On Deterministically Approximating Total Variation Distance

Weiming Feng[*]
UC Berkeley
fwm1994@gmail.com

Liqiang Liu
Peking University
lql@pku.edu.cn

Tianren Liu
Peking University
trl@pku.edu.cn

Abstract

Total variation distance (TV distance) is an important measure for the difference between two distributions. Recently, there has been progress in approximating the TV distance between product distributions: a deterministic algorithm for a restricted class of product distributions (Bhattacharyya, Gayen, Meel, Myrisiotis, Pavan and Vinodchandran 2023) and a randomized algorithm for general product distributions (Feng, Guo, Jerrum and Wang 2023). We give a *deterministic* fully polynomial-time approximation algorithm (FPTAS) for the TV distance between product distributions. Given two product distributions $\mathbb{P}$ and $\mathbb{Q}$ over $[q]^n$, our algorithm approximates their TV distance with relative error ε in time $O\big(\frac{qn^2}{\varepsilon} \log q \log \frac{n}{\varepsilon \, \Delta_{\mathrm{TV}}(\mathbb{P},\mathbb{Q})}\big)$.

Our algorithm is built around two key concepts: 1) The *likelihood ratio* as a distribution, which captures sufficient information to compute the TV distance. 2) We introduce a metric between likelihood ratio distributions, called the *minimum total variation distance*. Our algorithm computes a sparsified likelihood ratio distribution that is close to the original one w.r.t. the new metric. The approximated TV distance can be computed from the sparsified likelihood ratio.

Our technique also implies deterministic FPTAS for the TV distance between Markov chains.

1 Introduction

The *total variation distance* (TV distance), which is also known as the *statistical difference*, is a fundamental metric for measuring the difference between two probability distributions. For two distributions $\mathbb{P}$ and $\mathbb{Q}$ over the sample space Ω, their TV distance is defined by

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

The TV distance is essentially the L_1-norm of the difference, and can also be characterized as the discrepancy of the optimal coupling between $\mathbb{P}$ and $\mathbb{Q}$ [MU17]. The TV distance connects to many statistical distances including the relative entropy and other well-studied divergences [PW22].

This paper studies the computational problem of TV distance. The problem is particularly interesting when both $\mathbb{P}$ and $\mathbb{Q}$ are high-dimensional distributions that admit succinct descriptions. However, computing (even approximating with additive error) the TV distance is known to be hard for general distributions. The seminal work of Sahai and Vadhan [SV03] proved that deciding whether two distributions given by Boolean circuits are "close" or "far-apart" w.r.t. TV distance is a complete problem in **SZK** (Statistical Zero Knowledge). **SZK**-complete problems are often

---

believed to be computationally hard [BCH$^+$20]. Many works studied the complexity of computing the TV distance for specific classes of distributions [LP02, CMR07, Kie18, BGM$^+$23]. One of the simplest situations is when $\mathbb{P}$ and $\mathbb{Q}$ are both *product distributions* over the sample space $[q]^n$, where $[q] := \{1, 2, \ldots, q\}$ is a finite domain and n is the dimension. Recently, Bhattacharyya, Gayen, Meel, Myrisiotis, Pavan and Vinodchandran [BGM$^+$23] proved a surprising result: the exact computing of the TV distance between two product distributions is #**P**-complete, even in the Boolean domain ($q = 2$).

The hardness result motivates the study of approximation algorithms for the total variation distance between two product distributions. The algorithm is required to estimate the TV distance within ε *relative error*. Previous works mainly focused on randomized approximation algorithms, which find a good estimation with probability at least $1 - \delta$. The first polynomial-time randomized approximation algorithm was given in the preprint version of [BGM$^+$23], but it only works for a restricted class of product distributions. Later on, Feng, Guo, Jerrum and Wang [FGJW23] gave a randomized algorithm for general product distributions. The algorithm runs in time $O(\frac{qn^2}{\varepsilon^2} \log \frac{1}{\delta})$.

All algorithms mentioned above are based on the Monte Carlo method. One natural question is how to design deterministic approximation algorithms. Very recently, in the conference version[1] of [BGM$^+$23], authors strengthened their randomized algorithm into a deterministic one. However, the algorithm still requires some restrictions on the input distributions, i.e., the domain needs to be Boolean and $\mathbb{Q}$ has a constant number of distinct marginals (e.g., $\mathbb{Q}$ can be the uniform distribution). The major open question is to give a deterministic approximation algorithm that works for *general* product distributions. We answered this question affirmatively.

> **Theorem 1** (TV distance between product distributions) *There exists a deterministic algorithm such that given two product distributions $\mathbb{P}, \mathbb{Q}$ over $[q]^n$ and $\varepsilon > 0$, it outputs $\widehat{\Delta}$ satisfying $(1 - \varepsilon) \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \leq \widehat{\Delta} \leq \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$ in time $O\big(\frac{qn^2}{\varepsilon} \log q \log \frac{n}{\varepsilon \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})}\big)$.*

In the above theorem, both $\mathbb{P}$ and $\mathbb{Q}$ are described by their marginal distributions and each marginal distribution is given in binary. We remark that the term $\log \frac{1}{\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})}$ in running time is always a polynomial with respect to the input size. Moreover, our algorithm is faster for product distributions with larger total variation distances. The approximated TV distance $\widehat{\Delta}$ returned by our algorithm is always a lower bound of $\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$.

Our technique can be extended to more general high-dimensional distributions with "limited" dependency. One example is the distribution of the Markov chain trajectory. Let $\mathbb{P}$ be the joint distribution of a random vector $(X_1, X_2, \ldots, X_n) \in [q]^n$, where X_1 follows the initial distribution $\mathbb{P}_1$ and each X_i is generated based on a Markov kernel $\mathbb{P}_{i|i-1}$ that specifies the distribution of X_i conditional on X_{i-1}. We call $\mathbb{P}$ the distribution of an *n-step Markov chain*[2]. Given two n-step Markov chains $\mathbb{P}$ and $\mathbb{Q}$, we can also approximate their TV distance in polynomial time.

> **Theorem 2** (TV distance between n-step Markov chains) *There exists a deterministic algorithm such that given two n-step Markov chains $\mathbb{P}, \mathbb{Q}$ in state space $[q]$ and $\varepsilon > 0$, it outputs $\widehat{\Delta}$ satisfying $(1 - \varepsilon) \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \leq \widehat{\Delta} \leq \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$ in time $O\big(\frac{q^2 n^2}{\varepsilon} \log q \log \frac{n}{\varepsilon \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})}\big)$.*

We remark that our algorithm estimates the TV distance between the joint distributions of $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_n)$ rather than the TV distance between the distributions of X_n and Y_n. The latter is trivial because the marginal distribution of X_i (also Y_i) can be computed by simple matrix multiplications. But estimating $\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$ is non-trivial because the sample space is

---

[1]Link to the paper: https://www.cs.toronto.edu/~meel/Papers/ijcai23-bggmpv.pdf
[2]The Markov chain has the state space $[q]$ as each $X_i \in [q]$, but $\mathbb{P}$ is a distribution over the sample space $[q]^n$.

exponentially large. Our result is related to the problem of comparing two labelled Markov chains. See Section 1.2 for a detailed discussion.

1.1 Technical Overview

Let $\mathbb{P}, \mathbb{Q}$ be two distributions over a sample space Ω, their total variation distance is well known to equal the maximal *advantage* of distinguishing between $\mathbb{P}$ and $\mathbb{Q}$. Say a distinguisher is given a sample x from either $\mathbb{P}$ or $\mathbb{Q}$ and is asked to guess which distribution x is sampled from. The distinguisher D can be formalized as a randomized algorithm (or a Markov kernel) from Ω to $\{0, 1\}$. The probability that the distinguisher D outputs 0 under hypothesis $\mathbb{P}$ (resp. $\mathbb{Q}$) can be written as $\Pr_{X \sim \mathbb{P}}[D(X) \to 0]$ (resp. $\Pr_{X \sim \mathbb{Q}}[D(X) \to 0]$). The difference between them is called the advantage. The maximal advantage equals the total variation distance

$$\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \max_D \left(\Pr_{X \sim \mathbb{P}}[D(X) \to 0] - \Pr_{X \sim \mathbb{Q}}[D(X) \to 0] \right),$$

and the optimal distinguisher who maximizes the advantage, is the *likelihood ratio test*

$$D(x) = \begin{cases} 0, & \text{if } \frac{\mathbb{P}(x)}{\mathbb{Q}(x)} > 1 \\ 1, & \text{if } \frac{\mathbb{P}(x)}{\mathbb{Q}(x)} \leq 1 \end{cases}$$

where $\frac{\mathbb{P}(x)}{\mathbb{Q}(x)}$ is called the likelihood ratio. The likelihood ratio $\frac{\mathbb{P}(x)}{\mathbb{Q}(x)}$ is an important tool in information theory [PW22]. It is particularly useful when x is sampled from $\mathbb{P}$ or from $\mathbb{Q}$.

In this paper, we formally define the likelihood ratio as a distribution, denoted by $(\mathbb{P}\|\mathbb{Q})$, such that $(\mathbb{P}\|\mathbb{Q})(r) := \Pr_{X \sim \mathbb{Q}}[\frac{\mathbb{P}(X)}{\mathbb{Q}(X)} = r]$. In other words, $(\mathbb{P}\|\mathbb{Q})$ is the distribution of $\frac{\mathbb{P}(X)}{\mathbb{Q}(X)}$ where $X \sim \mathbb{Q}$.

As we will discussed in Section 3, the likelihood ratio distribution (or "ratio" in short) $\mathbb{R} = (\mathbb{P}\|\mathbb{Q})$ contains all the "useful" information about $(\mathbb{P}, \mathbb{Q})$. For example, $\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{R \sim \mathbb{R}} \max(1 - R, 0)$, so we can denote the distance by $\Delta_{\mathrm{TV}}(\mathbb{R})$. For the task of computing the total variation distance, it suffices to compute $\mathbb{R} = (\mathbb{P}\|\mathbb{Q})$.

Given two product distributions $\mathbb{P} = \mathbb{P}_1 \mathbb{P}_2 \ldots \mathbb{P}_n$ and $\mathbb{Q} = \mathbb{Q}_1 \mathbb{Q}_2 \ldots \mathbb{Q}_n$, their ratio $(\mathbb{P}\|\mathbb{Q})$

$$(\mathbb{P}\|\mathbb{Q})(r) = \Pr_{(X_1, \ldots, X_n) \sim \mathbb{Q}} \left[\frac{\mathbb{P}(X_1, \ldots, X_n)}{\mathbb{Q}(X_1, \ldots, X_n)} = r \right] = \Pr_{\substack{X_1 \sim \mathbb{Q}_1 \\ \cdots \\ X_n \sim \mathbb{Q}_n}} \left[\underbrace{\frac{\mathbb{P}_1(X_1)}{\mathbb{Q}_1(X_1)}}_{R_1} \cdots \underbrace{\frac{\mathbb{P}_n(X_n)}{\mathbb{Q}_n(X_n)}}_{R_n} = r \right]$$

is the distribution of $R_1 R_2 \ldots R_n$, where $R_1, \ldots, R_n$ are independent and $R_i \sim \mathbb{R}_i = (\mathbb{P}_i \| \mathbb{Q}_i)$. This suggests a naïve algorithm to compute the total variation distance.

- Compute $\mathbb{R}_i = (\mathbb{P}_i \| \mathbb{Q}_i)$ for each $i \in [n]$.
- Compute $\mathbb{R} = \mathbb{R}_1 \cdot_{\mathsf{indp}} \mathbb{R}_2 \cdot_{\mathsf{indp}} \cdots \mathbb{R}_n$.
- Output $\Delta_{\mathrm{TV}}(\mathbb{R})$.

Here $\mathbb{R}_1 \cdot_{\mathsf{indp}} \mathbb{R}_2$ denotes the distribution of $R_1 R_2$ where R_1, R_2 are independently sampled from $\mathbb{R}_1, \mathbb{R}_2$. The naïve algorithm computes the *exact* value of total variation distance, but has exponential time and space complexity, because the support of $\mathbb{R}$ can be exponentially large.

To contain the complexity blow-up, our actual algorithm computes an approximation of $\mathbb{R}$. The high-level framework of our algorithm looks as follows.

- Compute $\mathbb{R}_i = (\mathbb{P}_i \| \mathbb{Q}_i)$ for each $i \in [n]$.
- Compute $\mathbb{R}_1 \cdot_{\mathsf{indp}} \mathbb{R}_2$. Then sparsify it as $\tilde{\mathbb{R}}_{1:2} \approx \mathbb{R}_1 \cdot_{\mathsf{indp}} \mathbb{R}_2$.

- Compute $\tilde{\mathbb{R}}_{1:2} \cdot_{\mathsf{indp}} \mathbb{R}_3$. Then sparsify it as $\tilde{\mathbb{R}}_{1:3} \approx \tilde{\mathbb{R}}_{1:2} \cdot_{\mathsf{indp}} \mathbb{R}_3$.

  $\vdots$
- Compute $\tilde{\mathbb{R}}_{1:n-1} \cdot_{\mathsf{indp}} \mathbb{R}_n$. Then sparsify it as $\tilde{\mathbb{R}}_{1:n} \approx \tilde{\mathbb{R}}_{1:n-1} \cdot_{\mathsf{indp}} \mathbb{R}_n$.
- Output $\Delta_{\mathrm{TV}}(\mathbb{R}_{1:n})$ as an approximation of $\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$.

In our algorithm, the *sparsification* subroutine plays a central role: a given complicated ratio is sparsified into a simpler one that is "close to" the given ratio. The closeness is formalized as a new metric. For the correctness of our algorithm, the metric should satisfy a few properties: i) triangle inequality, if $\mathbb{R} \approx \mathbb{R}'$ and $\mathbb{R}' \approx \mathbb{R}''$, then $\mathbb{R} \approx \mathbb{R}''$; ii) if $\mathbb{R} \approx \tilde{\mathbb{R}}$, then $\Delta_{\mathrm{TV}}(\mathbb{R}) \approx \Delta_{\mathrm{TV}}(\tilde{\mathbb{R}})$; iii) if $\mathbb{R} \approx \tilde{\mathbb{R}}$, then $\mathbb{R} \cdot_{\mathsf{indp}} \mathbb{R}' \approx \tilde{\mathbb{R}} \cdot_{\mathsf{indp}} \mathbb{R}'$. Then correctness is guaranteed as follows:

$$\tilde{\mathbb{R}}_{1:i+1} \approx \tilde{\mathbb{R}}_{1:i} \cdot_{\mathsf{indp}} \mathbb{R}_{i+1} \implies \tilde{\mathbb{R}}_{1:i+1} \cdot_{\mathsf{indp}} \mathbb{R}_{i+2} \cdot_{\mathsf{indp}} \cdots \mathbb{R}_n \approx \tilde{\mathbb{R}}_{1:i} \cdot_{\mathsf{indp}} \mathbb{R}_{i+1} \cdot_{\mathsf{indp}} \mathbb{R}_{i+2} \cdot_{\mathsf{indp}} \cdots \mathbb{R}_n ;$$

then by triangle inequality, $\tilde{\mathbb{R}}_{1:n} \approx \mathbb{R}_1 \cdot_{\mathsf{indp}} \mathbb{R}_2 \cdot_{\mathsf{indp}} \cdots \mathbb{R}_n = \mathbb{R}$; therefore $\Delta_{\mathrm{TV}}(\mathbb{R}) \approx \Delta_{\mathrm{TV}}(\tilde{\mathbb{R}})$.

We say two ratios $\mathbb{R}, \mathbb{R}'$ are close, if there exist distributions $\mathbb{P}, \mathbb{Q}, \mathbb{P}', \mathbb{Q}'$ satisfying $(\mathbb{P}\|\mathbb{Q}) = \mathbb{R}$, $(\mathbb{P}'\|\mathbb{Q}') = \mathbb{R}'$ such that $\mathbb{P}$ (resp. $\mathbb{Q}$) is close to $\mathbb{P}'$ (resp $\mathbb{Q}'$) with respect to the TV distance. In such case, $\Delta_{\mathrm{TV}}(\mathbb{R})$ $(= \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}))$ must be close to $\Delta_{\mathrm{TV}}(\mathbb{R}')$ $(= \Delta_{\mathrm{TV}}(\mathbb{P}', \mathbb{Q}'))$ due to the triangle inequality of the TV distance. This inspires us to consider the *minimum total variation distance*,

$$\Delta_{\mathrm{MTV}}(\mathbb{R}, \mathbb{R}') = \min_{\substack{(\mathbb{P}\|\mathbb{Q})=\mathbb{R} \\ (\mathbb{P}'\|\mathbb{Q}')=\mathbb{R}'}} \max\left(\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{P}'), \Delta_{\mathrm{TV}}(\mathbb{Q}, \mathbb{Q}')\right).$$

In Section 3.1, we show that Δ_{MTV} is a metric between ratios.

In the rest of this overview, we briefly describe how the sparsification subroutine works, and why the sparsified ratio is close to the original ratio with respect to the minimum total variation distance.

The sparsification subroutine takes a ratio $\mathbb{R}$ as the input. The ratio $\mathbb{R}$ can be represented by a table of entries $(r_1, p_1), (r_2, p_2), \ldots$, each entry (r_i, p_i) represents $\mathbb{R}(r_i) = p_i$. If the table has too many entries, then some of them must be close together. Say $(r_1, p_1), \ldots, (r_k, p_k)$ are very close together in the sense that $r_1, \ldots, r_k \in [a, b)$, where $[a, b)$ is a sufficiently narrow interval. W.l.o.g., assume $[a, b) \subseteq [0, 1]$, and the case inside $[1, \infty]$ is symmetric. The sparsification subroutine "merges" these k entries into one entry. The sparsified ratio $\tilde{\mathbb{R}}$ is represented by table $(\frac{\sum_{i \in [k]} r_i p_i}{\sum_{i \in [k]} p_i}, \sum_{i \in [k]} p_i), (r_{k+1}, p_{k+1}), (r_{k+2}, p_{k+2}), \ldots$ We claim that $\tilde{\mathbb{R}}$ is close to $\mathbb{R}$ with respect to the minimum total variation distance Δ_{MTV}.

Define two distributions $\mathbb{P}, \mathbb{Q}$ over sample space $\Omega = \{\omega_0, \omega_1, \omega_2, \ldots\}$ as

$$\mathbb{P}(\omega_i) = r_i p_i, \quad \mathbb{Q}(\omega_i) = p_i, \quad \mathbb{P}(\omega_0) = 1 - \mathbb{E}_{R \sim \mathbb{R}}[R], \quad \mathbb{Q}(\omega_0) = 0,$$

so that $(\mathbb{P}\|\mathbb{Q}) = \mathbb{R}$. Consider distribution $\tilde{\mathbb{P}}$ that is the same as $\mathbb{P}$ except

$$\tilde{\mathbb{P}}(\omega_i) = r^* p_i \text{ for } i \in [k], \text{ where } r^* := \frac{\sum_{i \in [k]} r_i p_i}{\sum_{i \in [k]} p_i}.$$

Then $\tilde{\mathbb{R}} = (\tilde{\mathbb{P}}\|\mathbb{Q})$, this also ensures that $\tilde{\mathbb{R}}$ is a valid ratio. By definition, $\Delta_{\mathrm{MTV}}(\mathbb{R}, \tilde{\mathbb{R}}) \leq \Delta_{\mathrm{TV}}(\mathbb{P}, \tilde{\mathbb{P}}) \leq \sum_{i \in [k]} |r_i - r^*| p_i$. Since $r_1, \ldots, r_k, r^* \in [a, b]$, we have $|r_i - r^*|$ being small for all $i \in [k]$. Quantitatively, say interval $[a, b]$ is sufficiently narrow means $\frac{1-a}{1-b} \leq 1 + \varepsilon$. Thus we have $|r_i - r^*| \leq \varepsilon(1 - r_i)$ for all $i \in [k]$, and the error introduced by merging is bounded by $\Delta_{\mathrm{MTV}}(\mathbb{R}, \tilde{\mathbb{R}}) \leq \sum_{i \in [k]} \varepsilon(1 - r_i) p_i$.

4

The actual sparsification subroutine picks a collection of disjointed sufficiently narrow intervals[3] $[a_1, b_1), [a_2, b_2), \ldots, [a_m, b_m)$ that jointly cover $[0, 1)$, then merges all the entries that lie in the same interval. The total error introduced by merging is bounded by

$$\sum_{j=1}^{m-1} \sum_{i \text{ s.t. } r_i \in [a_j, b_j]} \varepsilon(1 - r_i) p_i \leq \sum_{i \text{ s.t. } r_i < 1} \varepsilon(1 - r_i) p_i = \varepsilon \mathop{\mathbb{E}}_{R \sim \mathbb{R}}[\max(1 - R, 0)] = \varepsilon \, \Delta_{\text{TV}}(\mathbb{R}) \,.$$

Symmetrically, the error introduced by merging within $(1, \infty]$ is also bounded by $\varepsilon \, \Delta_{\text{TV}}(\mathbb{R})$.

1.2 Related Works and Open Problems

The TV distance between two labelled Markov chains (LMC, a.k.a. hidden Markov chain) has been studied by previous works [LP02, CMR07, Kie18]. It was proved that for LMCs, computing TV distance within ε-additive error, where $\varepsilon > 0$ is given in binary, is #**P**-hard [Kie18]. The Markov chains we studied in Theorem 2 is *not hidden*, which is equivalent (by a polynomial-time reduction) to the *deterministic acyclic LMCs* in [Kie18]. Kiefer [Kie18, Theorem 10] gave a randomized algorithm that approximates the TV distance between (not necessarily deterministic) acyclic LCMs with *additive-error* ϵ, where the algorithm succeeds with probability at least $1 - \delta$ and the running time is $\text{poly}(\frac{1}{\epsilon}, \log \frac{1}{\delta}, \text{input size})$. Compared with our result, Kiefer's algorithm works for more general distributions but for deterministic acyclic LMCs, our algorithm is deterministic and achieves stronger relative-error approximation.

Some works [CDKS18, BGMV20] studied the problem of computing TV distances for structured high-dimensional distributions, e.g. Bayesian networks, Ising models and multivariate Gaussian distributions. Some randomized approximation algorithms were discovered, but they only achieved an additive-error approximation. One open problem is to find efficient deterministic approximation (with additive or relative error) algorithms for those distributions. Our technique relies on a strong conditional independence property of the distribution. We wonder whether one can relax this restriction and make our technique work for more general graphical models.

Our technique for approximating the TV distance is different from the previous ones. Previous randomized algorithms [Kie18, CDKS18, BGMV20, FGJW23] are all based on the Monte Carlo method. For the deterministic approximation of the TV distance between two product distributions, the algorithm in [BGM$^+$23] first reduces the problem to the approximate counting of knapsack solutions (#Knapsack) and then solves the #Knapsack by existing deterministic approximation algorithms [GKM10, SVV12]. We introduce a new sparsification technique together with a new metric called minimum total variation distance, which is of independent interest.

2 Notations

We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$ for any positive integer n. We use $\mathbb{1}[A] \in \{0, 1\}$ to indicate whether the condition (or event) A holds.

We use ∞ to denote the infinity point, which is an extended real number. For any $a \in (0, \infty)$, let $a/0 = \infty$, $a/\infty = 0$ and $a < \infty$.

This paper only considers *discrete* distributions. Distributions are denoted by calligraphic letters (e.g., $\mathbb{P}, \mathbb{Q}$). A discrete distribution $\mathbb{P}$ over a sample space Ω can be defined by its probability mass function $\mathbb{P} : \Omega \rightarrow [0, 1]$. For each $x \in \Omega$, the value $\mathbb{P}(x)$ is the probability that x is sampled from $\mathbb{P}$. We use $\text{Supp}(\mathbb{P})$ to denote the *support* of $\mathbb{P}$, which refers to $\{x \in \Omega \mid \mathbb{P}(x) > 0\}$. Let

---

[3]The last interval is *not* sufficiently narrow because its right endpoint is 1, and has to be analyzed separately.

$X \sim \mathbb{P}$ denote that *random variable X follows distribution $\mathbb{P}$*. For each subset $S \subseteq \Omega$, we use the conventional notation $\mathbb{P}(S)$ to denote the probability that a sampling from $\mathbb{P}$ falls in set S. That is, $\mathbb{P}(S) := \Pr_{X \leftarrow \mathbb{P}}[X \in S]$.

For any two distributions $\mathbb{P}, \mathbb{Q}$, let $\mathbb{P}\mathbb{Q}$ (or $\mathbb{P} \times \mathbb{Q}$) denotes the distribution of (X, Y), let $\mathbb{P} \cdot_{\mathsf{indp}} \mathbb{Q}$ denotes the distribution of $X \cdot Y$, where X, Y are independent random variables satisfying $\mathbb{P}, \mathbb{Q}$ respectively.

A Markov kernel κ from sample space Ω to sample space Ω' is a function $\kappa : \Omega' \times \Omega \to [0, 1]$, such that for any $x \in \Omega$, function $y \mapsto \kappa(y|x)$ is a distribution over Ω'. For any distribution $\mathbb{P}$ over Ω, let $\kappa\mathbb{P}$ denote the distribution over Ω' that $(\kappa\mathbb{P})(y) = \sum_{x \in \Omega} \mathbb{P}(x)\kappa(y|x)$. Their semidirect product, denoted by $\mathbb{P}\kappa$ or $\mathbb{P} \times \kappa$, is a distribution over $\Omega \times \Omega'$ that $(\mathbb{P}\kappa)(x, y) = \mathbb{P}(x)\kappa(y|x)$.

3 Likelihood Ratio as a Distribution

Given two distributions $\mathbb{P}, \mathbb{Q}$ over a sample space Ω, people call $\frac{\mathbb{P}(x)}{\mathbb{Q}(x)}$ the *likelihood ratio* for $x \in \Omega$. We introduce a notation for the distribution of the likelihood ratio.

Definition 1 (Ratio)**.** Let $\mathbb{P}, \mathbb{Q}$ be two discrete distributions over a sample space Ω. The likelihood ratio distribution (or *ratio*, in short) between $\mathbb{P}, \mathbb{Q}$, denoted by $(\mathbb{P}\|\mathbb{Q})$, is a distribution over $[0, \infty)$.

$$(\mathbb{P}\|\mathbb{Q})(r) := \Pr_{X \sim \mathbb{Q}}\Big[\frac{\mathbb{P}(X)}{\mathbb{Q}(X)} = r\Big].$$

Equivalently, the distribution $(\mathbb{P}\|\mathbb{Q})$ can be defined as follows: sample random variable $X \sim \mathbb{Q}$, define $(\mathbb{P}\|\mathbb{Q})$ as the distribution of $\frac{\mathbb{P}(X)}{\mathbb{Q}(X)}$. Note that the denominator is always non-zero because X is sampled from $\mathbb{Q}$. In the paper, we typically denote the ratio distribution by $\mathbb{R} = (\mathbb{P}\|\mathbb{Q})$.

Not every discrete distribution $\mathbb{R}$ over $[0, \infty)$ is the ratio between some two discrete distributions. We say $\mathbb{R}$ is a *valid ratio* if there exist two discrete distributions $\mathbb{P}, \mathbb{Q}$ such that $\mathbb{R} = (\mathbb{P}\|\mathbb{Q})$.

Proposition 2. A discrete distribution $\mathbb{R}$ over $[0, \infty)$ is a valid ratio if and only if $\mathbb{E}_{R \sim \mathbb{R}}[R] \leq 1$.

Proof. Suppose $\mathbb{R} = (\mathbb{P}\|\mathbb{Q})$, then $\mathbb{E}_{R \sim \mathbb{R}}[R] = \sum_{x \in \mathrm{Supp}(\mathbb{Q})} \mathbb{Q}(x)\frac{\mathbb{P}(x)}{\mathbb{Q}(x)} = \mathbb{P}(\mathrm{Supp}(\mathbb{Q})) \leq 1$.

On the other hand, if $\mathbb{E}_{R \sim \mathbb{R}}[R] \leq 1$, define distribution $\mathbb{R}^\dagger$ over $[0, \infty] = [0, \infty) \cup \{\infty\}$ as

$$\mathbb{R}^\dagger(r) = \begin{cases} r \cdot \mathbb{R}(r), & \text{if } r \in [0, \infty) \\ 1 - \mathbb{E}_{R \sim \mathbb{R}}[R], & \text{if } r = \infty. \end{cases} \tag{1}$$

It is easy to verify that $\mathbb{R}^\dagger$ is a distribution and $\mathbb{R} = (\mathbb{R}^\dagger\|\mathbb{R})$. $\qquad\square$

Definition 3 (Canonical Pair)**.** For any valid ratio $\mathbb{R}$, define its alternative ratio, denoted by $\mathbb{R}^\dagger$, as the distribution in (1). We call $(\mathbb{R}^\dagger, \mathbb{R})$ the *canonical pair* of $\mathbb{R}$.

We can define another natural distribution of likelihood ratio $(\mathbb{P}\|\mathbb{Q})^\dagger$

$$(\mathbb{P}\|\mathbb{Q})^\dagger(r) := \Pr_{X \sim \mathbb{P}}\Big[\frac{\mathbb{P}(X)}{\mathbb{Q}(X)} = r\Big].$$

Its sample space has to be extended to $[0, \infty]$, because the denominator can be zero. It is easy to verify that $\mathbb{R} = (\mathbb{P}\|\mathbb{Q})$ implies $\mathbb{R}^\dagger = (\mathbb{P}\|\mathbb{Q})^\dagger$.

Proposition 4. $(\mathbb{P}_1\mathbb{P}_2\|\mathbb{Q}_1\mathbb{Q}_2) = (\mathbb{P}_1\|\mathbb{Q}_1) \cdot_{\mathsf{indp}} (\mathbb{P}_2\|\mathbb{Q}_2)$ for any discrete distributions $\mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2$.

Proof. Let $(X_1, X_2) \sim \mathbb{Q}_1\mathbb{Q}_2$. Define $R_i := \frac{\mathbb{P}_i(X_i)}{\mathbb{Q}_i(X_i)}$, then $R_i \sim (\mathbb{P}_i \| \mathbb{Q}_i)$. Since X_1, X_2 are independent, R_1, R_2 are also independent. Define $R := R_1 R_2 = \frac{(\mathbb{P}_1\mathbb{P}_2)(X_1,X_2)}{(\mathbb{Q}_1\mathbb{Q}_2)(X_1,X_2)}$, then $R \sim (\mathbb{P}_1\mathbb{P}_2 \| \mathbb{Q}_1\mathbb{Q}_2)$. $\quad\square$

The likelihood ratio is a well-studied concept. Almost all statistical distances and divergences between $\mathbb{P}$ and $\mathbb{Q}$ can be derived from their ratio $\mathbb{R} = (\mathbb{P} \| \mathbb{Q})$. The derivation of the total variation distance is shown in the following proposition. The cases of other distances and divergences can be found in most information theory textbooks (e.g. [PW22]).

Proposition 5. Let $\mathbb{R} = (\mathbb{P} \| \mathbb{Q})$, then $\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \Delta_{\mathrm{TV}}(\mathbb{R})$, where

$$\Delta_{\mathrm{TV}}(\mathbb{R}) := \mathop{\mathbb{E}}_{R \sim \mathbb{R}}\big[(1 - R) \cdot \mathbb{1}[R < 1]\big] = \mathop{\mathbb{E}}_{R \sim \mathbb{R}}\big[\max(1 - R, 0)\big].$$

Proof. $\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \sum_{\substack{x \text{ s.t.} \\ \mathbb{Q}(x) > \mathbb{P}(x)}} \Big(\mathbb{Q}(x) - \mathbb{P}(x)\Big) = \sum_{\substack{x \text{ s.t.} \\ \mathbb{Q}(x) > \mathbb{P}(x)}} \mathbb{Q}(x)\Big(1 - \frac{\mathbb{P}(x)}{\mathbb{Q}(x)}\Big) = \mathop{\mathbb{E}}_{R \sim \mathbb{R}}\Big[(1 - R) \cdot \mathbb{1}[R < 1]\Big].$ $\quad\square$

It comes as no surprise how ratio can derive these distances and divergences. In some sense, ratio $\mathbb{R} = (\mathbb{P} \| \mathbb{Q})$ is a complete characterization of the decision problem between $\mathbb{P}$ and $\mathbb{Q}$.

Sufficient Statistic. Say random variable X is sampled from either $\mathbb{P}$ or $\mathbb{Q}$. Whether X is sampled from $\mathbb{P}$ or $\mathbb{Q}$ is the hidden parameter of the statistical model. Consider a statistic $R = \kappa(X)$ where $\kappa(x) := \frac{\mathbb{P}(x)}{\mathbb{Q}(x)}$. If the hidden parameter is $\mathbb{Q}$, then $R \sim \mathbb{R}$; otherwise, $R \sim \mathbb{R}^\dagger$. That is, $\mathbb{R} = \kappa\mathbb{Q}$, $\mathbb{R}^\dagger = \kappa\mathbb{P}$. It is well known that R is a *sufficient statistic* containing all the "useful" information of X. There exists a Markov kernel κ^{-1} who recovers the entire information of $(\mathbb{P}, \mathbb{Q})$ from $(\mathbb{R}^\dagger, \mathbb{R})$, such that $\mathbb{Q} = \kappa^{-1}\mathbb{R}$, $\mathbb{P} = \kappa^{-1}\mathbb{R}^\dagger$. So the problem of distinguishing $\mathbb{P}, \mathbb{Q}$ is *equivalent* to the problem of distinguishing $\mathbb{R}^\dagger, \mathbb{R}$.

Equivalent Decision Problems. In decision theory, the equivalence relationship between decision problems are formalized. Two decision problems $(\mathbb{P}, \mathbb{Q})$ and $(\mathbb{P}', \mathbb{Q}')$ are called equivalent, if there exist Markov kernels κ, κ^{-1} such that $\kappa\mathbb{P} = \mathbb{P}', \kappa\mathbb{Q} = \mathbb{Q}', \kappa^{-1}\mathbb{P}' = \mathbb{P}, \kappa^{-1}\mathbb{Q}' = \mathbb{Q}$ [LC12]. It is easy to verify that they are equivalent if and only if $(\mathbb{P}\|\mathbb{Q}) = (\mathbb{P}'\|\mathbb{Q}')$. If direction: let $\mathbb{R} = (\mathbb{P}\|\mathbb{Q}) = (\mathbb{P}'\|\mathbb{Q}')$, then both $(\mathbb{P}, \mathbb{Q})$ and $(\mathbb{P}', \mathbb{Q}')$ equivalent to the canonical pair $(\mathbb{R}^\dagger, \mathbb{R})$. Only if direction is ensured by a new metric defined in Section 3.1. Thus $\mathbb{R} = (\mathbb{P}\|\mathbb{Q})$ can represent the equivalence class where $(\mathbb{P}, \mathbb{Q})$ is.

The above discussion inspires us to consider the "data processing" relation. Say decision problem $(\mathbb{P}, \mathbb{Q})$ is stronger than $(\mathbb{P}', \mathbb{Q}')$, denoted by $(\mathbb{P}, \mathbb{Q}) \geq (\mathbb{P}', \mathbb{Q}')$, if and only if there exists a Markov kernel κ such that $\mathbb{P}' = \kappa\mathbb{Q}$, $\mathbb{Q}' = \kappa\mathbb{P}$. This relation is almost an order relation [LC12], satisfying reflexivity, transitivity and antisymmetry w.r.t. the equivalence relation between decision problems

$$(\mathbb{P}, \mathbb{Q}) \geq (\mathbb{P}', \mathbb{Q}') \ \wedge \ (\mathbb{P}', \mathbb{Q}') \geq (\mathbb{P}, \mathbb{Q}) \ \text{ if and only if } \ (\mathbb{P}\|\mathbb{Q}) = (\mathbb{P}'\|\mathbb{Q}').$$

Therefore, it induce an order relation between ratios.

Definition 6. For any two ratios $\mathbb{R}_1, \mathbb{R}_2$, we say "$\mathbb{R}_1$ is stronger than $\mathbb{R}_2$", denoted by $\mathbb{R}_1 \geq \mathbb{R}_2$, or "$\mathbb{R}_2$ is weaker than $\mathbb{R}_1$", denoted by $\mathbb{R}_2 \leq \mathbb{R}_1$, if there exists distributions $\mathbb{P}, \mathbb{Q}$ and a Markov kernel κ such that $\mathbb{R}_1 = (\mathbb{P}\|\mathbb{Q})$ and $\mathbb{R}_2 = (\kappa\mathbb{P}\|\kappa\mathbb{Q})$.

It captures the "data processing": κ is the process, $\kappa\mathbb{P}, \kappa\mathbb{Q}$ are the post-processing distributions. Therefore, by the data processing inequality for TV distance, $\mathbb{R}_1 \geq \mathbb{R}_2 \implies \Delta_{\mathrm{TV}}(\mathbb{R}_1) \geq \Delta_{\mathrm{TV}}(\mathbb{R}_2)$.

Proposition 7. For any ratios satisfying $\mathbb{R}_1 \geq \mathbb{R}_1'$, $\mathbb{R}_2 \geq \mathbb{R}_2'$, it holds that $\mathbb{R}_1 \cdot_{\mathsf{indp}} \mathbb{R}_2 \geq \mathbb{R}_1' \cdot_{\mathsf{indp}} \mathbb{R}_2'$.

Proof. There exist distributions $\mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2$ and Markov kernels κ_1, κ_2 such that $\mathbb{R}_i = (\mathbb{P}_i \| \mathbb{Q}_i)$ and $\mathbb{R}_i' = (\kappa_i \mathbb{P}_i \| \kappa_i \mathbb{Q}_i)$. Then $\mathbb{R}_1 \cdot_{\mathsf{indp}} \mathbb{R}_2 = (\mathbb{P}_1 \mathbb{P}_2 \| \mathbb{Q}_1 \mathbb{Q}_2)$ and $\mathbb{R}_1' \cdot_{\mathsf{indp}} \mathbb{R}_2' = ((\kappa_1 \mathbb{P}_1)(\kappa_2 \mathbb{P}_2) \| (\kappa_1 \mathbb{P}_1)(\kappa_2 \mathbb{P}_2))$. So it suffices to find a Markov kernel κ such that $\kappa(\mathbb{P}_1 \mathbb{P}_2) = (\kappa_1 \mathbb{P}_1)(\kappa_2 \mathbb{P}_2)$, $\kappa(\mathbb{Q}_1 \mathbb{Q}_2) = (\kappa_1 \mathbb{Q}_1)(\kappa_2 \mathbb{Q}_2)$. The required Markov kernel is $\kappa(x_1', x_2' | x_1, x_2) := \kappa_1(x_1' | x_1) \kappa_2(x_2' | x_2)$. □

3.1 Minimum Total Variation Distance

We introduce a metric between (valid) ratios, called the minimum total variation distance.

Definition 8 (MTV Distance). For two valid ratios $\mathbb{R}_1, \mathbb{R}_2$, the *minimum total variation distance* between them, denoted by $\Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_2)$, is defined as

$$\Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_2) := \inf_{\substack{discrete\ \mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2 \\ (\mathbb{P}_1 \| \mathbb{Q}_1) = \mathbb{R}_1 \\ (\mathbb{P}_2 \| \mathbb{Q}_2) = \mathbb{R}_2}} \max\big(\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2), \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_2)\big).$$

We define Δ_{MTV} as an infimum for safe. We cannot rule out the possibility that the minimum does not exist, unless the supports of $\mathbb{R}_1$ and $\mathbb{R}_2$ are finite (Lemma 9).

Lemma 9. *When the supports of ratios $\mathbb{R}_1$ and $\mathbb{R}_2$ are finite,*

$$\Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_2) = \min_{\substack{discrete\ \mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2 \\ (\mathbb{P}_1 \| \mathbb{Q}_1) = \mathbb{R}_1 \\ (\mathbb{P}_2 \| \mathbb{Q}_2) = \mathbb{R}_2}} \max\big(\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2), \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_2)\big).$$

The lemma is proved by enforcing the search space of $(\mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2)$ to be a tuple of distributions over a fixed finite domain. The proof is deferred to Appendix B.

The minimum total variance distance satisfies the following properties, as proved in Appendix B.

Lemma 10. *The minimum total variance distance is a metric.*

Lemma 11. *For any two valid ratios $\mathbb{R}_1, \mathbb{R}_2$, it holds that $|\Delta_{\mathrm{TV}}(\mathbb{R}_1) - \Delta_{\mathrm{TV}}(\mathbb{R}_2)| \leq 2\,\Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_2)$.*

Lemma 12. *For any valid ratios $\mathbb{R}_1, \mathbb{R}_2, \mathbb{R}_3, \mathbb{R}_4$,*

$$\Delta_{\mathrm{MTV}}(\mathbb{R}_1 \cdot_{\mathsf{indp}} \mathbb{R}_2, \mathbb{R}_3 \cdot_{\mathsf{indp}} \mathbb{R}_4) \leq \Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_3) + \Delta_{\mathrm{MTV}}(\mathbb{R}_2, \mathbb{R}_4).$$

A very similar concept called *Le Cam's distance* exists in decision theory. Here we present an equivalent definition restricting on the decision problems that has only two hypotheses.

Definition 13 (Deficiency and Le Cam's Distance [LC12]). For decision problems $(\mathbb{P}_1, \mathbb{Q}_1), (\mathbb{P}_2, \mathbb{Q}_2)$, the *deficiency* of $(\mathbb{P}_1, \mathbb{Q}_1)$ with respect to $(\mathbb{P}_2, \mathbb{Q}_2)$ is defined as

$$\mathrm{deficiency}((\mathbb{P}_1, \mathbb{Q}_1), (\mathbb{P}_2, \mathbb{Q}_2)) := \inf_{\substack{Markov\ kernel\ \kappa}} \max\big(\Delta_{\mathrm{TV}}(\kappa \mathbb{P}_1, \mathbb{P}_2), \Delta_{\mathrm{TV}}(\kappa \mathbb{Q}_1, \mathbb{Q}_2)\big);$$

their *Le Cam's distance* is $\max(\mathrm{deficiency}((\mathbb{P}_1, \mathbb{Q}_1), (\mathbb{P}_2, \mathbb{Q}_2)), \mathrm{deficiency}((\mathbb{P}_2, \mathbb{Q}_2), (\mathbb{P}_1, \mathbb{Q}_1)))$.

Le Cam's distance is a pseudo metric between decision problems. It is not hard to show that two decision problems $(\mathbb{P}_1, \mathbb{Q}_1), (\mathbb{P}_2, \mathbb{Q}_2)$ have zero Le Cam's distance if and only if $(\mathbb{P}_1 \| \mathbb{Q}_1) = (\mathbb{P}_2 \| \mathbb{Q}_2)$.

Therefore, Le Cam's distance is a metric between ratios. We can similarly define deficiency between ratios.

$$\text{deficiency}(\mathbb{R}_1, \mathbb{R}_2) := \inf_{\substack{discrete\ \mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2 \\ (\mathbb{P}_1 \| \mathbb{Q}_1) \leq \mathbb{R}_1 \\ (\mathbb{P}_2 \| \mathbb{Q}_2) = \mathbb{R}_2}} \max\left(\Delta_{\text{TV}}(\mathbb{P}_1, \mathbb{P}_2), \Delta_{\text{TV}}(\mathbb{Q}_1, \mathbb{Q}_2) \right),$$

$$\text{Le Cam's distance}(\mathbb{R}_1, \mathbb{R}_2) := \max(\text{deficiency}(\mathbb{R}_1, \mathbb{R}_2), \text{deficiency}(\mathbb{R}_2, \mathbb{R}_1)).$$

For any two ratios $\mathbb{R}_1, \mathbb{R}_2$, it holds that $\text{deficiency}(\mathbb{R}_1, \mathbb{R}_2) = 0$ if and only if $\mathbb{R}_1 \geq \mathbb{R}_2$, and Le Cam's distance$(\mathbb{R}_1, \mathbb{R}_2) \leq \Delta_{\text{MTV}}(\mathbb{R}_1, \mathbb{R}_2)$.

4   Sparsify the Likelihood Ratio

As discussed in Section 3, the ratio $\mathbb{R} = (\mathbb{P} \| \mathbb{Q})$ completely characterizes the problem of distinguishing $\mathbb{P}$ and $\mathbb{Q}$. Once the ratio is known, the distances between $\mathbb{P}, \mathbb{Q}$ (including the TV distance) follow easily. The bottleneck is the complexity of computing and representing $\mathbb{R}$. We will store ratio $\mathbb{R}$ as the table of values of its probability mass function, so the space complexity is proportional to the support size $|\text{Supp}(\mathbb{R})|$. When $\mathbb{P}, \mathbb{Q}$ are described as product distributions, the size of $\text{Supp}(\mathbb{R})$ can be exponentially large. Our solution is to simplify the ratio $\mathbb{R}$ without introducing too much error. The process of simplification is called the *sparsification* of the ratio. The amount as error is measured by MTV distance (Definition 8).

Lemma 14 (Sparsification). *There exists a deterministic algorithm* Sparsify *as defined in* (3). *Given a valid ratio $\mathbb{R}$ and two error bounds $\varepsilon_s, \delta_s > 0$ as inputs, it outputs a sparsified ratio $\tilde{\mathbb{R}} \leq \mathbb{R}$ in time $O(\frac{1}{\varepsilon_s} \log \frac{1}{\delta_s} + |\text{Supp}(\mathbb{R})|)$, such that $|\text{Supp}(\tilde{\mathbb{R}})| = O(\frac{1}{\varepsilon_s} \log \frac{1}{\delta_s})$ and $\Delta_{\text{MTV}}(\mathbb{R}, \tilde{\mathbb{R}}) \leq \frac{1}{2}(\varepsilon_s \Delta_{\text{TV}}(\mathbb{R}) + \delta_s)$.*

The intuition behind the sparsification has been sketched in the technical overview (Section 1.1). We divide $[0, \infty]$ into a collection of disjointed intervals, then "merge" all the probability masses within one interval together. We formalized this process as SparsifyWrtIntervals (Algorithm 1). The support size of the sparsified ratio is no more than the number of disjoint intervals we choose. The running time equals the input ratio support size plus the number of intervals, assuming the ratio is presented by a sorted table and the intervals is also sorted.

As a warm-up, consider how does the sparsification process work on a single interval. Let $(\mathbb{R}^\dagger, \mathbb{R})$ be the canonical pair of $\mathbb{R}$, let I be a sufficiently small interval. The sparsification of $\mathbb{R}$ with respect to an interval I works as follows:

$$\tilde{\mathbb{R}} = (\kappa \mathbb{R}^\dagger \| \kappa \mathbb{R}) \text{ where } \kappa \text{ is a deterministic Markov kernel } \kappa(r) = \begin{cases} \omega^*, & \text{if } r \in I \\ r, & \text{if } r \notin I \end{cases}$$

If $\mathbb{R}^\dagger(I) = \mathbb{R}(I) = 0$, the sparsification does nothing and $\tilde{\mathbb{R}} = \mathbb{R}$. Otherwise, the sparsified ratio and its alternative are

$$\tilde{\mathbb{R}}(r) = \begin{cases} \mathbb{R}(r), & \text{if } r \notin I \\ 0, & \text{if } r \in I \setminus \{r^*\} \\ \mathbb{R}(I), & \text{if } r = r^* \end{cases} \qquad \tilde{\mathbb{R}}^\dagger(r) = \begin{cases} \mathbb{R}^\dagger(r), & \text{if } r \notin I \\ 0, & \text{if } r \in I \setminus \{r^*\} \\ \mathbb{R}^\dagger(I), & \text{if } r = r^* \end{cases}$$

where $r^* := \dfrac{\mathbb{R}^\dagger(I)}{\mathbb{R}(I)}$. Apparently $r^* \in I$.

Algorithm 1: SparsifyWrtIntervals($\mathbb{R}, \{I_t\}_{1 \leq t \leq m}$)

Input: a ratio $\mathbb{R}$ with finite support and a collection of disjoint intervals
$\quad\quad I_1, \ldots, I_m \subseteq [0, \infty]$

1 Let $(\mathbb{R}^\dagger, \mathbb{R})$ be the canonical pair of $\mathbb{R}$;

2 Define distributions $\mathbb{P}^*, \mathbb{Q}^*$ over $\Omega^* = [0, \infty] - \bigcup_t I_t + \{\omega_1, \ldots, \omega_m\}$ as

$$\mathbb{P}^*(\omega_t) = \mathbb{R}^\dagger(I_t), \quad \mathbb{Q}^*(\omega_t) = \mathbb{R}(I_t) \quad \text{for } 1 \leq t \leq m;$$
$$\mathbb{P}^*(r) = \mathbb{R}^\dagger(r), \quad \mathbb{Q}^*(r) = \mathbb{R}(r) \quad \text{for } r \in [0, \infty] - \bigcup_t I_t.$$

Equivalently, $\mathbb{P}^* = \kappa\mathbb{R}^\dagger$, $\mathbb{Q}^* = \kappa\mathbb{R}$ for deterministic Markov kernel κ

$$\kappa(r) = \begin{cases} \omega_i, & \text{if } r \in I_i \text{ for some } i \in [m] \\ r, & \text{otherwise} \end{cases}$$

Output: a sparsified ratio $\tilde{\mathbb{R}} = (\mathbb{P}^* \| \mathbb{Q}^*)$

To bound $\Delta_{\mathrm{MTV}}(\mathbb{R}, \tilde{\mathbb{R}})$, consider the following two distributions $\tilde{\mathbb{P}}, \tilde{\mathbb{Q}}$

$$\tilde{\mathbb{Q}}(r) = \begin{cases} \mathbb{R}(r), & \text{if } r \notin I \\ \mathbb{R}^\dagger(r)/r^*, & \text{if } r \in I \end{cases} \quad\quad \tilde{\mathbb{P}}(r) = \begin{cases} \mathbb{R}^\dagger(r), & \text{if } r \notin I \\ \mathbb{R}(r) \cdot r^*, & \text{if } r \in I \end{cases}$$

They satisfy $\tilde{\mathbb{R}} = (\tilde{\mathbb{P}} \| \mathbb{R}) = (\mathbb{R}^\dagger \| \tilde{\mathbb{Q}})$. Therefore,

$$\Delta_{\mathrm{MTV}}(\mathbb{R}, \tilde{\mathbb{R}}) \leq \Delta_{\mathrm{TV}}(\tilde{\mathbb{P}}, \mathbb{R}^\dagger) = \frac{1}{2} \sum_{r \in I} \left| \mathbb{R}(r) \cdot r^* - \mathbb{R}^\dagger(r) \right| = \frac{1}{2} \sum_{r \in I} \mathbb{R}(r) \cdot |r^* - r|,$$

$$\Delta_{\mathrm{MTV}}(\mathbb{R}, \tilde{\mathbb{R}}) \leq \Delta_{\mathrm{TV}}(\tilde{\mathbb{Q}}, \mathbb{R}) = \frac{1}{2} \sum_{r \in I} \left| \mathbb{R}^\dagger(r)/r^* - \mathbb{R}(r) \right| = \frac{1}{2} \sum_{r \in I} \mathbb{R}^\dagger(r) \cdot |\tfrac{1}{r^*} - \tfrac{1}{r}|.$$

We should use the tighter one to bound $\Delta_{\mathrm{MTV}}(\mathbb{R}, \tilde{\mathbb{R}})$.

Intuitively, $\tilde{\mathbb{Q}}$ is obtained by modifying $\mathbb{R}$ at points $r \in I$; $\tilde{\mathbb{P}}$ is obtained by modifying $\mathbb{R}^\dagger$ at points $r \in I$. If the interval $I \subseteq [0, 1]$, then $\mathbb{R}^\dagger(r) \leq \mathbb{R}(r)$ for all $r \in I$, thus modifying $\mathbb{R}^\dagger$ introduces less error, in other words, $\Delta_{\mathrm{TV}}(\tilde{\mathbb{P}}, \mathbb{R}^\dagger)$ is a tighter bound of $\Delta_{\mathrm{MTV}}(\mathbb{R}, \tilde{\mathbb{R}})$. Symmetrically, if the interval $I \subseteq [1, \infty]$, $\Delta_{\mathrm{TV}}(\tilde{\mathbb{Q}}, \mathbb{R})$ is a tighter bound. This intuitive argument can be formalized by:

$$\Delta_{\mathrm{TV}}(\tilde{\mathbb{Q}}, \mathbb{R}) = \frac{1}{2} \sum_{r \in I} \mathbb{R}^\dagger(r) \cdot |\tfrac{1}{r^*} - \tfrac{1}{r}| = \frac{1}{2} \sum_{r \in I} \mathbb{R}(r) \cdot \frac{1}{r^*} \cdot |r^* - r| = \frac{1}{r^*} \cdot \Delta_{\mathrm{TV}}(\tilde{\mathbb{P}}, \mathbb{R}^\dagger).$$

Consider the case $I \subseteq [0, 1]$, let a, b be the endpoints of I, so $a < b \leq 1$. Then $\Delta_{\mathrm{TV}}(\tilde{\mathbb{P}}, \mathbb{R}^\dagger)$ can be upper bounded by one of the two following arguments:

- If $|r^* - r| \leq \varepsilon_{\mathrm{s}}(1 - r)$ for all $r \in I$, then

$$\Delta_{\mathrm{TV}}(\tilde{\mathbb{P}}, \mathbb{R}^\dagger) = \frac{1}{2} \sum_{r \in I} \mathbb{R}(r) \cdot |r^* - r| \leq \frac{1}{2} \varepsilon_{\mathrm{s}} \sum_{r \in I} \mathbb{R}(r) \cdot (1 - r).$$

  By its similarity with $\Delta_{\mathrm{TV}}(\mathbb{R}) = \sum_{r \leq 1} \mathbb{R}(r)(1 - r)$, even if we sparsify the ratio with respect to many disjointed intervals, the total error is bounded by $\frac{1}{2} \varepsilon_{\mathrm{s}} \Delta_{\mathrm{TV}}(\mathbb{R})$.

  To ensure $|r^* - r| \leq \varepsilon_{\mathrm{s}}(1 - r)$ for all $r \in I$, the two endpoints should satisfy $(b - a) \leq \varepsilon_{\mathrm{s}}(1 - b)$. Note that this cannot be satisfied if $b = 1$.

- If $|r^* - r| \leq \delta_s$ for all $r \in I$, then

$$\Delta_{\mathrm{TV}}(\tilde{\mathbb{P}}, \mathbb{R}^\dagger) = \frac{1}{2} \sum_{r \in I} \mathbb{R}(r) \cdot |r^* - r| \leq \frac{1}{2} \sum_{r \in I} \mathbb{R}(r) \cdot \delta_s \leq \frac{\delta_s}{2}.$$

To ensure $|r^* - r| \leq \delta_s$ for all $r \in I$, the two endpoints should satisfy $b - a \leq \delta_s$.

Based on the above observations, we choose the following partition of $[0, \infty]$, which is a collection of disjoint intervals

$$I_0, I_1, \ldots, I_{m-1}, I_m, \{1\}, J_m, J_{m-1}, \ldots, J_1, J_0,$$

in that order.

- Let $I_t := [a_t, a_{t+1})$ for every $t < m$, where $a_t := 1 - (1 + \varepsilon_s)^{-t}$. Thus $a_{t+1} - a_t \leq \varepsilon_s(1 - a_{t+1})$.

- Let $I_m := [a_m, 1)$ and let

$$m := \left\lceil \frac{-\log(\delta_s)}{\log(1 + \varepsilon_s)} \right\rceil. \tag{2}$$

Thus $1 - a_m \leq \delta_s$.

Symmetrically, let $J_t := (\frac{1}{a_{t+1}}, \frac{1}{a_t}]$ for $t < m$ and let $J_m := (1, \frac{1}{a_m}]$.

Putting the above intervals into SparsifyWrtIntervals (Algorithm 1) gives the Sparsify algorithm used in Lemma 14.

$$\mathsf{Sparsify}(\mathbb{R}, \varepsilon_s, \delta_s) = \mathsf{SparsifyWrtIntervals}(\mathbb{R}, \{I_0, I_1, \ldots, I_{m-1}, I_m, \{1\}, J_m, J_{m-1}, \ldots, J_1, J_0\}). \tag{3}$$

4.1 Proof of the Sparsification Lemma

This section proves the sparsification lemma (Lemma 14). The sparsified ratio $\tilde{\mathbb{R}} = (\kappa \mathbb{R}^\dagger \| \kappa \mathbb{R})$ for κ defined in Algorithm 1, thus $\tilde{\mathbb{R}} \leq \mathbb{R}$. The support size of the sparsified ratio is no more than the number of intervals, which is $O(\frac{1}{\varepsilon_s} \log \frac{1}{\delta_s})$. The running time is proportional to the number of intervals and the support size of the input ratio, which is $O(\frac{1}{\varepsilon_s} \log \frac{1}{\delta_s} + |\mathrm{Supp}(\mathbb{R})|)$. The rest of the section forces on bounding the error with respect to the minimum total variation distance.

For each $i \in [m]$, define $r_i^* \in I_i$ as

$$r_i^* := \begin{cases} \dfrac{\mathbb{R}^\dagger(I_i)}{\mathbb{R}(I_i)}, & \text{if } \mathbb{R}(I_i) \neq 0 \\ \text{any number in } I_i, & \text{if } \mathbb{R}(I_i) = 0 \end{cases}$$

Note that $\mathbb{R}(I_i) = 0$ implies $\mathbb{R}^\dagger(I_i) = 0$. Symmetrically, define $r_i^{**} \in J_i$ as

$$r_i^{**} := \begin{cases} \dfrac{\mathbb{R}^\dagger(J_i)}{\mathbb{R}(J_i)}, & \text{if } \mathbb{R}^\dagger(J_i) \neq 0 \\ \text{any number in } J_i, & \text{if } \mathbb{R}^\dagger(J_i) = 0 \end{cases}$$

The sparsified ratio $\tilde{\mathbb{R}}$ and its alternative $\tilde{\mathbb{R}}^\dagger$ can be written as

$$\tilde{\mathbb{R}}(r) = \begin{cases} \mathbb{R}(I_i), & \text{if } r = r_i^* \text{ for some } i \in [m] \\ \mathbb{R}(J_i), & \text{if } r = r_i^{**} \text{ for some } i \in [m] \\ \mathbb{R}(1), & \text{if } r = 1 \\ 0, & \text{otherwise} \end{cases} \qquad \tilde{\mathbb{R}}^\dagger(r) = \begin{cases} \mathbb{R}^\dagger(I_i), & \text{if } r = r_i^* \text{ for some } i \in [m] \\ \mathbb{R}^\dagger(J_i), & \text{if } r = r_i^{**} \text{ for some } i \in [m] \\ \mathbb{R}^\dagger(1), & \text{if } r = 1 \\ 0, & \text{otherwise} \end{cases}$$

11

To bound $\Delta_{\text{MTV}}(\mathbb{R}, \tilde{\mathbb{R}})$, consider the following two distributions, as inspired by the discussion of sparsification w.r.t. one interval,

$$\tilde{\mathbb{Q}}(r) = \begin{cases} \mathbb{R}(r), & \text{if } r \leq 1 \\ \mathbb{R}^\dagger(r)/r_i^{**}, & \text{if } r \in J_i \end{cases} \qquad \tilde{\mathbb{P}}(r) = \begin{cases} \mathbb{R}^\dagger(r), & \text{if } r \geq 1 \\ \mathbb{R}(r) \cdot r^*, & \text{if } r \in I_i \end{cases}$$

Claim 15. $\tilde{\mathbb{P}}, \tilde{\mathbb{Q}}$ are well-defined distributions and $\tilde{\mathbb{R}} = (\tilde{\mathbb{P}} \| \tilde{\mathbb{Q}})$.

Proof. For each $i \in [m]$, $\tilde{\mathbb{Q}}(J_i) = \mathbb{R}^\dagger(J_i)/r_i^{**} = \mathbb{R}(J_i)$. Thus

$$\tilde{\mathbb{Q}}([0, \infty]) = \tilde{\mathbb{Q}}([0,1]) + \sum_{i \in [m]} \tilde{\mathbb{Q}}(J_i) = \mathbb{R}([0,1]) + \sum_{i \in [m]} \mathbb{R}(J_i) = \mathbb{R}([0,\infty]) = 1 \,.$$

So $\tilde{\mathbb{Q}}$ is a distribution over $[0, \infty]$. Symmetrically, $\tilde{\mathbb{P}}(I_i) = \mathbb{R}(I_i) \cdot r_i^* = \mathbb{R}^\dagger(I_i)$, and $\tilde{\mathbb{P}}$ is a distribution. Say random variable $X \sim \tilde{\mathbb{Q}}$,

- $X \in I_i \iff \frac{\tilde{\mathbb{P}}(X)}{\tilde{\mathbb{Q}}(X)} = r_i^*$ and $\Pr[X \in I_i] = \tilde{\mathbb{Q}}(I_i) = \mathbb{R}(I_i)$.

- $X \in J_i \iff \frac{\tilde{\mathbb{P}}(X)}{\tilde{\mathbb{Q}}(X)} = r_i^{**}$ and $\Pr[X \in J_i] = \tilde{\mathbb{Q}}(J_i) = \mathbb{R}(J_i)$.

- $X = 1 \iff \frac{\tilde{\mathbb{P}}(X)}{\tilde{\mathbb{Q}}(X)} = 1$ and $\Pr[X = 1] = \tilde{\mathbb{Q}}(1) = \mathbb{R}(1)$.

Thus $\tilde{\mathbb{R}} = (\tilde{\mathbb{P}} \| \tilde{\mathbb{Q}})$. $\qquad \square$

Then $\Delta_{\text{MTV}}(\tilde{\mathbb{R}}, \mathbb{R})$ is upper bounded by $\Delta_{\text{MTV}}(\tilde{\mathbb{R}}, \mathbb{R}) \leq \max(\Delta_{\text{TV}}(\tilde{\mathbb{P}}, \mathbb{R}^\dagger), \Delta_{\text{TV}}(\tilde{\mathbb{Q}}, \mathbb{R}))$. Consider the two terms on the right hand side separately.

$$\Delta_{\text{TV}}(\tilde{\mathbb{P}}, \mathbb{R}^\dagger) = \frac{1}{2} \sum_{r<1} \left| \tilde{\mathbb{P}}(r) - \mathbb{R}^\dagger(r) \right| = \frac{1}{2} \sum_{i=1}^m \sum_{r \in I_i} \left| \tilde{\mathbb{P}}(r) - \mathbb{R}^\dagger(r) \right| = \frac{1}{2} \sum_{i=1}^m \sum_{r \in I_i} \mathbb{R}(r) \cdot \left| \tilde{r}_i^* - r \right|$$

By how we choose the intervals, for each $i \in [m]$ and any $\tilde{r}_i^*, r \in I_i$, we have

$$\left| r_i^* - r \right| \leq \begin{cases} \varepsilon_s(1 - r), & \text{if } i < m \\ \delta_s, & \text{if } i = m \end{cases}$$

Therefore,

$$\begin{aligned} \Delta_{\text{TV}}(\tilde{\mathbb{P}}, \mathbb{R}^\dagger) &\leq \frac{1}{2} \sum_{i=1}^{m-1} \sum_{r \in I_i} \mathbb{R}(r) \cdot \varepsilon_s(1 - r) + \frac{1}{2} \sum_{r \in I_m} \mathbb{R}(r) \cdot \delta_s \\ &\leq \frac{1}{2} \sum_{r<1} \mathbb{R}(r) \cdot \varepsilon_s(1 - r) + \frac{1}{2}\delta_s \\ &= \frac{1}{2}\varepsilon_s \Delta_{\text{TV}}(\mathbb{R}) + \frac{1}{2}\delta_s \,. \end{aligned}$$

Symmetrically, for each $i \in [m]$ and any $r_i^{**}, r \in J_i$, we have

$$\left| \frac{1}{r_i^{**}} - \frac{1}{r} \right| \leq \begin{cases} \varepsilon_s(1 - \frac{1}{r}), & \text{if } i < m \\ \delta_s, & \text{if } i = m \end{cases}$$

12

due to the choice of intervals. Then $\Delta_{\mathrm{TV}}(\tilde{\mathbb{Q}}, \mathbb{R})$ can be upper bounded by

$$\Delta_{\mathrm{TV}}(\tilde{\mathbb{Q}}, \mathbb{R}) \leq \frac{1}{2} \sum_{r>1} \mathbb{R}^{\dagger}(r) \cdot \varepsilon_{\mathrm{s}}\left(1 - \frac{1}{r}\right) + \frac{1}{2} \cdot \delta_{\mathrm{s}} = \frac{1}{2}\varepsilon_{\mathrm{s}}\,\Delta_{\mathrm{TV}}(\mathbb{R}) + \frac{1}{2}\delta_{\mathrm{s}}\,.$$

The last equality symbol relies on

$$\Delta_{\mathrm{TV}}(\mathbb{R}) = \Delta_{\mathrm{TV}}(\mathbb{R}^{\dagger}, \mathbb{R}) = \sum_{r \text{ s.t. } \mathbb{R}^{\dagger}(r) > \mathbb{R}(r)} (\mathbb{R}^{\dagger}(r) - \mathbb{R}(r)) = \sum_{r>1} \mathbb{R}^{\dagger}(r)\left(1 - \frac{1}{r}\right)\,.$$

As the conclusion, $\Delta_{\mathrm{MTV}}(\tilde{\mathbb{R}}, \mathbb{R}) \leq \max(\Delta_{\mathrm{TV}}(\tilde{\mathbb{P}}, \mathbb{R}^{\dagger}), \Delta_{\mathrm{TV}}(\tilde{\mathbb{Q}}, \mathbb{R})) \leq \frac{1}{2}(\varepsilon_{\mathrm{s}}\,\Delta_{\mathrm{TV}}(\mathbb{R}) + \delta_{\mathrm{s}})$.

5 Estimate TV Distance between Product Distributions

The two product distributions of interest are $\mathbb{P} = \mathbb{P}_1\mathbb{P}_2\ldots\mathbb{P}_n$ and $\mathbb{Q} = \mathbb{Q}_1\mathbb{Q}_2\ldots\mathbb{Q}_n$, where $\mathbb{P}_1$, $\mathbb{P}_2, \ldots, \mathbb{P}_n, \mathbb{Q}_1, \mathbb{Q}_2, \ldots, \mathbb{Q}_n$ are $2n$ distributions over sample space $[q]$. We present a deterministic algorithm that estimates the TV distance between $\mathbb{P}$ and $\mathbb{Q}$. The intuition has be explained in the technical overview (Section 1.1).

Theorem 1. *There exists a deterministic algorithm (Algorithm 2). Given two product distributions $\mathbb{P}, \mathbb{Q}$ over $[q]^n$ and $\varepsilon > 0$, it outputs an estimation $\widehat{\Delta}$ satisfying $(1-\varepsilon)\,\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \leq \widehat{\Delta} \leq \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$, in time $O\left(\frac{qn^2}{\varepsilon}\log q \log\frac{n}{\varepsilon\,\Delta_{\mathrm{TV}}(P,Q)}\right)$.*

Algorithm 2: A Deterministic FPTAS for the TV distance between product distributions

Input: $2n$ distributions $\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_n, \mathbb{Q}_1, \mathbb{Q}_2, \ldots, \mathbb{Q}_n$, and an error bound $\varepsilon > 0$;

1 Let $d_{\mathrm{LB}} \leftarrow \max_{1\leq i\leq n} \Delta_{\mathrm{TV}}(\mathbb{P}_i, \mathbb{Q}_i)$

2 Let $\mathbb{R}'_{1:1} \leftarrow \mathbb{R}_1$, where $\mathbb{R}_1 := (\mathbb{P}_1\|\mathbb{Q}_1)$

3 **for** k *from* 1 *to* $n-1$ **do**

4 $\quad$ $\tilde{\mathbb{R}}_{1:k} \leftarrow \mathsf{Sparsify}(\mathbb{R}'_{1:k}, \frac{\varepsilon}{2n}, \frac{\varepsilon}{2n}d_{\mathrm{LB}})$

5 $\quad$ $\mathbb{R}'_{1:k+1} \leftarrow \tilde{\mathbb{R}}_{1:k} \cdot_{\mathsf{indp}} \mathbb{R}_{k+1}$, where $\mathbb{R}_{k+1} := (\mathbb{P}_{k+1}\|\mathbb{Q}_{k+1})$

6 **return** $\widehat{\Delta} = \Delta_{\mathrm{TV}}(\mathbb{R}'_{1:n})$

The algorithm starts by computing a relatively tight lower bound of the $\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$.

Claim 16. $\frac{\Delta_{\mathrm{TV}}(\mathbb{P},\mathbb{Q})}{n} \leq d_{\mathrm{LB}} \leq \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$.

Proof. $\Delta_{\mathrm{TV}}(\mathbb{P}_i, \mathbb{Q}_i) \leq \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$ for every $i \in [n]$, thus $d_{\mathrm{LB}} \leq \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and (X_i, Y_i) is the optimal coupling of $(\mathbb{P}_i, \mathbb{Q}_i)$. Hence,

$$\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \leq \Pr[(X_1, \ldots, X_n) \neq (Y_1, \ldots, Y_n)] \leq \sum_{i=1}^{n} \Pr[X_i \neq Y_i] = \sum_{i=1}^{n} \Delta_{\mathrm{TV}}(\mathbb{P}_i, \mathbb{Q}_i) \leq nd_{\mathrm{LB}}\,. \quad \square$$

The algorithm then iteratively computes $\mathbb{R}'_{1:k}$ for $k = 1$ up to n. $\mathbb{R}'_{1:k}$ is an approximation of the "actual" ratio $\mathbb{R}_{1:k} := (\mathbb{P}_1\mathbb{P}_2\ldots\mathbb{P}_k\|\mathbb{Q}_1\mathbb{Q}_2\ldots\mathbb{Q}_k)$.

Claim 17. For each $k \in [n]$, it holds that $\mathbb{R}'_{1:k} \leq \mathbb{R}_{1:k}$ and $\Delta_{\mathrm{MTV}}(\mathbb{R}'_{1:k}, \mathbb{R}_{1:k}) \leq \frac{k-1}{2n}\varepsilon\,\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$.

Proof. It is proved by induction on k. The base case is satisfied as $\mathbb{R}'_{1:1} = (\mathbb{P}_1 \| \mathbb{Q}_1) = \mathbb{R}_{1:1}$.

For larger k, the inductive assumption is $\mathbb{R}'_{1:k-1} \leq \mathbb{R}_{1:k-1}$ and $\Delta_{\mathrm{MTV}}(\mathbb{R}'_{1:k-1}, \mathbb{R}_{1:k-1}) \leq \frac{k-2}{2n} \varepsilon \, \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$. By the data processing inequality for TV distance, $\Delta_{\mathrm{TV}}(\mathbb{R}'_{1:k-1}) \leq \Delta_{\mathrm{TV}}(\mathbb{R}_{1:k-1}) \leq \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$. By the nature of the sparsification process (Lemma 14), the sparsified ratio $\tilde{\mathbb{R}}_{1:k-1} \leq \mathbb{R}'_{1:k-1}$ and

$$\Delta_{\mathrm{MTV}}(\tilde{\mathbb{R}}_{1:k-1}, \mathbb{R}'_{1:k-1}) \leq \frac{1}{2}\Big(\frac{\varepsilon}{2n} \Delta_{\mathrm{TV}}(\mathbb{R}'_{1:k-1}) + \frac{\varepsilon}{2n} d_{\mathrm{LB}}\Big) \leq \frac{\varepsilon}{2n} \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \,. \tag{4}$$

Then the inductive proof is concluded by

$$\mathbb{R}'_{1:k} = \tilde{\mathbb{R}}_{1:k-1} \cdot_{\mathsf{indp}} \mathbb{R}_k \underset{\text{Proposition } 7}{\leq} \mathbb{R}_{1:k-1} \cdot_{\mathsf{indp}} \mathbb{R}_k = \mathbb{R}_{1:k} \,,$$

$$\Delta_{\mathrm{MTV}}(\mathbb{R}'_{1:k}, \mathbb{R}_{1:k}) \underset{\text{Lemma } 12}{\leq} \Delta_{\mathrm{MTV}}(\tilde{\mathbb{R}}_{1:k-1}, \mathbb{R}_{1:k-1})$$

$$\leq \Delta_{\mathrm{MTV}}(\tilde{\mathbb{R}}_{1:k-1}, \mathbb{R}'_{1:k-1}) + \Delta_{\mathrm{MTV}}(\mathbb{R}_{1:k-1}, \mathbb{R}'_{1:k-1}) \underset{\text{I.H. and (4)}}{\leq} \frac{1}{2n}\varepsilon \, \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) + \frac{k-2}{2n}\varepsilon \, \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \,.$$

$\square$

In the end, the algorithm outputs $\hat{\Delta} = \Delta_{\mathrm{TV}}(\mathbb{R}'_{1:n})$ as an estimation of the TV distance. Since $\Delta_{\mathrm{MTV}}(\mathbb{R}'_{1:n}, \mathbb{R}_{1:n}) \leq \frac{1}{2}\varepsilon \, \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$, we have

$$|\hat{\Delta} - \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})| = |\Delta_{\mathrm{TV}}(\mathbb{R}'_{1:n}) - \Delta_{\mathrm{TV}}(\mathbb{R}_{1:n})| \leq \varepsilon \, \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \,.$$

In the meanwhile, $\mathbb{R}'_{1:n} \leq \mathbb{R}_{1:n}$ implies

$$\hat{\Delta} = \Delta_{\mathrm{TV}}(\mathbb{R}'_{1:n}) \leq \Delta_{\mathrm{TV}}(\mathbb{R}_{1:n}) = \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \,.$$

The running time of our algorithm is dominated by the main loop. The support size of $\tilde{\mathbb{R}}_{1:k}$ is bounded by $O\big(\frac{n}{\varepsilon} \ln \frac{n}{\varepsilon d_{\mathrm{LB}}}\big) \leq O\big(\frac{n}{\varepsilon} \ln \frac{n}{\varepsilon \, \Delta_{\mathrm{TV}}(P,Q)}\big)$, by Lemma 14. The support size of $\mathbb{R}'_{1:k+1}$ is at most q times larger. The time complexity of computing $\mathbb{R}'_{1:k+1}$ is $O\big(\frac{qn}{\varepsilon} \log q \log \frac{n}{\varepsilon \, \Delta_{\mathrm{TV}}(P,Q)}\big)$, which is the complexity of merging q sorted list, each list of length $O\big(\frac{n}{\varepsilon} \log \frac{n}{\varepsilon \, \Delta_{\mathrm{TV}}(P,Q)}\big)$. The time complexity of computing $\tilde{\mathbb{R}}_{1:k+1}$ from $\mathbb{R}'_{1:k+1}$ is smaller. Thus the time complexity of each iteration is dominated by computing $\mathbb{R}'_{1:k+1}$, and the total time complexity is $O\big(\frac{qn^2}{\varepsilon} \log q \log \frac{n}{\varepsilon \, \Delta_{\mathrm{TV}}(P,Q)}\big)$.

6 Estimate TV Distance between Markov Chains

A Markov chain $\mathbb{P}$ over $[q]^n$ can be represented by its initial distribution $\mathbb{P}_1$ and $n-1$ Markov kernels $\kappa_2, \ldots, \kappa_n$ such that $\mathbb{P}(x_1, x_2, \ldots, x_n) = \mathbb{P}_1(x_1)\kappa_2(x_2|x_1)\ldots\kappa_n(x_n|x_{n-1})$. In the Markov chain setting, we use the conventional notation $\mathbb{P}_{k|k-1}$ to denote the Markov kernel κ_k. As the notation suggested, if $(X_1, \ldots, X_n) \sim \mathbb{P}$,

$$\mathbb{P}_{k|k-1}(x_k|x_{k-1}) = \Pr[X_k = x_k \mid X_{k-1} = x_{k-1}]$$

for all $x_{k-1}, x_k \in [q]$ such that $\Pr[X_{k-1} = x_{k-1}] > 0$.

Similarly, another Markov chain $\mathbb{Q}$ is represented by its initial distribution $\mathbb{Q}_1$ and Markov kernels $\mathbb{Q}_{2|1}, \ldots, \mathbb{Q}_{n|n-1}$. We give a deterministic FPTAS for the TV distance between $\mathbb{P}$ and $\mathbb{Q}$.

Theorem 2. *There is a deterministic algorithm (Algorithm 3) such that given two Markov chains $\mathbb{P}, \mathbb{Q}$ over $[q]^n$ and an error bound $\varepsilon > 0$, it outputs $\hat{\Delta}$ satisfying $(1-\varepsilon) \, \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \leq \hat{\Delta} \leq \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$ in time $O(\frac{n^2 q^2}{\varepsilon} \log q \log(\frac{n}{\varepsilon \, \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})}))$.*

14

To present our algorithm, we introduce a few more notations, in particular, the *conditional ratio between Markov kernels*, which plays the central role in our algorithm.

The conventional notation $\mathbb{P}_{k:n|k-1}$ denotes a Markov kernel from $[q]$ to $[q]^{n-k+1}$, is defined as

$$\mathbb{P}_{k:n|k-1}(x_k, x_{k+1}, \ldots, x_n | x_{k-1}) := \mathbb{P}_{n|n-1}(x_n|x_{n-1})\mathbb{P}_{n-1|n-2}(x_{n-1}|x_{n-2})\ldots\mathbb{P}_{k|k-1}(x_k|x_{k-1}).$$

As the notation suggested, if $(X_1, \ldots, X_n) \sim \mathbb{P}$,

$$\mathbb{P}_{k:n|k-1}(x_k, x_{k+1}, \ldots, x_n | x_{k-1}) = \Pr[(X_k, \ldots, X_n) = (x_k, \ldots, x_n) \mid X_{k-1} = x_{k-1}]$$

for all $x_{k-1}, x_k, \ldots, x_n \in [q]$ such that $\Pr[X_{k-1} = x_{k-1}] > 0$. We use $\mathbb{P}_{k|k-1}|x$ and $\mathbb{P}_{k:n|k-1}|x$ to denote the derived distributions $\mathbb{P}_{k|k-1}(\ldots|x)$ and $\mathbb{P}_{k:n|k-1}(\ldots|x)$ respectively. That is

$$(\mathbb{P}_{k|k-1}|x_{k-1})(x_k) := \mathbb{P}_{k|k-1}(x_k|x_{k-1}), \qquad (\mathbb{P}_{k:n|k-1}|x_{k-1})(x_{k:n}) := \mathbb{P}_{k:n|k-1}(x_{k:n}|x_{k-1}).$$

They can be viewed as the conditional distributions of X_k and $X_{k:n}$.

Then we can consider the ratio between $\mathbb{P}_{k:n|k-1}|x$ and $\mathbb{Q}_{k:n|k-1}|x$, defined as

$$\mathbb{R}_{k:n|k-1}|x := \left(\mathbb{P}_{k:n|k-1}|x \,\|\, \mathbb{Q}_{k:n|k-1}|x\right).$$

This implicitly defines a Markov kernel $\mathbb{R}_{k:n|k-1}$ from $[q]$ to $[0, \infty)$, as formalized below.

Definition 18 (Conditional Ratio between Markov kernels)**.** For two Markov kernels $\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}$ from Ω_X to Ω_Y, the *conditional ratio* between them, denoted as $(\mathbb{P}_{Y|X}\|\mathbb{Q}_{Y|X})$, is a Markov kernel from Ω_X to $[0, \infty)$, such that, if letting $\mathbb{R}_{Y|X} := (\mathbb{P}_{Y|X}\|\mathbb{Q}_{Y|X})$, for every $x \in \Omega_X$

$$\mathbb{R}_{Y|X}|x = \left(\mathbb{P}_{Y|X}|x \,\|\, \mathbb{Q}_{Y|X}|x\right).$$

A *valid conditional ratio* over Ω_X is Markov kernel κ from Ω_X to $[0, \infty)$, such that $\kappa(\cdot|x)$ is a valid ratio for every $x \in \Omega_X$,

Conditional ratio plays the central role in our algorithm, as presented in Algorithm 3. The core of the algorithm is an iterative process computing an approximation of the conditional ratio $\mathbb{R}_{k:n|k-1}$ for each k. In each iteration step, to compute an approximation of $\mathbb{R}_{k:n|k-1}|x$, the algorithm need an approximation of $\mathbb{R}_{k+1:n|k}$ from last iteration, and distributions $\mathbb{P}_{k|k-1}|x, \mathbb{Q}_{k|k-1}|x$ from the inputs. We call this subroutine Concatenate, and it is defined and analyzed in Section 6.1.

6.1 The Concatenation of Conditional Ratios

Given $\mathbb{P}_X, \mathbb{P}_{Y|X}$ and $\mathbb{Q}_X, \mathbb{Q}_{Y|X}$, to compute the ratio $\mathbb{R}_{XY} := (\mathbb{P}_X\mathbb{P}_{Y|X}\|\mathbb{Q}_X\mathbb{Q}_{Y|X})$, the naïve solution is to first compute the two joint distributions $\mathbb{P}_X\mathbb{P}_{Y|X}, \mathbb{Q}_X\mathbb{Q}_{Y|X}$. This section shows that there is an alternative approach: first compute $\mathbb{R}_{Y|X} := (\mathbb{P}_{Y|X}\|\mathbb{Q}_{Y|X})$, then

$$\mathbb{R}_{XY} \text{ can be computed from } \mathbb{P}_X, \mathbb{Q}_X \text{ and } \mathbb{R}_{Y|X}.$$

Let $(X, Y) \sim \mathbb{Q}_X\mathbb{Q}_{Y|X}$ and set

$$R = \frac{(\mathbb{P}_X\mathbb{P}_{Y|X})(X, Y)}{(\mathbb{Q}_X\mathbb{Q}_{Y|X})(X, Y)} = \frac{\mathbb{P}_X(X)}{\mathbb{Q}_X(X)}\frac{\mathbb{P}_{Y|X}(Y|X)}{\mathbb{Q}_{Y|X}(Y|X)}.$$

then $R \sim \mathbb{R}_{XY} = (\mathbb{P}_X\mathbb{P}_{Y|X}\|\mathbb{Q}_X\mathbb{Q}_{Y|X})$. What is the distribution of R conditioning on $X = x$? Let $\mathbb{R}_x$ denote this conditional distribution. Then $\mathbb{R}_x$ is the distribution of

$$\frac{\mathbb{P}_X(x)}{\mathbb{Q}_X(x)}\frac{\mathbb{P}_{Y|X}(Y_x|x)}{\mathbb{Q}_{Y|X}(Y_x|x)} \tag{5}$$

Algorithm 3: A Deterministic FPTAS for the TV distance between two Markov chains

Input: two Markov chains $\mathbb{P} = \mathbb{P}_1 \mathbb{P}_{2|1} \ldots \mathbb{P}_{n|n-1}$, $\mathbb{Q} = \mathbb{Q}_1 \mathbb{Q}_{2|1} \ldots \mathbb{Q}_{n|n-1}$,
   an error bound $\varepsilon > 0$

1 Compute d_{LB} such that $\frac{\Delta_{\mathrm{TV}}(\mathbb{P},\mathbb{Q})}{2n} \le d_{\mathrm{LB}} \le \Delta_{\mathrm{TV}}(\mathbb{P},\mathbb{Q})$ (Section 6.2)

2 Compute conditional ratio $\mathbb{R}'_{n:n|n-1} \leftarrow (\mathbb{P}_{n|n-1} \| \mathbb{Q}_{n|n-1})$

   That is, $\mathbb{R}'_{n:n|n-1}|x \leftarrow (\mathbb{P}_{n|n-1}|x \| \mathbb{Q}_{n|n-1}|x)$ for each $x \in [q]$

3 **for** k *from* $n-1$ *to* 1 **do**

4   Compute conditional ratio $\tilde{\mathbb{R}}_{k+1:n|k}$ as follows: **for** x *in* $[q]$ **do**

5          $\tilde{\mathbb{R}}_{k+1:n|k}|x \leftarrow \mathsf{Sparsify}(\mathbb{R}'_{k+1:n|k}|x,\ \frac{\varepsilon}{4n},\ \frac{\varepsilon}{2n}d_{\mathrm{LB}})$

6   **if** $k > 1$ **then**

7   Compute conditional ratio $\mathbb{R}'_{k:n|k-1}$ as follows: **for** x *in* $[q]$ **do**

8              $\mathbb{R}'_{k:n|k-1}|x \leftarrow \mathsf{Concatenate}(\mathbb{P}_{k|k-1}|x,\ \mathbb{Q}_{k|k-1}|x,\ \tilde{\mathbb{R}}_{k+1:n|k})$

9   **else**

10   Compute ratio $\mathbb{R}'_{1:n} \leftarrow \mathsf{Concatenate}(\mathbb{P}_1,\ \mathbb{Q}_1,\ \tilde{\mathbb{R}}_{2:n|1})$

Output: $\widehat{\Delta} = \Delta_{\mathrm{TV}}(\mathbb{R}'_{1:n})$

---

where Y_x is sampled from the distribution of Y conditioning on $X = x$. In other words, $Y_x \sim \mathbb{Q}_{Y|X}|x$. In such case, the distribution of the second fraction in (5) is $\mathbb{R}_{Y|X}|x$. Therefore, $\mathbb{R}_x$ can be simply computed from $\mathbb{R}_{Y|X}|x$, and the wanted ratio $\mathbb{R}_{XY}$ is a weighted average of $\mathbb{R}_x$.

If we let $\mathbb{D}_{X \sim \mathbb{Q}}[f(X)]$ denote the distribution of $f(X)$ when $X \sim \mathbb{Q}$, the above analysis can be written as one formula.

$$
(\mathbb{P}_X \mathbb{P}_{Y|X} \| \mathbb{Q}_X \mathbb{Q}_{Y|X}) = \mathop{\mathbb{D}}_{(X,Y) \sim \mathbb{Q}_X \mathbb{Q}_{Y|X}} \left[\frac{\mathbb{P}_X(X)}{\mathbb{Q}_X(X)} \frac{\mathbb{P}_{Y|X}(Y|X)}{\mathbb{Q}_{Y|X}(Y|X)} \right]
$$

$$
= \sum_x \mathbb{Q}_X(x) \mathop{\mathbb{D}}_{Y \sim \mathbb{Q}_{Y|X}|x} \left[\frac{\mathbb{P}_X(x)}{\mathbb{Q}_X(x)} \frac{\mathbb{P}_{Y|X}(Y|x)}{\mathbb{Q}_{Y|X}(Y|x)} \right] = \sum_x \mathbb{Q}_X(x) \mathop{\mathbb{D}}_{R \sim \mathbb{R}_{Y|X}|x} \left[\frac{\mathbb{P}_X(x)}{\mathbb{Q}_X(x)} \cdot R \right] \quad (6)
$$

Lemma 19 (Concatenation)**.** *There is a deterministic algorithm* $\mathsf{Concatenate}$ *inspired by* (6). *Given distributions* $\mathbb{P}_X, \mathbb{Q}_X$ *over* Ω_X *and a conditional ratio* $\mathbb{R}_{Y|X}$ *over* Ω_X, *the algorithm outputs* $\mathbb{R}_{XY} = \mathsf{Concatenate}(\mathbb{P}_X, \mathbb{Q}_X, \mathbb{R}_{Y|X})$ *such that*

$$
\mathbb{R}_{XY} = (\mathbb{P}_X \mathbb{P}_{Y|X} \| \mathbb{Q}_X \mathbb{Q}_{Y|X})
$$

for any sample space Ω_Y *and any Markov kernels* Ω_X *to* Ω_Y *satisfying* $(\mathbb{P}_{Y|X} \| \mathbb{Q}_{Y|X}) = \mathbb{R}_{Y|X}$.

If the support size of $\mathbb{R}_{Y|X}|x$ *is less than* N *for every* $x \in \Omega_X$, *the support size of* $\mathbb{R}_{XY}$ *is no more than* $N|\Omega_X|$, *the time complexity of* $\mathsf{Concatenate}$ *is bounded by* $N|\Omega_X| \log(|\Omega_X|)$.

Proof. The correctness is ensured by (6).

The complexity bottleneck is to merge $|\Omega_X|$ sorted lists, each list has up to N elements. $\qquad\square$

The concatenation process has a few nice properties. In particular, concatenation preserves the order relation (Definition 6) and does not increase minimum total variation distance.

Lemma 20. *For any conditional ratios* $\mathbb{R}_{Y|X}, \mathbb{R}'_{Y|X}$ *over* Ω_X *such that* $\mathbb{R}_{Y|X}|x \ge \mathbb{R}'_{Y|X}|x$ *for every* $x \in \Omega_X$, *for any distributions* $\mathbb{P}_X, \mathbb{Q}_X$ *over* Ω_X, *it holds that*

$$
\mathsf{Concatenate}(\mathbb{P}_X, \mathbb{Q}_X, \mathbb{R}_{Y|X}) \ge \mathsf{Concatenate}(\mathbb{P}_X, \mathbb{Q}_X, \mathbb{R}'_{Y|X}).
$$

Proof. Since $\mathbb{R}_{Y|X}|x \geq \mathbb{R}'_{Y|X}|x$ for every $x \in \Omega_X$, there exist a sample space $\Omega_Y^{(x)}$, distributions $\mathbb{P}^{(x)}, \mathbb{Q}^{(x)}$ over $\Omega_Y^{(x)}$, and a Markov kernel κ_x from $\Omega_Y^{(x)}$ to $\Omega_Y^{(x)}$ such that

$$\mathbb{R}_{Y|X}|x = \left(\mathbb{P}^{(x)} \| \mathbb{Q}^{(x)} \right), \qquad \mathbb{R}'_{Y|X}|x = \left(\kappa_x \mathbb{P}^{(x)} \| \kappa_x \mathbb{Q}^{(x)} \right).$$

Define sample space $\Omega_Y = \bigcup_x \Omega_Y^{(x)}$, Markov kernels $\mathbb{P}_{Y|X}, \mathbb{P}'_{Y|X}, \mathbb{Q}_{Y|X}, \mathbb{Q}'_{Y|X}$ from Ω_X to Ω_Y as

$$\mathbb{P}_{Y|X}|x := \mathbb{P}^{(x)}, \quad \mathbb{Q}_{Y|X}|x := \mathbb{Q}^{(x)}, \quad \mathbb{P}'_{Y|X}|x := \kappa_x \mathbb{P}^{(x)}, \quad \mathbb{Q}'_{Y|X}|x := \kappa_x \mathbb{Q}^{(x)}.$$

Then $\mathbb{R}_{Y|X} = (\mathbb{P}_{Y|X} \| \mathbb{Q}_{Y|X})$ and $\mathbb{R}'_{Y|X} = (\mathbb{P}'_{Y|X} \| \mathbb{Q}'_{Y|X})$. So the two sides of the wanted inequality equal $(\mathbb{P}_X \mathbb{P}_{Y|X} \| \mathbb{Q}_X \mathbb{Q}_{Y|X})$ and $(\mathbb{P}_X \mathbb{P}'_{Y|X} \| \mathbb{Q}_X \mathbb{Q}'_{Y|X})$ respectively.

Define a Markov kernel κ from $\Omega_X \times \Omega_Y$ to itself as

$$\kappa(x', y'|x, y) := \mathbb{1}[x' = x] \cdot \kappa_x(y'|y).$$

Intuitively, the distribution of κ conditioning on (x, y) is to set $X' = x$ and sample Y' according to $\kappa_x(\cdot|y)$. It is easy to verify that $\kappa(\mathbb{P}_X \mathbb{P}_{Y|X}) = \mathbb{P}_X \mathbb{P}'_{Y|X}$ and $\kappa(\mathbb{Q}_X \mathbb{Q}_{Y|X}) = \mathbb{Q}_X \mathbb{Q}'_{Y|X}$. Therefore,

$$\mathsf{Concatenate}(\mathbb{P}_X, \mathbb{Q}_X, \mathbb{R}_{Y|X}) = \left(\mathbb{P}_X \mathbb{P}_{Y|X} \,\|\, \mathbb{Q}_X \mathbb{Q}_{Y|X} \right)$$
$$\geq \left(\mathbb{P}_X \mathbb{P}'_{Y|X} \,\|\, \mathbb{Q}_X \mathbb{Q}'_{Y|X} \right) = \mathsf{Concatenate}(\mathbb{P}_X, \mathbb{Q}_X, \mathbb{R}'_{Y|X}). \quad \square$$

6.2 A Lower Bound for TV Distance

Similar to the case of product distributions (Section 5), our algorithm approximating the TV distance between Markov chains needs a relative tight lower bound to start with.

Lemma 21. *There exists an deterministic algorithm such that given two Markov chains $\mathbb{P}$ and $\mathbb{Q}$ over $[q]^n$, it returns d_{LB} satisfying $\frac{\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})}{2n} \leq d_{\mathrm{LB}} \leq \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$ in time $O(q^2 n)$.*

Proof. Define $\mathbb{D}_0 := \mathbb{P}$. For each $k \in [n]$, define distribution $\mathbb{D}_k := \mathbb{Q}_1 \cdots \mathbb{Q}_{k|k-1} \mathbb{P}_{k+1|k} \cdots \mathbb{P}_{n|n-1}$.

$$\mathbb{D}_k(x_1, \ldots, x_n) = \mathbb{Q}_{1:k}(x_1, \ldots, x_k) \mathbb{P}_{k+1:n|k}(x_{k+1}, \ldots, x_n|x_k)$$

Thus $\mathbb{D}_n = \mathbb{Q}$. By the triangle inequality,

$$\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \leq \sum_{k=1}^n \Delta_{\mathrm{TV}}(\mathbb{D}_{k-1}, \mathbb{D}_k) \leq n \max_{1 \leq k \leq n} \Delta_{\mathrm{TV}}(\mathbb{D}_{k-1}, \mathbb{D}_k).$$

For any $1 \leq k \leq n$, it holds that

$$\Delta_{\mathrm{TV}}(\mathbb{D}_{k-1}, \mathbb{D}_k) \leq \Delta_{\mathrm{TV}}(\mathbb{D}_{k-1}, \mathbb{P}) + \Delta_{\mathrm{TV}}(\mathbb{D}_k, \mathbb{P})$$
$$= \Delta_{\mathrm{TV}}(\mathbb{Q}_{1:k-1}, \mathbb{P}_{1:k-1}) + \Delta_{\mathrm{TV}}(\mathbb{Q}_{1:k}, \mathbb{P}_{1:k}) \leq 2 \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}).$$

Therefore, d_{LB} satisfies the requirement if it is defined as

$$d_{\mathrm{LB}} := \frac{1}{2} \max_{1 \leq k \leq n} \Delta_{\mathrm{TV}}(\mathbb{D}_{k-1}, \mathbb{D}_k).$$

Note that, d_{LB} can be computed in time $O(q^2 n)$ because

$$\Delta_{\mathrm{TV}}(\mathbb{D}_{k-1}, \mathbb{D}_k) = \Delta_{\mathrm{TV}}(\mathbb{Q}_{k-1} \mathbb{P}_{k|k-1}, \mathbb{Q}_{k-1} \mathbb{Q}_{k|k-1})$$

and we can compute all $\mathbb{Q}_{k-1} \mathbb{P}_{k|k-1}, \mathbb{Q}_{k-1} \mathbb{Q}_{k|k-1}$ in $O(q^2 n)$ time. (Here $\mathbb{Q}_k$ denote the k-th marginal distribution of $\mathbb{Q}$, which can be recursively computed by $\mathbb{Q}_k := \mathbb{Q}_{k|k-1} \mathbb{Q}_{k-1}$.) $\quad \square$

6.3 Analysis of our Algorithm

In this section, we analyze Algorithm 3 and prove Theorem 2.

First analyze the time complexity. Let $N = O(\frac{n}{\varepsilon} \log(\frac{n}{\varepsilon \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})}))$ denote an upper bound of the support size of every sparsified ratios generated by Sparsify. The support size of each concatenation ratio $\mathbb{R}'_{k:n|k-1}|x$ is bounded by Nq. The complexity is dominated by nq calls to Sparsify and nq calls to Concatenate. Each Sparsify call takes at most $O(Nq)$ time, each Concatenate call takes at most $O(Nq \log q)$ time. Thus the total time complexity is bound by $O(Nnq^2 \log q)$.

Let $\mathbb{R}_{1:n}$, $\mathbb{R}_{k:n|k-1}$ denote the actual ratio and conditional ratio

$$\mathbb{R}_{1:n} := (\mathbb{P}_{1:n} \| \mathbb{Q}_{1:n}), \qquad \mathbb{R}_{k:n|k-1} := \left(\mathbb{P}_{k:n|k-1} \,\|\, \mathbb{Q}_{k:n|k-1}\right).$$

Our algorithm recursively computes a good approximation of $\mathbb{R}_{k:n|k-1}$ for smaller and smaller k. In the end, the our algorithm computes $\mathbb{R}'_{1:n}$, which is an approximation of $\mathbb{R}_{1:n}$, satisfying $\mathbb{R}'_{1:n} \le \mathbb{R}_{1:n}$ and $\Delta_{\mathrm{MTV}}(\mathbb{R}'_{1:n}, \mathbb{R}_{1:n}) \le \frac{1}{2}\varepsilon \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$. Thus $\widehat{\Delta} = \Delta_{\mathrm{TV}}(\mathbb{R}'_{1:n})$ is a sufficiently good approximation of $\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$.

In the proof, we use the conventional notation $\mathbb{P}_k$ to denote the marginal distribution of the k-th coordinate of $\mathbb{P}$. It can be recursively defined by $\mathbb{P}_k = \mathbb{P}_{k|k-1}\mathbb{P}_{k-1}$. The conventional notation $\mathbb{P}_{t|k}$ is a Markov kernel specifying the distribution of the t-th coordinate conditioning on the k-th coordinate.

Claim 22. For every $1 < k \le n$, the conditional ratio $\mathbb{R}'_{k:n|k-1}$ satisfies $\mathbb{R}'_{k:n|k-1}|x \le \mathbb{R}_{k:n|k-1}|x$ for all $x \in [q]$. The ratio $\mathbb{R}'_{1:n}$ satisfies $\mathbb{R}'_{1:n} \le \mathbb{R}_{1:n}$.

Proof. There is an inductive proof from larger k to smaller k. The base case is when $k = n$. The base case follows trivially from the fact that $\mathbb{R}'_{n:n|n-1} = \mathbb{R}_{n:n|n-1}$.

For smaller k, by the inductive hypothesis, $\mathbb{R}'_{k+1:n|k}|x \le \mathbb{R}_{k+1:n|k}|x$ for each $x \in [q]$. Order is preserved by sparsification (Lemma 14), so $\tilde{\mathbb{R}}_{k+1:n|k}|x \le \mathbb{R}'_{k+1:n|k}|x \le \mathbb{R}_{k+1:n|k}|x$. Order is also preserved by concatenation (Lemma 20), so

$$\mathbb{R}'_{k:n|k-1}|x = \mathsf{Concatenate}(\mathbb{P}_{k|k-1}|x, \ \mathbb{Q}_{k|k-1}|x, \ \tilde{\mathbb{R}}_{k+1:n|k})$$
$$\le \mathsf{Concatenate}(\mathbb{P}_{k|k-1}|x, \ \mathbb{Q}_{k|k-1}|x, \ \mathbb{R}_{k+1:n|k}) = \mathbb{R}_{k:n|k-1}|x.$$

In the special case when $k = 1$, by the same argument (Lemma 14), $\tilde{\mathbb{R}}_{2:n|1}|x \le \mathbb{R}_{2:n|1}|x$. Then by a similar argument about concatenation (Lemma 20),

$$\mathbb{R}'_{1:n} = \mathsf{Concatenate}(\mathbb{P}_1, \mathbb{Q}_1, \tilde{\mathbb{R}}_{2:n|1}) \le \mathsf{Concatenate}(\mathbb{P}_1, \mathbb{Q}_1, \mathbb{R}_{2:n|1}) = \mathbb{R}_{1:n}. \qquad \square$$

Claim 23. The ratio $\mathbb{R}'_{1:n}$ in Algorithm 3 satisfies $\Delta_{\mathrm{MTV}}(\mathbb{R}'_{1:n}, \mathbb{R}_{1:n}) \le \frac{n-1}{2n}\varepsilon \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$.

We present a rather intuitive proof here. A more formal proof is given in Appendix A.

Proof. Image that there are n hybrid worlds, numbered by $i = 1, \ldots, n$. In the i-th hybrid world, we modify how Algorithm 3 works.

The i-th hybrid world consider a modified algorithm that skips the sparsification step in the main loop for every index $k < i$. In other words, let $\tilde{\mathbb{R}}_{k+1:n|k} = \mathbb{R}'_{k+1:n|k}$ for every $k < i$. To avoid confusion, we use $\mathbb{R}^{(i)}_{k+1:n|k}$ to denote this value for every $k < i$, we use $\mathbb{R}^{(i)}_{1:n}$ to denote the value of $\mathbb{R}'_{1:n}$ in the i-th hybrid world.

The 1st hybrid world is identical to the real world, thus $\mathbb{R}_{1:n}^{(1)} = \mathbb{R}'_{1:n}$. The last hybrid world is "ideal" in the sense that there is no error introduced by sparsification, thus $\mathbb{R}_{1:n}^{(n)} = \mathbb{R}_{1:n}$. To bound the MTV distance between $\mathbb{R}'_{1:n}$ and $\mathbb{R}_{1:n}$, it suffice to prove for every $1 \le i < n$,

$$\Delta_{\mathrm{MTV}}(\mathbb{R}_{1:n}^{(i)}, \mathbb{R}_{1:n}^{(i+1)}) \le \frac{1}{2n}\varepsilon\,\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})\,.$$

Comparing the i-th and $(i+1)$-th hybrid worlds, the only difference is whether the sparsification step is skipped when $k = i$. In the $(i + 1)$-th hybrid world, $\mathbb{R}_{i+1:n|i}^{(i+1)} = \mathbb{R}'_{i+1:n|i}$. In the i-th hybrid world, $\tilde{\mathbb{R}}_{i+1:n|i}|x$ is the sparsification of $\mathbb{R}'_{i+1:n|i}|x$. By Lemma 14,

$$\Delta_{\mathrm{MTV}}(\tilde{\mathbb{R}}_{i+1:n|i}|x, \mathbb{R}_{i+1:n|i}^{(i+1)}|x) \le \frac{\varepsilon}{4n}d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\,\Delta_{\mathrm{TV}}(\mathbb{R}_{i+1:n|i}^{(i+1)}|x) \le \frac{\varepsilon}{4n}d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\,\Delta_{\mathrm{TV}}(\mathbb{R}_{i+1:n|i}|x)\,.$$

So there exist Markov kernels $\mathbb{P}_{Y|i}^{(i)*}, \mathbb{Q}_{Y|i}^{(i)*}, \mathbb{P}_{Y|i}^{(i+1)**}, \mathbb{Q}_{Y|i}^{(i+1)**}$ from $[q]$ to a sample space $\Omega_Y^{(i)}$ satisfying

$$\tilde{\mathbb{R}}_{i+1:n|i} = \left(\mathbb{P}_{Y|i}^{(i)*} \,\|\, \mathbb{Q}_{Y|i}^{(i)*}\right), \qquad \mathbb{R}_{i+1:n|i}^{(i+1)} = \left(\mathbb{P}_{Y|i}^{(i+1)**} \,\|\, \mathbb{Q}_{Y|i}^{(i+1)**}\right), \tag{7}$$

$$\max\left(\Delta_{\mathrm{TV}}\left(\mathbb{P}_{Y|i}^{(i)*}|x, \mathbb{P}_{Y|i}^{(i+1)**}|x\right),\ \Delta_{\mathrm{TV}}\left(\mathbb{Q}_{Y|i}^{(i)*}|x, \mathbb{Q}_{Y|i}^{(i+1)**}|x\right)\right) \le \frac{\varepsilon}{4n}d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\,\Delta_{\mathrm{TV}}(\mathbb{R}_{i+1:n|i}|x)\,. \tag{8}$$

Inspired by these distributions, we define some extra artificial hybrid worlds:

The i* hybrid world is a truncated version of the i-th hybrid word. In this hybrid worlds, the algorithm sets $\tilde{\mathbb{R}}_{i+1:n|i}$ according to (7) and the rest of the computation is the same as the i-th hybrid word. In this hybrid world, the algorithm will also compute $\mathbb{R}_{1:n}^{(i)}$.

By its definition, the algorithm computes the exact ratio between

$$\mathbb{P}_1\mathbb{P}_{2|1}\ldots\mathbb{P}_{i|i-1}\mathbb{P}_{Y|i}^{(i)*} \text{ and } \mathbb{Q}_1\mathbb{Q}_{2|1}\ldots\mathbb{Q}_{i|i-1}\mathbb{Q}_{Y|i}^{(i)*}\,.$$

The $(i+1)$** hybrid world is a truncated version of the $(i + 1)$-th hybrid word. In this hybrid worlds, the algorithm sets $\mathbb{R}_{i+1:n|i}^{(i+1)}$ according to (7) and the rest of the computation is the same as the i-th hybrid word. In this hybrid world, the algorithm will also compute $\mathbb{R}_{1:n}^{(i+1)}$.

By its definition, the algorithm computes the exact ratio between

$$\mathbb{P}_1\mathbb{P}_{2|1}\ldots\mathbb{P}_{i|i-1}\mathbb{P}_{Y|i}^{(i+1)**} \text{ and } \mathbb{Q}_1\mathbb{Q}_{2|1}\ldots\mathbb{Q}_{i|i-1}\mathbb{Q}_{Y|i}^{(i+1)**}\,.$$

Therefore,

$$\Delta_{\mathrm{MTV}}(\mathbb{R}_{1:n}^{(i)}, \mathbb{R}_{1:n}^{(i+1)}) \le \max\Big(\Delta_{\mathrm{TV}}\Big(\mathbb{P}_1\mathbb{P}_{2|1}\ldots\mathbb{P}_{i|i-1}\mathbb{P}_{Y|i}^{(i)*},\ \mathbb{P}_1\mathbb{P}_{2|1}\ldots\mathbb{P}_{i|i-1}\mathbb{P}_{Y|i}^{(i+1)**}\Big),$$

$$\Delta_{\mathrm{TV}}\Big(\mathbb{Q}_1\mathbb{Q}_{2|1}\ldots\mathbb{Q}_{i|i-1}\mathbb{Q}_{Y|i}^{(i)*},\ \mathbb{Q}_1\mathbb{Q}_{2|1}\ldots\mathbb{Q}_{i|i-1}\mathbb{Q}_{Y|i}^{(i+1)**}\Big)\Big)\,.$$

The right-hand side of the inequality is bounded by

$$
\begin{aligned}
&\Delta_{\mathrm{TV}}\Big(\mathbb{P}_1\mathbb{P}_{2|1}\dots\mathbb{P}_{i|i-1}\mathbb{P}_{Y|i}^{(i)*},\ \mathbb{P}_1\mathbb{P}_{2|1}\dots\mathbb{P}_{i|i-1}\mathbb{P}_{Y|i}^{(i+1)**}\Big)\\
&=\Delta_{\mathrm{TV}}\Big(\mathbb{P}_i\mathbb{P}_{Y|i}^{(i)*},\ \mathbb{P}_i\mathbb{P}_{Y|i}^{(i+1)**}\Big)\\
&=\mathop{\mathbb{E}}_{x\sim\mathbb{P}_i}\Big[\Delta_{\mathrm{TV}}\Big(\mathbb{P}_{Y|i}^{(i)*}|x,\ \mathbb{P}_{Y|i}^{(i+1)**}|x\Big)\Big]\\
&\leq\mathop{\mathbb{E}}_{x\sim\mathbb{P}_i}\Big[\frac{\varepsilon}{4n}d_{\mathrm{LB}}+\frac{\varepsilon}{8n}\Delta_{\mathrm{TV}}(\mathbb{R}_{i+1:n|i}|x)\Big]\\
&=\frac{\varepsilon}{4n}d_{\mathrm{LB}}+\frac{\varepsilon}{8n}\mathop{\mathbb{E}}_{x\sim\mathbb{P}_i}\Big[\Delta_{\mathrm{TV}}(\mathbb{R}_{i+1:n|i}|x)\Big]\\
&=\frac{\varepsilon}{4n}d_{\mathrm{LB}}+\frac{\varepsilon}{8n}\mathop{\mathbb{E}}_{x\sim\mathbb{P}_i}\Big[\Delta_{\mathrm{TV}}(\mathbb{P}_{i+1:n|i}|x,\mathbb{Q}_{i+1:n|i}|x)\Big]\\
&=\frac{\varepsilon}{4n}d_{\mathrm{LB}}+\frac{\varepsilon}{8n}\Delta_{\mathrm{TV}}(\mathbb{P}_i\mathbb{P}_{i+1:n|i},\mathbb{P}_i\mathbb{Q}_{i+1:n|i})\\
&\leq\frac{\varepsilon}{2n}\Delta_{\mathrm{TV}}(\mathbb{P},\mathbb{Q})\,.
\end{aligned}
$$

The first inequality symbol follows from formula (8). The second inequality symbol relies on

$$
\begin{aligned}
\Delta_{\mathrm{TV}}(\mathbb{P}_i\mathbb{P}_{i+1:n|i},\mathbb{P}_i\mathbb{Q}_{i+1:n|i})&\leq\Delta_{\mathrm{TV}}(\mathbb{P}_i\mathbb{P}_{i+1:n|i},\mathbb{Q}_i\mathbb{Q}_{i+1:n|i})+\Delta_{\mathrm{TV}}(\mathbb{Q}_i\mathbb{Q}_{i+1:n|i},\mathbb{P}_i\mathbb{Q}_{i+1:n|i})\\
&=\Delta_{\mathrm{TV}}(\mathbb{P}_{i:n},\mathbb{Q}_{i:n})+\Delta_{\mathrm{TV}}(\mathbb{Q}_i,\mathbb{P}_i)\leq 2\,\Delta_{\mathrm{TV}}(\mathbb{P},\mathbb{Q})\,.\quad\square
\end{aligned}
$$

Acknowledgement

References

[BCH+20] Adam Bouland, Lijie Chen, Dhiraj Holden, Justin Thaler, and Prashant Nalini Vasude-van. On the power of statistical zero knowledge. *SIAM J. Comput.*, 49(4), 2020.

[BGM+23] Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, Aduri Pavan, and N. V. Vinodchandran. On approximating total variation distance. In *IJCAI*, 2023. (to appear, preprint version in *arXiv abs/2206.07209*).

[BGMV20] Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, and N. V. Vinodchandran. Efficient distance approximation for structured high-dimensional distributions via learning. In *NeurIPS*, 2020.

[CDKS18] Yu Cheng, Ilias Diakonikolas, Daniel Kane, and Alistair Stewart. Robust learning of fixed-structure bayesian networks. In *NeurIPS*, pages 10304–10316, 2018.

[CMR07] Corinna Cortes, Mehryar Mohri, and Ashish Rastogi. L_p distance and equivalence of probabilistic automata. *Int. J. Found. Comput. Sci.*, 18(4):761–779, 2007.

[FGJW23] Weiming Feng, Heng Guo, Mark Jerrum, and Jiaheng Wang. A simple polynomial-time approximation algorithm for the total variation distance between two product distributions. *TheoretiCS*, 2, 2023.

[GKM10] Parikshit Gopalan, Adam Klivans, and Raghu Meka. Polynomial-time approximation schemes for knapsack and related counting problems using branching programs. *arXiv preprint arXiv:1008.3187*, 2010.

[Kie18] Stefan Kiefer. On computing the total variation distance of hidden markov models. In *ICALP*, volume 107 of *LIPIcs*, pages 130:1–130:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

[LC12] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[LP02] Rune B. Lyngsø and Christian N. S. Pedersen. The consensus string problem and the complexity of comparing hidden markov models. *J. Comput. Syst. Sci.*, 65(3):545–569, 2002.

[MU17] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

[PW22] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Book draft, 2022.

[SV03] Amit Sahai and Salil P. Vadhan. A complete problem for statistical zero knowledge. *J. ACM*, 50(2):196–249, 2003.

[SVV12] Daniel Stefankovic, Santosh S. Vempala, and Eric Vigoda. A deterministic polynomial-time approximation scheme for counting knapsack solutions. *SIAM J. Comput.*, 41(2):356–366, 2012.

A Another Analysis of our Algorithm in the Markov Chain Setting

This section presents an alternative analysis for bounding the error of Algorithm 3. It starts by introducing a finer characterization of the "difference" between two ratios.

Definition 24 (Region). For any two ratios $\mathbb{R}_1, \mathbb{R}_2$, define the feasible region of TV distance pair (or "*region*", in short) between $\mathbb{R}_1, \mathbb{R}_2$, denoted by $\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}_1, \mathbb{R}_2)$, as

$$\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}_1, \mathbb{R}_2) := \left\{ \left(\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2), \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_2) \right) \middle| \begin{array}{l} \text{distributions } \mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2 \\ \text{such that } (\mathbb{P}_1 \| \mathbb{Q}_1) = \mathbb{R}_1, (\mathbb{P}_2 \| \mathbb{Q}_2) = \mathbb{R}_2 \end{array} \right\}.$$

Region has many nice properties. It is more expressive than the minimum total variation distance. The later can be defined as the L_1-distance between $(0,0)$ and the region.

$$\Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_2) = \inf_{(\delta_1, \delta_2) \in \mathcal{R}_{\mathrm{TVD}}(\mathbb{R}_1, \mathbb{R}_2)} \max(\delta_1, \delta_2).$$

We remark that, the current definition of Δ_{MTV} is not special. For example, consider an alternative distance

$$\Delta_{\mathrm{MTV}}'(\mathbb{R}_1, \mathbb{R}_2) = \inf_{(\delta_1, \delta_2) \in \mathcal{R}_{\mathrm{TVD}}(\mathbb{R}_1, \mathbb{R}_2)} (\delta_1 + \delta_2).$$

This alternative distance, which is the L_∞-distance from $(0,0)$, also works well with the paper. An early draft of this paper uses Δ_{MTV}' to analyze our algorithms.

For this paper, it is sufficient to have the following property. The proof is deferred to Appendix B.

Lemma 25 (Triangle Inequality). *For any ratios $\mathbb{R}_1, \mathbb{R}_2, \mathbb{R}$ such that $\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}_1, \mathbb{R}) \leq_\exists (\alpha_1, \beta_1)$ and $\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}_2, \mathbb{R}) \leq_\exists (\alpha_2, \beta_2)$, it holds that $\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}_1, \mathbb{R}_2) \leq_\exists (\alpha_1 + \alpha_2, \beta_1 + \beta_2)$.*

The description of this lemma uses a convenient notation $\leq_\exists$, defined as follows. For any two points $(v_1, v_2), (w_1, w_2)$ in the Euclidean plane, we say $(v_1, v_2) \leq (w_1, w_2)$ if $v_1 \leq w_1$ and $v_2 \leq w_2$. Apparently, this defines an order relation. Let $\mathcal{R}$ be a point set in the Euclidean plane, we say $\mathcal{R} \leq_\exists (v_1, v_2)$ if there exists $(u_1, u_2) \in \mathcal{R}$ such that $(u_1, u_2) \leq (v_1, v_2)$.

Using the new concept of region, we gives a fine characterization of how the error passes through the concatenation process.

Lemma 26. *Given conditional ratios $\mathbb{R}_{Y|X}, \mathbb{R}'_{Y|X}$ over Ω_X, for any mappings $\alpha, \beta : \Omega_X \to [0, \infty)$ such that $\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}_{Y|X}|x, \mathbb{R}'_{Y|X}|x) \leq_\exists (\alpha(x), \beta(x))$ for every $x \in \Omega_X$, for any distributions $\mathbb{P}_X, \mathbb{Q}_X$ over Ω_X, it holds that*

$$\mathcal{R}_{\mathrm{TVD}}(\mathsf{Concatenate}(\mathbb{P}_X, \mathbb{Q}_X, \mathbb{R}_{Y|X}), \mathsf{Concatenate}(\mathbb{P}_X, \mathbb{Q}_X, \mathbb{R}'_{Y|X})) \leq_\exists \left(\mathop{\mathbb{E}}_{X \sim \mathbb{P}_X}[\alpha(X)], \mathop{\mathbb{E}}_{X \sim \mathbb{Q}_X}[\beta(X)] \right).$$

Proof. Since $\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}_{Y|X}|x, \mathbb{R}'_{Y|X}|x) \leq_\exists (\alpha(x), \beta(x))$ for every $x \in \Omega_X$, there exist a sample space Ω_Y, Markov kernels $\mathbb{P}_{Y|X}, \mathbb{P}'_{Y|X}, \mathbb{Q}_{Y|X}, \mathbb{Q}'_{Y|X}$ from Ω_X to Ω_Y such that

$$\mathbb{R}_{Y|X} = (\mathbb{P}_{Y|X} \| \mathbb{Q}_{Y|X}), \qquad \mathbb{R}'_{Y|X} = (\mathbb{P}'_{Y|X} \| \mathbb{Q}'_{Y|X}),$$
$$\Delta_{\mathrm{TV}}(\mathbb{P}_{Y|X}|x, \mathbb{P}'_{Y|X}|x) \leq \alpha(x), \quad \Delta_{\mathrm{TV}}(\mathbb{Q}_{Y|X}|x, \mathbb{Q}'_{Y|X}|x) \leq \beta(x) \quad \text{for every } x \in \Omega_X.$$

The two ratios in the wanted inequality equal $(\mathbb{P}_X \mathbb{P}_{Y|X} \| \mathbb{Q}_X \mathbb{Q}_{Y|X})$ and $(\mathbb{P}_X \mathbb{P}'_{Y|X} \| \mathbb{Q}_X \mathbb{Q}'_{Y|X})$ respectively. Therefore[4]

$$\mathcal{R}_{\mathrm{TVD}}(\mathsf{Concatenate}(\mathbb{P}_X, \mathbb{Q}_X, \mathbb{R}_{Y|X}), \mathsf{Concatenate}(\mathbb{P}_X, \mathbb{Q}_X, \mathbb{R}'_{Y|X}))$$
$$\leq_\exists \left(\Delta_{\mathrm{TV}}(\mathbb{P}_X \mathbb{P}_{Y|X}, \mathbb{P}_X \mathbb{P}'_{Y|X}), \ \Delta_{\mathrm{TV}}(\mathbb{Q}_X \mathbb{Q}_{Y|X}, \mathbb{Q}_X \mathbb{Q}'_{Y|X}) \right)$$
$$= \left(\mathop{\mathbb{E}}_{x \sim \mathbb{P}_X}\left[\Delta_{\mathrm{TV}}(\mathbb{P}_{Y|X}|x, \mathbb{P}'_{Y|X}|x) \right], \mathop{\mathbb{E}}_{x \sim \mathbb{Q}_X}\left[\Delta_{\mathrm{TV}}(\mathbb{Q}_{Y|X}|x, \mathbb{Q}'_{Y|X}|x) \right] \right)$$
$$\leq \left(\mathop{\mathbb{E}}_{X \sim \mathbb{P}_X}[\alpha(X)], \mathop{\mathbb{E}}_{X \sim \mathbb{Q}_X}[\beta(X)] \right). \qquad \square$$

Now we are ready to present an alternative error analysis of Algorithm 3.

Claim 27. For every $1 < k \leq n$, the intermediate conditional ratio $\mathbb{R}'_{k:n|k-1}$ satisfies

$$\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}'_{k:n|k-1}|x, \mathbb{R}_{k:n|k-1}|x) \leq_\exists \left(\frac{n-k}{4n} \cdot \varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n} \sum_{t=k}^{n-1} \mathop{\mathbb{E}}_{z \sim \mathbb{P}_{t|k-1}|x}\left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z) \right], \right.$$
$$\left. \frac{n-k}{4n} \cdot \varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n} \sum_{t=k}^{n-1} \mathop{\mathbb{E}}_{z \sim \mathbb{Q}_{t|k-1}|x}\left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z) \right] \right),$$

for all $x \in [q]$. The ratio $\mathbb{R}'_{1:n}$ satisfies $\Delta_{\mathrm{MTV}}(\mathbb{R}'_{1:n}, \mathbb{R}_{1:n}) \leq \frac{n-1}{2n} \varepsilon \, \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$.

---

[4]The expectation in the 3rd line of the following formula is written using an inconsistent but intuitive notation $x \sim \mathbb{P}_X$. If using the same notation as the rest of the paper, it should be written as $\mathbb{E}_{X \sim \mathbb{P}_X}\left[\Delta_{\mathrm{TV}}(\mathbb{P}_{Y|X}|X, \mathbb{P}'_{Y|X}|X) \right]$.

Proof. There is an inductive proof from larger k to smaller k. The base case is when $k = n$. The base case follows trivially from the fact that $\mathbb{R}'_{n:n|n-1} = \mathbb{R}_{n:n|n-1}$.

For every smaller k, by Lemma 14, the error introduced by sparsifying $\mathbb{R}'_{k+1:n|k}|x$ is bounded by

$$\Delta_{\mathrm{MTV}}(\tilde{\mathbb{R}}_{k+1:n|k}|x, \mathbb{R}'_{k+1:n|k}|x) \le \frac{\varepsilon}{4n}d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\Delta_{\mathrm{TV}}(\mathbb{R}'_{k+1:n|k}|x) \le \frac{\varepsilon}{4n}d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\Delta_{\mathrm{TV}}(\mathbb{R}_{k+1:n|k}|x)\,,$$

which means

$$\mathcal{R}_{\mathrm{TVD}}(\tilde{\mathbb{R}}_{k+1:n|k}|x, \mathbb{R}'_{k+1:n|k}|x) \le_\exists \left(\frac{\varepsilon}{4n}d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\Delta_{\mathrm{TV}}(\mathbb{R}_{k+1:n|k}|x), \frac{\varepsilon}{4n}d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\Delta_{\mathrm{TV}}(\mathbb{R}_{k+1:n|k}|x)\right).$$

Combining with the inductive hypothesis that bounds $\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}'_{k+1:n|k}|x, \mathbb{R}_{k+1:n|k}|x)$ by

$$\mathcal{R}_{\mathrm{TVD}}(\tilde{\mathbb{R}}_{k+1:n|k}|x, \mathbb{R}_{k+1:n|k}|x) \le_\exists \left(\frac{n-k-1}{4n}\cdot\varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\sum_{t=k+1}^{n-1}\mathop{\mathbb{E}}_{z\sim\mathbb{P}_{t|k}|x}\left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z)\right],\right.$$

$$\left.\frac{n-k-1}{4n}\cdot\varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\sum_{t=k+1}^{n-1}\mathop{\mathbb{E}}_{z\sim\mathbb{Q}_{t|k}|x}\left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z)\right]\right),$$

using the triangle inequality (Lemma 25)

$$\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}'_{k+1:n|k}|x, \mathbb{R}_{k+1:n|k}|x) \le_\exists \left(\frac{n-k}{4n}\cdot\varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\sum_{t=k}^{n-1}\mathop{\mathbb{E}}_{z\sim\mathbb{P}_{t|k}|x}\left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z)\right],\right.$$

$$\left.\frac{n-k}{4n}\cdot\varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\sum_{t=k}^{n-1}\mathop{\mathbb{E}}_{z\sim\mathbb{Q}_{t|k}|x}\left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z)\right]\right). \quad (9)$$

Note that, in the above formula, $\mathbb{P}_{k|k}|x$ is the degenerated distribution that only samples x. Then the inductive step is concluded by Lemma 26

$$\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}'_{k:n|k-1}|x, \mathbb{R}_{k:n|k-1}|x) \le_\exists \left(\mathop{\mathbb{E}}_{y\sim\mathbb{P}_{k|k-1}|x}\left[\frac{n-k}{4n}\cdot\varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\sum_{t=k}^{n-1}\mathop{\mathbb{E}}_{z\sim\mathbb{P}_{t|k}|y}\left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z)\right]\right],\right.$$

$$\left.\mathop{\mathbb{E}}_{y\sim\mathbb{Q}_{k|k-1}|x}\left[\frac{n-k}{4n}\cdot\varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\sum_{t=k}^{n-1}\mathop{\mathbb{E}}_{z\sim\mathbb{Q}_{t|k}|y}\left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z)\right]\right]\right)$$

$$= \left(\frac{n-k}{4n}\cdot\varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\sum_{t=k}^{n-1}\mathop{\mathbb{E}}_{y\sim\mathbb{P}_{k|k-1}|x}\left[\mathop{\mathbb{E}}_{z\sim\mathbb{P}_{t|k}|y}\left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z)\right]\right],\right.$$

$$\left.\frac{n-k}{4n}\cdot\varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\sum_{t=k}^{n-1}\mathop{\mathbb{E}}_{y\sim\mathbb{Q}_{k|k-1}|x}\left[\mathop{\mathbb{E}}_{z\sim\mathbb{Q}_{t|k}|y}\left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z)\right]\right]\right)$$

$$= \left(\frac{n-k}{4n}\cdot\varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\sum_{t=k}^{n-1}\mathop{\mathbb{E}}_{z\sim\mathbb{P}_{t|k-1}|x}\left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z)\right],\right.$$

$$\left.\frac{n-k}{4n}\cdot\varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n}\sum_{t=k}^{n-1}\mathop{\mathbb{E}}_{z\sim\mathbb{Q}_{t|k-1}|x}\left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z)\right]\right). \quad (10)$$

In the special case when $k = 1$, by the same analysis as the first half of the inductive step, formula (9) holds. Then by an argument mostly similar[5] to (10),

$$\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}'_{1:n}, \mathbb{R}_{1:n}) \leq_\exists \left(\frac{n-1}{4n} \cdot \varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n} \sum_{t=1}^{n-1} \underset{z \sim \mathbb{P}_t}{\mathbb{E}} \left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z) \right], \right.$$
$$\left. \frac{n-1}{4n} \cdot \varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n} \sum_{t=1}^{n-1} \underset{z \sim \mathbb{Q}_t}{\mathbb{E}} \left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z) \right] \right).$$

For each $1 \leq t < n$, we have

$$\underset{z \sim \mathbb{Q}_t}{\mathbb{E}} \left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z) \right] = \underset{z \sim \mathbb{Q}_t}{\mathbb{E}} \left[\Delta_{\mathrm{TV}}(\mathbb{P}_{t+1:n|t}|z, \mathbb{Q}_{t+1:n|t}|z) \right]$$
$$= \Delta_{\mathrm{TV}}(\mathbb{Q}_t \mathbb{P}_{t+1:n|t}, \mathbb{Q}_t \mathbb{Q}_{t+1:n|t}) \leq \Delta_{\mathrm{TV}}(\mathbb{Q}_t \mathbb{P}_{t+1:n|t}, \mathbb{P}_t \mathbb{P}_{t+1:n|t}) + \Delta_{\mathrm{TV}}(\mathbb{P}_t \mathbb{P}_{t+1:n|t}, \mathbb{Q}_t \mathbb{Q}_{t+1:n|t})$$
$$= \Delta_{\mathrm{TV}}(\mathbb{Q}_t, \mathbb{P}_t) + \Delta_{\mathrm{TV}}(\mathbb{P}_{t:n}, \mathbb{Q}_{t:n}) \leq 2 \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}).$$

Symmetrically, $\mathbb{E}_{z \sim \mathbb{P}_t} \left[\Delta_{\mathrm{TV}}(\mathbb{R}_{t+1:n|t}|z) \right] \leq 2 \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$. Therefore,

$$\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}'_{1:n}, \mathbb{R}_{1:n}) \leq_\exists \left(\frac{n-1}{4n} \cdot \varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n}(n-1) \cdot 2 \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}), \text{ the same} \right).$$
$$\Delta_{\mathrm{MTV}}(\mathbb{R}'_{1:n}, \mathbb{R}_{1:n}) \leq \frac{n-1}{4n} \cdot \varepsilon d_{\mathrm{LB}} + \frac{\varepsilon}{8n}(n-1) \cdot 2 \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$$
$$\leq \frac{n-1}{2n} \cdot \varepsilon \Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}). \qquad \square$$

B Deferred Proofs

Proof of Lemma 9. The key step of the proof is to enforce the search space of $(\mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2)$ to be finite-dimensional. Concretely, we construct a *finite* set Ω such that we can assume w.l.o.g. that Ω is the sample space of distributions $\mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2$. That is,

$$\Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_2) = \inf_{\substack{discrete\ \mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2 \\ in\ sample\ space\ \Omega \\ (\mathbb{P}_1 \| \mathbb{Q}_1) = \mathbb{R}_1 \\ (\mathbb{P}_2 \| \mathbb{Q}_2) = \mathbb{R}_2}} \max\left(\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2), \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_2) \right). \tag{11}$$

Assume that we have found such a *finite* set Ω satisfying equation (11), then the infimum symbol can be replaced by minimum, because the search space is a compact set and the TV distance is a continuous function. Thus to finish the proof, it suffices to (explicitly) construct Ω and show (11).

The sample space Ω is defined as

$$\Omega = \left(\mathrm{Supp}(\mathbb{R}_1) \cup \{\infty\} \right) \times \left(\mathrm{Supp}(\mathbb{R}_2) \cup \{\infty\} \right).$$

The desired (11) is implied by the following statement: for any discrete distributions $\mathbb{P}'_1, \mathbb{Q}'_1, \mathbb{P}'_2, \mathbb{Q}'_2$ over any sample space Ω' such that $(\mathbb{P}'_1 \| \mathbb{Q}'_1) = \mathbb{R}_1$ and $(\mathbb{P}'_2 \| \mathbb{Q}'_2) = \mathbb{R}_2$, there exists distributions $\mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2$ over Ω such that $(\mathbb{P}_1 \| \mathbb{Q}_1) = \mathbb{R}_1$, $(\mathbb{P}_2 \| \mathbb{Q}_2) = \mathbb{R}_2$ and

$$\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2) \leq \Delta_{\mathrm{TV}}(\mathbb{P}'_1, \mathbb{P}'_2) \quad \text{and} \quad \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_2) \leq \Delta_{\mathrm{TV}}(\mathbb{Q}'_1, \mathbb{Q}'_2). \tag{12}$$

---

[5]The argument will be exactly the same, if we define degenerated distributions $\mathbb{P}_0, \mathbb{Q}_0$ over a size-1 sample space and define Markov kernels $\mathbb{P}_{1|0}, \mathbb{Q}_{1|0}$ such that as $\mathbb{P}_{1|0}\mathbb{P}_0 = \mathbb{P}_1$ and $\mathbb{Q}_{1|0}\mathbb{Q}_0 = \mathbb{Q}_1$.

The distributions $\mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2$ are constructed as follows. Define mapping $\pi : \Omega' \to \Omega$ as

$$\pi(z) = \Big(\frac{\mathbb{P}_1'(z)}{\mathbb{Q}_1'(z)}, \frac{\mathbb{P}_2'(z)}{\mathbb{Q}_2'(z)} \Big).$$

Let $\mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2$ be the distributions of $\pi(X_1), \pi(Y_1), \pi(X_2), \pi(Y_2)$ where $X_1 \sim \mathbb{P}_1', Y_1 \sim \mathbb{Q}_1'$, $X_2 \sim \mathbb{P}_2', Y_2 \sim \mathbb{Q}_2'$. By the data-processing principle, (12) holds. Now the only remaining task is to show $(\mathbb{P}_1 \| \mathbb{Q}_1) = \mathbb{R}_1$, $(\mathbb{P}_2 \| \mathbb{Q}_2) = \mathbb{R}_2$. By symmetry, it suffices to prove one of them.

For each $(r_1, r_2) \in \Omega$ such that $\mathbb{Q}_1(r_1, r_2) > 0$ (which implies $r_1 \neq \infty$),

$$\frac{\mathbb{P}_1(r_1, r_2)}{\mathbb{Q}_1(r_1, r_2)} = \frac{\sum_{z \in \Omega' \text{ s.t. } \pi(z) = (r_1, r_2)} \mathbb{P}_1'(z)}{\sum_{z \in \Omega' \text{ s.t. } \pi(z) = (r_1, r_2)} \mathbb{Q}_1'(z)} = \frac{\sum_{z \in \Omega' \text{ s.t. } \pi(z) = (r_1, r_2)} r_1 \cdot \mathbb{Q}_1'(z)}{\sum_{z \in \Omega' \text{ s.t. } \pi(z) = (r_1, r_2)} \mathbb{Q}_1'(z)} = r_1.$$

For each $r \in [0, \infty)$

$$(\mathbb{P}_1 \| \mathbb{Q}_1)(r) = \Pr_{Y \leftarrow \mathbb{Q}_1} \Big[\frac{\mathbb{P}_1(Y)}{\mathbb{Q}_1(Y)} = r \Big] = \Pr_{Y \leftarrow \mathbb{Q}_1} \Big[\text{first entry of } Y = r \Big]$$

$$= \Pr_{Y' \leftarrow \mathbb{Q}_1'} \Big[\text{first entry of } \pi(Y') = r \Big] = \Pr_{Y' \leftarrow \mathbb{Q}_1'} \Big[\frac{\mathbb{P}_1'(Y')}{\mathbb{Q}_1'(Y')} = r \Big] = (\mathbb{P}_1' \| \mathbb{Q}_1')(r).$$

Thus $(\mathbb{P}_1 \| \mathbb{Q}_1) = (\mathbb{P}_1' \| \mathbb{Q}_1') = \mathbb{R}_1$. $\qquad\square$

Proof of Lemma 10. It is easy to verify that $\Delta_{\mathrm{MTV}}(\mathbb{R}, \mathbb{R}) = 0$ and $\Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_2) = \Delta_{\mathrm{MTV}}(\mathbb{R}_2, \mathbb{R}_1)$.

Positivity: Next, we verify the positivity property: $\Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_2) > 0$ if $\mathbb{R}_1 \neq \mathbb{R}_2$. If $\mathbb{R}_1, \mathbb{R}_2$ have finite-size supports, the positivity property is very easy to prove. Here we briefly present a proof without assuming finite-size supports.

The proof requires some basic knowledge of decision theory. In the decision problem between two distribution $\mathbb{P}$ and $\mathbb{Q}$, a distinguisher is given a sample x that is sampled from either $\mathbb{P}$ or $\mathbb{Q}$, and is supposed to guess which distribution x is sampled from. The distinguisher can be formalized as a randomized algorithm from the sample space to $\{0, 1\}$. It is well-known that TV distance can be defined as

$$\Delta_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \max_D \Big| \Pr_{X \sim \mathbb{P}} \big[D(X) \to 0 \big] - \Pr_{X \sim \mathbb{Q}} \big[D(X) \to 0 \big] \Big|.$$

The *Neyman-Pearson region* of the decision problem, denoted by $\mathcal{R}_{\mathrm{NP}}(\mathbb{P}, \mathbb{Q})$, is defined as

$$\mathcal{R}_{\mathrm{NP}}(\mathbb{P}, \mathbb{Q}) := \Big\{ \big(\Pr_{X \sim \mathbb{P}}[D(X) \to 0], \Pr_{X \sim \mathbb{Q}}[D(X) \to 0] \big) \;\Big|\; D \Big\} \subseteq [0, 1] \times [0, 1].$$

Every Neyman-Pearson region satisfies the following properties.

- The region is close and convex.

- The region contains $(0, 0)$ and $(1, 1)$.

- The region is centrally symmetric: $(\alpha, \beta) \in \mathcal{R}_{\mathrm{NP}}(\mathbb{P}, \mathbb{Q})$ if and only if $(1 - \alpha, 1 - \beta) \in \mathcal{R}_{\mathrm{NP}}(\mathbb{P}, \mathbb{Q})$.

- The region is determined by the ratio, and vice verse. Thus we can define the Neyman-Pearson region of ratio $\mathbb{R}$, denoted by $\mathcal{R}_{\mathrm{NP}}(\mathbb{R})$, so that $\mathcal{R}_{\mathrm{NP}}(\mathbb{R}) = \mathcal{R}_{\mathrm{NP}}(\mathbb{P}, \mathbb{Q})$ as long as $\mathbb{R} = (\mathbb{P} \| \mathbb{Q})$.

Given any two distinct ratios $\mathbb{R}_1 \neq \mathbb{R}_2$, we have $\mathcal{R}_{\mathrm{NP}}(\mathbb{R}_1) \neq \mathcal{R}_{\mathrm{NP}}(\mathbb{R}_2)$. Assume w.l.o.g. that $(\alpha, \beta) \in \mathcal{R}_{\mathrm{NP}}(\mathbb{R}_1) \setminus \mathcal{R}_{\mathrm{NP}}(\mathbb{R}_2)$. Since $\mathcal{R}_{\mathrm{NP}}(\mathbb{R}_2)$ is close, there exists a constant $\delta > 0$ such that $\max(|\alpha - \alpha'|, |\beta - \beta'|) \geq \delta$ for any $(\alpha', \beta') \in \mathcal{R}_{\mathrm{NP}}(\mathbb{R}_2)$.

For any $\mathbb{P}_1, \mathbb{Q}_1$ satisfying $(\mathbb{P}_1 \| \mathbb{Q}_1) = \mathbb{R}_1$, since $(\alpha, \beta) \in \mathcal{R}_{\mathrm{NP}}(\mathbb{P}_1, \mathbb{Q}_1)$, there exists a distinguisher D such that

$$\Pr_{X \sim \mathbb{P}_1}\big[D(X) \to 0\big] = \alpha, \qquad \Pr_{X \sim \mathbb{Q}_2}\big[D(X) \to 0\big] = \beta.$$

For any $\mathbb{P}_2, \mathbb{Q}_2$ satisfying $(\mathbb{P}_2 \| \mathbb{Q}_2) = \mathbb{R}_2$, let

$$(\alpha', \beta') := \Big(\Pr_{X \sim \mathbb{P}_2}\big[D(X) \to 0\big], \Pr_{X \sim \mathbb{Q}_2}\big[D(X) \to 0\big]\Big) \in \mathcal{R}_{\mathrm{NP}}(\mathbb{R}_2).$$

Then

$$\begin{aligned}
&\max(\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2), \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_2)) \\
&\geq \max\Big(\Big|\Pr_{X \sim \mathbb{P}_1}\big[D(X) \to 0\big] - \Pr_{X \sim \mathbb{P}_2}\big[D(X) \to 0\big]\Big|, \Big|\Pr_{X \sim \mathbb{Q}_1}\big[D(X) \to 0\big] - \Pr_{X \sim \mathbb{Q}_2}\big[D(X) \to 0\big]\Big|\Big) \\
&= \max(|\alpha - \alpha'|, |\beta - \beta'|) \geq \delta.
\end{aligned}$$

Therefore $\Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_2) \geq \delta > 0$.

Triangle inequality: We will prove that for any ratios $\mathbb{R}, \mathbb{R}_1, \mathbb{R}_2$,

$$\Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_2) \leq \Delta_{\mathrm{MTV}}(\mathbb{R}, \mathbb{R}_1) + \Delta_{\mathrm{MTV}}(\mathbb{R}, \mathbb{R}_2).$$

By the definition of MTV distance, there exist[6] distributions $\mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_3, \mathbb{Q}_3$ over sample space Ω_1 such that

$$(\mathbb{P}_1 \| \mathbb{Q}_1) = \mathbb{R}_1, \quad (\mathbb{P}_3 \| \mathbb{Q}_3) = \mathbb{R}, \quad \max(\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_3), \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_3)) = \Delta_{\mathrm{MTV}}(\mathbb{R}, \mathbb{R}_1).$$

Similarly, there exist distributions $\mathbb{P}_2, \mathbb{Q}_2, \mathbb{P}_4, \mathbb{Q}_4$ over sample space Ω_2 such that

$$(\mathbb{P}_2 \| \mathbb{Q}_2) = \mathbb{R}_2, \quad (\mathbb{P}_4 \| \mathbb{Q}_4) = \mathbb{R}, \quad \max(\Delta_{\mathrm{TV}}(\mathbb{P}_2, \mathbb{P}_4), \Delta_{\mathrm{TV}}(\mathbb{Q}_2, \mathbb{Q}_4)) = \Delta_{\mathrm{MTV}}(\mathbb{R}, \mathbb{R}_2).$$

We are going to construct distributions $\mathbb{P}', \mathbb{Q}', \mathbb{P}'_1, \mathbb{Q}'_1, \mathbb{P}'_2, \mathbb{Q}'_2$ such that

$$\begin{aligned}
\mathbb{R}_1 = (\mathbb{P}'_1 \| \mathbb{Q}'_1), \quad \mathbb{R}_2 &= (\mathbb{P}'_2 \| \mathbb{Q}'_2), \quad \mathbb{R} = (\mathbb{P}' \| \mathbb{Q}'), \\
\Delta_{\mathrm{TV}}(\mathbb{P}'_1, \mathbb{P}') = \Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_3), \quad &\Delta_{\mathrm{TV}}(\mathbb{Q}'_1, \mathbb{Q}') = \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_3), \\
\Delta_{\mathrm{TV}}(\mathbb{P}'_2, \mathbb{P}') = \Delta_{\mathrm{TV}}(\mathbb{P}_2, \mathbb{P}_4), \quad &\Delta_{\mathrm{TV}}(\mathbb{Q}'_2, \mathbb{Q}') = \Delta_{\mathrm{TV}}(\mathbb{Q}_2, \mathbb{Q}_4).
\end{aligned} \tag{13}$$

Then the proof will be concluded by

$$\begin{aligned}
\Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_2) &\leq \max(\Delta_{\mathrm{TV}}(\mathbb{P}'_1, \mathbb{P}'_2), \Delta_{\mathrm{TV}}(\mathbb{Q}'_1, \mathbb{Q}'_2)) \\
&\leq \max(\Delta_{\mathrm{TV}}(\mathbb{P}'_1, \mathbb{P}') + \Delta_{\mathrm{TV}}(\mathbb{P}'_2, \mathbb{P}'), \ \Delta_{\mathrm{TV}}(\mathbb{Q}'_1, \mathbb{Q}') + \Delta_{\mathrm{TV}}(\mathbb{Q}'_2, \mathbb{Q}')) \\
&\leq \max(\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_3), \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_3)) + \max(\Delta_{\mathrm{TV}}(\mathbb{P}_2, \mathbb{P}_4), \Delta_{\mathrm{TV}}(\mathbb{Q}_2, \mathbb{Q}_4)) \\
&= \Delta_{\mathrm{MTV}}(\mathbb{R}, \mathbb{R}_1) + \Delta_{\mathrm{MTV}}(\mathbb{R}, \mathbb{R}_2).
\end{aligned}$$

---

[6] For this paper, we can assume $\mathbb{R}_1, \mathbb{R}$ have finite-size supports, then such distributions $\mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_3, \mathbb{Q}_3$ are guaranteed to exist (Lemma 9). Otherwise, the proof need a small modification.

Intuitively, the distribution $\mathbb{Q}'$ samples a pair of dependent random variables $(X, Y) \sim \mathbb{Q}'$, such that $X \sim \mathbb{Q}_3$ and the distribution Y conditional on $X = x$ is the same as the distribution of Y conditional on $\frac{\mathbb{P}_3(x)}{\mathbb{Q}_3(x)} = \frac{\mathbb{P}_4(Y)}{\mathbb{Q}_4(Y)}$. To formalize the intuition, we define a joint distribution $\mathbb{Q}_{3,4}$.

$$\mathbb{Q}_{3,4}(x, y) = \begin{cases} \mathbb{Q}_3(x) \dfrac{\mathbb{Q}_4(y)}{\mathbb{Q}_4(\Omega_{2,r})}, & \text{if } \exists r \text{ s.t. } x \in \Omega_{1,r} \wedge y \in \Omega_{2,r} \wedge \mathbb{Q}_4(\Omega_{2,r}) > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\Omega_{1,r}, \Omega_{2,r}$ are defined as

$$\Omega_{1,r} := \left\{ x \in \Omega_1 \;\middle|\; \frac{\mathbb{P}_3(x)}{\mathbb{Q}_3(x)} = r \right\}, \qquad \Omega_{2,r} := \left\{ y \in \Omega_2 \;\middle|\; \frac{\mathbb{P}_4(y)}{\mathbb{Q}_4(y)} = r \right\},$$

$$\Omega_{1,\infty} := \left\{ x \in \Omega_1 \;\middle|\; \mathbb{Q}_3(x) = 0 \right\}, \qquad \Omega_{2,\infty} := \left\{ y \in \Omega_2 \;\middle|\; \mathbb{Q}_4(y) = 0 \right\}.$$

Note that the probability mass function of $\mathbb{Q}_{3,4}$ equals

$$\mathbb{Q}_3(x) \frac{\mathbb{Q}_4(y)}{\mathbb{Q}_4(\Omega_{2,r})} = \mathbb{R}(r) \frac{\mathbb{Q}_3(x)}{\mathbb{Q}_3(\Omega_{1,r})} \frac{\mathbb{Q}_4(y)}{\mathbb{Q}_4(\Omega_{2,r})} = \mathbb{Q}_4(y) \frac{\mathbb{Q}_3(x)}{\mathbb{Q}_3(\Omega_{1,r})}.$$

Thus there are two other equivalent interpretations:

- First sample $Y \sim \mathbb{Q}_4$. Conditioning on $Y = y$, the distribution of X is the same as the distribution of X conditional on $\frac{\mathbb{P}_3(X)}{\mathbb{Q}_3(X)} = \frac{\mathbb{P}_4(y)}{\mathbb{Q}_4(y)}$.

- First sample $R \sim \mathbb{R}$. Conditioning on $R = r$, X and Y are independent, and the conditional distribution of X (resp. Y) is the same as the distribution of $X \sim \mathbb{Q}_3$ (resp. $Y \sim \mathbb{Q}_4$) conditional on $\frac{\mathbb{P}_3(X)}{\mathbb{Q}_3(X)} = r$ (resp. $\frac{\mathbb{P}_4(Y)}{\mathbb{Q}_4(Y)} = r$).

Similarly, we can define a joint distribution $\mathbb{P}_{3,4}$ as

$$\mathbb{P}_{3,4}(x, y) = \begin{cases} \mathbb{P}_3(x) \dfrac{\mathbb{P}_4(y)}{\mathbb{P}_4(\Omega_{2,r})}, & \text{if } \exists r \text{ s.t. } x \in \Omega_{1,r} \wedge y \in \Omega_{2,r} \wedge \mathbb{P}_4(\Omega_{2,r}) > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Note that, if letting $(\mathbb{R}^c, \mathbb{R})$ be the canonical pair of $\mathbb{R}$, the probability mass function of $\mathbb{P}_{3,4}$ equals

$$\mathbb{P}_3(x) \frac{\mathbb{P}_4(y)}{\mathbb{P}_4(\Omega_{2,r})} = \mathbb{R}^c(r) \frac{\mathbb{P}_3(x)}{\mathbb{P}_3(\Omega_{1,r})} \frac{\mathbb{P}_4(y)}{\mathbb{P}_4(\Omega_{2,r})} = \mathbb{P}_4(y) \frac{\mathbb{P}_3(x)}{\mathbb{P}_3(\Omega_{1,r})}.$$

We can define the conditional distributions $\mathbb{P}_{3|4}, \mathbb{P}_{4|3}, \mathbb{Q}_{3|4}, \mathbb{Q}_{4|3}$ so that $\mathbb{P}_{3,4} = \mathbb{P}_3 \mathbb{P}_{4|3} = \mathbb{P}_4 \mathbb{P}_{3|4}$ and $\mathbb{Q}_{3,4} = \mathbb{Q}_3 \mathbb{Q}_{4|3} = \mathbb{Q}_4 \mathbb{Q}_{3|4}$. That is, for any $(x, y) \in \Omega_1 \times \Omega_2$

$$\mathbb{P}_{3,4}(x, y) = \mathbb{P}_3(x) \cdot \mathbb{P}_{4|3}(y|x) = \mathbb{P}_4(y) \cdot \mathbb{P}_{3|4}(x|y),$$
$$\mathbb{Q}_{3,4}(x, y) = \mathbb{Q}_3(x) \cdot \mathbb{Q}_{4|3}(y|x) = \mathbb{Q}_4(y) \cdot \mathbb{Q}_{3|4}(x|y).$$

Moreover, for any $x \in \text{Supp}(\mathbb{P}_3) \cap \text{Supp}(\mathbb{Q}_3)$, we have $x \in \Omega_{1,r}$ for some $r \in (0, \infty)$, then

$$\mathbb{P}_{4|3}(y|x) = \begin{cases} \dfrac{\mathbb{P}_4(y)}{\mathbb{P}_4(\Omega_{2,r})}, & \text{if } y \in \Omega_{2,r} \\ 0, & \text{otherwise} \end{cases} = \begin{cases} \dfrac{\mathbb{Q}_4(y)}{\mathbb{Q}_4(\Omega_{2,r})}, & \text{if } y \in \Omega_{2,r} \\ 0, & \text{otherwise} \end{cases} = \mathbb{Q}_{4|3}(y|x).$$

Thus we can define them so that $\mathbb{P}_{4|3} = \mathbb{Q}_{4|3}$. Similarly, we can let $\mathbb{P}_{3|4} = \mathbb{Q}_{3|4}$.

Define distributions $\mathbb{P}', \mathbb{Q}', \mathbb{P}'_1, \mathbb{Q}'_1, \mathbb{P}'_2, \mathbb{Q}'_2$ over $\Omega = \Omega_1 \times \Omega_2$ as

$$\mathbb{P}' = \mathbb{P}_{3,4}\,, \quad \mathbb{P}'_1 = \mathbb{P}_1 \mathbb{P}_{4|3}\,, \quad \mathbb{P}'_2 = \mathbb{P}_2 \mathbb{P}_{3|4}\,,$$
$$\mathbb{Q}' = \mathbb{Q}_{3,4}\,, \quad \mathbb{Q}'_1 = \mathbb{Q}_1 \mathbb{Q}_{4|3}\,, \quad \mathbb{Q}'_2 = \mathbb{Q}_2 \mathbb{Q}_{3|4}\,.$$

We conclude the proof by verifying the properties in (13) holds.

- Let $(X, Y) \sim \mathbb{Q}'_1$, then $\frac{\mathbb{P}'_1(X,Y)}{\mathbb{Q}'_1(X,Y)} \sim (\mathbb{P}'_1 \| \mathbb{Q}'_1)$. In the meanwhile,

$$\frac{\mathbb{P}'_1(X, Y)}{\mathbb{Q}'_1(X, Y)} = \frac{(\mathbb{P}_1 \mathbb{Q}_{4|3})(X, Y)}{(\mathbb{Q}_1 \mathbb{Q}_{4|3})(X, Y)} = \frac{\mathbb{P}_1(X) \mathbb{Q}_{4|3}(Y|X)}{\mathbb{Q}_1(X) \mathbb{Q}_{4|3}(Y|X)} = \frac{\mathbb{P}_1(X)}{\mathbb{Q}_1(X)} \sim \mathbb{R}_1\,.$$

  Thus $(\mathbb{P}'_1 \| \mathbb{Q}'_1) = \mathbb{R}_1$.

- By the data processing inequality for TV distance (use it twice)

$$\Delta_{\mathrm{TV}}(\mathbb{P}'_1, \mathbb{P}') = \Delta_{\mathrm{TV}}(\mathbb{P}_1 \mathbb{P}_{4|3}, \mathbb{P}_{3,4}) = \Delta_{\mathrm{TV}}(\mathbb{P}_1 \mathbb{P}_{4|3}, \mathbb{P}_3 \mathbb{P}_{4|3}) = \Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_3)\,.$$

The rest of properties in (13) can be verified similarly. $\qquad\qquad\square$

Proof of Lemma 11. For any $\mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_2, \mathbb{Q}_2$ such that $\mathbb{R}_1 = (\mathbb{P}_1 \| \mathbb{Q}_1)$, $\mathbb{R}_2 = (\mathbb{P}_2 \| \mathbb{Q}_2)$, we have

$$|\Delta_{\mathrm{TV}}(\mathbb{R}_1) - \Delta_{\mathrm{TV}}(\mathbb{R}_2)| = |\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{Q}_1) - \Delta_{\mathrm{TV}}(\mathbb{P}_2, \mathbb{Q}_2)| \leq \Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2) + \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_2)$$

by triangle inequality. Thus

$$|\Delta_{\mathrm{TV}}(\mathbb{R}_1) - \Delta_{\mathrm{TV}}(\mathbb{R}_2)| \leq \inf_{\substack{(\mathbb{P}_1 \| \mathbb{Q}_1) = \mathbb{R}_1 \\ (\mathbb{P}_2 \| \mathbb{Q}_2) = \mathbb{R}_2}} \left(\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2) + \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_2) \right) \leq 2\, \Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_2). \qquad \square$$

Proof of Lemma 12. Consider any distributions $\mathbb{P}_1, \mathbb{Q}_1, \ldots, \mathbb{P}_4, \mathbb{Q}_4$ satisfying $\mathbb{R}_i = (\mathbb{P}_i \| \mathbb{Q}_i)$ for all $i \in [4]$. By Proposition 4, $\mathbb{R}_1 \cdot_{\mathsf{indp}} \mathbb{R}_2 = (\mathbb{P}_1 \mathbb{P}_2 \| \mathbb{Q}_1 \mathbb{Q}_2)$ and $\mathbb{R}_3 \cdot_{\mathsf{indp}} \mathbb{R}_4 = (\mathbb{P}_3 \mathbb{P}_4 \| \mathbb{Q}_3 \mathbb{Q}_4)$. Thus

$$\Delta_{\mathrm{MTV}}(\mathbb{R}_1 \cdot_{\mathsf{indp}} \mathbb{R}_2, \mathbb{R}_3 \cdot_{\mathsf{indp}} \mathbb{R}_4) \leq \max(\Delta_{\mathrm{TV}}(\mathbb{P}_1 \mathbb{P}_2, \mathbb{P}_3 \mathbb{P}_4), \ \Delta_{\mathrm{TV}}(\mathbb{Q}_1 \mathbb{Q}_2, \mathbb{Q}_3 \mathbb{Q}_4)).$$

By the the triangle inequality,

$$\Delta_{\mathrm{TV}}(\mathbb{P}_1 \mathbb{P}_2, \mathbb{P}_3 \mathbb{P}_4) \leq \Delta_{\mathrm{TV}}(\mathbb{P}_1 \mathbb{P}_2, \mathbb{P}_3 \mathbb{P}_2) + \Delta_{\mathrm{TV}}(\mathbb{P}_3 \mathbb{P}_2, \mathbb{P}_3 \mathbb{P}_4) = \Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_3) + \Delta_{\mathrm{TV}}(\mathbb{P}_2, \mathbb{P}_4)\,.$$

Similarly, $\Delta_{\mathrm{TV}}(\mathbb{Q}_1 \mathbb{Q}_2, \mathbb{Q}_3 \mathbb{Q}_4) \leq \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_3) + \Delta_{\mathrm{TV}}(\mathbb{Q}_2, \mathbb{Q}_4)$. So

$$\Delta_{\mathrm{MTV}}(\mathbb{R}_1 \cdot_{\mathsf{indp}} \mathbb{R}_2, \mathbb{R}_3 \cdot_{\mathsf{indp}} \mathbb{R}_4) \leq \max(\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_3), \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_3)) + \max(\Delta_{\mathrm{TV}}(\mathbb{P}_2, \mathbb{P}_4), \Delta_{\mathrm{TV}}(\mathbb{Q}_2, \mathbb{Q}_4)).$$

Since the inequality holds for any $\mathbb{P}_1, \mathbb{Q}_1, \ldots, \mathbb{P}_4, \mathbb{Q}_4$ satisfying $\mathbb{R}_i = (\mathbb{P}_i \| \mathbb{Q}_i)$,

$$\Delta_{\mathrm{MTV}}(\mathbb{R}_1 \cdot_{\mathsf{indp}} \mathbb{R}_2, \mathbb{R}_3 \cdot_{\mathsf{indp}} \mathbb{R}_4)$$
$$\leq \inf_{\substack{(\mathbb{P}_1 \| \mathbb{Q}_1) = \mathbb{R}_1 \\ (\mathbb{P}_3 \| \mathbb{Q}_3) = \mathbb{R}_3}} \max\left(\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_3), \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_3) \right) + \inf_{\substack{(\mathbb{P}_2 \| \mathbb{Q}_2) = \mathbb{R}_2 \\ (\mathbb{P}_4 \| \mathbb{Q}_4) = \mathbb{R}_4}} \max\left(\Delta_{\mathrm{TV}}(\mathbb{P}_2, \mathbb{P}_4) + \Delta_{\mathrm{TV}}(\mathbb{Q}_2, \mathbb{Q}_4) \right)$$
$$= \Delta_{\mathrm{MTV}}(\mathbb{R}_1, \mathbb{R}_3) + \Delta_{\mathrm{MTV}}(\mathbb{R}_2, \mathbb{R}_4)\,. \qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$$

Proof of Lemma 25. In the proof of Lemma 10, we have shown the following statement.

For any distributions $\mathbb{P}_1, \mathbb{Q}_1, \mathbb{P}_3, \mathbb{Q}_3$ over sample space Ω_1 and distributions $\mathbb{P}_2, \mathbb{Q}_2, \mathbb{P}_4, \mathbb{Q}_4$ over sample space Ω_2 satisfying $(\mathbb{P}_1\|\mathbb{Q}_1) = \mathbb{R}_1, (\mathbb{P}_2\|\mathbb{Q}_2) = \mathbb{R}_2, (\mathbb{P}_3\|\mathbb{Q}_3) = (\mathbb{P}_4\|\mathbb{Q}_4) = \mathbb{R}$, there exist distributions $\mathbb{P}', \mathbb{Q}', \mathbb{P}'_1, \mathbb{Q}'_1, \mathbb{P}'_2, \mathbb{Q}'_2$ satisfying (13).

Lemma 25 follows almost directly from the statement, by letting

$$\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}_1, \mathbb{R}) \ni (\Delta_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_3), \Delta_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_3)) \leq (\alpha_1, \beta_1),$$
$$\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}_2, \mathbb{R}) \ni (\Delta_{\mathrm{TV}}(\mathbb{P}_2, \mathbb{P}_4), \Delta_{\mathrm{TV}}(\mathbb{Q}_2, \mathbb{Q}_4)) \leq (\alpha_2, \beta_2),$$

then the existing of $\mathbb{P}', \mathbb{Q}', \mathbb{P}'_1, \mathbb{Q}'_1, \mathbb{P}'_2, \mathbb{Q}'_2$ satisfying (13) implies

$$\mathcal{R}_{\mathrm{TVD}}(\mathbb{R}_1, \mathbb{R}_2) \ni (\Delta_{\mathrm{TV}}(\mathbb{P}'_1, \mathbb{P}'_2), \Delta_{\mathrm{TV}}(\mathbb{Q}'_1, \mathbb{Q}'_2))$$
$$\leq (\Delta_{\mathrm{TV}}(\mathbb{P}'_1, \mathbb{P}') + \Delta_{\mathrm{TV}}(\mathbb{P}', \mathbb{P}'_2), \Delta_{\mathrm{TV}}(\mathbb{Q}'_1, \mathbb{Q}') + \Delta_{\mathrm{TV}}(\mathbb{Q}', \mathbb{Q}'_2)) \leq (\alpha_1 + \alpha_2, \beta_1 + \beta_2). \quad \square$$

C   Visualize the Sparsification Process

This section provides an informal intuitive visualization of the sparsification process. The Neyman-Pearson region of a decision problem $(\mathbb{P}, \mathbb{Q})$ is considered when proving the positivity of the minimum total variation distance (Lemma 10). In that proof, we claim that the Neyman-Pearson region contains exactly the same information as the ratio $\mathbb{R} := (\mathbb{P}\|\mathbb{Q})$.

Say the ratio $\mathbb{R} := (\mathbb{P}\|\mathbb{Q})$ is represented by a sorted table $(r_1, p_1), (r_2, p_2), (r_3, p_3), \ldots$, that is, $r_1 < r_2 < \ldots$ and $\mathbb{R}(r_i) = p_i$. By the Neyman-Pearson lemma, the (upper) boundary of the Neyman-Pearson region is the polygonal chain that connects

$$(0,0), \underbrace{(\mathbb{P}([0, r_1]), \mathbb{Q}([0, r_1]))}_{=(r_1 p_1, p_1)}, \underbrace{(\mathbb{P}([0, r_2]), \mathbb{Q}([0, r_2]))}_{=(r_1 p_1 + r_2 p_2, p_1 + p_2)}, \ldots, \underbrace{(\mathbb{P}([0, r_t], \mathbb{Q}([0, r_t])))}_{=(\sum_{t<i} r_t p_t, \sum_{t<i} p_t)}, \ldots, (1, 1),$$

as illustrated in Figure 1. The i-th direct segment in the polygonal chain is the vector $(r_i p_i, p_i)$. This also explains how the Neyman-Pearson region contains exactly the same information as the ratio.

The goal of our sparsification process is to find a simpler ratio (the sparsified ratio) that is close to the given ratio. From another point of view, the goal is to find a region with simpler boundary that is close to given Neyman-Pearson region.

If a few consecutive direct segments in the boundary have similar slopes, they can be replaced by a single direct segment, causing minor change on the shape of the region. This is exactly how our sparsification algorithm works: identifying a cluster of probability masses then merging them together.
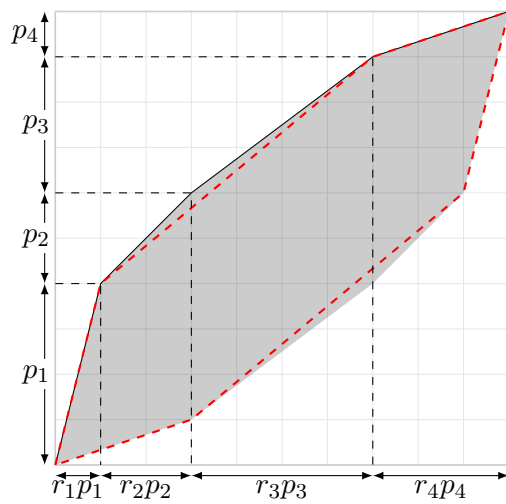
Figure 1: Illustration of the Neyman-Pearson region when ratio has support size 4. The red dashed line marks the boundary of the new Neyman-Pearson region of the sparsified ratio.