Approximating the Total Variation Distance between Gaussians

Arnab Bhattacharyya
The University of Warwick

Weiming Feng
The University of Hong Kong

Piyush Srivastava
Tata Institute of Fundamental Research

Abstract

The total variation distance is a metric of central importance in statistics and probability theory. However, somewhat surprisingly, questions about computing it *algorithmically* appear not to have been systematically studied until very recently. In this paper, we contribute to this line of work by studying this question in the important special case of multivariate Gaussians. More formally, we consider the problem of approximating the total variation distance between two multivariate Gaussians to within an ϵ-relative error. Previous works achieved a *fixed* constant relative error approximation via closed-form formulas. In this work, we give algorithms that given any two n-dimensional Gaussians D_1, D_2, and any error bound $\epsilon > 0$, approximate the total variation distance $D := d_{TV}(D_1, D_2)$ to ϵ-relative accuracy in $\mathrm{poly}(n, \frac{1}{\epsilon}, \log \frac{1}{D})$ operations. The main technical tool in our work is a reduction that helps us extend the recent progress on computing the TV-distance between *discrete* random variables to our continuous setting.

1 INTRODUCTION

A fundamental problem in applied probability is to understand how different two probability distributions are from each other. Depending upon the applications, various measures of divergence are used for this purpose; one of the most important of these is the *total variation (TV) distance*. Given two probability distributions P and Q on the same set Ω, the total variation distance $d_{\mathrm{TV}}(P, Q)$ is the largest possible difference, over all *tests* $S \subset \Omega$, between the probabilities $P(S)$ and $Q(S)$ of the test "succeeding" on a sample drawn from P and Q respectively.

A natural question then is: given the specifications of P and Q, can one compute $d_{\mathrm{TV}}(P, Q)$? A standard alternative expression for the TV distance in terms of the ℓ_1-norm shows that when the sample space Ω is *finite*, $d_{\mathrm{TV}}(P, Q)$ can be computed exactly in $O(|\Omega|)$ arithmetic operations, given the values of $P(\omega)$ and $Q(\omega)$ for each $\omega \in \Omega$. However, even in the slightly more general (and much more important!) setting of *product distributions*, the problem becomes much more interesting. Now, we are given distributions (P_i, Q_i) on the same finite set Ω, for $1 \leq i \leq n$, and we wish to compute the TV distance between the product distributions $P := P_1 \times \cdots \times P_n$ and $Q := Q_1 \times \cdots \times Q_n$ on Ω^n. The obvious algorithm from above now has a prohibitive, exponentially growing cost of $\Theta(|\Omega|^n)$.

Somewhat surprisingly, it seems that the computational complexity of computing the TV distance, even in such simple settings, had not been systematically studied until recently. Bhattacharyya et al. (2023) provided evidence that even in the setting of product distributions, we probably cannot do much better than the above obvious algorithm if we want to compute d_{TV} exactly. They did so by proving a #P-hardness result: a polynomial time algorithm for the above problem would imply a polynomial time algorithm for counting satisfying assignments to Boolean formulas. We therefore turn to *approximation*, and require instead that given a further input $\epsilon \in (0, 1)$, the algorithm should output a number Z such that $|Z - d_{\mathrm{TV}}(P, Q)| \leq \epsilon d_{\mathrm{TV}}(P, Q)$ (i.e., the algorithm makes small relative error), in time that is polynomial in $n, |\Omega|, 1/\epsilon$ and $\log(1/d_{\mathrm{TV}}(P, Q))$. We note here that the dependence on $\log(1/d_{\mathrm{TV}}(P, Q))$ is reasonable. First, we may assume that $d_{\mathrm{TV}}(P, Q) \neq 0$, since otherwise, the problem is trivial. Further, if the final answer is a binary number, then the output bit length itself is $\Omega(\log(1/d_{\mathrm{TV}}(P, Q)))$. In addition, if all the input numbers are rationals specified as ratios of binary integers, then $\log(1/d_{\mathrm{TV}}(P, Q))$ is at most a

Authors are listed in alphabetical order of last names.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

polynomial in the input size.[1] We refer to this problem as DISPRODTV; see Problem 2.4 for a formal description. Bhattacharyya et al. (2023) gave algorithms for DISPRODTV in certain special cases. An elegant randomized algorithm for DISPRODTV that covers all cases was then given by Feng et al. (2023). More recently, Feng et al. (2024) gave a deterministic polynomial time algorithm for DISPRODTV.

1.1 Our Contribution

In this paper, we consider the problem of computing the TV distance between P and Q when they are Gaussian distributions in $\mathbb{R}^n$, specified by their mean vectors and covariance matrices. Towards this goal, we make the following contributions.

A General (and Simpler) Analysis Feng et al. (2024) developed a framework for studying TV distance estimation between products of discrete distributions on finite sets. We simplify and also extend this framework to the setting of general product distributions (including those without a probability density function). Our framework allows us to bypass one of the main innovations of Feng et al. (2024), which they referred to as the Δ_{MTV} distance, by recasting the analysis in terms of a more direct quantity that we call the *extension distance*. As an application, we give an arguably simpler proof of correctness of the algorithm of Feng et al. (2024) in Section 3.

TV Distance between Gaussians We then apply our framework to the problem of estimating the TV distance between P and Q when P and Q are arbitrary Gaussian distributions in $\mathbb{R}^n$. Our first step is to use the linear algebraic properties of Gaussians to reduce the problem to the case where P and Q are both products of one-dimensional Gaussians. Assuming that linear algebraic operations such as eigen-decomposition of symmetric matrices can be performed exactly, this part of the reduction is exact; however, we comment on the bit-complexity issues arising in this part in the supplementary material. After this reduction, we are *almost* in the setting of prior work (Feng et al., 2023, 2024) that studied the problem of estimating the TV distance between product distributions, *except* that the time complexity in these works has a polynomial dependence on the size of the state space Ω of each component of the product. For us, however, this set Ω is the uncountable set $\mathbb{R}$ of all real numbers. We thus need to perform a careful discretization step to bring the problem to the setting of prior work. Happily, it turns out that this discretization can be analyzed using our extension (described above) of the framework of Feng et al. (2024). Combining everything, we obtain an algorithm for approximating TV distance between arbitrary Gaussian distributions P and Q in $\mathbb{R}^n$ to a relative accuracy of $\pm\epsilon$ in time polynomial in $n, 1/\epsilon$ and $\log(1/d_{\mathrm{TV}}(P,Q))$, in a computational model where eigenvalues and eigenvectors of symmetric matrices are assumed to be exactly computable. Below, we give the statements of our results regarding TV distance between Gaussians; the proofs of these results are given in Section 4. The problem we are interested in is the following.

Problem 1.1 (**MultGaussianTV**)**. INPUT**: Two n-dimensional mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_1 \in \mathbb{R}^n$ and two $n \times n$ positive semi-definite covariance matrices $\Sigma_1, \Sigma_2 \in \mathbb{R}^{n\times n}_{\geq 0}$, and a rational number $\epsilon \in (0,1)$. **OUTPUT**: A number z such that $(1-\epsilon)D \leq z \leq (1+\epsilon)D$, where $D := d_{\mathrm{TV}}\left(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1), \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)\right)$.

The case $D = 0$ occurs only if $\boldsymbol{\mu_1} = \boldsymbol{\mu_2}$ and $\Sigma_1 = \Sigma_2$, which can be easily detected. Thus, for ease of presentation, we assume that the algorithm only has to consider inputs for which $D > 0$.

Computational model For ease of presentation, in the main body of the paper, we present our results using the real-RAM model, where basic arithmetic operations (comparison, addition, subtraction, multiplication, division, logarithm, and square root) on real numbers take one unit of time. This is one of the standard cost models in numerical analysis and computational geometry (see, e.g. Preparata and Shamos (1985, Section 1.4) and Bürgisser and Cucker (2013, Section 5.1)). For the randomized version of our algorithm (Theorem 1.3), we also assume that the algorithm can access randomness in the following way: for any $p \in [0,1]$ that the algorithm constructs, it can obtain, in one operation, a random bit sampled according to Bernoulli(p) and independent of all previous bits it has used. However, our algorithm does *not* require the ability to sample from continuous distributions (e.g. Gaussian distributions).[2] However, as stated above, we do assume the ability to perform diagonalization of symmetric matrices. Our results assuming this computation model are stated below in Theorems 1.2 and 1.3.

Theorem 1.2. *There exists a deterministic algorithm that solves* MULTGAUSSIANTV *using at most*

[1] However, if we measure only the number of *arithmetic* operations, rather than the number of *bit* operations, then this dependence on $\log(1/d_{\mathrm{TV}}(P,Q))$ is not necessarily essential. We comment further on this point later in the paper.

[2] Note, however, that in the supplementary material, we describe how to implement our algorithms given access only to an approximate diagonalization oracle; and the best known approximate diagonalization algorithms do require access to approximate samples from uniform or standard normal distributions.

$O\left(\frac{n^3}{\epsilon^2}\log^2\frac{n}{\epsilon D}\left(\log\frac{n}{\epsilon}+\log\log\frac{3}{D}\right)+\frac{n^2}{\epsilon}\log^3\frac{n}{\epsilon D}\right)$ *arithmetic operations along with diagonalizations of two square matrices of size at most* $n\times n$.

Theorem 1.3. *There exists a randomized algorithm that solves* MultGaussianTV *with probability at least* $1-\delta$ *using at most* $O\left(\frac{n^3}{\epsilon^3}\log\frac{n}{\epsilon D}\log\frac{1}{\delta}+\frac{n^2}{\epsilon}\log^3\frac{n}{\epsilon D}\right)$ *arithmetic operations along with diagonalizations of two square matrices of size at most* $n\times n$.

The bit complexity model Although the real-RAM model we use above is convenient to work with (and also popular), a more realistic cost model is to assume that the inputs are *rational numbers* encoded as ratios of integers specified in binary, and to account for costs using the total number of *bit* operations required to produce the desired output. This is one of the standard cost models in computational complexity theory and in optimization (Grötschel et al., 1993). Note that in this model, operations such as logarithm, square root, and matrix diagonalization can only be carried out approximately.[3] An algorithm in this model is said to run in polynomial time if the total number of bit operations it performs is at most a polynomial function of the total representation length of the input.

In Sections 6 and 7 of the supplementary material, we show how our algorithms continue to work efficiently even when steps requiring exact computation of functions such as logarithm, square root and matrix diagonalization are replaced by polynomial-time but approximate versions of these functions implemented in the bit complexity model. While we focus on the real-RAM model in the main body of the paper, we include pointers to the supplementary material at points where such replacements may be required when analyzing the algorithm in the bit complexity model.

1.2 Discussion and Related Work

The problem of approximating $d_{\mathrm{TV}}(P,Q)$ to within additive error of $\pm 1/10$, given access only to efficient samplers for P and Q is believed to be hard even for discrete distributions on finite sets (Sahai and Vadhan, 2003): solving it efficiently would break certain cryptosystems (see, e.g., Bouland et al. (2020, Section 1.1)). This hardness result holds even if one is also given access to descriptions of the circuits implementing the samplers. If one is also in addition given access to the probability density functions of P and Q, then one can attempt a Monte Carlo approach, but such an approach would require about $1/\epsilon^2$ samples for an estimate with an *additive* error ϵ. Such additive error algorithms for the TV distance have also been studied by Kiefer (2018), Bhattacharyya et al. (2020), and Tao et al. (2024) in various settings. However, such additive error results can become inefficient when transferred to the relative error setting because the TV distance itself can be exponentially small in the underlying dimension (so that, e.g. in the Monte Carlo approach, the required number of samples would also grow exponentially). One is therefore forced to look for other methods in the relative error setting, and this has been done in various discrete settings in papers by Bhattacharyya et al. (2023, 2024) and Feng et al. (2023, 2024) (some of which were already discussed above).

Prior work on approximating TV distance between Gaussians has focused on achieving a *fixed* relative error, but by means of closed form formulas; see, e.g. Devroye et al. (2023) and references therein, Barabesi and Pratelli (2024) for some recent progress in a special case, and Davies et al. (2022) for similar results in the setting of mixtures of Gaussians. The quantitatively best such result we are aware of appears in the work of Arbas et al. (2023, Theorem 1.8), who give a $100\sqrt{2}$-factor approximation for the TV distance between multivariate Gaussians via a closed form formula. (In contrast, note that Theorems 1.2 and 1.3 produce $(1+\epsilon)$-factor multiplicative approximations for any $\epsilon>0$.) The case of one-dimensional Gaussians is related to the numerical evaluation of the *error function*, a problem of much interest in numerical analysis (Chevillard, 2012). In our analysis, we do use some of the above results (see, e.g., Proposition 2.3 and Lemma 4.3).

2 PRELIMINARIES

Total Variation Distance Let P and Q be two probability measures on a measurable space $(\Omega,\mathcal{F})$. Their *total variation distance* (TV distance) is defined as $d_{\mathrm{TV}}(P,Q):=\sup_{A\in\mathcal{F}}|P(A)-Q(A)|$. We often denote $d_{\mathrm{TV}}(P,Q)$ by $d_{\mathrm{TV}}(X,Y)$, where $X\sim P$ and $Y\sim Q$ are random variables with laws P and Q respectively.

For real vector valued random variables with $\Omega=\mathbb{R}^n$, we assume $\mathcal{F}=\mathcal{B}(\mathbb{R}^n)$ is the usual Borel σ-field. In particular, if P and Q are two probability measures in $\mathbb{R}^d$ with probability density functions (PDFs) p and q respectively, then $d_{\mathrm{TV}}(P,Q)=\frac{1}{2}\int_{x\in\mathbb{R}^d}|p(x)-q(x)|dx$.

We will use the following standard property of d_{TV}.

Proposition 2.1. *Let* X *and* Y *be two random variables. For any measurable* f*, it holds that* $d_{\mathrm{TV}}(X,Y)\geq d_{\mathrm{TV}}(f(X),f(Y))$*. If* f *is further an invertible function,*

[3] The floating point model (see, e.g., Demmel (1997)), perhaps the most widely used in practice, can be seen as a version of this model in which one fixes a precision k and a range parameter R, and must approximate the result of every intermediate operation by a number of the form $s\times 2^m$ where s,m are integers satisfying $|s|<2^k$ and $|m|\leq R$.

then $d_{\mathrm{TV}}(X,Y) = d_{\mathrm{TV}}(f(X), f(Y))$.

Gaussian Distribution A real valued random variable X is said to be *Gaussian* with *mean* μ and *variance* σ^2 if it has the probability density function (PDF) $f = f_{\mu,\sigma^2}$ (shortened to f when μ and σ^2 are clear from the context) given by $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for all $x \in \mathbb{R}$. The probability distribution of X is denoted by $\mathcal{N}(\mu, \sigma^2)$. $\mathcal{N}(0,1)$ is called the *standard* Gaussian distribution.

A vector valued random variable $\boldsymbol{X}$ taking values in $\mathbb{R}^n$ is said to be Gaussian with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and (symmetric, positive semi-definite) *covariance matrix* Σ in $\mathbb{R}^n$ if for every vector $\boldsymbol{a} \in \mathbb{R}^n$, the real valued random variable $\boldsymbol{a}^T\boldsymbol{X}$ has law $\mathcal{N}(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\Sigma\boldsymbol{a})$. The distribution of $\boldsymbol{X}$ is denoted $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. When $\boldsymbol{\mu} = \boldsymbol{0}$ and $\Sigma = I_n$ (the $n \times n$ identity matrix), the distribution is said to be the *standard* Gaussian in n-dimensions. In general, when Σ is diagonal, the individual components X_i of $\boldsymbol{X}$ are independent (real) Gaussian random variables with variance $\Sigma_{i,i}$. We collect below some standard facts about the Gaussian distribution.

Fact 2.2. *The random variable* $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ *is identically distributed as* $C\boldsymbol{Z} + \boldsymbol{\mu}$*, where* $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, I_r)$ *is a standard* r*-dimensional Gaussian,* $C = \mathbb{R}^{n\times r}$ *is a matrix such that* $CC^T = \Sigma$*, and* $r \leq n$ *is the rank of the matrix* Σ*. Thus,* $\boldsymbol{X}$ *is supported on the affine space* $\mathbf{Range}(\Sigma) + \boldsymbol{\mu}$*, where* $\mathbf{Range}(\Sigma)$ *is the column space of* Σ*. Further, if* $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$*, where* $A \in \mathbb{R}^{m\times n}$ *and* $\boldsymbol{b} \in \mathbb{R}^m$*, then* $\boldsymbol{Y}$ *has the law* $\mathcal{N}(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^T)$*.*

The distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is said to be *non-degenerate* if Σ is invertible, and hence positive definite. The probability density function (PDF) $f = f_{\boldsymbol{\mu},\Sigma} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ (shortened to f when $\boldsymbol{\mu}$ and Σ are clear from the context) of a non-degenerate $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is given by $f(\boldsymbol{x}) := \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\mu)\right)}{\sqrt{(2\pi)^n \det(\Sigma)}}$.

The Error Function The error function, $\mathrm{erf}(\cdot)$, defined as $\mathrm{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\, dt$, for $x \in [-\infty, \infty]$, is a useful primitive arising in several numerical applications. Note that if F denotes the cumulative distribution function *CDF* of $\mathcal{N}(0,1)$, then for all $x \in [-\infty, \infty]$, $F(x) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$. The computation of the error function has been extensively studied. We will need the following result about the additive approximation of the error function, which follows from methods reported by Chevillard (2012) (we include the proof, along with an analysis in the bit complexity model, in Section 6 of the supplementary material).

Proposition 2.3. *There exists an algorithm such that given* μ, σ^2*,* $-\infty \leq a \leq b \leq \infty$ *and* $0 < \epsilon \leq 1/2$*, it returns a real number* $y \geq 0$ *such that* $|y - \int_a^b f(t)dt| \leq \epsilon$*, where* $f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$*. The algorithm uses* $O(\log^2(1/\epsilon))$ *arithmetic operations.*

Discrete Product Distributions We collect here known results on approximating TV-distances between products of discrete distributions, that were discussed in the introduction.

Problem 2.4 (DisProdTV). INPUT: $2n$ discrete distributions P_i, Q_i for $1 \leq i \leq n$ over a finite domain $[M] = \{1, 2, \ldots, M\}$, and a rational number $\epsilon \in (0,1)$. **OUTPUT**: A number z such that $(1-\epsilon)d_{\mathrm{TV}}(P,Q) \leq z \leq (1+\epsilon)d_{\mathrm{TV}}(P,Q)$, where $P = P_1 \times \ldots \times P_n$ and $Q = Q_1 \times \ldots \times Q_n$ are two product distributions over $[M]^n$.

As mentioned in the introduction, the case $\epsilon = 0$ of the above problem is #P-complete (Bhattacharyya et al., 2023). The following two approximation algorithms are known.

Theorem 2.5 (Feng et al. (2024))**.** *There exists a deterministic algorithm that solves* DisProdTV *in time* $O(\frac{Mn^2}{\epsilon} \log M \log \frac{n}{\epsilon d_{\mathrm{TV}}(P,Q)})$.

Theorem 2.6 (Feng et al. (2023))**.** *There exists a randomized algorithm that solves* DisProdTV *in time* $O(\frac{Mn^2}{\epsilon^2} \log \frac{1}{\delta})$ *with probability at least* $1 - \delta$*.*

The running time in the above theorem is in terms of the number of arithmetic operations. Further, the randomized algorithm underlying Theorem 2.6 is assumed to have the following access to randomness: for any $p \in [0,1]$ that the algorithm constructs, it can obtain, in one operation, a random bit sampled according to Bernoulli(p) and independent of all previous bits it has used.

We will use both these algorithms as subroutines in order to derive Theorems 1.2 and 1.3. However, in Section 3, we give a new (and arguably simpler) proof of correctness of the algorithm underlying Theorem 2.5, as a corollary of our framework for analyzing TV distance between product distributions.

Remarks on bit complexity The proofs in Feng et al. (2023, 2024) show that the algorithms in Theorems 2.5 and 2.6 are also polynomial-time in terms of bit complexity. (In particular, note that when the inputs $(P_i, Q_i)_{i=1}^n$ to DisProdTV are rational numbers specified as ratios of integers denoted in binary, $d_{\mathrm{TV}}(P,Q)$ has a representation length that is linear in the total input size.) Further, the model of access to randomness for the algorithm in Theorem 2.6 can be relaxed so that the algorithm can only access a sequence of i.i.d. Bernoulli(1/2) random bits: this is essentially because for any p of the form $k/2^\ell$, where

ℓ is a positive integer and $k \leq 2^l$ is an odd integer, a sample from Bernoulli($k/2^\ell$) can be obtained using ℓ i.i.d. bits sampled from Bernoulli(1/2) (in fact, in expectation, one only needs to sample two bits, even if p is not of the form $k/2^\ell$).

Note, however, that in the algorithm of Theorem 2.5, not the just the number of *bit* operations, but even the number of *arithmetic operations* depends upon both the *number* of input parameters and their total representation length. In other words, while this algorithm is polynomial-time in term of bit complexity, it is not "strongly polynomial" (Grötschel et al., 1993, p. 32). In the context of Theorem 2.5, a strongly polynomial algorithm would have to be polynomial-time in terms of bit complexity, and in addition, the number of arithmetic operations it performs would be bounded above by a polynomial *only* of n, M and $1/\epsilon$. To the best of our knowledge, it is not known whether a strongly polynomial deterministic algorithm for DISPRODTV exists.

3 THE EXTENSION DISTANCE

In this section, we present a proof of correctness of the algorithm of Feng et al. (2024) for discrete distributions by recasting its analysis in the general framework of Radon-Nikodym derivatives. For our application to Gaussians, we do not need the results at this level of generality, but we believe that presenting the ideas of Feng et al. (2024) in this general framework both simplifies the presentation and can also be helpful for future extensions of their algorithmic ideas. Further, even in the case of discrete distributions, our proof of correctness of their algorithm is different (and arguably simpler) because it dispenses with one of the main innovations of Feng et al. (2024) (which they referred to as the Δ_{MTV} distance) and shows instead that their algorithm for product distributions can be analyzed in terms of a more direct quantity which we call the *extension distance* (see Definition 3.9).

We begin with some preliminary notions. Our notations and conventions for the standard notions described below follows the standard references Williams (1991) and Durrett (2019, Appendix A.4). In particular, following these references, we will agree that two random variables are said to be the same (with respect to a measure) if they differ only on a set of measure 0.

Definition 3.1 (Radon-Nikodym derivative (Durrett, 2019, Theorems A.4.7 and A.4.8)). Let P and Q be probability measures on a measurable space $(\Omega, \mathcal{F})$. There exists a unique decomposition of P into two finite non-negative measures P_1 and P_2 such that $P = P_Q + P_Q^\perp$, P_Q is *absolutely continuous* with respect to Q (denoted $P \ll Q$, and meaning that for every event $A \in \mathcal{F}$, $Q(A) = 0$ implies $P_Q(A) = 0$), while $P_Q^\perp$ and Q are mutually singular (meaning that there exists an event $A \in \mathcal{F}$ such that $P_Q^\perp(A) = 0$ and $Q(A^c) = 0$). Further, there exists a unique (with respect to the probability measure Q) $\mathcal{F}$-measurable non-negative random variable R, denoted $\frac{\mathrm{d}P_Q}{\mathrm{d}Q}$, such that for any event $A \in \mathcal{F}$, $P(A) = P_Q^\perp(A) + \mathbb{E}_Q[RI_A]$, where I_A denotes the indicator function of A. In case $P \ll Q$, we have $P_Q^\perp \equiv 0$, so that the first term on the right hand side of the above equation is zero, and $R = \frac{\mathrm{d}P}{\mathrm{d}Q}$ is referred to as the *Radon-Nikodym derivative* of P with respect to Q. Also, $\mathbb{E}_Q[R] \leq 1$, with equality holding if and only if $P \ll Q$.

Remark 3.2. The random variable $R = \frac{\mathrm{d}P_Q}{\mathrm{d}Q}$, together with the law induced on it by the probability measure Q, was called the *ratio* in the work of Feng et al. (2024), and was denoted in that paper by the notation $P\|Q$. We also adopt this notation. Note that the ratio is the same as the usual Radon-Nikodym derivative when $P \ll Q$. The following definition (a rephrasing of similar definitions given by Feng et al. (2024)) will be useful.

Definition 3.3 (Valid ratio and independent products). A non-negative random variable R defined on a probability space $(\Omega, \mathcal{F}, Q)$ is said to be a *valid ratio* if $\mathbb{E}[R] \leq 1$. Given valid ratios R_i on $(\Omega_i, \mathcal{F}_i, Q_i)$ for $i \in 1, 2$, the *independent product* $R_1 \underset{\text{ind}}{\circ} R_2$ of R_1 and R_2 is the valid ratio on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, Q_1 \times Q_2)$ defined by $(R_1 \underset{\text{ind}}{\circ} R_2)(\omega_1, \omega_2) := R_1(\omega_1)R_2(\omega_2)$. Note that the binary operation $\underset{\text{ind}}{\circ}$ is associative, and that the random variables $R_1 \underset{\text{ind}}{\circ} R_2$ and $R_2 \underset{\text{ind}}{\circ} R_1$ have the same cumulative distribution function.

We record the following standard observation about the Radon-Nikodym derivatives in the setting of product measures.

Fact 3.4. *Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega, \mathcal{F}_2)$ be measurable spaces, and consider the product measurable space $(\Omega_1 \times \Omega_2, \mathcal{F})$ where $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$ denotes the product σ-algebra of $\mathcal{F}_1$ and $\mathcal{F}_2$. Let π_1 and π_2 be probability measures on $(\Omega_1, \mathcal{F}_1)$ and, similarly, let Q and P be probability measures on $(\Omega_2, \mathcal{F}_2)$. Then for the product measures $\pi_1 \times Q$ and $\pi_2 \times P$ on $(\Omega, \mathcal{F})$ we have $((\pi_2 \times P) \,\|\, (\pi_1 \times Q))(\omega_1, \omega_2) = (\pi_2 \| \pi_1)(\omega_1) \cdot (P\|Q)(\omega_2)$. In terms of Definition 3.3, this can be written as $(\pi_2 \times P) \,\|\, (\pi_1 \times Q) = (\pi_2 \| \pi_1) \underset{\text{ind}}{\circ} (P\|Q)$.*

It is well known that $d_{\mathrm{TV}}(P, Q) = \mathbb{E}[(1 - R)_+]$ when $R = P\|Q$ (where $|x|_+ := \max\{x, 0\}$; see the supplementary material for a proof). This expression suggests defining the following functional.

Definition 3.5 (**The TV functional**)**.** Let R be a valid ratio on the probability space $(\Omega, \mathcal{F}, Q)$. Then, we define $TV(R) := \mathbb{E}[(1-R)_+] = \mathbb{E}[(R-1)_+] + 1 - \mathbb{E}[R] = \frac{1}{2}\left(\mathbb{E}[|1-R|] + 1 - \mathbb{E}[R]\right)$, where the equality of the three expressions follows by simple algebra. Thus, if P and Q are probability measures on $(\Omega, \mathcal{F})$ such that $R = P\|Q$, then $d_{\mathrm{TV}}(P, Q) = TV(R)$.

An important property of the TV functional is that $TV(R)$ depends only upon the cumulative distribution function (c.d.f.) of the random variable R, and it does not otherwise depend on the probability space on which R has been defined. In particular, if R and S are valid ratios defined on possibly different probability spaces, but if R and S have the same c.d.f., then $TV(R) = TV(S)$. The following monotonicity properties of the TV functional are useful.

Lemma 3.6. *Let R be a valid ratio defined on a probability space $(\Omega, \mathcal{F}, P)$. Then, (1) For every $c \in [0, 1]$, the valid ratio cR satisfies $TV(R) \leq TV(cR)$. (2) If $\mathcal{G} \subset \mathcal{F}$ is a sub-σ-algebra, then the valid ratio $R_{\mathcal{G}} := \mathbb{E}[R|\mathcal{G}]$ satisfies $TV(R_{\mathcal{G}}) \leq TV(R)$. Further, if the event $\{R < 1\}$ is $\mathcal{G}$-measurable, then the events $\{R < 1\}$ and $\{R_{\mathcal{G}} < 1\}$ are almost surely equivalent (i.e., the probability that exactly one of them happens is 0), and $TV(R_{\mathcal{G}}) = TV(R)$. (3) If S is a valid ratio defined on some probability space $(\Omega', \mathcal{F}', P')$ then $TV(R) \leq TV(R \underset{\text{ind}}{\circ} S)$.*

At the heart of the algorithmic framework of Feng et al. (2024) is a discretization scheme for ratios. Consider the following partition of the non-negative reals.

Definition 3.7 (**(γ, δ)-partition**)**.** Fix $\gamma, \delta \in (0, 1)$, and let $m = 1 + \lceil \ln(1/\gamma)/\ln(1+\delta) \rceil$. Define $a_0 = 1$, $a_m = 0$ and $a_k := 1 - (1+\delta)^{k-1}\gamma$ for integers $1 \leq k < m$, and the intervals $I_0 := [1, 1] = \{1\}$ and $I_k := [a_k, a_{k-1})$ for $1 \leq k \leq m$, which partition $[0, 1]$ into disjoint intervals. Similarly, define $J_k := (1/a_{k-1}, 1/a_k]$ for $1 \leq k < m$ and $J_m := (1/a_{m-1}, \infty)$. The collection of disjoint intervals

$$\mathcal{I}_{\gamma,\delta} := \{I_0\} \cup \{I_i \,:\, 1 \leq k \leq m\} \cup \{J_i \,:\, 1 \leq k \leq m\}$$

is called a (γ, δ)-*partition* of $\mathbb{R}_{\geq 0}$. By a slight overloading of notation, we will also view $\mathcal{I}_{\gamma,\delta}$ as a function from $\mathbb{R}_{\geq 0}$ to the set $\mathcal{I}_{\gamma,\delta}$ by setting $\mathcal{I}_{\gamma,\delta}(x) = I$ for $x \in \mathbb{R}_{\geq 0}$ when I is the unique interval in $\mathcal{I}_{\gamma,\delta}$ satisfying $x \in I$.

We can now describe the discretization procedure of Feng et al. (2024) in our setting.

Definition 3.8 (**(γ, δ)-discretization**)**.** Let R be a valid ratio on some probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{I}_{\gamma,\delta}$ be the partition of $\mathbb{R}_{\geq 0}$ in Definition 3.7. The (γ, δ)-*discretization* of R is the $\sigma(\mathcal{I}_{\gamma,\delta}(R))$-measurable random variable $\tilde{R}_{\gamma,\delta}$ defined by

$$\tilde{R} = \tilde{R}_{\gamma,\delta} := \mathbb{E}[R|\mathcal{I}_{\gamma,\delta}(R)]. \tag{1}$$

In particular, it follows from item 2 of Lemma 3.6 that $\tilde{R}$ is also a valid ratio satisfying $TV(\tilde{R}) = TV(R)$, and that, almost surely, the events $\{R < 1\}$ and $\{\tilde{R} < 1\}$ are equivalent. In fact, a very similar argument as in the proof of that lemma (given in the supplementary material) shows that for each $I \in \mathcal{I}_{\gamma,\delta}$, the events $\{R \in I\}$ and $\{\tilde{R} \in I\}$ are almost surely equivalent.

Our point of departure from the work of Feng et al. (2024) is in how the "error" introduced by the above discretization is measured. Feng et al. (2024) defined a quantity which they call Δ_{MTV}, and which they relate to Le Cam's deficiency, and measure the discretization error in terms of this quantity. However, this quantity is defined in terms of total variation distances between distributions realizing the appropriate ratios. We follow instead a more direct approach, and show that the discretization error can be measured in terms of the following, arguably simpler, definition.

Definition 3.9 (**Extension distance**)**.** Let R_1 and R_2 be valid ratios defined on some probability space $(\Omega, \mathcal{F}, P)$. The *extension distance* $\mathrm{ext}(R_1, R_2)$ between R_1 and R_2 is defined as

$$\sup_S \left| TV(S \underset{\text{ind}}{\circ} R_1) - TV(S \underset{\text{ind}}{\circ} R_2) \right|, \tag{2}$$

where the supremum is over all valid ratios S. Recall that since $TV(R)$ depends only upon the c.d.f. of R, we can restrict S to the set of non-negative Borel measurable random variables with expectation at most 1. From the definition itself, it is apparent that $\mathrm{ext}(\cdot, \cdot)$ is at least a *pseudo-metric*: it is non-negative, symmetric, and satisfies the triangle inequality. These properties will be sufficient for our purposes, and so we do not consider the question of whether it is a metric as well.

The utility of the extension distance comes from the following result, which relates it to the discretization procedure of Feng et al. (2024). Some of the computations in the (short) proof of this result (given in the supplement) are similar in spirit to parts of the analysis by Feng et al. (2024), though there they were performed with the goal of controlling their Δ_{MTV} distance.

Theorem 3.10. *Let $\tilde{R}$ be the (γ, δ)-discretization of a valid ratio R defined on some probability space $(\Omega, \mathcal{F}, P)$. Then* $\mathrm{ext}(R, \tilde{R}) \leq \gamma + \delta TV(R)$.

As an application of the extension distance framework, we now analyze the correctness of the algorithm of Feng et al. (2024) for computing the total variation distance between products of discrete distributions. We restate their algorithm for reference as Algorithm 1.

Theorem 3.11. *The algorithm* DisProdTV *is correct.*

Proof. We follow the notation used in Algorithm 1. For $1 \le i \le n-1$ define $S_i := R_{i+1} \underset{\text{ind}}{\circ} R_{i+2} \underset{\text{ind}}{\circ} \cdots \underset{\text{ind}}{\circ} R_n$, and set $S_n \equiv 1$ (i.e., the trivial ratio that takes value 1 with probability 1). Note that $\Delta \le d_{TV}(P,Q) = TV(Y_1 \underset{\text{ind}}{\circ} S_1)$. Further, for each $1 \le i \le n-1$, $\tilde{Y_i} \underset{\text{ind}}{\circ} S_i = Y_{i+1} \underset{\text{ind}}{\circ} S_{i+1}$ (from line 6 of Algorithm 1). The algorithm's output is $Z = TV(\tilde{Y_n}) = TV(\tilde{Y_n} \underset{\text{ind}}{\circ} S_n)$. Thus, we get

$$\begin{aligned} |d_{TV}(P,Q) - Z| &= \left| TV(Y_1 \underset{\text{ind}}{\circ} S_1) - TV(\tilde{Y_n} \underset{\text{ind}}{\circ} S_n) \right| \\ &= \left| \sum_{i=1}^{n} \left(TV(Y_i \underset{\text{ind}}{\circ} S_i) - TV(\tilde{Y_i} \underset{\text{ind}}{\circ} S_i) \right) \right|. \quad (3) \end{aligned}$$

We can now use the definition of the extension distance, followed by Theorem 3.10 to estimate the ith term above as follows: $\left| TV(Y_i \underset{\text{ind}}{\circ} S_i) - TV(\tilde{Y_i} \underset{\text{ind}}{\circ} S_i) \right| \le \text{ext}(Y_i, \tilde{Y_i}) \le \gamma + \delta TV(Y_i)$. Plugging in the values of γ and δ, we thus obtain

$$|d_{TV}(P,Q) - Z| \le \frac{\epsilon\Delta}{2} + \frac{\epsilon}{2n} \sum_{1 \le i \le n} TV(Y_i). \quad (4)$$

A direct inductive argument then shows that each Y_i is a conditional expectation of $R_1 \underset{\text{ind}}{\circ} R_2 \underset{\text{ind}}{\circ} \cdots \underset{\text{ind}}{\circ} R_i$. Applying items 2 and 3 of Lemma 3.6, we thus obtain, for each $1 \le i \le n$:

$$\begin{aligned} TV(Y_i) &\le TV(R_1 \underset{\text{ind}}{\circ} \cdots \underset{\text{ind}}{\circ} R_i) \\ &\le TV(R_1 \underset{\text{ind}}{\circ} \cdots \underset{\text{ind}}{\circ} R_i \underset{\text{ind}}{\circ} S_i) \\ &= TV(R_1 \underset{\text{ind}}{\circ} \cdots \underset{\text{ind}}{\circ} R_i \underset{\text{ind}}{\circ} R_{i+1} \underset{\text{ind}}{\circ} \cdots \underset{\text{ind}}{\circ} R_n) \\ &= d_{TV}(P,Q). \end{aligned}$$

Substituting this in eq. (4) and recalling that $\Delta \le d_{TV(P,Q)}$, we obtain the claimed result. □

Given its correctness, a runtime bound for Algorithm 1 can be estimated easily, and is as given in Theorem 2.5. An argument similar to (but simpler than) the proof of Theorem 3.11 also gives the following.

Lemma 3.12. *Fix $\gamma, \delta \in (0,1)$. For $1 \le i \le n$, let R_i be valid ratios defined on probability spaces $(\Omega_i, \mathcal{F}_i, P_i)$, and let $\tilde{R}_i$ be their corresponding (γ,δ)-discretizations. Then $\left| TV(R_1 \underset{\text{ind}}{\circ} R_2 \underset{\text{ind}}{\circ} \cdots \underset{\text{ind}}{\circ} R_n) - TV(\tilde{R}_1 \underset{\text{ind}}{\circ} \tilde{R}_1 \underset{\text{ind}}{\circ} \cdots \underset{\text{ind}}{\circ} \tilde{R}_n) \right| \le n\delta\kappa + n\gamma$, where $\kappa := \max_{1 \le i \le n}^{n} TV(R_i)$.*

Input: n pairs of probability distributions $(P_i, Q_i)_{i=1}^n$ on the finite set $[M]$, and a rational number $\epsilon \in (0,1)$.
Output: A number Z satisfying $(1-\epsilon)d_{TV}(P,Q) \le Z \le (1+\epsilon)d_{TV}(P,Q)$, where $P := P_1 \times P_2 \times \cdots \times P_n$ and $Q := Q_1 \times Q_2 \times \cdots \times Q_n$.

1 Set $\Delta \leftarrow \max_{1 \le i \le n} d_{TV}(P_i, Q_i)$;
2 Initialize $\gamma \leftarrow \frac{\epsilon\Delta}{2n}$ and $\delta \leftarrow \frac{\epsilon}{2n}$;
3 Set $R_i \leftarrow P_i \| Q_i$ for $1 \le i \le n$, and $Y_1 \leftarrow R_1$;
4 **for** $i \leftarrow 1$ **to** n **do**
5 Set $\tilde{Y_i}$ to be the (γ,δ)-discretization of Y_i;
6 If $i < n$, set $Y_{i+1} \leftarrow \tilde{Y_i} \underset{\text{ind}}{\circ} R_{i+1}$;
7 **end**
8 **return** $TV(\tilde{Y_n})$;

Algorithm 1: DisProdTV (Feng et al., 2024)

4 TV-DISTANCE BETWEEN GAUSSIANS

Consider Problem 1.1. Given two n-dimensional Gaussians $\mathcal{N}(\boldsymbol{\mu_1}, \Sigma_1), \mathcal{N}(\boldsymbol{\mu_2}, \Sigma_2)$, we want to approximate $D = d_{\text{TV}}(\mathcal{N}(\boldsymbol{\mu_1}, \Sigma_1), \mathcal{N}(\boldsymbol{\mu_2}, \Sigma_2))$ within relative error $(1 \pm \epsilon)$. Without loss of generality, we may further assume that both Σ_1 and Σ_2 have the full rank n. Otherwise, if $\mathbf{Rank}(\Sigma_1) \ne \mathbf{Rank}(\Sigma_2)$, then $D = 1$, since one of the distributions then has zero probability mass on the support of the other. If $\mathbf{Rank}(\Sigma_1) = \mathbf{Rank}(\Sigma_2) = r < n$, then consider the TV-distance between $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \Sigma_1)$ and $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu_2} - \boldsymbol{\mu_1}, \Sigma_2)$, which is the same as D. By Fact 2.2, $\boldsymbol{X}$ is supported on $\mathbf{Range}(\Sigma_1)$. There are two subcases: (1) If $\mathbf{Range}(\Sigma_2) + (\boldsymbol{\mu_2} - \boldsymbol{\mu}) \ne \mathbf{Range}(\Sigma_1)$, then $D = 1$, as in the previous case. (2) If $\mathbf{Range}(\Sigma_2) + (\boldsymbol{\mu_2} - \boldsymbol{\mu_1}) = \mathbf{Range}(\Sigma_1)$, let Π be an $r \times n$ matrix whose row vectors form an orthogonal basis of $\mathbf{Range}(\Sigma_1)$. Since Π is a bijection between $\mathbf{Range}(\Sigma_1)$ and $\mathbb{R}^r$, by Proposition 2.1, $d_{\text{TV}}(\boldsymbol{X}, \boldsymbol{Y}) = d_{\text{TV}}(\Pi\boldsymbol{X}, \Pi\boldsymbol{Y})$. However, we then have $\Pi\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \Pi\Sigma_1\Pi^T)$ and $\Pi\boldsymbol{Y} \sim \mathcal{N}(\Pi(\boldsymbol{\mu_2} - \boldsymbol{\mu_1}), \Pi\Sigma_2\Pi^T)$, so that they are r-dimensional Gaussian vectors whose covariance matrices have full rank r. Using (unnormalized) Gram-Schmidt orthogonalization, the process described above can be implemented using $O(n^3)$ arithmetic operations (and also in polynomial-time in terms of bit complexity: see, e.g., Grötschel et al. (1993, Section 1.4)). From now on, we always assume that Σ_1 and Σ_2 have the full rank n.

Reduction to Product Distributions We now use standard linear algebraic properties to reduce to the case of product distributions. This is the only part of our algorithm that needs access to a diagonalization sub-routine. (As discussed in the introduction, Section 7 in the supplementary material describes how to implement a version of the next lemma in the bit complexity model.)

Lemma 4.1. *There exists an algorithm which given* $(\boldsymbol{\mu_1}, \Sigma_1)$ *and* $(\boldsymbol{\mu_2}, \Sigma_2)$ *(where* Σ_1, Σ_2 *are* $n \times n$ *positive definite matrices and* $\boldsymbol{\mu_1}, \boldsymbol{\mu_2}$ *are vectors in* $\mathbb{R}^n$*), outputs* $(\boldsymbol{\mu}, \Sigma)$*, using* $O(n^3)$ *arithmetic operations and two diagonalizations of* $n \times n$ *symmetric matrices, such that* Σ *is a diagonal matrix and* $d_{\mathrm{TV}}\left(\mathcal{N}(\boldsymbol{\mu_1}, \Sigma_1), \mathcal{N}(\boldsymbol{\mu_2}, \Sigma_2)\right) = d_{\mathrm{TV}}\left(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}(\mathbf{0}, I_n)\right)$.

Proof. Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ and $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$. We use diagonalization to decompose Σ_2 as $Q_2 \Lambda_2 Q_2^T$, where Q_2 is an orthogonal matrix and Λ_2 is a diagonal matrix. Define the matrix $A := Q_2 \Lambda_2^{-1/2} Q_2^T$. Now, we diagonalize the matrix $A \Sigma_1 A^T$ to obtain the decomposition $Q_1 \Lambda_1 Q_1^T$ (where Q_1 is orthogonal and Λ_1 is diagonal). Define $\hat{\boldsymbol{X}} = Q_1^T A(\boldsymbol{X} - \boldsymbol{\mu_2})$ and $\hat{\boldsymbol{Y}} = Q_1^T A(\boldsymbol{Y} - \boldsymbol{\mu_2})$. Note that the transformation $\boldsymbol{v} \mapsto Q_1^T A(\boldsymbol{v} - \boldsymbol{\mu_2})$ is invertible. Thus, by Proposition 2.1, $d_{\mathrm{TV}}(\boldsymbol{X}, \boldsymbol{Y}) = d_{\mathrm{TV}}\left(\hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}}\right)$. Finally, $\hat{\boldsymbol{X}} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $\hat{\boldsymbol{Y}} \sim \mathcal{N}(\mathbf{0}, I_n)$, where $\boldsymbol{\mu} = Q_1^T A(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ and $\Sigma = \Lambda_1$. Apart from the two diagonalization operations, the algorithm only performs a constant number of matrix multiplications that cost a total of $O(n^3)$ arithmetic operations. □

Discretizing Product Gaussian Distributions With Lemma 4.1, we only need to approximate the TV-distance between the Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with diagonal covariance matrix Σ and the standard Gaussian $\mathcal{N}(0, I_n)$; both of which are continuous product distributions. We give the following reduction algorithm that transforms this task into the task of estimating the TV-distance between two *discrete* product distributions.

Lemma 4.2. *There exists a deterministic algorithm, which given a vector* $\boldsymbol{\mu} \in \mathbb{R}^n$*, a positive* $n \times n$ *diagonal matrix* Σ*, and* $\epsilon > 0$*, outputs* $2n$ *discrete distributions* $(\tilde{P}_i, \tilde{Q}_i)_{1 \le i \le n}$*, using at most* $O(\frac{n^2}{\epsilon} \log^3 \frac{n}{\epsilon d_{\mathrm{TV}}(\boldsymbol{X}, \boldsymbol{Y})})$ *arithmetic operations, and satisfying the following properties (here* $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ *and* $\boldsymbol{Y} \sim \mathcal{N}(\mathbf{0}, I)$ *are Gaussian random vectors):*

- *every* $\tilde{P}_i$ *and* $\tilde{Q}_i$ *is defined over the domain* $[M]$ *with* $M = O(\frac{n}{\epsilon} \log \frac{n}{\epsilon d_{\mathrm{TV}}(X, Y)})$*;*
- *the product distributions* $\tilde{P} = \tilde{P}_1 \times \ldots \times \tilde{P}_n$ *and* $\tilde{Q} = \tilde{Q}_1 \times \ldots \times \tilde{Q}_n$ *satisfy* $\left| d_{\mathrm{TV}}\left(\tilde{P}, \tilde{Q}\right) - d_{\mathrm{TV}}(X, Y) \right| \le \frac{\epsilon}{3} d_{\mathrm{TV}}(\boldsymbol{X}, \boldsymbol{Y})$.

We now proceed to prove this lemma. We start with the following estimate. For each $i \in [n]$, $X_i \sim \mathcal{N}(\mu_i, \Sigma_{i,i})$ and $Y_i \sim \mathcal{N}(0, 1)$, define $\Delta_i := \frac{1}{200} \min\{1, \max\{|\Sigma_{i,i} - 1|, 40|\mu_i|\}\}$ and $\Delta := \max_{1 \le i \le n} \Delta_i$. Then, one has the following estimate from previous works (see the supplementary material for a proof).

Lemma 4.3. $\Delta \le d_{\mathrm{TV}}(\boldsymbol{X}, \boldsymbol{Y}) \le 10^4 n \Delta$.

Define R_i to be the ratio $\mathcal{N}(\mu_i, \Sigma_{ii}) \| \mathcal{N}(0, 1)$ defined on the probability space $(\mathbb{R}, \mathcal{B}, \mathcal{N}(0, 1))$. We set

$$\gamma = \frac{\epsilon \Delta}{50n}, \qquad \delta = \frac{\epsilon}{50n}, \tag{5}$$

and let $\tilde{R}_i$ denote the (γ, δ)-discretization of R_i, for each $1 \le i \le n$. Now, let P_i (respectively, Q_i) denote the probability distribution on the finite set (see Definition 3.7) $\mathcal{I}_{\gamma,\delta}$ defined by $P_i(J) = \mathbb{E}[R_i \cdot I[R_i \in J]]$ (respectively, $Q_i(J) = \mathbb{P}[R_i \in J]$) for each interval $J \in \mathcal{I}_{\gamma,\delta}$. Since $\tilde{R}_i = \mathbb{E}[R_i | \mathcal{I}_{\gamma,\delta}(R_i)]$, it then follows that $\tilde{R}_i = P_i \| Q_i$.

The exact computation of the discrete distribution P_i and Q_i involves computing probabilities and expectations of Gaussian variables, which, in turn, amounts to evaluation of the error function. The following lemma shows that we can compute an approximate version of these distributions.

Lemma 4.4. *There exists an algorithm which, for each* $i \in [n]$*, computes a pair of discrete distributions* $(\tilde{P}_i, \tilde{Q}_i)$ *over* $[M]$*, where* $M = O(\frac{n}{\epsilon} \log \frac{n}{\epsilon \Delta})$*, using* $O(\frac{n}{\epsilon} \log^3 \frac{n}{\epsilon \Delta})$ *arithmetic operations such that* $d_{\mathrm{TV}}(\tilde{P}_i, P_i) \le \frac{\epsilon \Delta}{50n}$ *and* $d_{\mathrm{TV}}(\tilde{Q}_i, Q_i) \le \frac{\epsilon \Delta}{50n}$.

Proof. Fix an index $1 \le i \le n$. Let $f(x) = \frac{1}{\sqrt{2\pi \Sigma_{i,i}}} e^{-(x-\mu_i)^2/(2\Sigma_{i,i})}$ denote the density function of $\mathcal{N}(\mu_i, \Sigma_{i,i})$ and $g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ denote the density function of $\mathcal{N}(0, 1)$. Recall that $\gamma = \frac{\epsilon \Delta}{50n}, \delta = \frac{\epsilon}{50n}$.

Let $M = |\mathcal{I}_{\gamma,\delta}|$, so that, by definition of γ, δ and $\mathcal{I}_{\gamma,\delta}$,

$$M = O(\log(1/\gamma)/\log(1+\delta)) = O\left(\frac{n}{\epsilon} \log \frac{n}{\epsilon \Delta}\right).$$

For any $J \in \mathcal{I}_{\gamma,\delta}$, we then have, by definition,

$$P_i(J) = \mathbb{E}[R_i \cdot I[R_i \in J]] = \int\limits_{x \in D(J)} f(x) dx, \text{ and} \tag{6}$$

$$Q_i(J) = \mathbb{P}[R_i \in J] = \int\limits_{x \in D(J)} g(x) dx, \tag{7}$$

where the integration domain $D(J)$ is defined by $D(J) := \{R_i \in J\} = \{z \mid \frac{f(z)}{g(z)} \in J\}$. If $a \le b$ are

the endpoints of J then the set[4] $D(Z)$ is the set of all z satisfying $\ln\left(a\sqrt{\Sigma_{i,i}}\right) < -\frac{(z-\mu_i)^2}{2\Sigma_{i,i}} + \frac{z^2}{2} < \ln\left(b\sqrt{\Sigma_{i,i}}\right)$. By solving the quadratic inequality, we see that $D(J)$ is a union of at most two intervals. From eqs. (6) and (7), we thus obtain that the distributions P_i and Q_i can be computed exactly if one has an exact oracle for computing the error function $\mathrm{erf}(\cdot)$. However, since we can only compute such integrals approximately, we compute two distributions $\tilde{P}_i$ and $\tilde{Q}_i$ over $[M]$ that approximate P_i and Q_i respectively.

We show how to compute $\tilde{P}_i$. The distribution $\tilde{Q}_i$ can be computed similarly. Let $\zeta = \frac{1}{10M} \cdot \frac{\epsilon\Delta}{50n}$. For each $J \in \mathcal{I}_{\gamma,\delta}$, we need to compute $\int_{x\in D(J)} f(x)dx$ where $D(J)$ is a union of at most two intervals. We use Proposition 2.3 to compute the integration over each interval with additive error $\zeta/2$. By adding the results together, we obtain a number $p_J \geq 0$ such that $|p_J - P_i(J)| \leq \zeta$. One may consider defining $\tilde{P}_i$ as the distribution such that for any $J \in \mathcal{I}_{\gamma,\delta}$, $\tilde{P}_i(J) = p_J$. However, it may not hold that $\sum_{J\in\mathcal{I}_{\gamma,\delta}} p_J = 1$. Note, however, that $|\sum_J p_J - 1| \leq M\zeta$. We thus make the following adjustments on the vector $(p_J)_{J\in\mathcal{I}_{\gamma,\delta}}$ to define the distribution $\tilde{P}_i$: (1) If $\sum_J p_J < 1$, we add $1 - \sum_J p_J$ probability mass to $\tilde{P}_i([1,1])$. (2) If $\sum_J p_J > 1$, we remove $\sum_J p_J - 1$ probability mass in total for $\tilde{P}_i$ to make $\tilde{P}_i$ a valid distribution. We can then bound the TV-distance as follows $d_{\mathrm{TV}}\left(\tilde{P}_i, P_i\right) = \frac{1}{2}\sum_{J\in\mathcal{I}_{\gamma,\delta}} \left|\tilde{P}_i(J) - P_i(J)\right| \leq \frac{1}{2}\sum_{J\in\mathcal{I}_{\gamma,\delta}} \left(|\tilde{P}_i(J) - p_J| + |p_J - P_i(J)|\right) \leq M\zeta < \frac{\epsilon\Delta}{50n}$. By Proposition 2.3, the number of arithmetic operations needed for computing each p_J is $O(\log^2(1/\zeta))$. The total number of arithmetic operations is therefore $O(M\log^2(1/\zeta)) = O(\frac{n}{\epsilon}\log^3\frac{n}{\epsilon\Delta})$. □

Lemma 4.5. *Let $\boldsymbol{X}, \boldsymbol{Y}, P_i, Q_i$ be as above. For any distributions $(\tilde{P}_i, \tilde{Q}_i)_{1\leq i\leq n}$ with $d_{\mathrm{TV}}\left(\tilde{P}_i, P_i\right) \leq \frac{\epsilon\Delta}{50n}$ and $d_{\mathrm{TV}}\left(\tilde{Q}_i, Q_i\right) \leq \frac{\epsilon\Delta}{50n}$, define the product distributions $\tilde{P} = \tilde{P}_1 \times \ldots \times \tilde{P}_n$ and $\tilde{Q} = \tilde{Q}_1 \times \ldots \times \tilde{Q}_n$. Then*

$$\left|d_{\mathrm{TV}}\left(\tilde{P}, \tilde{Q}\right) - d_{\mathrm{TV}}\left(\boldsymbol{X}, \boldsymbol{Y}\right)\right| \leq \frac{\epsilon}{3} d_{\mathrm{TV}}\left(\boldsymbol{X}, \boldsymbol{Y}\right).$$

Proof. The following proof uses the extension distance based generalization of the framework of Feng et al. (2024) that was developed in Section 3 above. Define the product distributions $P = P_1 \times \cdots \times P_n$ and $Q = Q_1 \times \cdots \times Q_n$. A simple coupling argument then shows that $D_P := d_{\mathrm{TV}}\left(P, \tilde{P}\right) \leq \frac{\epsilon\Delta}{50}$ and $D_Q := d_{\mathrm{TV}}\left(Q, \tilde{Q}\right) \leq \frac{\epsilon\Delta}{50}$. By triangle inequality for d_{TV},

$$\left|d_{\mathrm{TV}}\left(\tilde{P}, \tilde{Q}\right) - d_{\mathrm{TV}}\left(P, Q\right)\right| \leq D_P + D_Q \leq \frac{\epsilon\Delta}{25}. \quad (8)$$

[4]Depending on the definition of J, some $<$ signs should be replaced with $\leq$, but that does not affect the result of integration in this case.

Next, we claim the following result,

$$\left|d_{\mathrm{TV}}\left(P, Q\right) - d_{\mathrm{TV}}\left(X, Y\right)\right| \leq \frac{\epsilon}{25} d_{\mathrm{TV}}\left(X, Y\right). \quad (9)$$

To see this, recall that each $\tilde{R}_i = P_i \| Q_i$ is a (γ,δ)-discretization (with γ, δ as above) of $R_i = \mathcal{L}_{Y_i} \| \mathcal{L}_{X_i}$ (where $\mathcal{L}_Z$ denotes the law of the random variable Z). It thus follows from Fact 3.4 and Definition 3.5 that $d_{\mathrm{TV}}(P, Q) = TV(\tilde{R}_1 \underset{\text{ind}}{\circ} \tilde{R}_2 \underset{\text{ind}}{\circ} \ldots \underset{\text{ind}}{\circ} \tilde{R}_n)$ while $d_{\mathrm{TV}}(\boldsymbol{X}, \boldsymbol{Y}) = TV(R_1 \underset{\text{ind}}{\circ} R_2 \underset{\text{ind}}{\circ} \ldots \underset{\text{ind}}{\circ} R_n)$. Eq. (9) then is a direct consequence of Lemma 3.12, combined with the facts that $\Delta \leq d_{\mathrm{TV}}\left(\boldsymbol{X}, \boldsymbol{Y}\right)$ and that for each i, $TV(R_i) = d_{\mathrm{TV}}(X_i, Y_i) \leq d_{\mathrm{TV}}\left(\boldsymbol{X}, \boldsymbol{Y}\right)$. The lemma then follows by combining (8) and (9) with $\Delta \leq d_{\mathrm{TV}}(\boldsymbol{X}, \boldsymbol{Y})$. □

Proof of Lemma 4.2. We apply the algorithm of Lemma 4.4 to each dimension to obtain the discretized distributions $\tilde{P}_i$ and $\tilde{Q}_i$. The running time can be verified by Lemma 4.3 and Lemma 4.4, while the error bound is given by Lemma 4.5. □

The Final Algorithm By Lemma 4.1 and Lemma 4.2, to solve MultGaussianTV, we only need to approximate the TV-distance between the discrete product distributions $\tilde{P}, \tilde{Q}$ constructed in Lemma 4.2. Since $\tilde{P}, \tilde{Q}$ are discrete, the resulting problem can now be solved using existing algorithms in Theorem 2.5 and Theorem 2.6. Note that our reduction algorithm is deterministic and uses $O(n^3 + \frac{n^2}{\epsilon}\log^3\frac{n}{\epsilon D})$ operations (in addition to diagonalizations of two $n \times n$ matrices), where D is the TV-distance between the two input Gaussian distributions. This gives both Theorems 1.2 and 1.3.

5 CONCLUSION

We give an efficient algorithm for approximating the TV-distance between two multivariate Gaussians with relative error. Along the way, we extend the analysis framework of Feng et al. (2024) from discrete probability distributions to general probability measures. Several directions remain open; including TV distance estimation for general log-concave distributions, graphical models, and Gaussian-perturbed distributions; and approximations for other notions of distance such as the Wasserstein distance. A more open-ended question is to find as yet unexplored applications of algorithms for estimating total variation distances, especially given the role constant factor approximations for total variation distance between Gaussians (discussed above) have played in various applications: see, e.g., Arbas et al. (2023) for an example and Davies et al. (2022) for references.

Acknowledgments

PS acknowledges Rikhav Shah for helping with technical questions about the results of Shah (2024). AB and PS acknowledge the Mathematics of Data program at the Institute for Mathematical Sciences, National University of Singapore, where this project was initiated. PS acknowledgements support from the Department of Atomic Energy, Government of India, under project no. RTI4001, from the Infosys-Chandrasekharan virtual center for Random Geometry, and from SERB MATRICS grant number MTR/2023/001547. The contents of this paper do not necessarily reflect the views of the funding agencies listed above.

References

Arbas, J., Ashtiani, H., and Liaw, C. (2023). Polynomial time and private learning of unbounded Gaussian mixture models. In *Proc. 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 1018–1040. PMLR.

Armentano, D., Beltrán, C., Bürgisser, P., Cucker, F., and Shub, M. (2018). A stable, polynomial-time algorithm for the eigenpair problem. *Journal of the European Mathematical Society*, 20(6):1375–1437.

Banks, J., Garza-Vargas, J., Kulkarni, A., and Srivastava, N. (2023). Pseudospectral shattering, the sign function, and diagonalization in nearly matrix multiplication time. *Foundations of Computational Mathematics*, 23(6):1959–2047.

Barabesi, L. and Pratelli, L. (2024). An inequality for the total variation distance between high-dimensional centered Gaussian laws. *Statistics & Probability Letters*, 211:110148.

Bhattacharyya, A., Gayen, S., Meel, K. S., Myrisiotis, D., Pavan, A., and Vinodchandran, N. V. (2023). On approximating total variation distance. In *Proc. 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3479–3487. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Bhattacharyya, A., Gayen, S., Meel, K. S., Myrisiotis, D., Pavan, A., and Vinodchandran, N. V. (2024). Total variation distance meets probabilistic inference. In *Proc. 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 3776–3794. PMLR.

Bhattacharyya, A., Gayen, S., Meel, K. S., and Vinodchandran, N. V. (2020). Efficient distance approximation for structured high-dimensional distributions via learning. In *Proc. 2020 Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 33 of *Advances in Neural Information Processing Systems*, pages 14699–14711. Curran Associates, Inc.

Borwein, J. M. and Borwein, P. B. (1987). *Pi and the AGM: A study in analytic number theory and computational complexity*. Canadian Mathematical Society Series of Monographs and Advanced Texts = Monographies et Études de La Société Mathématique Du Canada. Wiley, New York.

Bouland, A., Chen, L., Holden, D., Thaler, J., and Vasudevan, P. N. (2020). On the power of statistical zero knowledge. *SIAM Journal on Computing*, 49(4):FOCS17–1.

Bürgisser, P. and Cucker, F. (2013). *Condition: The geometry of numerical algorithms*, volume 349 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg.

Chan, H. H., Chan, S. H., and Liu, Z. (2004). Domb's numbers and Ramanujan–Sato type series for $1/\pi$. *Advances in Mathematics*, 186(2):396–410.

Chevillard, S. (2012). The functions erf and erfc computed with arbitrary precision and explicit error bounds. *Information and Computation*, 216:72–95.

Davies, S., Mazumdar, A., Pal, S., and Rashtchian, C. (2022). Lower bounds on the total variation distance between mixtures of two Gaussians. In *Proc. 33rd International Conference on Algorithmic Learning Theory (COLT)*, pages 319–341. PMLR.

Demmel, J. W. (1997). *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics.

Devroye, L., Mehrabian, A., and Reddad, T. (2023). The total variation distance between high-dimensional Gaussians with the same mean. arXiv:1810.08693.

Dey, P., Kannan, R., Ryder, N., and Srivastava, N. (2023). Bit complexity of Jordan normal form and polynomial spectral factorization. In Kalai, Y. T., editor, *Proc. 14th Innovations in Theoretical Computer Science Conference (ITCS)*, volume 251 of *LIPIcs*, pages 42:1–42:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Durrett, R. (2019). *Probability: Theory and examples*. Cambridge University Press, Cambridge, UK, fifth edition.

Feng, W., Guo, H., Jerrum, M., and Wang, J. (2023). A simple polynomial-time approximation algorithm for the total variation distance between two product distributions. *TheoretiCS*, 2. URL: https://doi.org/10.46298/theoretics.23.7.

Feng, W., Liu, L., and Liu, T. (2024). On deterministically approximating total variation distance. In

Proc. 2024 ACM-SIAM Symposium on Discrete Algorithms, (SODA), pages 1766–1791. SIAM.

Grötschel, M., Lovász, L., and Schrijver, A. (1993). *Geometric algorithms and combinatorial optimization*, volume 2 of *Algorithms and Combinatorics*. Springer Berlin Heidelberg, Berlin, Heidelberg, second corrected edition edition.

Kiefer, S. (2018). On computing the total variation distance of hidden Markov models. In *Proc. 45th International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 107 of *LIPIcs*, pages 130:1–130:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Preparata, F. P. and Shamos, M. I. (1985). *Computational Geometry: An Introduction*. Texts and Monographs in Computer Science. Springer-Verlag.

Ramanujan, S. (1914). Modular equations and approximations to π. *Quarterly Journal of Pure and Applied Mathematics*, 45:350–372.

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

Shah, R. (2024). Fast Hermitian diagonalization with nearly optimal precision. arXiv:2408.09880.

Shah, R. (2025). Hermitian diagonalization in linear precision. In *Proc. 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Proceedings, pages 5599–5615. SIAM.

Srivastava, N. (2023). The complexity of diagonalization. In *Proc. 2023 International Symposium on Symbolic and Algebraic Computation (ISSAC)*, pages 1–6, New York, NY, USA. ACM.

Tao, L., Xu, S., Wang, C.-H., Suh, N., and Cheng, G. (2024). Discriminative estimation of total variation distance: A fidelity auditor for generative data. arXiv:2405.15337.

Williams, D. (1991). *Probability with martingales*. Cambridge University Press.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Not Applicable**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**
   (b) Complete proofs of all theoretical results. **Yes**
   (c) Clear explanations of any assumptions. **Yes**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Not Applicable**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Not Applicable**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Not Applicable**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Not Applicable**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Not Applicable**
   (b) The license information of the assets, if applicable. **Not Applicable**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable**
   (d) Information about consent from data providers/curators. **Not Applicable**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**

SUPPLEMENTARY MATERIAL

Structure of the Supplementary Material For clarity of reference, the section numbers in the supplementary material follow on from the section numbers in the main paper. Section 6 provides the omitted proof of Proposition 2.3 on computational estimation of the error function, and also includes a sketch of how to carry out the proof of Lemma 4.4 in a weaker computational model where logarithm and square root cannot be computed exactly. Section 7 shows how to carry out the reduction to the product case (described in Section 4) assuming access only to an approximate diagonaliztion oracle that can be implemented using floating point arithmetic. Finally, Section 8 contains proofs omitted from Section 3, while Section 9 contains proofs omitted from Section 4.

6 PROPERTIES OF THE ERROR FUNCTION

We start with the following proposition, which follows from ideas reported by Chevillard (2012). Here, we denote by $\mathrm{len}\,(x)$ the *representation length* in bits of a rational x, when it is specified by its numerator and denominator. Later, we also use the same notation $\mathrm{len}\,(S)$ to denote the total representation length of a collection of rational numbers.

Proposition 6.1. *There exists an algorithm such that given any small $0 < \epsilon \le 1/2$ and $x > 0$, it returns a real number $y \ge 0$ satisfying $|y - \mathrm{erf}\,(x)\,| \le \epsilon$ using $O(\log^2(1/\epsilon))$ arithmetic operations. Further, for rational inputs, these operations may be assumed to be performed on rational numbers of representation length at most $\tilde{O}(\mathrm{len}\,(x) \cdot \log^2(1/\epsilon))$.*

Proof of Proposition 6.1. Without loss of generality, we assume $\epsilon \le 10^{-4}$ (if the input ϵ is larger, we replace it with 10^{-4}). If $x > \ln(1/\epsilon)$, we simply return the number $y = \mathrm{erf}\,(+\infty) = 1$. The error can be bounded as follows

$$\begin{aligned} |y - \mathrm{erf}\,(x)\,| &\le \frac{2}{\sqrt{\pi}} \int_{t > \ln \frac{1}{\epsilon}} e^{-t^2} dt \\ &\le \frac{2}{\sqrt{\pi}} \int_{t > \ln \frac{1}{\epsilon}} e^{-t \ln \frac{1}{\epsilon}} dt = \frac{2}{\sqrt{\pi}} \frac{1}{\ln \frac{1}{\epsilon}} \exp\left(-\ln^2 \frac{1}{\epsilon}\right). \end{aligned}$$

Since ϵ is a small constant, the last term is at most ϵ.

Now, we assume $x < \ln(1/\epsilon)$. The following equation can be found, e.g. in the work of Chevillard (2012, eq. (1)):

$$\mathrm{erf}\,(x) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)n!}.$$

Let $N = 10\lceil \ln(1/\epsilon) \rceil^2$. We compute $y = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{N} (-1)^n \frac{x^{2n+1}}{(2n+1)n!}$ using $O(\log^2(1/\epsilon))$ arithmetic operations on rational numbers of representation length at most $\tilde{O}(\mathrm{len}\,(x) \cdot \log^2(1/\epsilon))$ (at present, we ignore the small contribution to the error that comes from having to use rational approximations of $\sqrt{\pi}$). To bound the error, we note that $n! \ge (n/e)^n$ so that

$$\begin{aligned} |y - \mathrm{erf}\,(x)\,| &\le \frac{2}{\sqrt{\pi}} \sum_{n > N} \frac{x^{2n+1}}{(2n+1)n!} \\ &< \frac{2}{\sqrt{\pi}} \sum_{n \ge N} \frac{x^{2n+1} e^n}{n^n} \\ &< \frac{2x}{\sqrt{\pi}} \sum_{n > N} \left(\frac{x^2 e}{n}\right)^n. \end{aligned}$$

By the choice of N and the fact $x < \ln(1/\epsilon)$, $\frac{x^2 e}{n} \le \frac{x^2 e}{N} < \frac{1}{e}$. Hence,

$$|y - \mathrm{erf}\,(x)\,| \le \frac{2 \ln(1/\epsilon)}{\sqrt{\pi}} \sum_{n > N} e^{-n} < \epsilon/2,$$

where the last inequality holds because ϵ is sufficiently small. Finally, if $y < 0$, we let $y = 0$, which only decreases the error. This proves the first claim in the proposition.

It remains only to account for the error arising from having to use a rational approximation to $1/\sqrt{\pi}$, when working with rational numbers only. Propagating such an error through the above analysis shows that it is sufficient for our purpose to compute $1/\pi$ up to an *additive* error of $\epsilon/10$. This can be accomplished using $O(\log^2(1/\epsilon))$ arithmetic operations on rational numbers of representation length at most $\tilde{O}(\log^2(1/\epsilon))$ using well-known series for the approximation of $1/\pi$; see, e.g., eq. (1) of Chan et al. (2004) and the work of Ramanujan (1914) and Borwein and Borwein (1987) for the original results. □

Proof of Proposition 2.3. First, let $a \leftarrow (a-\mu)/\sqrt{2\sigma^2}$, $b \leftarrow (b-\mu)/\sqrt{2\sigma^2}$ and $f(t) = \frac{1}{\sqrt{\pi}} e^{-t^2}$. The value of the integration remains unchanged. If $ab \leq 0$, then $\int_a^b f(t)dt = \frac{1}{2}(\operatorname{erf}(|a|) + \operatorname{erf}(|b|))$. If $ab > 0$, say $a > 0$ and $b > 0$, then $\int_a^b f(t)dt = \frac{1}{2}(\operatorname{erf}(b) - \operatorname{erf}(a))$.[5] Using Proposition 6.1, we compute the value of each error function with additive error ϵ, which, by substitution in the above, gives us an additive error of at most ϵ for the integrals. Finally, if our answer $y < 0$, we let $y = 0$, which only decreases the error. From Proposition 6.1, the runtime is as claimed. □

Note that the proof of Proposition 2.3 using Proposition 6.1 also gives the corresponding bit complexity result, except for the extra factor of $1/\sqrt{2}$ that appears in the computation of the approximation y as defined in Proposition 6.1. This factor can be handled as the same (in fact simpler) way as the factor of $1/\sqrt{\pi}$ was handled in the proof of Proposition 6.1.

6.1 Proof of Lemma 4.4 in the bit complexity model

The algorithm in Lemma 4.4 needs to solve the following quadratic inequality.

$$\ln\left(a\sqrt{\Sigma_{i,i}}\right) < -\frac{(z-\mu_i)^2}{2\Sigma_{i,i}} + \frac{z^2}{2} < \ln\left(b\sqrt{\Sigma_{i,i}}\right).$$

The solution is a union of at most two intervals. One can solve for the endpoints of these intervals exactly if the computational model can exactly compute square roots and logarithms.

In the bit complexity model, we can only find these endpoints approximately. Let c and d be the two endpoints of one of the intervals. Then, using standard approximations for the logarithm, we can obtain two approximate solutions $\tilde{c}$ and $\tilde{d}$ such that $|c - \tilde{c}| \leq \delta$ and $|d - \tilde{d}| \leq \delta$ for some error bound $\delta \in (0,1)$, using $\operatorname{poly}(\operatorname{len}(\delta), \operatorname{len}(\Sigma_{ii}), \operatorname{len}(\mu_i))$ bit operations. Let $f(x) = \frac{1}{\sqrt{2\pi\Sigma_{i,i}}} e^{-(x-\mu_i)^2/(2\Sigma_{i,i})}$ be the density function in the proof of Lemma 4.4. The error can be controlled as

$$\left|\int_{\tilde{c}}^{\tilde{d}} f(x)dx - \int_c^d f(x)dx\right| \leq 2\delta \sup_x f(x)$$
$$= 2\delta f(\mu_i) = \delta\sqrt{\frac{2}{\pi\Sigma_{i,i}}}.$$

Similarly, for $g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, the error is at most $\delta\sqrt{\frac{2}{\pi}}$. One can take δ such that $\delta\sqrt{\frac{2}{\pi \min(1,\Sigma_{i,i})}} = O(\frac{\epsilon\Delta}{Mn})$. Then the rest of the proof of Lemma 4.4 works exactly as before even in this computational model. The running time for solving the inequality in this model is $\operatorname{poly}(\log(Mn), \operatorname{len}(\epsilon), \operatorname{len}(\Sigma_{ii}), \operatorname{len}(\mu_i), \operatorname{len}(\Delta)))$.

7 ALGORITHM WITH APPROXIMATE DIAGONALIZATION ORACLE

In the main paper (specifically, in Lemma 4.1) we assumed the existence of an oracle that computes the *exact* orthogonal diagonalization of a real symmetric matrix. It is common to work in such an exact real arithmetic model for ease of presentation and in order to understand the high-level properties of numerical algorithms.

[5] The case $a < 0$ and $b < 0$ follows by the symmetry of $\operatorname{erf}(\cdot)$.

However, such a model of computation is not realistic (see, e.g., Dey et al. (2023) for a discussion). More realistic models include floating point arithmetic, or a model of exact arithmetic with rationals represented as ratios of integers.

In this section, we show that the algorithm in Lemma 4.1 remains efficient even in the bit complexity model. We do so by showing that this algorithm can be implemented given access only to an appropriate *approximate* diagonalization oracle.

The oracle we work with is the following. Given a small enough positive δ and a real symmetric positive definite matrix $A \in \mathbb{Q}^{n\times n}$ such that the input size is $\operatorname{len}(A)$, we assume that the approximate diagonalization oracle returns, using at most $\operatorname{poly}(n, \operatorname{len}(A), \operatorname{len}(\delta))$ bit operations, an $n \times n$ diagonal matrix $\Lambda^{1/2}$ and an $n \times n$ matrix Q, all whose singular values lie in the interval $[1-\delta/3, 1+\delta/3]$, such that

$$\|A - Q\Lambda Q^T\| \leq \delta\|A\|. \tag{10}$$

The existence of such an oracle in the floating point model of computation (and hence also in the bit complexity model) is known: see Shah (2024, Theorem 4.2) for a state-of-the-art randomized algorithm implementing such an oracle in the complex Hermitian case, and the earlier paper by Banks et al. (2023, Remark 6.1) for an explicit instantiation in the case of real Hermitian matrices.[6] Note also that it follows from the condition on Q that $\|Q^TQ - I\| \leq \delta$ (when δ is small enough).

We now show that there exists a polynomial-time algorithm that performs a similar reduction as in Lemma 4.1, while only accessing the above approximate diagonalization oracle.

For any real symmetric positive definite matrix A, define the condition number

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)},$$

where $\lambda_{\min}(A), \lambda_{\max}(A) > 0$ are the minimum and maximum eigenvalues of A.

Theorem 7.1. *There exists an algorithm such that given $\epsilon > 0$, $(\boldsymbol{\mu_1}, \Sigma_1)$, $(\boldsymbol{\mu_2}, \Sigma_2)$ (where Σ_1, Σ_2 are $n\times n$ positive definite matrices in $\mathbb{Q}^{n\times n}$ and $\boldsymbol{\mu_1}, \boldsymbol{\mu_2}$ are vectors in $\mathbb{Q}^n$),a lower bound Δ with $D = d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1), \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)) \geq \Delta$, and an upper bound $\kappa > 1$ with $\kappa(\Sigma_1), \kappa(\Sigma_2) \leq \kappa$, it returns $\boldsymbol{\mu} \in \mathbb{Q}^n$ and a diagonal positive definite $\Sigma \in \mathbb{Q}^{n\times n}$ such that*

$$|D - d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}(0, I_n))| \leq \frac{\epsilon}{100} D,$$

The algorithm uses $O(n^3)$ arithmetic operations and two approximate diagonalizations of $n \times n$ symmetric positive definite matrices of representation length $\operatorname{poly}\left(\frac{n}{\epsilon}, \operatorname{len}((\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \boldsymbol{\Sigma_1}, \boldsymbol{\Sigma_2}))\right)$ with error bound $\delta = \operatorname{poly}\left(\frac{\epsilon\Delta}{n\kappa}\right)$.

Given the $\boldsymbol{\mu}$ and Σ returned by the algorithm in Theorem 7.1, one can use the discretization procedure in the main paper to approximate the TV-distance between $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $\mathcal{N}(0, I_n)$ within relative error $\epsilon/2$, which, by Theorem 7.1, approximates D within relative error ϵ.

The algorithm in Theorem 7.1 needs to know a lower bound parameter Δ and an upper bound parameter κ in advance. If, however, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and positive definite Σ_1, Σ_2 (assume $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ or $\Sigma_1 \neq \Sigma_2$, otherwise $D = 0$) have rational entries with numerators and denominators that are all integers represented by at most $\operatorname{poly}(n)$ bits, then $\kappa = 2^{\operatorname{poly}(n)}$ (see Srivastava (2023); Dey et al. (2023)) and $\Delta = 2^{-\operatorname{poly}(n)}$ (see Lemma 7.6 below). The overall running time is thus $\operatorname{poly}\left(\frac{n}{\epsilon}, \operatorname{len}((\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \boldsymbol{\Sigma_1}, \boldsymbol{\Sigma_2})), \operatorname{len}(\delta)\right) = \operatorname{poly}\left(\frac{n}{\epsilon}, \operatorname{len}((\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \boldsymbol{\Sigma_1}, \boldsymbol{\Sigma_2}))\right)$.

The following standard upper bound on TV-distance will be used in the proof of Theorem 7.1. It follows from the exact expression for the KL divergence between Gaussians, and Pinkser's inequality.

Fact 7.2 (see, e.g., Devroye et al. (2023, Proposition 2.1)). *Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^n$ and $\Sigma_1, \Sigma_2 \in \mathbb{R}^{n\times n}$ be two covariance matrices with rank n. The TV-distance between $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$ is at most*

$$\frac{1}{2}\sqrt{\operatorname{tr}(\Sigma_1\Sigma_2^{-1}) - n + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \ln\frac{\det(\Sigma_1)}{\det(\Sigma_2)}}.$$

[6] The oracles in the papers by Shah (2024) and Banks et al. (2023) return the diagonal matrix Λ directly. However, for ease of presentation, we include the entry-wise square root of the diagonal matrix Λ into the oracle itself. In the floating point model, this entry-wise square root only needs to be computed to an additive accuracy of $O(\delta\|A\|)$, and this can be done without significantly modifying the runtime characteristics of the oracle.

In particular, if $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\Sigma_2 = I_n$, then the TV-distance is at most $\frac{1}{2}\sqrt{\mathrm{tr}(\Sigma_1) - n + \ln\det(\Sigma_1)}$.

In the proof of Theorem 7.1, we will repeatedly use the following standard facts from linear algebra.

Fact 7.3 (Weyl's inequality, see, e.g., Demmel (1997, Theorem 5.1)). *If A and B are $n \times n$ real symmetric matrices with eigenvalues $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$ and $\beta_1 \geq \beta_2 \geq \ldots \beta_n$, then for each $1 \leq i \leq n$, $|\alpha_i - \beta_i| \leq \|A - B\|$.*

Fact 7.4. *If A is an $n \times n$ matrix then $|\mathrm{tr}(A)| \leq n\|A\|$.*

In particular, we record the following simple consequence of these facts.

Claim 7.5. *Let A be an $n \times n$ positive definite matrix. Let $\delta < 1/\kappa(A)$ be a small enough positive number, and suppose that Q, Λ are obtained by calling the above approximate diagonalization oracle on A with a small error parameter $\delta > 0$. Then,*

- $\lambda_{\max}(\Lambda) \leq (1+\delta)^3\lambda_{\max}(A) = (1+\delta)^3\|A\|$, *and*
- $\lambda_{\min}(\Lambda) \geq \frac{(1-\delta\kappa(A))\lambda_{\min}(A)}{(1+\delta)^2}$.

Thus, $\kappa(\Lambda) \leq (1+\delta)^5\frac{\kappa(A)}{1-\delta\kappa(A)}$.

Proof. Using Fact 7.3, $\lambda_{\max}(Q\Lambda Q^T) - \lambda_{\max}(A) \leq \|A - Q\Lambda Q^T\| \leq \delta\|A\| = \delta\lambda_{\max}(A)$. Note that $(1 - \delta/3)^2\lambda_{\max}(\Lambda) \leq \lambda_{\max}(Q\Lambda Q^T)$. The first item thus holds when $\delta > 0$ is small. For the second item, by Fact 7.3, $\lambda_{\min}(A) - \lambda_{\min}(Q\Lambda Q^T) \leq \|A - Q\Lambda Q^T\| \leq \delta\|A\| = \delta\lambda_{\max}(A)$. Note that $(1 + \delta/3)^2\lambda_{\min}(\Lambda) \geq \lambda_{\min}(Q\Lambda Q^T)$. Putting the two inequalities together proves the second item. □

Proof of Theorem 7.1. Let $X \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ and $Y \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$. We use the approximate diagonalization oracle on Σ_2 with error bound $\delta = \mathrm{poly}\left(\frac{\epsilon\Delta}{n\kappa}\right)$ chosen small enough, to compute $\tilde{Q}_2$ and $\tilde{\Lambda}_2^{1/2}$ with the guarantees of eq. (10). Define $\tilde{A} := \tilde{Q}_2\tilde{\Lambda}_2^{-1/2}\tilde{Q}_2^T$; note that $\tilde{A}$ is symmetric. Consider the random variable $\bar{X} = \tilde{A}X - \tilde{A}\boldsymbol{\mu}_2$, which follows the law of $\mathcal{N}(\tilde{A}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tilde{A}\Sigma_1\tilde{A}^T)$, and the random variable $\bar{Y}' = \tilde{A}Y - \tilde{A}\boldsymbol{\mu}_2$, which follows the law of $\mathcal{N}(0, \tilde{A}\Sigma_2\tilde{A}^T)$. Since $\tilde{A}$ is invertible, by Proposition 2.1, we have

$$D = d_{\mathrm{TV}}(X, Y) = d_{\mathrm{TV}}\left(\bar{X}, \bar{Y}'\right). \tag{11}$$

However, to mimic the proof of Lemma 4.1, we replace $\bar{Y}'$ with $\bar{Y} \sim \mathcal{N}(0, I_n)$. Note that $\bar{Y}'$ may not be a standard Gaussian because we use only an approximate diagonalization oracle. We thus need to bound the total variation distance between $\bar{Y}$ and $\bar{Y}'$. Using Fact 7.2,

$$d_{\mathrm{TV}}\left(\bar{Y}, \bar{Y}'\right) \leq \frac{1}{2}\sqrt{\mathrm{tr}(\tilde{A}\Sigma_2\tilde{A}^T) - n + \ln\det(\tilde{A}\Sigma_2\tilde{A}^T)}. \tag{12}$$

Using the invariance of trace under cyclic shifts followed by the guarantee of the approximation diagonalization oracle along with Fact 7.4 above, we get (we also use that δ is small enough so that $(1+\delta)^k \leq 1.1$ for $k \leq 10$).

$$\begin{aligned}
&\mathrm{tr}(\tilde{A}\Sigma_2\tilde{A}^T) && (13)\\
&= \mathrm{tr}(\tilde{A}^T\tilde{A}\Sigma_2) && (14)\\
&\leq \mathrm{tr}(\tilde{A}^T\tilde{A}\tilde{Q}_2\tilde{\Lambda}_2\tilde{Q}_2^T) + n\delta\|\Sigma_2\|\|\tilde{A}\|^2 && (15)\\
&\leq \mathrm{tr}(\tilde{A}^T\tilde{A}\tilde{Q}_2\tilde{\Lambda}_2\tilde{Q}_2^T) + n\delta(1+\delta)^4\kappa(\Sigma_2)\frac{\lambda_{\min}(\Sigma_2)}{\lambda_{\min}(\tilde{\Lambda}_2)} && (16)\\
&\leq \mathrm{tr}(\tilde{Q}_2^T\tilde{A}^T\tilde{A}\tilde{Q}_2\tilde{\Lambda}_2) + \frac{2n\delta\kappa}{1-\delta\kappa}. && (17)
\end{aligned}$$

Here, in the first inequality, we use $\mathrm{tr}(\tilde{A}^T\tilde{A}\Sigma_2) \leq \mathrm{tr}(\tilde{A}^T\tilde{A}\tilde{Q}_2\tilde{\Lambda}_2\tilde{Q}_2^T) + |\mathrm{tr}(\tilde{A}^T\tilde{A}(\Sigma_2 - \tilde{Q}_2\tilde{\Lambda}_2\tilde{Q}_2^T))|$, where we bound the second term by Fact 7.4. In the last inequality, we use Claim 7.5 along with the approximate diagonalization

oracle's guarantees. Now, to estimate the first term, we expand $\tilde{A} = \tilde{Q}_2 \tilde{\Lambda}_2^{-1/2} \tilde{Q}_2^T$, and note that $\|\tilde{Q}_2^T \tilde{Q}_2 - I\| \leq \delta$. Repeatedly using this last estimate along with Fact 7.4 above then gives

$$\begin{aligned}\operatorname{tr}(\tilde{A}\Sigma_2\tilde{A}_2^T) &\leq \operatorname{tr} I + \frac{7n\delta\lambda_{\max}(\tilde{\Lambda}_2)}{\lambda_{\min}(\tilde{\Lambda}_2)} + \frac{2n\delta\kappa}{1-\delta\kappa} \\ &\leq n + \frac{10n\delta\kappa}{1-\delta\kappa}, \end{aligned} \tag{18}$$

where we again use Claim 7.5 along with the approximate diagonalization oracle's guarantee in eq. (10) to get the last inequality.

Next, we need to bound the determinant of $\tilde{A}\Sigma_2\tilde{A}^T$, where $\tilde{A} = \tilde{Q}_2\tilde{\Lambda}_2^{-1/2}\tilde{Q}_2^T$. Let λ_i for $1 \leq i \leq n$ denote the eigenvalues of Σ_2 arranged in descending order. Recall that $\kappa(\Sigma_2) \leq \kappa$. We have

$$\begin{aligned}\det(\tilde{A}\Sigma_2\tilde{A}^T) &= \det(\tilde{A})^2 \det(\Sigma_2) \\ &= \det(\tilde{Q}_2)^6 \frac{\det(\Sigma_2)}{\det(\tilde{Q}_2\tilde{\Lambda}_2\tilde{Q}_2^T)} \\ &\overset{(*)}{\leq} (1+\delta/3)^{6n} \prod_{i=1}^{n} \frac{\lambda_i}{\lambda_i - \delta\lambda_1} \\ &\leq \exp\left(\frac{3n\delta\kappa}{1-\delta\kappa}\right), \end{aligned} \tag{19}$$

where $(*)$ holds because of Weyl's inequality in Fact 7.3 (with $A = \Sigma_2$ and $B = \tilde{Q}_2\tilde{\Lambda}_2\tilde{Q}_2^T$) and the guarantees from the oracle. As our $\delta = \operatorname{poly}\left(\frac{\epsilon\Delta}{n\kappa}\right)$ is chosen small enough, we thus get (by substituting eqs. (18) and (19) into eq. (12))

$$d_{\mathrm{TV}}\left(\bar{Y}, \bar{Y}'\right) \leq \frac{1}{2}\sqrt{\frac{13n\delta\kappa}{1-\delta\kappa}} \leq \frac{\epsilon\Delta}{200} \leq \frac{\epsilon D}{200}. \tag{20}$$

The final step in the proof of Lemma 4.1 is to transform $\bar{X}$ into a product distribution. Let $\bar{\Sigma}_1 = \tilde{A}\Sigma_1\tilde{A}^T$ be the covariance matrix of $\bar{X}$. We use the approximate oracle on $\bar{\Sigma}_1$ with a small error bound $\delta = \operatorname{poly}\left(\frac{\epsilon\Delta}{n\kappa}\right)$ to obtain $\tilde{Q}_1$ and $\tilde{\Lambda}_1^{1/2}$ satisfying the guarantees of eq. (10): in particular, we have $\|\bar{\Sigma}_1 - \tilde{Q}_1\tilde{\Lambda}_1\tilde{Q}_1^T\| \leq \delta\|\bar{\Sigma}_1\|$. We then let $\hat{X}' = \tilde{Q}_1^T\bar{X}$ and $\hat{Y}' = \tilde{Q}_1^T\bar{Y}$. Then, as δ is small enough, $\tilde{Q}_1$ is invertible so that

$$d_{\mathrm{TV}}(\hat{X}', \hat{Y}') = d_{\mathrm{TV}}(\bar{X}, \bar{Y}). \tag{21}$$

Define first $\hat{Y}$ to have the same distribution as $\bar{Y} \sim \mathcal{N}(0, I)$. Note that $\hat{Y}' \sim \mathcal{N}(0, \tilde{Q}_1^T\tilde{Q}_1)$. The guarantees from the approximate diagonalization oracle yield that $\|\tilde{Q}_1^T\tilde{Q}_1 - I\| \leq \delta$. Thus, Fact 7.2 yields that

$$\begin{aligned} d_{\mathrm{TV}}(\hat{Y}, \hat{Y}') &\leq \frac{1}{2}\sqrt{\operatorname{tr}(\tilde{Q}_1^T\tilde{Q}_1) - n + \log\det(\tilde{Q}_1^T\tilde{Q}_1)} \\ &\leq \sqrt{n\delta} \\ &\leq \frac{\epsilon\Delta}{200} \leq \frac{\epsilon D}{200}, \end{aligned} \tag{22}$$

where the last two inequalities come from the small choice of δ.

We now consider $\hat{X}'$. The covariance matrix of $\hat{X}'$ is $\Sigma' = \tilde{Q}_1^T\bar{\Sigma}_1\tilde{Q}_1$ and the mean vector of $\hat{X}'$ is $\tilde{Q}_1^T\tilde{A}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Since $\tilde{Q}_1$ is obtained using the approximate diagonalization oracle, Σ' may not be diagonal. Define a product Gaussian $\hat{X}$ with the same mean vector $\tilde{Q}_1^T\tilde{A}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ but a diagonal covariance matrix $\tilde{\Lambda}_1$. By Fact 7.2,

$$d_{\mathrm{TV}}(\hat{X}', \hat{X}) \leq \frac{1}{2}\sqrt{\operatorname{tr}(\Sigma'\tilde{\Lambda}_1^{-1}) - n + \ln\frac{\det(\Sigma')}{\det(\tilde{\Lambda}_1)}}. \tag{23}$$

In a manner analogous to eqs. (15), (17) and (18) above, we then get that

$$\operatorname{tr}(\Sigma'\tilde{\Lambda}_1^{-1}) = \operatorname{tr}(\tilde{Q}_1^T\bar{\Sigma}_1\tilde{Q}_1\tilde{\Lambda}_1^{-1}) \leq n + \frac{10n\delta\kappa(\bar{\Sigma}_1)}{1-\delta\kappa(\bar{\Sigma}_1)}, \tag{24}$$

provided $\delta < 1/\kappa(\bar{\Sigma}_1)$ is small enough. We thus need to bound the condition number of $\bar{\Sigma}_1 = \tilde{A}\Sigma_1\tilde{A}^T$. Note that by the small enough choice of δ, all of the matrices $\Sigma_1, \bar{\Sigma}_1$, and $\tilde{A}$ are positive definite. We then have

$$\begin{aligned}
\lambda_{\max}(\bar{\Sigma}_1) &= \|\bar{\Sigma}_1\| \\
&\leq \|\tilde{A}\|^2 \cdot \|\Sigma_1\| \\
&\leq (1+\delta)^4 \lambda_{\min}^{-1}(\tilde{\Lambda}_2)\lambda_{\max}(\Sigma_1), \text{ and} \\
\lambda_{\min}^{-1}(\bar{\Sigma}_1) &= \|\bar{\Sigma}_1^{-1}\| \\
&\leq \|\tilde{A}^{-1}\|^2 \cdot \|\Sigma_1^{-1}\| \\
&\leq (1+\delta)^4 \lambda_{\max}(\tilde{\Lambda}_2)\lambda_{\min}^{-1}(\Sigma_1).
\end{aligned}$$

Hence, $\kappa(\bar{\Sigma}_1) \leq (1+\delta)^8 \kappa(\tilde{\Lambda}_2)\kappa(\Sigma_1)$. By assumption, $\kappa(\Sigma_1) \leq \kappa$. By Claim 7.5 applied with A in the statement of the claim set to Σ_2 and Λ to $\tilde{\Lambda}_2$, we thus get that (for δ small enough) $\kappa(\bar{\Sigma}_1) \leq \frac{4\kappa^2}{1-\delta\kappa}$ (since $\kappa(\Sigma_1)$ and $\kappa(\Sigma_2)$ are both at most κ). Substituting this bound in eq. (24), we get (again for $\delta = \text{poly}\left(\frac{\epsilon\Delta}{n\kappa}\right)$ small enough)

$$\operatorname{tr}(\Sigma'\tilde{\Lambda}_1^{-1}) \leq n + 200n\delta\kappa^2. \tag{25}$$

Next, we need to bound the ratio of determinants, which is (by a calculation similar to eq. (19), and using the estimate on $\kappa(\bar{\Sigma}_1)$ above)

$$\begin{aligned}
\frac{\det(\Sigma')}{\det(\tilde{\Lambda}_1)} &= \det(\tilde{Q}_1)^4 \frac{\det(\bar{\Sigma}_1)}{\det(\tilde{Q}_1\tilde{\Lambda}_1\tilde{Q}_1^T)} \\
&\leq \exp\left(\frac{3n\delta\kappa(\bar{\Sigma}_1)}{1-\delta\kappa(\bar{\Sigma}_1)}\right) \\
&\leq \exp(24n\delta\kappa^2).
\end{aligned} \tag{26}$$

We now substitute the bounds in eqs. (25) and (26) into the bound in eq. (23) for the total variation distance between $\hat{X}$ and $\hat{X}'$. Since $\delta = \text{poly}\left(\frac{\epsilon\Delta}{n\kappa}\right)$ is small enough, the total variation distance between $\hat{X}$ and $\hat{X}'$ is thus bounded as

$$d_{\mathrm{TV}}\left(\hat{X}, \hat{X}'\right) \leq \frac{1}{2}\sqrt{224n\delta\kappa^2} \leq \frac{\epsilon\Delta}{200} \leq \frac{\epsilon D}{200}. \tag{27}$$

Let $\boldsymbol{\mu} = \tilde{Q}_1^T\tilde{A}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $\Sigma = \tilde{\Lambda}_1$ be the mean vector and covariance matrix of the product Gaussian $\hat{X}$, as computed above. Recall that $\hat{Y} \sim \mathcal{N}(0, I_n)$. By triangle inequality,

$$\begin{aligned}
&d_{\mathrm{TV}}(\hat{X}, \hat{Y}) \\
&\leq d_{\mathrm{TV}}\left(\hat{X}, \hat{X}'\right) + d_{\mathrm{TV}}\left(\hat{X}', \hat{Y}'\right) + d_{\mathrm{TV}}\left(\hat{Y}', \hat{Y}\right) && \text{(by triangle inequality)} \\
&\leq \frac{\epsilon D}{100} + d_{\mathrm{TV}}(\bar{X}, \bar{Y}) && \text{(by eqs. (21), (22) and (27))} \\
&\leq \frac{\epsilon D}{100} + d_{\mathrm{TV}}(\bar{X}, \bar{Y}') + d_{\mathrm{TV}}(\bar{Y}, \bar{Y}') && \text{(by triangle inequality)} \\
&\leq \frac{\epsilon D}{50} + D. && \text{(by eqs. (11) and (20))}
\end{aligned}$$

A similar triangle inequality calculation starting with $D = d_{\mathrm{TV}}(X, Y) = d_{\mathrm{TV}}(\bar{X}, \bar{Y}')$ (see eq. (11)) shows that

$$d_{\mathrm{TV}}(\hat{X}, \hat{Y}) \geq D - \frac{\epsilon D}{50},$$

which completes the proof. The running time of the algorithm is almost the same as the running time of the algorithm in Lemma 4.1 except that the two algorithms use different oracles. □

Finally, we give a lower bound Δ if all inputs are given as rational numbers. Following the notation of Dey et al. (2023), let $\mathbb{Z}\langle\langle a\rangle\rangle$ denote all integers of bit length at most a, and let $\mathbb{Q}\langle\langle a\rangle\rangle$ denote all rational numbers p/q with $p, q \in \mathbb{Z}\langle\langle a\rangle\rangle$.

Lemma 7.6. *Suppose that $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{Q}\langle\langle a \rangle\rangle^n$ and $\Sigma_1, \Sigma_2 \in \mathbb{Q}\langle\langle a \rangle\rangle^{n \times n}$, where $a = \operatorname{poly}(n)$. Assume that either $\boldsymbol{\mu_1} \neq \boldsymbol{\mu_2}$ or else $\Sigma_1 \neq \Sigma_2$, and let $D = d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1), \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2))$. Then, $D \geq \Delta = 2^{-\operatorname{poly}(n)}$.*

Proof. Let $X \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ and $Y \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$. Note that if $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$ then $D = 0$.

Consider the first case that there exists $1 \leq i \leq n$ such that $\boldsymbol{\mu}_1(i) \neq \boldsymbol{\mu}_2(i)$ or $\Sigma_1(i,i) \neq \Sigma_2(i,i)$. The TV-distance between X, Y can be lower bounded by the TV-distance between X_i and Y_i. Using the lower bound of Devroye et al. (2023, Theorem 1.3; see Lemma 9.1), we have the following lower bound

$$\begin{aligned} D &\geq d_{\mathrm{TV}}(X_i, Y_i) \\ &\geq \frac{1}{200} \min\left\{1, \max\left\{\frac{|\Sigma_2(i,i) - \Sigma_1(i,i)|}{\Sigma_1(i,i)}, \frac{40|\mu_1(i) - \mu_2(i)|}{\sqrt{\Sigma_1(i,i)}}\right\}\right\} \\ &=: \Delta_1 \geq 2^{-\operatorname{poly}(n)}, \end{aligned}$$

because each rational number is in $\mathbb{Q}\langle\langle a \rangle\rangle$ with $a = \operatorname{poly}(n)$.

Assume the first case does not hold, which implies $\boldsymbol{\mu}_1(i) = \boldsymbol{\mu}_2(i)$ and $\Sigma_1(i,i) = \Sigma_2(i,i)$ for all $1 \leq i \leq n$. There must exist $i \neq j$ such that $\Sigma_1(i,j) \neq \Sigma_2(i,j)$. Define the vector $a \in \mathbb{R}^n$ such that $a_i = 1, a_j = 1$ and $a_k = 0$ for all $k \neq i, j$. By Proposition 2.1 , $D \geq d_{\mathrm{TV}}\left(a^T X, a^T Y\right)$, where $a^T X \sim \mathcal{N}(\boldsymbol{\mu}_1(i) + \boldsymbol{\mu}_1(j), \Sigma_1(i,i) + 2\Sigma_1(i,j) + \Sigma_1(j,j))$ and $a^T Y \sim \mathcal{N}(\boldsymbol{\mu}_2(i) + \boldsymbol{\mu}_2(j), \Sigma_2(i,i) + 2\Sigma_2(i,j) + \Sigma_2(j,j))$. The same proof as above shows that $D \geq d_{\mathrm{TV}}\left(a^T X, a^T Y\right) \geq \Delta_2 = 2^{-\operatorname{poly}(n)}$.

Let $\Delta = \Delta_1$ for the first case and $\Delta = \Delta_2$ for the second case. □

In passing, we note that several fundamental problems regarding the computational complexity of diagonalization and related procedures remain open. We refer to the recent survey by Srivastava (2023) for a discussion and to the papers by Armentano et al. (2018), Banks et al. (2023), Dey et al. (2023), and Shah (2024, 2025) for recent progress and further references: this list of references is by no means meant to be exhaustive.

8 PROOFS OMITTED FROM SECTION 3

In this section, we often use the definition and standard properties of conditional expectation, which may be found, e.g., in Williams (1991, Sections 9.2 and 9.7).

The following alternative expressions for the total variation distance are well known, but we include the proofs for completeness.

Fact 8.1. *Let P and Q be probability measures on a measurable space $(\Omega, \mathcal{F})$. Let the decomposition $P_Q, P_Q^\perp$ of P and the random variable R be as in Definition 3.1. Then, we have*

$$d_{\mathrm{TV}}(P, Q) = \mathbb{E}_Q[(1 - R)_+] \tag{28}$$

$$= \mathbb{E}_Q[(R - 1)_+] + P_Q^\perp(\Omega) \tag{29}$$

$$= \frac{1}{2}\mathbb{E}_Q[|1 - R|] + \frac{1}{2}P_Q^\perp(\Omega). \tag{30}$$

In particular, if $P \ll Q$ then $d_{\mathrm{TV}}(P, Q) = \mathbb{E}_Q[(1 - R)_+] = \mathbb{E}_Q[(R - 1)_+] = \frac{1}{2}\mathbb{E}_Q[|1 - R|]$. (Here, we use the standard notation $x_+ := \max\{x, 0\}$ for any real number x, so that $|x| = x_+ + (-x)_+$.)

Proof. Equation (30) follows by averaging eqs. (28) and (29), so we focus on proving the latter two equations. Let $B \in \mathcal{F}$ be such that $Q(B^c) = 0$ and $P_Q^\perp(B) = 0$ (such a B exists as Q and $P_Q^\perp$ are mutually singular). Denote by C the subset $\{\omega \in \Omega : R(\omega) \leq 1\} \cap B$ of B on which $R \leq 1$, and by D the subset $\{\omega \in \Omega : R(\omega) > 1\} \cap B$ on which $R > 1$: both these sets are $\mathcal{F}$-measurable since R is $\mathcal{F}$-measurable. We then compute (using $Q(B^c) = P_Q^\perp(B) = 0$

and the decomposition in Definition 3.1) that

$$Q(C) - P(C) = \mathbb{E}_Q[(1-R)_+], \text{ and} \tag{31}$$

$$\begin{aligned} P(D \cup B^c) - Q(D \cup B^c) & \\ &= \mathbb{E}_Q[(R-1)_+] + P_Q^{\perp}(B^c) \\ &= \mathbb{E}_Q[(R-1)_+] + P_Q^{\perp}(\Omega). \end{aligned} \tag{32}$$

Since $C^c = D \cup B^c$, this calculation also shows that the right hand sides of eqs. (28) and (29) are equal. We thus focus on proving eq. (28). To do this, it is sufficient to observe that for any $A \in \mathcal{F}$, we have $Q(A \cap C^c) - P(A \cap C^c) \leq 0$ and that $Q(A \cap C) - P(A \cap C) = \mathbb{E}_Q[(1-R)_+ I_A]$. Adding these two equations and taking the supremum over all $A \in \mathcal{F}$ then shows that $d_{\mathrm{TV}}(P, Q) \leq \mathbb{E}_Q[(1-R)_+]$, which combined with eq. (31) completes the proof. □

8.1 Proof of Lemma 3.6

Proof of Lemma 3.6. Item 1 follows from the definition of the TV functional since $x \mapsto (1-x)_+$ is a non-increasing function (and since $R \geq 0$). For item 2, it follows from standard properties of the conditional expectation (Williams, 1991, Section 9.7 (a, d)) that the conditional expectation of a valid ratio is also a valid ratio. The first part of item 2 then follows from Jensen's inequality for conditional expectations (Williams, 1991, Section 9.7 (h)), since the function $x \mapsto (1-x)_+$ is convex:

$$\mathbb{E}[(1 - \mathbb{E}[R|\mathcal{G}])_+] \leq \mathbb{E}[\mathbb{E}[(1-R)_+|\mathcal{G}]] = \mathbb{E}[(1-R)_+]. \tag{33}$$

For the second part, consider the events $E_1 = \{R < 1 \text{ and } R_{\mathcal{G}} \geq 1\}$ and $E_2 = \{R \geq 1 \text{ and } R_{\mathcal{G}} < 1\}$. We wish to show that $P(E_1 \cup E_2) = 0$. By the hypothesis for this item, both E_1 and E_2 are $\mathcal{G}$-measurable sets, so that (by the definition of the conditional expectation) the integral of the random variable $R - R_{\mathcal{G}}$ is zero on both E_1 and E_2. However, this random variable is (by definition) strictly negative on E_1 and strictly positive on E_2. It then follows by a standard argument (e.g., by considering probabilities of events of the form $E_1 \cap \{R - R_{\mathcal{G}} < -1/n\}$ for positive integers n for the case of E_1) that $P(E_1) = P(E_2) = 0$, which implies the claim that the events $\{R < 1\}$ and $\{R_{\mathcal{G}} < 1\}$ are almost surely equivalent. For the equality, we now compute using the above equivalence and the definition of the conditional expectation:

$$\begin{aligned} TV(R) &= \mathbb{E}[(1-R) \cdot I[R < 1]] \\ &= \mathbb{E}[(1-R) \cdot I[R_{\mathcal{G}} < 1]] \\ &= \mathbb{E}[(1-R_{\mathcal{G}}) \cdot I[R_{\mathcal{G}} < 1]] \\ &= TV(R_{\mathcal{G}}). \end{aligned}$$

For the third part, we note that $\mathbb{E}[R \underset{\text{ind}}{\circ} S|\sigma(R)] = \mathbb{E}[S] \cdot R$. Since $0 \leq \mathbb{E}[S] \leq 1$, we get the result by applying item 1 followed by item 2. □

8.2 Proof of Remark in Definition 3.8

Towards the end of Definition 3.8, we remarked that for each interval $J \in \mathcal{I}_{\gamma,\delta}$, the events $\{R \in J\}$ and $\{\tilde{R} \in J\}$ are almost surely equivalent (i.e., the probability that exactly one of them occurs is zero). Given the definition of $\tilde{R}$, this is intuitively obvious, and the formal proof is very similar to the argument in the proof of Lemma 3.6. We include this argument for completeness.

We show first that for any two distinct intervals $J \neq J' \in \mathcal{I}_{\gamma,\delta}$ the probability of the event $\left\{R \in J, \tilde{R} \in J'\right\}$ is zero. Let E denote this event. By definition of $\mathcal{I}_{\gamma,\delta}$, any two distinct intervals in $\mathcal{I}_{\gamma,\delta}$ are disjoint. Thus, either all points in J are smaller than all points in J', or else all points in J' are smaller than all points in J. Define $Z = \tilde{R} - R$ in the former case and $Z = R - \tilde{R}$ in the latter case. We then have that Z is *strictly* positive when E occurs. In particular, if we define $E_k := E \cap \{Z > 1/k\}$ for each positive integer k, then $E = \bigcup_{k \geq 1} E_k$, so that if $\mathbb{P}[E_k] = 0$ for each positive integer k then $\mathbb{P}[E] = 0$ follows from the countable sub-additivity of probabilities. We now show that $\mathbb{P}[E_k] = 0$ for each positive integer k.

Note that by definition of the σ-field $\mathcal{G} := \sigma(\mathcal{I}_{\gamma,\delta}(R))$, the event $\{R \in J\}$ is $\mathcal{G}$-measurable (since $\{\mathbb{R} \in J\}$ is by definition the same as $\{\mathcal{I}_{\gamma,\delta}(R) = J\}$). $\tilde{R} = \mathbb{E}[R|\mathcal{G}]$ is by definition $\mathcal{G}$-measurable, and therefore so is the event

$\left\{\tilde{R} \in J'\right\}$. Thus, E is also $\mathcal{G}$-measurable. It thus follows from the definition of conditional expectation that $\mathbb{E}[R \cdot I[E]] = \mathbb{E}[\mathbb{E}[R|\mathcal{G}] \cdot I[E]] = \mathbb{E}[\tilde{R} \cdot I[E]]$. We thus get that $\mathbb{E}[Z \cdot I[E]] = 0$. Thus, since $E_k \subseteq E$ and Z is strictly positive on E, we get that $\mathbb{E}[Z \cdot I[E_k]] = 0$ as well. On the other hand, when the event E_k occurs, $Z > 1/k$. Thus, we also have $\mathbb{E}[Z \cdot I[E_k]] \geq \mathbb{P}[E_k]/k$. We thus get that $\mathbb{P}[E_k] = 0$, for each positive integer k.

We thus have $\mathbb{P}\left[R \in J, \tilde{R} \in J'\right] = 0$ whenever J and J' are distinct intervals in $\mathcal{I}_{\gamma,\delta}$. The claim in the remark now follows since $R, \tilde{R}$ are non-negative so that for any interval $J \in \mathcal{I}_{\gamma,\delta}$,

$$\mathbb{P}\left[R \in J, \tilde{R} \notin J\right] = \sum_{\substack{J' \in \mathcal{I}_{\gamma,\delta} \\ J' \neq J}} \mathbb{P}\left[R \in J, \tilde{R} \in J'\right] = 0, \text{ and}$$
$$\mathbb{P}\left[R \notin J, \tilde{R} \in J\right] = \sum_{\substack{J' \in \mathcal{I}_{\gamma,\delta} \\ J' \neq J}} \mathbb{P}\left[\tilde{R} \in J, R \in J'\right] = 0.$$

8.3 Properties of the (γ, δ)-Discretization

The following elementary property of (γ, δ)-partitions turns out to be very useful.

Observation 8.2. *Suppose $x, y \in [0, 1]$ lie in the same partition in $\mathcal{I}_{\gamma,\delta}$, i.e., $\mathcal{I}_{\gamma,\delta}(x) = \mathcal{I}_{\gamma,\delta}(y)$. Then*

$$|x - y| \leq \begin{cases} \gamma & \text{if } x, y \in I_1. \\ (1-x)\delta & \text{otherwise.} \end{cases} \tag{34}$$

Proof. We use the notation of Definition 3.7. If $x, y \in I_0 = \{1\}$, there is nothing to prove. Otherwise, if $x, y \in I_k$ for $k \geq 1$, we have $|x - y| \leq a_{k-1} - a_k$, which is equal to γ when $k = 1$. Otherwise, when $k > 1$, it equals $(1 - a_{k-1})\delta$, which is at most $(1 - x)\delta$. □

We record the following simple properties of the (γ, δ)-discretization.

Observation 8.3. *Let $\tilde{R}$ be the (γ, δ)-discretization of a valid ratio R defined on some probability space $(\Omega, \mathcal{F}, P)$. Then*

1. $\mathbb{E}[\left|R - \tilde{R}\right| I[R < 1]] \leq \gamma + \delta TV(R)$.
2. $\mathbb{E}\left[\left|\frac{1}{R} - \frac{1}{\tilde{R}}\right| RI[R \geq 1]\right] \leq \gamma + \delta TV(R)$.

Proof. Let the intervals $I_0, I_1, J_1, \ldots, I_m, J_m$ be as in the definition of $\mathcal{I}_{\gamma,\delta}$ in Definition 3.7. Then $I[R < 1] = \sum_{i=1}^{m} I[R \in I_i]$ and $I[R \geq 1] = I[R \in I_0] + \sum_{i=1}^{m} I[R \in J_i]$. As noted in Definition 3.8, the random variables $I[R \in I]$ and $I[\tilde{R} \in I]$ are almost surely equal. Thus, almost surely, $R \in I \implies \tilde{R} \in I$ for each $I \in \mathcal{I}_{\gamma,\delta}$. Thus, by the construction of $\mathcal{I}_{\gamma,\delta}$ (see Observation 8.2), we have $\left|R - \tilde{R}\right| \leq \gamma$ when $R \in I_1$, and $\left|R - \tilde{R}\right| \leq a_{k-1} - a_k = \gamma\delta(1+\delta)^{k-2} = \delta(1 - a_{k-1}) \leq \delta(1-R)_+$ when $R \in I_k$ for $2 \leq k \leq m$. Combining these gives item 1. Similarly, we have $R = \tilde{R}$ when $R \in I_0$, $\left|1/R - 1/\tilde{R}\right| R \leq \gamma R$ when $R \in J_1$ and $\left|1/R - 1/\tilde{R}\right| R \leq \delta(R-1)_+$ when $R \in J_k$ for $2 \leq k \leq m$. Combining these and recalling that $\mathbb{E}[R] \leq 1$ gives item 2. □

8.4 Proof of Theorem 3.10

Proof of Theorem 3.10. Consider any valid ratio S defined on some probability space $(\Omega', \mathcal{F}', P')$. Then the definitions of $S \underset{\text{ind}}{\circ} R$ and $S \underset{\text{ind}}{\circ} \tilde{R}$ show that we can assume that $S, R, \tilde{R}, S \underset{\text{ind}}{\circ} R$ and $S \underset{\text{ind}}{\circ} \tilde{R}$ are all defined on the

space $(\Omega' \times \Omega, \mathcal{F}' \times \mathcal{F}, P' \times P)$, in such a way that S is independent of R and $\tilde{R}$. We now note that

$$\begin{aligned}
&\left|TV(S \underset{\text{ind}}{\circ} R) - TV(S \underset{\text{ind}}{\circ} \tilde{R})\right| \\
&= \frac{1}{2} \cdot \left|\mathbb{E}\left[|1 - SR| - |1 - S\tilde{R}|\right] + \mathbb{E}[S(\tilde{R} - R)]\right| \\
&= \frac{1}{2} \cdot \Big|\mathbb{E}\left[(|1 - SR| - |1 - S\tilde{R}|)I[R < 1]\right] \\
&\qquad + \mathbb{E}\left[(|1 - SR| - |1 - S\tilde{R}|)I[R \geq 1]\right] \\
&\qquad + \mathbb{E}[S(\tilde{R} - R)]\Big|.
\end{aligned} \tag{35}$$

Note that the last term is zero since by the independence of S with R and $\tilde{R}$,

$$\mathbb{E}[S(\tilde{R} - R)] = \mathbb{E}[S] \cdot \mathbb{E}[\tilde{R} - R] = 0, \tag{36}$$

since $\tilde{R}$ is a conditional expectation of R. For the first term we have

$$\begin{aligned}
&\left|\mathbb{E}\left[(|1 - SR| - |1 - S\tilde{R}|)I[R < 1]\right]\right| \\
&\leq \mathbb{E}\left[\left||1 - SR| - |1 - S\tilde{R}|\right| \cdot I[R < 1]\right] \\
&\leq \mathbb{E}[S \cdot |R - \tilde{R}| \cdot I[R < 1]] \\
&= \mathbb{E}[S] \cdot \mathbb{E}[|R - \tilde{R}| \cdot I[R < 1]] \leq \gamma + \delta TV(R).
\end{aligned} \tag{37}$$

Here, the first inequality is Jensen's inequality, the second inequality uses $||a| - |b|| \leq |a - b|$ for real a and b, the equality is by the independence of S from R and $\tilde{R}$, and the last inequality is from item 1 of Observation 8.3 and the fact that S being a valid ratio satisfies $\mathbb{E}[S] \leq 1$.

For the second term, we begin by noting that for any *fixed* $\omega' \in \Omega'$, the random variable $\left|\frac{1}{\tilde{R}} - S(\omega')\right| \cdot I[R \geq 1]$ is a random variable on the probability space $(\Omega, \mathcal{F}, P)$ that is measurable with respect to the sub σ-algebra $\mathcal{G} \coloneqq \sigma(\mathcal{I}_{\gamma,\delta}(R))$ (since so are $\tilde{R}$ and $I[R \geq 1]$). We then have

$$\begin{aligned}
&\int_{\omega \in \Omega} \left|\frac{1}{\tilde{R}} - S(\omega')\right| \cdot I[R \geq 1] \cdot R \, P(d\omega) \\
&= \int_{\omega \in \Omega} \mathbb{E}\left[\left|\frac{1}{\tilde{R}} - S(\omega')\right| \cdot I[R \geq 1] \cdot R \middle| \mathcal{G}\right] P(d\omega) \\
&= \int_{\omega \in \Omega} \left|\frac{1}{\tilde{R}} - S(\omega')\right| \cdot I[R \geq 1] \cdot \mathbb{E}[R|\mathcal{G}] \, P(d\omega) \\
&= \int_{\omega \in \Omega} \left|\frac{1}{\tilde{R}} - S(\omega')\right| \cdot I[R \geq 1] \cdot \tilde{R} \, P(d\omega),
\end{aligned} \tag{38}$$

where the first and second equalities are by standard properties of the conditional expectation (Williams, 1991, Section 9.7 (a, j)), while the third is by the definition of $\tilde{R}$ as $E[R|\mathcal{G}]$. Integrating this equation over ω' using the measure P', we then have from Fubini's theorem that

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{\tilde{R}} - S\right| \cdot I[R \geq 1] \cdot R\right] \\
&\qquad = \mathbb{E}\left[\left|\frac{1}{\tilde{R}} - S\right| \cdot I[R \geq 1] \cdot \tilde{R}\right].
\end{aligned} \tag{39}$$

We can now estimate the second term on the RHS in eq. (35) as follows.

$$\begin{aligned}
&\left|\mathbb{E}\left[(|1-SR|-|1-S\tilde{R}|)I[R\geq 1]\right]\right| \\
&= \left|\mathbb{E}\left[\left(\left|\frac{1}{R}-S\right|R-\left|\frac{1}{\tilde{R}}-S\right|\tilde{R}\right)\cdot I[R\geq 1]\right]\right| \\
&= \left|\mathbb{E}\left[\left(\left|\frac{1}{R}-S\right|R-\left|\frac{1}{\tilde{R}}-S\right|R\right)\cdot I[R\geq 1]\right]\right| \\
&\leq \mathbb{E}\left[\left|\left|\frac{1}{R}-S\right|-\left|\frac{1}{\tilde{R}}-S\right|\right|\cdot R\cdot I[R\geq 1]\right] \\
&\leq \mathbb{E}\left[\left|\frac{1}{R}-\frac{1}{\tilde{R}}\right|\cdot R\cdot I[R\geq 1]\right] \leq \gamma+\delta TV(R).
\end{aligned} \tag{40}$$

Here, the first equality is valid since on the event $\{R\geq 1\}$ we also have $\left\{\tilde{R}\geq 1\right\}$ so that $1/R$ and $1/\tilde{R}$ are bounded random variables on this event. The second equality uses eq. (39). The two inequalities after that follow from Jensen's inequality and the triangle inequality, exactly as in the estimation of the first term of eq. (35) as carried out in eq. (37). The final inequality uses item 2 of Observation 8.3. The claim now follows by substituting the estimates in eqs. (36), (37) and (40) into eq. (35), and then taking supremum over all possible choices of S. □

8.5 Omitted Arguments in the Proof of Theorem 3.11

In the proof of Theorem 3.11, we noted that a direct inductive argument shows that each Y_i is a conditional expectation of $R_1 \underset{\text{ind}}{\circ} R_2 \underset{\text{ind}}{\circ} \ldots \underset{\text{ind}}{\circ} R_i$ (where the notation is as in the proof of Theorem 3.11). While this is an easy and standard argument, we include the details here for completeness.

All the following arguments are performed on the probability space $([M]^n, \mathcal{H}, Q)$, where $\mathcal{H}$ is the σ-field of all subsets of $[M]^n$, and Q is the product distribution on this space defined in Algorithm 1. For $1\leq i\leq n$, let $\mathcal{F}_i$ denote the σ-field $\sigma(R_i)$ generated by R_i, and define the product σ-fields $\mathcal{H}_i := \mathcal{F}_1\times\cdots\times\mathcal{F}_i$. Under the underlying product probability distribution Q, note that for all $j>i$, any event in $\mathcal{H}_i$ is independent of any event in $\mathcal{F}_j$.[7] We will show by induction that for every $1\leq i\leq n$, there is a sub σ-field $\mathcal{G}_i\subseteq\mathcal{H}_i$ such that $Y_i=\mathbb{E}[R_1R_2\ldots R_i|\mathcal{G}_i]$.

The claim is trivially true in the base case $i=1$, since $Y_1=R_1$, so that we can take $\mathcal{G}_1=\mathcal{F}_1=\mathcal{H}_1$. For the inductive case, suppose that $Y_i=\mathbb{E}[R_1R_2\ldots R_i|\mathcal{G}_i]$ for some σ-field $\mathcal{G}_i\subseteq\mathcal{H}_i$. Note that this implies that Y_i is $\mathcal{G}_i$ measurable (in other words, $\sigma(Y_i)\subseteq\mathcal{G}_i$). Thus, the σ-field $\mathcal{G}'_i := \sigma(\mathcal{I}_{\gamma,\delta}(Y_i))\subseteq\sigma(Y_i)$ is a sub-σ-field of $\mathcal{G}_i$ as well. Since $\tilde{Y}_i=\mathbb{E}[Y_i|\mathcal{G}'_i]$, we thus see from the tower property of conditional expectations (Williams, 1991, Section 9.7 (i)) that

$$\tilde{Y}_i=\mathbb{E}\big[\mathbb{E}[R_1R_2\ldots R_i|\mathcal{G}_i]|\mathcal{G}'_i\big]=\mathbb{E}[R_1R_2\ldots R_i|\mathcal{G}'_i], \tag{41}$$

where $\mathcal{G}'_i\subseteq\mathcal{H}_i$. Now, define $\mathcal{G}_{i+1} := \mathcal{G}'_i\times\mathcal{F}_{i+1}$, which by definition, is a sub-σ-field of $\mathcal{H}_{i+1}$. Note also that R_{i+1} is measurable with respect to $\mathcal{F}_{i+1}$ (by definition) and hence also with respect to $\mathcal{G}_{i+1}$. Thus, by a standard property of conditional expectations (Williams, 1991, Section 9 (j)),

$$\mathbb{E}[R_1R_2\ldots R_iR_{i+1}|\mathcal{G}_{i+1}]=\mathbb{E}[R_1R_2\ldots R_i|\mathcal{G}_{i+1}]R_{i+1}. \tag{42}$$

Now, as noted above, by construction of the product distribution Q, every event in $\mathcal{G}'_i\subseteq\mathcal{H}_i$ is independent of every event in $\mathcal{F}_{i+1}$. It then follows from the construction of the product σ-field and the interaction between independence and conditional expectation (Williams, 1991, Section 9.7 (k)) that

$$\tilde{Y}_i \overset{\text{eq. (41)}}{=} \mathbb{E}[R_1R_2\ldots R_i|\mathcal{G}'_i]=\mathbb{E}[R_1R_2\ldots R_i|\mathcal{G}'_i\times\mathcal{F}_{i+1}]. \tag{43}$$

Combining eqs. (42) and (43), we thus see that $Y_{i+1}=\tilde{Y}_iR_{i+1}$ satisfies

$$Y_{i+1}=\mathbb{E}[R_1R_2\ldots R_iR_{i+1}|\mathcal{G}_{i+1}], \tag{44}$$

which completes the induction.

[7] Note that since Theorem 3.11 is in the setting of finite sets, the "σ-fields" here are just finite set systems closed under complement and union, and containing the universal set. However, it is convenient to do the argument in this general setting. Note, however, that $\times$ denotes the product operation on σ-fields, and not merely the Cartesian product: for example, $\mathcal{F}_1\times\mathcal{F}_2$ is the set system generated by sets of the form $\{A\times B : A\in\mathcal{F}_1 \text{ and } B\in\mathcal{F}_2\}$.

8.6 Proof of Lemma 3.12

Proof of Lemma 3.12. Define $Z_i = \tilde{R}_1 \underset{\text{ind}}{\circ} \tilde{R}_2 \underset{\text{ind}}{\circ} \dots \underset{\text{ind}}{\circ} \tilde{R}_{i-1}$, and $S_i = R_{i+1} \underset{\text{ind}}{\circ} R_{i+2} \underset{\text{ind}}{\circ} \dots \underset{\text{ind}}{\circ} R_n$ (so that $Z_1 = S_n = 1$). Then we have

$$\begin{aligned}
&\left|TV(R_1 \underset{\text{ind}}{\circ} R_2 \underset{\text{ind}}{\circ} \dots \underset{\text{ind}}{\circ} R_n) - TV(\tilde{R}_1 \underset{\text{ind}}{\circ} \tilde{R}_1 \underset{\text{ind}}{\circ} \dots \underset{\text{ind}}{\circ} \tilde{R}_n)\right| \\
&\leq \sum_{i=1}^{n} \left|TV(Z_i \underset{\text{ind}}{\circ} S_i \underset{\text{ind}}{\circ} R_i) - TV(Z_i \underset{\text{ind}}{\circ} S_i \underset{\text{ind}}{\circ} \tilde{R}_i)\right| \\
&\leq \sum_{i=1}^{n} \text{ext}(R_i, \tilde{R}_i), \qquad (45)
\end{aligned}$$

where the last inequality is by the definition of the extension distance. The claim now follows from Theorem 3.10. (Note that we also use implicitly the fact that the $\underset{\text{ind}}{\circ}$ operation is "commutative" in the sense that $R_1 \underset{\text{ind}}{\circ} R_2$ has the same law as $R_2 \underset{\text{ind}}{\circ} R_1$.) □

9 PROOFS OMITTED FROM SECTION 4

We need to use the following result of Devroye et al. (2023).

Lemma 9.1 (Devroye et al. (2023, Theorem 1.3)). *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(0,1)$ be two one-dimensional Gaussians, then*

$$\begin{aligned}
&\frac{1}{200} \min\left\{1, \max\{|1-\sigma^2|, 40|\mu|\}\right\} \\
&\quad \leq d_{\text{TV}}(X, Y) \\
&\quad \leq \frac{3|\sigma^2 - 1|}{2} + \frac{|\mu|}{2}.
\end{aligned}$$

9.1 Proof of Lemma 4.3

Proof of Lemma 4.3. Using Lemma 9.1, we know that $\Delta_i \leq d_{\text{TV}}(X_i, Y_i)$. Furthermore, if $\max\{|1-\Sigma_{i,i}|, 40|\mu_i|\} \geq 1$, then $\Delta_i = 1/200$ and $10^4\Delta_i > 1 \geq d_{\text{TV}}(X_i, Y_i)$. If $\max\{|1-\Sigma_{i,i}|, 40|\mu_i|\} < 1$, we have

$$\begin{aligned}
10^4\Delta_i &\geq \frac{25}{2}|1-\Sigma_{i,i}| + \frac{25}{2} \times 40|\mu_i| \\
&\geq \frac{3}{2}|1-\Sigma_{i,i}| + \frac{1}{2}|\mu_i| \\
&\geq d_{\text{TV}}(X_i, Y_i).
\end{aligned}$$

Hence, for any $1 \leq i \leq n$, we have

$$\Delta_i \leq d_{\text{TV}}(X_i, Y_i) \leq 10^4\Delta_i.$$

Recall $\Delta = \max_i \Delta_i$. We have $\Delta \leq \max_i d_{\text{TV}}(X_i, Y_i) \leq d_{\text{TV}}(X, Y)$. In the other direction, $n10^4\Delta \geq 10^4 \sum_{i=1}^n \Delta_i \geq \sum_{i=1}^n d_{\text{TV}}(X_i, Y_i) \geq d_{\text{TV}}(X, Y)$, where the last inequality holds due to $\sum_{i=1}^n d_{\text{TV}}(X_i, Y_i) \geq d_{\text{TV}}(X, Y)$. To verify the last inequality, consider a coupling of X, Y that couples each dimensional independently and optimally, and the inequality follows from the coupling characterization of the total variation distance. □