{D_i+1}{2}}\eta_i$, $\eta_i (r/\eta_i)^{-\frac{2}{D_i+1}} \log(r/\eta_i)$ is an increasing function of η_i, summing up the inequalities for all $i \in \{1, \cdots, n\}$ gives,

$$\sum_{i=1}^{n} \varepsilon_i(r) \leq nC(D_{\max})\eta_{\max} \left(\frac{r}{\eta_{\max}}\right)^{-\frac{2}{D_{\max}+1}} \log\left(\frac{r}{\eta_{\max}}\right) \qquad \forall r \geq R(D_{\max})$$

where $D_{\max} = \max\{D_i\} = (k_{\max} + 1)d$ and $\eta_{\max} = \max\{\eta_i\}$.

Now, we sum up the inequalities $\text{SKL}(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}}) \leq 2\varepsilon_i(r) + r d_{\mathcal{F}}(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}})$ for $r \geq R(D_{\max})$ for all $i \in \{1, \cdots, n\}$. Because of the subadditivity of Symmetric KL divergence on Bayes-nets P, Q (Corollary 4), we get,

$$\text{SKL}(P, Q) - 2\sum_{i=1}^{n} \varepsilon_i(r) \leq r \sum_{i=1}^{n} d_{\mathcal{F}_i}(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}}) \qquad \forall r \geq R(D_{\max})$$

That is, the neural distances defined by $\mathcal{F}_1, \ldots, \mathcal{F}_n$ satisfy r-linear subadditivity for,

$$r \geq R(D_{\max})$$

with error,

$$\epsilon = 2\sum_{i=1}^{n} \varepsilon_i(r) = \mathcal{O}\left(n r^{-\frac{2}{D_{\max}+1}} \log r\right)$$

with respect to the Symmetric KL divergence on Bayes-nets.

Note that r and ϵ are constants independent of the Bayes-nets P, Q and the sets of discriminator classes $\{\mathcal{F}_1, \cdots, \mathcal{F}_n\}$. And $D_{\max} = (k_{\max} + 1)d$ where $k_{\max}$ is the maximum in-degree of G and d is the dimensionality of each variable of the Bayes-nets. □

A.8 Proof of Theorem 9

Proof. We first give a proof when condition (2) holds. For a pair of MRFs P and Q with the same factorization (thus with the same underlying graph G),

$$P(x) = \prod_{C \in \mathcal{C}} \psi_C^P(X_C) \qquad Q(x) = \prod_{C \in \mathcal{C}} \psi_C^Q(X_C)$$

The Symmetric KL divergence between P and Q,

$$\mathrm{SKL}(P, Q) \coloneqq \mathrm{KL}(P, Q) + \mathrm{KL}(Q, P) = \mathbb{E}_{x \sim P}\left[\log(P/Q)\right] - \mathbb{E}_{x \sim Q}\left[\log(P/Q)\right]$$

can be decomposed into,

$$\mathrm{SKL}(P, Q) = \sum_{C \in \mathcal{C}} \left(\mathbb{E}_{x_C \sim P_{X_C}} \left[\log(\psi_C^P/\psi_C^Q)\right] - \mathbb{E}_{x_C \sim Q_{X_C}} \left[\log(\psi_C^P/\psi_C^Q)\right] \right)$$

Where each term in the summation is upper-bounded by the 1-Wasserstein distance between P_{X_C} and Q_{X_C} up to a constant factor,

$$\begin{aligned} &\mathbb{E}_{x_C \sim P_{X_C}} \left[\log(\psi_C^P/\psi_C^Q)\right] - \mathbb{E}_{x_C \sim Q_{X_C}} \left[\log(\psi_C^P/\psi_C^Q)\right] \\ &\qquad \leq \eta_C \mathrm{W}_1(P_{X_C}, Q_{X_C}) \coloneqq \eta_C \sup_{\phi \text{ 1-Lipschitz}} \left\{ \mathbb{E}_{x_C \sim P_{X_C}}[\phi(x)] - \mathbb{E}_{x_C \sim Q_{X_C}}[\phi(x)] \right\} \end{aligned}$$

if $\log(\psi_C^P/\psi_C^Q)$ is Lipschitz continuous with Lipschitz constant η_C. Summing up the inequalities for all maximal cliques $C \in \mathcal{C}$, we get,

$$\mathrm{SKL}(P, Q) \leq \eta_{\max} \sum_{C \in \mathcal{C}} \mathrm{W}_1(P_{X_C}, Q_{X_C})$$

where $\eta_{\max} = \max\{\eta_C | C \in \mathcal{C}\}$ is the maximum Lipschitz constant. That is, 1-Wasserstein distance satisfies $\eta_{\max}$-linear subadditivity with respect to the Symmetric KL Divergence on MRFs.

We conclude the proof by showing that condition (1) implies condition (2). For a discrete and finite space Ω, each $\log(\psi_C^P/\psi_C^Q)$ maps any configuration x_C in $\Omega_C \subseteq \mathbb{R}^{|C|d}$ (the space of variables X_C) to a real number, where $|C|$ is the size of clique C and d is the dimensionality of each variable of the MRFs. We can always extend the domain of $\log(\psi_C^P/\psi_C^Q)$ to $\mathbb{R}^{|C|d}$, so that the extended function is Lipschitz continuous with Lipschitz constant,

$$\eta_C = \max \left\{ \frac{\left| \log(\psi_C^P(x_C^1)/\psi_C^Q(x_C^1)) - \log(\psi_C^P(x_C^2)/\psi_C^Q(x_C^2)) \right|}{\|x_C^1 - x_C^2\|} \middle| x_C^1 \neq x_C^2 \in \Omega_C \right\}$$

The rest of the proof follows from the proof above. □

A.9 Proof of Corollary 10

Proof. The proof is similar to the proof of Theorem 8 (in Appendix A.7) with a few differences. For a pair of MRFs P and Q with the same factorization (thus with the same underlying graph G), the Symmetric KL divergence between P and Q can be decomposed into,

$$\mathrm{SKL}(P, Q) = \sum_{C \in \mathcal{C}} \left(\mathbb{E}_{x_C \sim P_{X_C}} \left[\log(\psi_C^P/\psi_C^Q)\right] - \mathbb{E}_{x_C \sim Q_{X_C}} \left[\log(\psi_C^P/\psi_C^Q)\right] \right)$$

For two distributions P, Q and a set of discriminators $\mathcal{F}$ satisfying all the three conditions, by Theorem 28 we know that for any $C \in \mathcal{C}$, $\log(\psi_C^P/\psi_C^Q)$ is inside the closure of the linear span of $\mathcal{F}_C$, i.e. $\log(\psi_C^P/\psi_C^Q) \in \mathrm{cl}(\mathrm{span}\mathcal{F}_C)$. Moreover, each $\log(\psi_C^P/\psi_C^Q)$ is approximated by the corresponding $\mathcal{F}_C$ with an error decay function, denoted by $\varepsilon_C(r)$. Using Theorem 27 and assign $g = \log(\psi_C^P/\psi_C^Q)$ (instead of $\log(P_{X_C}/Q_{X_C})$), we get,

$$\mathbb{E}_{x_C \sim P_{X_C}} \left[\log(\psi_C^P/\psi_C^Q)\right] - \mathbb{E}_{x_C \sim Q_{X_C}} \left[\log(\psi_C^P/\psi_C^Q)\right] \leq 2\varepsilon_C(r) + r d_{\mathcal{F}_C}(P_{X_C}, Q_{X_C}) \qquad \forall r \geq 0, \forall C \in \mathcal{C}$$

Because of the condition (3): each $\log(\psi_C^P/\psi_C^Q)$ is bounded and Lipschitz continuous, there exists a constant $\eta_C > 0$, such that $\left|\log(\psi_C^P/\psi_C^Q)\right| < \eta_C$, and for any $x, y \in \Omega_C$ (which is the space of variables X_C), it holds that $\left|\log(\psi_C^P(x)/\psi_C^Q(x)) - \log(\psi_C^P(y)/\psi_C^Q(y))\right| \leq \frac{\eta_C}{\text{diam}(\Omega_C)}\|x-y\|$.

Again, by Theorem 28, we get an efficient upper-bound on $\varepsilon_C(r)$,

$$\varepsilon_C(r) \leq C(D_C)\eta_C \left(\frac{r}{\eta_C}\right)^{-\frac{2}{D_C+1}} \log\left(\frac{r}{\eta_C}\right) \qquad \forall r \geq R(D_C) > \mathrm{e}^{\frac{D_C+1}{2}}\eta_C, \forall C \in \mathcal{C}$$

where $C(D_C)$ and $R(D_C)$ are constants that only depend on the dimensionality, D_C, of variables X_C. More specifically, $D_C = |C|d \leq c_{\max}d$, where $|C|$ is the size of clique C, d is the dimensionality of each variable of the MRFs, and $c_{\max} = \max\{|C|\big|C \in \mathcal{C}\}$ is the maximum size of the cliques in G.

Because $C(D_C)$ and $R(D_C)$ are increasing functions of the dimensionality D_C, and for $r \geq R(D_C) > \mathrm{e}^{\frac{D_C+1}{2}}\eta_C$, $\eta_C\left(r/\eta_C\right)^{-\frac{2}{D_C+1}}\log\left(r/\eta_C\right)$ is an increasing function of η_C, summing up the inequalities for all $C \in \mathcal{C}$ gives,

$$\sum_{C\in\mathcal{C}} \varepsilon_C(r) \leq |\mathcal{C}|C(D_{\max})\eta_{\max}\left(\frac{r}{\eta_{\max}}\right)^{-\frac{2}{D_{\max}+1}} \log\left(\frac{r}{\eta_{\max}}\right) \qquad \forall r \geq R(D_{\max})$$

where $|\mathcal{C}|$ is the number of maximal cliques in G, $D_{\max} = \max\{D_C|C \in \mathcal{C}\} = c_{\max}d$, and $\eta_{\max} = \max\{\eta_C|C \in \mathcal{C}\}$.

Now, we sum up the inequalities $\mathbb{E}_{x_C\sim P_{X_C}}\left[\log(\psi_C^P/\psi_C^Q)\right] - \mathbb{E}_{x_C\sim Q_{X_C}}\left[\log(\psi_C^P/\psi_C^Q)\right] \leq 2\varepsilon_C(r) + r d_{\mathcal{F}_C}(P_{X_C}, Q_{X_C})$ for $r \geq R(D_{\max})$ for all $C \in \mathcal{C}$. Because of the decomposed form of the Symmetric KL divergence on MRFs P, Q, we get,

$$\text{SKL}(P,Q) - 2\sum_{C\in\mathcal{C}}\varepsilon_C(r) \leq r\sum_{C\in\mathcal{C}} d_{\mathcal{F}_C}(P_{X_C}, Q_{X_C}) \qquad \forall r \geq R(D_{\max})$$

That is, the neural distances defined by $\{\mathcal{F}_C|C \in \mathcal{C}\}$ satisfy r-linear subadditivity for,

$$r \geq R(D_{\max})$$

with error,

$$\epsilon = 2\sum_{C\in\mathcal{C}}\varepsilon_C(r) = \mathcal{O}\left(|\mathcal{C}|r^{-\frac{2}{D_{\max}+1}}\log r\right)$$

with respect to the Symmetric KL divergence on MRFs.

Note that r and ϵ are constants independent of the MRFs P, Q and the sets of discriminator classes $\{\mathcal{F}_C|C \in \mathcal{C}\}$. $|\mathcal{C}|$ is the number of maximal cliques in G and $D_{\max} = c_{\max}d$ where $c_{\max} = \max\{|C|\big|C \in \mathcal{C}\}$ is the maximum size of the cliques in G and d is the dimensionality of each variable of the MRFs. □

B f-Divergences and Inequalities

For two probability distributions P and Q on the same sample space Ω, the f-divergence of P from Q, denoted $D_f(P,Q)$, is defined as,

$$D_f(P,Q) \coloneqq \int_\Omega f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q$$

If densities exist, $D_f(P,Q) = \int_\Omega f\left(\frac{P(x)}{Q(x)}\right)Q(x)\mathrm{d}x$. In this definition, the function $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex, lower-semi-continuous function satisfying $f(1) = 0$. We can define $f(0) = \lim_{t\downarrow 0} f(t) \in \mathbb{R}\cup\{\infty\}$. Every convex, lower semi-continuous function f has a convex conjugate function f^*, defined as $f^* = \sup_{u\in\text{dom}_f}\{ut - f(u)\}$.

B.1 Common f-Divergences

All commonly-used f-divergences are listed in Table 2.

Name	Notation	Generator $f(t)$
Kullback–Leibler	KL	$t\log(t)$
Reverse KL	RKL	$-\log(t)$
Symmetric KL	SKL	$(t-1)\log(t)$
Jensen-Shannon	JS	$\frac{t}{2}\log\frac{2t}{t+1}+\frac{1}{2}\log\frac{2}{t+1}$
Squared Hellinger	H^2	$\frac{1}{2}\left(\sqrt{t}-1\right)^2$
Total Variation	TV	$\frac{1}{2}\lvert t-1\rvert$
Pearson χ^2	χ^2	$(t-1)^2$
Reverse Pearson χ^2	$R\chi^2$	$\frac{1}{t}-t$
α-Divergence	$\mathcal{H}_\alpha$	$\begin{cases}\frac{t^\alpha-1}{\alpha(\alpha-1)} & \alpha\neq 0,1\\ t\ln t & \alpha=1\\ -\ln t & \alpha=0\end{cases}$

Table 2: List of common f-divergences with generator functions.

We always adopt the most widely-accepted definitions. Note the $\frac{1}{2}$ coefficients in the definitions of squared Hellinger distance and Total Variation distance, in the spirit of normalizing their ranges to $[0,1]$.

The α-divergences $\mathcal{H}_\alpha$ ($\alpha\in\mathbb{R}$), popularized by (Liese and Vajda, 2006), generalize many f-divergences including KL divergence, reverse KL divergence, χ^2 divergence, reverse χ^2 divergence, and Hellinger distances. More specifically, they satisfy the following relations: $\mathcal{H}_1=\mathrm{KL}$, $\mathcal{H}_0=R\mathrm{KL}$, $\mathcal{H}_2=\frac{1}{2}\chi^2$, $\mathcal{H}_{-1}=\frac{1}{2}R_{\chi^2}$, and $\mathcal{H}_{\frac{1}{2}}=4\mathrm{H}^2$.

B.2 Inequalities between f-Divergences

First, we show a general approach to obtain inequalities between f-divergences. Then, we prove the inequalities between squared Hellinger distance and Jensen-Shannon divergence. We also list the well-known Pinsker's inequality for completeness.

Lemma 15. *Consider two f-divergences D_{f_1} and D_{f_2} with generator functions $f_1(\cdot)$ and $f_2(\cdot)$. If there exist two positive constants $0<A<B$, such that for any $t\in[0,\infty)$, it holds that,*

$$Af_2(t)\le f_1(t)\le Bf_2(t)$$

Then, for any two densities P and Q (such that $P\ll Q$), we have,

$$AD_{f_2}(P,Q)\le D_{f_1}(P,Q)\le BD_{f_2}(P,Q)$$

Proof. Note that we extend the domain of f_1 and f_2 by defining $f_1(0)=\lim_{t\downarrow 0}f_1(t)$ (and similar for f_2). We require $P\ll Q$ so that f-divergences are well-defined. In this sense, for any $x\in\Omega$, $P(x)/Q(x)\in[0,\infty)$ is defined, and we have $Af_2(P(x)/Q(x))\le f_1(P(x)/Q(x))\le Bf_2(P(x)/Q(x))$. Multiply non-negative $Q(x)$ and integrate over Ω. We obtain the desired inequality: $AD_{f_2}(P,Q)\le D_{f_1}(P,Q)\le BD_{f_2}(P,Q)$. □

Theorem 16 (Theorem 11 of (Sason and Verdu, 2016))**.** *For any two densities P and Q, (assume natural logarithm is used in the definition of Jensen-Shannon divergence), we have*

$$(\ln 2)\mathrm{H}^2(P,Q)\le \mathrm{JS}(P,Q)\le \mathrm{H}^2(P,Q)$$

Proof. Given Lemma 15, we only need to prove that for any $t\in[0,\infty)$, the following inequality holds,

$$(\ln 2)f_{\mathrm{H}^2}(t)\le f_{\mathrm{JS}}(t)\le f_{\mathrm{H}^2}(t)$$

where the definitions of f_{H^2} and f_{JS} can be found in Table 2.

Note that when $t = 1$, all terms are 0 and the inequalities hold trivially. For $t \neq 1$, as $f_{\mathrm{H}^2}(t) > 0$, we define,

$$\xi(t) = \frac{f_{\mathrm{JS}}(t)}{f_{\mathrm{H}^2}(t)} = \frac{t \ln \frac{2t}{t+1} + \ln \frac{2}{t+1}}{\left(\sqrt{t}-1\right)^2}$$

$\xi(t)$ is defined on $[0,1) \cup (1,\infty)$, We want to prove that $\ln 2 \leq \xi(t) \leq 1$ always holds. Its derivative is,

$$\xi'(t) = \frac{\sqrt{t} \ln \frac{2t}{t+1} + \ln \frac{2}{t+1}}{\sqrt{t}\left(1-\sqrt{t}\right)^3}$$

Denote the numerator above by $\xi_{(1)}(t)$. Its derivative is,

$$\xi'_{(1)}(t) = \frac{(t+1)\ln \frac{2t}{t+1} + 2\left(1-\sqrt{t}\right)}{2\sqrt{t}(t+1)}$$

Again, denote the numerator above by $\xi_{(2)}(t)$. Its derivative is,

$$\xi'_{(2)}(t) = \frac{1}{t} - \frac{1}{\sqrt{t}} + \ln \frac{2t}{t+1}$$

Using the well-known logarithm inequality: for any $x > 0$, $\ln x > 1 - \frac{1}{x}$, we have,

$$\xi'_{(2)}(t) \geq \frac{1}{t} - \frac{1}{\sqrt{t}} + 1 - \frac{t+1}{2t} = \frac{\left(\sqrt{t}-1\right)^2}{2t} \geq 0$$

Also, since $\xi_{(2)}(1) = 0$, and the denominator of $\xi'_{(1)}(t)$ is always positive, hence,

$$\xi'_{(1)}(t) \begin{cases} < 0 & t \in [0,1) \\ > 0 & t \in (1,\infty) \end{cases}$$

Because $\xi_{(1)}(1) = 0$, this implies $\xi_{(1)}(t) \geq 0$. Thus,

$$\xi'(t) \begin{cases} > 0 & t \in [0,1) \\ < 0 & t \in (1,\infty) \end{cases}$$

That is, $\xi(t)$ is strictly increasing on $[0,1)$, and is strictly decreasing on $(1,\infty)$. To determine its range, we only need to compute these limits: $\lim_{t\downarrow 0} \xi(t)$, $\lim_{t\uparrow 1} \xi(t)$, $\lim_{t\downarrow 1} \xi(t)$, and $\lim_{t\to+\infty} \xi(t)$:

$$\begin{aligned}
&\lim_{t\downarrow 0} \xi(t) = \ln 2 \\
&\lim_{t\uparrow 1} \xi(t) = \lim_{t\downarrow 1} \xi(t) = \lim_{t\to 1} \frac{\sqrt{t} \ln \frac{2t}{t+1}}{\sqrt{t}-1} = \lim_{t\to 1} \frac{2\sqrt{t^3}}{t+1} = 1 \\
&\lim_{t\to+\infty} \xi(t) = \lim_{t\to+\infty} \frac{t \ln \frac{2t}{t+1}}{\left(\sqrt{t}-1\right)^2} = \lim_{t\to+\infty} \frac{\ln \frac{2t}{t+1} + \frac{1}{t+1}}{\frac{\sqrt{t}-1}{\sqrt{t}}} = \ln 2
\end{aligned}$$

Together with the monotonic properties of $\xi(t)$, we know

$$\ln 2 \leq \xi(t) \leq 1$$

□

Theorem 17 (Pinsker's Inequality, Eq. (1) of (Sason and Verdu, 2016))**.** *For any two densities P and Q, we have,*

$$\mathrm{TV}(P,Q) \leq \sqrt{\frac{1}{2}\mathrm{KL}(P,Q)}$$

It is a well-known result. See for example Theorem 2.16 of (Massart, 2007) for a proof.

C Wasserstein Distances: Formulas and Inequalities

Suppose Ω is a metric space with distance $d(\cdot,\cdot)$. The p-Wasserstein distance W_p is defined as,

$$\mathrm{W}_p(P,Q) \coloneqq \left(\inf_{\gamma\in\Gamma(P,Q)}\int_{\Omega\times\Omega} d(x,y)^p \mathrm{d}\gamma(x,y)\right)^{\frac{1}{p}}$$

where $\gamma \in \Gamma(P,Q)$ denotes the set of all possible couplings of P and Q.

C.1 Formulas for Wasserstein Distances

We list the algorithm and the formula to calculate the Wasserstein distance when space Ω is finite or the distributions P and Q are Gaussians.

Theorem 18. *For any two discrete distributions P, Q on a finite space $\Omega = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, the p-Wasserstein distance W_p can be computed by the following linear program:*

$$\begin{aligned}
\mathrm{W}_p(P,Q)^p = \quad & \min \ \textstyle\sum_{i=1}^n \sum_{j=1}^n d^p(\mathbf{x}_i,\mathbf{x}_j)\pi_{ij} & \\
& \textit{subject to } \textstyle\sum_{j=1}^n \pi_{ij} = P(\mathbf{x}_i) & i = 1,\cdots,n \\
& \textstyle\sum_{i=1}^n \pi_{ij} = Q(\mathbf{x}_j) & j = 1,\cdots,n \\
& \textit{and } \pi_{ij} > 0 & i = 1,\cdots,n \textit{ and } j = 1,\cdots,n
\end{aligned}$$

Useful discussions can be found in (Oberman and Ruan, 2015).

Theorem 19. *For any two non-degenerate Gaussians $P = \mathcal{N}(m_1, C_1)$ and $Q = \mathcal{N}(m_2, C_2)$ on $\mathbb{R}^n$, with respective means $m_1, m_2 \in \mathbb{R}^n$ and (symmetric positive semi-definite) covariance matrices $C_1, C_2 \in \mathbb{R}^{n\times n}$. The square of 2-Wasserstein distance W_2 between P, Q is,*

$$\mathrm{W}_2(P,Q)^2 = \|m_1 - m_2\|_2^2 + \mathrm{Tr}\left(C_1 + C_2 - 2\left(C_2^{1/2} C_1 C_2^{1/2}\right)^{1/2}\right)$$

where $\|\cdot\|_2$ is the Euclidean norm.

See (Olkin and Pukelsheim, 1982) for a proof.

C.2 Inequalities between p-Wasserstein Distance and Total Variation Distance

Both Wasserstein distances and Total Variation distance can be regarded as optimal transportation costs. More specifically,

$$\begin{aligned}
\mathrm{W}_p(P,Q) &\coloneqq \left(\inf_{\gamma\in\Gamma(P,Q)}\int_{\Omega\times\Omega} d(x,y)^p \mathrm{d}\gamma(x,y)\right)^{\frac{1}{p}} \\
\mathrm{TV}(P,Q) &\coloneqq \inf_{\gamma\in\Gamma(P,Q)}\int_{\Omega\times\Omega} \mathbf{1}_{x\neq y}\mathrm{d}\gamma(x,y)
\end{aligned}$$

where $\Gamma(P,Q)$ denotes the set of all measures on $\Omega\times\Omega$ with marginals P and Q on variable x and y respectively, (also called the set of all possible couplings of P and Q). Bounding the distance $d(x,y)$ directly leads to inequalities between p-Wasserstein distance and Total Variation distance.

Theorem 20. *For any two distributions P and Q on a space Ω, if Ω is bounded with diameter $\mathrm{diam}(\Omega) = \max\{d(x,y)|x,y\in\Omega\}$, then,*

$$\mathrm{W}_p(P,Q)^p \le \mathrm{diam}(\Omega)^p \mathrm{TV}(P,Q)$$

Moreover, if Ω is finite, let $d_{\min} = \min_{x\neq y} d(x,y)$ be the minimum mutual distance between pairs of distinct points in Ω, then,

$$\mathrm{W}_p(P,Q)^p \ge d_{\min}^p \mathrm{TV}(P,Q)$$

Proof. This theorem is a generalization of Theorem 4 of (Gibbs and Su, 2002). Since $d(\cdot,\cdot)$ is a metric of space Ω, $d(x,y)=0$ if and only if $x=y$. Thus $d(x,y)\equiv d(x,y)\mathbf{1}_{x\neq y}$, and we have,

$$\mathrm{W}_p(P,Q)^p=\inf_{\gamma\in\Gamma(P,Q)}\int_{\Omega\times\Omega}d(x,y)^p\mathbf{1}_{x\neq y}\mathrm{d}\gamma(x,y)$$

If Ω is bounded, then for any x,y in Ω, it holds that $d(x,y)\leq\mathrm{diam}(\Omega)$. Applying this inequality to the formula above leads to $\mathrm{W}_p(P,Q)^p\leq\mathrm{diam}(\Omega)^p\mathrm{TV}(P,Q)$.

Similarly, if Ω is finite, then for any distinct $x\neq y$ in Ω, it holds that $d(x,y)\geq d_{\min}$. We can generalize it to: for any x,y in Ω, we have $d(x,y)\mathbf{1}_{x\neq y}\geq d_{\min}\mathbf{1}_{x\neq y}$. Applying this inequality to the formula above leads to $\mathrm{W}_p(P,Q)^p\geq d_{\min}^p\mathrm{TV}(P,Q)$. □

D "Breadth First Search"-Subadditivity on MRFs

Most of our theoretical results in this paper are for the subadditivity of divergences on Bayes-nets. However, following the same recursive approach as in the proof of Theorem 1, we can develop a different version of subadditivity on MRFs that depends on a Breadth-First Search (BFS) ordering $(1,\ldots,n)$ on the undirected graph G, which we call *BFS-Subadditivity on MRFs* (to distinguish it from the version we defined in Definition 1).

For BFS-Subadditivity on MRFs, each local neighborhood is the union of a node $k\in\{1,\ldots,n\}$ and a subset $\Sigma_k=\cup_{i=1}^k N_i\setminus\{1,\ldots,k\}$, where N_i is the set of nodes adjacent to node i, and Σ_k is a separating subset between $\{1,\ldots,k\}$ and $\{k+1,\ldots,n\}\setminus\Sigma_k$. The construction of BFS-Subadditivity of a divergence δ requires exactly the same two properties as in Theorem 1, i.e. δ is subadditive with respect to product measures and length-3 Markov Chains. In this sense, it is not hard to verify that all the divergences we prove to satisfy subadditivity on Bayes-net in the paper, satisfy BFS-Subadditivity on MRFs as well.

D.1 Constructing Subadditivity Upper-Bound on Generic Graphical Models

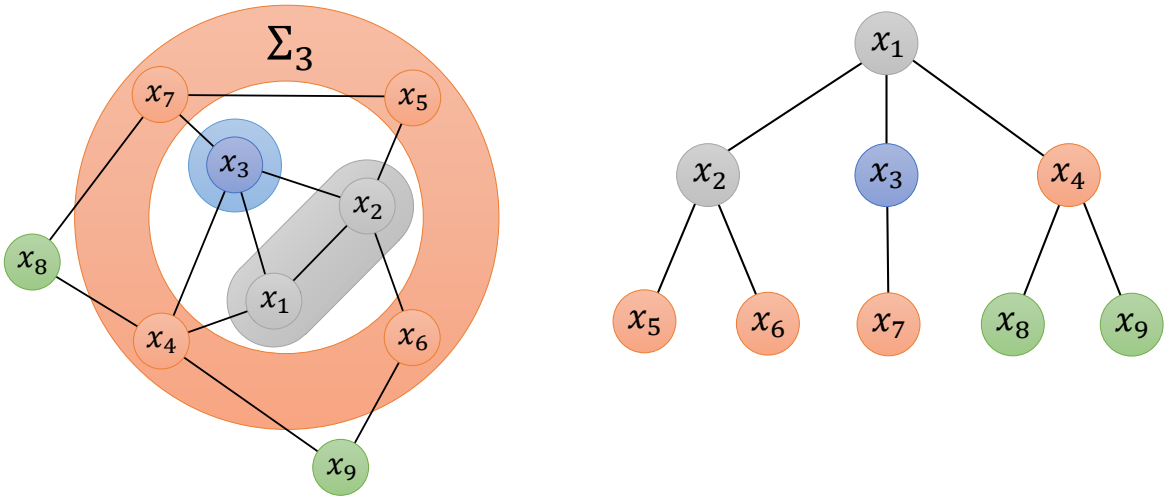
From the proof of Theorem 1 in Appendix A.1, we obtain the subadditivity upper-bound on Bayes-nets by repeatedly applying the subadditivity inequality on Markov Chain $X\to Y\to Z$. Moreover, we allow $X=\varnothing$ or $Y=\varnothing$ (i.e., X and Z are conditional independent), as addressed by the second and third cases in the proof. In general, for a generic probability graphical model with an underlying graph G (there may be directed and undirected edges in G), let P and Q be two distributions characterized by such graphical model. If δ satisfy subadditivity on Markov Chain $X\to Y\to Z$ with conditionally independent variables X and Y, we can obtain a subadditivity upper-bound on $\delta(P,Q)$ by the following procedure:

1. Choose an ordering of nodes $(1,\cdots,n)$. The ordering is valid if the induction can be proceeded form start to end.
2. For node $k=1,\cdots,n-1$, let Σ_k be the smallest set of nodes such that $\Sigma_k\subsetneqq\{k+1,\cdots,n\}$ and X_k is conditionally independent of $\cup_{i=k+1}^n X_i$ given X_{Σ_k}, which can be written as $X_k\perp\!\!\!\perp\cup_{i=k+1}^n X_i\mid X_{\Sigma_k}$. If we cannot find such Σ_k, the ordering $(1,\cdots,n)$ is invalid and the induction cannot be proceeded. Applying the subadditivity of δ on the Markov Chain of super-nodes $X_{\{k+1,\cdots,n\}\setminus\Sigma_k}\to X_{\Sigma_k}\to X_k$ gives an inequality $\delta(P_{\cup_{i=k}^n X_i},Q_{\cup_{i=k}^n X_i})\leq\delta(P_{\cup_{i=k+1}^n X_i},Q_{\cup_{i=k+1}^n X_i})+\delta(P_{X_{\Sigma_k}\cup X_k},Q_{X_{\Sigma_k}\cup X_k})$.
3. By combining all the inequalities obtained, we get a subadditivity upper-bound $\sum_{i=1}^n\delta(P_{X_{\Sigma_i}\cup X_i},Q_{X_{\Sigma_i}\cup X_i})\geq\delta(P,Q)$.

This process is identical to the proof of Theorem 1 for Bayes-nets, except that (1) we have to manually choose a valid ordering of nodes, and (2) the set of parents Π_k is replaced by the smallest set of nodes $X_{\Sigma_k}\subsetneqq\{k+1,\cdots,n\}$ such that $X_k\perp\!\!\!\perp\cup_{i=k+1}^n X_i\mid X_{\Sigma_k}$, which depends on the ordering we choose. For Bayes-nets, the ordering we use is the reversed topological ordering, and for each k, we have $\Sigma_k=\Pi_k$.

D.2 BFS-Subadditivity on MRFs and its Application to Sequences of Words

Let us now illustrate this process on MRFs, whose underlying probability structure is described by undirected graphs. An enumeration of the nodes of a graph G is said to be a BFS ordering if it is a possible output of the



(a) A MRF with 9 variables.

(b) A BFS tree of the MRF.

Figure 7: A local neighborhood according to BFS-subadditivity, $\{3\} \cup \Sigma_3$, of a MRF with 9 variables, if the BFS ordering $(1, \cdots, 9)$ is used. Where (a) is the MRF and (b) is the corresponding BFS tree. It is a snapshot of the induction process at $k = 3$. Where the gray nodes have been processed, the blue node is the current focus, the orange nodes represent the separating subset Σ_3, which is the smallest subset such that $X_3 \perp\!\!\!\perp \cup_{i=4}^{9} X_i \mid X_{\Sigma_3}$, and the green nodes are the rest.

BFS algorithm on this graph. If we use a BFS ordering $(1, \cdots, n)$, then it is not hard to prove that for any $k \in \{1, \cdots, n\}$, we have $\Sigma_k = \cup_{i=1}^{k} N_i \setminus \{1, \cdots, k\}$, where N_i is the set of nodes adjacent to node i (i.e. the set of nearest neighbors). As shown in Fig. 7, if we choose a BFS ordering, Σ_k is actually the smallest set of nodes that surround the current and processed nodes $\{1, \cdots, k\}$. Σ_k is called a separating subset between $\{1, \cdots, k\}$ and $\{k+1, \cdots, n\} \setminus \Sigma_k$, as every path from a node in $\{1, \cdots, k\}$ to a node in $\{k+1, \cdots, n\} \setminus \Sigma_k$ passes through Σ_k. By the global Markov property of MRFs, we indeed have $X_k \perp\!\!\!\perp \cup_{i=k+1}^{n} X_i \mid X_{\Sigma_k}$.

As an example, we may consider a particular type of MRFs: sequences with local dependencies but no natural directionality, e.g., sequences of words. If we assume the distribution of a word depends on both the pre- and post- context, and consider up to $(2p+1)$-grams (i.e. consider the distribution of up to $2p+1$ consecutive words), the corresponding MRF is an undirected graph G, where each node i is connected to its p previous nodes and p subsequent nodes. Let $(1, \cdots, n)$ be the natural ordering of these n words. Clearly, both $(1, \cdots, n)$ and $(n, \cdots, 1)$ are valid BFS orderings. Following the method above, and if we truncate the induction at step $k = n - p$ (see Appendix I.1 for details), these two orderings result in an identical subadditivity upper bound $\sum_{k=1}^{n-p} \delta(P_{\cup_{i=k}^{k+p} X_i}, Q_{\cup_{i=k}^{k+p} X_i})$. Each local neighborhoods contains $p+1$ consecutive words. Equipped with this theoretical-justified subadditivity upper-bound, we can use a set of local discriminators in GANs, each on a subsequence of $p+1$ consecutive words. This is how we apply local discriminators to sequences of words.

E A Counter-Example for the Subadditivity of 2-Wasserstein Distance

In this section, we report a counter-example for the subadditivity of 2-Wasserstein distance using Gaussian distributions in $\mathbb{R}^3$. Note that as we shown in Corollary 7, in a finite space Ω, 2-Wasserstein distance satisfies $(\sqrt{2}\text{diam}(\Omega)/d_{\min})$-linear subadditivity on Bayes-nets, where $\text{diam}(\Omega)$ is the diameter and $d_{\min}$ is the smallest distance between pairs of distinct points in Ω. However the counter-example in this section shows that, in an arbitrary metric space Ω, 2-Wasserstein distance does not satisfy subadditivity (with linear coefficient $\alpha = 1$) on Bayes-nets and MRFs.

Consider an non-degenerate 3-dimensional Gaussian with zero mean $P = \mathcal{N}(\mathbf{0}, C)$ on variables (X, Y, Z) ($C \in \mathbb{R}^{3\times 3}$ is the covariance matrix), which are also Bayes-nets with structure $X \to Y \to Z$. From the definition of Bayes-nets: each variable is conditionally independent of its non-descendants given its parents, we know P is a Bayes-net if and only if for any $x, y, z \in \mathbb{R}$, it holds that $P_{Z|X,Y}(z|x, y) = P_{Z|Y}(z|y)$. Let C_{ij} denote the element of C at the

i-th row and j-th column. It is not hard to compute that,

$$P_{Z|Y}(z|y) = \mathcal{N}\left(\frac{C_{32}}{C_{22}}y, C_{33} - \frac{C_{32}C_{23}}{C_{22}}\right)$$
$$P_{Z|X,Y}(z|x,y) = \mathcal{N}\left(\begin{bmatrix}C_{31} & C_{32}\end{bmatrix}\begin{bmatrix}C_{11} & C_{12}\\ C_{21} & C_{12}\end{bmatrix}^{-1}\begin{bmatrix}x\\ y\end{bmatrix}, C_{33} - \begin{bmatrix}C_{31} & C_{32}\end{bmatrix}\begin{bmatrix}C_{11} & C_{12}\\ C_{21} & C_{12}\end{bmatrix}^{-1}\begin{bmatrix}C_{13}\\ C_{23}\end{bmatrix}\right)$$

Matching the means and variances of these two 1-dimensional Gaussians of z, we know that the two conditional distributions coincide, and therefore P is a Bayes-net, if and only if $C_{32}C_{21} = C_{31}C_{22}$, i.e. the 2×2 upper-right (or equivalently, the lower-left) sub-matrix of C has zero determinant. This condition can also be written as $\mathrm{Var}[Y]\mathrm{Cov}[X,Z] = \mathrm{Cov}[X,Y]\mathrm{Cov}[Y,Z]$.

It is clear that this condition on the covariance matrix C is symmetric under switching variables X and Z. This means $P_{X|Y,Z}(x|y,z) = P_{X|Y}(x|y)$ holds simultaneously, and the most appropriate graphical model to describe P is the MRF. However, as long as the Markov property $P_{Z|X,Y}(z|x,y) = P_{Z|Y}(z|y)$ holds, P is a valid Bayes-net. These 3-dimensional Gaussians are special, as they satisfy the definitions of both Bayes-nets and MRFs.

Based on the discussions above, we construct two 3-dimensional Gaussians P and Q that are valid Bayes-nets and MRFs, as follows.

Counter-Example 1. *Consider two 3-dimensional Gaussians $P^x = \mathcal{N}(\mathbf{0}, C_1)$ and $Q^{xy} = \mathcal{N}(\mathbf{0}, C_2)$ in $\Omega = \mathbb{R}^3$ parametrized by $(x,y) \in \{(x,y)\in\mathbb{R}^2 | 0 < x, y < 1\}$, where,*

$$C_1 = \begin{bmatrix}1 & x & 0\\ x & 1 & 0\\ 0 & 0 & 1\end{bmatrix} \qquad C_2 = \begin{bmatrix}1 & x & xy\\ x & 1 & y\\ xy & y & 1\end{bmatrix}$$

and $\mathbf{0}\in\mathbb{R}^3$ is the zero vector. The two distributions are valid Bayes-nets and MRFs with structure $X\to Y\to Z$ (when considered as Bayes-nets) or X–Y–Z (when considered as MRFs), since the 2×2 upper-right (or lower-left) sub-matrices of C_1 and C_2 has zero determinants. The 2-Wasserstein distance between them, $W_2(P^x, Q^{xy})$, depends on parameters (x,y). For any $(x,y)\in\{(x,y)\in\mathbb{R}^2 | 0<x,y<1\}$, it holds that $\mathrm{W}_2(P^x_{XYZ}, Q^{xy}_{XYZ}) > \mathrm{W}_2(P^x_{XY}, Q^{xy}_{XY}) + \mathrm{W}_2(P^x_{YZ}, Q^{xy}_{YZ})$, which violets the subadditivity inequality (with linear coefficient $\alpha = 1$) of 2-Wasserstein distance on Bayes-nets and MRFs.

Counter-Example 1 can be numerically verified, as the 2-Wasserstein distance between Gaussians can be exactly computed using the formula in Theorem 19 in Appendix C.1. As shown in Fig. 8, the subadditivity gap $\Delta = \mathrm{W}_2(P^x_{XY}, Q^{xy}_{XY}) + \mathrm{W}_2(P^x_{YZ}, Q^{xy}_{YZ}) - \mathrm{W}_2(P^x_{XYZ}, Q^{xy}_{XYZ})$ is negative for any $(x,y)\in\{(x,y)\in\mathbb{R}^2 | 0<x,y<1\}$, thus the subadditivity inequality is violated.

This straightforward but fundamental counter-example shows that Wasserstein's subadditivity does not hold even if all distributions are Gaussians. For many common divergences including Jensen-Shannon divergence, Total Variation distance, and p-Wasserstein distance, the best we can prove is linear subadditivity.

F Local Subadditivity

In this section, we consider the case when two distributions P, Q are close to each other. This can happen after some training steps in a GAN. We consider two notions of "closeness" for distributions.

Definition 3 (One- and Two-Sided ϵ-Close Distributions)**.** *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 $P \ll\gg Q$.*

F.1 Local Subadditivity under Perturbation

For the sake of theoretical simplicity, we consider the limit $\epsilon\to 0$ for two-sided ϵ-close distributions. We call Q a perturbation of P (Makur, 2015).

Theorem 21. *For two-sided ϵ-close distributions P, Q with $\epsilon\to 0$ on a common Bayes-net G, any f-divergence*

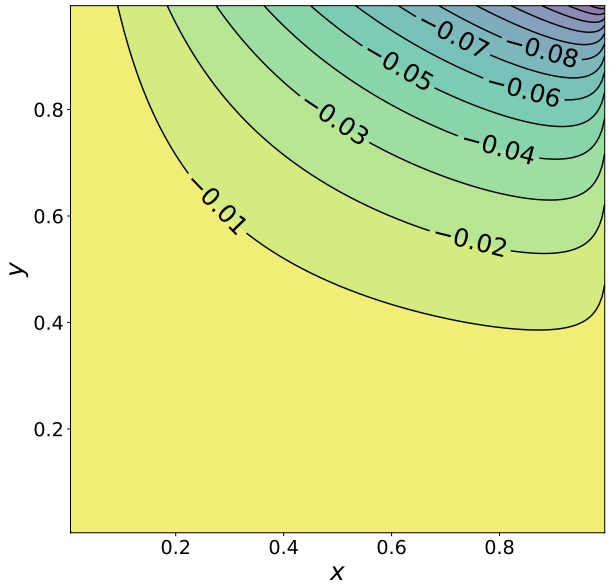


Figure 8: Contour maps showing the counter-example for the subadditivity of 2-Wasserstein distance. The two distributions P^x, Q^{xy} are 3-dimensional Gaussians $P^x = \mathcal{N}(\mathbf{0}, C_1)$, $Q^{xy} = \mathcal{N}(\mathbf{0}, C_2)$ which are valid Bayes-nets and MRFs. The contours and colors indicate the subadditivity gap $\Delta = \mathrm{W}_2(P^x_{XY}, Q^{xy}_{XY}) + \mathrm{W}_2(P^x_{YZ}, Q^{xy}_{YZ}) - \mathrm{W}_2(P^x_{XYZ}, Q^{xy}_{XYZ})$.

$D_f(P, Q)$ such that $f''(1) > 0$ has subadditivity up to $\mathcal{O}(\epsilon^3)$. That is,

$$D_f(P, Q) \leq \sum_{i=1}^{n} D_f(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}}) + \mathcal{O}(\epsilon^3)$$

Moreover, the subadditivity gap is proportional to the sum of χ^2 divergences between marginals on the set of parents of each node, up to $\mathcal{O}(\epsilon^3)$. That is,

$$\Delta = \sum_{i=1}^{n} D_f(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}}) - D_f(P, Q) = \frac{f''(1)}{2} \sum_{i=1}^{n} \chi^2(P_{\Pi_i}, Q_{\Pi_i}) + \mathcal{O}(\epsilon^3)$$

Theorem 21 indicates that when P, Q are very close, the focus of the set of local discriminators falls on the differences between the marginals on the set of parents. We make use of the Taylor expansion of $f(\cdot)$ in the proof. To prove Theorem 21, we first prove the following lemma describing the approximation behavior of nearly all f-divergences when P, Q are perturbations with respect to each other.

Lemma 22. *For two-sided ϵ-close distributions P, Q with $\epsilon \to 0$, any f-divergence $D_f(P, Q)$ with $f(t)$ twice differentiable at $t = 1$ and $f''(1) > 0$, is proportional to $\chi^2(P, Q)$ up to $\mathcal{O}(\epsilon^3)$, i.e.*

$$D_f(P, Q) = \frac{f''(1)}{2} \chi^2(P, Q) + \mathcal{O}(\epsilon^3)$$

And χ^2 is now symmetric up to $\mathcal{O}(\epsilon^3)$, i.e. $\chi^2(P, Q) = \chi^2(Q, P) + \mathcal{O}(\epsilon^3)$.

Proof. Since $f(t)$ twice differentiable at $t = 1$, and $P(x)/Q(x) \in (1 - \epsilon, 1 + \epsilon)$ with $0 < \epsilon \ll 1$, by Taylor's theorem we get,

$$f(\frac{P}{Q}) = f'(1) \left(\frac{P}{Q} - 1 \right) + \frac{1}{2} f''(1) \left(\frac{P}{Q} - 1 \right)^2 + \mathcal{O}(\epsilon^3)$$

Multiply by Q and integrate over $\Omega \in \mathbb{R}^{nd}$ gives,

$$\begin{aligned} D_f(P,Q) &= \frac{f''(1)}{2}\int Q\left(\frac{P}{Q}-1\right)^2 \mathrm{d}x + \mathcal{O}(\epsilon^3) \\ &= \frac{f''(1)}{2}\chi^2(P,Q) + \mathcal{O}(\epsilon^3) \end{aligned}$$

Where the first order term vanishes because $\int P\mathrm{d}x = \int Q\mathrm{d}x = 1$. This equation implies that all f-divergences such that $f''(1) > 0$ behave similarly when the two distributions P and Q are sufficiently close.

Meanwhile, because $P/Q = 1 + \mathcal{O}(\epsilon)$, we have,

$$\begin{aligned} \chi^2(P,Q) &= \int \frac{(P-Q)^2}{P}\frac{P}{Q}\mathrm{d}x \\ &= \int \frac{(P-Q)^2}{P}(1+\mathcal{O}(\epsilon))\mathrm{d}x \\ &= \chi^2(Q,P) + \mathcal{O}(\epsilon^3) \end{aligned}$$

Thus we can exchange P and Q freely in any $\mathcal{O}(\epsilon^2)$ terms (e.g. $(P-Q)^2/Q$), while preserving the equality up to $\mathcal{O}(\epsilon^3)$. □

Based on Lemma 22, Theorem 21 can be proved by comparing an f-divergence with the squared Hellinger distance.

Proof of Theorem 21: We first prove that the subadditivity inequality holds using Lemma 22. Define $R(x) = \frac{1}{2}\left(\sqrt{PQ} + \frac{P+Q}{2}\right)$ as the average of the geometric and arithmetic means of P and Q. Clearly for any $x \in \Omega$, it holds that $|R(x) - Q(x)| < |P(x) - Q(x)| < \epsilon$. Thus $R/Q = 1 + \mathcal{O}(\epsilon)$, and by Lemma 22, we have,

$$\begin{aligned} D_f(P,Q) &= \frac{f''(1)}{2}\chi^2(P,Q) + \mathcal{O}(\epsilon^3) \\ &= \frac{f''(1)}{2}\int \frac{(P-Q)^2}{R}\frac{R}{Q}\mathrm{d}x + \mathcal{O}(\epsilon^3) \\ &= \frac{f''(1)}{2}\int \frac{(P-Q)^2}{R}\mathrm{d}x + \mathcal{O}(\epsilon^3) \\ &= 2f''(1)\int \left(\sqrt{P}-\sqrt{Q}\right)\mathrm{d}x + \mathcal{O}(\epsilon^3) \\ &= 4f''(1)\mathrm{H}^2(P,Q) + \mathcal{O}(\epsilon^3) \end{aligned}$$

Since $f''(1) > 0$, we can re-write this equation as $\mathrm{H}^2(P,Q) = \frac{1}{4f''(1)}D_f(P,Q) + \mathcal{O}(\epsilon^3)$. Applying this formula to both sides of the subadditivity inequality of H^2 (Theorem 2): $\mathrm{H}^2(P,Q) \le \sum_{i=1}^n \mathrm{H}^2(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}})$, we conclude that the subadditivity inequality holds up to $\mathcal{O}(\epsilon^3)$:

$$D_f(P,Q) \le \sum_{i=1}^n D_f(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}}) + \mathcal{O}(\epsilon^3)$$

Then, we prove that the subadditivity gap $\Delta \coloneqq \sum_{i=1}^n D_f(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}}) - D_f(P,Q)$ is proportional to $\sum_{i=1}^n \chi^2(P_{\Pi_i}, Q_{\Pi_i})$ up to $\mathcal{O}(\epsilon^3)$ using a different approach. Let us start from the simple case when P, Q are Markov Chains with structure $X \to Y \to Z$. The Markov property $P_{Z|XY} = P_{Z|Y}$ holds (and the same for Q). Since the joint distributions P_{XYZ} and Q_{XYZ} are two-sided ϵ-close, so are the marginal and conditional distributions. We define the differences between the marginals and conditionals of P and Q as follows,

$$\begin{aligned} Q_{X|Y} &= P_{X|Y} + \epsilon J_{X|Y} \\ Q_Y &= P_Y + \epsilon J_Y \\ Q_{Z|Y} &= P_{Z|Y} + \epsilon J_{Z|Y} \end{aligned}$$

Clearly $\int J_{X|Y}\mathrm{d}x = \int J_Y \mathrm{d}y = \int J_{Z|Y}\mathrm{d}z = 0$. Using Lemma 22, we have,

$$\begin{aligned}
&\frac{2}{\epsilon^2 f''(1)} D_f(P_{XYZ}, Q_{XYZ}) + \mathcal{O}(\epsilon) \\
&= \frac{1}{\epsilon^2}\int \frac{(P_{XYZ} - Q_{XYZ})^2}{P_{XYZ}}\mathrm{d}x\mathrm{d}y\mathrm{d}z \\
&= \frac{1}{\epsilon^2}\int \frac{\left(P_{X|Y}P_Y P_{Z|Y} - Q_{X|Y}Q_Y Q_{Z|Y}\right)^2}{P_{X|Y}P_Y P_{Z|Y}}\mathrm{d}x\mathrm{d}y\mathrm{d}z \\
&= \int \Bigg(\frac{J_Y^2 P_{X|Y}P_{Z|Y}}{P_Y} + \frac{J_{X|Y}^2 P_Y P_{Z|Y}}{P_{X|Y}} + \frac{J_{Z|Y}^2 P_{X|Y}P_Y}{P_{Z|Y}} \\
&\qquad + 2J_{X|Y}J_Y P_{Z|Y} + 2J_Y J_{Z|Y} P_{X|Y} + 2J_{X|Y}J_{Z|Y}P_Y \Bigg)\mathrm{d}x\mathrm{d}y\mathrm{d}z \\
&= \int \frac{J_Y^2}{P_Y}\mathrm{d}y + \int \frac{J_{X|Y}^2 P_Y}{P_{X|Y}}\mathrm{d}x\mathrm{d}y + \int \frac{J_{Z|Y}^2 P_Y}{P_{Z|Y}}\mathrm{d}y\mathrm{d}z
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{2}{\epsilon^2 f''(1)} D_f(P_{XY}, Q_{XY}) + \mathcal{O}(\epsilon) &= \frac{1}{\epsilon^2}\int \frac{(P_{X|Y}P_Y - Q_{X|Y}Q_Y)^2}{P_{X|Y}P_Y}\mathrm{d}x\mathrm{d}y \\
&= \int \left(\frac{J_{X|Y}^2 P_Y}{P_{X|Y}} + \frac{J_Y^2 P_{X|Y}}{P_Y} + 2J_{X|Y}J_Y \right)\mathrm{d}x\mathrm{d}y \\
&= \int \frac{J_{X|Y}^2 P_Y}{P_{X|Y}}\mathrm{d}x\mathrm{d}y + \int \frac{J_Y^2}{P_y}\mathrm{d}y
\end{aligned}$$

And,

$$\frac{2}{\epsilon^2 f''(1)} D_f(P_{YZ}, Q_{YZ}) + \mathcal{O}(\epsilon) = \int \frac{J_{Z|Y}^2 P_Y}{P_{Z|Y}}\mathrm{d}y\mathrm{d}z + \int \frac{J_Y^2}{P_Y}\mathrm{d}y$$

Thus, the subadditivity gap on the Markov Chain $X \to Y \to Z$ is,

$$\begin{aligned}
\Delta_{\text{Markov Chain}} &= D_f(P_{XY}, Q_{XY}) + D_f(P_{YZ}, Q_{YZ}) - D_f(P_{XY}, Q_{XY}) \\
&= \frac{f''(1)}{2}\int \frac{J_Y^2}{P_Y}\mathrm{d}y + \mathcal{O}(\epsilon^3) \\
&= \frac{f''(1)}{2}\chi^2(P_Y, Q_Y) + \mathcal{O}(\epsilon^3)
\end{aligned}$$

Moreover, consider the special case when $Y = \varnothing$, thus P, Q are product measures on conditionally independent variables X and Z. Similarly, we have,

$$\frac{2}{\epsilon^2 f''(1)} D_f(P_{XZ}, Q_{XZ}) + \mathcal{O}(\epsilon) = \chi^2(P_X, Q_X) + \chi^2(P_Z, Q_Z)$$

Hence the subadditivity gap is,

$$\Delta_{\text{Product Measure}} = D_f(P_X, Q_X) + D_f(P_Z, Q_Z) - D_f(P_{XZ}, Q_{XZ}) = 0 + \mathcal{O}(\epsilon^3)$$

Now, for any pair of generic Bayes-nets P and Q, following the approach in the proof of Theorem 1 in Appendix A.1, we repeatedly apply the subadditivity inequality on Markov Chains of super-nodes $X_{\{1,\cdots,n-k-1\}\setminus\Pi_{n-k}} \to X_{\Pi_{n-k}} \to X_{n-k}$, for $k = 0, 1, \cdots, n-2$. Consider three cases:

1. $\Pi_{n-k} \neq \varnothing$ and $\Pi_{n-k} \subsetneqq \{1, \cdots, n-k-1\}$: In this case, the subadditivity gap is $\frac{f''(1)}{2}\chi^2(P_{\Pi_{n-k}}, Q_{\Pi_{n-k}}) + \mathcal{O}(\epsilon^3)$.

2. $\Pi_{n-k} = \{1, \cdots, n-k-1\}$: In this case, as discussed in Appendix A.1, we add a redundant term $\delta(P_{\cup_{i=1}^{n-k-1} X_i}, Q_{\cup_{i=1}^{n-k-1} X_i}) \equiv \delta(P_{\Pi_{n-k}}, Q_{\Pi_{n-k}})$ into the subadditivity upper-bound. Thus, by Lemma 22, the subadditivity gap is $\frac{f''(1)}{2}\chi^2(P_{\Pi_{n-k}}, Q_{\Pi_{n-k}}) + \mathcal{O}(\epsilon^3)$
3. $\Pi_{n-k} = \varnothing$: In this case, X_{n-k} is independent from $(X_1, \ldots, X_{n-k-1})$ in both Bayes-nets. Thus the subadditivity gap is 0.

For all the three cases, the subadditivity gap at an induction step k is $\frac{f''(1)}{2}\chi^2(P_{\Pi_{n-k}}, Q_{\Pi_{n-k}}) + \mathcal{O}(\epsilon^3)$ (note that $\chi^2(P_{\Pi_{n-k}}, Q_{\Pi_{n-k}}) = 0$ when $\Pi_{n-k} = \varnothing$). Along with the induction process for $k = 0, 1, \cdots, n-2$, the subadditivity gaps accumulate, and we finally get,

$$\Delta \coloneqq \sum_{i=1}^{n} D_f(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}}) - D_f(P, Q) = \frac{f''(1)}{2} \sum_{i=1}^{n} \chi^2(P_{\Pi_i}, Q_{\Pi_i}) + \mathcal{O}(\epsilon^3)$$

□

F.2 Linear Subadditivity for Close Distributions

Now, we consider distributions that are one or two-sided ϵ-close with a non-infinitesimal $\epsilon > 0$. This is a more realistic setup compared to the setup in Appendix F.1. The Taylor expansion approach used there is no longer applicable. However, using the methodology to prove general f-divergence inequalities (Lemma 15), and a technique of equivalent f-divergences, we are able to obtain linear subadditivity for both cases, under very mild conditions.

We first prove a lemma which reveals the connection between the notion of closeness and linear subadditivity.

Lemma 23. *Consider two f-divergences D_{f_1} and D_{f_2} with generator functions $f_1(t)$ and $f_2(t)$, where f_2 has subadditivity on Bayes-nets with respect to Definition 1. Let $I \subseteq (0, \infty)$ be an interval. If there exists two positive constants $A < B$, such that for any $t \in I$, it holds that $f_2(t) \geq 0$ and $A \leq f_1(t)/f_2(t) \leq B$. Then, for any pair of distributions P and Q, such that for any $x \in \Omega$, $P(x)/Q(x) \in I$, the linear subadditivity inequality of D_{f_1} holds with coefficient $0 < \alpha = A/B < 1$.*

Proof. For any $t \in I$, multiplying $f_2(t) \geq 0$ to the inequalities $A \leq f_1(t)/f_2(t) \leq B$ gives,

$$Af_2(t) \leq f_1(t) \leq Bf_2(t) \qquad \forall t \in I$$

Similar to the proof of Lemma 15 in Appendix B.2, since for any $x \in \Omega$, it holds that $P(x)/Q(x) \in I$, we have,

$$Af_2(P(x)/Q(x)) \leq f_1(P(x)/Q(x)) \leq Bf_2(P(x)/Q(x)) \qquad \forall x \in \Omega$$

Multiply non-negative $Q(x)$ and integrate over Ω. Thus, for such pairs of P, Q, we obtain,

$$AD_{f_2}(P, Q) \leq D_{f_1}(P, Q) \leq BD_{f_2}(P, Q)$$

Now consider P, Q are Bayes-nets such that for any $x \in \Omega$, $P(x)/Q(x) \in I = [a, b]$, i.e. $a \leq P(x)/Q(x) \leq b$. For any non-empty set $S \subsetneqq \{X_1, \cdots, X_n\}$, let $\Omega_{\{X_1, \cdots, X_n\} \setminus S}$ be the space of the variables not in S. Then, multiplying non-negative $Q(x)$ to $a \leq P(x)/Q(x) \leq b$ and integrating over $\Omega_{\{X_1, \cdots, X_n\} \setminus S}$ gives $aQ_S \leq P_S \leq bQ_S$. Moreover, Q_S is positive because Q is positive. Thus, for any pair of marginal distributions P_S and Q_S of such distributions, they also satisfy that for any $x \in \Omega_S$, $P_S(x)/Q_S(x) \in I = [a, b]$.

Applying the first inequality to pairs of marginals $P_{X_i \cup X_{\Pi_i}}$ and $Q_{X_i \cup X_{\Pi_i}}$ gives,

$$D_{f_2}(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}}) \leq \frac{1}{A} D_{f_1}(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}}) \qquad \forall i \in \{1, \cdots, n\}$$

Similarly, applying the second inequality to P and Q gives,

$$\frac{1}{B} D_{f_1}(P, Q) \leq D_{f_2}(P, Q)$$

Combine them with the subadditivity inequality of D_{f_2}, i.e. $D_{f_2}(P,Q) \le \sum_{i=1}^n D_{f_2}(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}})$, we have,

$$\frac{A}{B} D_{f_1}(P,Q) \le \sum_{i=1}^n D_{f_1}(P_{X_i \cup X_{\Pi_i}}, Q_{X_i \cup X_{\Pi_i}})$$

This proves that D_{f_1} satisfy A/B-linear subadditivity for such pairs of Bayes-nets P and Q. □

Now, we list the two theorems characterizing the linear subadditivity of f-divergences when the distributions are one- or two-sided ϵ-close.

Theorem 24. *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 $\lim_{\epsilon \downarrow 0} \alpha = 1$.*

Proof. Following Lemma 23, we consider the quotient $f(t)/f_{\mathrm{H}^2}(t)$, where f_{H^2} is the generator function of squared Hellinger distance, and $f_{\mathrm{H}^2}(t) \coloneqq \frac{1}{2}\left(\sqrt{t}-1\right)^2 \ge 0$ is always non-negative. If we can bound this quotient by positive numbers on an interval $t \in (1-\epsilon, 1+\epsilon)$ for some $0 < \epsilon < 1$, then by Lemma 23, we prove that D_f satisfies linear subadditivity when the distributions P and Q are two-sided ϵ-close.

Because $f(t)$ and $f_{\mathrm{H}^2}(t)$ are continuous functions on $(0,\infty)$, the quotient $f(t)/f_{\mathrm{H}^2}(t)$ is also continuous on $(0,\infty)$. To bound the quotient in the neighborhood around $t=1$, we need to prove $\lim_{t\to 1} f(t)/f_{\mathrm{H}^2}(t)$ exists and is positive. For f_{H^2}, we know $f'_{\mathrm{H}^2}(1) = \frac{1}{2}\left(1 - 1/\sqrt{t}\right)\big|_{t=1} = 0$ and $f''_{\mathrm{H}^2}(1) = \frac{1}{4}t^{-3/2}\Big|_{t=1} = \frac{1}{4} > 0$. Thus, since $f(t)$ is twice differentiable at $t=1$, the limit of the quotient at $t=1$ exists and is positive if and only if $f'(1) = 0$ and $f''(1) > 0$. That is,

$$0 < \lim_{t\to 1} f(t)/f_{\mathrm{H}^2}(t) < \infty \iff f'(1) = 0 \text{ and } f''(1) > 0$$

The latter condition is given, but the former condition, $f'(1) = 0$, does not hold even for some f-divergences which satisfy subadditivity on any Bayes-nets, e.g. for KL divergence, $f'_{\mathrm{KL}}(1) = 1 + \log(t)\big|_{t=1} = 1 \ne 0$.

However a trick can be used to rewrite the generator function $f(t)$ without changing the definition of D_f, so that the modified generator function satisfies the desired condition. For any $k \in \mathbb{R}$, the modified generator $\hat{f}(t) = f(t) + k(t-1)$ defines the same f-divergence,

$$\begin{aligned} D_{\hat{f}}(P,Q) &= \int Q \hat{f}\left(\frac{P}{Q}\right) \mathrm{d}x = \int Q\left(f\left(\frac{P}{Q}\right) + k\left(\frac{P}{Q} - 1\right)\right) \mathrm{d}x \\ &= \int Q f\left(\frac{P}{Q}\right) \mathrm{d}x + k\int (P-Q)\mathrm{d}x = D_f(P,Q) \end{aligned}$$

Thus, for any $f(t)$ twice differentiable at $t=1$ with $f''(1) > 0$, we can define $\hat{f}(t) \coloneqq f(t) - f'(1)(t-1)$. It is easy to verify that $\hat{f}(t)$ has zero first derivative $\hat{f}'(1) = 0$ and positive second derivative $\hat{f}''(1) > 0$ at $t=1$. The modified generator satisfies the two required conditions. As a consequence, we have $0 < \lim_{t\to 1} \hat{f}(t)/f_{\mathrm{H}^2}(t) < \infty$, and the quotient can be bounded by positive numbers in the neighborhood of $t=1$, because of the continuity of $f(t)$. Applying Lemma 23 to interval $I = (1-\epsilon, 1+\epsilon)$ concludes the proof. □

Theorem 11 applies to all practical f-divergences, including KL, reverse KL, χ^2, reverse χ^2, and squared Hellinger H^2 divergences.

In addition to the requirements of Theorem 11, if $f(\cdot)$ is also strictly convex and $f(0) = \lim_{t\downarrow 0} f(t)$ is finite, $\forall t \in [0,1)$, we have the following subadditivity result for one-sided close distributions.

Theorem 25. *An f-divergence whose $f(\cdot)$ is continuous and strictly convex on $(0,\infty)$, twice differentiable at $t=1$, and has finite $f(0) = \lim_{t\downarrow 0} f(t)$, has linear subadditivity with coefficient $\alpha > 0$, when P, Q are one-sided $\epsilon(\alpha)$-close with $\epsilon > 0$, where $\epsilon(\alpha)$ is an non-increasing function and $\lim_{\epsilon\downarrow 0} \alpha > 0$.*

Proof. From the proof of Theorem 11, let $\hat{f}(t) \coloneqq f(t) - f'(1)(t-1)$ be the modified generator function. We know the quotient $\hat{f}(t)/f_{\mathrm{H}^2}(t)$ can be bounded by positive numbers for any $t \in (1-\epsilon, 1+\epsilon)$ for some $0 < \epsilon < 1$. It remains to prove that $\hat{f}(t)/f_{\mathrm{H}^2}(t)$ can be bounded by positive numbers on the interval $[0, 1-\epsilon)$.

The generator $f(t)$ is a strictly convex function on $(0,\infty)$, so is the modified generator $\hat{f}(t)$, since their difference is a linear function of t. Because $\hat{f}'(1)=0$, the tangent line of the curve of $\hat{f}(t)$ at $t=1$ coincides with the x-axis. Since $\hat{f}(t)$ is strictly convex on $(0,\infty)$, the graph of $\hat{f}(t)$ lies above the x-axis, i.e. for any $t\in(0,\infty)$ we have $\hat{f}(t)\geq 0$, where the equality holds if and only if $t=1$. Hence, for any $t\in[0,1-\epsilon)$, it holds that $\hat{f}(t)>0$. Moreover, $\hat{f}(0)=f(0)+f'(1)$ and we know $f(0)=\lim_{t\downarrow 0} f(t)$ is finite. In this sense, $\hat{f}(0)$ is finite and positive. By the continuity of the modified generator $\hat{f}(t)$, we know $\hat{f}(t)$ can be bounded by positive numbers on $[0,1-\epsilon)$. Moreover, clearly $f_{\mathrm{H}^2}(t)\coloneqq\frac{1}{2}\left(\sqrt{t}-1\right)^2$ can be bounded by positive numbers $[0,1-\epsilon)$. This implies that the quotient $\hat{f}(t)/f_{\mathrm{H}^2}(t)$ can be bounded by positive numbers on $[0,1-\epsilon)$. Applying Lemma 23 to the combined interval $I=[0,1+\epsilon)=[0,1-\epsilon)\cup\{\epsilon\}\cup(1-\epsilon,1+\epsilon)$ concludes the proof. □

Using Theorem 12, we can relax the condition $P\gg Q$, as long as $f(0)<\infty$ and $f(\cdot)$ is strictly convex. A broad class of f-divergences satisfy this; see Appendix G below.

G Examples of Local Subadditivity

In this section, we discuss a notable class of f-divergences that satisfy local subadditivity, namely the α-divergences. α-Divergences are f-divergences whose generator functions $f_{\mathcal{H}_\alpha}(\cdot)$ generalize power functions (see Table 2 in Appendix B.1). We show that all α-divergences satisfy linear subadditivity when the distributions are two-sided close, and α-divergences with $\alpha>0$ satisfy linear subadditivity when the distributions are only one-sided close.

Since for any $\alpha\in\mathbb{R}$, $f_{\mathcal{H}_\alpha}(t)$ is continuous with respect to t, and its second order derivative at $t=1$, i.e. $f''_{\mathcal{H}_\alpha}(1)=t^{\alpha-2}\big|_{t=1}=1$ is positive, by Theorem 11 we conclude the following result.

Example 2. *α-divergences,*

$$\mathcal{H}_\alpha(P,Q)\coloneqq\begin{cases}\frac{1}{\alpha(\alpha-1)}\int Q\left((P/Q)^\alpha-1\right)\mathrm{d}x & \alpha\neq 0,1\\ \mathrm{KL}(P,Q) & \alpha=1\\ \mathrm{KL}(Q,P) & \alpha=0\end{cases}$$

which generalize KL and reverse KL divergences, χ^2 and reverse χ^2 divergences, and squared Hellinger distance (see Appendix B.1 for details), satisfy linear subadditivity when the two distributions P and Q are two-sided ϵ-close for some $\epsilon>0$.

For α-divergences with $\alpha>0$, apart from the above-mentioned properties, $f_{\mathcal{H}_\alpha}(t)$ is strictly convex since for any $t\in(0,\infty)$, we have $f''_{\mathcal{H}_\alpha}(t)=t^{\alpha-2}>0$. And $f(0)=\lim_{t\downarrow 0}$ is always finite, because when $\alpha=1$, we have $\lim_{t\downarrow 0} f(t)=0$, and when $\alpha>0$ and $\alpha\neq 1$, the limit $\lim_{t\downarrow 0} f(t)=-\frac{1}{\alpha(\alpha-1)}$ exists. By Theorem 12, we obtain the following.

Example 3. *α-divergences with $\alpha>0$, which generalize KL divergence, χ^2 divergence, and squared Hellinger distance, satisfy linear subadditivity when the two distributions P and Q are one-sided ϵ-close for some $\epsilon>0$.*

H Prior Work on Bounding the IPMs

We list some of the prior work on bounding the Integral Probability Metrics (IPMs). All the concepts and theorems introduced here are used to prove the generalized subadditivity of neural distances on Bayes-nets (Theorem 8) and on MRFs (Corollary 10).

H.1 Preliminaries and Notations

Firstly, we introduce some concepts that help us characterize the set of discriminators $\mathcal{F}$. Consider $\mathcal{F}$ as a set of some functions $\phi:\Omega\to\mathbb{R}$, where $\Omega\subseteq\mathbb{R}^D$. The Banach space of bounded continuous functions is denoted by $C_b(\Omega)\coloneqq\{\phi:\Omega\to\mathbb{R}|\phi \text{ is continuous and } \|\phi\|_\infty<\infty\}$, where $\|\phi\|_\infty=\sup_{x\in X}|\phi(x)|$ is the uniform norm. The linear span of $\mathcal{F}$ is defined as,

$$\mathrm{span}\mathcal{F}\coloneqq\left\{\alpha_0+\sum_{i=1}^n\alpha_i\phi_i\middle|\alpha_i\in\mathbb{R},\phi_i\in\mathcal{F},n\in\mathbb{N}\right\}$$

For a function $g \in \text{span}\mathcal{F}$, we define the $\mathcal{F}$-variation norm $\|g\|_{\mathcal{F}}$ as the infimum of the L_1 norm of the expansion coefficients of g over $\mathcal{F}$, that is,

$$\|g\|_{\mathcal{F}} = \inf \left\{ \sum_{i=1}^{n} |\alpha_i| \Big| g = \alpha_0 + \sum_{i=1}^{n} \alpha_i \phi_i, \forall \alpha_i \in \mathbb{R}, \phi_i \in \mathcal{F}, n \in \mathbb{N} \right\}$$

Let $\text{cl}(\text{span}\mathcal{F})$ be the closure of the linear span of $\mathcal{F}$. We say $g \in \text{cl}(\text{span}\mathcal{F})$ is approximated by $\mathcal{F}$ with an error decay function $\varepsilon(r)$ for $r \geq 0$, if there exists a $\phi_r \in \text{span}\mathcal{F}$, such that $\|\phi_r\|_{\mathcal{F}} \leq r$ and $\|\phi - \phi_r\|_\infty \leq \varepsilon(r)$. In this sense, it is not hard to show that $g \in \text{cl}(\text{span}\mathcal{F})$ if and only if $\inf_{r \geq 0} \varepsilon(r) = 0$.

H.2 The Universal Approximation Theorems

From Theorem 2.2 of (Zhang et al., 2018), we know that $d_{\mathcal{F}}(P, Q)$ is discriminative, i.e. $d_{\mathcal{F}}(P, Q) = 0 \iff P = Q$, if and only if $C_b(X)$ is contained in the closure of $\text{span}\mathcal{F}$, i.e. $C_b(X) \subseteq \text{cl}(\text{span}\mathcal{F})$. In other words, it means that we require $\text{span}\mathcal{F}$ to be dense in $C_b(X)$, so that $d_{\mathcal{F}}(P, Q) \to 0$ implies the weak converge of the fake distribution Q to the real distribution P.

By the famous universal approximator theorem (e.g. Theorem 1 of (Leshno et al., 1993)), the discriminative criteria $C_b(X) \subseteq \text{cl}(\text{span}\mathcal{F})$ can be satisfied by small discriminator sets such as the neural networks with only a single neuron, $\mathcal{F} = \{\sigma(w^T x + b) | w \in \mathbb{R}^D, b \in \mathbb{R}\}$, if the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is continuous but not a polynomial. Later, (Bach, 2017) proves that the set of single-neuron neural networks with rectified linear unit (ReLU) activation also satisfies the criteria.

Theorem 26 (Theorem 1 of (Leshno et al., 1993), (Bach, 2017))**.** *For the set of neural networks with a single neuron, i.e. $\mathcal{F} = \{\sigma(w^T x + b) | w \in \mathbb{R}^D, b \in \mathbb{R}\}$. The linear span of $\mathcal{F}$ is dense in the Banach space of bounded continuous functions $C_b(X)$, i.e. $C_b(x) \subseteq \text{cl}(\text{span}\mathcal{F})$, if the activation function $\sigma(\cdot)$ is continuous but not a polynomial, or if $\sigma(u) = \max\{u, 0\}^\alpha$ for some $\alpha \in \mathbb{N}$ (when $\alpha = 1$, $\sigma(u) = \max\{u, 0\}$ is the ReLU activation).*

See (Leshno et al., 1993) and (Bach, 2017) for further details and the proofs.

H.3 IPMs Upper-Bounding the Symmetric KL Divergence

(Zhang et al., 2018) explains how IPMs can control the likelihood function, so that along with the training of an IPM-based GAN, the training likelihood should generally increase. More specifically, they prove that if the densities P and Q exist, and $\log(P/Q)$ is inside the closure of the linear span of $\mathcal{F}$, i.e. $\log(P/Q) \in \text{cl}(\text{span}\mathcal{F})$, a function of the IPM $d_{\mathcal{F}}(P, Q)$ can upper-bound the Symmetric KL divergence $\text{SKL}(P, Q)$. In this sense, minimizing the IPM leads to the minimization of Symmetric KL divergence (and thus KL divergence), which is equivalent to the maximization of the training likelihood.

Theorem 27 (Proposition 2.7 and 2.9 of (Zhang et al., 2018))**.** *Any function g inside the closure of the linear span of $\mathcal{F}$, i.e. $g \in \text{cl}(\text{span}\mathcal{F})$, is approximated by $\mathcal{F}$ with an error decay function $\varepsilon(r)$. It satisfies,*

$$\Big| \mathbb{E}_{x \sim P}[g(x)] - \mathbb{E}_{x \sim Q}[g(x)] \Big| \leq 2\varepsilon(r) + r d_{\mathcal{F}}(P, Q) \qquad \forall r \geq 0$$

Moreover, consider two distributions with positive densities P and Q, if $g = \log(P/Q) \in \text{cl}(\text{span}\mathcal{F})$, we have,

$$\text{SKL}(P, Q) \equiv \Big| \mathbb{E}_{x \sim P}[\log(P(x)/Q(x))] - \mathbb{E}_{x \sim Q}[\log(P(x)/Q(x))] \Big| \leq 2\varepsilon(r) + r d_{\mathcal{F}}(P, Q) \qquad \forall r \geq 0$$

Proof. The proof is in Appendix C of (Zhang et al., 2018). We repeat the proof here for completeness.

Since g is approximated by $\mathcal{F}$ with error decay function $\varepsilon(r)$, for any $r \geq 0$, there exist some $\phi_r \in \text{span}\mathcal{F}$, which can be represented as $\phi_r = \sum_{i=1}^n \alpha_i \phi_i + \alpha_0$ with some $\alpha_i \in \mathbb{R}$ and $\phi_i \in \mathcal{F}$, such that $\sum_{i=1}^n |\alpha_i| = \|\phi_r\|_{\mathcal{F}} \leq r$ and

$\|g - \phi_r\|_\infty < \varepsilon(r)$. In this sense, we have,

$$\begin{aligned}
&\Big|\mathbb{E}_{x\sim P}[g(x)] - \mathbb{E}_{x\sim Q}[g(x)]\Big| \\
&= \Big|\big(\mathbb{E}_{x\sim P}[g(x)] - \mathbb{E}_{x\sim P}[\phi_r(x)]\big) - \big(\mathbb{E}_{x\sim Q}[g(x)] - \mathbb{E}_{x\sim Q}[\phi_r(x)]\big) + \big(\mathbb{E}_{x\sim P}[\phi_r(x)] - \mathbb{E}_{x\sim Q}[\phi_r(x)]\big)\Big| \\
&\le \Big|\mathbb{E}_{x\sim P}[g(x) - \phi_r(x)]\Big| + \Big|\mathbb{E}_{x\sim Q}[g(x) - \phi_r(x)]\Big| + \Big|\mathbb{E}_{x\sim P}[\phi_r(x)] - \mathbb{E}_{x\sim Q}[\phi_r(x)]\Big| \\
&\le \mathbb{E}_{x\sim P}\big|g(x) - \phi_r(x)\big| + \mathbb{E}_{x\sim Q}\big|g(x) - \phi_r(x)\big| + \Big|\sum_{i=1}^n \alpha_i\big(\mathbb{E}_{x\sim P}[\phi_i(x)] - \mathbb{E}_{x\sim Q}[\phi_i(x)]\big)\Big| \\
&\le 2\varepsilon(r) + \sum_{i=1}^n |\alpha_i|\Big|\mathbb{E}_{x\sim P}[\phi_i(x)] - \mathbb{E}_{x\sim Q}[\phi_i(x)]\Big| \\
&\le 2\varepsilon(r) + r d_{\mathcal{F}}(P, Q)
\end{aligned}$$

Applying this inequality to $g = \log(P/Q)$ proves that, for any $r \ge 0$, this linear function of IPM $2\varepsilon(r) + r d_{\mathcal{F}}(P, Q)$ upper-bounds the Symmetric KL divergence $\mathrm{SKL}(P, Q)$. □

The upper-bounds obtained by Theorem 27 are a set linear functions of the IPM, $\{2\varepsilon(r) + r d_{\mathcal{F}}(P, Q)\big|r \ge 0\}$. In order to prove that the IPM $d_{\mathcal{F}}(P, Q)$ can upper-bound the Symmetric KL divergence $\mathrm{SKL}(P, Q)$ up to some constant coefficient and additive error, i.e. $\alpha\mathrm{SKL}(P, Q) - \epsilon \le d_{\mathcal{F}}(P, Q)$ for some constants $\alpha, \varepsilon > 0$, we have to control both $\varepsilon(r)$ and r simultaneously. Because $\lim_{r\to\infty} \varepsilon(r) = 0$, all we need is an efficient upper-bound on $\varepsilon(r)$ for large enough r, which is provided in (Bach, 2017).

Theorem 28 (Proposition 6 of (Bach, 2017))**.** *For a bounded space Ω, let $g : \Omega \to \mathbb{R}$ be a bounded and Lipschitz continuous function (i.e. there exists a constant $\eta > 0$ such that $\|g\|_\infty < \eta$ and for any $x, y \in \Omega \subseteq \mathbb{R}^D$, it holds that $\|g(x) - g(y)\|_\infty \le \frac{1}{\mathrm{diam}\Omega}\eta\|x - y\|_2$), and let $\mathcal{F}$ be a set of neural networks with a single neuron, which have ReLU activation and bounded parameters (i.e. $\mathcal{F} = \{\max\{w^T x + b, 0\}\big|w \in \mathbb{R}^D, b \in \mathbb{R}, \|[w, b]\|_2 = 1\}$). Then, we have $g \in \mathrm{cl}(\mathrm{span}\mathcal{F})$, and g is approximated by $\mathcal{F}$ with error decay function $\varepsilon(r)$, such that,*

$$\varepsilon(r) \le C(D)\eta\left(\frac{r}{\eta}\right)^{-\frac{2}{D+1}} \log\left(\frac{r}{\eta}\right) \qquad \forall r \ge R(D)$$

where $C(D), R(D)$ are constants which only depend on the number of dimensions, D.

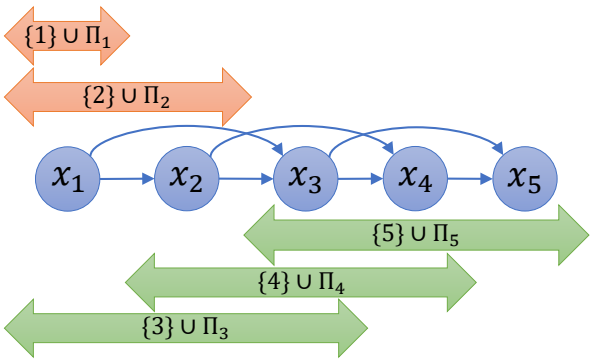
See Proposition 3, Appendix C.3, and Appendix D.4 of (Bach, 2017) for the proof.

I Subadditivity Upper-Bounds at Different "Levels of Detail" on Sequences

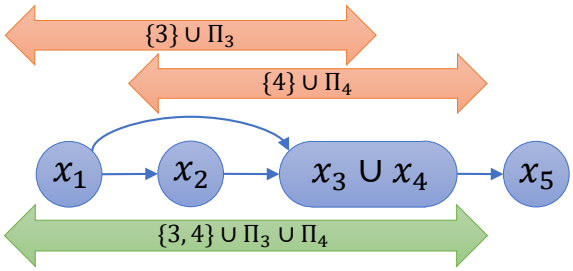
The subadditivity upper-bound on a Bayes-net, $\sum_{i=1}^n \delta(P_{X_i\cup X_{\Pi_i}}, Q_{X_i\cup X_{\Pi_i}})$, depends on the structure of the Bayes-net. More specifically, the underlying DAG G determines the set of local neighborhoods $\{\{1\}\cup\Pi_1, \cdots, \{n\}\cup\Pi_n\}$, and consequently, determines how we construct the set of local discriminators. In this section, we discuss that the set of local neighborhoods can be change either by truncating the induction process when deriving the subadditivity upper-bound (see the proof of Theorem 1 in Appendix A.1 for example), or by contracting the neighboring nodes of the Bayes-net. Both methods result in a tighter subadditivity upper-bound at a coarser level-of-detail (i.e., with larger local neighborhoods). For the sake of simplicity, we limit the scope to sequences describing auto-regressive time series. For such graph G, there are T nodes ($\{1, \cdots, T\}$), and each node depends on its p previous nodes; see Fig. 9(a) for an example.

I.1 Truncation of Induction

For a probability divergence δ which satisfies subadditivity on Bayes-nets, the subadditivity upper-bound $\sum_{t=1}^T \delta(P_{X_t\cup X_{\Pi_t}}, Q_{X_t\cup X_{\Pi_t}})$ of $\delta(P, Q)$ is obtained by repeatedly applying the subadditivity of δ on Markov Chains of super-nodes $X_{\{1,\cdots,s\}\setminus\Pi_{s+1}} \to X_{\Pi_{s+1}} \to X_{s+1}$, for $s = T-1, T-2, \cdots, 1$. We can truncate the induction process and get an alternative upper-bound: $\delta(P, Q) < \delta(P_{\cup_{t=1}^s X_t}, Q_{\cup_{t=1}^s X_t}) + \sum_{t=s+1}^T \delta(P_{X_t\cup X_{\Pi_t}}, Q_{X_t\cup X_{\Pi_t}})$. This new upper-bound is tighter, but it does not encode the conditional independence information of the sub-sequence



(a) Local neighborhoods of auto-regressive time series with $T = 5$ and $p = 2$. The original set of local neighborhoods is represented by the red and green bars. Two local neighborhoods $\{1\} \cup \Pi_1$ and $\{2\} \cup \Pi_2$ (red bars) can be safely removed by truncating the induction process.



(b) Change of the local neighborhoods of auto-regressive time series with $T = 5$ and $p = 2$, if contracting neighboring nodes 3 and 4 to form a super-node $\{3, 4\}$. Two local neighborhoods $\{3\} \cup \Pi_3$ and $\{4\} \cup \Pi_4$ (red bars) are replaced by a new neighborhood $\{3, 4\} \cup \Pi_3 \cup \Pi_4$ (green bar).

Figure 9: Changes of the local neighborhoods of a Bayes-net representing auto-regressive time series with $T = 2$ and $p = 2$, if we (a) truncate the induction process, or (b) contract a pair of neighboring nodes. In each case, the subadditivity upper-bound becomes tighter and characterize the Bayes-net at a coarser level-of-detail.

$(X_1, \cdots, X_s)$. However, this alternative upper-bound is preferable if we choose s to be the largest number where its set of parents is exactly its previous nodes, i.e. $\Pi_s = \{1, \cdots, s-1\}$. The subadditivity inequality that we combined at induction step s is $\delta(P_{\cup_{t=1}^{s} X_t}, Q_{\cup_{t=1}^{s} X_t}) \equiv \delta(P_{X_{\Pi_s} \cup X_s}, Q_{X_{\Pi_s} \cup X_s}) \leq \delta(P_{\cup_{t=1}^{s-1} X_t}, Q_{\cup_{t=1}^{s-1} X_t}) + \delta(P_{X_{\Pi_s} \cup X_s}, Q_{X_{\Pi_s} \cup X_s})$ (corresponding to the second case in the proof of Theorem 1 in Appendix A.1). Truncating at such s avoids introducing the redundant term $\delta(P_{\cup_{t=1}^{s-1} X_t}, Q_{\cup_{t=1}^{s-1} X_t})$ into the upper-bound. As shown in Fig. 9(a), for this specific example $s = p + 1 = 3$ is the largest number such that $\Pi_3 = \{1, 2\}$. Truncating at $s = 3$ removes $\{1\} \cup \Pi_1$ and $\{2\} \cup \Pi_2$ from the set of local neighborhoods, resulting in a more efficient subadditivity upper-bound $\sum_{t=3}^{5} \delta(P_{X_t \cup X_{\Pi_t}}, Q_{X_t \cup X_{\Pi_t}})$. This is helpful for time series data, since it makes all local neighborhoods have the same number of dimensions. If all $X_t \in \mathbb{R}^d$, then for $t = 3, 4$ and 5, $X_t \cup X_{\Pi_t} \in \mathbb{R}^{3d}$. In this sense, we can share the same neural network architecture among all the local discriminators.

I.2 Neighboring Nodes Contraction

The set of local neighborhoods is determined by the structure G of the Bayes-net. Network contraction not only simplifies the Bayes-net but also leads to a tighter subadditivity upper-bound at a lower level-of-detail. Here, we only consider the contraction of neighboring nodes in a time series $(X_1, \cdots, X_T)$. If we merge node s with $s + 1$ $(s = 1, \cdots, T - 1)$, and form a super-node $\{s.s+1\}$, local neighborhoods $\{s\} \cup \Pi_s$ and $\{s+1\} \cup \Pi_{s+1}$ are replaced by $\{s, s+1\} \cup \Pi_s \cup \Pi_{s+1}$, and the total number of neighborhoods decreases by one. As shown in Fig. 9(b), when nodes 3 and 4 are merged, local neighborhoods $\{3\} \cup \Pi_3$ and $\{4\} \cup \Pi_4$ are replaced by $\{3, 4\} \cup \Pi_3 \cup \Pi_4$. We omit the conditional dependence between nodes 3 and 4, but reduce one local discriminator in the GAN. Neighboring nodes contraction allows us to control the level-of-detail that the subadditivity upper-bound encodes flexibly. This can be useful when the variables in the Bayes-net have non-uniform dimensionalities.

J More Experiment Results

J.1 Experiments on Synthetic MRFs

We perform some additional experiments to demonstrate the benefits of exploiting the underlying MRF structure of the data in the design of model-based GANs. We generate a synthetic Gaussian MRF dataset, the graph of which is a 4-cycle (namely a graph with 4 nodes, each of which has degree 2). We train our model-based GAN (we call them MRF GANs) and the standard model-free GAN on five thousand samples. The network architecture (except for the number of inputs) and the hyper-parameters are the same as what we used on the EARTHQUAKE

dataset (see Appendix K). We evaluate the performance of the GANs by the Energy Statistics score (for the definition, see Section 8). We observe that the average Energy Statistics of MRF GAN is $(1.3 \pm 0.2) \times 10^{-3}$, which is smaller (thus better) than the standard GAN's, $(7.9 \pm 1.2) \times 10^{-3}$. This simple experiment confirms that the benefits of exploiting the conditional independence structure apply to MRFs as well. This is consistent with our theory. Such benefits can be reaped even in low dimensions.

J.2 Sensitivity Analysis of Bayes-net GANs

In this part, we analyze how sensitive Bayes-net GAN is to its causal structure. In Section 8.2, we know from the experiments on two real Bayes-nets that the Bayes-net GANs with the ground truth causal DAGs consistently outperform the standard GANs in terms of the generation quality and the convergence speed. Here, we perturb the ground truth DAG of *the EARTHQUAKE dataset* by randomly rewiring 1 of its 4 edges. The graph editing distance between the resulting noisy causal DAG and the ground truth is therefore 1. We train a Bayes-net GAN with local discriminators constructed using this noisy DAG, and evaluate it by the four metrics as in Section 8.2 (Table 3), as well as using the causal structure predicted from its generated data (Fig. 10). In Table 3, we observe that the energy statistics and the detection AUC scores are not sensitive to the noise in the casual structure used by the Bayes-net GAN. The BIC score and the predicted structure are more affected. The Bayes-net GAN with the noisy causal graph learns some redundant indirect dependence (Fig. 10(b)).

DAG 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)
Ground truth	0.24 ± 0.04	0.523 ± 0.005	$+1.68 \pm 0.17$	0.4 ± 0.7
Noisy	0.27 ± 0.11	0.528 ± 0.002	-2.02 ± 0.15	2.4 ± 1.1

Table 3: Quality metrics of samples generated by the Bayes-net GANs with the ground truth and the noisy causal structure.

(a) Bayes-net GAN with ground truth DAG

(b) Bayes-net GAN with noisy DAG

Figure 10: Causal structures predicted from the data generated by the Bayes-net GAN with (a) the ground truth causal graph and (b) the noisy causal graph.

K Experimental Setups

In this section, we report the detailed setups of the experiments in Section 8. Unless otherwise stated, model comparisons between our model-based GANs and the standard GANs are conducted using exactly the same set of training hyper-parameters. All experiments are repeated five or ten times.

K.1 Datasets

- **Experiment 1: synthetic ball throwing trajectories**: *The ball throwing trajectory dataset* is synthetic, 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.

- **Experiment 2: real Bayes-nets**: The *EARTHQUAKE dataset* 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). *The CHILD dataset* is a Bayes-net with 20 nodes and 25 edges for diagnosing congenital heart disease in a newborn "blue baby" (Spiegelhalter, 1992). The underlying causal DAGs of two Bayes-nets are known, and are obtained from `https://www.bnlearn.com/bnrepository/`. Both Bayes-nets have categorical features. We simulate samples from the Bayes-nets and then train GANs on them.

K.2 Local Discriminators

- **Experiment 1: synthetic ball throwing trajectories**: Each local discriminator measures the Jensen-Shannon divergence in the local neighborhood, following (Nowozin et al., 2016). The use of local discriminators is justified by the $(1/\ln 2)$-linear subadditivity of Jensen-Shannon divergence on Bayes-nets (Corollary 5).

- **Experiment 2: real Bayes-nets**: Each local discriminator measures the Wasserstein distance in the local neighborhood, following (Arjovsky et al., 2017). The use of local discriminators is justified by the generalized subadditivity of neural distances on Bayes-nets (Theorem 8). We use Gumbel-Softmax (Jang et al., 2016) in the output layer of the generator, so that the generator produces categorical data while allowing (approximate) back propagation.

K.3 Network Architectures

- **Experiment 1: synthetic ball throwing trajectories**: For the GANs on *the ball throwing trajectory* dataset, we use a 5-layer fully connected network (FCN) for the generator where the number of hidden dimensions is set to 8 for all layers. We take a hybrid design for each local discriminator. Each local discriminator is a combination of a 4-layer FCN and a 3-layer convolutional neural network (CNN) so that it can penalize both inaccurate global distributions (via FCN) and inaccurate local dynamics (via CNN). The discriminator of the standard GAN on *the ball throwing trajectory* dataset is the local discriminator with localization width 15 (which is equal to the length of the time-series). The local discriminators share the same architecture (except the input layer) even if the localization width varies, but they do not share parameters. We make the discriminators' architecture powerful enough such that adding more neurons/layers cannot bring us any further performance gain.

- **Experiment 2: real Bayes-nets**: For the GANs on *the EARTHQUAKE dataset*, we use a 5-layer FCN for the generator where the number of hidden dimensions is set to 32 for all layers. We also use a 4-layer FCN for each local discriminator where the number of hidden dimensions is set to 8. For the GANs on *the CHILD dataset*, we use a 7-layer FCN with 256 hidden dimensions for the generator, and a 6-layer FCN with 32 hidden dimensions for the discriminator. There is no parameter sharing among the local discriminators. We apply Batch Normalization after each hidden layer in the generator, but not in the discriminator. All ReLUs are leaky, with slope 0.2. In the experiments, we keep the architecture of the generator and the other hyper-parameters the same, and compare our model-based GANs with the standard GANs on Bayes-nets.

K.4 Training Setups and Hyper-Parameters

The networks are implemented using the *PyTorch* framework. All networks are trained from scratch on one *NVIDIA RTX 2080 Ti* GPU with 11GB memory.

- **Experiment 1: synthetic ball throwing trajectories**: We train GANs with local discriminators (with localization width equals to $1, 2, 3, 5, 8, 11, 15$) for 500 epochs, with learning rate 0.0001 and batch size 128. We repeat each experiment 10 times and report the averages with uncertainties.

- **Experiment 2: real Bayes-nets**: We train the standard GANs and the Bayes-net GANs for 100 epochs, with learning rate 0.001 and batch size 128. We repeat each experiment 5 times and report the averages with uncertainties.

K.5 Evaluation Setups

- **Experiment 1: ball throwing trajectories**: We estimate the gravitational acceleration g learned by the GANs, via degree-2 polynomial regression on the generated trajectories.
- **Experiment 2: real Bayes-nets**: The energy statistics are calculated using the standard *torch-two-sample* package (available at `https://github.com/josipd/torch-two-sample`). The fake detection AUC scores are obtained by training binary classifiers to distinguish the fake samples from the real ones. The binary classifier is a 3-layer FCN with hidden dimensions 16 on *the EARTHQUAKE dataset* and a 5-layer FCN with hidden dimension 32 on *the CHILD dataset*. We train the classifiers for 100 epochs, with learning rate 0.001 and batch size 128.

L Empirical Verification of Subadditivity

In this section, we verify the subadditivity of squared Hellinger distance, KL divergence, Symmetric KL divergence, and the linear subadditivity of Jensen-Shannon divergence, Total Variation distance, 1-Wasserstein distance, and 2-Wasserstein distance on binary auto-regressive sequences in a finite space Ω.

To construct a simple Bayes-net P on a sequence of bits $(X_1, \cdots, X_n) \in \{0,1\}^n$, consider the auto-regressive sequence defined by,

$$P(X_t = 1 | X_{t-1}, \cdots, X_{t-p}) = \sigma(\sum_{i=1}^{p} \varphi_i X_{t-i})$$

where $p \in \mathbb{N}$ such that $0 < p < n$ is called the order of this auto-regressive sequence, and $[\varphi_1, \cdots, \varphi_n]$ are the coefficients. The marginal distributions of the initial variables $X_1, \cdots, X_p$ have to be pre-defined. We assume they are conditionally independent, and define,

$$P(X_i = 1) = \psi_i \qquad \forall i \in \{1, \cdots, p\}$$

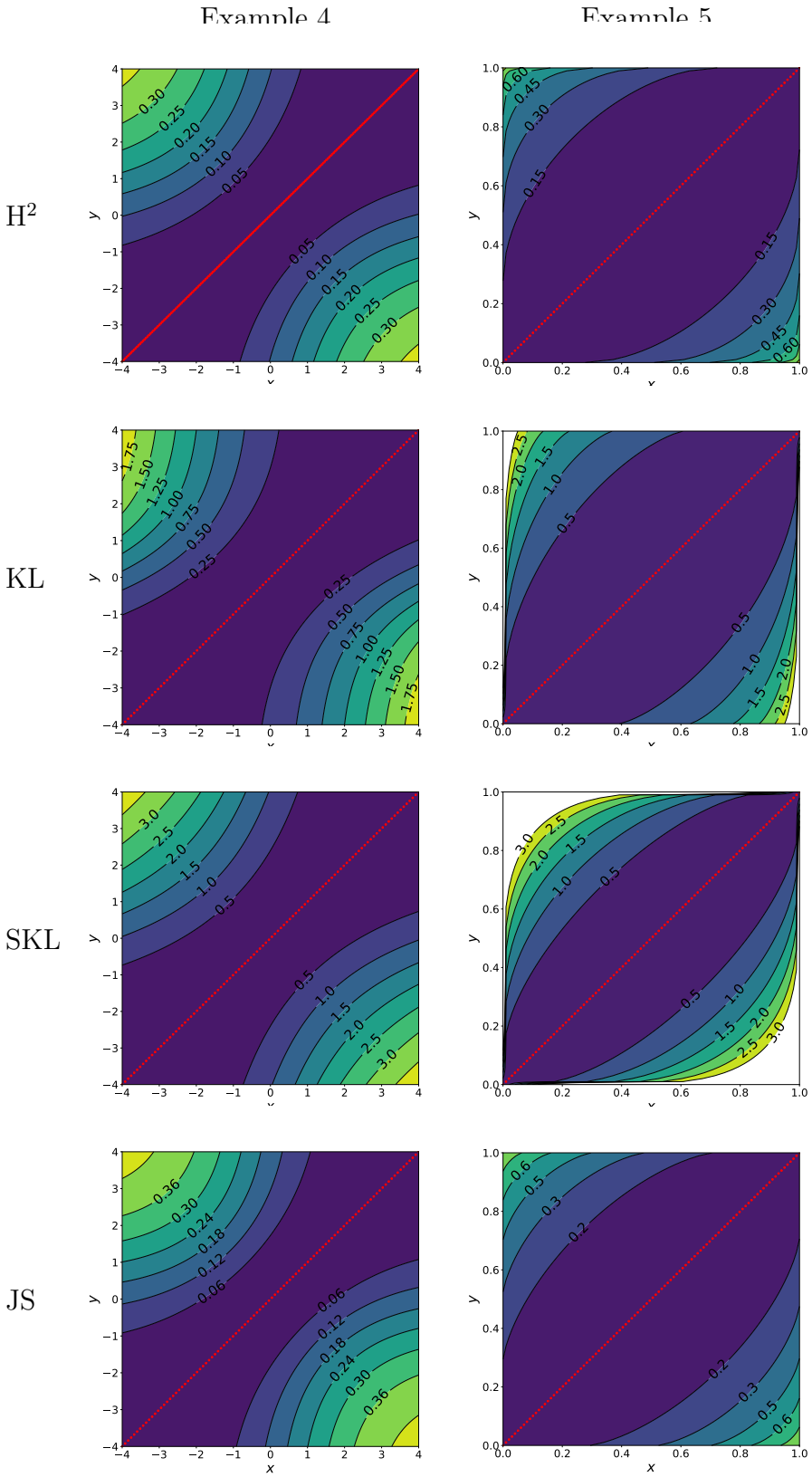
where for any $i \in \{1, \cdots, p\}$, $\psi_i \in [0,1]$. If the distribution of a binary sequence $(X_1, \cdots, X_n)$ follows the definitions above, we say it is a binary auto-regressive sequence of order p with coefficients $[\varphi_1, \cdots, \varphi_n]$ and initials $[\psi_1, \cdots, \psi_n]$.

Binary auto-regressive sequences are Bayes-nets, because each variable X_t is conditionally independent of its non-descendants given its parent variables $X_{t-1}, \cdots, X_{t-p}$. The probabilistic graph G is determined by the length n and the order p. For a statistical divergence δ satisfying subadditivity, as described in Appendix I.1, we truncate the induction process and get a subadditivity upper-bound $\sum_{t=p+1}^{n} \delta(P_{\cup_{i=t-p}^{t} X_i}, Q_{\cup_{i=t-p}^{t} X_i})$. We verify that the subadditivity inequality (or linear subadditivity inequality) holds for various statistical divergences, on two specific examples.

Example 4 (Binary Auto-Regressive Sequences with Different Local Dependencies)**.** *Consider binary auto-regressive sequences $(X_1, X_2, X_3, X_4) \in \{0,1\}^4$ of order $p = 2$ with initials $[\psi_1, \psi_2] = [\frac{1}{2}, \frac{1}{2}]$. Two distributions P^x (with coefficients $[\varphi_1, \varphi_2] = [0, x]$) and Q^y (with coefficients $[\varphi_1, \varphi_2] = [0, y]$) are Bayes-nets with identical underlying structure. Divergence $\delta(P^x, Q^y)$ is a function of the parameters (x, y). For all $(x, y) \in \{(x, y) \in \mathbb{R}^2 | x \neq y\}$, we have $\delta(P^x, Q^y) < \sum_{t=p+1}^{n} \delta(P^x_{\cup_{i=t-p}^{t} X_i}, Q^y_{\cup_{i=t-p}^{t} X_i})$ if δ satisfies subadditivity, or $\alpha \cdot \delta(P^x, Q^y) < \sum_{t=p+1}^{n} \delta(P^x_{\cup_{i=t-p}^{t} X_i}, Q^y_{\cup_{i=t-p}^{t} X_i})$ if δ satisfies α-linear subadditivity.*

Example 5 (Binary Auto-Regressive Sequences with Different Initial Distributions)**.** *Consider binary auto-regressive sequences $(X_1, X_2, X_3, X_4) \in \{0,1\}^4$ of order $p = 2$ with coefficients $[\varphi_1, \varphi_2] = [1, -1]$. Two distributions P^x (with initials $[\psi_1, \psi_2] = [\frac{1}{2}, x]$) and Q^y (with initials $[\psi_1, \psi_2] = [\frac{1}{2}, y]$) are Bayes-nets with identical underlying structure. Divergence $\delta(P^x, Q^y)$ is a function of the parameters (x, y). For all $(x, y) \in \{(x, y) \in \mathbb{R}^2 | 0 < x \neq y < 1\}$, we have $\delta(P^x, Q^y) < \sum_{t=p+1}^{n} \delta(P^x_{\cup_{i=t-p}^{t} X_i}, Q^y_{\cup_{i=t-p}^{t} X_i})$ if δ satisfies subadditivity, or $\alpha \cdot \delta(P^x, Q^y) < \sum_{t=p+1}^{n} \delta(P^x_{\cup_{i=t-p}^{t} X_i}, Q^y_{\cup_{i=t-p}^{t} X_i})$ if δ satisfies α-linear subadditivity.*

We verify the subadditivity of H^2, KL, SKL, and the linear subadditivity of JS, TV, W_1 and W_2 on these two examples, as shown in Fig. 11. We draw contour plots of the subadditivity gap $\Delta = \sum_{t=p+1}^{n} \delta(P^x_{\cup_{i=t-p}^{t} X_i}, Q^y_{\cup_{i=t-p}^{t} X_i}) - \delta(P^x, Q^y)$ (if δ satisfies subadditivity) or $\Delta = \sum_{t=p+1}^{n} \delta(P^x_{\cup_{i=t-p}^{t} X_i}, Q^y_{\cup_{i=t-p}^{t} X_i}) - \alpha \cdot \delta(P^x, Q^y)$ (if δ satisfies α-linear subadditivity). All the inequalities are verified as we can visually confirm all contours are positive.


Example 4
Example 5
H^2
KL
SKL
JS
x
y

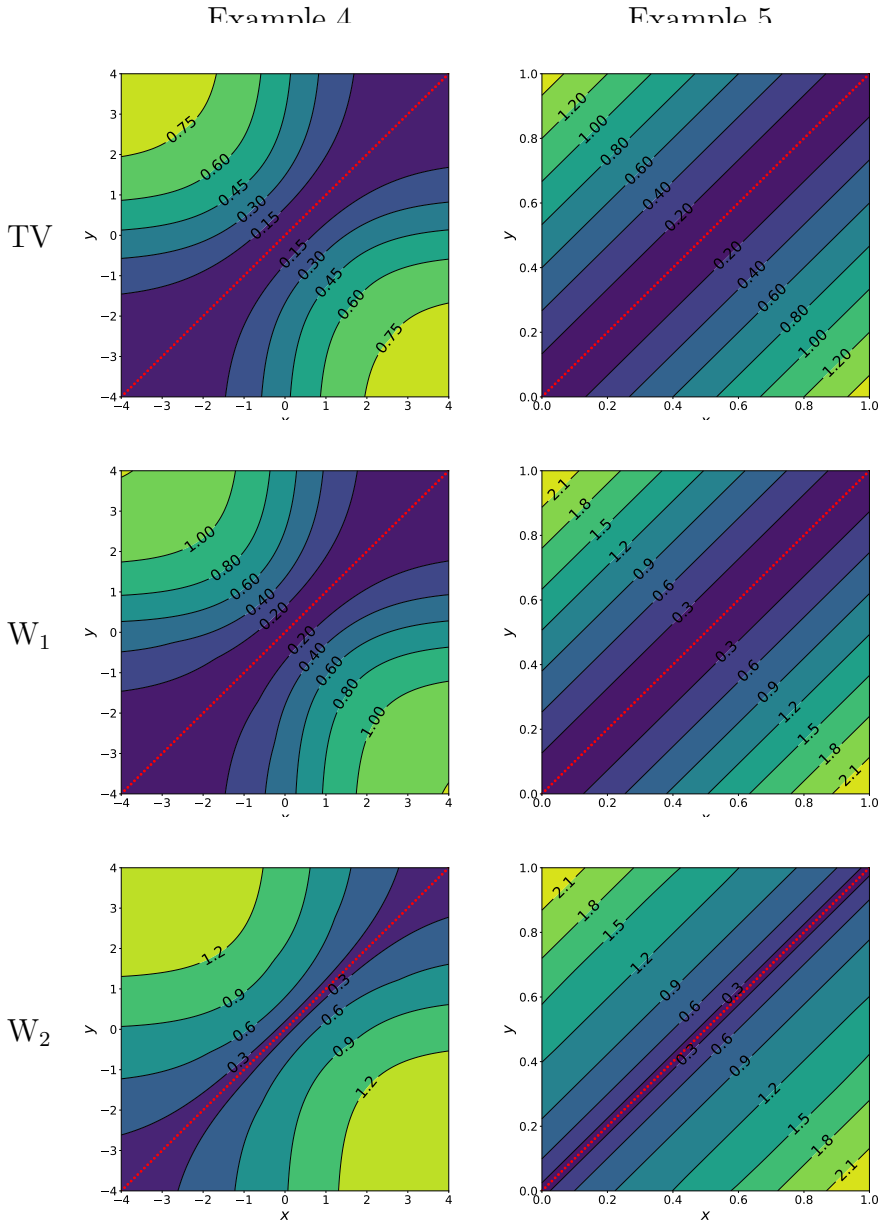


Figure 11: Contour maps showing the binary auto-regressive sequence examples of subadditivity or linear subadditivity of H^2, KL, SKL, JS, TV, W_1, and W_2. The two distributions P^x, Q^y are distributions of binary auto-regressive sequences with length $n = 4$ and order $p = 2$, following definitions in Example 4 and Example 5. The contours and colors indicate the subadditivity gap $\Delta = \sum_{t=p+1}^{n} \delta(P^x_{\cup_{i=t-p}^{t} X_i}, Q^y_{\cup_{i=t-p}^{t} X_i}) - \delta(P^x, Q^y)$ (if δ satisfies subadditivity) or $\Delta = \sum_{t=p+1}^{n} \delta(P^x_{\cup_{i=t-p}^{t} X_i}, Q^y_{\cup_{i=t-p}^{t} X_i}) - \alpha \cdot \delta(P^x, Q^y)$ (if δ satisfies α-linear subadditivity). The red dotted line indicates places where the subadditivity gap is 0. White regions have too large subadditivity gap to be colored.

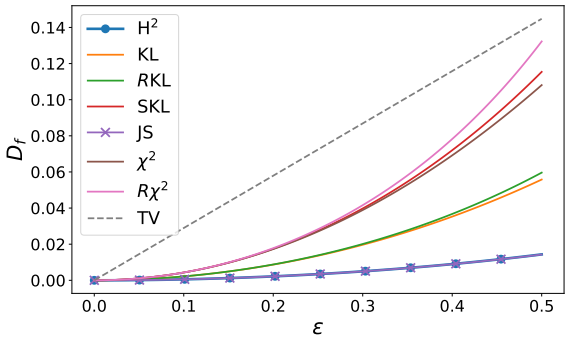
M Empirical Verification of the Local Approximations of f-Divergences

In this section, we observe the local behavior of common f-divergences when the two distributions P and Q are sufficiently close. And we verify the conclusion of Lemma 22: all f-divergences D_f with a generator function $f(t)$ that is twice differentiable at $t = 1$ and satisfies $f''(1) > 0$ have similar local approximations up to a constant factor up to $\mathcal{O}(\epsilon^3)$. More specifically, for a pair of two-sided ϵ-close distributions P and Q, we verify all such f-divergences satisfy:

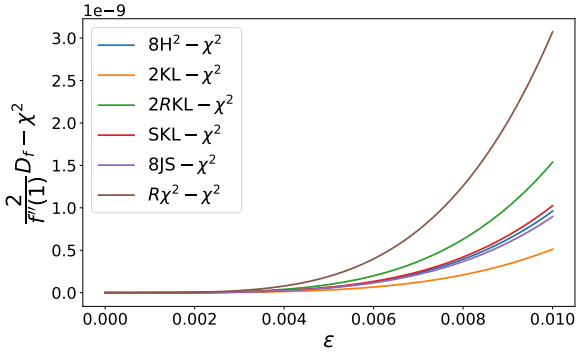
$$D_f(P, Q) = \frac{f''(1)}{2}\chi^2(P, Q) + \mathcal{O}(\epsilon^3)$$

Let us consider a simple example of two-sided close distributions on $\Omega = \mathbb{R}$. Suppose $Q = \mathcal{N}(0, 1)$ is the 1-dimensional unit Gaussian. Let $P(x) = (1 + \epsilon \sin(x))\, Q(x)$ for some $\epsilon \in (0, 1)$. It is easy to verify that P is a valid probability distribution: $\int_{-\infty}^{\infty} P(x)\mathrm{d}x = \int_{-\infty}^{\infty} Q(x)\mathrm{d}x + \epsilon \int_{-\infty}^{\infty} \sin(x)Q(x)\mathrm{d}x = 1$, where the term $\int_{-\infty}^{\infty} \sin(x)Q(x)\mathrm{d}x$ vanishes because $Q(x)$ is an even function and $\sin(x)$ is odd. Since for any $x \in \Omega = \mathbb{R}$, it holds that $P(x)/Q(x) = 1 + \epsilon \sin(x) \in [1 - \epsilon, 1 + \epsilon]$, we know P and Q are two-sided ϵ-close.

We compute several common f-divergences between such P and Q, for different $\epsilon \in [0, 0.5]$, as shown in Fig. 12(a). We can see that, except for Total Variation distance which has a generator f_{TV} not differentiable at 1, all common f-divergences behave similarly up to a constant factor. Actually, these curves cluster into three groups according to $f''(1)$. In the first cluster: $f''_{\mathrm{SKL}}(1) = f''_{\chi^2}(1) = f''_{R\chi^2}(1) = 2$. In the second cluster: $f''_{\mathrm{KL}}(1) = f''_{R\mathrm{KL}}(1) = 1$. While in the third cluster: $f''_{\mathrm{H}^2}(1) = f''_{\mathrm{JS}}(1) = \frac{1}{4}$. Moreover, we visualize the differences between f-divergences normalized with respect to $f''(1)$ and χ^2 divergence, for $\epsilon \in [0, 0.01]$. We can see in Fig. 12(b), all the differences are very small. This verifies that all f-divergences such that $f''(1) > 0$ satisfy $\frac{2}{f''(1)} D_f(P, Q) = \chi^2(P, Q)$ up to $\mathcal{O}(\epsilon^3)$.



(a) Common f-divergences between such P and Q for $\epsilon \in [0, 0.5]$.



(b) Differences between f-divergences normalized with respect to $f''(1)$ and χ^2 divergence for $\epsilon \in [0, 0.01]$.

Figure 12: Common f-divergences between two-sided ϵ-close distributions P, Q, where Q is the 1-dimensional unit Gaussian and $P(x) = (1 + \epsilon \sin(x))\, Q(x)$. In (a), we compare these f-divergences for $\epsilon \in [0, 0.5]$. In (b), we verify the conclusion of Lemma 22: $\frac{2}{f''(1)} D_f(P, Q) = \chi^2(P, Q) + \mathcal{O}(\epsilon^3)$.