Proceedings of Machine Learning Research vol 65:1–7, 2017

Square Hellinger Subadditivity for Bayesian Networks and its Applications to Identity Testing

Constantinos Daskalakis COSTIS@MIT.EDU
EECS and CSAIL, MIT

Qinxuan Pan QINXUAN@MIT.EDU
EECS and CSAIL, MIT

Extended Abstract[1]

At the heart of scientific activity lies the practice of formulating models about observed phenomena, and developing tools to test the validity of these models. Oftentimes, the models are probabilistic; for example, one may model the effectiveness of a drug in a population as a truncated Normal, or the waiting times in a queuing system as exponential random variables. When a model is probabilistic, testing its validity becomes a distribution testing problem. In our drug example, one would like to measure the effectiveness of the drug in a sample of the population, and somehow determine whether these samples are "consistent" with a truncated Normal distribution. As humans delve into the study of more and more complex phenomena, they quickly face high-dimensional distributions. *The goal of this paper is to advance our understanding of high-dimensional hypothesis testing.*

Consider the task of testing whether a high-dimensional distribution P, to which we have sample access, is identical to some model distribution $Q \in \Delta(\Sigma^n)$, where Σ is some alphabet and n is the dimension. A natural goal, which we will call *goodness-of-fit testing* in the tradition of Statistics, is to distinguish

$$P = Q \quad \text{from} \quad d(P, Q) > \epsilon,$$

where $d(\cdot, \cdot)$ is some distance between distributions and $\epsilon > 0$ some accuracy parameter. In this paper, we will take $d(\cdot, \cdot)$ to be the total variation distance, although all our results hold if one considers Hellinger distance instead.

Sometimes we do not have a model distribution Q, but sample access to two distributions $P, Q \in \Delta(\Sigma^n)$, and we want to determine if they are equal. Again, a natural goal is to distinguish

$$P = Q \quad \text{from} \quad d(P, Q) > \epsilon.$$

We will call this latter problem, where both distributions are unknown, *identity testing*.

As our access to P or to both P and Q in the above problems is via samples, we cannot hope to always solve them correctly. So our goal is actually probabilistic. We want to be correct with probability at least $1 - \delta$, for some parameter δ. For ease of presentation, let us take $\delta = 1/3$ for the remainder of this paper. Clearly, this probability can be then boosted to arbitrary δ's at a cost of a factor of $O(\log 1/\delta)$ in the sample complexity.

1. Full version appears as arXiv preprint arXiv:1612.03164, v2 Daskalakis and Pan (2016).

Both goodness-of-fit and identity testing have received tremendous attention in Statistics. In the above formulation of these problems, they have received a fair amount of attention over the last decade in both Theoretical Computer Science and Information Theory; see e.g. Batu et al. (2001, 2004); Paninski (2008); Valiant and Valiant (2014); Acharya et al. (2015); Canonne et al. (2016a) and their references. Despite intense research, the high-dimensional (large n) version of the problems has received much smaller attention Batu et al. (2001); Alon et al. (2007); Rubinfeld and Xie (2010); Bhattacharyya et al. (2011); Acharya et al. (2015), despite its importance for applications. In part, this is due to the fact that the problem, as stated above, is hopeless for large n. For example, if Q is the uniform distribution over $\{0,1\}^n$, it is known that $\Theta(2^{n/2}/\epsilon^2)$ samples are necessary (and sufficient) for goodness-of-fit testing Batu et al. (2001); Paninski (2008); Valiant and Valiant (2014).

Our goal in this paper is to leverage combinatorial structure in the specification of P and Q to get around these exponential lower bounds. We are motivated by prior work of Daskalakis, Dikkala and Kamath Daskalakis et al. (2016), which initiated the study of testing problems for structured distributions. They considered testing problems for Ising models, showing that goodness-of-fit and independence testing (testing if an Ising model is a product measure over $\{0,1\}^n$) can be solved efficiently from $\mathrm{poly}(n/\epsilon)$ samples. Their bounds hold for Ising models defined on arbitrary graphs, and for the stronger notion of symmetric Kullback-Leibler divergence (which upper bounds (the square of) total variation distance). In particular, their results are able to side-step the afore-described exponential lower bounds for a broad and important class of probability distributions.

Motivated by this recent work on the Ising model, in this paper we study testing problems on *Bayesian networks*, which is a versatile and widely used probabilistic framework for modeling high-dimensional distributions with structure. A Bayesian network specifies a probability distribution in terms of a DAG G whose nodes V are random variables taking values in some alphabet Σ. To describe the probability distribution, one specifies conditional probabilities $P_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v})$, for all vertices v in G, and configurations $x_v \in \Sigma$ and $x_{\Pi_v} \in \Sigma^{\Pi_v}$, where Π_v represents the set of parents of v in G, taken to be $\emptyset$ if v has no parents. In terms of these conditional probabilities, a probability distribution over Σ^V is defined as follows:

$$P(x) = \prod_v P_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}), \text{for all } x \in \Sigma^V. \tag{1}$$

A special case of a Bayesian network is, of course, a Markov chain, where the graph G is a directed line graph. But Bayesian networks are much more versatile and are in fact universal. They can interpolate between product measures and arbitrary distributions over Σ^V as the DAG becomes denser and denser. Because of their versatility they have found myriad applications in diverse fields of application and study, ranging from probability theory to engineering, computational biology, and law. Our goal is to determine whether goodness-of-fit and identity for these fundamental distributions are actually testable. To achieve this, we develop a deeper understanding into the statistical distance between Bayesian networks.

Results and Techniques. Given sample access to two Bayes nets P and Q on n variables taking values in some set Σ, we would like to decide whether $P = Q$ vs $\delta(P,Q) \geq \epsilon$, where $\delta(P,Q)$ denotes the total variation distance between P and Q. To build intuition, suppose that P and Q are defined on the same DAG, and Q is given. Our goal is to test the equality of P and Q, with fewer than $O(|\Sigma|^{n/2}/\epsilon^2)$ samples required by standard methods, by exploiting the structure of the DAG.

A natural way to exploit the structure of the DAG is to "localize the distance" between P and Q. It's easy to prove that the total variation distance between P and Q can be bounded as follows:

$$\delta(P,Q) \leq \sum_v \delta(P_{\{v\}\cup\Pi_v}, Q_{\{v\}\cup\Pi_v}) + \sum_v \delta(P_{\Pi_v}, Q_{\Pi_v}),$$

where, as above, Π_v denotes the parents of v in the DAG, if any. The sub-additivity of total variation distance with respect to the neighborhoods of the DAG allows us to distinguish between $P = Q$ and $\delta(P,Q) \geq \epsilon$ by running n tests, distinguishing $P_{\{v\}\cup\Pi_v} = Q_{\{v\}\cup\Pi_v}$ vs $\delta(P_{\{v\}\cup\Pi_v}, Q_{\{v\}\cup\Pi_v}) \geq \epsilon/2n$, for all v. We output "$P = Q$" if and only if all these tests output equality. Importantly the distributions $P_{\{v\}\cup\Pi_v}$ and $Q_{\{v\}\cup\Pi_v}$ are supported on $|\{v\}\cup\Pi_v|$ variables. Hence if our DAG has maximum in-degree d, each of these tests requires $O(|\Sigma|^{(d+1)/2}n^2/\epsilon^2)$ samples. An extra $O(\log n)$ factor in the sample complexity can guarantee that each test succeeds with probability at least $1 - 1/3n$, hence all tests succeed simultaneously with probability at least $2/3$. Unfortunately the quadratic dependence of the sample complexity on n is sub-optimal.

A natural approach to improve the sample complexity is to consider instead the *Kullback-Leibler divergence* between P and Q. Pinsker's inequality gives us that $\mathrm{KL}(P||Q) \geq 2\delta^2(P,Q)$. Hence, $\mathrm{KL}(P||Q) = 0$, if $P = Q$, while $\mathrm{KL}(P||Q) \geq 2\epsilon^2$, if $\delta(P,Q) \geq \epsilon$. Moreover, we can now exploit the chain rule of the Kullback-Leibler divergence to localize the distance. Hence, to distinguish $P = Q$ vs $\delta(P,Q) \geq \epsilon$ it suffices to run n tests, distinguishing $P_{\{v\}\cup\Pi_v} = Q_{\{v\}\cup\Pi_v}$ vs $\mathrm{KL}(P_{\{v\}\cup\Pi_v}||Q_{\{v\}\cup\Pi_v}) \geq 2\epsilon^2/n$, for all v. We output "$P = Q$" if and only if all these tests output equality. Unfortunately, goodness-of-fit with respect to the Kullback-Leibler divergence requires infinitely many samples. On the other hand, if every element in the support of $Q_{\{v\}\cup\Pi_v}$ has probability $\Omega\left(\frac{\epsilon^2/n}{|\Sigma|^{d+1}}\right)$, it follows from the χ^2-test of Acharya et al. (2015) that $P_{\{v\}\cup\Pi_v} = Q_{\{v\}\cup\Pi_v}$ vs $\mathrm{KL}(P_{\{v\}\cup\Pi_v}||Q_{\{v\}\cup\Pi_v}) \geq 2\epsilon^2/n$ can be distinguished from $O(|\Sigma|^{\frac{d+1}{2}}n/\epsilon^2)$ samples. An extra $O(\log n)$ factor in the sample complexity can guarantee that each test succeeds with probability at least $1 - 1/3n$, hence all tests succeed simultaneously with probability at least $2/3$. So we managed to improve the sample complexity by a factor of n. This requires, however, preprocessing the Bayes-nets so that there are no low-probability elements in the supports of the marginals. We do not know how to do this pre-processing unfortunately.

So, to summarize, total variation distance is subadditive in the neighborhoods of the DAG, resulting in $O(n^2/\epsilon^2)$ sample complexity. Kullback-Leibler is also subadditive and importantly bounds the square of total variation distance. This is a key to a $O(n/\epsilon^2)$ sample complexity, but it requires no low probability elements in the support of the marginals, which we do not know how to enforce. Looking for a middle ground to address these issues, we study *Hellinger distance*, which relates to total variation distance and Kullback-Leibler as follows:

$$\delta(P,Q) \leq \sqrt{2}\cdot H(P,Q) \leq \sqrt{\mathrm{KL}(P||Q)}.$$

One of our main technical contributions is to show that the square Hellinger distance between two Bayesian networks on the same DAG is subadditive on the neighborhoods, namely:

Theorem 1 (Square Hellinger Subadditivity) *Suppose P and Q are Bayesian networks with the same underlying DAG G, i.e. both factorize as in (1). Then,*

$$H^2(P,Q) \leq \sum_v H^2(P_{\{v\}\cup\Pi_v}, Q_{\{v\}\cup\Pi_v}),$$

where Π_v denotes the set of parents of v in G, if any.

The above bound, given as Corollary 7 in the full version, follows from a slightly more general statement given there in Section 2 as Theorem 2. Given the sub-additivity of the Hellinger distance and its relation to total variation, we can follow the same rationale as above to localize the distance. Hence, to distinguish $P = Q$ vs $\delta(P, Q) \geq \epsilon$ it suffices to run n tests, distinguishing $P_{\{v\}\cup\Pi_v} = Q_{\{v\}\cup\Pi_v}$ vs $H^2(P_{\{v\}\cup\Pi_v}, Q_{\{v\}\cup\Pi_v}) \geq \epsilon^2/2n$, for all v. Importantly goodness-of-fit testing with respect to the square Hellinger distance can be performed from $O(n/\epsilon^2)$ samples. This is the key to our testing results.

While we presented our intuition for goodness-of-fit testing and when the structure of the Bayes-nets is known, we actually do not need to know the structure and can handle sample access to both distributions. Our results are summarized below. All results below hold if we replace total variation with Hellinger distance.

- **Testing Result 1:** Given sample access to two Bayes-nets P, Q on the same but unknown structure of maximum in-degree d, $\tilde{O}(|\Sigma|^{3/4(d+1)} \cdot \frac{n}{\epsilon^2})$ samples suffice to test $P = Q$ vs $\delta(P, Q) \geq \epsilon$. See Theorem 12 of the full version. The running time is quasi-linear in the sample size times $O(n^{d+1})$. The dependence of our sample complexity on n and ϵ is tight in this case, as shown by Daskalakis et al. (2016). If the DAG is known, the running time is quasi-linear in the sample size times $O(n)$.

- **Testing Result 2:** Given sample access to two Bayes-nets P, Q on possibly different and unknown trees, $\tilde{O}(|\Sigma|^{4.5} \cdot \frac{n}{\epsilon^2})$ samples suffice to test $P = Q$ vs $\delta(P, Q) \geq \epsilon$. See Theorem 13 of the full version. The running time is quasi-linear in the sample size times $O(n^6)$. The dependence of our sample complexity on n and ϵ is optimal up to logarithmic factors, even when one of the two distributions is given explicitly, appealing to the same result of Daskalakis et al. (2016) mentioned above.

  Proving this result presents the additional analytical difficulty that two Bayes-nets on different trees have different factorizations, hence it is unclear if their square Hellinger distance can be localized to subsets of nodes involving a small number of variables. In Section 3 of the full version, we prove that given any pair of tree-structured Bayes-nets P and Q, there exists a common factorization of P and Q so that every factor involves up to 6 variables. This implies a useful subadditivity bound for square Hellinger distance into n subsets of 6 nodes. See Theorem 10, and the underlying combinatorial lemma, Lemma 9 of the full version.

- **Testing Result 3:** Finally, our results above were ultimately based on localizing the distance between two Bayes-nets on neighborhoods of small size, as dictated by the Bayes-net structure. As we have already mentioned, even if the Bayes-nets are known to be trees, and one of the Bayes-nets is given explicitly, $O(n/\epsilon^2)$ samples are necessary. Pushing the simplicitly of the problem to the extreme, we consider the case where both P and Q are Bayes-nets on the empty graph, Q is given, and $\Sigma = \{0, 1\}$. Using a non-localizing test, we show that the identity of P and Q can be tested from $O(\sqrt{n}/\epsilon^2)$ samples, which is optimal up to constant factors, as shown by Daskalakis et al. (2016). See Theorem 14 of the full version.

  The proof of this theorem also exploits the subadditivity of the square Hellinger distance. Suppose $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$ are the expectations of the marginals of P and Q on the different coordinates, and without loss of generality suppose that $q_i \leq \frac{1}{2}$, for all i. We use the subadditivity of square Hellinger to show that, if $\delta(P, Q) \geq \epsilon$, then $\sum_i \frac{(p_i - q_i)^2}{q_i} \geq \epsilon^2/2$.

Noticing that $\sum_i \frac{(p_i-q_i)^2}{q_i}$ is an identical expression to the χ^2 divergence applied to vectors $(p_1,\ldots,p_n)$ and $(q_1,\ldots,q_n)$, we reduce the problem to a χ^2-test, mimicking the approach of Acharya et al. (2015). We only need to be careful that $\sum_i p_i$ and $\sum_i q_i$ do not necessarily equal 1, but this does not create any issues.

Learning vs Testing. A natural approach to testing the equality between two Bayes-nets P and Q is to first use samples from P and Q to learn Bayes nets $\hat{P}$ and $\hat{Q}$ that are respectively close to P and Q, then compare $\hat{P}$ and $\hat{Q}$ offline, i.e. without drawing further samples from P and Q. While this approach has been used successfully for single-dimensional hypothesis testing, see e.g. Acharya et al. (2015), it presents analytical and computational difficulties in the high-dimensional regime. While learning of Bayes nets has been a topic of intense research, including the celebrated Chow-Liu algorithm for tree-structured Bayes-nets Chow and Liu (1968), we are aware of no computationally efficient algorithms that operate with $\tilde{O}(n/\epsilon^2)$ samples without assumptions. In particular, using net-based techniques Devroye and Lugosi (2001); Daskalakis and Kamath (2014); Acharya et al. (2014), standard calculations show that any Bayes-net on n variables and maximum in-degree d can be learned from $\tilde{O}(\frac{n\cdot|\Sigma|^d}{\epsilon^2})$ samples, but this algorithm is highly-inefficient computationally (exponential in n). Our algorithms are both efficient, and beat the sample complexity of this inefficient algorithm. On the efficient algorithms front, we are only aware of efficient algorithms that provide guarantees when the number of samples is $>> \frac{n\cdot|\Sigma|^d}{\epsilon^2}$ or that place assumptions on the parameters or the structure of the Bayes-net to be able to learn it (see e.g. Anandkumar et al. (2012); Bresler (2015) and their references), even when the structure is a tree Chow and Liu (1968). Our algorithms do not need any assumptions on the parameters or the structure of the Bayes-net.

Comparison to Canonne et al. (2016b). Testing problems on Bayes-nets similar to the ones we study here are also considered, independently and contemporaneously, in Canonne et al. (2016b) for binary alphabets. However, their emphasis is quite different. Instead of trying to formulate efficient sample-optimal algorithms that work for all cases, they try to identify assumptions under which testing can be done from sub-linear in the number of nodes, n, samples. For example, for the goodness-of-fit problem where the structure of the unknown distribution P is known to be the same as that of the known distribution Q, they require that, for all nodes v, the conditional probability of v taking any value, conditioning on any assignment to the parent nodes, be bounded away from 0 by at least $\Omega(1/\sqrt{n})$. They also assume that the parents of every node attain any configuration with probability at least $\Omega(1/\sqrt{n})$. (See Definition 5.1 and Theorem 5.2 in Canonne et al. (2016b).) In the case the structure of P is unknown, they further assume that its structure contains less edges than that of Q. Moreover, they need that Q satisfies a "γ-non-degeneracy" condition, which, roughly speaking, prohibits Q from being close to any distribution that satisfies additional conditional independence beyond that already implied by Q's graphical structure. (See Definition 6.10 and Theorem 6.11 in Canonne et al. (2016b).) In contrast, we do not make any assumptions, and obtain optimal results for our unconditional testing questions. As we show the testing complexity depends linearly in n.

References

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Sorting with adversarial comparators and application to density estimation. In *Proceedings of the 2014 IEEE*

International Symposium on Information Theory, ISIT '14, pages 1682–1686, Washington, DC, USA, 2014. IEEE Computer Society.

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 3577–3598. Curran Associates, Inc., 2015.

Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing (STOC)*, 2007.

Animashree Anandkumar, Vincent YF Tan, Furong Huang, and Alan S Willsky. High-dimensional structure estimation in ising models: Local separation criterion. *The Annals of Statistics*, pages 1346–1375, 2012.

Tugkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of FOCS*, 2001.

Tuğkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the 36th Annual ACM Symposium on the Theory of Computing*, STOC '04, New York, NY, USA, 2004. ACM.

Arnab Bhattacharyya, Eldar Fischer, Ronitt Rubinfeld, and Paul Valiant. Testing monotonicity of distributions over general partial orders. In *Innovations in Computer Science (ICS)*, 2011.

Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15, pages 771–782, New York, NY, USA, 2015. ACM.

Clément Canonne, Ilias Diakonikolas, Themistoklis Gouleakis, and Ronitt Rubinfeld. Testing Shape Restrictions of Discrete Distributions. In *the 33rd International Symposium on Theoretical Aspects of Computer Science (STACS)*, 2016a.

Clement Canonne, Ilias Diakonikolas, Daniel Kane, and Alistair Stewart. Testing bayesian networks. *arXiv preprint arXiv:1612.03156*, 2016b.

CK Chow and CN Liu. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467, 1968.

Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians. In *Proceedings of the 27th Annual Conference on Learning Theory*, COLT '14, pages 1183–1213, 2014.

Constantinos Daskalakis and Qinxuan Pan. Square Hellinger Subadditivity for Bayesian Networks and its Applications to Identity Testing. *arXiv*, 2016. http://arxiv.org/abs/1612.03164.

Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing Ising Models. *arXiv*, 2016.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer, 2001.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *Information Theory, IEEE Transactions on*, 54(10):4750–4755, 2008.

Ronitt Rubinfeld and Ning Xie. Testing non-uniform k-wise independent distributions over product spaces. In *the 37th International Colloquium on Automata, Languages and Programming (ICALP)*, 2010.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 51–60, Washington, DC, USA, 2014. IEEE Computer Society.