Learnability of Parameter-Bounded Bayes Nets

Arnab Bhattacharyya[1] Davin Choo[1]
Sutanu Gayen[2] Dimitrios Myrisiotis[3]
[1] National University of Singapore
[2] Indian Institute of Technology Kanpur
[3] CNRS@CREATE LTD.

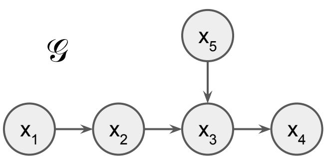
Scan QR for full paper

Central Question

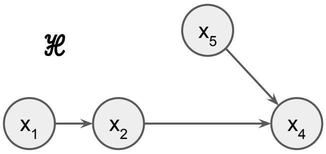
Given a "succinct" description of a distribution $\mathbb{P}$ and a number p, how easy is it to find a Bayes net $\mathcal{G}$ of at most p parameters such that $\mathcal{G}$ represents $\mathbb{P}$?

Bayesian networks (Bayes Nets)

- Any distribution $\mathbb{P}$ on n nodes $\mathbf{X} = \{X_1, \dots, X_n\}$ can be described by 2^n-1 entries in a lookup table.
- Bayes nets provide a succinct way of representing high-dimensional distributions and are defined by
  - a directed acyclic graph (DAG);
  - a collection of conditional probability distributions, one for each node in the DAG.



- Any distribution represented by $\mathcal{G}$ can be described by $10 < 2^5 - 1 = 31$ parameters:
  - 1 number for $\mathbb{P}(X_1)$;
  - 1 number for $\mathbb{P}(X_5)$;
  - 2 numbers for $\mathbb{P}(X_2 \mid X_1)$;
  - 4 numbers for $\mathbb{P}(X_3 \mid X_2, X_5)$;
  - 2 numbers for $\mathbb{P}(X_4 \mid X_3)$;
  - e.g., we can deduce $\mathbb{P}(X_1 = 1)$ from $\mathbb{P}(X_1 = 0)$.



- If $\mathbb{Q}(x_1,x_2,x_4,x_5) = \sum_{x3} \mathbb{P}(x_1,\dots,x_5)$ is a distribution on $\{X_1, X_2, X_4, X_5\}$ obtained by marginalizing out X_3 from $\mathcal{G}$, then $\mathbb{Q}$ is represented by $\mathcal{H}$.

In-degree versus parameters

- While one can upper bound complexity of a Bayes net by its maximum in-degree d, the number of parameters is more fine-grained.
  - A star: $O(n + 2^d)$ parameters;
  - A clique: $\Omega(n \cdot 2^d)$ parameters.
- "Succinct representation" of [CHM04]:
  - Distribution $\mathbb{P}$ is a marginal of a Bayes net of small maximum in-degree.

Some related work

- [CH92, SDLC93, HGC95] studied the problem of learning the underlying DAG of a Bayes net from data, by focusing on maximizing certain scoring criterion by the underlying DAG.
- This task was later shown to be NP-hard [Chi96].
- [CHM04] showed that deciding whether a given distribution $\mathbb{P}$ can be represented by some Bayes net of at most p parameters or not is NP-hard.
- There are well-known algorithms for learning the underlying DAG of a Bayes net from distributional samples such as the PC [SGS00] and GES [Chi02] algorithms.
- More recently, [BCD20] gave finite sample guarantees of learning Bayes nets that have n nodes, each taking values over an alphabet Σ, using samples from $\mathbb{P}$.

NP-hardness result of [CHM04]

- The `DBFAS` decision problem:
  - Given a directed graph $\mathcal{G} = (\mathbf{X}, \mathbf{E})$ with maximum vertex degree of 3, and a positive integer $k \leq |\mathbf{E}|$, determine whether there is a subset of edges $\mathbf{E}' \subseteq \mathbf{E}$ with of size $|\mathbf{E}'| \leq k$ such that $\mathbf{E}'$ contains at least one directed edge from every directed cycle in $\mathcal{G}$.
  - [Gav77] showed that `DBFAS` is NP-hard.
- The `LEARN` decision problem:
  - Given variables $\mathbf{X} = (X_1, \dots, X_n)$, a probability distribution $\mathbb{P}$ over $\mathbf{X}$, and a parameter bound p, determine whether there exists a Bayes net $\mathcal{G}$ with at most p parameters such that $\mathcal{G}$ represents $\mathbb{P}$.
  - [CHM04] showed that `LEARN` is NP-hard via reduction from `DBFAS`.
- Note that any distribution can be represented by some Bayes net over the complete DAG, since there are no d-separations implied by this kind of DAG; such a Bayes net over a complete DAG requires $2^{|\mathbf{X}|}$ - 1 parameters to describe.
- We define `LEARN-DBFAS` as the set of instances of `LEARN` that are in the range of the reduction of [CHM04] from `DBFAS` to `LEARN`.
- [CHM04] showed that `LEARN-DBFAS` is NP-hard, even when given access to an independence oracle for $\mathbb{P}$.
  - "Succinct representation" of [CHM04]: Distribution $\mathbb{P}$ is a marginal of a Bayes net of small maximum in-degree.

The `REALIZABLE-LEARN` problem

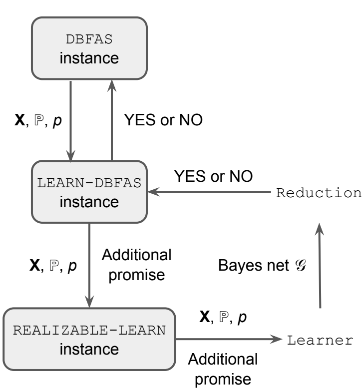
The `LEARN-DBFAS` decision problem with the <u>additional promise</u> that there exists a Bayes net $\mathcal{G}$ with at most p parameters such that $\mathcal{G}$ represents $\mathbb{P}$.

Result 1

`REALIZABLE-LEARN` is NP-hard

- Technical overview (see diagram below):

We show that *if* there exists some blackbox polynomial time algorithm `Learner` for `REALIZABLE-LEARN`, *then* there is a polynomial time algorithm `Reduction` that correctly answers `LEARN-DBFAS`. Therefore, `REALIZABLE-LEARN` is also NP-hard.



Result 2

Fix any accuracy and confidence parameters $\varepsilon > 0$ and $\delta > 0$. Given sample access to a distribution $\mathbb{P}$ over n variables, each defined on the alphabet Σ, and the promise that there is a Bayes net with at most p parameters that represents $\mathbb{P}$,

$$\mathcal{O}\left(\frac{\log \frac{1}{\delta}}{\varepsilon^2}\left(p \log\left(\frac{n\,|\Sigma|}{\varepsilon}\right) + n \frac{\log\left(\frac{p}{n(|\Sigma|-1)}\right)}{\log|\Sigma|} \log n\right)\right)$$

IID samples from $\mathbb{P}$ suffice to learn a distribution $\mathbb{Q}$ defined on DAG with $\leq p$ parameters such that $d_{TV}(\mathbb{P},\mathbb{Q}) \leq \varepsilon$, with success probability $\geq 1 - \delta$.

- Result 2 generalizes the finite sample result of [BCD20] from the degree-bounded setting to a parameter-bounded setting.
- Technical overview:
  - Construct an ε-net over all possible DAGs that satisfy the parameter upper bound p.
  - Apply a well-known technique from the density estimation literature called "Scheffé tournament;" see [DK14].
  - By a counting argument, there are not many possible DAGs that give rise to some Bayes net of at most p parameters.
  - By a counting argument, there are only a few conditional distributions that can be represented by a Bayes net $\mathcal{G}$ over a DAG that realizes a given in-degree sequence.
  - Thus, we can bound the number of distributions that cover all conditional distributions which can be represented by a Bayes net over the DAG of $\mathcal{G}$.
- Note that this result is only sample-efficient but not time-efficient since there are exponentially many candidates in the tournament.

Open problem

Suppose we are given sample access to a distribution $\mathbb{P}$ and are promised that there exists a Bayes net on $\mathcal{G}$ with at most p parameters such that $\mathcal{G}$ represents $\mathbb{P}$. Is it hard to find a Bayes net $\mathcal{G}'$ that has $\alpha \cdot p$ parameters such that $\mathcal{G}'$ represents $\mathbb{P}$ (where $\mathcal{G}'$ may not be $\mathcal{G}$), for some constant $\alpha > 1$?

References

[Gav77] Fanica Gavril. *Some NP-complete problems on graphs.* [CH92] Gregory F Cooper and Edward Herskovits. *A Bayesian method for the induction of probabilistic networks from data.* [SDLC93] David J Spiegelhalter, A Philip Dawid, Steffen L Lauritzen, and Robert G Cowell. *Bayesian analysis in expert systems.* [HGC95] David Heckerman, Dan Geiger, and David Maxwell Chickering. *Learning Bayesian networks: The combination of knowledge and statistical data.* [Chi96] David Maxwell Chickering. *Learning Bayesian networks is NP-complete.* [SGS00] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search.* [Chi02] David Maxwell Chickering. *Optimal structure identification with greedy search.* [CHM04] Max Chickering, David Heckerman, and Chris Meek. *Large-sample learning of Bayesian networks is NP-hard.* [DK14] Constantinos Daskalakis and Gautam Kamath. *Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians.* [BCD20] Johannes Brustle, Yang Cai, and Constantinos Daskalakis. *Multi-item mechanisms without item-independence: Learnability via robustness.*

Learnability of Parameter-Bounded Bayes Nets

Arnab Bhattacharyya[1] Davin Choo[1]
Sutanu Gayen[2] Dimitrios Myrisiotis[3]
[1] National University of Singapore
[2] Indian Institute of Technology Kanpur
[3] CNRS@CREATE

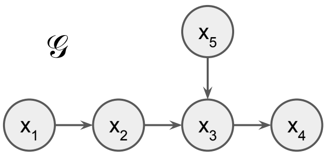
Scan QR for full paper

Central Question

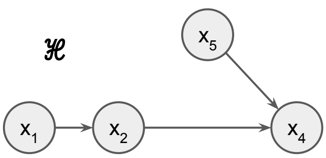
Given a "succinct" description of a distribution $\mathbb{P}$ and a number *p*, how easy is it to find a Bayes net $\mathcal{G}$ of at most *p* parameters such that $\mathcal{G}$ represents $\mathbb{P}$?

Bayesian networks (Bayes Nets)

- Any distribution $\mathbb{P}$ on *n* nodes $\mathbf{X} = \{X_1, \dots, X_n\}$ can be described by 2^n-1 entries in a lookup table
- Bayes nets provide a succinct way of representing high-dimensional distributions and are defined by:
  - A directed acyclic graph (DAG)
  - A collection of conditional probability distributions, one for each node in the DAG



- Any distribution represented by $\mathcal{G}$ can be described by 10 < 2^5-1 = 31 parameters
  - 1 number for $\mathbb{P}(X_1)$
  - 1 number for $\mathbb{P}(X_5)$
  - 2 numbers for $\mathbb{P}(X_2 \mid X_1)$
  - 4 numbers for $\mathbb{P}(X_3 \mid X_2, X_5)$
  - 2 numbers for $\mathbb{P}(X_4 \mid X_3)$
  - e.g., can deduce $\mathbb{P}(X_1 = 1)$ from $\mathbb{P}(X_1 = 0)$



- If $\mathbb{Q}(x_1,x_2,x_4,x_5) = \sum_{x_3} \mathbb{P}(x_1,\dots,x_5)$ is a distribution on $\{X_1, X_2, X_4, X_5\}$ obtained by marginalizing out X_3 from $\mathcal{G}$, then $\mathbb{Q}$ is represented by $\mathcal{H}$

In-degree versus parameters

- While one can upper bound complexity of a Bayes net by its maximum in-degree d, the number of parameters is more fine-grained
  - A star: $O(n + 2^d)$ parameters
  - A clique: $O(n \cdot 2^d)$ parameters
- "Succinct representation" of [CHM04]
  - Distribution $\mathbb{P}$ is a marginal of a Bayes net of small maximum in-degree

Some related work

- [CH92, SDLC93, HGC95] studied the problem of learning the underlying DAG of a Bayes net from data, by focusing on maximizing certain scoring criterion by the underlying DAG.
- This task was later shown to be NP-hard [Chi96].
- [CHM04] showed that deciding whether a given distribution $\mathbb{P}$ can be represented by some Bayes net of at most p parameters or not is NP-hard.
- There are well-known algorithms for learning the underlying DAG of a Bayes net from distributional samples such as the PC [SGS00] and GES [Chi02] algorithms.
- More recently, [BCD20] gave finite sample guarantees of learning Bayes nets that have n nodes, each taking values over an alphabet Σ, using samples from $\mathbb{P}$.

NP-hardness result of [CHM04]

- The `DBFAS` decision problem
  - Given a directed graph $\mathcal{G}$ = ($\mathbf{X}$, $\mathbf{E}$)$ with maximum vertex degree of 3, and a positive integer k ≤ |$\mathbf{E}$|, determine whether there is a subset of edges $\mathbf{E}' \subseteq \mathbf{E}$ with of size $|\mathbf{E}'| \leq k$ such that $\mathbf{E}'$ contains at least one directed edge from every directed cycle in $\mathcal{G}$.
  - [Gav77] showed that `DBFAS` is NP-hard.
- The `LEARN` decision problem
  - Given variables $\mathbf{X} = (X_1, \dots, X_n)$, a probability distribution $\mathbb{P}$ over $\mathbf{X}$, and a parameter bound p, determine whether there exists a Bayes net $\mathcal{G}$ with at most p parameters such that $\mathbb{P}$ is Markov with respect to $\mathcal{G}$.
  - Markov = ??
  - [CHM04] showed that `LEARN` is NP-hard via reduction from `DBFAS`
- What is "Markov"?
  - A probability distribution $\mathbb{P}$ is said to be Markov with respect to a DAG $\mathcal{G}$ if d-separation (some graphical notion) in $\mathcal{G}$ implies conditional independence in $\mathbb{P}$.
  - Note that any distribution is Markov with respect to some Bayes net over the complete DAG, since there are no d-separations implied by this kind of DAG; such a Bayes net over a complete DAG requires $2^{|\mathbf{X}|}$ - 1 parameters to describe.
- We define `LEARN-DBFAS` as the set of instances of `LEARN` that are in the range of the reduction of [CHM04] from `DBFAS` to `LEARN`
- [CHM04] showed that `LEARN-DBFAS` is NP-hard, even when given access to an independence oracle for $\mathbb{P}$.
  - "Succinct representation" of [CHM04]: Distribution $\mathbb{P}$ is a marginal of a Bayes net of small maximum in-degree

The `REALIZABLE-LEARN` problem

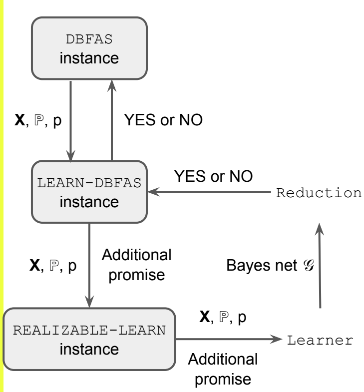
The `LEARN-DBFAS` decision problem with the <u>additional promise</u> that there exists a Bayes net $\mathcal{G}$ with at most p parameters such that $\mathbb{P}$ is Markov with respect to $\mathcal{G}$

Result 1

`REALIZABLE-LEARN` is NP-hard

- Technical overview (see diagram below)

We show that *if* there exists some blackbox polynomial time algorithm `Learner` for `REALIZABLE-LEARN`, then there is a polynomial time algorithm `Reduction` that correctly answers `LEARN-DBFAS`. Therefore, `REALIZABLE-LEARN` is also NP-hard.



Result 2

Fix any accuracy and confidence parameters ε > 0 and δ > 0. Given sample access to a distribution $\mathbb{P}$ over n variables, each defined on the alphabet Σ, and the promise that there is a Bayes net with at most p parameters that represents $\mathbb{P}$,

$$\mathcal{O}\left(\frac{\log\frac{1}{\delta}}{\varepsilon^2}\left(p\log\left(\frac{n|\Sigma|}{\varepsilon}\right)+n\frac{\log\left(\frac{p}{n(|\Sigma|-1)}\right)}{\log|\Sigma|}\log n\right)\right)$$

IID samples from $\mathbb{P}$ suffice to learn a distribution $\mathbb{Q}$ defined on DAG with ≤ p parameters such that $d_{TV}(\mathbb{P},\mathbb{Q}) \leq \varepsilon$, with success probability ≥ 1 - δ.

- Result 2 generalizes the finite sample result of [BCD20] from the degree-bounded setting to a parameter-bounded setting
- Technical overview
  - Construct an ε-net over all possible DAGs that satisfy the parameter upper bound p
  - Apply a well-known technique from the density estimation literature called "Scheffé tournament"; see [DK14].
  - By counting argument, there are not many possible DAGs that give rise to some Bayes net of at most p parameters.
  - By counting argument, there are only a few conditional distributions that are Markov with respect to a Bayes net $\mathcal{G}$ over a DAG that realizes a given in-degree sequence.
  - Thus, one can bound the number of distributions that cover all possible conditional distributions which are Markov with respect to $\mathcal{G}$.
- Note that this result is only sample-efficient but not time-efficient since there are exponentially many candidates in the tournament.

Open problem

Suppose we are given sample access to a distribution $\mathbb{P}$ and are promised that there exists a Bayes net on $\mathcal{G}$ with at most p parameters such that $\mathbb{P}$ is Markov with respect to $\mathcal{G}$. Is it hard to find a Bayes net $\mathcal{G}'$ that has α · p parameters such that $\mathbb{P}$ is Markov with respect to $\mathcal{G}'$ (where $\mathcal{G}'$ may not be $\mathcal{G}$), for some constant α > 1?

References

[Gav77] Fanica Gavril. *Some NP-complete problems on graphs.* [CH92] Gregory F Cooper and Edward Herskovits. *A Bayesian method for the induction of probabilistic networks from data.* [SDLC93] David J Spiegelhalter, A Philip Dawid, Steffen L Lauritzen, and Robert G Cowell. *Bayesian analysis in expert systems.* [HGC95] David Heckerman, Dan Geiger, and David Maxwell Chickering. *Learning Bayesian networks: The combination of knowledge and statistical data.* [Chi96] David Maxwell Chickering. *Learning Bayesian networks is NP-complete.* [SGS00] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search.* [Chi02] David Maxwell Chickering. *Optimal structure identification with greedy search.* [CHM04] Max Chickering, David Heckerman, and Chris Meek. *Large-sample learning of Bayesian networks is NP-hard.* [DK14] Constantinos Daskalakis and Gautam Kamath. *Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians.* [BCD20] Johannes Brustle, Yang Cai, and Constantinos Daskalakis. *Multi-item mechanisms without item-independence: Learnability via robustness.*