A Short Note on the Sample Complexity of Learning Probabilistic Circuits

John Leland[1] **YooJung Choi**[1]

[1]School of Computing and Augmented Intelligence, Arizona State University, Tempe, Arizona, USA

Abstract

The expressive capabilities of probabilistic circuits (PCs) in both the exact and approximate case have been the subject of extensive study. However, sample complexity, which provides approximation guarantees probabilistically remains relatively unstudied for PCs. In pursuit of better understanding the capacity of PCs to approximate distributions, we investigate the sample complexity of learning the parameters of deterministic and decomposable PCs. By restricting to the case where the PC structure of the true distribution is known, we are able to achieve tight bounds for the Hellinger distance and reverse KL divergence, as well as providing results for the total variation distance.

1 INTRODUCTION

Probabilistic circuits (PCs) are deep generative models that compactly represent probability distributions through the composition of sum, product, and leaf gates. PCs can be seen as encompassing many families of tractable probabilistic models—models that support exact and efficient inference of various queries—and the expressive efficiency of PC classes has been studied extensively in the case of exact representation, as well as, to a lesser degree, approximate case [11, 4, 8, 22, 23, 12, 15, 17]. However, even though PCs are often learned from data in practice, analysis of the sample complexity of learning PCs remains underexplored; with the sole existing work on the matter covering only the class of tree-shaped sum product networks [1].

In pursuit of further understanding the capabilities of PCs to approximate distributions, this work derives the sample complexity of learning the parameters of *deterministic and decomposable PCs*, in the restricted case of known structure. We assess the sample complexity under a number of f-divergences—namely the total variation distance, the Hellinger distance, and the reverse Kullback-Leibler (KL) divergence. For Hellinger and reverse KL divergence we obtain tight upper and lower bounds, resulting in a sample complexity of $\Theta(\frac{S+\log(1/\delta)}{\epsilon^2})$ and $\Theta(\frac{S+\log(1/\delta)}{\epsilon})$ respectively, where S is the size of the PC. For the total variation distance we obtain an upper bound on the sample complexity of $O(\frac{S+\log(1/\delta)}{\epsilon^2})$ and a lower bound of $\Omega(\frac{S+\log(1/\delta)}{\epsilon})$.

2 PRELIMINARIES

Notation In this work we will use $\mathbf{x}$ to denote instantiations of variables. Our error parameter will always be denoted by ϵ, and our confidence parameter will be written as δ. To distinguish the two, we will have $\hat{P}$ be our distribution learned from samples and P be the true distribution.

2.1 SAMPLE COMPLEXITY

For a given distance measure D, we wish to characterize the total number of samples from a distribution P that is needed to learn it within distance $\epsilon > 0$, with error probability $\delta \in (0, 1]$:

$$\Pr[D(\hat{P}||P) \leq \epsilon] \geq 1 - \delta$$

We will write the sample complexity in shorthand as $\Phi(D, \epsilon, \delta)$; this work specifically considers the cases where D can be total variation distance, reverse KL divergence, or the Hellinger distance.

Definition 2.1. The *total variation distance (TVD), reverse KL divergence, and squared Hellinger distance* between two probability distributions P and $\hat{P}$ over a set of Boolean

variables $\mathbf{X}$ are defined respectively as:

$$D_{\mathsf{TV}}(\hat{P}\|P) = \frac{1}{2}\sum_{\mathbf{x}}\left|\hat{P}(x) - P(x)\right|,$$

$$D_{\mathsf{KL}}(\hat{P}\|P) = \sum_{\mathbf{x}}\hat{P}(x)\log(\frac{\hat{P}(x)}{P(x)}),$$

$$D_{\mathsf{H}}(\hat{P}\|P)^2 = 1 - \sum_{\mathbf{x}}\sqrt{\hat{P}(x)P(x)}.$$

Note here that we use the *squared* Hellinger for the majority of our proofs due to its properties such as tensorization, but the bounds trivially extend to the Helliger distance.

2.2 PROBABILISTIC CIRCUITS

Probabilistic circuits (PCs) [8] provide a unifying framework for a wide class of tractable probabilistic models, including arithmetic circuits [10], sum-product networks [19], cutset networks [20], probabilistic sentential decision diagrams [14], and bounded-treewidth graphical models [13, 9].

Definition 2.2 (Probabilistic circuits)**.** A probabilistic circuit (PC) $\mathcal{C} := (\mathcal{G}, \theta)$ represents a joint probability distribution $p(\mathbf{X})$ over random variables $\mathbf{X}$ through a directed acyclic graph (DAG) $\mathcal{G}$ parameterized by θ. The DAG is composed of 3 types of nodes: leaf, product $\otimes$, and sum $\oplus$ nodes. Every leaf node in $\mathcal{G}$ is an input, and every internal node receives inputs from its children $\mathrm{in}(n)$. The number of children at node n is denoted as $\mathsf{ch}(n)$. The scope of a given node, $\phi(n)$, is a recursively defined function which associates to each unit n a subset of $\mathbf{X}$: for each non-input unit n, $\phi(n) = \cup_{c\in\mathrm{in}(n)}\phi(c)$, and the scope of a leaf node is a single variable in $\mathbf{X}$. Naturally, the scope of the root node is $\mathbf{X}$. Each node n of a PC is then recursively defined as:

$$p_n(x) := \begin{cases} l(x), & \text{if } n \text{ is a leaf} \\ \prod_{c\in\mathrm{in}(n)} p_c(x) & \text{if } n \text{ is a } \otimes \\ \sum_{c\in\mathrm{in}(n)} \theta_{n,c} p_c(x) & \text{if } n \text{ is a } \oplus \end{cases} \quad (1)$$

where $\theta_{n,c} \in [0, 1]$ is the parameter associated with the edge connecting nodes n, c in $\mathcal{G}$, and $\sum_{c\in\mathrm{in}(n)}\theta_{n,c} = 1$. In this paper, we assume $l(x)$ at a leaf node is a Boolean indicator function: i.e., $\mathbb{1}[x = 1]$ or $\mathbb{1}[x = 0]$. The distribution represented by the circuit is the output at its root node.

Probabilistic circuits ensure the tractability of a variety of queries by composing sum and product gates under specific conditions. In this work we focus on circuits that are deterministic and decomposable [1].

[1]Smoothness is necessary, but we can trivially smooth our circuit in $O(n^2)$ if necessary [8]

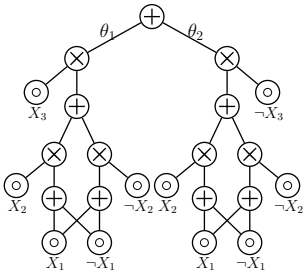

Figure 1: A smooth, decomposable, and deterministic PC (weights shown only for the root for conciseness).

Definition 2.3 (Smoothness and decomposability)**.** A sum unit is *smooth* if its children have identical scopes: $\phi(c) = \phi(n)$, $\forall c \in \mathrm{in}(n)$. A product unit is *decomposable* if its children have disjoint scopes: $\phi(c_i) \cap \phi(c_j) = \emptyset$, $\forall c_i \neq c_j \in \mathrm{in}(n)$. A PC is smooth and decomposable iff every sum unit is smooth and every product unit is decomposable.

Definition 2.4 (Determinism)**.** A sum node is *deterministic* if, for any fully-instantiated input, the output of at most one of its children is nonzero. In other words, the supports of its children are mutually disjoint. A PC is deterministic iff all of its sum nodes are deterministic.

Figure 1 depicts an example PC that is smooth, decomposable, and deterministic.

Deterministic and decomposable PCs have a very useful property when looking at their joint probability distribution. See that for a given variable instantiation $\mathbf{X} = x$, we have that the circuit admits only a single tree-shaped path through the PC [7]. Thus, to compute the mass function[2], we can simply multiply together the parameters associated with the tree defined by x. We denote the tree defined by x as $\mathsf{tree}(x)$, and thus $P(x) = \prod_{e\in\mathsf{tree}(x)}\theta_e$.

Deterministic and decomposable PCs also admit closed-form maximum-likelihood estimation of their parameters [16]. This closed form can be thought of simply as counting the number of samples that reach a child edge c at a sum node n, divided by the total number of samples to reach that same sum node n. Mathematically, this can be formalized with the following notions:

Definition 2.5 (Flows)**.** The flow $F_{n,c}(\mathbf{x})$ of any edge (n, c) in a PC given variables assignments $\mathbf{x}$ is defined as $\mathbb{I}[\mathbf{x} \in \gamma_n \cap \gamma_c]$ where $\mathbb{I}[\cdot]$ is the indicator function. The flow $F_{n,c}(D)$ with respect to a dataset $D = \{\mathbf{x}^{(i)}\}_{i=1}^N$ is the sum of the flows of all samples: $F_{n,c}(D) = \sum_{i=1}^N F_{n,c}(\mathbf{x}^{(i)})$.

It has been shown that the flow for all edges in a PC can be computed by a forward and backwards pass, which are both linear time in the size of the circuit. The MLE parameters

[2]Or density function

are computed using as follows [14]:

$$\forall(n,c) : \theta^*_{n,c} = \frac{F_{n,c}(D)}{\sum_{c \in \text{in}(n)} F_{n,c}(D)} \tag{2}$$

3 LEARNING PROBABILISTIC CIRCUITS

3.1 MLE REDUCES TO EMPIRICAL ESTIMATION

Given that we can learn the MLE parameters in closed form for a deterministic and decomposable PC, we will use the definition to reduce to learning an empirical distribution over a set of categories. Essential to this argument is the latent variable interpretation of PCs [18], which directly corresponds to learning a discrete (categorical) distribution over the choices of children at each sum node. This is formalized in the following,

Lemma 3.1 (Sum nodes as CPTs)**.** *Suppose we have a sum node n with latent variable Z, latent variable assignment z, and a set of parameters $\{\theta_1, \theta_2, \ldots, \theta_m\}$. The probability of a given weight is defined as follows:*

$$n(Z = k|z) = \begin{cases} \theta_k & \text{if } z \in Z \\ \bar{\theta}_k & \text{if } z \in \bar{Z} \end{cases}$$

Now, let us suppose the optimal distribution over the parameters as the true discrete distribution, which will be denoted as P_n, for a given sum node n. Under this setting, see that updating parameters via the MLE algorithm in a deterministic and decomposable PC is *equivalent* to learning the empirical estimator of a discrete distribution.

Proposition 3.2. *Given a deterministic sum node n with parameters $\{\theta_1, \ldots, \theta_s\}$ and samples $D = \{\mathbf{x}^{(j)}\}_{j=1}^M$, learning the optimal sum node n' using the maximum log-likelihood algorithm is equivalent to learning a discrete distribution P with the empirical estimator $\hat{P}(i) = \frac{1}{M}\sum_{j=1}^M \mathbb{I}[\mathbf{x}^{(j)} = c]$ for $c \in \{1, 2, \ldots, \mathsf{ch}(n)\}$.*

Proof. Let n be a deterministic sum node with parameters $\{\theta_1, \ldots, \theta_s\}$. For a given dataset D, we learn the parameters as

$$\begin{aligned} \theta^*_{n,c} &= \frac{F_{n,c}(D)}{\sum_{c \in \text{in}(n)} F_{n,c}(D)} \\ &= \frac{\mathbb{I}[\mathbf{x} \in \gamma_n \cap \gamma_c]}{\sum_{c \in \text{in}(n)} \mathbb{I}[\mathbf{x} \in \gamma_n \cap \gamma_c]} \\ &= \frac{\sum_{j=1}^M \mathbb{I}[\mathbf{x}^{(j)} \in \gamma_n \cap \gamma_c]}{\sum_{c \in \text{in}(n)} \sum_{j=1}^M \mathbb{I}[\mathbf{x}^{(j)} \in \gamma_n \cap \gamma_c]}. \end{aligned}$$

As we are only concerned with one sum node, we fix n and see that under the latent variable interpretation the above is equivalent to a sum node choosing one category in a discrete distribution:

$$\frac{\sum_{j=1}^M \mathbb{I}[\mathbf{x}^{(j)} = c]}{\sum_{c \in \text{in}(n)} \sum_{j=1}^M \mathbb{I}[\mathbf{x}^{(j)} = c]} = \frac{\sum_{j=1}^M \mathbb{I}[\mathbf{x}^{(j)} = c]}{M}.$$

Thus, learning the optimal sum parameters using the maximum log-likelihood algorithm is equivalent to learning a discrete distribution with the empirical estimator. □

This connection immediately allows us to derive sample complexity results for a single deterministic sum node; however, PCs can be so much more than just a single sum node.

4 TIGHT SAMPLE COMPLEXITY RESULTS

Extending the sample complexity results for learning discrete distributions with the empirical estimator is non-trivial to extend beyond a single sum node for the following reasons: (1) the number of samples that reach a given sum node n, denoted m_n, is distributed as $\mathcal{B}(M, p(\gamma_n))$, where M is the total number of samples taken at the root and $p(\gamma_n)$ is the probability of reaching node n, (2) the number of samples is dependent on the number of samples to reach a given nodes parent, and (3) the distance between nodes with a smaller $p(\gamma)$ have a smaller effect on the overall sample complexity, and thus using the naive method of approximating each sum within ϵ can result in far more samples than truly required.

Extending beyond one sum node, we are able to find the tight sample complexity bound to be as follows,

Theorem 4.1 (Sample Complexity of Learning a Deterministic and Decomposable PC)**.** *Let $\hat{P}$ be a deterministic and decomposable probabilistic circuit of size S. Then,*

$$\Phi(D_{\mathsf{KL}}(\hat{P}\|P), S, \epsilon, \delta) = \Theta(\frac{S + \log(1/\delta)}{\epsilon}),$$
$$\Phi(D_{\mathsf{H}}(\hat{P}\|P), S, \epsilon, \delta) = \Theta(\frac{S + \log(1/\delta)}{\epsilon^2}).$$

We also obtain results for the total variation distance; though they are not tight,

Corollary 4.2. *Let $\hat{P}$ be a deterministic and decomposable probabilistic circuit of size S. Then,*

$$\Phi(D_{\mathsf{TV}}(\hat{P}\|P), S, \epsilon, \delta) = \Omega(\frac{S + \log(1/\delta)}{\epsilon}),$$
$$\Phi(D_{\mathsf{TV}}(\hat{P}\|P), S, \epsilon, \delta) = O(\frac{S + \log(1/\delta)}{\epsilon^2}).$$

Note here that the form of our bounds looks quite similar to the upper and lower bounds on learning discrete distributions from [5].

4.1 THE UPPER BOUND

Here, we would like to remind that we are using the **reverse KL divergence**: $D_{\mathsf{KL}}(\hat{P}\|P)$. This is necessary as enforcing a learning guarantee less than ϵ with the standard KL divergence requires learning arbitrarily small probabilities without approximating with a trivial 0 [5]. Thankfully the following inequalities still hold,

Lemma 4.3 (Divergence Inequalities [6])**.** *Let $\hat{P}$ and P be probability measures, then the following inequalities hold (1) $D_{\mathsf{TV}}(\hat{P}\|P) \leq \sqrt{\frac{1}{2}D_{\mathsf{KL}}(\hat{P}\|P)}$, (2) $\frac{1}{2}D_{\mathsf{TV}}(\hat{P}\|P)^2 \leq D_{\mathsf{H}}(\hat{P}\|P)^2 < D_{\mathsf{TV}}(\hat{P}\|P)$, (3) $D_{\mathsf{H}}(\hat{P}\|P)^2 < \frac{1}{2}D_{\mathsf{KL}}(\hat{P}\|P)$.*

Furthermore, the reverse KL divergence between deterministic and decomposable PCs have the following useful property,

Lemma 4.4 (Deterministic reverse KLD Decomposition)**.** *Let $\hat{P}$ be a deterministic and decomposable PC with the same structure as the true distribution P. The reverse KL divergence $D_{\mathsf{KL}}(\hat{P}\|P)$ can be decomposed as*

$$D_{\mathsf{KL}}(\hat{P}\|P) = \sum_{n=1}^{S} \hat{P}(\gamma_n) D_{\mathsf{KL}}(\hat{\theta}_{n,c}\|\theta_{n,c}),$$

where $\hat{P}(\gamma_n)$ is the probability of reaching sum node n under the learned circuit, S is the total number of sum nodes, and $D_{\mathsf{KL}}(\hat{\theta}_{n,c}\|\theta_{n,c})$ is the reverse KL divergence of the children of node n.

Proof sketch. For the full proof see Appendix A.2. First use the fact that as $\hat{P}, P$ are deterministic and thus for a given $\mathbf{x}$ follow a tree through the circuit. This allows us to turn the logarithmic term into $\sum_{e\in\mathsf{tree}(\mathbf{x})} \log(\frac{\hat{\theta}_e}{\theta_e})$. Using indicator functions we can change this to be sums over the nodes and children. Plugging this reformulated version into the expectation definition of the reverse KL divergence we use simple algebra to achieve the final result. □

As the reverse KL divergence provides a simple formulation (only having to look at the distance between each pair of sum nodes), we will utilize this along with Lemma 4.3 to construct all 3 of our upper bounds as

$$\begin{aligned} &\Pr[D_{\mathsf{KL}}(\hat{P}\|P) \leq \epsilon] \leq 1-\delta \\ \implies &\Pr[D_{\mathsf{TV}}(\hat{P}\|P) \leq \sqrt{\epsilon/2}] \leq 1-\delta \\ \implies &\Pr[D_{\mathsf{H}}(\hat{P}\|P)^2 \leq \epsilon/2] \leq 1-\delta. \end{aligned}$$

We will take inspiration from Agrawal [2] to bound $\Pr[D_{\mathsf{KL}}(\hat{P}\|P) \leq \epsilon] \leq 1-\delta$ by first showing that,

Theorem 4.5. *For P a deterministic and decomposable PC and $\hat{P}$ learned with MLE for the same structure, for all $\epsilon > \frac{S(w-1)}{M}$ it holds that*

$$\Pr[D_{\mathsf{KL}}(\hat{P}\|P) \geq \epsilon] \leq e^{-\epsilon M}\left(\frac{\epsilon e M}{S(w-1)}\right)^{S(w-1)}$$

where S is the number of sum nodes and w is the maximum width (or maximum number of children for a sum node).

To prove this theorem, we follow a simple road-map, (1) use Lemma 4.4 to decompose into only upper bounding each sum node, (2) use the Law of Total Expectation to fix the number of samples to each sum node n, (3) take the upper bound from [2] for each sum node with a fixed number of samples, and (4) take a Chernoff bound.

The upper bound required for this strategy comes directly from [2],

Theorem 4.6. *For P a categorical distribution over k categories and $\hat{P}$ the empirical estimator with n samples, it holds that for all $0 \leq t < n$,*

$$\mathbb{E}[\exp(tD_{\mathsf{KL}}(\hat{P}\|P))] \leq \left(\frac{1}{1-t/n}\right)^{k-1}.$$

This allows us to trivially upper bound the MGF for each sum node if the number of samples to reach this sum node is fixed, as such we are ready for our proof.

Proof sketch. We provide only a proof sketch here, deferring the completed proof to Appendix A.3. See that by Lemma 4.4 we can write $\mathbb{E}[\exp(tD_{\mathsf{KL}}(\hat{P}\|P))] = \mathbb{E}[\exp(t\sum_n \hat{P}(\gamma_n)D_{\mathsf{KL}}(\hat{P}_{n,c}\|P)_{n,c})]$. Then, using the law of total expectation, we can show that

$$\begin{aligned} &\mathbb{E}[\exp(tD_{\mathsf{KL}}(\hat{P}\|P))] \\ &= \mathbb{E}_m[\prod_n \mathbb{E}[\exp(t\frac{m_n}{M}D_{\mathsf{KL}}(\hat{\theta}_{n,c}\|\theta_{n,c})|m_n]] \end{aligned}$$

Then, upper bounding each individual $\mathbb{E}[\exp(t\frac{m_n}{M}D_{\mathsf{KL}}(\hat{\theta}_{n,c}\|\theta_{n,c})|m_n]$ using Theorem 4.6 we yield an upper bound which can be plugged into a Chernoff bound. This gives us our final result by taking the complement of $D_{\mathsf{KL}}(\hat{P}\|P)) \geq \epsilon$. □

Using this probabilistic inequality, we can follow the argument from Canonne [5] to finally prove our upper bound from Theorem 4.1,

Proof of Upper Bound in Theorem 4.1. Using the change of exponent rules, we rewrite Theorem 4.5 as

$$\begin{aligned} \Pr[D_{\mathsf{KL}}(\hat{P}\|P) \geq \epsilon] &\leq e^{S(w-1)\log(\frac{\epsilon e M}{S(w-1)})}e^{-\epsilon M} \\ &= e^{-\epsilon M + S(w-1)\log(\frac{\epsilon e M}{S(w-1)})} \end{aligned}$$

Now using the fact that for $M \geq \frac{15S(w-1)}{e\epsilon}$ we have $S(w-1)\log(\frac{e\epsilon M}{S(w-1)}) < \frac{M\epsilon}{2}$. Thus, for $M \geq \frac{15S(w-1)}{e\epsilon}$

$$\Pr[D_{\mathsf{KL}}(\hat{P}\|P) \geq \epsilon] \leq e^{-\frac{M\epsilon}{2}}$$

Finally, setting $e^{-\frac{M\epsilon}{2}} \leq \delta$ we get that $M \geq \frac{2\log(1/\delta)}{\epsilon}$ implies that $\Pr[D_{\mathsf{KL}}(\hat{P}\|P) \geq \epsilon] < \delta$. Thus, we guarantee learning $D_{\mathsf{KL}}(\hat{P}\|P) \leq \epsilon$ with probability $1-\delta$ using at least $O(\frac{S+\log(1/\delta)}{\epsilon})$ samples. The extension to the KL divergence and squared Hellinger distance follows from Lemma 4.3, which results a learning guarantee for $D_{\mathsf{TV}}(\hat{P}\|P) \leq \epsilon$ and $D_{\mathsf{H}}(\hat{P}\|P)^2 < \epsilon$ with probability $1-\delta$ using at least $O(\frac{S+\log(1/\delta)}{\epsilon^2})$ samples. □

4.2 THE LOWER BOUND

As we have seen with the upper bound, proving the sample complexity of a deterministic and decomposable PC uses remarkably similar techniques to the simple categorical case. The lower bound is no exception to this. The standard way of proving the sample complexity lower bound for a k category discrete distribution is outlined in [5], where the goal is to use Assouad's lemma [3] in conjunction with the hardness of estimating the bias of a coin. Notice that this technique does not rely on any specific learning algorithm for a deterministic and decomposable PC, and thus is truly a lower bound.

The construction will proceed as follows: (1) construct a family of circuits by perturbing pairs of sum node parameters (2) get upper and lower bounds on the squared Hellinger distance, (3) use Assouad's lemma to get a lower bound on the expected Hellinger distance, and (4) finish off with the hardness of estimating a coin's bias.

Lemma 4.7 (PC Family Construction)**.** *There exists a family of $2^{S(w-1)/2}$ distributions over deterministic and decomposable PCs such that for all $v, v' \in \{-1,1\}^{S(w-1)/2}$*

$$D_{\mathsf{H}}(P_v\|P_{v'})^2 \geq \frac{8\epsilon^2}{S(w-1)}H(v,v')$$

where $H(v,v')$ is the Hamming distance between v, v'. Furthermore, for all $v, v' \in \{-1,1\}^{S(w-1)/2}$ such that $H(v,v') = 1$,

$$D_{\mathsf{H}}(P_v\|P_{v'})^2 \leq \frac{64\epsilon^2}{S}.$$

Proof sketch. The complete proof can be found in Appendix A.4. The construction starts from a simple idea, let P be the distribution that is uniform over all sum nodes, i.e. $\theta_{n,c} = \frac{1}{\mathsf{ch}(n)}$. Then, define $P_v = \frac{1}{\mathsf{ch}(n)} + v\tau_n$ and $P_{v'} = \frac{1}{\mathsf{ch}(n)} + v'\tau_n$. Using a number of basic inequalities, we find that the resulting function has a component which is equivalent to a partition function. Evaluating this and using another set of inequalities establishes the lower bound on $D_{\mathsf{H}}(P_v\|P_{v'})^2$.

The construction yields the upper bound in a far more intuitive manner. See that flipping one bit to be different results in $\sqrt{P_{v_c}(\mathbf{x})P_{v_c'}(\mathbf{x})} = \sqrt{(\frac{1}{\mathsf{ch}(n)}^2 - \tau_n^2)}$ for the bit where $v_c \neq v_c'$. □

See that the following definition of Assouad's lemma when combined with our constructed circuits will directly provide a lower bound:

Theorem 4.8 (Assouad's Lemma [3])**.** *Let $\mathcal{T} \subseteq \tau_n(\Omega)$ be a family of probability distributions. Suppose there exists a family of $\mathcal{H} \subseteq \mathcal{T}$ of 2^r distributions and constants $\alpha, \beta > 0$ such that, writing $\mathcal{H} = \{P_z\}_{z\in\{-1,1\}^r}$,*

1. *for all $v, v' \in \{-1,1\}^r$, the distance between $P_v, P_{v'}$ is at least proportional to the Hamming distance:*
$$D_{\mathsf{H}}(P_v\|P_{v'})^2 \geq \alpha H(v,v')$$
2. *for all $v, v' \in \{-1,1\}^r$ with $H(v,v') = 1$, the squared Hellinger distance of $P_v, P_{v'}$ is small:*
$$D_{\mathsf{H}}(P_v\|P_{v'})^2 \leq \beta$$

Then, for all $M \geq 1$,

$$\inf_{\hat{P}} \sup_{P} \mathbb{E}_{s_1,\ldots,s_M}[D_{\mathsf{H}}(\hat{P}\|P)^2] \geq \frac{\alpha}{4}r(1-\beta)^{2M}.$$

Note that this is a modified version of Assouad's lemma from what appears in [6] to instead lower bound the expected squared Hellinger, we provide an overview of this adjusted version in Appendix A.5. Using this version of Assouad's lemma in conjunction with the construction from Lemma 4.7 we obtain the following lower bound proof.

Proof of Lower Bound in Theorem 4.1. Plugging in $\alpha = \frac{8\epsilon^2}{S(w-1)}$, $r = \frac{S\mathsf{ch}(n)}{2}$, and $\beta = \frac{64\epsilon^2}{S}$ we get the following:

$$\begin{aligned}
&\inf_{A\in\mathcal{A}_M} \sup_{D\in\mathcal{T}} \mathbb{E}_{s_1,\ldots,s_M\sim D}[D_{\mathsf{H}}(D\|\hat{D}_A)^2] \\
&\geq \frac{1}{4}\alpha r(1-\beta)^{2M} \\
&= \frac{1}{4}\frac{8\epsilon^2}{S(w-1)}re^{-2\frac{4\epsilon^2}{S}M}
\end{aligned}$$

Now, see that to have any hope of achieving an error of ϵ^2, we must have that $e^{-2M\frac{4\epsilon^2}{S}} \leq 1/2$, (or really any constant less than 1

$$\begin{aligned}
e^{-2M\frac{4\epsilon^2}{S}} &\leq log(1/2) \\
-2M\left(\frac{4\epsilon^2}{S}\right) &\leq -\log(2) \\
M &\geq \frac{S\log(2)}{8\epsilon^2}
\end{aligned}$$

which in turn shows that $M = \Omega(\frac{S}{\epsilon^2})$. Additionally, in our current problem setup we can be seen as attempting to learn the bias of a coin for a single pair of our perturbed variables, $\frac{1}{\mathsf{ch}(n)} + \tau_n, \frac{1}{\mathsf{ch}(n)} - \tau_n$. In the learning theory literature this is a well studied task, and a lower bound on the sample complexity of learning this with probability $1 - \delta$ is $\Omega(\frac{\log(1/\delta)}{\epsilon^2})$. Thus, to account for the hardness via Assouad's lemma and the hardness of learning the bias of our perturbed variables, we have that

$$M = \Omega(\max(\frac{S}{\epsilon^2}, \frac{\log(1/\delta)}{\epsilon^2})) = \Omega(\frac{S + \log(1/\delta)}{\epsilon^2}).$$

Taking our distance metrics to then be the Hellinger distance and the reverse KL divergence, we see that if $D_{\mathsf{H}}(\hat{P}\|P)^2 \leq \epsilon^2$ then $D_{\mathsf{H}}(\hat{P}\|P) \leq \epsilon$ and $D_{\mathsf{KL}}(\hat{P}\|P) \leq \frac{\epsilon}{2}$. Ignoring constants, this proves that

$$\Phi(D_{\mathsf{KL}}(\hat{P}\|P), S, \epsilon, \delta) = \Omega(\frac{S + \log(1/\delta)}{\epsilon}),$$
$$\Phi(D_{\mathsf{H}}(\hat{P}\|P), S, \epsilon, \delta) = \Omega(\frac{S + \log(1/\delta)}{\epsilon^2}).$$

□

The above proof also provides the map for proving the lower bound in Corollary 4.2.

Lower bound in Corollary 4.2. As $D_{\mathsf{KL}}(\hat{P}\|P)^2 \leq D_{\mathsf{TV}}(\hat{P}\|P)$ we cannot simply propagate our bound above as desired. However, see that the exact same argument holds in setting $\tau_n^2 = \frac{16\epsilon^2}{S\mathsf{ch}(n)^2}$, which implies that $M = \Omega(\frac{S}{\epsilon^2})$ and thus

$$\Phi(D_{\mathsf{TV}}(\hat{P}\|P), S, \epsilon, \delta) = \Omega(\frac{S + \log(1/\delta)}{\epsilon}),$$
$$\Phi(D_{\mathsf{TV}}(\hat{P}\|P), S, \epsilon, \delta) = O(\frac{S + \log(1/\delta)}{\epsilon^2}).$$

□

It is worthwhile noting here that using the construction from Lemma 4.7 this is the best lower bound for total variation distance we can achieve by lower bounding $D_{\mathsf{H}}(\hat{P}\|P)^2$. In order to improve our lower bound on the total variation distance, we would need to find a different constant α such that $D_{\mathsf{TV}}(P_v\|P_{v'}) \geq \alpha H(v, v')$. This is due to the fact that improving our current α with respect to the squared Hellinger such that we could derive $\Phi(D_{\mathsf{TV}}(\hat{P}\|P), \epsilon, \delta) = \Omega(\frac{S+\log(1/\delta)}{\epsilon^2})$ would also imply that $\Phi(D_{\mathsf{TV}}(\hat{P}\|P), \epsilon, \delta) = \Omega(\frac{S+\log(1/\delta)}{\epsilon^4})$! Which is an obvious contradiction to the upper bound we have derived in Theorem 4.1.

5 CONCLUSION

For deterministic and decomposable PCs over Boolean indicator leaves, we have proven tight lower bounds on the sample complexity when utilizing the Hellinger and KL divergence; as well as weaker bounds for the total variation distance. However, this does not resolve all possible questions in this area. Future work would be wise to investigate dropping determinism, as while our current lower bound technique would hold, the upper bound technique would need to radically change due to a lack of closed form MLE. Further work in this area should also be dedicated to learning the structure of the PC with samples, as well as finding more sophisticated techniques to deal with the numerous types of leaves.

REFERENCES

[1] Ishaq Aden-Ali and Hassan Ashtiani. On the Sample Complexity of Learning Sum-Product Networks, February 2020. URL `http://arxiv.org/abs/1912.02765`. arXiv:1912.02765.

[2] Rohit Agrawal. Finite-sample concentration of the multinomial in relative entropy. *IEEE Transactions on Information Theory*, 66(10):6297–6302, 2020.

[3] Patrice Assouad. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

[4] Simone Bova, Florent Capelli, Stefan Mengel, and Friedrich Slivovsky. Knowledge compilation meets communication complexity. In *IJCAI*, volume 16, pages 1008–1014, 2016.

[5] Clément L Canonne. A short note on learning discrete distributions. *arXiv preprint arXiv:2002.11457*, 2020.

[6] Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020.

[7] Arthur Choi and Adnan Darwiche. On relaxing determinism in arithmetic circuits. In *International Conference on Machine Learning*, pages 825–833. PMLR, 2017.

[8] YooJung Choi, Antonio Vergari, and Guy Van den Broeck. Probabilistic circuits: A unifying framework for tractable probabilistic models. *UCLA. URL: http://starai. cs. ucla. edu/papers/ProbCirc20. pdf*, page 6, 2020.

[9] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968. doi: 10.1109/TIT.1968.1054142.

[10] Adnan Darwiche. A logical approach to factoring belief networks. *KR*, 2:409–420, 2002.

[11] Adnan Darwiche. A differential approach to inference in bayesian networks. *J. ACM*, 50(3):280–305, May 2003. ISSN 0004-5411. doi: 10.1145/765568.765570. URL https://doi.org/10.1145/765568.765570.

[12] Alexis de Colnet and Stefan Mengel. A compilation of succinctness results for arithmetic circuits. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 18, pages 205–215, 2021.

[13] Gal Elidan and Stephen Gould. Learning bounded treewidth bayesian networks. *Advances in neural information processing systems*, 21, 2008.

[14] Doga Kisa, Guy Van den Broeck, Arthur Choi, and Adnan Darwiche. Probabilistic sentential decision diagrams. In *KR*, 2014.

[15] John Leland and YooJung Choi. On the hardness of approximating distributions with tractable probabilistic models. *Advances in Neural Information Processing Systems*, 38:42977–43000, 2026.

[16] Anji Liu and Guy Van den Broeck. Tractable regularization of probabilistic circuits. *Advances in Neural Information Processing Systems*, 34:3558–3570, 2021.

[17] James Martens and Venkatesh Medabalimi. On the expressive efficiency of sum product networks. *arXiv preprint arXiv:1411.7717*, 2014.

[18] Robert Peharz, Robert Gens, Franz Pernkopf, and Pedro Domingos. On the latent variable interpretation in sum-product networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(10):2030–2044, 2016.

[19] Hoifung Poon and Pedro Domingos. Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 689–690. IEEE, 2011.

[20] Tahrima Rahman, Prasanna Kothalkar, and Vibhav Gogate. Cutset networks: A simple, tractable, and scalable approach for improving the accuracy of chow-liu trees. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14*, pages 630–645. Springer, 2014.

[21] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.

[22] Lang Yin and Han Zhao. On the expressive power of tree-structured probabilistic circuits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=suYAAOI5bd.

[23] Honghua Zhang, Steven Holtzen, and Guy Broeck. On the relationship between probabilistic circuits and determinantal point processes. In *Conference on Uncertainty in Artificial Intelligence*, pages 1188–1197. PMLR, 2020.

A PROOFS

A.1 PROOF OF PROPOSITION 3.2

Proof. Let $\hat{S}$ be a deterministic sum node with parameters $\{\theta_1, \ldots, \theta_m\}$. For a given dataset D, we learn the parameters as

$$\begin{aligned}\theta^*_{n,c} &= \frac{F_{n,c}(D)}{\sum_{c \in \mathrm{in}(n)} F_{n,c}(D)} \\ &= \frac{\mathbb{I}[\mathbf{x} \in \gamma_n \cap \gamma_c]}{\sum_{c \in \mathrm{in}(n)} \mathbb{I}[\mathbf{x} \in \gamma_n \cap \gamma_c]} \\ &= \frac{\sum_{j=1}^{N} \mathbb{I}[\mathbf{x}^{(j)} \in \gamma_n \cap \gamma_c]}{\sum_{c \in \mathrm{in}(n)} \sum_{j=1}^{N} \mathbb{I}[\mathbf{x}^{(j)} \in \gamma_n \cap \gamma_c]}.\end{aligned}$$

As we are only concerned with one sum node, we fix n and see that under the latent variable interpretation the above is equivalent to a sum node choosing one category in a discrete distribution:

$$\frac{\sum_{j=1}^{N} \mathbb{I}[\mathbf{x}^{(j)} = c]}{\sum_{c \in \mathrm{in}(n)} \sum_{j=1}^{N} \mathbb{I}[\mathbf{x}^{(j)} = c]} = \frac{\sum_{j=1}^{N} \mathbb{I}[\mathbf{x}^{(j)} = c]}{N}.$$

Thus, learning the optimal sum parameters using the maximum log-likelihood algorithm is equivalent to learning a discrete distribution with the empirical estimator. □

A.2 PROOF OF LEMMA 4.4

Proof.

$$\begin{aligned}\forall \mathbf{x} : \log(\frac{\hat{P}(\mathbf{x})}{P(\mathbf{x})}) &= \log(\frac{\prod_{e \in \mathsf{tree}(\mathbf{x})} \hat{\theta}_e}{\prod_{e \in \mathsf{tree}(\mathbf{x})} \theta_e}) \\ &= \sum_{e \in \mathsf{tree}(\mathbf{x})} \log(\frac{\hat{\theta}_e}{\theta_e}) \\ &= \sum_n \sum_c \log(\frac{\hat{\theta}_{n,c}}{\theta_{n,c}}) \mathbb{I}((n,c) \in \mathsf{tree}(\mathbf{x}))\end{aligned}$$

Next, we can use the expectation definition of the reverse KL divergence:

$$\begin{aligned}D_{\mathsf{KL}}(\hat{P} \| P) &= \mathbb{E}_{\mathbf{x} \in \hat{P}}[\log(\frac{\hat{P}(\mathbf{x})}{P(\mathbf{x})})] \\ &= \mathbb{E}_{\mathbf{x} \sim \hat{P}}[\sum_n \sum_c \log(\frac{\hat{\theta}_{n,c}}{\theta_{n,c}}) \mathbb{I}((n,c) \in \mathsf{tree}(\mathbf{x}))] \\ &= \sum_n \sum_c \log(\frac{\hat{\theta}_{n,c}}{\theta_{n,c}}) \mathbb{E}_{\mathbf{x} \sim \hat{P}}[\mathbb{I}((n,c) \in \mathsf{tree}(\mathbf{x}))]\end{aligned}$$

It is a commonly used fact that for an event A, $\mathbb{E}[\mathbb{I}(A)] = P(A)$. Our event in this case can be seen as two parts: (1) n is in

$\mathsf{tree}(\mathbf{x})$ for a given $\mathbf{x}$, (2) c is the next child chosen by the sum node. Thus,

$$\begin{aligned}
\sum_n \sum_c \log(\frac{\hat{\theta}_{n,c}}{\theta_{n,c}}) \mathbb{E}_{\mathbf{x}\sim\hat{P}}[\mathbb{I}((n,c) \in \mathsf{tree}(\mathbf{x}))] &= \sum_n \sum_c \log(\frac{\hat{\theta}_{n,c}}{\theta_{n,c}}) \hat{P}(n \in \mathsf{tree}(\mathbf{x}) \wedge c \text{ is chosen}) \\
&= \sum_n \sum_c \log(\frac{\hat{\theta}_{n,c}}{\theta_{n,c}}) \hat{P}(n \in \mathsf{tree}(\mathbf{x})) \hat{P}(c \text{ is chosen } | n \in \mathsf{tree}(\mathbf{x})) \\
&= \sum_n \sum_c \log(\frac{\hat{\theta}_{n,c}}{\theta_{n,c}}) \hat{P}(\gamma_n) \hat{\theta}_{n,c} \\
&= \sum_n \hat{P}(\gamma_n) \sum_c \hat{\theta}_{n,c} \log(\frac{\hat{\theta}_{n,c}}{\theta_{n,c}}) \\
&= \sum_n \hat{P}(\gamma_n) D_{\mathsf{KL}}(\hat{P}_c \| P_c).
\end{aligned}$$

□

A.3 PROOF OF THEOREM 4.5

Proof of Theorem 4.5. Let $\hat{P}$ be a deterministic and decomposable PC learned using the MLE from the true distribution P which is a PC of the same structure. Then by Lemma 4.4, we have that $\mathbb{E}[\exp(tD_{\mathsf{KL}}(\hat{P}\|P))] = \mathbb{E}[\exp(t\sum_n \hat{P}(\gamma_n) D_{\mathsf{KL}}(\hat{P}_{n,c}\|P_{n,c}))]$. Next, use the law of total expectation and condition to fix the m_n's to some value.

$$\begin{aligned}
&\mathbb{E}[\exp(t\sum_n \hat{P}(\gamma_n) D_{\mathsf{KL}}(\hat{P}_{n,c}\|P_{n,c}))] \\
&= \mathbb{E}_m[\mathbb{E}[\exp(t\sum_n \frac{m_n}{M} D_{\mathsf{KL}}(\hat{P}_{n,c}\|P_{n,c}))|(m_1,\ldots,m_s)]] \\
&= \mathbb{E}_m[\mathbb{E}[\prod_n \exp(t\frac{m_n}{M} D_{\mathsf{KL}}(\hat{\theta}_{n,c}\|\theta_{n,c})|(m_1,\ldots,m_s)]] \\
&= \mathbb{E}_m[\prod_n \mathbb{E}[\exp(t\frac{m_n}{M} D_{\mathsf{KL}}(\hat{\theta}_{n,c}\|\theta_{n,c})|m_n]]
\end{aligned}$$

As such, it is easy to see that we can view each $D_{\mathsf{KL}}(\hat{\theta}_{n,c}\|\theta_{n,c})$ as P being a categorical distribution over $|\mathsf{ch}(n)|$ categories conditioned on having m_n samples each. Which fixes m_n, and allows us to use Theorem 4.6,

$$\begin{aligned}
&\mathbb{E}[\exp(t\frac{m_n}{M} D_{\mathsf{KL}}(\hat{\theta}_{n,c}\|\theta_{n,c})|m_n] \\
&\leq \left(\frac{1}{1-\frac{t\frac{m_n}{M}}{m_n}}\right)^{|\mathsf{ch}(n)|-1} \\
&= \left(\frac{1}{1-\frac{t}{M}}\right)^{|\mathsf{ch}(n)|-1}.
\end{aligned}$$

Thus we can see that setting $w = \max_n |\mathsf{ch}(n)|$,

$$\begin{aligned}
&\mathbb{E}_m[\prod_n \mathbb{E}[\exp(t\frac{m_n}{M} D_{\mathsf{KL}}(\hat{\theta}_{n,c}\|\theta_{n,c})|m_n]] \\
&\leq \mathbb{E}_m[\prod_n \left(\frac{1}{1-\frac{t}{M}}\right)^{|\mathsf{ch}(n)|-1}] \\
&\leq \mathbb{E}_m[\left(\frac{1}{1-\frac{t}{M}}\right)^{S(w-1)}] \\
&= \left(\frac{1}{1-\frac{t}{M}}\right)^{S(w-1)}.
\end{aligned}$$

Using a standard Chernoff bound, we then have that for $0 \leq t < M$ we have that

$$\begin{aligned}\Pr[D_{\mathsf{KL}}(\hat{P}\|P) \geq \epsilon] &\leq \inf_{t\in[0,M)]} e^{-t\epsilon}\mathbb{E}[e^{tD_{\mathsf{KL}}(\hat{P}\|P)}] \\ &\leq \inf_{t\in[0,M)]} e^{-t\epsilon}\left(\frac{1}{1-\frac{t}{M}}\right)^{S(w-1)},\end{aligned}$$

Choose $\frac{t}{M} = 1 - \frac{S(w-1)}{\epsilon M}$, then for all $\epsilon > \frac{S(w-1)}{M}$,

$$\begin{aligned}\Pr[D_{\mathsf{KL}}(\hat{P}\|P) \geq \epsilon] &\leq e^{-t\epsilon}\left(\frac{1}{1-\frac{t}{M}}\right)^{S(w-1)} \\ &\leq e^{-\epsilon M}\left(\frac{\epsilon e M}{S(w-1)}\right)^{S(w-1)}\end{aligned}$$

which completes the proof of Theorem 4.5. □

A.4 PROOF OF LEMMA 4.7

Proof. Let P be a distribution such that at each sum node n, $\theta_{n,c} = \frac{1}{\mathsf{ch}(n)}$ and each sum node has an even number of children. We provide the perturbed distribution notation as P_v where at each sum node, half of the children will be perturbed and the other half will be dependent variables; the value of said sum node parameters will be $\theta_{n,c}[v_c] = \frac{1}{\mathsf{ch}(n)} + v_c\tau_n$ where $\forall n, \tau_n = \sqrt{\frac{16\epsilon^2}{S(w-1)^2}}$ and $v_c \in \{-1, 1\}$, and remember w is the maximum sum node width. Then $v \in \{-1, 1\}^\rho$ where ρ is the total number of permuted variables in the circuit. We first look to lower bound the squared Hellinger distance for all $v, v' \in \{-1, 1\}^\rho$.

$$\begin{aligned}D_{\mathsf{H}}(P_v\|P_{v'}) &= 1 - \sum_{\mathbf{x}} \sqrt{P_v(\mathbf{x})P_{v'}(\mathbf{x})} \\ &= 1 - \sum_{\mathbf{x}} \sqrt{\prod_n \prod_c (\theta_{n,c}[v_c]\theta_{n,c}[v'_c])^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))}} \\ &= 1 - \sum_{\mathbf{x}} \sqrt{\prod_n \prod_{c|v_c=v'_c} (\theta_{n,c}[v_c]\theta_{n,c}[v'_c])^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))}} \sqrt{\prod_n \prod_{c|v_c\neq v'_c} (\theta_{n,c}[v_c]\theta_{n,c}[v'_c])^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))}} \\ &= 1 - \sum_{\mathbf{x}} \prod_n \prod_{c|v_c=v'_c} \theta_{n,c}[v]^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))} \sqrt{\prod_n \prod_{c|v_c\neq v'_c} (\theta_{n,c}[v_c]\theta_{n,c}[v'_c])^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))}} \\ &= 1 - \sum_{\mathbf{x}} \prod_n \prod_{c|v_c=v'_c} (\frac{1}{\mathsf{ch}(n)} + v_c\tau_n)^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))} \sqrt{\prod_n \prod_{c|v_c\neq v'_c} \left((\frac{1}{\mathsf{ch}(n)} + \tau_n)(\frac{1}{\mathsf{ch}(n)} - \tau_n)\right)^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))}} \\ &= 1 - \sum_{\mathbf{x}} \prod_n \prod_{c|v_c=v'_c} (\frac{1}{\mathsf{ch}(n)} + v_c\tau_n)^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))} \prod_n \prod_{c|v_c\neq v'_c} \sqrt{\left(\frac{1}{\mathsf{ch}(n)^2} - \tau_n^2\right)^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))}} \\ &\geq 1 - \sum_{\mathbf{x}} \prod_n \prod_{c|v_c=v'_c} (\frac{1}{\mathsf{ch}(n)} + v_c\tau_n)^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))} \prod_n \prod_{c|v_c\neq v'_c} \frac{1}{\mathsf{ch}(n)} \sqrt{\left(1 - \tau_n^2\mathsf{ch}(n)^2\right)^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))}}\end{aligned}$$

Using the trivial inequality $\sqrt{1-\mathbf{x}} \leq 1 - \frac{\mathbf{x}}{2}$,

$$\begin{aligned}
D_{\mathsf{H}}(P_v \| P_{v'}) &\geq 1 - \sum_{\mathbf{x}} \prod_n \prod_{c|v_c=v'_c} \left(\frac{1}{\mathsf{ch}(n)} + v_c\tau_n\right)^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))} \prod_n \prod_{c|v_c\neq v'_c} \frac{1}{\mathsf{ch}(n)} \left(1 - \frac{\tau_n^2 \mathsf{ch}(n)^2}{2}\right)^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))} \\
&= 1 - \sum_{\mathbf{x}} \prod_n \prod_{c|v_c=v'_c} \left(\frac{1}{\mathsf{ch}(n)} + v_c\tau_n\right)^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))} \prod_n \prod_{c|v_c\neq v'_c} \left(\frac{1}{\mathsf{ch}(n)} - \frac{\tau_n^2 \mathsf{ch}(n)}{2}\right)^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))} \\
&= 1 - \sum_{\mathbf{x}} \prod_n \prod_{c|v_c=v'_c} \left(\frac{1}{\mathsf{ch}(n)} + v_c\tau_n\right)^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))} \prod_{c|v_c\neq v'_c} \left(\frac{1}{\mathsf{ch}(n)} - \frac{\tau_n^2 \mathsf{ch}(n)}{2}\right)^{\mathbb{I}((n,c)\in\mathsf{tree}(\mathbf{x}))}
\end{aligned}$$

Now, for the same structured circuit define a different set of weights! This time, define each sum node parameter as follows:

$$\psi_{n,c} = \begin{cases} \frac{1}{\mathsf{ch}(n)} + v_c\tau_n & \text{if } v_c = v'_c \\ \frac{1}{\mathsf{ch}(c)} - \frac{\tau_n^2 \mathsf{ch}(n)}{2} & \text{if } v_c \neq v'_c \end{cases}$$

Then the above becomes more simple,

$$D_{\mathsf{H}}(P_v \| P_{v'}) \geq 1 - \sum_{\mathbf{x}} \prod_{(n,c)\in\mathsf{tree}(\mathbf{x})} \psi_{n,c}.$$

We have just returned to doing a simple forward pass through our circuit structure, albeit with modified weights this time! First, let's consider what happens at a single sum node n. If $v_c = v'_c$ then we know that our weight is of the form $\frac{1}{\mathsf{ch}(n)} + v_c\tau_n$; however, we set up the original weights in the PC as pairs, and as such we have a remaining $\frac{1}{\mathsf{ch}(n)} - v_c\tau_n$ term. Thus for all matching v_c, v'_c we have that

$$\sum_{v_c=v'_c} \frac{1}{\mathsf{ch}(n)} + v_c\tau_n + \frac{1}{\mathsf{ch}(n)} - v_c\tau_n = \sum_{v_c=v'_c} \frac{2}{\mathsf{ch}(n)}.$$

In the case that $v_c \neq v'_c$ the weight of our edge is $\frac{1}{\mathsf{ch}(c)} - \frac{\tau_n^2 \mathsf{ch}(n)}{2}$, which we have again paired off in order to maintain normalization in the original circuit. But, the pair is also equivalent to $\frac{1}{\mathsf{ch}(c)} - \frac{\tau_n^2 \mathsf{ch}(n)}{2}$. Thus for all non-matched v_c, v'_c we have:

$$\sum_{v_c\neq v'_c} \frac{1}{\mathsf{ch}(c)} - \frac{\tau_n^2 \mathsf{ch}(n)}{2} + \frac{1}{\mathsf{ch}(c)} - \frac{\tau_n^2 \mathsf{ch}(n)}{2} = \sum_{v_c\neq v'_c} \frac{2}{\mathsf{ch}(n)} - \tau_n^2 \mathsf{ch}(n).$$

Now, the total sum for this circuit is as follows:

$$\begin{aligned}
\sum_c \psi_{n,c} &= \sum_{v_c=v'_c} \frac{2}{\mathsf{ch}(n)} + \sum_{v_c\neq v'_c} \frac{2}{\mathsf{ch}(n)} - \tau_n^2 \mathsf{ch}(n) \\
&= \sum_{v_c=v'_c} \frac{2}{\mathsf{ch}(n)} + \sum_{v_c\neq v'_c} \frac{2}{\mathsf{ch}(n)} - \sum_{v_c\neq v'_c} \tau_n^2 \mathsf{ch}(n) \\
&= \left(\sum_{v_c=v'_c} \frac{2}{\mathsf{ch}(n)} + \sum_{v_c\neq v'_c} \frac{2}{\mathsf{ch}(n)}\right) - \sum_{v_c\neq v'_c} \tau_n^2 \mathsf{ch}(n)
\end{aligned}$$

Remember that we only perturbed $\mathsf{ch}(n)/2$ parameters, as such $\left(\sum_{v_c=v'_c} \frac{2}{\mathsf{ch}(n)} + \sum_{v_c\neq v'_c} \frac{2}{\mathsf{ch}(n)}\right) = \frac{\mathsf{ch}(n)}{2} \frac{2}{\mathsf{ch}(n)} = 1$. Thus, for any given sum node we have that:

$$\begin{aligned}
\sum_c \psi_{n,c} &= 1 - \sum_{v_c\neq v'_c} \tau_n^2 \mathsf{ch}(n) \\
&= 1 - \tau_n^2 \mathsf{ch}(n) \sum_c \mathbb{I}(v_c \neq v'_c)
\end{aligned}$$

This is not entirely surprising as we are only looking at a single sum node, which is effectively a categorical distribution. To then see how this propagates through the circuit we must deal with the product nodes.

$$\prod_n \psi_{n,c} \leq \prod_n (1 - \tau_n^2 \mathsf{ch}(n) \sum_c \mathbb{I}(v_c \neq v_c')).$$

Use the following helpful inequality: $1 - z \leq e^{-z}$.

$$\begin{aligned}\prod_n (1 - \tau_n^2 \mathsf{ch}(n) H_n(v_c, v_c')) &\leq \prod_n e^{-\tau_n^2 \mathsf{ch}(n) H_n(v_c, v_c')} \\ &= e^{-\sum_n \tau_n^2 \mathsf{ch}(n) H_n(v_c, v_c')}\end{aligned}$$

Now use our original assignment of $\tau_n^2 = \frac{16\epsilon^2}{S(w-1)^2}$, which results in,

$$\begin{aligned}e^{-\sum_n \tau_n^2 \mathsf{ch}(n) H_n(v_c, v_c')} &= e^{-\sum_n \frac{16\epsilon^2}{S(w-1)^2} \mathsf{ch}(n) H_n(v_c, v_c')} \\ &\geq e^{-\frac{16\epsilon^2}{S(w-1)} \sum_n H_n(v_c, v_c')}\end{aligned}$$

See now that $\sum_n H_n(v_c, v_c')$ is the total hamming distance across all sum nodes! Now enforcing that $\frac{16\epsilon^2}{S(w-1)} < 1$ we have that $1 - e^{-z} \geq \frac{z}{2}$ which gives us,

$$\begin{aligned}D_{\mathsf{H}}(P_v \| P_v')^2 &\geq e^{-\frac{16\epsilon^2}{S(w-1)} \sum_n H_n(v_c, v_c')} \\ &\geq \frac{8\epsilon^2}{S(w-1)} H(v, v').\end{aligned}$$

For the upper bound on $D_{\mathsf{H}}(P_v \| P_{v'})^2$, see the following:

$$\begin{aligned}D_{\mathsf{H}}(P_v \| P_{v'})^2 &= 1 - \sum_{\mathbf{x}} \sqrt{P_v(\mathbf{x}) P_{v'}(\mathbf{x})} \\ &= 1 - \sum_{\mathbf{x}} \sqrt{\prod_{v_c = v_c'} \theta_e[v]^2} \sqrt{(\frac{1}{\mathsf{ch}(n)^2} - \tau_n^2)} \\ &= 1 - \sum_{\mathbf{x}} \prod_{v_c = v_c'} \theta_e[v] \frac{1}{\mathsf{ch}(n)} \sqrt{(1 - \frac{16\epsilon^2}{S(w-1)^2} \mathsf{ch}(n)^2)}\end{aligned}$$

Suppose we have the worst case scenario and all $\mathsf{ch}(n) = w$, then as $\mathsf{ch}(n) \geq 2$, $\frac{w^2}{(w-1)^2} \leq 4$.

$$1 - \sum_{\mathbf{x}} \prod_{v_c = v_c'} \theta_e[v] \frac{1}{\mathsf{ch}(n)} \sqrt{(1 - \frac{64\epsilon^2}{S})}$$

Now, since

$$\sum_{\mathbf{x}} \prod_{v_c = v_c'} \theta_e[v] \frac{1}{\mathsf{ch}(n)} = 1,$$

we have that

$$1 - \sum_{\mathbf{x}} \prod_{v_c = v_c'} \theta_e[v] \frac{1}{\mathsf{ch}(n)} \sqrt{(1 - \frac{64\epsilon^2}{S})} = 1 - \sqrt{(1 - \frac{64\epsilon^2}{S})}$$

See that for $z \in [0, 1]$, $1 - \sqrt{(1 - z)} < z$, thus

$$D_{\mathsf{H}}(P_v \| P_{v'})^2 \leq \frac{64\epsilon^2}{S}.$$

□

A.5 SHORT EXPLANATION OF ASSOUAD'S LEMMA

In Tsybakov [21] see that Assouad's Lemma is as follows,

Lemma A.1 (Assouad's Lemma [21]). *Let $\Omega = \{0,1\}^r$ be the set of all binary sequences of length m. Let $\{P_\omega, \omega \in \Omega\}$ be a set of 2^r probability measures, and let the corresponding expectations be denoted by $\mathbb{E}_\omega$. Then,*

$$\inf_{\hat{\omega}} \max_{\omega \in \Omega} \mathbb{E}_\omega[H(\omega, \hat{\omega})] \geq \frac{r}{2} \min_{\omega,\omega': H(\omega,\omega')=1} \inf_{\psi} \left(P_\omega(\psi \neq 0) + P_{\omega'}(\psi \neq 1) \right)$$

where $H(\omega, \omega')$ is the Hamming distance between ω and ω', $\inf_{\hat{\omega}}$ denotes the infimum over all estimators taking values in Ω and where $\inf_\psi$ denotes the infimum over all tests taking values in $\{0,1\}$.

Tsybakov [21] makes the observation that

$$\inf_{\psi} \left(P_\omega(\psi \neq 0) + P_{\omega'}(\psi \neq 1) \right) = \int \min(dP_\omega, dP_{\omega'})$$

in the continuous case. Thus for the discrete case, this is equivalent to

$$\inf_{\psi} \left(P_\omega(\psi \neq 0) + P_{\omega'}(\psi \neq 1) \right) = \sum_x \min(P_\omega(\mathbf{x}), P_{\omega'}(\mathbf{x})).$$

From [21] see that using Le Cam's inequalities

$$\sum_{\mathbf{x}} \min(P_\omega(\mathbf{x}), P_{\omega'}(\mathbf{x})) \geq \frac{1}{2} (\sum_{\mathbf{x}} \sqrt{P_\omega(\mathbf{x}) P_{\omega'}(\mathbf{x})})^2.$$

Now, we are interested in the squared Hellinger distance, which is equivalent to $D_{\mathsf{H}}(P_\omega \| P_{\omega'})^2 = 1 - \sum_{\mathbf{x}} \sqrt{P_\omega(\mathbf{x}) P_{\omega'}(\mathbf{x})}$. As such, if for all $\omega, \omega' \in \{0,1\}^r$ we have that $D_{\mathsf{H}}(P_\omega \| P_{\omega'})^2 < \beta$ then we have that $\sum_{\mathbf{x}} \sqrt{P_\omega(\mathbf{x}) P_{\omega'}(\mathbf{x})} > (1-\beta)$. If we take this over $M \geq 1$ samples, we then have $\sum_{\mathbf{x}} \sqrt{P_\omega^{\otimes M}(\mathbf{x}) P_{\omega'}^{\otimes M}(\mathbf{x})} \geq (1-\beta)^M$

Now, to connect directly to the statement in Theorem 4.8 map $\omega, \omega' \in \{0,1\}^r$ to $v, v' \in \{-1,1\}^r$, and if we guarantee

1. for all $v, v' \in \{-1,1\}^r$, the distance between $P_v, P_{v'}$ is at least proportional to the Hamming distance:

$$D_{\mathsf{H}}(P_v \| P_{v'})^2 \geq \alpha H(v, v')$$

2. for all $v, v' \in \{-1,1\}^r$ with $H(v, v') = 1$, the squared Hellinger distance of $P_v, P_{v'}$ is small:

$$D_{\mathsf{H}}(P_v \| P_{v'})^2 \leq \beta,$$

then using the second condition,

$$\begin{aligned} \inf_{\hat{P}} \max_{P \in \Omega} \mathbb{E}_v[H(v, \hat{v})] &\geq \frac{r}{4} \left(\sum_{\mathbf{x}} \sqrt{P_\omega^{\otimes M}(\mathbf{x}) P_{\omega'}^{\otimes M}(\mathbf{x})} \right)^2 \\ &\geq \frac{r}{4} (1-\beta)^{2M}. \end{aligned}$$

Finally, using the first condition,

$$\inf_{\hat{P}} \max_{P \in \Omega} \mathbb{E}_v[D_{\mathsf{H}}(P_v \| P_{v'})^2] \geq \frac{\alpha r}{4} (1-\beta)^{2M}.$$