Total Variation Distance Meets Probabilistic Inference

Arnab Bhattacharyya[1], Sutanu Gayen[2], Kuldeep S. Meel[3], Dimitrios Myrisiotis[4], A. Pavan[5], and N. V. Vinodchandran[6]

[1]National University of Singapore; [2]Indian Institute of Technology Kanpur; [3]University of Toronto; [4]CNRS@CREATE LTD.; [5]Iowa State University; [6]University of Nebraska-Lincoln

July 15, 2024

What Is All About

What Is All About

We give a novel reduction from total variation distance estimation (for Bayes nets) to probabilistic inference (for Bayes nets).

Bayes Nets

Bayes Nets

Bayes nets (Pearl, 1989) offer a succinct way of representing high-dimensional distributions.

They are defined by a DAG and a collection of conditional probability distributions, one for each DAG node.

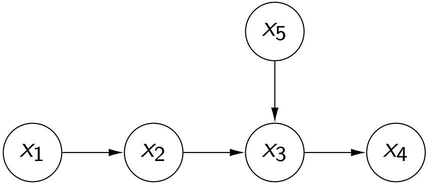

Figure: A Bayes net $\mathcal{G}$.

Bayes Nets

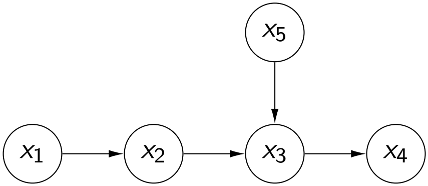

Figure: A Boolean Bayes net $\mathcal{G}$.

> The distribution represented by $\mathcal{G}$ can be described by a look-up table consisting of $2^5 - 1 = 31$ numbers,

while

> the description of $\mathcal{G}$ uses only 10 numbers (that is, 1 number for each of the distributions of x_1 and x_5, 2 numbers for each of the conditional probability distributions of x_2 and x_4, and 4 numbers for that of x_3).

Total Variation (TV) Distance

Total Variation (TV) Distance

There are many notions of distance between distributions, such as f-divergences (Hellinger, KL, χ^2, etc.) or integral probability metrics (Wasserstein, TV, etc.).

We focus on TV distance.

Total Variation (TV) Distance

Definition

For distributions P, Q over a common domain D, the *TV distance between P and Q* is

$$d_{\mathrm{TV}}(P, Q) := \sup_{A \subseteq D} |P(A) - Q(A)| .$$

Total Variation (TV) Distance

TV distance is important, because

1. it is natural: $d_{\mathrm{TV}}(P, Q)$ is equal to the maximum gap between the probabilities assigned by P and Q to a single event;
2. it has many desirable properties: It is a metric, it is bounded in $[0, 1]$, and is invariant with respect to bijections.

Probabilistic Inference

Probabilistic Inference

The following notion is a fundamental computational task with a wide range of applications.

Definition

Given random variables $X_1, \ldots, X_n$ and sets $S_1, \ldots, S_n$, such that for all $1 \leq i \leq n$ the set S_i is a subset of the range of X_i, compute

$$\mathbf{Pr}[X_1 \in S_1, \ldots, X_n \in S_n] .$$

Some Related Work

Some Related Work

- Goldreich, Sahai, and Vadhan (CRYPTO 1999, JACM 2003) showed that TV distance is hard to additively estimate for distributions samplable by Boolean circuits.

Some Related Work

- On an algorithmic note, Bhattacharyya, Gayen, Meel, and Vinodchandran (NeurIPS 2022) designed efficient algorithms to additively estimate the TV distance between distributions efficiently samplable and efficiently computable.

Some Related Work

- Bhattacharyya, Gayen, Meel, Myrisiotis, Pavan, and Vinodchandran (IJCAI 2023) proved that exact computation of TV distance between product distributions is $\#$P-hard.

  They also gave an FPTAS for estimating $d_{\mathrm{TV}}(P, Q)$ for the case where Q has a bounded number of distinct marginals (for example, when Q is the uniform distribution).

Some Related Work

- Feng, Guo, Jerrum, and Wang (TheoretiCS 2023) designed an FPRAS for multiplicatively estimating the TV distance between any two product distributions, and Feng, Liu, and Liu (SODA 2024) gave an FPTAS for the same task.

Our Results

Our Results: Enter Probabilistic Inference

Theorem

For any class $\mathcal{C}$ of Bayes nets for which there is an efficient probabilistic inference algorithm, there is an FPRAS for estimating the TV distance between any two Bayes nets from $\mathcal{C}$ defined over the same DAG.

Our Results: Enter Probabilistic Inference

We immediately get the following, by the (folklore) fact that probabilistic inference is efficient for Bayes nets of small treewidth (using the Variable Elimination algorithm).

Corollary

There is an FPRAS for estimating the TV distance between any two Bayes nets of treewidth $O(\log n)$ defined over the same DAG of n nodes.

Techniques:
Power From Couplings

Techniques: Power From Couplings

Definition

A *coupling* $\mathcal{C}$ *between distributions* P, Q is a joint distribution (X, Y) such that $X \sim P$ and $Y \sim Q$. We say that *a coupling* $\mathcal{O}$ *is optimal* if $\mathcal{O}$ is a coupling and

$$\Pr_{\mathcal{O}}[X = Y = w] = \min(P(w), Q(w))$$

for all w.

Techniques: Power From Couplings

Couplings and TV distance

A straightforward way of estimating TV distance is to make use of its characterization that uses optimal couplings.

That is, for $X \sim P$, $Y \sim Q$, and optimal coupling $\mathcal{O}$, we have

$$d_{\mathrm{TV}}(P, Q) = \Pr_{\mathcal{O}}[X \neq Y] \,.$$

Techniques: Power From Couplings

Problem

What is $\mathcal{O}$? It is not clear how to find it!

Solution

Circumvent this issue by using partial couplings!

Techniques: Power From Couplings

Definition

A *partial coupling* $\mathcal{L}$ *between distributions* P, Q is a joint distribution (X, Y) such that $X \sim P$ and

$$\Pr_{\mathcal{L}}[X = Y = w] = \min(P(w)\,, Q(w))$$

for all w.

That is, it is not required that $Y \sim Q$.

Techniques: Power From Couplings

Solution (cont.)

It would suffice to define an efficiently computable estimator function f (bounded in $[0, 1]$) and efficiently samplable distribution π such that

$$\mathop{\mathbf{E}}_{w\sim\pi}[f(w)] = \frac{\mathbf{Pr}_{\mathcal{O}}[X \neq Y_1]}{\mathbf{Pr}_{\mathcal{L}}[X \neq Y_2]} = \frac{d_{\mathrm{TV}}(P, Q)}{Z},$$

for $X \sim P$, $Y_1 \sim Q$, and some sufficiently small and easy to compute $Z := \mathbf{Pr}_{\mathcal{L}}[X \neq Y_2]$.

Techniques: Power From Couplings

Solution (cont.)

Then we can estimate $\mathbf{E}_{w \sim \pi}[f(w)]$ by a Monte Carlo approach and therefore get an estimate of

$$Z \cdot \mathop{\mathbf{E}}_{w \sim \pi}[f(w)] = d_{\mathrm{TV}}(P, Q)\,.$$

Techniques: Power From Couplings

Where is the probabilistic inference algorithm used?

The probabilistic inference algorithm is used

(a) in the computation of Z and

(b) to sample from π.

Open Problems

Open Problems

We outline these questions:

1. For what other classes of probabilistic models do there exist TV distance approximation schemes?
2. What can we say about other notions of distance or similarity between probabilistic models?

Our Work on arXiv

Our Work on arXiv

Thank You!