RC22558 (W0209-042) September 11, 2002
Computer Science

IBM Research Report

Approximability and the Effective Width of Probability Distributions

Alina Beygelzimer
Department of Computer Science
University ofRochester
Rochester, NY 14627

Irina Rish
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Approximability and the Effective Width of Probability Distributions

Alina Beygelzimer
Department of Computer Science
University of Rochester
Rochester, NY 14627
Email: beygel@cs.rochester.edu

Irina Rish
IBM T. J. Watson Research Center
Hawthorne, NY 10532
Email: rish@us.ibm.com

Abstract

Our work is motivated by the goal of learning probabilistic graphical models that are both accurate and efficient for inference. In this paper, we investigate the "degree of approximability" of joint probability distributions by introducing the parameter $k(\delta)$, called the *effective treewidth*. This parameter captures the tradeoff between the accuracy of approximation δ, measured as the information divergence from the true distribution, and model's inference complexity, which is exponential in its treewidth k. We show that both treewidth and information divergence exhibit a threshold behavior. The relative location of such thresholds is an inherent property of the underlying distribution. Finally, we propose an efficient sampling algorithm for estimating these thresholds from data, thereby predicting the effective treewidth (the approximability) of the underlying distribution. This provides a principled approach to model selection when learning bounded-treewidth models.

1 Introduction

This paper is motivated by the goal of learning accurate probabilistic graphical models that also yield *efficient inference*, instead of learning models first without consideration of their inference complexity, and then coping with this complexity via approximate inference methods. Since we usually have only a finite set of samples from some unknown distribution, the best we can do is to learn is an approximation of this distribution. Thus it seems only natural to prefer approximations that are more efficient for inference, as the ability to perform inference is one of the main reasons for building probabilistic models in the first place. In [3], both theoretical and empirical arguments where provided suggesting that learning algorithms should incorporate some measure of the inference complexity as a part of their model selection criteria. It was shown that models representing statistically close distributions may have quite different graph structures, in particular, a large difference in treewidth which leads to an exponential difference in their inference complexity (since traditional exact inference algorithms on graphical models are exponential in the graph's treewidth). Thus, the following questions arise: If we tolerate a certain inaccuracy in our model, what is the best inference complexity we can hope to achieve? Or, what is the best achievable approximation accuracy given a constraint on the complexity of inference (e.g., a bound on the treewidth)? Formally, we wish to solve the following variational problem:

$$\min_{\hat{P}} D(P||\hat{P}) + \lambda \,\mathrm{width}(\hat{P}),$$

where P is our target distribution, $D(P||\hat{P})$ is the information divergence of the approximation $\hat{P}$ from this target P, and $\mathrm{width}(\hat{P})$ denotes the treewidth of a graphical model encoding $\hat{P}$. Here λ is a Lagrange multiplier for the treewidth. While we are not providing an analytical solution to this problem here, we will be able to determine achievable values of treewidth given the constraint $D(P||P') \leq \delta$. We investigate the "degree of approximability" of probability distributions using a new parameter, called the *effective treewidth*, that captures the tradeoff between the approximation accuracy and inference complexity.

The notion of effective treewidth is formalized in a way similar to the rate-distortion theory in lossy source compression. We call a pair (k, δ)-*achievable* for a distribution P, if there exists a distribution with information divergence from P at most δ that is decomposable on a graph with treewidth at most k. In this case P is said to *admit* treewidth $k(\delta)$. Alternatively, P is said to be $\delta(k)$-*approximable*. The *effective treewidth* of P, in respect to a given δ, is defined to be the largest $k(\delta)$ such that the pair (k, δ) is achievable, i.e. if all distributions at distance at most δ from P are not decomposable on graphs with treewidth less than $k(\delta)$. This formulation gives the level of inevitable complexity (i.e. treewidth) k, given the desired closeness (i.e. sum of clique weights) δ.

We will also be interested in the "average-case" analogs of these definitions. Namely we will say that a pair (k, δ) is ϵ-achievable for P if at least an ϵ-fraction of graphs with treewidth k certify that (k, δ) is achievable. The other notions are defined similarly. Thus we not only care about the *existence* of an approximation with given δ and k, but also in the *number* of such approximations.

Our formulations and results are largely motivated by the threshold behavior found in monotone graph properties (see, for example, [9, 10, 4]). Namely, the probability that a random graph with a given edge probability satisfies such property jumps from zero to one around some critical value of edge probability, and this transition from the property being very unlikely to the property being very likely is very sharp. Both the property of "having treewidth at most some fixed integer k", and the property of "having information divergence from the target distribution at most some fixed value δ" have such thresholds. Observe that the first property is independent of the target distribution. Now, given the desired closeness to the target distribution, δ, we want to find the smallest k such that the probability that the learned distribution is at distance at most δ (or rather the sum of clique weights is at least δ) is still very high (i.e., has not reached its critical value), yet at the same time the probability that models with this edge density have treewidth at most k is *already* high. In other words, we want to restrict ourselves to a class of treewidth-k bounded models, such that the models in the class are *still* simple, yet this class *already* contains a sufficiently good approximation to the target distribution. We formalize the problem and give a polynomial-time algorithm for finding k. In the average-case (parameterized by ϵ), the class not only contains a good approximation, but at least an ϵ-fraction of approximations is good. Thus there is a simple probabilistic algorithm for finding one; indeed, we just need to sample $1/\epsilon$ distributions in the class and choose the closest one.

The idea of our algorithm is very simple: we sample random graphs in order to approximate the information divergence threshold $p_c = p_c(\delta)$. Then, for each k, we estimate the probability that a random graph in $G(n, p_c)$ has treewidth at most k, where $G(n, p_c)$ is the probability space of graphs with n nodes and edge probability p_c. Based on these estimates, we either suggest an appropriate family of bounded-treewidth graphs as a search space of candidate models, or say that (with high probability) the target distribution is inherently complex and cannot be adequately approximated using low-treewidth models.

This paper makes the following contributions. First, we introduce a new parameter, effective treewidth, that captures the "approximability" of a distribution as a trade-off between approximation accuracy (measured by the information divergence from the true distribution) and efficiency of the resulting model (measured by its treewidth). We show that both treewidth and information divergence exhibit a threshold behavior when . The relative position of such thresholds is an inherent property of the underlying distribution. Finally, we propose an efficient algorithm for estimating these thresholds from data, thereby predicting the effective treewidth (the approximability) of the underlying distribution. This provides a principled approach to model selection when learning bounded-treewidth approximations, which has recently been an active research area [5, 14, 2] since the famous paper by Chow and Liu [5], who showed that the best tree projection of an arbitrary distribution can be found efficiently.

2 Preliminaries

Let P be a probability distribution on n discrete random variables $X_1, X_2, \ldots, X_n$, and let G be an undirected graph on a set of n vertices corresponding to the random variables in P. We will say that P is *decomposable* (also *factorizable*) in G, if P can be factored over the cliques of G, i.e. expressed as a product of clique potentials,

$$P(x_1, ..., x_n) = \prod_C \phi(x_C),$$

where the potential $\phi(x_C)$ depends only on the values of the variables in X_C. Here C ranges over all maximal cliques in G, and X_C denotes the set of variables corresponing to the nodes in C. For example, a tree-decomposable distribution P can be written as $P(x_1, ..., x_n) = \prod_{i,j} \phi(x_i, x_j)$, where the product is taken over all pairs of nodes (x_i, x_j) that correspond to an edge (i.e., a maximal clique) in the tree. A graph with an associated set of potential functions is called a probabilistic graphical model.

Probabilistic graphical models (e.g., Markov and Bayesian networks) are used for answering queries such as finding the posterior probability of an unknown set of variables given evidence (i.e., values of some other variables), or finding the most-likely variable assignment. However, (exact) probabilistic inference in graphical models is NP-hard [6]. Commonly used inference algorithms, such as the *junction tree* algorithm [13], or closely related *variable-elimination* techniques (e.g., [8]), are exponential in the size (i.e., dimensionality) of largest dependency created by the algorithm, which is equivalent to the size of largest clique (minus 1) induced in the graph. This graph parameter is called the *treewidth*, and is formally defined as follows. Given a graph G, the *width* of variable X_i along ordering o is the number of X_i's neighbors preceding X_i in o. The *width of G along o* is the maximum width along o. The *induced graph* of G along o is obtained by connecting the preceding neighbors of each X_i, going from $i = n$ to $i = 1$. The treewidth (induced width) along o, is the width of the induced graph along o, while the treewidth (induced width) of G is the minimum treewidth over all possible orderings. Another interpretation of the treewidth is the size of the largest clique (minus 1) in a triangulation of a Markov network (or a moralized Bayesian network), or, equivalently, the size (in the number of variables) of the largest function computed during inference.

Given a distribution $P(X)$ and its approximation $\hat{P}(X)$, the *information divergence* ([7]) between of $\hat{P}$ from P is defined as $D(P||\hat{P}) = \sum_x P(x) \log \frac{P(x)}{\hat{P}(x)}$. This information-theoretic measure can be interpreted as the average number of extra bits one has to use to encode a message X drawn from distribution $P(X)$ if a "wrong" distribution $\hat{P}(X)$ is assumed instead of $P(X)$. Note that information divergence is not a metric as it is not necessarily symmetric.

Given a set of independent samples from some unknown probability distribution P on n discrete random variables $X_1, X_2, \ldots, X_n$, the general goal is to learn a model of this distribution that involves dependencies only on limited subsets of the variables. Restricting the size of dependencies controls both overfitting and the complexity of inference in the resulting model. Clearly, there is a tradeoff between the complexity of the model and the accuracy it provides. Thus in order to control the dependency structure of the learned approximation, it seems appropriate to consider a parameterized family of classes of distributions decomposable on graphs of treewidth at most k, with $k = 1$ corresponding to the set of tree-decomposable distributions. We will denote the class of treewidth-k decomposable distributions by $\mathcal{D}_k$.

Given a restricted class of distributions $\mathcal{D}$, one naturally wants to find the distribution within $\mathcal{D}$ minimizing the information divergence from the target distribution P, which is called the *projection* of P onto $\mathcal{D}$, and which is also the distribution within the class that maximizes the probability of observing the data. For each k, let $\hat{P}_k$ denote a projection of P onto $\mathcal{D}_k$.

2.1 Learning bounded-treewidth models

Chow and Liu [5] showed that the projection onto the set of trees takes a particularly simple form. They showed that for a specific tree T, the projection of P onto the set of Markov networks factorizable over T is uniquely given by the distribution in which marginal probabilities along the edges of T coincide with those computed from P. Due to the product form of tree-decomposable distributions, the formula for the information divergence decomposes into a term that depends only on the true distribution P (and thus is the same for all approximations; in fact, it is just the information contained in P), and a term that additively decomposes to the edges of T, and is maximized by the distribution whose marginals along the edges agree with those of computed from P. (Since for a fixed structure G, the projection onto the set of distributions over G is uniquely defined, we will identify G with this projection, ignoring some abuse of terminology.) Due to this additive decomposition, the tree yielding the closest (in terms of information divergence) projection is simply given by any maximum weight spanning tree, where the edge weights are the mutual informations between the corresponding variables. Notice that candidate spanning trees can be compared without any knowledge of P beyond that given by pairwise statistics.

Srebro [14] showed that a similar decomposition holds for k-width bounded Markov networks. For a fixed triangulated graph G, the projection of P onto G is given by the distribution that agrees with P on all clique marginals. The projection itself is then just the product (over all, not necessarily maximal, cliques) of the clique factors. The main problem is of course to find a graph with treewidth k giving the closest projection among all treewidth-k graphs. Since adding edges can only bring the distribution closer to the target, all graphs are assumed to be already triangulated. Srebro reduced this problem to the problem of finding the maximum weight hypertree, and showed that although finding the best tree projection is computationally simple, it is NP-hard to find projections for any $k \geq 2$. The main technical contribution of [14, 11] is an approximation algorithm for the problem. For a fixed width k, the algorithm finds a treewidth-k spanning subgraph of a given n-node graph such that the weight of this subgraph is within some k-dependent factor from the maximum. The time complexity is $n^{O(k)}$.

It is important to note that Srebro [14] considered approximation in the context of density estimation, and not model selection. The task there is to approximate a distribution, and the quality of the approximating distribution is determined by its closeness to the true one (in terms of its information divergence). Thus the choice of k is directly driven by the size of the sample space – the only rationale for limiting the class of hypothesis distributions is to prevent overfitting. With an infinite amount of data, they would learn a clique, since adding edges would always decrease the divergence. The number of available samples, however, implicitly puts an information theoretic lower bound on the treewidth of the learned approximation. Indeed, the information about the distribution comes only from the sample, thus the length of the minimum representation of the disitrbution is upper bounded by the sample size, which implicitely puts a bound on the treewidth.

We, on the other hand, are concerned with finding the width that is *characteristic* of the target distribution. In other words, given a distribution, the goal is to find the most appropriate treewidth-bounded class $\mathcal{D}_k$ onto which to project this distribution. We introduce the notion of *effective treewidth* which attempts to capture this intuition. The treewidth is the worst-case measure of the complexity of a distribution in the sense that it is totally oblivious to the actual probability values of the distribution on top of the graph (i.e., all distributions consistent with the same set of independence assumptions have the same treewidth). Clearly, the strength of probabilistic dependencies among variables may be quite different for different distributions over the same graph, and what we need is a more refined measure that captures the effective width of the target distribution given a desired level of divergence from it.

2.2 Threshold behavior of random graphs

Consider the probability space $G(n,p)$ of random graphs on n nodes with edge probability p (i.e., every pair of nodes is connected with probability p, independently of every other pair). Let $G_{n,p}$ stand for a random graph from this probability space. We will also sometimes use $G_{n,m}$ to denote a random graph with n nodes and m edges. When $p = m/\binom{n}{2}$, the two models are are practically identical [4]. A *graph property* is just a subset of graphs closed under isomorphism. A property is *monotone increasing* if it is preserved under edge addition; that is, if a graph G satisfies the propery, then every graph on the same set of nodes containing G as a subgraph must satisfy it as well. For a property $\mathcal{P}$ with a fixed set of n nodes, let $\mu_p(\mathcal{P})$ denote the probability that $G_{n,p}$ satisfies $\mathcal{P}$. We say that almost every (a.e.) $G_{n,p}$ has the property $\mathcal{P}$ if $\mu_p(\mathcal{P})$ converges to 1 as n approaches infinity. Of course, we can analogously define a monotone *decreasing* property and the corresponding definition of a.e. graph satisfying it. For example, the property of having treewidth at most k is monotone decreasing. The direction of monotonicity will be clear from the context.

The theory of random graphs was founded by Erdős and Rényi [9], and one of the main observations they made was that many natural monotone properties appear rather suddenly, i.e., as we increase p, there is sharp transition from a property being very unlikely to it being very likely in $G(n,p)$. Friedgut [10] proved that *every* monotone graph property has such a threshold behavior.

Given a parameter $\epsilon > 0$ and a monotone property $\mathcal{P}$, let p_0 be given by $\mu_{p_0}(\mathcal{P}) = \epsilon$, and p_1 be such that $\mu_{p_1}(\mathcal{P}) = 1 - \epsilon$; in other words, ϵ quantifies what it means for $\mathcal{P}$ to "be likely". The length α of the threshold interval is defined as $p_1 - p_0$. The threshold interval contains the critical value of p, denoted p_c, such that $\mu_{p_c}(\mathcal{P}) = 1/2$, thus $\mu_p(\mathcal{P})$ jumps from nearly 0 to nearly 1 around $p = p_c$. Both the length of the interval $\alpha(n)$ and the threshold function $p_c(n)$ shrink with n. The relative rate at which they do so determines the sharpness of the threshold. Friedgut [10] gave an upper bound on the length of the threshold interval for any monotone graph property, $\alpha(n) \leq O(\log(1/\epsilon)/\log n)$.

3 Formulation of the effective width

Recall that we identify each graph with the projection of the target distribution onto the graph. We call a pair (k,δ)-*achievable* for a distribution P, if there exists a treewidth-k graph decomposable distribution whose information divergence from P is at most δ. In this case P is said to *admit* treewidth $k(\delta)$. Alternatively, P is said to be $\delta(k)$-*approximable* (as if its projection onto $\mathcal{D}_k$ is at distance at most δ). The *effective treewidth* of P, in respect to a given δ, is defined as the largest $k(\delta)$ such that the pair (k,δ) is achievable, i.e. if all distributions at distance at most δ from P are not decomposable on graphs with treewidth less than $k(\delta)$. This formulation gives the level of inevitable complexity (i.e. treewidth) k, given the desired closeness (i.e. sum of clique weights) δ.

We will also be interested in the "average-case" analogs of these definitions. Namely we will say that a pair (k,δ) is ϵ-achievable for P if at least $\epsilon|\mathcal{D}_k|$ graphs in $\mathcal{D}_k$ certify that (k,δ) is achievable. Similarly, P is $\delta_\epsilon(k)$-*approximable* if its projection onto a random graph in $\mathcal{D}_k$ is at distance at most δ from P with probability at least ϵ. The effective treewidth $k_\epsilon(\delta)$ of P is defined similarly. Thus we not only care about the existence of an approximation with a given δ and k, but also in the *number* of such approximations.

4 Two properties of distributions

In this section we introduce and discuss two graph properties essential for the rest of the paper.

Property $\mathcal{P}_\delta$ Using Srebro's decomposition [14], the information divergence of (a set of distributions over) a given graph G from the target distribution P can be expressed as

$$D(P\|G) = D(P\|\emptyset) - W(G),$$

where $W(G)$ is the sum of weights [1] over all (not necessarily maximal) cliques of G, not including singleton cliques. Singletons are accounted for in the divergence from the empty graph (the first term), which is the same for all approximations, and can be ignored in the quest for one minimizing the divergence from P. Hence the best approximation is the one maximizing $W(G)$. Notice that $W(G)$ can be determined without any knowledge of the actual distribution beyond its low order marginals necessary to compute clique potentials.

Fix a target distribution P and some $\delta > 0$, and consider the property of n-node graphs G of having $W(G) \geq \delta$. Call this property $\mathcal{P}_\delta$. Notice that this is not a *graph* property, because it is not necessarily closed under automorphisms (relabellings) of G. However, it can easily be made a graph property without increasing the size of G more than polynomially via the following transformation: Substitute every node $i \in \{1, \ldots, n\}$ by a clique on $(n+i)$ nodes (replacing every edge (i,j) by $(n+i)(n+j)$ new edges), and appropriately scale the weight of every clique. More precisely, if G' is the modified G, let $W(G') = \sum_C [w(C)/\prod_{j\in C} l(j)]$, where $l(j)$ is the multiplicity of node j in G', and the sum is over all non-singleton cliques of size at most n. It is not hard to see that $W(G) = W(G')$.

Observe that $\mathcal{P}_\delta$ is monotone increasing. Adding edges to a graph can only bring the graph closer to the true distribution, since any distribution decomposable on the original graph is also decomposable on the augmented one. Thus if G is a subgraph of G', then $W(G) \geq \delta$ only if $W(G') \geq \delta$.

Property $\mathcal{P}_k$ Fix an integer k, and consider the property of n-node graphs of having treewidth at most k. Call this property $\mathcal{P}_k$ and observe that it is a structural property of a graph, which does *not* depend on the target distribution and its projection onto the graph. It is also a monotone decreasing property, since if a graph has treewidth at most k, then certainly any of its subgraphs does.

When $k = 1$, $\mathcal{P}_k$ is just the property of being acyclic (triangle-free). One of the observations made by Erdős and Rényi [9] was that both the crtical p and the length of the threshold interval for $k = 1$ are approximately $1/n$. Kloks [12] showed that actually very few graphs have bounded treewidth. In particular, he showed that for any $\tau > 0$, and any $\epsilon < \frac{\tau-1}{\tau+1}$, almost all n-node graphs with edge probability $\sim\tau/n$ have treewidth at least n^ϵ. Furthermore, if τ is slightly larger ($\tau \geq 1.18$), then the treewidth is linear in n. Thus all transitions from k to $k+1$ for k sublinear in n happen in a very narrow initial interval of p, bounded above by $O(1/n)$.

5 An overview of the algorithm

Given a target distribution P and a distance parameter δ, the threshold curve $\mu_p(\mathcal{P}_\delta) = \mathbf{Pr}[W(G_{n,p}) \geq \delta]$ and its critical value p_c^δ are uniquely defined (although unknown). Recall that p_c^δ is given by $\mu_{p_c^\delta}(\mathcal{P}_\delta) = 1/2$. Our goal is to find the smallest treewidth k such that the pair (k, δ) is ϵ-achievable for P for some large enough ϵ to be defined later.

The main idea is very simple. Suppose that we know the critical value p_c^δ. For each width k, let p_c^k denote the critical value of the property $\mathcal{P}_k$ of having treewidth at most k, i.e., p_c^k is such that $\mu_{p_c^k}(\mathcal{P}_k) = 1/2$.

For reasons that will become clear in a moment, we want to find the smallest k such that $p_c^k > p_c^\delta$. In other words, we want the smallest k satisfying

$$\mathbf{Pr}[\text{width}(G_{n,p_c^\delta}) \leq k] > \frac{1}{2}.$$

Let k_c be this k. To find k_c, we will generate a sample of m independent random graphs in G_{n,p_c^δ}, denoted $G_1, \ldots, G_m$, and use it to simultaneously approximate $\mathbf{Pr}[\text{width}(G_{n,p_c^\delta}) \leq k]$ for $k = 1, \ldots, n-1$. More accurately, for each i and k, let U_i^k be an indicator random variable, which is equal to 1 if width$(G_i) \leq k$, and 0 otherwise. The analysis of the sampler in Section 6 says that $\mathbf{E}[U_i^k]$ provides a sufficiently good estimate of $\mathbf{Pr}[\text{width}(G_{n,p_c^\delta}) \leq k]$. Moreover, m (the size of the sample space) is independent of n, it depends only on the accuracy of approximation and the error with which this accuracy is guaranteed. Now k_c is just the smallest k such that $\mathbf{E}[U_i^k] > 1/2$.

We know that at least half of $G(n, p_c^\delta)$ satisfies $\mathcal{P}_\delta$. On the other hand, more than a half of $G(n, p_c^\delta)$ satisfies $\mathcal{P}_{k_c}$. By a trivial counting argument, there *must* exist graphs in $G(n, p_c^\delta)$ satisfying both. In fact, many

[1] The weight of clique S is defined by $w(S) = \mathbf{E}_{X\sim P}[\log\phi(X_S)]$, where the clique factor $\phi(X_S)$ is given by $P(X_S)/\prod_{S'\subset S}\phi(X_{S'})$, and the product is over all, not necessarily maximal cliques.

graphs will; more precisely, at least a $(p_c^{k_c} - p_c^{\delta})$-fraction of all graphs in $\mathcal{D}_k$. Clearly we are overcounting, since the same graphs may contribute to both probabilities. However, not absurdly, since the graphs in $G(n, p_c^{\delta})$ with treewidth bounded by k are (at least intuitively) are not likely to be the onces closest to the target distribution, since the remaining graphs have larger treewidth and thus are more likely to fit the distribution better. Although, as shown in [3], a difference in treewidth does not necessarily imply a difference in the information divergence.

The only problem is that we do not actually know the critical value p_c^{δ}. However, we can estimate it using random sampling. As mentioned in Section 4, when $p > 2\tau/n$ for $\tau \geq 1.18$, almost every $G_{n,p}$ has treewidth already linear in n. On the other hand, almost every $G_{n,1/n}$ is a tree. Suppose that we are given a constant β such that if $p > \beta/n$, then almost every $G_{n,p}$ has treewith that is too large for us to handle. We emphasize that the values related to treewidth do *not* depend on the target distribution, and can in principle be precomputed for a given n. In order to find p_c^{δ}, we will just do a binary search in the interval $[1/n, \beta/n]$. More precisely, during each step of the search with the current edge probability p, we will approximate $\mathbf{Pr}[G_{n,p}]$ using the sampler from Section 6, and based on the estimate decide which half of the current interval to prune. The search is continued until p gets sufficiently close to satisfying $\mu_{\delta}(G_{n,p}) = 1/2$.

It may be useful to note that if we have a precomputed expectation $\mathbf{E}[\text{width}(G_{n,p_c^{\delta}})]$, then Markov's inequality immediately gives an upper bound on k_c. Indeed, we have

$$\mathbf{Pr}[\text{width}(G_{n,p}) \leq k] \geq 1 - \frac{\mathbf{E}[\text{width}(G_{n,p})]}{k}.$$

Thus k_c must be such that $1 - \mathbf{E}[\text{width}(G_{n,p_c^{\delta}})]/k \geq 1/2$, or $k \geq 2\mathbf{E}[\text{width}(G_{n,p_c^{\delta}})]$. The upper bound on k_c is given by the smallest integral k satisfying the above inequality.

Example A small example should help make the goals clear. Consider Figure 1, which shows threshold curves for a 3-wise independent distribution on 8 random variables ($n = 8$). A distribution is k-wise independent if every subset of k variables is mutually independent (however, there may exist dependencies on larger subsets). Thus we could hardly expect treewidth-2 graphs to do well on this distribution, since all triples are independent, and their marginals do not reveal any higher-order structure; as we shall see this is indeed the case. The x-axis in Figure 1 corresponds to the number of edges, i.e. $p\binom{n}{2}$, the y-axis denotes the probability that $G_{n,p}$ satisfies the property corresponding to a given curve. The monotone decreasing curves correspond to the properties $\mathcal{P}_k$ for $k = \{1, \ldots, 6\}$ (from left to right respectively). For $k = 7$, the curve is just $\mu_p(\mathcal{P}_k) = 1$. We emphasize that these curves do *not* depend on the target distribution. The two monotone increasing curves correspond to the property $\mathcal{P}_{\delta}$ for two values of δ such that $\delta_1 > \delta_2$ (the smaller δ, the worse the quality of approximation is, thus the higher the probability of attaing it). As p increases, the probability of having small treewidth decreases, while the probability of getting close to the target increases. We want to find to capture the situation when the first probability is *still* high, while the second is *already* high. Given a level of closeness δ, we formalize this desired situation as follows: we want to find the smallest treewidth k such that the corresponding curves meet at the probability above 1/2. Of course, the algorithm could be parameterized by any probability greater than 1/2, but we opted for the simplicity of presentation. Suppose that δ_1 (corresponding to information divergence 0.1) is our desired level of "closeness" to the true distribution. Then $k(\delta_1) = 3$, which is expected, since the target distribution is 3-wise independent, thus we need at least 4-cliques to approximate it reasonably well. Similarly, $k(\delta_2) = 5$, where δ_2 corresponds to information divergence 0.06, thus we essentially need a clique to get within the allowed divergence.

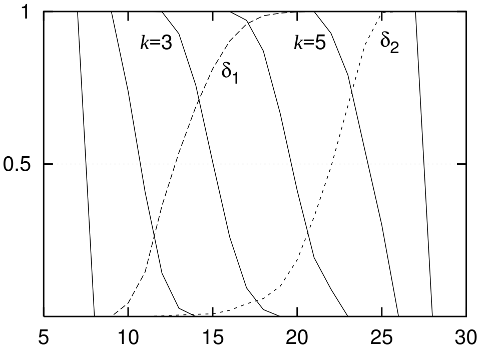

Figure 1: Thresholds for a 3-wise independent distribution on 8 random variables.

6 The sampler

We show that the straightforward sampling is sufficient for our purposes. To estimate $\mathbf{Pr}[W(G_{n,p}) \geq \delta]$ for given p and δ, the algorithm just generates m independent copies of $G_{n,p}$, denoted $G_1, \ldots, G_m$, and outputs $S_m = \frac{1}{m}\sum_{i=1}^{m} U_i$ as its estimate of $\mu = \mathbf{Pr}[W(G_{n,p}) \geq \delta]$, where U_i is a random variable indicating whether $W(G_i) \geq \delta$, and the expectation above is over the uniform choice of $G_{n,p}$. The approximation is parameterized by the error probability γ and accuracy ρ, namely we want to get a value at most ρ away from the true one with probability at least $1-\gamma$ (over the choice of $G_1, \ldots, G_m$). By the Chernoff Bound,

$$\mathbf{Pr}\left[|S_n - \mu| > \rho\right] < 2e^{-2m\rho^2}.$$

Setting $m = \frac{\ln(2/\gamma)}{2\rho^2}$ makes the above error probability at most γ, as required. Thus $O(\frac{\ln(1/\gamma)}{\rho^2})$ samples suffice. The same argument works for estimating the probability that width$(G_n, p) \leq k$, given p and k.

Notice that for estimating the expectation $\mu = \mathbf{E}[W(G_{n,p})]$, the number of samples required by the above algorithm is no longer independent of n. Although the sum of weights is always bounded (which is crucial, since otherwise no reasonable sampling method could guarantee any approximation), this bound is not independent of n. Indeed, for any target distribution P and graph G, we have $W(G) \leq D(P\|\emptyset) \leq n - H(P) \leq n$. The sampler outputs $S_m = \frac{1}{m}\sum_{i=1}^{m} W(G_i)$ as its estimate of μ, and the Chernoff Bound gives $m = \frac{n^2 \ln(2/\gamma)}{2\epsilon^2}$. The same is true for estimating $\mathbf{E}[\text{width}(G_{n,p})]$, since the range of width$(G_{n,p})$ is in general of size linear in n.

7 Putting it together

In this section we describe the algorithm. The input parameters are the following:

(a) an upper bound t on the allowed treewidth (and the corresponding value $\beta \geq 1$ such that almost all graphs with edge probability at least β/n have treewidth greater than t);

(b) clique factors (for subsets of size at most t) determined from the target distribution P; (for a subset S, $|S| \leq t$, its projected factor [14] is given by $w(S) = \mathbf{E}_{X\sim P}[\log \phi(X_S)]$, and can be estimated from the data);

(c) accuracy parameter δ (i.e. for a model to be acceptable, its graph G must satisfy $W(G) \geq \delta$);

(d) error parameter γ (i.e. the learned model must satisfy the guarantees with probability at least $1-\gamma$).

8 Approximating the thresholds

In this section we derive bounds on the threshold function for $\mathcal{P}_\delta$. The main idea is as follows. We will express $W(G_{n,p})$ as a sum of some random variables in such a way that $\mathbf{E}[W(G_{n,p})]$ is easily computable using linearity of expectation. Then by Markov's inequality, $\mathbf{Pr}[W(G_{n,p}) \geq \delta] \leq \mathbf{E}[W(G_{n,p})]/\delta$. To get a lower bound for $\mathcal{P}_\delta$ satisfiability, we want to find p such that the above expectation approaches 0.

Algorithm $(P, \delta, t, \beta, \gamma)$

```
Approximating $p_c^\delta$:
  1. Estimate $\mu = \mathbf{Pr}[W(G_{n,\beta/n}) \geq \delta]$ with accuracy $\rho$ and error $\gamma$ (using the sampler from Section 6). If $\mu + \rho < 1/2$, output "$P$ is can not be $\delta$-approximated by models with treewidth at most $t$".
  2. Otherwise, let $l = 1/n$, $u = \beta/n$. Until the value of $p_c^\delta$ is not set:
     (a) Let $p = (u + l)/2$.
     (b) Estimate $\mu_p = \mathbf{Pr}[W(G_{n,p}) \geq \delta]$ with accuracy $\rho$ and error $\gamma/2$.
     (c) If $|\mu_p - 1/2| \leq \rho$, set $p_c^\delta = p$.
     (d) Else if $\mu_p > 1/2$, set $u = p$; otherwise set $l = p$.
Approximating $k_c$:
  1. For each $k = 1, \ldots, t$ estimate
         $\mu_k = \mathbf{Pr}[\text{width}(G_{n,p_c^\delta}) \leq k]$
     with accuracy $\rho$ and error $\gamma/2$.
  2. Let $k_c$ be the smallest $k$ such that $\mu_k - 2\rho > \mu_{p_c^\delta}$. Set $\epsilon = (\mu_{k_c} - \mu_{p_c^\delta} - 2\rho)$.
  3. Output "$(k_c, \delta)$ is $\epsilon$-achievable for $P$ with probability $\gamma$".
```

For each subset S, let U_S be a binary random variable indicating whether S is a clique. We have $\mathbf{E}[U_S] = p^{\binom{|S|}{2}}$. Then $W(G_{n,p})$ can be written as

$$W(G_{n,p}) = \sum_S w(S)U_S = \sum_l \sum_{S:|S|=l} w(S)U_S,$$

where l (the size of a subset) ranges in $1 < l \leq n$. By linearity of expectation,

$$\mathbf{E}[W(G_{n,p})] = \sum_l \sum_{S:|S|=l} w(S)\mathbf{E}[U_S] = \sum_l p^{\binom{l}{2}} \sum_{S:|S|=l} w(S).$$

The last equality is due to the fact that the clique weights do *not* depend on the graph, they depend only on the target distribution P. Since $p \leq 1$, the expression above yields $\mathbf{E}[W(G_{n,p})] < p\sum_S w(S)$, where the sum is over all non-singleton subsets of the nodes. Recall that $w(S) = \mathbf{E}[\log \phi(X_S)]$. It is not hard to see that the above sum is just $D(P\|\emptyset)$. Indeed, by linearity of expectation (where the expectation is taken in respect to P),

$$\begin{aligned}\sum_S w(S) &= \mathbf{E} \log \prod_S \phi(X_S) \\ &= \mathbf{E} \log \frac{P(X_1, \ldots, X_n)}{\prod_{S' \subset S} \phi(X_{S'})} \prod_{\substack{S' \subset S \\ |S|>1}} \phi(X'_S) \\ &= \mathbf{E} \log \frac{P(X_1, \ldots, X_n)}{P(X_1) \cdots P(X_n)} = D(P \| \emptyset).\end{aligned}$$

Let $D = D(P\|\emptyset)$. Then we have $\mathbf{E}[W(G_{n,p})] < pD$, and by Markov's inequality

$$\mathbf{Pr}[W(G_{n,p}) \geq \delta] \leq \frac{\mathbf{E}[W(G_{n,p})]}{\delta} < \frac{pD}{\delta}.$$

Now if $p \ll \delta/D$, then almost certainly $G_{n,p}$ does not satisfy $\mathcal{P}_\delta$. If $p \sim \frac{\delta}{2D}$, then the above probability μ_p^δ is less than $1/2$. This gives a lower bound on the critical value of p, namely $p_c^\delta > \delta/2D$.

This argument, however, does not give any upper bound on p_c, since the values of p making $\mathbf{E}[W(G_{n,p})] \to \infty$ do not necessarily imply that $G_{n,p}$ satisfies $\mathcal{P}^\delta$ almost always. Such an upper bound, however, can be obtained using the second moment method (see, for example, [1]), which we omit here. Notice that this would also require a *lower* bound on $\mathbf{E}[W(G_{n,p})]$, since we basically want to show that if this expectation goes to infinity and the variance of $W(G_{n,p})$ satisfies certain properties, then $G_{n,p}$ satisfies $\mathcal{P}_\delta$ almost always (from an application of Chebyshev's inequality). The lower bound on the expectation then immediately gives an upper bound on the $\mathcal{P}_\delta$ satisfiability.

9 Conclusions

We introduce a new parameter, effective treewidth, that captures the "approximability" of a distribution as a trade-off between approximation accuracy (measured by the information divergence from the true distribution) and efficiency of the resulting model (measured by its treewidth). We show that both the treewidth and information divergence exhibit a threshold behavior. The relative position of such thresholds is an inherent property of the underlying distribution. Finally, we propose an efficient algorithm for estimating these thresholds from data, thereby predicting the effective treewidth (the approximability) of the underlying distribution. This provides a principled approach to model selection when learning bounded-treewidth approximations, which has recently been an active research area [5, 14, 2] since the famous paper by Chow and Liu [5], who showed that the best tree projection of an arbitrary distribution can be found efficiently.

10 Further directions

A complementary goal to that discussed in the paper is to determine the width *with respect to a query distribution.* Another important avenue of research focuses on problem classes that have particular structure. For example, certain applications often yield probabilistic problems that are characterized by nearly-deterministic (low-entropy) distributions (e.g. in noisy channel coding, which mainly involves deterministic encoding and nearly-deterministic (low-noise transmission) dependencies, and yields an excellent performance of local approximation schemes).

References

[1] Noga Alon and Joel H. Spencer. *The probabilistic method.* Wiley-Interscience [John Wiley & Sons], New York, second edition, 2000. With an appendix on the life and work of Paul Erdős.

[2] F. R. Bach and M. I. Jordan. Thin junction trees. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, Cambridge, MA, 2002. MIT Press.

[3] A. Beygelzimer and I. Rish. Inference complexity as a model-selection criterion for learning bayesian networks. In *Proceedings of the Eighth International Conference on Principles of Knowledge Representation and Reasoning (KR2002), Toulouse, France*, 2002.

[4] B. Bollobás. *Random graphs.* Cambridge University Press, Cambridge, second edition, 2001.

[5] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Information Theory*, 14:462–467, 1968.

[6] G.F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2–3):393–405, 1990.

[7] Thomas M. Cover and Joy A. Thomas. *Elements of information theory.* John Wiley & Sons Inc., New York, 1991. A Wiley-Interscience Publication.

[8] R. Dechter. Bucket elimination: A unifying framework for probabilistic inference. In *Proc. Twelfth Conf. on Uncertainty in Artificial Intelligence*, pages 211–219, 1996.

[9] P. Erdős and A. Rényi. On the evolution of random graphs. *Bull. Inst. Internat. Statist.*, 38:343–347, 1961.

[10] E. Friedgut and G. Kalai. Every monotone graph property has a sharp threshold. *Proceedings of the American Mathematical Society*, 124(10):2993–3002, 1996.

[11] D. Karger and N. Srebro. Learning markov networks: maximum bounded tree-width graphs. In *Proceedings of the 12th Annual Symposium on Discrete Algorithms (SODA)*, pages 392–401, 2001.

[12] T. Kloks. *Treewidth. Computations and Approximations*, volume 842 of *Lecture Notes in Computer Science.* Springer-Verlag, Berlin, 1994.

[13] S.L. Lauritzen and D.J. Spiegelhalter. Local computation with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50(2):157–224, 1988.

[14] N. Srebro. Maximum likelihood bounded Tree-Width markov networks. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 504–511, 2001.