Approximate Modeling with Probabilistic Circuits

John Leland

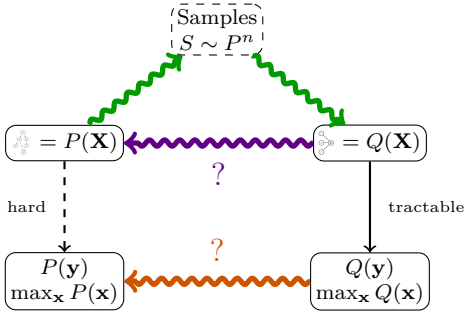
Introduction

- Exact representation of functions and probability distributions is well studied.
  - Size lower bounds deterministic and decomposable PCs [Bova et al., 2016, Choi et al., 2017].
  - NP-hardness
- Can we try to get around these hardness results by allowing for some error?
  - Previous function/distribution approximation work:
    - d-DNNFs [de Colnet et al., 2021]
    - OBDDs [Bollig et al., 2002]
    - Sum-Product Networks [Martens et al., 2014]
  - There is a **lot** of work on approximating inference/queries:
    - probabilistic models [Dagum et al., 1993, Conaty et al., 2017, Roth, 1996, Kwisthout, 2015]
    - approximate model counting [Meel et al., 2024, Chakraborty et al., 2021]
    - many more....

Questions

Lets expand the definition of approximation from Martens et al. [2014] from convergence to allowing bounded error, then the big questions are:

1. Does approximation reduce the complexity of modeling distributions?
2. If we use a tractable model, Q, to approximate a distribution, does inference on Q guarantee approximate inference?
3. Can we provide some approximation guarantees on learned circuits?



NeurIPS 2025

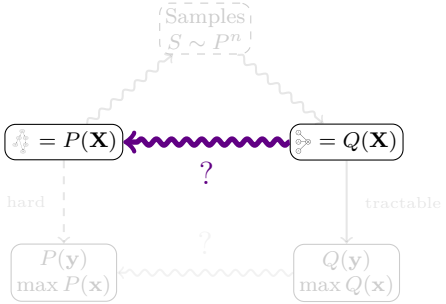
On the Hardness of Approximating Distributions with Tractable Probabilistic Models

John Leland **YooJung Choi**
School of Computing and Augmented Intelligence
Arizona State University
jslelan1@asu.edu, yj.choi@asu.edu

- Covers questions #1 and #2!

Questions

1. Does approximation reduce the complexity of modeling distributions?



Approximate Modeling

Definition (ϵ-D-Approximation)

Let P, Q be some probability measures and D be some distance we want to minimize between P and Q. We say that Q is an ϵ-D-Approximator of P if $D(P||Q) < \epsilon$ for some $\epsilon > 0$.

Approximate Modeling

Definition (ϵ-D-Approximation)

Let P, Q be some probability measures and D be some distance we want to minimize between P and Q. We say that Q is an ϵ-D-Approximator of P if $D(P||Q) < \epsilon$ for some $\epsilon > 0$.

- Choose the family of k-convex f-divergences to be D.
  - Includes total variation distance, squared Hellinger, KL-divergence, etc.

Hardness of D_f-Approximation

Theorem 1

Given a (potentially unnormalized) probability distribution $\hat{P}$ and a k-convex f-divergence D_f, for any $0 < \epsilon < \frac{1}{4}$, it is NP-hard to model the $k\epsilon^2$-D_f-approximation of its normalized distribution P as a model that can tractably compute marginals.

- This is proved by a reduction from $\mathsf{SAT}(\hat{P})$

Hardness of D_f-Approximation

Theorem 1

Given a (potentially unnormalized) probability distribution $\hat{P}$ and a k-convex f-divergence D_f, for any $0 < \epsilon < \frac{1}{4}$, it is NP*-hard to model the $k\epsilon^2$-D_f-approximation of its normalized distribution P as a model that can tractably compute marginals.*

- This is proved by a reduction from $\mathsf{SAT}(\hat{P})$
  - Define another Boolean function $\hat{P}'$ such that $\hat{P}' = (Y \wedge \hat{P}) \vee (\neg Y \wedge X_1 \wedge \cdots \wedge X_n)$.

Hardness of D_f-Approximation

Theorem 1

Given a (potentially unnormalized) probability distribution $\hat{P}$ and a k-convex f-divergence D_f, for any $0 < \epsilon < \frac{1}{4}$, it is NP*-hard to model the $k\epsilon^2$-D_f-approximation of its normalized distribution P as a model that can tractably compute marginals.*

- This is proved by a reduction from $\mathsf{SAT}(\hat{P})$
  - Define another Boolean function $\hat{P}'$ such that $\hat{P}' = (Y \wedge \hat{P}) \vee (\neg Y \wedge X_1 \wedge \cdots \wedge X_n)$.
  - Place a uniform distribution, P, over the models of $\hat{P}'$.

Hardness of D_f-Approximation

Theorem 1

Given a (potentially unnormalized) probability distribution $\hat{P}$ and a k-convex f-divergence D_f, for any $0 < \epsilon < \frac{1}{4}$, it is NP-hard to model the $k\epsilon^2$-D_f-approximation of its normalized distribution P as a model that can tractably compute marginals.

- This is proved by a reduction from $\mathsf{SAT}(\hat{P})$
  - Define another Boolean function $\hat{P}'$ such that $\hat{P}' = (Y \wedge \hat{P}) \vee (\neg Y \wedge X_1 \wedge \cdots \wedge X_n)$.
  - Place a uniform distribution, P, over the models of $\hat{P}'$.
  - Let Q be a model that can tractably marginalize and is a $1/4$-D_{TV}-estimator, then we can solve $\mathsf{SAT}(\hat{P})$ by thresholding $Q(Y = 1) \geq 1/4$.

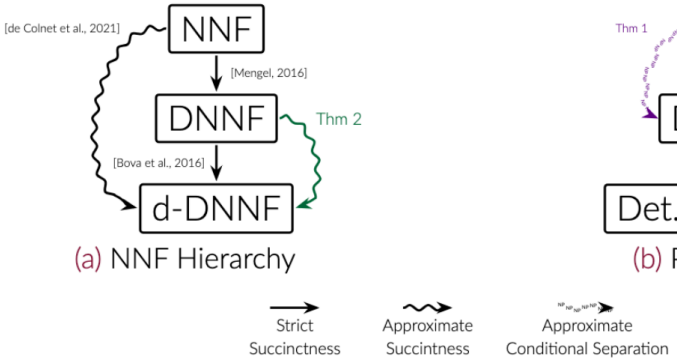
Hardness of D_f-Approximation

Theorem 1

Given a (potentially unnormalized) probability distribution $\hat{P}$ and a k-convex f-divergence D_f, for any $0 < \epsilon < \frac{1}{4}$, it is NP-hard to model the $k\epsilon^2$-D_f-approximation of its normalized distribution P as a model that can tractably compute marginals.

- This is proved by a reduction from $\mathsf{SAT}(\hat{P})$
  - Define another Boolean function $\hat{P}'$ such that $\hat{P}' = (Y \wedge \hat{P}) \vee (\neg Y \wedge X_1 \wedge \cdots \wedge X_n)$.
  - Place a uniform distribution, P, over the models of $\hat{P}'$.
  - Let Q be a model that can tractably marginalize and is a $1/4$-D_{TV}-estimator, then we can solve $\mathsf{SAT}(\hat{P})$ by thresholding $Q(Y = 1) \geq 1/4$.
- This result applies to **any model that can tractably marginalize**!

Large Deterministic and Decomposable Approximators



Large Deterministic and Decomposable Approximators

Theorem 2

A deterministic, decomposable PC that is a ϵ-D_{TV}-Approximator of P_n, where $\epsilon = \frac{1}{16} - \Omega(1/Poly(n^2))$, has size $2^{\Omega(n)}$.

- Proof Roadmap:

Large Deterministic and Decomposable Approximators

Theorem 2

A deterministic, decomposable PC that is a ϵ-D_{TV}-Approximator of P_n, where $\epsilon = \frac{1}{16} - \Omega(1/Poly(n^2))$, has size $2^{\Omega(n)}$.

- Proof Roadmap:
  - Deterministic, decomposable D_{TV}-approximator $\implies$ weak approximator of Boolean function.

Large Deterministic and Decomposable Approximators

Theorem 2

A deterministic, decomposable PC that is a ϵ*-*D_{TV}*-Approximator of* P_n*, where* $\epsilon = \frac{1}{16} - \Omega(1/Poly(n^2))$*, has size* $2^{\Omega(n)}$*.*

- Proof Roadmap:
  - Deterministic, decomposable D_{TV}-approximator $\implies$ weak approximator of Boolean function.
  - Exists a family of functions P_n (derived from the Sauerhoff function) that are modeled by decomposable PCs in size $O(n^2)$

Large Deterministic and Decomposable Approximators

Theorem 2

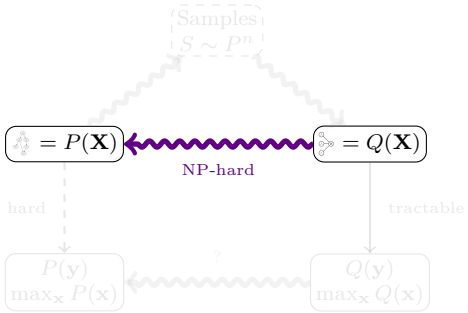
A deterministic, decomposable PC that is a ϵ-D_{TV}-Approximator of P_n, where $\epsilon = \frac{1}{16} - \Omega(1/Poly(n^2))$, has size $2^{\Omega(n)}$.

- Proof Roadmap:
  - Deterministic, decomposable D_{TV}-approximator $\implies$ weak approximator of Boolean function.
  - Exists a family of functions P_n (derived from the Sauerhoff function) that are modeled by decomposable PCs in size $O(n^2)$
  - Any deterministic, decomposable PC approximating P_n must be exponentially sized!

Questions

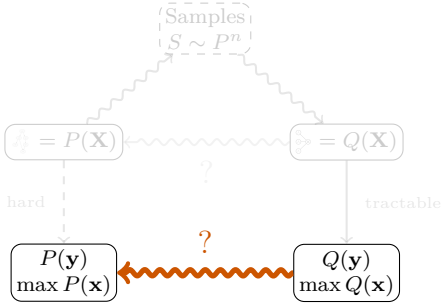
1. Does approximation reduce the complexity of modeling distributions?

   A: No :(



Questions

2. If we use a tractable model, Q, to approximate a distribution, does inference on Q guarantee approximate inference?



Inferences Guarantees in Approximate Modeling

Guaranteed:

- Absolute approximation of marginals and MAP

Not guaranteed:

- Relative approximation of marginals and MAP
- Relative and absolute approximation of conditionals

Absolute approximation: $\forall x, y : |x - y| < \epsilon$
Relative approximation: $\forall x, y : \frac{1}{1+\epsilon} < \frac{x}{y} < 1$

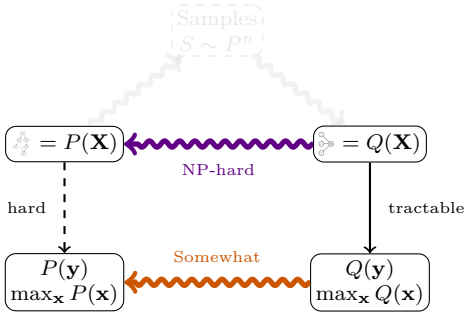
Questions

1. Does approximation reduce the complexity of modeling distributions?

   A: No :(

2. If we use a tractable model, Q, to approximate a distribution, does inference on Q guarantee approximate inference?

   A: Sometimes



Other directions for Approximation

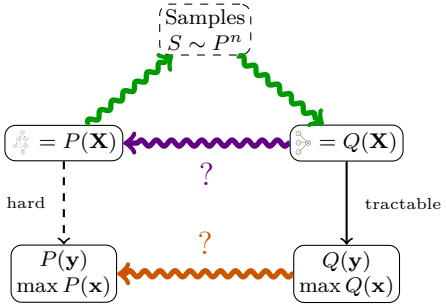
- Other circuit classes?
- (ϵ, δ) guarantees?
- Learning PCs with approximation guarantees?

Other directions for Approximation

- Other circuit classes?
- (ϵ, δ) guarantees?
- Learning PCs with approximation guarantees?

Questions

3. Can we provide some approximation guarantees on learned circuits?



Arxiv Version

- Preliminary work that covers question #3!

Short Introduction to Sample Complexity

- Given samples from a true distribution P
- Want to learn a PC $\hat{P}$ from these samples
- With enough samples can we guarantee $\Pr[D(\hat{P}\|P) \leq \epsilon] > 1 - \delta$?

Upper Bounds: **There exists** an algorithm that must have at least M samples to learn such that $\Pr[D(\hat{P}\|P) \leq \epsilon] > 1 - \delta$.

Lower Bounds: **Any algorithm** must have at least M samples to learn such that $\Pr[D(\hat{P}\|P) \leq \epsilon] > 1 - \delta$.

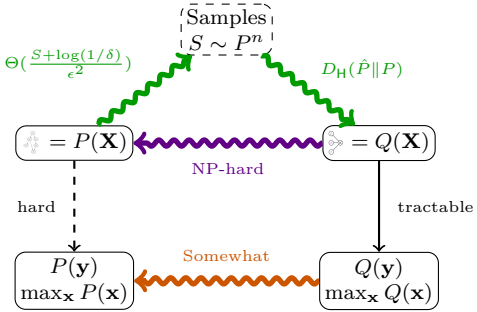
Results

- Set $\hat{P}$ and P to be deterministic and decomposable PCs
  - Boolean indicator leaves and exact same **fixed** structure
  - Only a different parameterization
- Upper bound comes from the closed form MLE for deterministic and decomposable sum nodes.

Can guarantee that learning $\hat{P}$ such that $\Pr[D_{\mathsf{H}}(\hat{P} \| P) \leq \epsilon] > 1 - \delta$ requires $\Theta(\frac{S + \log(1/\delta)}{\epsilon^2})$ samples, where S is the number of sum nodes in the circuit.

Wrap

1. Does approximation reduce the complexity of modeling distributions?
   A: No :(
2. If we use a tractable model, Q, to approximate a distribution, does inference on Q guarantee approximate inference?
   A: Sometimes
3. Can we provide some approximation guarantees on learned circuits?
   A: Yes! Depending on the number of samples and distance.



Future Directions for Approximate Modeling

- Drop determinism.
  - There are no lower bound techniques for DNNF's approximating functions.
  - Also would be interesting for structure decomposable.
- Learning the structures using samples?
- Continuous distributions?
- Does there exist a distance measure which guarantees relative approximation?

References I

Bollig et al. On the nonapproximability of boolean functions by obdds and read-k-times branching programs. *Information and Computation*, 2002.

Bova et al. Knowledge compilation meets communication complexity. In *IJCAI*, 2016.

Chakraborty et al. Chapter 26. Approximate Model Counting. In *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2021.

Choi et al. On relaxing determinism in arithmetic circuits. In *ICML*, 2017.

Conaty et al. Approximation complexity of maximum a posteriori inference in sum-product networks. *arXiv preprint arXiv:1703.06045*, 2017.

Dagum et al. Approximating probabilistic inference in bayesian belief networks is np-hard. *Artificial intelligence*, 60, 1993.

References II

de Colnet et al. Lower bounds for approximate knowledge compilation. In *KR*, 2021.

Kwisthout. Tree-Width and the Computational Complexity of MAP Approximations in Bayesian Networks. *Journal of Artificial Intelligence Research*, 53, 2015. ISSN 1076-9757.

Martens et al. On the expressive efficiency of sum product networks. *arXiv preprint arXiv:1411.7717*, 2014.

Meel et al. An fpras for model counting for non-deterministic read-once branching programs. *arXiv preprint arXiv:2406.16515*, 2024.

Roth. On the hardness of approximate reasoning. *Artificial intelligence*, 1996.

Sample Complexity Intuition

Upper Bound

- Decompose KL-divergence into KL-divergences of individual sum nodes.
- Upper bound $\Pr[D_{\mathsf{KL}}(\hat{P}\|P) \geq \epsilon]$ using a Chernoff Bound.
- Using $D_{\mathsf{H}}(\hat{P}\|P) \leq \sqrt{\frac{1}{2} D_{\mathsf{KL}}(\hat{P}\|P)}$ and $D_{\mathsf{TV}}(\hat{P}\|P) \leq \sqrt{\frac{1}{2} D_{\mathsf{KL}}(\hat{P}\|P)}$ we get upper bounds for $D_{\mathsf{H}}(\hat{P}\|P), D_{\mathsf{TV}}(\hat{P}\|P)$, and $D_{\mathsf{KL}}(\hat{P}\|P)$.

Lower Bound

- Construct a family of PCs using the Hamming cube
- Apply Assouad's lemma to get a lower bound
- Use the hardness of estimating the bias of a coin for the remaining term
- Doing this for squared Hellinger gets us tight bounds on $D_{\mathsf{H}}(\hat{P}\|P)$ and $D_{\mathsf{KL}}(\hat{P}\|P)$, with a slightly loose bound on $D_{\mathsf{TV}}(\hat{P}\|P)$.