A SIMPLE POLYNOMIAL-TIME APPROXIMATION ALGORITHM FOR THE TOTAL VARIATION DISTANCE BETWEEN TWO PRODUCT DISTRIBUTIONS

WEIMING FENG, HENG GUO, MARK JERRUM, JIAHENG WANG

Abstract. We give a simple polynomial-time approximation algorithm for the total variation distance between two product distributions.

1. Introduction

The total variation (TV) distance is a fundamental metric to measure the difference between two distributions. It is essentially the L^1 distance. Unlike many other quantities for similar uses, such as the relative entropy and the χ^2-divergence, the TV distance does not tensorise over product distributions. In fact, it was discovered recently that, somewhat surprisingly, exact computation of the total variation distance, even between product distributions over the Boolean domain, is #**P**-hard [1].

This leaves open the question of approximation complexity of the TV distance. In [1], the authors give polynomial-time randomised approximation algorithms in two special cases over the Boolean domain, when one of the distribution has marginals over $1/2$ and dominates the other, or when one of the distribution has a constant number of distinct marginals. Their method is based on Dyer's dynamic programming algorithm for approximating the number of knapsack solutions [2].

In this note, we give a simple polynomial-time approximation algorithm for total variation distance between two product distributions. Our algorithm is based on the Monte Carlo method and does not have further restrictions.

Theorem 1.1. *Let $[q] = \{1, 2, \ldots, q\}$ be a finite set. There exists an algorithm such that given two product distributions P, Q over $[q]^n$ and parameters $\varepsilon > 0$ and $0 < \delta < 1$, it outputs a random value $\widehat{d}$ in time $O(\frac{n^2}{\varepsilon^2} \log \frac{1}{\delta})$ such that $(1-\varepsilon)d_{\mathrm{TV}}(P,Q) \le \widehat{d} \le (1+\varepsilon)d_{\mathrm{TV}}(P,Q)$ holds with probability at least $1-\delta$.*

Our algorithm can also handle the case where each coordinate has a different domain size without any change. In Theorem 1.1, the input product distributions are given by the marginal probability for each coordinate and each $c \in [q]$ in binary. The stated running time assumes that all arithmetic operations can be done in $O(1)$ time.

To approximate the TV distance, the naïve Monte Carlo algorithm works well when the two distributions are sufficiently far away. However, when the TV distance is exponentially small, naïve Monte Carlo may require exponentially many samples to return an accurate estimate. Our idea is to consider a distribution that can be efficiently sampled from and yet boosts the probability that the two distributions are different. Ideally, we would want to use the optimal coupling, but that is difficult to compute. We use instead the coordinate-wise greedy coupling as a proxy, where each coordinate is coupled optimally independently. We further condition on the (potentially very unlikely) event that the two samples are different. Normally, conditioning on an unlikely event is a bad move since computational tasks would have become hard. However, here they are still easy thanks to the independence of the coordinates under the coupling. With this conditional distribution, our estimator is the ratio between the probabilities of the assignment in the optimal coupling and in the greedy coupling. We show that this estimator is always bounded from above by 1 and its expectation is at least $1/n$. This means that the standard Monte Carlo method will succeed with high probability using only polynomially many samples.

(Weiming Feng, Heng Guo, Jiaheng Wang) School of Informatics, University of Edinburgh, Informatics Forum, Edinburgh, EH8 9AB, United Kingdom. E-mail: wfeng@ed.ac.uk, hguo@inf.ed.ac.uk, jiaheng.wang@ed.ac.uk

(Mark Jerrum) School of Mathematical Sciences, Queen Mary, University of London, Mile End Road, London, E1 4NS, United Kingdom. E-mail: m.jerrum@qmul.ac.uk

One remaining question is if a deterministic approximation algorithm exists for the TV distance. The answer might be positive, because of the connection with counting knapsack solutions established by Bhattacharyya, Gayen, Meel, Myrisiotis, Pavan, and Vinodchandran [1], and the deterministic approximation algorithm for the latter problem by Štefankovič, Vempala, and Vigoda [4] and by Gopalan, Klivans, and Meka [3], independently.

2. Preliminaries

Let Ω be a (finite) state space, and P and Q be two distributions over Ω. The total variation distance is defined by

$$d_{\mathrm{TV}}(P,Q) := \frac{1}{2}\sum_{\omega\in\Omega}|P(\omega)-Q(\omega)|.$$

It satisfies the following:

- for any event $A \subseteq \Omega$, $d_{\mathrm{TV}}(P,Q) \geq |P(A)-Q(A)|$;
- for any coupling C between P and Q, $d_{\mathrm{TV}}(P,Q) \leq \mathbf{Pr}_C[X \neq Y]$, where $X \sim P$ and $Y \sim Q$.

In particular, there exists an event A_O and an optimal coupling O such that $d_{\mathrm{TV}}(P,Q) = |P(A_O)-Q(A_O)| = \mathbf{Pr}_O[X \neq Y]$. Optimal couplings are not necessarily unique. For any optimal coupling O, it holds that

$$\forall \omega \in \Omega, \quad \mathbf{Pr}_O[X = Y = \omega] = \min\{P(\omega), Q(\omega)\}. \tag{1}$$

The above equation holds because (1) for any valid coupling C, it holds that $\mathbf{Pr}_C[X = Y = \omega] \leq \min\{P(\omega), Q(\omega)\}$; (2) to achieve the optimal coupling, every ω must achieve the equality. We have

$$\mathbf{Pr}_O[X = \omega \wedge Y \neq X] = \mathbf{Pr}_O[X=\omega] - \mathbf{Pr}_O[X=Y=\omega] = \max\{0, P(\omega)-Q(\omega)\}. \tag{2}$$

3. Algorithm

From now on we consider only product distributions. Let $\Omega = [q]^n$ be the state space, where $[q] = \{1,\dots,q\}$ is a finite set. Let $P = P_1 \otimes P_2 \otimes \cdots \otimes P_n$ and $Q = Q_1 \otimes Q_2 \otimes \cdots \otimes Q_n$ be two product distributions. Let O be an (arbitrary) optimal coupling between P and Q.

Let C be the coordinate-wise greedy coupling. Namely, for each coordinate i and $c \in [q]$, $\mathbf{Pr}_C[X_i = Y_i = c] = \min\{P_i(c), Q_i(c)\}$, and the remaining probability can be assigned arbitrarily as long as C is a valid coupling (but each coordinate is independent). In other words, for each $i \in [n]$, C couples P_i and Q_i optimally and independently. Note that

$$\mathbf{Pr}_C[X \neq Y] = 1 - \mathbf{Pr}_C[X=Y] = 1 - \prod_{i=1}^{n}(1 - d_{\mathrm{TV}}(P_i,Q_i)) \tag{3}$$

can be computed exactly.

Consider the distribution π such that

$$\pi(\omega) := \mathbf{Pr}_C[X = \omega \mid X \neq Y]. \tag{4}$$

We may assume P and Q are not identical, as otherwise the algorithm just outputs 0. This makes sure that the distribution π is well-defined. The following lemma shows that we can draw random samples from π efficiently.

Lemma 3.1. *We can sample from the distribution π in $O(n)$ time.*

Proof. We draw a random sample $\omega \in [q]^n$ from π index by index. In the k-th step, where $1 \leq k \leq n$, we sample $\omega_k \in [q]$ from $\pi_k(\cdot \mid \omega_1, \omega_2, \dots, \omega_{k-1})$, which is the marginal distribution on the k-th variable conditional on the values of the first $k-1$ variables being $\omega_1, \omega_2, \dots, \omega_{k-1}$. By definition,

$$\pi_k(\omega_k \mid \omega_1, \omega_2, \dots, \omega_{k-1}) = \frac{\mathbf{Pr}_{X\sim\pi}[\forall 1 \leq i \leq k, X_i = \omega_i]}{\mathbf{Pr}_{X\sim\pi}[\forall 1 \leq i \leq k-1, X_i = \omega_i]}.$$

As $\omega_1, \ldots, \omega_{k-1}$ are sampled from the marginal distribution of π, the denominator is positive. We show how to compute the numerator next, and the denominator can be computed similarly. By definition

$$\begin{aligned}
&\mathbf{Pr}_{X\sim\pi}[\forall 1 \le i \le k, X_i = \omega_i] = \mathbf{Pr}_{(X,Y)\sim C}[\forall 1 \le i \le k, X_i = \omega_i \mid X \ne Y] \\
\text{(by Bayes' law)} \quad &= \left(1 - \mathbf{Pr}_{(X,Y)\sim C}[X = Y \mid \forall 1 \le i \le k, X_i = \omega_i]\right) \cdot \frac{\prod_{i=1}^{k} P_i(\omega_i)}{1 - \prod_{i=1}^{n}(1 - d_{\mathrm{TV}}(P_i, Q_i))}.
\end{aligned}$$

In the coupling C, every pair of X_i and Y_i is coupled optimally and independently. We have

$$\begin{aligned}
\mathbf{Pr}_{(X,Y)\sim C}[X = Y \mid \forall 1 \le i \le k, X_i = \omega_i] &= \prod_{i=1}^{k} \frac{\mathbf{Pr}_C[X_i = Y_i = \omega_i]}{\mathbf{Pr}_C[X_i = \omega_i]} \prod_{i=k+1}^{n} \mathbf{Pr}_C[X_i = Y_i] \\
\text{(by (1))} \quad &= \prod_{i=1}^{k} \frac{\min\{P_i(\omega_i), Q_i(\omega_i)\}}{P_i(\omega_i)} \prod_{i=k+1}^{n} (1 - d_{\mathrm{TV}}(P_i, Q_i)). \qquad (5)
\end{aligned}$$

Combining the two equations, we can compute $\mathbf{Pr}_{X\sim\pi}[\forall 1 \le i \le k, X_i = \omega_i]$, and thus we can compute and sample from $\pi_k(\cdot \mid \omega_1, \omega_2, \ldots, \omega_{k-1})$. When sampling from the distribution π, we preprocess $\prod_{i=k+1}^{n}(1 - d_{\mathrm{TV}}(P_i, Q_i))$ for all k, and maintain the prefix products $\prod_{i=1}^{k} \min\{P_i(\omega_i), Q_i(\omega_i)\}$ and $\prod_{i=1}^{k} P_i(\omega_i)$. This way, each conditional marginal distribution can be computed with $O_q(1)$ incremental cost. Hence, the total running time is $O_q(n)$, where $O_q(\cdot)$ hides a factor linear in q. □

Let ω be a random sample from π. Now consider the following estimator:

$$f(\omega) := \frac{\mathbf{Pr}_O[X = \omega \wedge X \ne Y]}{\mathbf{Pr}_C[X = \omega \wedge X \ne Y]} = \frac{\max\{0, P(\omega) - Q(\omega)\}}{\mathbf{Pr}_C[X = \omega \wedge X \ne Y]}, \qquad (6)$$

where the second equality is due to (2). This estimator f is well-defined, because when $\mathbf{Pr}_C[X = \omega \wedge X \ne Y] = 0$, $\pi(\omega) = 0$ as well and ω will not be drawn.

In fact, if $\pi(\omega) = 0$, or equivalently $\mathbf{Pr}_C[X = \omega \wedge X \ne Y] = 0$, it must be that $\max\{0, P(\omega) - Q(\omega)\} = 0$. This is because $\mathbf{Pr}_C[X = \omega \wedge X \ne Y] = 0$ implies that either $\mathbf{Pr}_C[X = \omega] = P(\omega) = 0$ or $\mathbf{Pr}_C[X \ne Y \mid X = \omega] = 0$. In the first case, $\max\{0, P(\omega) - Q(\omega)\} = 0$. In the second case $\mathbf{Pr}_C[Y = \omega \mid X = \omega] = 1$, which implies that $Q(\omega) \ge P(\omega)$, and $\max\{0, P(\omega) - Q(\omega)\} = 0$ as well.

Lemma 3.2. *For any $\omega \in \Omega$ with $\pi(\omega) > 0$, $f(\omega)$ can be computed in $O(n)$ time.*

Proof. Note that

$$\mathbf{Pr}_C[X = \omega \wedge X \ne Y] = P(\omega)\,\mathbf{Pr}_C[X \ne Y \mid X = \omega] = P(\omega)(1 - \mathbf{Pr}_C[X = Y \mid X = \omega]).$$

Since $\pi(\omega) > 0$, it holds that $P(\omega) > 0$. Using (5), we have

$$f(\omega) = \max\left\{0, \frac{1 - \frac{Q(\omega)}{P(\omega)}}{\frac{1}{P(\omega)}\mathbf{Pr}_C[X = \omega \wedge X \ne Y]}\right\} = \max\left\{0, \frac{1 - \prod_{i=1}^{n} \frac{Q_i(\omega_i)}{P_i(\omega_i)}}{1 - \prod_{i=1}^{n} \frac{\min\{P_i(\omega_i), Q_i(\omega_i)\}}{P_i(\omega)}}\right\},$$

which can be computed in $O(n)$ time. □

Lemma 3.3. *We have the following:*

$$\mathbf{E}_\pi f = \frac{\mathbf{Pr}_O[X \ne Y]}{\mathbf{Pr}_C[X \ne Y]}; \qquad (7)$$

$$\frac{1}{n} \le \mathbf{E}_\pi f \le 1. \qquad (8)$$

Moreover, for any $\omega \in \Omega$ with $\pi(\omega) > 0$,

$$0 \le f(\omega) \le 1, \qquad (9)$$

and it holds that

$$\mathbf{Var}_\pi f \le \mathbf{E}_\pi f. \qquad (10)$$

Proof. For (7), Let $\Omega_+ = \{\omega \in \Omega \mid \pi(\omega) > 0\}$. Then,

$$\begin{aligned}\mathbf{E}_\pi f &= \sum_{\omega\in\Omega_+} \pi(\omega) \times \frac{\mathbf{Pr}_O[X = \omega \wedge X \neq Y]}{\mathbf{Pr}_C[X = \omega \wedge X \neq Y]} \\ &= \sum_{\omega\in\Omega_+} \frac{\mathbf{Pr}_C[X = \omega \wedge X \neq Y]}{\mathbf{Pr}_C[X \neq Y]} \times \frac{\mathbf{Pr}_O[X = \omega \wedge X \neq Y]}{\mathbf{Pr}_C[X = \omega \wedge X \neq Y]} \\ &= \frac{\sum_{\omega\in\Omega_+} \mathbf{Pr}_O[X = \omega \wedge X \neq Y]}{\mathbf{Pr}_C[X \neq Y]} = \frac{\mathbf{Pr}_O[X \neq Y]}{\mathbf{Pr}_C[X \neq Y]},\end{aligned}$$

where in the last equation we used the aforementioned fact that $\pi(\omega) = 0$ implies $\max\{0, P(\omega) - Q(\omega)\} = 0$.

For (8), as O is the optimal coupling, $\mathbf{Pr}_O[X \neq Y] \leq \mathbf{Pr}_C[X \neq Y]$. For the other direction, notice that O projected to coordinate i, denoted O_i, is a coupling between P_i and Q_i. Thus,

$$\mathbf{Pr}_O[X \neq Y] \geq \max_{1\leq i\leq n} \mathbf{Pr}_{O_i}[X_i \neq Y_i] \geq \max_{1\leq i\leq n} d_{\mathrm{TV}}(P_i, Q_i),$$

On the other hand, by the union bound,

$$\mathbf{Pr}_C[X \neq Y] \leq \sum_{i=1}^n \mathbf{Pr}_{C_i}[X_i \neq Y_i] = \sum_{i=1}^n d_{\mathrm{TV}}(P_i, Q_i) \leq n \max_{1\leq i\leq n} d_{\mathrm{TV}}(P_i, Q_i).$$

For (9), the lower bound is trivial. For the upper bound, we only need to consider $\omega \in \Omega_+$ such that $P(\omega) > Q(\omega)$. In this case

$$\begin{aligned}f(\omega) &= \frac{\max\{0, P(\omega) - Q(\omega)\}}{\mathbf{Pr}_C[X = \omega \wedge X \neq Y]} = \frac{P(\omega) - Q(\omega)}{\mathbf{Pr}_C[X = \omega]\,\mathbf{Pr}_C[X \neq Y \mid X = \omega]} \\ &= \frac{P(\omega) - Q(\omega)}{P(\omega)(1 - \mathbf{Pr}_C[X = Y \mid X = \omega])} = \frac{1 - \frac{Q(\omega)}{P(\omega)}}{1 - \mathbf{Pr}_C[X = Y \mid X = \omega]}.\end{aligned}$$

Since C couples each coordinate independently,

$$\mathbf{Pr}_C[X = Y \mid X = \omega] = \prod_{i=1}^n \frac{\min\{P_i(\omega_i), Q_i(\omega_i)\}}{P_i(\omega_i)} \leq \prod_{i=1}^n \frac{Q_i(\omega_i)}{P_i(\omega_i)} = \frac{Q(\omega)}{P(\omega)}.$$

This finishes the proof of (9).

For (10), since $0 \leq f(\omega) \leq 1$ for all $\Omega \in \Omega_+$, $f(\omega)^2 \leq f(\omega)$ and thus $\mathbf{E}_\pi f^2 \leq \mathbf{E}_\pi f$. We have

$$\mathbf{Var}_\pi f = \mathbf{E}_\pi f^2 - (\mathbf{E}_\pi f)^2 \leq \mathbf{E}_\pi f^2 \leq \mathbf{E}_\pi f.$$ □

Lemma 3.3 implies that standard Monte Carlo method can be used to accurately estimate $\mathbf{E}_\pi f = \frac{\mathbf{Pr}_O[X\neq Y]}{\mathbf{Pr}_C[X\neq Y]}$. To implement the Monte Carlo algorithm, we use Lemma 3.1 and Lemma 3.2.

To be more specific, our approximate algorithm is to compute the median of means. The input contains the descriptions of $2n$ distributions $P_1, P_2, \ldots, P_n, Q_1, Q_2, \ldots, Q_n$ together with two parameters $\varepsilon > 0$ and $0 < \delta < 1$. The algorithm proceeds as follows:

- for each i from 1 to $m = \lceil \frac{10n}{\varepsilon^2} \rceil$, independently sample $\omega_i \sim \pi$ and let
$$F = \frac{1}{m} \sum_{i=1}^m f(\omega_i);$$
- use independent samples to compute F for $s = 10\lceil \log \frac{1}{\delta} \rceil$ times to get $F_1, F_2, \ldots, F_s$ and let
$$\widehat{F} = \mathrm{Median}\{F_1, F_2, \ldots, F_s\};$$
- output the value $\widehat{d} = (1 - \prod_{i=1}^n (1 - d_{\mathrm{TV}}(P_i, Q_i)))\widehat{F}$.

We claim that

$$\mathbf{Pr}\left[|F - \mathbf{E}_\pi f| \geq \varepsilon\, \mathbf{E}_\pi f\right] \leq \frac{1}{10}. \tag{11}$$

Assuming that (11) holds, by the Chernoff bound, it holds that

$$\mathbf{Pr}\left[\left|\widehat{F} - \mathbf{E}_\pi f\right| \geq \varepsilon \mathbf{E}_\pi f\right] \leq \delta.$$

Using (7) in Lemma 3.3 and (3), we have

$$\mathbf{Pr}\left[\left|\widehat{d} - d_{\mathrm{TV}}(P,Q)\right| \geq \varepsilon d_{\mathrm{TV}}(P,Q)\right] = \mathbf{Pr}\left[\left|\widehat{F} - \mathbf{E}_\pi f\right| \geq \varepsilon \mathbf{E}_\pi f\right] \leq \delta.$$

By Lemma 3.1 and Lemma 3.2, the total running time is $O(nms) = O(\frac{n^2}{\varepsilon^2}\log\frac{1}{\delta})$. This proves Theorem 1.1.

Finally, we prove the claim (11). Note that the expectation and the variance of the random variable F satisfy that $\mathbf{E}F = \mathbf{E}_\pi f$ and $\mathbf{Var}F = \frac{1}{m}\mathbf{Var}_\pi f$. By Chebyshev's inequality,

$$\mathbf{Pr}\left[|F - \mathbf{E}_\pi f| \geq \varepsilon \mathbf{E}_\pi f\right] = \mathbf{Pr}\left[|F - \mathbf{E}F| \geq \varepsilon \mathbf{E}F\right] \leq \frac{\mathbf{Var}F}{\varepsilon^2(\mathbf{E}F)^2} = \frac{\mathbf{Var}_\pi f}{m\varepsilon^2(\mathbf{E}_\pi f)^2}$$

(by (10), (8), and $m = \lceil\frac{10n}{\varepsilon^2}\rceil$) $\leq \frac{1}{m\varepsilon^2\,\mathbf{E}_\pi f} \leq \frac{n}{m\varepsilon^2} \leq \frac{1}{10}.$

Acknowledgement

MJ was supported by grant EP/S016694/1 'Sampling in hereditary classes' from the Engineering and Physical Sciences Research Council (EPSRC) of the UK. FWM, GH, and WJH have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 947778). WJH has also received financial support from an Informatics Global PhD Scholarship at The University of Edinburgh.

References

[1] Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, Aduri Pavan, and N. V. Vinodchandran. On approximating total variation distance. *arXiv*, abs/2206.07209, 2022.

[2] Martin E. Dyer. Approximate counting by dynamic programming. In *STOC*, pages 693–699. ACM, 2003.

[3] Parikshit Gopalan, Adam R. Klivans, and Raghu Meka. Polynomial-time approximation schemes for knapsack and related counting problems using branching programs. *arXiv*, abs/1008.3187, 2010.

[4] Daniel Štefankovič, Santosh S. Vempala, and Eric Vigoda. A deterministic polynomial-time approximation scheme for counting knapsack solutions. *SIAM J. Comput.*, 41(2):356–366, 2012.