On the Properties of Kullback-Leibler Divergence Between Multivariate Gaussian Distributions

Yufeng Zhang, Wanwei Liu, Zhenbang Chen, Ji Wang, Kenli Li

Abstract—Kullback-Leibler (KL) divergence is one of the most important divergence measures between probability distributions. In this paper, we prove several properties of KL divergence between multivariate Gaussian distributions. First, for any two n-dimensional Gaussian distributions $\mathcal{N}_1$ and $\mathcal{N}_2$, we give the supremum of $KL(\mathcal{N}_1||\mathcal{N}_2)$ when $KL(\mathcal{N}_2||\mathcal{N}_1) \leq \varepsilon$ $(\varepsilon > 0)$. For small ε, we show that the supremum is $\varepsilon + 2\varepsilon^{1.5} + O(\varepsilon^2)$. This quantifies the approximate symmetry of small KL divergence between Gaussians. We also find the infimum of $KL(\mathcal{N}_1||\mathcal{N}_2)$ when $KL(\mathcal{N}_2||\mathcal{N}_1) \geq M$ $(M > 0)$. We give the conditions when the supremum and infimum can be attained. Second, for any three n-dimensional Gaussians $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$, we find an upper bound of $KL(\mathcal{N}_1||\mathcal{N}_3)$ if $KL(\mathcal{N}_1||\mathcal{N}_2) \leq \varepsilon_1$ and $KL(\mathcal{N}_2||\mathcal{N}_3) \leq \varepsilon_2$ for $\varepsilon_1, \varepsilon_2 \geq 0$. For small ε_1 and ε_2, we show the upper bound is $3\varepsilon_1 + 3\varepsilon_2 + 2\sqrt{\varepsilon_1\varepsilon_2} + o(\varepsilon_1) + o(\varepsilon_2)$. This reveals that KL divergence between Gaussians follows a relaxed triangle inequality. Importantly, all the bounds in the theorems presented in this paper are independent of the dimension n. Finally, We discuss the applications of our theorems in explaining counterintuitive phenomenon of flow-based model, deriving deep anomaly detection algorithm, and extending one-step robustness guarantee to multiple steps in safe reinforcement learning.

Index Terms—Kullback-Leibler divergence, statistical divergence, multivariate Gaussian distribution, mathematical optimization, Lambert W function, deep learning, flow-based model, reinforcement learning

1 Introduction

A statistical divergence measures the "distance" between probability distributions. Let X be a space of probability distributions with the same support. A statistical divergence $D : X \times X \to \mathbb{R}^+$ ($\mathbb{R}^+$ is the set of non-negative real numbers) should satisfy (a) non-negativity: $D(p, q) \geq 0$ and (b) identity of indiscernibles: $D(p, p) = 0$, where p, q are probability densities. Another stronger concept, statistical distance, also measures the distance between probability distributions. A statistical distance should satisfy two extra properties including (c) symmetry: $D(p, q) = D(q, p)$ and (d) triangle inequality: $D(p, q) \leq D(p, g) + D(g, q)$, where p, q and g are probability densities.

Kullback-Leibler (KL) divergence, also referred to as relative entropy [1], plays a key role in many fields including machine learning [2], [3], information theory [4], and statistics [5], *etc*. The KL divergence between two continuous probability densities $p(x)$ and $q(x)$ is defined as

$$KL(p(x)||q(x)) = \int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{1}$$

KL divergence is not a proper distance [1]. Firstly, KL divergence is not symmetric. It might happen that the forward KL divergence[1] $KL(p||q)$ is very small but the reverse KL divergence $KL^*(p||q) = KL(q||p)$ is very large. Secondly, KL divergence dose not satisfy the triangle inequality. This brings obstacles in applying KL divergence in many circumstances.

KL divergence is one member of more generalized divergence families including f-divergence (also called ϕ-divergence) [6], Bregman divergence [7], and Rényi divergence [8]. For example, the widely used f-divergence includes many commonly used measures including KL divergence, Jensen-Shannon divergence, and squared Hellinger distance [5]. Many f-divergence are not proper distance metrics. KL divergence also has a deep connection with other information measures. For example, the second derivative of KL divergence is Fisher information metric. By taking the second-order Taylor expansion, KL divergence between two nearly distributions can be approximated with fisher information matrix [1]. Furthermore, forward and reverse KL divergence have the same second derivatives at the point where two distributions are equal. Therefore, KL divergence is locally approximately symmetric when two distributions are close to each other.

Meanwhile, Gaussian distribution is one of the most important distributions and central to statistics. It is also pervasive in many fields including machine learning and information theory. The probability density function of an

- *Yufeng Zhang and Kenli Li are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. E-mail: yufengzhang@hnu.edu.cn, lkl@hnu.edu.cn*
- *Wanwei Liu, Zhenbang Chen and Ji Wang are with the College of Computer, National University of Defense Technology, Changsha, China. Ji Wang is also with State Key Laboratory of High Performance Computing, National University of Defense Technology. E-mail: wwliu@nudt.edu.cn, zbchen@nudt.edu.cn, wj@nudt.edu.cn*
- *Kenli Li and Ji Wang are the corresponding authors.*

Manuscript received xx xx, 2022; revised xx xx, 2022.

1. Here we can choose to call $KL(p||q)$ or $KL(q||p)$ as forward KL divergence. The terminologies "forward" and "reverse" is just for convenience.

n-dimensional Gaussian distribution is given by

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \quad (2)$$

Here $\boldsymbol{\mu} \in \mathbb{R}^n$ is the mean and $\boldsymbol{\Sigma} \in \boldsymbol{S}_{++}^n$ is the covariance matrix, where $\boldsymbol{S}_{++}^n$ is the space of symmetric positive definite $n \times n$ matrices. Gaussian distribution constitutes the basis for more complicated distributions. For example, the mixture of Gaussians, namely Gaussian Mixture Model (GMM) has a wide range of applications due to its power of approximation [2].

The KL divergence between two n-dimensional Gaussians $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ has the following closed form [5]

$$\begin{aligned} &KL(\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \\ =&\frac{1}{2}\left(\log\frac{|\boldsymbol{\Sigma_2}|}{|\boldsymbol{\Sigma}_1|} + \mathrm{Tr}(\boldsymbol{\Sigma_2}^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) - n\right) \end{aligned} \quad (3)$$

where the logarithm is taken to base e and Tr is the trace of matrix. Due to the good form of Gaussians, one may expect that KL divergence between Gaussians may have some good properties. However, as like many other distributions, KL divergence between Gaussians is not symmetric and does not satisfy the triangle inequality either.

The concept of KL divergence has been proposed for seventy years [1]. It is surprising that the properties of KL divergence between Gaussians have not been investigated thoroughly.

In this paper, we investigate the following research problems.

1) For any two Gaussians $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, when forward KL divergence $KL(\mathcal{N}_1||\mathcal{N}_2)$ is bounded by a small number ε, what is the supremum of reverse KL divergence $KL(\mathcal{N}_2||\mathcal{N}_1)$? Although KL divergence is locally approximately symmetric, we want to step further in the investigation on such approximate symmetry in a Gaussian case.
2) For any two Gaussians $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, when forward KL divergence $KL(\mathcal{N}_1||\mathcal{N}_2)$ is not smaller than a number M, what is the infimum of reverse KL divergence $KL(\mathcal{N}_2||\mathcal{N}_1)$? This problem is dual to the first problem.
3) Does the KL divergence between Gaussians follow some property similar to the triangle inequality? Precisely, for any three Gaussians $\mathcal{N}_i(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ $(i \in 1, 2, 3)$, when $KL(\mathcal{N}_1||\mathcal{N}_2)$ and $KL(\mathcal{N}_2||\mathcal{N}_3)$ are bounded by small numbers $\varepsilon_1, \varepsilon_2$, respectively, how large $KL(\mathcal{N}_1||\mathcal{N}_3)$ can be?

Note that the third research problem is different from the several existing general Pythagoras theorems satisfied by KL divergence [4], [9], [10], [11]. In the existing general Pythagoras theorems, the bounds are dependent on the given distributions. In this paper, we want to obtain bounds that are independent of the parameters of Gaussians and hold for any Gaussian distributions. We will discuss this point in Section 4.

As far as we know, we are the first to propose and investigate the above research problems. These research problems are motivated by our research on deep anomaly detection with flow-based model [12], [13], [14]. Flow-based model is capable of provide explicit likelihood for input. Researchers have found that flow-based model may assign higher likelihoods for out-of-distribution (OOD) data than in-distribution (ID) data. For example, Glow [14] trained on CIFAR-10 assigns higher likelihoods for SVHN. However, we can not sample OOD data from the model with prior although they are assigned higher likelihoods by the model. This counterintuitive phenomenon has not been explained satisfactorily. In our research, we find that the KL divergence between several Gaussian(-like) distributions are vital in explaining the behavior of flow-based model. This inspires us to conduct research on the properties of KL divergence between Gaussians. In this context, the parameters of distributions are learned from data or dependent on the given inputs. It is impossible to identify the parameters or the KL divergence before the model is trained. We only know that some bound is guaranteed. Therefore, the existing general Pythagoras theorems are not applicable due to their dependence on the parameters of distributions. The theorems proved in this paper provide a solid foundation for explaining above phenomenon as well as our anomaly detection method using flow-based model. Our theorems can also be used as general conclusions in related fields including machine learning, information theory and statistics. For example, our theorems have been used as key support in safe reinforcement learning framework [15] after we post our last version of this manuscript on Arxiv [16]. We will elaborate these applications in Section 5.

Contributions. The contributions of this paper are as follows.

Given any three n-dimensional Gaussians $\mathcal{N}_1, \mathcal{N}_2$ and $\mathcal{N}_3$,

1) We prove that when $KL(\mathcal{N}_2||\mathcal{N}_1) \leq \varepsilon$ for $\varepsilon > 0$ the supremum of $KL(\mathcal{N}_1||\mathcal{N}_2)$ is $\frac{1}{2}\left((-W_0(-e^{-(1+2\varepsilon)}))^{-1} + \log(-W_0(-e^{-(1+2\varepsilon)})) - 1\right)$, where W_0 is the principal branch of Lambert W function. We give the condition when the supremum can be attained. For small ε, we show the supremum is $\varepsilon + 2\varepsilon^{1.5} + O(\varepsilon^2)$. This quantifies the approximate symmetry of small KL divergence between Gaussians.
2) We find the infimum of $KL(\mathcal{N}_1||\mathcal{N}_2)$ if $KL(\mathcal{N}_2||\mathcal{N}_1) \geq M$ for $M > 0$. We give two proofs for this result. The first proof has the similar structure with that for the above supremum. The second proof is based on the proof for the supremum. We also give the condition when the infimum can be attained.
3) We find an upper bound of $KL(\mathcal{N}_1||\mathcal{N}_3)$ if $KL(\mathcal{N}_1||\mathcal{N}_2) \leq \varepsilon_1$ and $KL(\mathcal{N}_2||\mathcal{N}_3) \leq \varepsilon_2$ for $\varepsilon_1, \varepsilon_2 \geq 0$. For small ε_1 and ε_2, we show the upper bound is $3\varepsilon_1 + 3\varepsilon_2 + 2\sqrt{\varepsilon_1\varepsilon_2} + o(\varepsilon_1) + o(\varepsilon_2)$. This indicates that KL divergence between Gaussians follows a relaxed triangle inequality.
4) All the bounds in our theorems are independent of the dimension n. This is a critical property especially in contexts where dimensionality has a fundamental impact.
5) We show several applications of the theorems proved in this paper including explaining counterintuitive phenomenon in flow-based model, deriving anomaly detection algorithm, and extending robustness guarantee in safe reinforcement learning.

The remaining part of this paper is organized as follows. In Section 2 we prepare lemmas that will be used in all theorems. In Section 3 we investigate the supremum

Research questions related to properties of KL divergence

Contributions of the article

(infimum) of reverse KL divergence between Gaussians when forward KL divergence is bounded. In Section 4 we investigate the relaxed triangle inequality of KL divergence between Gaussians. In Section 5 we discuss the applications of the theorems proved in this paper. In Section 6 we discuss related work. Finally, we conclude in Section 7.

2 Lemmas and Notations

Before presenting our results, we introduce the famous transcendental function, the Lambert W function, which occurs almost everywhere in this paper.

Definition 1. **Lambert W Function [17], [18].** The inverse function of function $y = xe^x$ is called Lambert W function $y = W(x)$.

When $x \in \mathbb{R}$, W is a multivalued function with two branches W_0, W_{-1}, where W_0 is the principal branch (also called branch 0) and W_{-1} is the branch -1. The derivative of W is

$$W'(x) = \frac{1}{x + e^{W(x)}} = \frac{W(x)}{x(1 + W(x))} \ (x \neq 0, -e^{-1}) \tag{4}$$

Function $f(x) = x - \log x$ lies in the core of our problems. In the following, we prove some useful lemmas related to $f(x)$.

Lemma 1. Given function $f(x) = x - \log x$ ($x \in \mathbb{R}^{++}$) ($\mathbb{R}^{++}$ is the set of positive real numbers), the following propositions hold.

(a) $f(x)$ is strictly convex and takes the minimum value 1 at $x = 1$.

(b) $f(x) > f(1/x)$ for $x > 1$ and $f(x) < f(1/x)$ for $0 < x < 1$.

(c) The inverse function of f is $f^{-1}(x) = -W(-e^{-x})$ $(x \geq 1)$.

(d) The solutions of equation $x - \log x = 1 + t$ $(t \geq 0)$ are $w_1(t) = -W_0(-e^{-(1+t)}) \in (0, 1]$ and $w_2(t) = -W_{-1}(-e^{-(1+t)}) \in [1, +\infty)$. It is easy to know $w_1(0) = w_2(0) = 1$. We treat $w_1(t), w_2(t)$ as functions of t.

(e) The derivatives of $w_1(t)$ and $w_2(t)$ are

$$w_1'(t) = \frac{-w_1(t)}{1 - w_1(t)} = \frac{W_0(-e^{-(1+t)})}{W_0(-e^{-(1+t)}) + 1} < 0 \tag{5}$$

$$w_2'(t) = \frac{-w_2(t)}{1 - w_2(t)} = \frac{W_1(-e^{-(1+t)})}{W_1(-e^{-(1+t)}) + 1} > 0 \tag{6}$$

(f) For $t > 0$, $f(w_1(t)) < f(\frac{1}{w_1(t)}), f(\frac{1}{w_2(t)}) < f(w_2(t))$.

(g) If $f(x) \leq 1 + t$ $(t \geq 0)$, then $w_1(t) \leq x \leq w_2(t)$ and

$$S(t) = \sup_{\substack{t \geq 0 \\ f(x) \leq 1+t}} f(\frac{1}{x}) = f(\frac{1}{w_1(t)}) \tag{7}$$

(h) If $f(x) \geq 1 + t$ $(t \geq 0)$, then $0 < x \leq w_1(t) \vee x \geq w_2(t)$ and

$$I(t) = \inf_{\substack{t \geq 0 \\ f(x) \geq 1+t}} f(\frac{1}{x}) = f(\frac{1}{w_2(t)}) \tag{8}$$

(i) For $t \geq 0$, $f'(w_2(t)) \leq -f'(\frac{1}{w_2(t)})$.

(j) For $t_1, t_2 \geq 0$,

$$f(w_1(t_1)w_2(t_2)) = t_1 + t_2 + 2 + w_1(t_1)w_1(t_2) - w_1(t_1) - w_1(t_2) \tag{9}$$

$$f(w_2(t_1)w_2(t_2)) = t_1 + t_2 + 2 + w_2(t_1)w_2(t_2) - w_2(t_1) - w_2(t_2) \tag{10}$$

Proof 1. The details of the proof are shown in Appendix A. □

The notations used in this paper are summarized in Table 1.

TABLE 1
Notations.

$f(x)$	$x - \log x$ $(x \in \mathbb{R}^{++})$
$W(x)$	the Lambert W function
$W_0(x)$	the principal branch (branch 0) of $W(x)$
$W_{-1}(x)$	the branch -1 of $W(x)$
$w_1(t)$	the smaller solution of $f(x) = 1 + t$ $(t \geq 0)$
$w_2(t)$	the larger solution of $f(x) = 1 + t$ $(t \geq 0)$
$\bar{f}(x_1, \ldots, x_n)$	$\sum_{i=1}^n f(x_i)$
λ	the eigenvalue of matrix
λ^*	the largest eigenvalue of matrix
λ_*	the least eigenvalue of matrix
$f_l(x)$	$f(1 - x) - 1$ $(0 \leq x < 1)$
$f_r(x)$	$f(x + 1) - 1$ $(x \geq 0)$
$g_l(\varepsilon)$	$f_l^{-1}(\varepsilon)$, the inverse function of f_l
$g_r(\varepsilon)$	$f_r^{-1}(\varepsilon)$, the inverse function of f_r
$\mathcal{N}(0, I)$	standard Gaussian distribution

3 Bounds of Forward and Reverse KL Divergence Between Gaussians

In this section, we give the supremum of reverse KL divergence when forward KL divergence is less than or equal to a positive number ε. We also show that the supremum is small if ε is small. These conclusions quantify the approximate symmetry of small KL divergence between Gaussians. We also give the infimum of reverse KL divergence when forward divergence is greater than or equal to a positive number M. Furthermore, we give the conditions when the supremum and infimum can be attained.

3.1 Supremum of Reverse KL Divergence Between Gaussians

We want to know how large the reverse KL divergence can be when forward KL divergence is bounded by a number ε. The following Theorem 1 gives the supremum of reverse KL divergence.

Theorem 1. For any two n-dimensional Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, if $KL(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \leq \varepsilon (\varepsilon \geq 0)$, then

$$KL(\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)||\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) \leq \frac{1}{2}\left(\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - \log \frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - 1\right)$$

The supremum is attained when the following two conditions hold.

(1) There exists only one eigenvalue λ_j of $B_2^{-1}\boldsymbol{\Sigma}_1(B_2^{-1})^\top$ or $B_1^{-1}\boldsymbol{\Sigma}_2(B_1^{-1})^\top$ equal to $-W_0(-e^{-(1+2\varepsilon)})$ and all other eigenvalues λ_i $(i \neq j)$ are equal to 1, where $B_1 = P_1 D_1^{1/2}$, P_1 is an orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{\Sigma}_1$, $D_1 = diag(\lambda_1, \ldots, \lambda_n)$ whose diagonal elements are the corresponding eigenvalues, B_2 is defined in the same way as B_1 except on $\boldsymbol{\Sigma}_2$.

Conditions for supremum

(2) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

Overview of proof of Theorem 1.

Theorem 1 can be seen as the following optimization problem $\boldsymbol{P_1}$.

$$\text{maximize } KL(\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)||\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) \tag{11}$$

$$s.t.\ KL(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \leq \varepsilon \tag{12}$$

Our aim is to solve problem $\boldsymbol{P_1}$ analytically. The proof consists of the following several steps.

1) *Invertible linear transformation*. We use a linear transformation to turn one of $\mathcal{N}_1$ and $\mathcal{N}_2$ into standard Gaussian. Since diffeomorphism preserves KL divergence [19], both the objective function and the constraints in $\boldsymbol{P_1}$ can be simplified.
2) *Reducing to new optimization problem*. We reduce $\boldsymbol{P_1}$ to the following core problem $\boldsymbol{P_2}$.

$$\text{maximize } \bar{\boldsymbol{f}}(\frac{1}{x_1}, \ldots, \frac{1}{x_n}) \tag{13}$$

$$s.t.\ \bar{\boldsymbol{f}}(x_1, \ldots, x_n) \leq n + \varepsilon' \tag{14}$$

where $\bar{\boldsymbol{f}}(x_1, \ldots, x_n) = \sum_{i=1}^n f(x_i) = \sum_{i=1}^n x_i - \log x_i$ ($x_i \in (0, \infty)$).

3) *Investigating $f(x)$*. $f(x)$ lies in the core of the problem. We have proven several properties of $f(x)$. The inverse function of $f(x)$ is $f^{-1} = -W(-e^{-x})$ ($x \geq 1$). This allows us to conduct further analysis in all other parts of this paper. Another foundamental property is the relation between $f(x)$ and $f(\frac{1}{x})$, which provides a base for subsequent steps.
4) *Concentrating ε'*. In problem $\boldsymbol{P_2}$, the supremum of $\bar{\boldsymbol{f}}(\frac{1}{x_1}, \ldots, \frac{1}{x_n})$ is affected by the domain of each dimension, which is in turn determined by how ε' is allocated to these dimensions. We call $(\varepsilon_1, \cdots, \varepsilon_n)$ where $\sum_1^n \varepsilon_i = \varepsilon$ as an *allocation*. We prove that $\bar{\boldsymbol{f}}(\frac{1}{x_1}, \ldots, \frac{1}{x_n})$ takes its maximum when ε' is allocated to only one dimension (*i.e.*, an "extreme" allocation). In other words, there exists one $\varepsilon_j = \varepsilon$ and $\varepsilon_i = 0 (i \neq j)$. The key idea is to prove the convexity of function $\Delta(\varepsilon) = f(\frac{1}{w_1(\varepsilon)}) - f(w_1(\varepsilon))$.

We put the key steps of proof of Theorem 1 into Lemma 2 and Lemma 3. After that, we give the main proof.

Lemma 2. Given n-ary function $\bar{\boldsymbol{f}}(\boldsymbol{x}) = \bar{\boldsymbol{f}}(x_1, \ldots, x_n) = \sum_{i=1}^n x_i - \log x_i$ ($x_i \in \mathbb{R}^{++}$), if $\bar{\boldsymbol{f}}(x_1, \ldots, x_n) \leq n + \varepsilon (\varepsilon > 0)$, then

$$\begin{aligned}&\sup \bar{\boldsymbol{f}}(\frac{1}{x_1}, \ldots, \frac{1}{x_n}) \\ =&\frac{1}{-W_0(-e^{-(1+\varepsilon)})} - \log \frac{1}{-W_0(-e^{-(1+\varepsilon)})} + n - 1\end{aligned} \tag{15}$$

The supremum is attained when there exists only one j such that $f(x_j) = 1 + \varepsilon$ and $f(x_i) = 1$ for $i \neq j$.

Proof 2. We want to solve the following optimization problem analytically.

$$\text{maximize } \bar{\boldsymbol{f}}(\frac{1}{x_1}, \ldots, \frac{1}{x_n}) \tag{16}$$

$$s.t.\ \bar{\boldsymbol{f}}(x_1, \ldots, x_n) \leq n + \varepsilon \tag{17}$$

Since $f(x) \geq 1$, the constraint $\bar{\boldsymbol{f}}(x_1, \ldots, x_n) = \sum_{i=1}^n f(x_i) = \sum_{i=1}^n x_i - \log x_i \leq n + \varepsilon$ can be replaced by the following constraints

$$\left(\bigwedge_{i=1}^n f(x_i) = x_i - \log x_i \leq 1 + \varepsilon_i\right) \wedge \left(\bigwedge_{i=1}^n \varepsilon_i \geq 0\right) \wedge \sum_{i=1}^n \varepsilon_i \leq \varepsilon \tag{18}$$

Given fixed $\varepsilon_1, \ldots, \varepsilon_n$ such that $\bigwedge_{i=1}^n \varepsilon_i \geq 0 \wedge \sum_{i=1}^n \varepsilon_i \leq \varepsilon$, we define

$$\begin{aligned}\bar{\boldsymbol{S}}(\varepsilon_1, \ldots, \varepsilon_n) &= \sup_{\bigwedge_{i=1}^n f(x_i) \leq 1 + \varepsilon_i} \bar{\boldsymbol{f}}(\frac{1}{x_1}, \ldots, \frac{1}{x_n}) \\ &= \sum_{i=1}^n \sup_{f(x_i) \leq 1 + \varepsilon_i} f(\frac{1}{x_i}) = \sum_{i=1}^n S(\varepsilon_i)\end{aligned} \tag{19}$$

So we have

$$\sup \bar{\boldsymbol{f}}(\frac{1}{x_1}, \ldots, \frac{1}{x_n}) = \sup_{\substack{\bigwedge_{i=1}^n \varepsilon_i \geq 0 \\ \sum_{i=1}^n \varepsilon_i \leq \varepsilon}} \bar{\boldsymbol{S}}(\varepsilon_1, \ldots, \varepsilon_n) \tag{20}$$

It is easy to know that $\bar{\boldsymbol{S}}(\varepsilon_1, \ldots, \varepsilon_n)$ is continuous and strictly increasing with $\varepsilon_1, \ldots, \varepsilon_n$. So the condition $\sum_{i=1}^n \varepsilon_i \leq \varepsilon$ in Equation (20) can be changed to $\sum_{i=1}^n \varepsilon_i = \varepsilon$. The remaining proof consists of two steps. In step 1, we find $\bar{\boldsymbol{S}}(\varepsilon_1, \ldots, \varepsilon_n)$ for fixed $\varepsilon_1, \ldots, \varepsilon_n$. In step 2, we find $\sup \bar{\boldsymbol{S}}(\varepsilon_1, \ldots, \varepsilon_n)$ for any $\varepsilon_1, \ldots, \varepsilon_n$ satisfying $\bigwedge_{i=1}^n \varepsilon_i \geq 0 \wedge \sum_{i=1}^n \varepsilon_i = \varepsilon$.

Step 1: According to Lemma 1g, for fixed ε_i we get

$$S(\varepsilon_i) = \sup_{f(x) \leq 1 + \varepsilon_i} f(\frac{1}{x}) = f(\frac{1}{w_1(\varepsilon_i)}) \tag{21}$$

Plugging Equation (21) into Equation (19), we get

$$\bar{\boldsymbol{S}}(\varepsilon_1, \ldots, \varepsilon_n) = \sum_{i=1}^n f(\frac{1}{w_1(\varepsilon_i)}) \tag{22}$$

Step 2: We define function

$$\Delta(\varepsilon) = f(\frac{1}{w_1(\varepsilon)}) - f(w_1(\varepsilon)) = \frac{1}{w_1(\varepsilon)} - w_1(\varepsilon) + 2\log w_1(\varepsilon) \tag{23}$$

Now we prove

$$\Delta(t\varepsilon) \leq t\Delta(\varepsilon)\ (0 \leq t < 1) \tag{24}$$

When $\varepsilon = 0$, it is trivial to verify that $\Delta(0) = 0$. In the following we show that $\Delta(\varepsilon)$ is strictly increasing and strictly convex. It is easy to know $\frac{\mathrm{d}\Delta(\varepsilon)}{\mathrm{d}w_1} = -\frac{1}{w_1^2} + \frac{2}{w_1} - 1$. Combining Lemma 1e, the derivative of $\Delta(\varepsilon)$ is

$$\begin{aligned}\frac{\mathrm{d}\Delta(\varepsilon)}{\mathrm{d}\varepsilon} &= \frac{\mathrm{d}\Delta(\varepsilon)}{\mathrm{d}w_1} \times \frac{\mathrm{d}w_1(\varepsilon)}{\mathrm{d}\varepsilon} \\ &= \left(-\frac{1}{w_1(\varepsilon)^2} + \frac{2}{w_1(\varepsilon)} - 1\right) \times \frac{-w_1(\varepsilon)}{1 - w_1(\varepsilon)} = \frac{1}{w_1(\varepsilon)} - 1\end{aligned}$$

The second order derivative of $\Delta(\varepsilon)$ is

$$\frac{\mathrm{d}^2\Delta(\varepsilon)}{\mathrm{d}\varepsilon^2} = -\frac{1}{w_1(\varepsilon)^2} \frac{-w_1(\varepsilon)}{1 - w_1(\varepsilon)} = \frac{1}{w_1(\varepsilon)(1 - w_1(\varepsilon))}$$

Since $w_1(\varepsilon) \in (0, 1)$ for $\varepsilon > 0$, it is easy to know $\frac{\mathrm{d}\Delta(\varepsilon)}{\mathrm{d}\varepsilon} > 0$, $\frac{\mathrm{d}^2\Delta(\varepsilon)}{\mathrm{d}\varepsilon^2} > 0$ for $\varepsilon > 0$. This indicates that $\Delta(\varepsilon)$ is strictly increasing and strictly convex on $(0, +\infty)$. Thus, for any $\varepsilon', \varepsilon'' > 0$, we have $\Delta((1-t)\varepsilon' + t\varepsilon'') < (1-t)\Delta(\varepsilon') + t\Delta(\varepsilon'')$

for any $0 < t < 1$. Remember that we have known $\Delta(0) = 0$. Since $\Delta(\varepsilon)$ is continuous, it is easy to know

$$\begin{aligned}\Delta(t\varepsilon'') &= \lim_{\varepsilon'\to 0}\Delta((1-t)\varepsilon' + t\varepsilon'') \\ &\le \lim_{\varepsilon'\to 0}(1-t)\Delta(\varepsilon') + t\Delta(\varepsilon'') = t\Delta(\varepsilon'') \end{aligned} \tag{25}$$

Thus, we can obtain Equation (24).

Therefore, for any $\varepsilon_1, \ldots, \varepsilon_n$ satisfying $\bigwedge_{i=1}^n \varepsilon_i \ge 0 \wedge \sum_{i=1}^n \varepsilon_i = \varepsilon$, we have

$$\begin{aligned}\bar{\mathbf{\Delta}}(\varepsilon_1, \ldots, \varepsilon_n) &= \sum_{i=1}^n f(\frac{1}{w_1(\varepsilon_i)}) - f(w_1(\varepsilon_i)) = \sum_{i=1}^n \Delta(\varepsilon_i) \\ &= \sum_{i=1}^n \Delta(\frac{\varepsilon_i}{\varepsilon}\varepsilon) \le \sum_{i=1}^n \frac{\varepsilon_i}{\varepsilon}\Delta(\varepsilon) = \Delta(\varepsilon) \end{aligned} \tag{26}$$

Inequality (26) is tight when there exists only one j such that $\varepsilon_j = \varepsilon$ and $\varepsilon_i = 0$ for all $i \ne j$. This means that for any $\varepsilon_1, \ldots, \varepsilon_n$ satisfying $\bigwedge_{i=1}^n \varepsilon_i \ge 0 \wedge \sum_{i=1}^n \varepsilon_i = \varepsilon$, the following inequality holds.

$$\begin{aligned}&\bar{\boldsymbol{S}}(\varepsilon_1, \ldots, \varepsilon_n) \\ &= \textstyle\sum_{i=1}^n f(\frac{1}{w_1(\varepsilon_i)}) && \text{(22)} \\ &\le \Delta(\varepsilon) + \textstyle\sum_{i=1}^n f(w_1(\varepsilon_i)) && \text{(26)} \\ &\le \frac{1}{w_1(\varepsilon)} - \log\frac{1}{w_1(\varepsilon)} - (w_1(\varepsilon) - \log w_1(\varepsilon)) && \text{Lemma 1d (23)} \\ &\quad + \textstyle\sum_{i=1}^n (1+\varepsilon_i) \\ &= \frac{1}{w_1(\varepsilon)} - \log\frac{1}{w_1(\varepsilon)} - (1+\varepsilon) + n + \varepsilon \\ &= \frac{1}{w_1(\varepsilon)} - \log\frac{1}{w_1(\varepsilon)} + n - 1 \\ &= \frac{1}{-W_0(-e^{-(1+\varepsilon)})} - \log\frac{1}{-W_0(-e^{-(1+\varepsilon)})} + n - 1 \end{aligned} \tag{27}$$

Finally, we have

$$\begin{aligned}\sup \bar{\boldsymbol{f}}(\frac{1}{x_1}, \ldots, \frac{1}{x_n}) &= \sup_{\substack{\bigwedge_{i=1}^n \varepsilon_i \ge 0 \\ \sum_{i=1}^n \varepsilon_i \le \varepsilon}} \bar{\boldsymbol{S}}(\varepsilon_1, \ldots, \varepsilon_n) \\ &= \frac{1}{-W_0(-e^{-(1+\varepsilon)})} - \log\frac{1}{-W_0(-e^{-(1+\varepsilon)})} + n - 1 \end{aligned} \tag{28}$$

$\bar{\boldsymbol{f}}(1/x_1, \ldots, 1/x_n)$ reaches its supremum when there exists only one j such that $f(x_j) = 1 + \varepsilon$ and $f(x_i) = 1$ for $i \ne j$.

□

In the following Lemma 3, we deal with the case when one Gaussian is standard. Then we extend Lemma 3 to general case.

Lemma 3. Let $\mathcal{N}(0, I)$ be standard Gaussian, ε be a positive number. For any n-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

(a) If $KL(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})||\mathcal{N}(0, I)) \le \varepsilon$, then

$$\begin{aligned}&KL(\mathcal{N}(0, I)||\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \\ &\le \frac{1}{2}\left(\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - \log\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - 1\right)\end{aligned}$$

(b) If $KL(\mathcal{N}(0, I)||\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \le \varepsilon$, then

$$\begin{aligned}&KL(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})||\mathcal{N}(0, I)) \\ &\le \frac{1}{2}\left(\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - \log\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - 1\right)\end{aligned}$$

Proof 3. (a) According to the definition of KL divergence, we have

$$\begin{aligned}KL(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})||\mathcal{N}(0, I)) &= \frac{1}{2}\left(-\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top\boldsymbol{\mu} - n\right) \\ KL(\mathcal{N}(0, I)||\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) &= \frac{1}{2}\left(\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - n\right)\end{aligned}$$

where n is the dimension of the distribution. The positive definite matrix $\boldsymbol{\Sigma}$ has factorization $\boldsymbol{\Sigma} = PDP^\top$ where P is an orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{\Sigma}$, $D = diag(\lambda_1, \ldots, \lambda_n)$ ($\lambda_i > 0$) whose diagonal elements are the corresponding eigenvalues. We also have

$$|\boldsymbol{\Sigma}| = |P||D||P^\top| = |D| = \prod_{i=1}^n \lambda_i \tag{29}$$

$$\log|\boldsymbol{\Sigma}| = \sum_{i=1}^n \log\lambda_i, -\log|\boldsymbol{\Sigma}| = \sum_{i=1}^n \log\frac{1}{\lambda_i} \tag{30}$$

$$\mathrm{Tr}(\boldsymbol{\Sigma}) = \mathrm{Tr}(PDP^\top) = \mathrm{Tr}(P^\top PD) = \mathrm{Tr}(D) = \sum_{i=1}^n \lambda_i \tag{31}$$

$$\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) = \sum_{i=1}^n \lambda_i' = \sum_{i=1}^n \frac{1}{\lambda_i} \tag{32}$$

where $\lambda_i' = 1/\lambda_i$ are eigenvalues of $\boldsymbol{\Sigma}^{-1}$.

If $KL(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})||\mathcal{N}(0, I)) \le \varepsilon$, we have $-\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top\boldsymbol{\mu} - n \le 2\varepsilon$. This condition is equal to the following conditions

$$-\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^n \lambda_i - \log\lambda_i \le n + \varepsilon_1 \tag{33}$$

$$\boldsymbol{\mu}^\top\boldsymbol{\mu} \le 2\varepsilon - \varepsilon_1 \tag{34}$$

$$0 \le \varepsilon_1 \le 2\varepsilon \tag{35}$$

In the following, we find the maximum of $\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1})$ and $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, respectively. From Equation (33), we have

$$\sum_{i=1}^n \lambda_i - \log\lambda_i \le n + \varepsilon_1 \tag{36}$$

Applying Lemma 2 on Inequality (36), we can obtain

$$\begin{aligned}&\sum_{i=1}^n \frac{1}{\lambda_i} - \log\frac{1}{\lambda_i} = \log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) \\ &\le \frac{1}{-W_0(-e^{-(1+\varepsilon_1)})} - \log\frac{1}{-W_0(-e^{-(1+\varepsilon_1)})} + n - 1\end{aligned} \tag{37}$$

Moreover, since $f(x) = x - \log x$ takes the minimum value $f(1) = 1$ at $x = 1$, it is easy to know $\lambda_i - \log\lambda_i \le 1 + \varepsilon_1$ from Inequality (36). According to Lemma 1g, we know

$$w_1(\varepsilon_1) \le \lambda_i \le w_2(\varepsilon_1),\ \frac{1}{w_2(\varepsilon_1)} \le \lambda_i' = \frac{1}{\lambda_i} \le \frac{1}{w_1(\varepsilon_1)} \tag{38}$$

We also have $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \le \lambda'^*\boldsymbol{\mu}^\top\boldsymbol{\mu}$ where λ'^* is the maximum eigenvalue of $\boldsymbol{\Sigma}^{-1}$. Combining Equation (34) and (38), we can know

$$\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \le \lambda'^*(2\varepsilon - \varepsilon_1) \le \frac{2\varepsilon - \varepsilon_1}{w_1(\varepsilon_1)} \tag{39}$$

Now note that Inequalities (37) and (39) are tight simultaneously when there exists one $\lambda_j = w_1(\varepsilon_1)$ and all other

$\lambda_i = 1$ for $i \neq j$. Thus, we can add the two sides of Inequalities (37) and (39) and get

$$\begin{aligned} &KL(\mathcal{N}(0,I)||\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})) \\ =&\frac{1}{2}\left(\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - n\right) \\ \leq&\frac{1}{2}\left(\frac{1}{-W_0(-e^{-(1+\varepsilon_1)})} - \log\frac{1}{-W_0(-e^{-(1+\varepsilon_1)})} + n - 1\right. \\ &\left.+\frac{2\varepsilon-\varepsilon_1}{w_1(\varepsilon_1)} - n\right) \\ =&\frac{1}{2}\left(\frac{1+2\varepsilon-\varepsilon_1}{w_1(\varepsilon_1)} - \log\frac{1}{w_1(\varepsilon_1)} - 1\right) \\ =&U(\varepsilon_1)\ (0 \leq \varepsilon_1 \leq 2\varepsilon) \end{aligned} \tag{40}$$

Notice that the derivative of $U(\varepsilon_1)$ is

$$\begin{aligned} U'(\varepsilon_1) =&\frac{1}{2}\left(\frac{w_1(\varepsilon_1)+2\varepsilon-\varepsilon_1}{w_1(\varepsilon_1)(1-w_1(\varepsilon_1))} - \frac{1}{1-w_1(\varepsilon_1)}\right) \\ =&\frac{1}{2}\frac{2\varepsilon-\varepsilon_1}{w_1(\varepsilon_1)(1-w_1(\varepsilon_1))} \end{aligned} \tag{41}$$

Since $w_1(\varepsilon_1) \in (0,1)$ for $\varepsilon_1 > 0$ and $0 \leq \varepsilon_1 \leq 2\varepsilon$, we can know $U'(\varepsilon_1) \geq 0$ for $\varepsilon_1 > 0$. Thus, $U(\varepsilon_1)$ takes the maximum value at $\varepsilon_1 = 2\varepsilon$. Finally, we have

$$\begin{aligned} &KL(\mathcal{N}(0,I)||\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})) \\ \leq&U(2\varepsilon) = \frac{1}{2}\left(\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - \log\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - 1\right) \end{aligned} \tag{42}$$

Inequality (42) is tight only when there exists one $\lambda_j = -W_0(-e^{-(1+2\varepsilon)})$ and all other $\lambda_i = 1$ for $i \neq j$, and $|\boldsymbol{\mu}| = 0$. We can see that when ε is small, the right hand side of Equation (42) is also small.

(b) The proof of Theorem 3b is similar. See Appendix B for the details.

□

In the following, we extend Lemma 3 to general Gaussians. Before our generalized theorem, we recall the following proposition which states that diffeomorphism preserves KL divergence (f-divergence) [19].

Proposition 1. (See [19]) Let $\boldsymbol{z} = f(\boldsymbol{x})$ be a diffeomorphism, $X_1 \sim p_X$ and $X_2 \sim q_X$ be two random variables and $Z_1 = f(X_1) \sim p_Z$, $Z_2 = f(X_2) \sim q_Z$. Then $KL(p_X||q_X) = KL(p_Z||q_Z)$.

Main Proof of Theorem 1

Proof 4. With the help of Proposition 1, it is not hard to extend Lemma 3 to general Gaussians. The key idea is to use an invertible linear transformation to transform one Gaussian to standard Gaussian, and then apply Lemma 3. Please see Appendix C for details.

□

To investigate the bound in Theorem 1 further, we can expand Lambert W function using the series presented in [18], [20] for small ε. This is expressed by the following Theorem.

Theorem 2. For any two n-dimensional Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$, $\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$, and a small positive number ε, if $KL(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)) \leq \varepsilon$, then

$$KL(\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)||\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)) \leq \varepsilon + 2\varepsilon^{1.5} + O(\varepsilon^2) \tag{43}$$

Proof 5. Please see Appendix D for the details of the proof.

□

Theorem 1 holds for any two Gaussians $\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$. According to the proof of Theorem 1 (Lemma 3), one of $\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ can be fixed. It is not hard to extend Lemma 3 to case where the fixed one Gaussian is not standard. We can apply linear transformation (see Equation (155)) as what we have done in the main proof of Theorem 1 (see Appendix C). Therefore, we have the following corollary.

Corollary 1. Theorem 1 and Theorem 2 hold when one of $\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ is fixed.

Remark 1. The supremum in Theorem 1 has the following properties.

1) The supremum is small (zero) when ε is small (zero). Figure 1 shows some values of the supremum of KL divergence.
2) The supremum increases rapidly when $\varepsilon > 2$ due to the rapid increase of term $\frac{1}{-W_0(-e^{-(1+2\varepsilon)})}$.
3) It is hard to reach the supremum in typical applications (*e.g.*, in machine learning practice) due to the strict conditions.
4) The bound is independent of the dimension n. This is a critical property in high-dimensional problems.

Properties of supremum

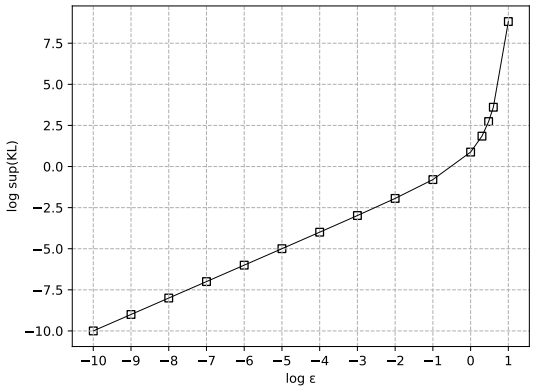

Fig. 1. Values of supremum of KL divergence shown on a logarithmic scale.

3.2 Infimum of Reverse KL Divergence Between Gaussians

We also want to know how small the reverse KL divergence could be when forward KL divergence is not less than a given number. In this subsection, we give the infimum of $KL(\mathcal{N}_2||\mathcal{N}_1)$ when $KL(\mathcal{N}_1||\mathcal{N}_2) \geq M$ ($M > 0$). The main result is shown in Theorem 3.

Theorem 3. For any two n-dimensional Gaussian distributionss $\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$, if $KL(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)) \geq M (M > 0)$, then

$$\begin{aligned} &KL(\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)||\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)) \\ \geq&\frac{1}{2}\left(\frac{1}{-W_{-1}(-e^{-(1+2M)})} - \log\frac{1}{-W_{-1}(-e^{-(1+2M)})} - 1\right) \end{aligned} \tag{44}$$

The infimum is attained when the following two conditions hold.

(1) There exists only one eigenvalue λ_j of $B_2^{-1}\Sigma_1(B_2^{-1})^\top$ or $B_1^{-1}\Sigma_2(B_1^{-1})^\top$ equal to $-W_{-1}(-e^{-(1+2M)})$ and all other eigenvalues λ_i $(i \neq j)$ are equal to 1, where $B_1 = P_1 D_1^{1/2}$, P_1 is an orthogonal matrix whose columns are the eigenvectors of Σ_1, $D_1 = diag(\lambda_1, \ldots, \lambda_n)$ whose diagonal elements are the corresponding eigenvalues, B_2 is defined in the same way as B_1 except on Σ_2.
(2) $\mu_1 = \mu_2$.

Proof of Theorem 3
Intuitively, the problems in Theorem 1 and 3 should has a tight relation. In this paper, we give two proofs of Theorem 3. The first proof has the similar structure as that of Theorem 1, except that Theorem 3 needs W_{-1}. We put the first proof in Appendix E. The second proof can be drawn from Theorem 1 directly by analyzing the supremum. We put the second proof below. These two proofs can verify each other.

Proof 6. In the following, we give the second proof which is drawn from Theorem 1.
Suppose $KL(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) \leq t$ $(t > 0)$, according to Theorem 1, we know

$$\begin{aligned} &KL(\mathcal{N}(\mu_2, \Sigma_2)||\mathcal{N}(\mu_1, \Sigma_1)) \\ \leq &\frac{1}{2}\left(\frac{1}{-W_0(-e^{-(1+2t)})} - \log\frac{1}{-W_0(-e^{-(1+2t)})} - 1\right) && (45) \\ = &\frac{1}{2}\left(\frac{1}{w_1(2t)} - \log\frac{1}{w_1(2t)} - 1\right) && (46) \\ = &\bar{S}(t) && (47) \end{aligned}$$

Since $\frac{1}{w_1(2t)}$ is strictly increasing with t, $\bar{S}(t)$ is continuous and strictly increasing with t. Besides, the range of function $\bar{S}(t)$ for $(t > 0)$ is $(0, +\infty)$.
Given positive number M, according to Theorem 1, there exists $\mathcal{N}(\mu_1, \Sigma_1)$, $\mathcal{N}(\mu_2, \Sigma_2)$ and m such that

$$\begin{aligned} \bar{S}(m) = &M && (48) \\ KL(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) = &M && (49) \\ KL(\mathcal{N}(\mu_2, \Sigma_2)||\mathcal{N}(\mu_1, \Sigma_1)) = &m && (50) \end{aligned}$$

Thus, given the precondition $KL(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) \geq M$, we can know that

$$\inf KL(\mathcal{N}(\mu_2, \Sigma_2)||\mathcal{N}(\mu_1, \Sigma_1)) \leq m \quad (51)$$

In the following, we show

$$\inf KL(\mathcal{N}(\mu_2, \Sigma_2)||\mathcal{N}(\mu_1, \Sigma_1)) = m \quad (52)$$

must holds. Otherwise, there exists an $m' < m$ and $\mathcal{N}(\mu_1, \Sigma_1)$, $\mathcal{N}(\mu_2, \Sigma_2)$ such that

$$\begin{aligned} KL(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) \geq &M && (53) \\ KL(\mathcal{N}(\mu_2, \Sigma_2)||\mathcal{N}(\mu_1, \Sigma_1)) = &m' && (54) \end{aligned}$$

Applying Theorem 1 on Equation (54), it is easy to know

$$\sup KL(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) = \bar{S}(m') \quad (55)$$

This contradicts with the precondition $KL(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) \geq M$ because $\bar{S}(m') < \bar{S}(m) = M$. Thus, Equation (52) holds.

Now we can solve m from $\bar{S}(m) = M$ as follows.

$$\begin{aligned} &\frac{1}{2}\left(\frac{1}{-W_0(-e^{-(1+2m)})} - \log\frac{1}{-W_0(-e^{-(1+2m)})} - 1\right) = M \\ \Leftrightarrow &\frac{1}{-W_0(-e^{-(1+2m)})} - \log\frac{1}{-W_0(-e^{-(1+2m)})} = 1 + 2M \\ \Leftrightarrow &\frac{1}{-W_0(-e^{-(1+2m)})} = -W_{-1}(-e^{-(1+2M)}) \\ \Leftrightarrow &\frac{1}{-W_{-1}(-e^{-(1+2M)})} = -W_0(-e^{-(1+2m)}) \\ \Leftrightarrow &\frac{1}{-W_{-1}(-e^{-(1+2M)})} - \log\frac{1}{-W_{-1}(-e^{-(1+2M)})} = 1 + 2m \\ \Leftrightarrow &m = \frac{1}{2}\left(\frac{1}{-W_{-1}(-e^{-(1+2M)})} - \log\frac{1}{-W_{-1}(-e^{-(1+2M)})} - 1\right) \end{aligned} \quad (56)$$

where the third and fifth equations follow from Lemma 1d. Plugging Equation (56) into (52), we can prove Theorem 3.

□

Remark 2. The bound in Theorem 3 has the similar form with that in Theorem 1. In fact, Theorem 1 and Theorem 3 forms a duality. Firstly, these two theorems can be proved independently in the similar way. Secondly, these two theorems can be derived from each other.

4 Relaxed Triangle Inequality

Until now, we have quantified the approximate symmetry of KL divergence between Gaussians. A natural question is how large can $KL(\mathcal{N}_1||\mathcal{N}_3)$ be when $KL(\mathcal{N}_1||\mathcal{N}_2)$ and $KL(\mathcal{N}_2)||\mathcal{N}_3)$ are small for three Gaussians $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$. In this section, we give a bound of $KL(\mathcal{N}_1||\mathcal{N}_3)$ that is also independent of the dimension n. Proving the relaxed triangle inequality is more difficult. The main result is presented in Theorem 4. We put the key steps of proof of Theorem 4 in Lemma 4 ~ 7 and Lemma 9.

Theorem 4. For any three n-dimensional Gaussians $\mathcal{N}(\mu_i, \Sigma_i)(i \in \{1, 2, 3\})$ such that $KL(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) \leq \varepsilon_1$ and $KL(\mathcal{N}(\mu_2, \Sigma_2)||\mathcal{N}(\mu_3, \Sigma_3)) \leq \varepsilon_2$ for $\varepsilon_1, \varepsilon_2 \geq 0$, then

$$\begin{aligned} &KL((\mathcal{N}(\mu_1, \Sigma_1)||\Sigma(\mu_3, \Sigma_3)) \\ <&\varepsilon_1 + \varepsilon_2 + \frac{1}{2}\Bigg(W_{-1}(-e^{-(1+2\varepsilon_1)})W_{-1}(-e^{-(1+2\varepsilon_2)}) && (57) \\ &+ W_{-1}(-e^{-(1+2\varepsilon_1)}) + W_{-1}(-e^{-(1+2\varepsilon_2)}) + 1 && (58) \\ &-W_{-1}(-e^{-(1+2\varepsilon_2)})\left(\sqrt{2\varepsilon_1} + \sqrt{\frac{2\varepsilon_2}{-W_0(-e^{-(1+2\varepsilon_2)})}}\right)^2\Bigg) && (59) \end{aligned}$$

Overview of proof of Theorem 4
We want to solve the following optimzation problem $\boldsymbol{P}_3$ analytically.

$$\begin{aligned} \text{maximize } &KL(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\Sigma(\mu_3, \Sigma_3)) \\ s.t. \ &KL(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) \leq \varepsilon_1 \\ &KL(\mathcal{N}(\mu_2, \Sigma_2)||\mathcal{N}(\mu_3, \Sigma_3)) \leq \varepsilon_2 \end{aligned}$$

Unfortunately, it is hard to find the supremum due to the complexity caused by Lambert W function. So we relax the

constraints to simplify the problem. Our proof consists of the following several steps.

1) *Invertible linear transformation*. The first step is similar to that of Theorem 1. We use a linear transformation on $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$ to simply the problem. After transformation, $\mathcal{N}_2$ is converted to standard Gaussian.
2) *Relaxing constraints*. In this step, we relax the constraints to get a simpler problem, which is in turn reduced to the following core problem $\boldsymbol{P}_4$.

Steps to derive the formula

$$\begin{aligned}\text{maximize } & \sum_{i=1}^{n} \lambda_{1,[i]}\lambda'_{2,[i]} - \log \lambda_{1,[i]}\lambda'_{2,[i]} \qquad (60)\\ s.t.\ & \lambda_{1,[i]} - \log \lambda_{1,[i]} = 1 + \varepsilon_{1,[i]}\ (1 \le i \le n)\\ & \bigwedge_{i=1}^{n} \varepsilon_{1,[i]} \ge 0 \wedge \sum_{i=1}^{n} \varepsilon_{1,[i]} = 2\varepsilon_1\\ & \lambda'_{2,[i]} - \log \lambda'_{2,[i]} = 1 + \varepsilon_{2,[i]}\ (1 \le i \le n)\\ & \bigwedge_{i=1}^{n} \varepsilon_{2,[i]} \ge 0 \wedge \sum_{i=1}^{n} \varepsilon_{2,[i]} = 2\varepsilon_2\end{aligned}$$

where $\lambda_{1,[i]}, \lambda'_{2,[i]}$ are the eigenvalues of $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2^{-1}$ arranged in decreasing order, respectively, and $\varepsilon_{1,[i]}, \varepsilon_{2,[i]}$ are arranged in decreasing order too.

3) *Concentrating ε_1 and ε_2*. The objective function (60) is determined by how ε_1 and ε_2 are allocated to $(\varepsilon_{1,[1]}, \cdots, \varepsilon_{1,[n]})$ and $(\varepsilon_{2,[1]}, \cdots, \varepsilon_{2,[n]})$. We prove that an "extreme allocation" can make the objective function maximized. In other words, Equation (60) takes its maximum when $\varepsilon_{1,[1]} = \varepsilon_1$ and $\varepsilon_{2,[1]} = \varepsilon_2$. We use a key Lemma 7 to deal with the two dimensional case (*i.e.*, $n = 2$). Then, we use Lemma 8 to extend the conclusion to high dimensional problems.
Lemma 7 is the most tricky part in this paper. In the proof, concentrating ε_1 and ε_2 is much harder than that in last section for Theorem 1. $f(x) = x - \log x$ is a transcendental function whose inverse function is expressed by Lambert W function. This makes even a 2-dimensional case of problem $\boldsymbol{P}_4$ hard to solve. We use an iterated way to prove Lemma 7. We prove that, for any fixed "non-extreme allocation" $(\varepsilon_{1,[1]}, \varepsilon_{1,[2]})$ (*i.e.*, $\varepsilon_{1,[2]} > 0$), there is a "more extreme" allocation $(\varepsilon_{2,[1]}, \varepsilon_{2,[2]})$ that can make the objective function maximized. Then we fix $(\varepsilon_{2,[1]}, \varepsilon_{2,[2]})$ and find a more extreme allocation $(\varepsilon'_{1,[1]}, \varepsilon'_{1,[2]})$ to lift the objective function further. Using these iterations, we can find an infinite sequence of allocations which can make the objective function reach its supremum when the allocation is an extreme one.

Notations. Before the proof, we define the following helper functions based on $f(x) = x - \log x$.

$$f_l(x) = f(1-x) - 1\ (0 \le x < 1),\ f_r(x) = f(x+1) - 1\ (x \ge 0) \qquad (61)$$

The derivatives of $f_l(x), f_r(x)$ are

$$f'_l(x) = -f'(1-x) = \frac{1}{1-x} - 1,\ f'_r(x) = f'(1+x) = 1 - \frac{1}{x+1} \qquad (62)$$

So both $f_l(x)$ and $f_r(x)$ are strictly increasing. We note the inverse functions of f_l, f_r as g_l, g_r, respectively. Combining Lemma 1c, it is not hard to verify that g_l, g_r are

$$g_l(\varepsilon) = f_l^{-1}(\varepsilon) = 1 - w_1(\varepsilon) = 1 + W_0(-e^{-(1+\varepsilon)})\ (\varepsilon \ge 0) \qquad (63)$$

$$g_r(\varepsilon) = f_r^{-1}(\varepsilon) = w_2(\varepsilon) - 1 = -W_{-1}(-e^{-(1+\varepsilon)}) - 1\ (\varepsilon \ge 0) \qquad (64)$$

According to Lemma 1e, the derivatives of g_l, g_r are

$$g'_l(\varepsilon) = f_l^{-1\prime}(\varepsilon) = \frac{w_1(\varepsilon)}{1 - w_1(\varepsilon)} = \frac{1 - f_l^{-1}(\varepsilon)}{f_l^{-1}(\varepsilon)} = \frac{1}{1 - w_1(\varepsilon)} - 1 \qquad (65)$$

$$g'_r(\varepsilon) = f_r^{-1\prime}(\varepsilon) = \frac{w_2(\varepsilon)}{w_2(\varepsilon) - 1} = \frac{f_r^{-1}(\varepsilon) + 1}{f_r^{-1}(\varepsilon)} = 1 + \frac{1}{w_2(\varepsilon) - 1} \qquad (66)$$

Specially, since $\lim_{\varepsilon \to 0} w_2(\varepsilon) = w_2(0) = 1$, it is easy to know

$$\lim_{\varepsilon \to 0} g'_r(\varepsilon) = +\infty \qquad (67)$$

In the following, we note $g'_r(0) = +\infty$ for convenience.

Lemma 4 gives two useful conclusions for subsequent analysis. They hold apparently.

Lemma 4. Let a, b, a^+, b^- be positive real numbers.
(a) if $a > b, a < a^+, b > b^-$, then $\frac{a+1}{b+1} < \frac{a^+ + 1}{b^- + 1}$.
(b) if $a \le b$, then $\frac{a(b+1)}{b(a+1)} \le 1$.

Lemma 5. Given $f(x) = x - \log x$ and $\varepsilon \ge 0$, then $w_2(\varepsilon) - 1 \ge 1 - w_1(\varepsilon)$ holds and the inequality is tight when $\varepsilon = 0$;

Proof 7. The details of the proof are shown in Appendix F. □

Lemma 6. Given $f(x) = x - \log x$ and $\varepsilon_x, \varepsilon_y \ge 0$, if $f(x) \le 1 + \varepsilon_x$ and $f(y) \le 1 + \varepsilon_y$, then

$$f(xy) \le f(w_2(\varepsilon_x) w_2(\varepsilon_y)) \qquad (68)$$

Proof 8. The details of the proof are shown in Appendix G. □

In Lemma 2 in the last section (and Lemma 10 in Section E in Supplementary material), we eliminate the dimension n from the bound by showing the convexity of constructed function. Unfortunately, the relaxed triangle inequality involves three Gaussians which make the analysis more complex. The following Lemma 7 is the core of proof of the relaxed triangle inequality theorem. It is the most techinical part in this paper. We will use Lemma 7 to make the bound in Theorem 4 independent of the dimension n.

Lemma 7. Given $f(x) = x - \log x$, let $\varepsilon_{x,1}, \varepsilon_{x,2}, \varepsilon_{y,1}, \varepsilon_{y,2}$ be four non-negative numbers such that $\varepsilon_{x,1} \ge \varepsilon_{x,2}, \varepsilon_{y,1} \ge \varepsilon_{y,2}$. Then

$$\begin{aligned}& f(w_2(\varepsilon_{x,1}) w_2(\varepsilon_{y,1})) + f(w_2(\varepsilon_{x,2}) w_2(\varepsilon_{y,2}))\\ \le & f(w_2(\varepsilon_{x,1} + \varepsilon_{x,2}) w_2(\varepsilon_{y,1} + \varepsilon_{y,2})) + 1\end{aligned} \qquad (69)$$

Overview of proof of Lemma 7

In the overview of proof of Theorem 4 in the beginning of Section 4, we have mentioned Lemma 7. In the left hand side of Inequality (69), $\varepsilon_{x,2}$ and $\varepsilon_{y,2}$ stay in the second term. Intuitively, we use Inequality (69) to move $\varepsilon_{x,2}, \varepsilon_{y,2}$ into the first item. It is hard to prove Inequality (69) directly due to the lack of conclusions relating to Lambert W function. In

the proof, We use an iterative way to absorb $\varepsilon_{x,2}, \varepsilon_{y,2}$ into the first term gradually.

We treat

$$(\varepsilon_{x,1}+\theta_x\varepsilon_{x,2},\ \varepsilon_{x,2}-\theta_x\varepsilon_{x,2}) \text{ and } (\varepsilon_{y,1}+\theta_y\varepsilon_{y,2},\ \varepsilon_{y,2}-\theta_y\varepsilon_{y,2})$$

as two allocations, where θ_x and θ_y control how $\varepsilon_{x,2}$ and $\varepsilon_{y,2}$ are allocated among the two terms. The whole proof can be seen as an variation of coordinate ascent. In each iteration, we fix one of θ_x and θ_y (*i.e.*, one allocation) and make another one vary. The goal is to maximize the objective function (Equation (71)). In this way, we will construct an infinite sequence of allocations. The procedure is much harder than a simple coordinate ascent algorithm. The proof mainly consists of the following four aspects.

- **A1** In each step, once we fix one allocation and make another one vary, we prove there exists one and only one supremum.
- **A2** We find an equation to express above supremum implicitly.
- **A3** We prove the procedure is really lifting the objective function.
- **A4** We construct an infinit sequence of allocations. Then we prove the limit of the allocation sequence will make the objective function reach its supremum.

In this procedure, the hardest part is how to find a more extreme allocation based on the last one. There is no analytical solution to express these allocations. Luckily, we find a key equation to express the property of these allocations implicitly (see Equations (83), (97), (101)). Based on our analysis on such equation, we succeed to construct a sequence of allocations and finally prove Lemma 7.

Proof 9. Inequality (69) is equal to

$$\begin{aligned}&f(w_2(\varepsilon_{x,1})w_2(\varepsilon_{y,1}))+f(w_2(\varepsilon_{x,2})w_2(\varepsilon_{y,2}))\\ \le& f(w_2(\varepsilon_{x,1}+\varepsilon_{x,2})w_2(\varepsilon_{y,1}+\varepsilon_{y,2}))+f(w_2(0)w_2(0))\end{aligned} \tag{70}$$

We define function

$$\begin{aligned}S(\theta_x,\theta_y)=&f(w_2(\varepsilon_{x,1}+\theta_x\varepsilon_{x,2})w_2(\varepsilon_{y,1}+\theta_y\varepsilon_{y,2}))\\ &+f(w_2(\varepsilon_{x,2}-\theta_x\varepsilon_{x,2})w_2(\varepsilon_{y,2}-\theta_y\varepsilon_{y,2}))\end{aligned} \tag{71}$$

for $-\frac{\varepsilon_{x,1}}{\varepsilon_{x,2}} \le \theta_x \le 1, -\frac{\varepsilon_{y,1}}{\varepsilon_{y,2}} \le \theta_y \le 1$. The domains of θ_x, θ_y are restricted to make $w_2(\cdot)$ in the definition of $S(\theta_x,\theta_y)$ meaningful. Inequation (69) states that $S(0,0) \le S(1,1)$. We can prove $S(0,0) \le S(1,1)$ in the following three cases.

Case 1 $\varepsilon_{x,2}=\varepsilon_{y,2}=0$.

Case 2 $\varepsilon_{x,2}>0, \varepsilon_{y,2}>0$. In this case, we have $\varepsilon_{x,1} \ge \varepsilon_{x,2}>0, \varepsilon_{y,1}\ge\varepsilon_{y,2}>0$.

Case 3 only one of $\varepsilon_{x,2}$ and $\varepsilon_{y,2}$ equals to 0.

It is easy to verify that $S(0,0)=S(1,1)$ for **Case 1**. In the following, we discuss **Case 2** first and deal with **Case 3** at the end of the proof.

Case 2:

In $S(\theta_x,\theta_y)$, θ_x, θ_y are symmetric. Without loss of generality, we choose any $0<\theta_{x,0}<1$ at the beginning. The following proof consists of two steps. In **Step 1**, we prove that for any fixed $0<\theta_{x,0}<1$, there exists one and only one $-\frac{\varepsilon_{y,1}}{\varepsilon_{y,2}} < \theta_{y,1} < 1$ such that $S(\theta_{x,0},\theta_y)$ takes its maximum. This accomplishes aspects **A1** and **A2** in the first iteration. In **Step 2**, we prove $S(1,1)\ge S(0,0)$. The key idea is finding a strictly increasing sequence $\{S[i]\}$ such that $S[0]$ can be arbitrarily close to $S(0,0)$ and $\lim_{i\to\infty} S[i]=S(1,1)$. **Step 2** will accomplish aspects **A1** ~ **A4** in all iterations.

Step 1. At the beginning, we select any $0<\theta_{x,0}<1$. For brevity, we note

$$\tilde{\varepsilon}_{x,1}[0]=\varepsilon_{x,1}+\theta_{x,0}\varepsilon_{x,2},\ \tilde{\varepsilon}_{x,2}[0]=\varepsilon_{x,2}-\theta_{x,0}\varepsilon_{x,2} \tag{72}$$

$$\tilde{\varepsilon}_{y,1}=\varepsilon_{y,1}+\theta_y\varepsilon_{y,2},\ \tilde{\varepsilon}_{y,2}=\varepsilon_{y,2}-\theta_y\varepsilon_{y,2} \tag{73}$$

where we use $\tilde{\varepsilon}_{x,(\cdot)}[0]$ to denote the variable is computed with $\theta_{x,0}$.

Note that $g_r(\varepsilon)$ (defined in Equation (64)) is strictly increasing with ε. Combining the precondition $\varepsilon_{x,1}\ge\varepsilon_{x,2}$, we can know

$$\frac{g_r(\tilde{\varepsilon}_{x,1}[0])}{g_r(\tilde{\varepsilon}_{x,2}[0])}=\frac{g_r(\varepsilon_{x,1}+\theta_{x,0}\varepsilon_{x,2})}{g_r(\varepsilon_{x,2}-\theta_{x,0}\varepsilon_{x,2})}>\frac{g_r(\varepsilon_{x,1})}{g_r(\varepsilon_{x,2})}\ge 1 \tag{74}$$

We note this condition as $\boldsymbol{C_1}[0]$ as follows.

$$\boldsymbol{C_1}[0]:\frac{g_r(\tilde{\varepsilon}_{x,1}[0])}{g_r(\tilde{\varepsilon}_{x,2}[0])}>1 \tag{75}$$

Now given the fixed $\theta_{x,0}$, the derivative of $S(\theta_{x,0},\theta_y)$ is

$$\begin{aligned}&\frac{\mathrm{d}S(\theta_{x,0},\theta_y)}{\mathrm{d}\theta_y}\\ =&\varepsilon_{y,2}\left(w_2(\tilde{\varepsilon}_{x,1}[0])\frac{w_2(\tilde{\varepsilon}_{y,1})}{w_2(\tilde{\varepsilon}_{y,1})-1}-\frac{1}{w_2(\tilde{\varepsilon}_{y,1})}\frac{w_2(\tilde{\varepsilon}_{y,1})}{w_2(\tilde{\varepsilon}_{y,1})-1}\right)\\ &-\varepsilon_{y,2}\left(w_2(\tilde{\varepsilon}_{x,2}[0])\frac{w_2(\tilde{\varepsilon}_{y,2})}{w_2(\tilde{\varepsilon}_{y,2})-1}-\frac{1}{w_2(\tilde{\varepsilon}_{y,2})}\frac{w_2(\tilde{\varepsilon}_{y,2})}{w_2(\tilde{\varepsilon}_{y,2})-1}\right)\\ =&\varepsilon_{y,2}\left(\frac{w_2(\tilde{\varepsilon}_{x,1}[0])w_2(\tilde{\varepsilon}_{y,1})-1}{w_2(\tilde{\varepsilon}_{y,1})-1}-\frac{w_2(\tilde{\varepsilon}_{x,2}[0])w_2(\tilde{\varepsilon}_{y,2})-1}{w_2(\tilde{\varepsilon}_{y,2})-1}\right)\\ =&\varepsilon_{y,2}\left(\frac{w_2(\tilde{\varepsilon}_{x,1}[0])w_2(\tilde{\varepsilon}_{y,1})-w_2(\tilde{\varepsilon}_{x,1}[0])+w_2(\tilde{\varepsilon}_{x,1}[0])-1}{w_2(\tilde{\varepsilon}_{y,1})-1}\right.\\ &\left.-\frac{w_2(\tilde{\varepsilon}_{x,2}[0])w_2(\tilde{\varepsilon}_{y,2})-w_2(\tilde{\varepsilon}_{x,2}[0])+w_2(\tilde{\varepsilon}_{x,2}[0])-1}{w_2(\tilde{\varepsilon}_{y,2})-1}\right)\\ =&\varepsilon_{y,2}\left(\left(w_2(\tilde{\varepsilon}_{x,1}[0])+\frac{w_2(\tilde{\varepsilon}_{x,1}[0])-1}{w_2(\tilde{\varepsilon}_{y,1})-1}\right)\right.\\ &\left.-\left(w_2(\tilde{\varepsilon}_{x,2}[0])+\frac{w_2(\tilde{\varepsilon}_{x,2}[0])-1}{w_2(\tilde{\varepsilon}_{y,2})-1}\right)\right)\end{aligned} \tag{76}$$

The second order derivative is

$$\begin{aligned}&\frac{\mathrm{d}^2S(\theta_{x,0},\theta_y)}{\mathrm{d}\theta_y^2}\\ =&\frac{-(w_2(\tilde{\varepsilon}_{x,1}[0])-1)}{(w_2(\tilde{\varepsilon}_{y,1})-1)^2}\frac{w_2(\tilde{\varepsilon}_{y,1})}{w_2(\tilde{\varepsilon}_{y,1})-1}(\varepsilon_{y,2})^2\\ &-\frac{-(w_2(\tilde{\varepsilon}_{x,2}[0])-1)}{(w_2(\tilde{\varepsilon}_{y,2})-1)^2}\frac{w_2(\tilde{\varepsilon}_{y,2})}{w_2(\tilde{\varepsilon}_{y,2})-1}(-(\varepsilon_{y,2})^2)\\ =&-\frac{(w_2(\tilde{\varepsilon}_{x,1}[0])-1)w_2(\tilde{\varepsilon}_{y,1})(\varepsilon_{y,2})^2}{(w_2(\tilde{\varepsilon}_{y,1})-1)^3}\\ &-\frac{(w_2(\tilde{\varepsilon}_{x,2}[0])-1)w_2(\tilde{\varepsilon}_{y,2})(\varepsilon_{y,2})^2}{(w_2(\tilde{\varepsilon}_{y,2})-1)^3}\end{aligned} \tag{77}$$

Since $w_2(\varepsilon)>1$ for $\varepsilon>0$, it is easy to know $\frac{\mathrm{d}^2S(\theta_{x,0},\theta_y)}{\mathrm{d}\theta_y^2}<0$ for $\theta_y<1$. Thus we get the following proposition.

Proposition 2. $S(\theta_{x,0},\theta_y)$ is strictly concave and has at most one maximum for $\theta_y<1$.

Remember that we are discussing **Case 2**, so $\varepsilon_{y,2} > 0$. Now letting $\frac{\mathrm{d}S(\theta_{x,0},\theta_y)}{\mathrm{d}\theta_y} = 0$ (*i.e.*, Equation (76)= 0), we can obtain

$$\frac{\mathrm{d}S(\theta_{x,0},\theta_y)}{\mathrm{d}\theta_y} = 0 \Leftrightarrow$$
$$w_2(\tilde{\varepsilon}_{x,1}[0]) + \frac{w_2(\tilde{\varepsilon}_{x,1}[0]) - 1}{w_2(\tilde{\varepsilon}_{y,1}) - 1} = w_2(\tilde{\varepsilon}_{x,2}[0]) + \frac{w_2(\tilde{\varepsilon}_{x,2}[0]) - 1}{w_2(\tilde{\varepsilon}_{y,2}) - 1} \tag{78}$$

Now, it seems that the proof is stuck because we can not solve Equation (78) analytically. However, we succeed to go further by analyzing Equation (78). Our analysis starts from the following transformation in Equations (79) ~ (83), which is hard to obtain but easy to verify.

Using the notations of helper functions $g_r(\varepsilon) = f_r^{-1}(\varepsilon), g_r'(\varepsilon) = f_r^{-1\prime}(\varepsilon)$ in Equations (64) and (66), we can rewrite Equation (78) as follows.

$$\begin{aligned}
&\text{Equation (78)}\\
\Leftrightarrow& w_2(\tilde{\varepsilon}_{x,1}[0]) - 1 + \frac{w_2(\tilde{\varepsilon}_{x,1}[0]) - 1}{w_2(\tilde{\varepsilon}_{y,1}) - 1}\\
&= w_2(\tilde{\varepsilon}_{x,2}[0]) - 1 + \frac{w_2(\tilde{\varepsilon}_{x,2}[0]) - 1}{w_2(\tilde{\varepsilon}_{y,2}) - 1} &(79)\\
\Leftrightarrow&(w_2(\tilde{\varepsilon}_{x,1}[0]) - 1)\left(1 + \frac{1}{w_2(\tilde{\varepsilon}_{y,1}) - 1}\right)\\
&= (w_2(\tilde{\varepsilon}_{x,2}[0]) - 1)\left(1 + \frac{1}{w_2(\tilde{\varepsilon}_{y,2}) - 1}\right)\\
\Leftrightarrow& g_r(\tilde{\varepsilon}_{x,1}[0]) g_r'(\tilde{\varepsilon}_{y,1}) = g_r(\tilde{\varepsilon}_{x,2}[0]) g_r'(\tilde{\varepsilon}_{y,2})\\
\Leftrightarrow& \frac{g_r(\tilde{\varepsilon}_{x,1}[0])}{g_r(\tilde{\varepsilon}_{x,2}[0])} = \frac{g_r'(\tilde{\varepsilon}_{y,2})}{g_r'(\tilde{\varepsilon}_{y,1})} &(80)\\
\Leftrightarrow& \frac{g_r(\tilde{\varepsilon}_{x,1}[0])}{g_r(\tilde{\varepsilon}_{x,2}[0])} = \frac{\left(\frac{1}{g_r'(\tilde{\varepsilon}_{y,1})}\right)}{\left(\frac{1}{g_r'(\tilde{\varepsilon}_{y,2})}\right)}\\
\Leftrightarrow& \frac{g_r(\tilde{\varepsilon}_{x,1}[0])}{g_r(\tilde{\varepsilon}_{x,2}[0])} = \frac{\left(\frac{g_r(\tilde{\varepsilon}_{y,1})}{g_r(\tilde{\varepsilon}_{y,1})+1}\right)}{\left(\frac{g_r(\tilde{\varepsilon}_{y,2})}{g_r(\tilde{\varepsilon}_{y,2})+1}\right)} &(81)\\
\Leftrightarrow& \frac{g_r(\tilde{\varepsilon}_{x,1}[0])}{g_r(\tilde{\varepsilon}_{x,2}[0])} = \frac{g_r(\tilde{\varepsilon}_{y,1})}{g_r(\tilde{\varepsilon}_{y,2})}\frac{g_r(\tilde{\varepsilon}_{y,2})+1}{g_r(\tilde{\varepsilon}_{y,1})+1} &(82)\\
\Leftrightarrow& \frac{g_r(\tilde{\varepsilon}_{y,1})}{g_r(\tilde{\varepsilon}_{y,2})} = \frac{g_r(\tilde{\varepsilon}_{x,1}[0])}{g_r(\tilde{\varepsilon}_{x,2}[0])}\frac{g_r(\tilde{\varepsilon}_{y,1})+1}{g_r(\tilde{\varepsilon}_{y,2})+1} &(83)
\end{aligned}$$

where Equation (81) follows from Equation (66).

Up to now, we transform the condition $\frac{\mathrm{d}S(\theta_{x,0},\theta_y)}{\mathrm{d}\theta_y} = 0$ in Equation (78) to Equation (83). In the following, Equation (83) will be used to investigate the property of the maximum for $S(\theta_{x,0},\theta_y)$. The goal is to accomplish aspect **A2** of the proof.

In the following **Substep 1.1**, we show that Equation (78) must have one and only one solution. In other words, there must be one and only one point making $\frac{\mathrm{d}S(\theta_{x,0},\theta_y)}{\mathrm{d}\theta_y} = 0$. Unfortunately, it is hard to get an analytical solution from Equation (78) due to the complexity brought by Lambert W function. Therefore, in **Substep 1.2**, we analyze Equations (79) ~ (83) to investigate the properties of the solution. Overall, the analysis in **Step 1** will be used as a basic step in **Step 2**.

Substep 1.1. According to the definition of $g_r'(\varepsilon)$ in Equation (66), $g_r'(\varepsilon)$ is strictly decreasing with ε. So $g_r'(\tilde{\varepsilon}_{y,2}) = g_r'(\varepsilon_{y,2} - \theta_y\varepsilon_{y,2})$ is strictly increasing and $g_r'(\tilde{\varepsilon}_{y,1}) = g_r'(\varepsilon_{y,1} + \theta_y\varepsilon_{y,2})$ is strictly decreasing with θ_y. Thus, the right hand side of Equation (80) is continous and strictly increasing with θ_y. Besides, according to Equation (67) and the definition of $\tilde{\varepsilon}_{y,1}, \tilde{\varepsilon}_{y,2}$ in Equation (72) and (72), it is easy to know

$$\begin{aligned}
\lim_{\theta_y \to -\frac{\varepsilon_{y,1}}{\varepsilon_{y,2}}} \frac{g_r'(\tilde{\varepsilon}_{y,2})}{g_r'(\tilde{\varepsilon}_{y,1})} &= \lim_{\theta_y \to -\frac{\varepsilon_{y,1}}{\varepsilon_{y,2}}} \frac{g_r'(\varepsilon_{y,2} - \theta_y\varepsilon_{y,2})}{g_r'(\varepsilon_{y,1} + \theta_y\varepsilon_{y,2})}\\
&= \frac{g_r'(\varepsilon_{y,2} + \varepsilon_{y,1})}{g_r'(0)} = \frac{g_r'(\varepsilon_{y,2} + \varepsilon_{y,1})}{+\infty} = 0 &(84)\\
\lim_{\theta_y \to 1} \frac{g_r'(\tilde{\varepsilon}_{y,2})}{g_r'(\tilde{\varepsilon}_{y,1})} &= \lim_{\theta_y \to 1} \frac{g_r'(\varepsilon_{y,2} - \theta_y\varepsilon_{y,2})}{g_r'(\varepsilon_{y,1} + \theta_y\varepsilon_{y,2})}\\
&= \frac{g_r'(0)}{g_r'(\varepsilon_{y,1} + \varepsilon_{y,2})} = \frac{+\infty}{g_r'(\varepsilon_{y,1} + \varepsilon_{y,2})} = +\infty &(85)
\end{aligned}$$

So the range of the right hand side of Equation (80) is $(0, +\infty)$ when $-\frac{\varepsilon_{y,1}}{\varepsilon_{y,2}} < \theta_y < 1$.

Remember that we start from $0 < \theta_{x,0} < 1$, combining the precondition $\varepsilon_{x,1} \geq \varepsilon_{x,2}$ and the definitions of $\tilde{\varepsilon}_{x,1}[0], \tilde{\varepsilon}_{x,2}[0]$ in Equation (72) and (72), it is easy to know that the left hand side of Equation (80) is a positive constant number. Therefore, Equation (80) must has one and only one solution. We note such solution as $\theta_{y,1}$. Combining with Proposition 2, we can know that for any fixed $0 < \theta_{x,0} < 1$, there exists one and only one $-\frac{\varepsilon_{y,1}}{\varepsilon_{y,2}} < \theta_{y,1} < 1$ that maximize $S(\theta_{x,0},\theta_y)$.

Here note that we still have no guarantee for $\theta_{y,1} > 0$ currently.

Substep 1.2. We can investigate the property of the solution $\theta_{y,1}$ by analyzing Equation (80), (82) and (83). Firstly, we can show

$$\frac{g_r(\tilde{\varepsilon}_{y,1}[1])}{g_r(\tilde{\varepsilon}_{y,2}[1])} = \frac{g_r(\varepsilon_{y,1} + \theta_{y,1}\varepsilon_{y,2})}{g_r(\varepsilon_{y,2} - \theta_{y,1}\varepsilon_{y,2})} > 1 \tag{86}$$

by contradiction. Assume to the contrary that $\frac{g_r(\tilde{\varepsilon}_{y,1}[1])}{g_r(\tilde{\varepsilon}_{y,2}[1])} \leq 1$, then $\frac{g_r(\tilde{\varepsilon}_{y,1}[1])+1}{g_r(\tilde{\varepsilon}_{y,2}[1])+1} \geq \frac{g_r(\tilde{\varepsilon}_{y,1}[1])}{g_r(\tilde{\varepsilon}_{y,2}[1])}$. Combining condition $\boldsymbol{C_1}[0]$ in Equation (75), we can deduce

$$\frac{g_r(\tilde{\varepsilon}_{x,1}[0])}{g_r(\tilde{\varepsilon}_{x,2}[0])}\frac{g_r(\tilde{\varepsilon}_{y,1}[1])+1}{g_r(\tilde{\varepsilon}_{y,2}[1])+1} > \frac{g_r(\tilde{\varepsilon}_{y,1}[1])+1}{g_r(\tilde{\varepsilon}_{y,2}[1])+1} \geq \frac{g_r(\tilde{\varepsilon}_{y,1}[1])}{g_r(\tilde{\varepsilon}_{y,2}[1])} \tag{87}$$

This contradicts with Equation (83). Therefore, we can obtain the following condition $\boldsymbol{C_1}[1]$.

$$\boldsymbol{C_1}[1]: \frac{g_r(\tilde{\varepsilon}_{y,1}[1])}{g_r(\tilde{\varepsilon}_{y,2}[1])} > 1 \tag{88}$$

Secondly, according to condition $\boldsymbol{C_1}[1]$, it is easy to know

$$\frac{g_r(\tilde{\varepsilon}_{y,1}[1])}{g_r(\tilde{\varepsilon}_{y,2}[1])} > \frac{g_r(\tilde{\varepsilon}_{y,1}[1])+1}{g_r(\tilde{\varepsilon}_{y,2}[1])+1} > 1 \tag{89}$$

Now combining Equation (83) and (89), we can know the following condition[2] $\boldsymbol{C_2}[1]$ holds.

$$\boldsymbol{C_2}[1]: \frac{g_r(\tilde{\varepsilon}_{y,1}[1])}{g_r(\tilde{\varepsilon}_{y,2}[1])} > \frac{g_r(\tilde{\varepsilon}_{x,1}[0])}{g_r(\tilde{\varepsilon}_{x,2}[0])} \tag{90}$$

In summary, we start from any fixed $0 < \theta_{x,0} < 1$ making condition $\boldsymbol{C_1}[0]$ in Equation (75) hold. Using **Step 1**, we find the only one $-\frac{\varepsilon_{y,1}}{\varepsilon_{y,2}} < \theta_{y,1} < 1$ such that $S(\theta_{x,0}, \theta_y)$ takes its maximum at $\theta_{y,1}$ and conditions $\boldsymbol{C_1}[1]$ and $\boldsymbol{C_2}[1]$ hold.

Step 2. The deduction in **Step 1** can be iterated repeatedly due to the symmetry of θ_x and θ_y in $S(\theta_x, \theta_y)$. For consistency, at the begining of the iterations, we can choose any $\theta_y \neq \theta_{y,1}$ as $\theta_{y,0}$.

In the following, we use notations

$$\begin{aligned}
&\tilde{\varepsilon}_{x,1}[i] = \varepsilon_{x,1} + \theta_{x,i}\varepsilon_{x,2},\ \tilde{\varepsilon}_{x,2}[i] = \varepsilon_{x,2} - \theta_{x,i}\varepsilon_{x,2} \\
&\tilde{\varepsilon}_{y,1}[i] = \varepsilon_{y,1} + \theta_{y,i}\varepsilon_{y,2},\ \tilde{\varepsilon}_{y,2}[i] = \varepsilon_{y,2} - \theta_{y,i}\varepsilon_{y,2}
\end{aligned} \tag{91}$$

where $\tilde{\varepsilon}_{(\cdot),k}[i]$ $(k \in \{1,2\})$ is computed with $\theta_{(\cdot),i}$. Now we fix $\theta_{y,1}$ and make θ_x vary, then we repeat **Step 1** on θ_x. Note that, in the second iteration the condition $\boldsymbol{C_1}[1]$ plays the same role as condition $\boldsymbol{C_1}[0]$ plays in the first iteration. Therefore, we can find a $\theta_{x,2}$ such that the following conditions hold

$$S(\theta_{x,2}, \theta_{y,1}) > S(\theta_{x,0}, \theta_{y,1}) \tag{92}$$

$$\begin{aligned}
\frac{g_r(\tilde{\varepsilon}_{x,1}[2])}{g_r(\tilde{\varepsilon}_{x,2}[2])} &= \frac{g_r(\tilde{\varepsilon}_{y,1}[1])}{g_r(\tilde{\varepsilon}_{y,2}[1])} \frac{g_r(\tilde{\varepsilon}_{x,1}[2]) + 1}{g_r(\tilde{\varepsilon}_{x,2}[2]) + 1} \\
&= \frac{g_r(\tilde{\varepsilon}_{x,1}[0])}{g_r(\tilde{\varepsilon}_{x,2}[0])} \frac{g_r(\tilde{\varepsilon}_{y,1}[1]) + 1}{g_r(\tilde{\varepsilon}_{y,2}[1]) + 1} \frac{g_r(\tilde{\varepsilon}_{x,1}[2]) + 1}{g_r(\tilde{\varepsilon}_{x,2}[2]) + 1}
\end{aligned} \tag{93}$$

$$\boldsymbol{C_1}[2]: \frac{g_r(\tilde{\varepsilon}_{x,1}[2])}{g_r(\tilde{\varepsilon}_{x,2}[2])} > 1 \tag{94}$$

$$\boldsymbol{C_2}[2]: \frac{g_r(\tilde{\varepsilon}_{x,1}[2])}{g_r(\tilde{\varepsilon}_{x,2}[2])} > \frac{g_r(\tilde{\varepsilon}_{y,1}[1])}{g_r(\tilde{\varepsilon}_{y,2}[1])} \tag{95}$$

where the first equation is by Equation (83). Note that, combing conditions $\boldsymbol{C_1}[0]$, $\boldsymbol{C_1}[1]$, $\boldsymbol{C_2}[1]$, $\boldsymbol{C_1}[2]$ and Equation (93), we know it is impossible that $\tilde{\varepsilon}_{x,1}[2] = \tilde{\varepsilon}_{x,1}[0]$ and $\theta_{x,2} = \theta_{x,0}$. So it is impossible $S(\theta_{x,2}, \theta_{y,1}) = S(\theta_{x,0}, \theta_{y,1})$.

We can repeat **Step 1** on θ_x and θ_y alternatively and construct a sequence $\{\theta_{x,0}, \theta_{y,1}, \theta_{x,2}, \theta_{y,3}, \dots\}$ such that the following conditions hold.

For $i \in \mathbb{N}$ we have

$$S(\theta_{x,2i+2}, \theta_{y,2i+1}) > S(\theta_{x,2i}, \theta_{y,2i+1}) \tag{96}$$

$$\begin{aligned}
\frac{g_r(\tilde{\varepsilon}_{x,1}[2i+2])}{g_r(\tilde{\varepsilon}_{x,2}[2i+2])} &= \frac{g_r(\tilde{\varepsilon}_{y,1}[2i+1])}{g_r(\tilde{\varepsilon}_{y,2}[2i+1])} \frac{g_r(\tilde{\varepsilon}_{x,1}[2i+2]) + 1}{g_r(\tilde{\varepsilon}_{x,2}[2i+2]) + 1} \\
&= \frac{g_r(\tilde{\varepsilon}_{x,1}[2i])}{g_r(\tilde{\varepsilon}_{x,2}[2i])} \frac{g_r(\tilde{\varepsilon}_{y,1}[2i+1]) + 1}{g_r(\tilde{\varepsilon}_{y,2}[2i+1]) + 1} \frac{g_r(\tilde{\varepsilon}_{x,1}[2i+2]) + 1}{g_r(\tilde{\varepsilon}_{x,2}[2i+2]) + 1}
\end{aligned} \tag{97}$$

$$\boldsymbol{C_1}[2i+2]: \frac{g_r(\tilde{\varepsilon}_{x,1}[2i+2])}{g_r(\tilde{\varepsilon}_{x,2}[2i+2])} > 1 \tag{98}$$

$$\boldsymbol{C_2}[2i+2]: \frac{g_r(\tilde{\varepsilon}_{x,1}[2i+2])}{g_r(\tilde{\varepsilon}_{x,2}[2i+2])} > \frac{g_r(\tilde{\varepsilon}_{y,1}[2i+1])}{g_r(\tilde{\varepsilon}_{y,2}[2i+1])} \tag{99}$$

2. In this context, $\boldsymbol{C_2}[1]$ is stronger than $\boldsymbol{C_1}[1]$. We separate $\boldsymbol{C_1}[1]$ and $\boldsymbol{C_2}[1]$ away for clarity.

For $i \in \mathbb{N} \wedge i > 0$ we have

$$S(\theta_{x,2i}, \theta_{y,2i+1}) > S(\theta_{x,2i}, \theta_{y,2i-1}) \tag{100}$$

$$\begin{aligned}
\frac{g_r(\tilde{\varepsilon}_{y,1}[2i+1])}{g_r(\tilde{\varepsilon}_{y,2}[2i+1])} &= \frac{g_r(\tilde{\varepsilon}_{x,1}[2i])}{g_r(\tilde{\varepsilon}_{x,2}[2i])} \frac{g_r(\tilde{\varepsilon}_{y,1}[2i+1]) + 1}{g_r(\tilde{\varepsilon}_{y,2}[2i+1]) + 1} \\
&= \frac{g_r(\tilde{\varepsilon}_{y,1}[2i-1])}{g_r(\tilde{\varepsilon}_{y,2}[2i-1])} \frac{g_r(\tilde{\varepsilon}_{x,1}[2i]) + 1}{g_r(\tilde{\varepsilon}_{x,2}[2i]) + 1} \frac{g_r(\tilde{\varepsilon}_{y,1}[2i+1]) + 1}{g_r(\tilde{\varepsilon}_{y,2}[2i+1]) + 1}
\end{aligned} \tag{101}$$

$$\boldsymbol{C_1}[2i+1]: \frac{g_r(\tilde{\varepsilon}_{y,1}[2i+1])}{g_r(\tilde{\varepsilon}_{y,2}[2i+1])} > 1 \tag{102}$$

$$\boldsymbol{C_2}[2i+1]: \frac{g_r(\tilde{\varepsilon}_{y,1}[2i+1])}{g_r(\tilde{\varepsilon}_{y,2}[2i+1])} > \frac{g_r(\tilde{\varepsilon}_{x,1}[2i])}{g_r(\tilde{\varepsilon}_{x,2}[2i])} \tag{103}$$

Now combining conditions $\boldsymbol{C_2}[2i+2]$, $\boldsymbol{C_2}[2i+1]$ and Equations (91), we can also obtain

$$\theta_{x,2i+2} > \theta_{x,2i},\ \theta_{y,2i+3} > \theta_{y,2i+1}\ (i \in \mathbb{N}) \tag{104}$$

Up to now, we have constructed the following strictly increasing sequences

$$\Theta_x[i] = \theta_{x,2i} \tag{105}$$

$$\Theta_y[i] = \theta_{y,2i+1} \tag{106}$$

$$R_x[i] = \frac{g_r(\tilde{\varepsilon}_{x,1}[2i])}{g_r(\tilde{\varepsilon}_{x,2}[2i])} \tag{107}$$

$$R_y[i] = \frac{g_r(\tilde{\varepsilon}_{y,1}[2i+1])}{g_r(\tilde{\varepsilon}_{y,2}[2i+1])} \tag{108}$$

for $i \in \mathbb{N}$. According to conditions $\boldsymbol{C_1}[0], \boldsymbol{C_1}[1], \cdots$, it is easy to know $R_x[i] > 1, R_y[i] > 1$ for $i \in \mathbb{N}$. Besides, we note

$$R_x^+[i] = \frac{g_r(\tilde{\varepsilon}_{x,1}[2i]) + 1}{g_r(\tilde{\varepsilon}_{x,2}[2i]) + 1},\ R_y^+[i] = \frac{g_r(\tilde{\varepsilon}_{y,1}[2i+1]) + 1}{g_r(\tilde{\varepsilon}_{y,2}[2i+1]) + 1} \tag{109}$$

for $i \in \mathbb{N}$. According to Lemma 4a, it is easy to know both $R_x^+[i], R_y^+[i]$ are strictly increasing. Importantly, we have constructed the following strictly increasing sequence for $i \in \mathbb{N}$.

$$S[i] = \begin{cases} S(\theta_{x,0}, \theta_{y,0}),\ i = 0 \\ S(\theta_{x,i-1}, \theta_{y,i}),\ i\%2 = 1 \\ S(\theta_{x,i}, \theta_{y,i-1}),\ i\%2 = 0 \wedge i > 0 \end{cases} \tag{110}$$

This accomplishes aspect **A3**.

Here we note that $\boldsymbol{C_1}[i] (i \geq 0)$ plays an important role in each iteration. When $\boldsymbol{C_1}[i]$ holds, we let the derivative of S equal to 0. Then we get the maximum of S and make $\boldsymbol{C_2}[i] (i \geq 1)$ hold in each iteration. Importantly, $\boldsymbol{C_2}[i]$ guarantees $R_x[i]$ and $R_y[i]$ are strictly increasing.

In the following, we prove

$$\lim_{i \to +\infty} R_x[i] = +\infty,\ \lim_{i \to +\infty} R_y[i] = +\infty \tag{111}$$

in order to accomplish aspect **A4** finally. Now let's observe how $R_x[i]$ increases. Using the notations of $R_x[i]$ and $R_x^+[i]$, we rewrite Equation (97) and get the following relation

$$R_x[i+1] = R_x[i] R_y^+[i] R_x^+[i+1] \tag{112}$$

This indicates that

$$R_x[i+1] - R_x[i] = R_x[i](R_y^+[i] R_x^+[i+1] - 1) \tag{113}$$

Here $R_x[i], R_y^+[i], R_x^+[i]$ are all strictly increasing and larger than 1. The relation in Equation (113) indicates that the difference between neighbouring elements of $\{R_x[i]\}$ is strictly increasing. This violates the Cauchy's criterion for convergence. Thus, we can conclude $\lim_{i\to+\infty} R_x[i] = +\infty$. Similarly, we can also conclude $\lim_{i\to+\infty} R_y[i] = +\infty$.
Now from $\Theta_x[i] < 1$, we can know

$$\begin{aligned} R_x[i] =& \frac{g_r(\tilde{\varepsilon}_{x,1}[2i])}{g_r(\tilde{\varepsilon}_{x,2}[2i])} = \frac{w_2(\varepsilon_{x,1}+\theta_{x,2i}\varepsilon_{x,2})-1}{w_2(\varepsilon_{x,2}-\theta_{x,2i}\varepsilon_{x,2})-1} \\ <& \frac{w_2(\varepsilon_{x,1}+\varepsilon_{x,2})-1}{w_2(\varepsilon_{x,2}-\theta_{x,2i}\varepsilon_{x,2})-1} \end{aligned} \tag{114}$$

The numerator of the rightmost item of Equation (114) is a constant. From $\lim_{i\to+\infty} R_x[i] = +\infty$, we can conclude $\lim_{i\to+\infty} w_2(\varepsilon_{x,2}-\theta_{x,2i}\varepsilon_{x,2})-1 = 0$ and $\lim_{i\to+\infty} \varepsilon_{x,2}-\theta_{x,2i}\varepsilon_{x,2} = 0$. Thus, we obtain

$$\lim_{i\to+\infty} \Theta_x[i] = \lim_{i\to+\infty} \theta_{x,2i} = 1 \tag{115}$$

Similarly, we can also obtain

$$\lim_{i\to+\infty} \Theta_y[i] = \lim_{i\to+\infty} \theta_{y,2i+1} = 1 \tag{116}$$

Now combining Equations (71), (105), (106), (110), (115) and (116), we can know

$$\begin{aligned} &\lim_{i\to+\infty} S[i] \\ =& S(1,1) = f(w_2(\varepsilon_{x,1}+\varepsilon_{x,2})w_2(\varepsilon_{y,1}+\varepsilon_{y,2})) \\ &+ f(w_2(\varepsilon_{x,2}-\varepsilon_{x,2})w_2(\varepsilon_{y,2}-\varepsilon_{y,2})) \\ =& f(w_2(\varepsilon_{x,1}+\varepsilon_{x,2})w_2(\varepsilon_{y,1}+\varepsilon_{y,2})) + 1 \end{aligned} \tag{117}$$

Since $S[i]$ is strictly increasing, we can conclude

$$S(\theta_{x,0},\theta_{y,0}) < f(w_2(\varepsilon_{x,1}+\varepsilon_{x,2})w_2(\varepsilon_{y,1}+\varepsilon_{y,2})) + 1 \tag{118}$$

Remember that, we can take any $\theta_{x,0} > 0$ and any $\theta_{y,0} \neq \theta_{y,1}$ as the start point of above iterations. This means that Equation (118) holds for any point $(\theta_{x,0},\theta_{y,0})$ satisfying $\theta_{x,0} > 0$ in the neighborhood of $(0,0)$ on the $\theta_x\theta_y$ plane. Now we can show

$$S(0,0) \leq f(w_2(\varepsilon_{x,1}+\varepsilon_{x,2})w_2(\varepsilon_{y,1}+\varepsilon_{y,2})) + 1 = S(1,1) \tag{119}$$

by contradiction. Assume to the contrary that $S(0,0) > f(w_2(\varepsilon_{x,1}+\varepsilon_{x,2})w_2(\varepsilon_{y,1}+\varepsilon_{y,2})) + 1$, due to continuity of $S(\theta_x,\theta_y)$, we can find a neighbour (θ'_x,θ'_y) $(\theta'_x > 0)$ of $(0,0)$ such that $S(\theta'_x,\theta'_y) > f(w_2(\varepsilon_{x,1}+\varepsilon_{x,2})w_2(\varepsilon_{y,1}+\varepsilon_{y,2})) + 1$. This contradicts with Inequlity (118).

Case 3:

Finally, we can discuss **Case 3** when one of $\varepsilon_{x,2}$ and $\varepsilon_{y,2}$ equals 0. Without loss of generality, we suppose that $\varepsilon_{y,2} = 0$. We can discuss this in two subcases.

- **Subcase 3.1**: $\varepsilon_{y,1} = \varepsilon_{y,2} = 0$. By Lemma 1d and Equation (71), it is easy to know

$$\begin{aligned} S(1,1) =& f(w_2(\varepsilon_{x,1}+\varepsilon_{x,2})w_2(0)) + f(w_2(0)w_2(0)) \\ =& 1+\varepsilon_{x,1}+\varepsilon_{x,2}+1 \\ S(0,0) =& f(w_2(\varepsilon_{x,1})w_2(0)) + f(w_2(\varepsilon_{x,2})w_2(0)) \\ =& 1+\varepsilon_{x,1}+1+\varepsilon_{x,2} \end{aligned}$$

This satisfies $S(0,0) \leq S(1,1)$.

- **Subcase 3.2**: $\varepsilon_{y,1} > \varepsilon_{y,2} = 0$.
Here we treat $S(\theta_x,\theta_y)$ as a function of three variables $\theta_x,\theta_y,\varepsilon_{y,2}$ as follows.

$$\begin{aligned} &S(\theta_x,\theta_y,\varepsilon_{y,2}) \\ =& f(w_2(\varepsilon_{x,1}+\theta_x\varepsilon_{x,2})w_2(\varepsilon_{y,1}+\theta_y\varepsilon_{y,2})) \\ &+ f(w_2(\varepsilon_{x,2}-\theta_x\varepsilon_{x,2})w_2(\varepsilon_{y,2}-\theta_y\varepsilon_{y,2})) \end{aligned} \tag{120}$$

It is easy to know $S(\theta_x,\theta_y,\varepsilon_{y,2})$ is continuous. Note that, we have proven $S(0,0,\varepsilon_{y,2}) \leq S(1,1,\varepsilon_{y,2})$ for any $\varepsilon_{y,2} > 0$ in **Case 2**. Therefore, we have

$$S(0,0,0) = \lim_{\varepsilon_{y,2}\to 0} S(0,0,\varepsilon_{y,2}) \leq \lim_{\varepsilon_{y,2}\to 0} S(1,1,\varepsilon_{y,2}) = S(1,1,0) \tag{121}$$

so $S(0,0) \leq S(1,1)$ for $\varepsilon_{y,1} > \varepsilon_{y,2} = 0$.

This concludes the proof of Lemma 7. □

Up to now, we have resolved the 2-dimensional case of the core of the proof. To extend to high-dimensional case, we need the following Lemma 8.

Lemma 8. (See [21]) For any two Hermition positive semidefinite $n\times n$-matrices A, B

$$\mathrm{Tr}(AB) \leq \sum_{i=1}^{n} \lambda_{A,[i]}\lambda_{B,[i]} \tag{122}$$

where $\lambda_{A,[i]}, \lambda_{B,[i]}$ are the eigenvalues of A, B arranged in decreasing order, respectively.

Now we present our theorem on the relaxed triangle inequality of KL divergences between Gaussians. Firstly, we deal with the case when one of the Gaussians is standard Gaussian. Then, we generalize the conclusion to general case.

Lemma 9. For any two n-dimensional Gaussian distributionss $\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ such that $KL(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(0,I)) \leq \varepsilon_1$, $KL(\mathcal{N}(0,I)||\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)) \leq \varepsilon_2$ $(\varepsilon_1,\varepsilon_2 \geq 0)$, then

$$\begin{aligned} &KL((\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)) \\ <& \varepsilon_1+\varepsilon_2+\frac{1}{2}\Bigg(W_{-1}(-e^{-(1+2\varepsilon_1)})W_{-1}(-e^{-(1+2\varepsilon_2)}) \\ &+ W_{-1}(-e^{-(1+2\varepsilon_1)}) + W_{-1}(-e^{-(1+2\varepsilon_2)}) + 1 \\ &-W_{-1}(-e^{-(1+2\varepsilon_2)})\left(\sqrt{2\varepsilon_1}+\sqrt{\frac{2\varepsilon_2}{-W_0(-e^{-(1+2\varepsilon_2)})}}\right)^2\Bigg) \end{aligned} \tag{123}$$

Proof 10. In the proofs of Lemma 2 (and Lemma 10 in Appendix), we construct equivalent optimization problems by introducing new variables in the constraints. Unfortunately, in the proof of Lemma 9, we cannot use the same step. Otherwise, the bound would be too complicated to resolve. To obtain a bound independent of the dimension n, we need to relax the constraint in the beginning.
Our aim is to find an upper bound of $KL((\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2))$ under the constraints $KL(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(0,I)) \leq \varepsilon_1$, $KL(\mathcal{N}(0,I)||\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)) \leq \varepsilon_2$. In the following, we first relax the constraints and then find an upper bound under the relaxed constraints.

KL divergence formula for 2 gaussians

According to the definition of KL divergence, we have

$$KL((\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2))$$
$$=\frac{1}{2}\left(\log\frac{|\boldsymbol{\Sigma_2}|}{|\boldsymbol{\Sigma}_1|}+\mathrm{Tr}(\boldsymbol{\Sigma_2}^{-1}\boldsymbol{\Sigma}_1)+(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)-n\right)$$

In the following two steps, we first find an upper bound for the first two items, then we find an upper bound for the rest items.

Step 1. According to Lemma 8, we have

$$\begin{aligned}
&\log\frac{|\boldsymbol{\Sigma_2}|}{|\boldsymbol{\Sigma}_1|}+\mathrm{Tr}(\boldsymbol{\Sigma_2}^{-1}\boldsymbol{\Sigma}_1)\\
=&\mathrm{Tr}(\boldsymbol{\Sigma_2}^{-1}\boldsymbol{\Sigma}_1)-\log\frac{|\boldsymbol{\Sigma_1}|}{|\boldsymbol{\Sigma}_2|}\\
=&\mathrm{Tr}(\boldsymbol{\Sigma_2}^{-1}\boldsymbol{\Sigma}_1)-\log(|\boldsymbol{\Sigma_2}^{-1}||\boldsymbol{\Sigma_1}|)\\
=&\mathrm{Tr}(\boldsymbol{\Sigma_2}^{-1}\boldsymbol{\Sigma}_1)-\log\prod_{i=1}^{n}\lambda_{1,i}\lambda'_{2,i}\\
\le&\sum_{i=1}^{n}\lambda_{1,[i]}\lambda'_{2,[i]}-\log\prod_{i=1}^{n}\lambda_{1,[i]}\lambda'_{2,[i]}\\
=&\sum_{i=1}^{n}\lambda_{1,[i]}\lambda'_{2,[i]}-\log\lambda_{1,[i]}\lambda'_{2,[i]} \qquad (124)
\end{aligned}$$

where $\lambda_{1,i},\lambda'_{2,i}$ are the eigenvalues of $\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2^{-1}$ arranged in decreasing order, respectively. In the following, we find an upper bound for Equation (124).

By the definition of KL divergence, the constraint $KL(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(0,I))\le\varepsilon_1$ is equal to

$$-\log|\boldsymbol{\Sigma}_1|+\mathrm{Tr}(\boldsymbol{\Sigma}_1)+\boldsymbol{\mu}_1^\top\boldsymbol{\mu}_1-n\le 2\varepsilon_1 \qquad (125)$$

Combining Lemma 1a, Equation (29) and (32), we relax the constraint in Inquality (125) as follows.

$$-\log|\boldsymbol{\Sigma}_1|+\mathrm{Tr}(\boldsymbol{\Sigma}_1)=\sum_{i=1}^{n}\lambda_{1,i}-\log\lambda_{1,i}\le n+2\varepsilon_1 \qquad (126)$$

$$\boldsymbol{\mu}_1^\top\boldsymbol{\mu}_1\le 2\varepsilon_1 \qquad (127)$$

where $\lambda_{1,i}$ are the eigenvalues of $\boldsymbol{\Sigma}_1$. For simplicity, we modify the constraint in Inquality (126) to the following constraint.

$$-\log|\boldsymbol{\Sigma}_1|+\mathrm{Tr}(\boldsymbol{\Sigma}_1)=\sum_{i=1}^{n}\lambda_{1,i}-\log\lambda_{1,i}=n+2\varepsilon_1 \qquad (128)$$

In the following, we find the upper bound for Equation (124) under constraints (128). Then we will see that the upper bound is increasing with ε_1. So there is no difference between constraints (126) and (128).

Form the perspective of optimization, the constraint in Inequality (128) can be replaced by the following constraints

$$\lambda_{1,i}-\log\lambda_{1,i}=1+\varepsilon_{1,i}\ (1\le i\le n) \qquad (129)$$

$$\bigwedge_{i=1}^{n}\varepsilon_{1,i}\ge 0\wedge\sum_{i=1}^{n}\varepsilon_{1,i}=2\varepsilon_1 \qquad (130)$$

Similarly, the constraint $KL(\mathcal{N}(0,I)||\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2))\le\varepsilon_2$ is equal to

$$\log|\boldsymbol{\Sigma}_2|+\mathrm{Tr}(\boldsymbol{\Sigma}_2^{-1})+\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2-n\le 2\varepsilon_2 \qquad (131)$$

which implies the following constraints

$$\log|\boldsymbol{\Sigma}_2|+\mathrm{Tr}(\boldsymbol{\Sigma}_2^{-1})=\sum_{i=1}^{n}\lambda'_{2,i}-\log\lambda'_{2,i}\le n+2\varepsilon_2 \qquad (132)$$

$$\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\le 2\varepsilon_2 \qquad (133)$$

where $\lambda'_{2,i}$ are the eigenvalues of $\boldsymbol{\Sigma}_2^{-1}$. We also modify the constraint in Inequality (132) to the following constraint which does not affect the upper bound.

$$\log|\boldsymbol{\Sigma}_2|+\mathrm{Tr}(\boldsymbol{\Sigma}_2^{-1})=\sum_{i=1}^{n}\lambda'_{2,i}-\log\lambda'_{2,i}=n+2\varepsilon_2 \qquad (134)$$

Furthermore, constraint (134) can be replaced by the following constraints.

$$\lambda'_{2,i}-\log\lambda'_{2,i}=1+\varepsilon_{2,i}\ (1\le i\le n) \qquad (135)$$

$$\bigwedge_{i=1}^{n}\varepsilon_{2,i}\ge 0\wedge\sum_{i=1}^{n}\varepsilon_{2,i}=2\varepsilon_2 \qquad (136)$$

In the following, we find an upper bound of Equation (124) under constraints (129), (130), (135), and (136).

Applying Lemma 6 to Equation (124) with conditions (129) and (135), we can obtain

$$\sum_{i=1}^{n}\lambda_{1,[i]}\lambda'_{2,[i]}-\log\lambda_{1,[i]}\lambda'_{2,[i]}\le\sum_{i=1}^{n}f(w_2(\varepsilon_{1,[i]})w_2(\varepsilon_{2,[i]})) \qquad (137)$$

where $\varepsilon_{1,[i]}$ and $\varepsilon_{2,[i]}$ are also arranged in decreasing order.

Now we apply Lemma 7 to the right hand side of Inequality (137) repeatedly on the first two dimensions as follows. Here we use notations $E_{1,k}=\sum_{i=1}^{k}\varepsilon_{1,[i]}$, $E_{2,k}=\sum_{i=1}^{k}\varepsilon_{2,[i]}$ for brevity.

$$\begin{aligned}
&\log\frac{|\boldsymbol{\Sigma_2}|}{|\boldsymbol{\Sigma}_1|}+\mathrm{Tr}(\boldsymbol{\Sigma_2}^{-1}\boldsymbol{\Sigma}_1) && \text{by (124)}\\
\le&\textstyle\sum_{i=1}^{n}\lambda_{1,[i]}\lambda'_{2,[i]}-\log\lambda_{1,[i]}\lambda'_{2,[i]} && \text{by (137)}\\
\le&\textstyle\sum_{i=1}^{n}f(w_2(\varepsilon_{1,[i]})w_2(\varepsilon_{2,[i]}))\\
=&f(w_2(\varepsilon_{1,[1]})w_2(\varepsilon_{2,[1]}))+f(w_2(\varepsilon_{1,[2]})w_2(\varepsilon_{2,[2]}))\\
&+\textstyle\sum_{i=3}^{n}f(w_2(\varepsilon_{1,[i]})w_2(\varepsilon_{2,[i]})) && \text{Lemma 7}\\
\le&f(w_2(\varepsilon_{1,[1]}+\varepsilon_{1,[2]})w_2(\varepsilon_{2,[1]}+\varepsilon_{2,[2]}))+1\\
&+\textstyle\sum_{i=3}^{n}f(w_2(\varepsilon_{1,[i]})w_2(\varepsilon_{2,[i]}))\\
=&f(w_2(E_{1,2})w_2(E_{2,2}))+f(w_2(\varepsilon_{1,[3]})w_2(\varepsilon_{2,[3]}))\\
&+\textstyle\sum_{i=4}^{n}f(w_2(\varepsilon_{1,[i]})w_2(\varepsilon_{2,[i]}))+1 && \text{Lemma 7}\\
\le&f(w_2(E_{1,2}+\varepsilon_{1,[3]})w_2(E_{2,2}+\varepsilon_{2,[3]}))+1\\
&+\textstyle\sum_{i=4}^{n}f(w_2(\varepsilon_{1,[i]})w_2(\varepsilon_{2,[i]}))+1\\
=&f(w_2(E_{1,3})w_2(E_{2,3}))+\textstyle\sum_{i=4}^{n}f(w_2(\varepsilon_{1,[i]})w_2(\varepsilon_{2,[i]}))\\
&+2\\
&\cdots\\
\le&f(w_2(E_{1,n})w_2(E_{2,n}))+n-1\\
=&f(w_2(\textstyle\sum_{i=1}^{n}\varepsilon_{1,[i]})w_2(\sum_{i=1}^{n}\varepsilon_{2,[i]}))+n-1\\
=&f(w_2(2\varepsilon_1)w_2(2\varepsilon_2))+n-1\\
=&2\varepsilon_1+2\varepsilon_2+2+w_2(2\varepsilon_1)w_2(2\varepsilon_2)-w_2(2\varepsilon_1) && \text{Lemma 1j}\\
&-w_2(2\varepsilon_2)+n-1\\
=&2\varepsilon_1+2\varepsilon_2+w_2(2\varepsilon_1)w_2(2\varepsilon_2)-w_2(2\varepsilon_1)-w_2(2\varepsilon_2)\\
&+n+1 \qquad (138)
\end{aligned}$$

The bound in Equation (138) is increasing with ε_1 and ε_2. Therefore, the constraints (128) and (134) can be modified back to (126) and (132), respectively .

Step 2. from Equation (127), we know

$$|\boldsymbol{\mu}_1| \leq \sqrt{2\varepsilon_1} \tag{139}$$

where $|\cdot|$ denotes the L_2 norm of vector. From Inequlity (133), we also know $\lambda'_{2*}\boldsymbol{\mu}_2^\top\boldsymbol{\mu}_2 \leq \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 \leq 2\varepsilon_2$, where λ'_{2*} is the minimum eigenvalue of $\boldsymbol{\Sigma}_2^{-1}$. Now combining the condition (132) and Lemma 1g, we get

$$\boldsymbol{\mu}_2^\top\boldsymbol{\mu}_2 \leq \frac{2\varepsilon_2}{\lambda'_{2*}} \leq \frac{2\varepsilon_2}{w_1(2\varepsilon_2)} \implies |\boldsymbol{\mu}_2| \leq \sqrt{\frac{2\varepsilon_2}{w_1(2\varepsilon_2)}} \tag{140}$$

Combining Inequalities (139), (140) and using the triangle inequality for norms of vectors, we have

$$|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1| \leq |\boldsymbol{\mu}_2| + |\boldsymbol{\mu}_1| \leq \sqrt{2\varepsilon_1} + \sqrt{\frac{2\varepsilon_2}{w_1(2\varepsilon_2)}} \tag{141}$$

Again, we have $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \leq \lambda'^*_2|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|^2$, where λ'^*_2 is the maximum eigenvalue of $\boldsymbol{\Sigma}_2^{-1}$. From Lemma 1g and condition (132), we know $\lambda'^*_2 \leq w_2(2\varepsilon_2)$. Thus, we can conclude that

$$\begin{aligned}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \leq & w_2(2\varepsilon_2)|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|^2 \\ \leq & w_2(2\varepsilon_2)\left(\sqrt{2\varepsilon_1} + \sqrt{\frac{2\varepsilon_2}{w_1(2\varepsilon_2)}}\right)^2\end{aligned} \tag{142}$$

Finally, combining Inequalities (138) and (142), we can conclude that

$$\begin{aligned}&KL((\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \\ <&\frac{1}{2}\Bigg(2\varepsilon_1 + 2\varepsilon_2 + w_2(2\varepsilon_1)w_2(2\varepsilon_2) - w_2(2\varepsilon_1) - w_2(2\varepsilon_2) \\ &+ n + 1 + w_2(2\varepsilon_2)\left(\sqrt{2\varepsilon_1} + \sqrt{\frac{2\varepsilon_2}{w_1(2\varepsilon_2)}}\right)^2 - n\Bigg) \\ =&\varepsilon_1 + \varepsilon_2 + \frac{1}{2}\Bigg(W_{-1}(-e^{-(1+2\varepsilon_1)})W_{-1}(-e^{-(1+2\varepsilon_2)}) \\ &+ W_{-1}(-e^{-(1+2\varepsilon_1)}) + W_{-1}(-e^{-(1+2\varepsilon_2)}) + 1 \\ &- W_{-1}(-e^{-(1+2\varepsilon_2)})\left(\sqrt{2\varepsilon_1} + \sqrt{\frac{2\varepsilon_2}{-W_0(-e^{-(1+2\varepsilon_2)})}}\right)^2\Bigg)\end{aligned} \tag{143}$$

□

Remark 3. The bound in Equation (143) has the following properties.

1) The bound becomes 0 when $\varepsilon_1 = \varepsilon_2 = 0$.
2) When both ε_1 and ε_2 are small, all the items in the bound are small and hence the bound is small.
3) The bound is independent of the dimension n because we have eliminated the impact of dimension n by Lemma 7. This is the most tricky part in this proof.
4) When ε_2 is large, the bound is mostly dominated by the last item in the branket.
5) In fact, we can distribute $2\varepsilon_1$ into two parts in Equation (126) and (127). However, this will lead to a complicated expression which is very hard to solve a supremum as like what we do on Equation (40). Numerical experiments show that the supremum varies with how $2\varepsilon_1$ ($2\varepsilon_2$) are allocated into two parts in the left hand sides of Equation (126) and (127) ((132) and (133)). Therefore, in constraints (126), (127), (132), and (133), we relax the conditions and get an relaxed upper bound with a simpler form in Inequality (143).

Proof of Theorem 4.

Proof 11. Theorem 4 extends Lemma 9 to three general Gaussians. We can use linear invertible transformation to convert one Gaussian into standard Gaussian and then apply Lemma 9. Please see Appendix H for details.

□

In Theorem 4, we try to find an upper bound as tight as possible. So the bound seems a little complicated. We can expand Lambert W function by series [18], [20] and simplify the bound as follows [15] [3].

Theorem 5. For any three n-dimensional Gaussians $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)(i \in \{1, 2, 3\})$ such that $KL(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \leq \varepsilon_1$ and $KL(\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)||\mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)) \leq \varepsilon_2$ for small $\varepsilon_1, \varepsilon_2 \geq 0$, then

$$\begin{aligned}&KL(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)) \\ <&3\varepsilon_1 + 3\varepsilon_2 + 2\sqrt{\varepsilon_1\varepsilon_2} + o(\varepsilon_1) + o(\varepsilon_2)\end{aligned} \tag{144}$$

Proof 12. See Appendix I for the details of the proof.

□

Finally, in the proof of Theorem 4, we use invertible linear transformation to convert $\mathcal{N}_2$ to standard Gaussian with preserving KL divergence. This still holds when $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ is fixed. So we get the the following corollary.

Corollary 2. Theorem 4 and 5 hold when $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ is fixed.

Remark 4. **Comparison with existing general Pythagoras inequalities**

It is known that KL divergence satisfies some general Pythagoras inequalities which seem similar to our relaxed triangle inequality. We note that they are different in the follows.

The bound in our relaxed triangle inequality is independent of the parameters of Gaussians and only related to ε_1 and ε_2. Our theorem is different from the several existing generalized Pythagoras inequalities satisfied by KL divergence, where the bounds are functions of the given distributions. We list them as follows.

1) The generalized Pythagoras inequality for KL divergence [4], [11] states that for a convex set of distributions $\mathcal{P}$, any distribution Q not in $\mathcal{P}$, and $D_{min} = \inf_{P\in\mathcal{P}} KL(P||Q)$, there exists a distribution P^* such that

$$KL(P||Q) \geq KL(P||P^*) + D_{min} \quad \text{for all} \quad P \in \mathcal{P}$$

2) Erven *et al.* generalize the Pythagoras inequality for KL divergence to Rényi divergence which includes KL divergence with order 1. See [11] for details.

3. After we post our last version of manuscript on Arxiv [16], Liu *et al.* cited our manuscript in their work [15] in which they simplify the bound in Theorem 4 by using series in simpler case $\varepsilon_1 = \varepsilon_2$.

3) Functional Bregman divergence also satisfies a generalized Pythagoras theorem [10]. Let $(\mathbb{R}^d, \Omega, v)$ be a measure space, where d is a positive integer and v is a Borel measure. Let $\mathcal{A}$ be a convex subset of $L^p(v)$. For any $f, g, h \in \mathcal{A}$, functional Bregman divergence d_ϕ satisfies

$$d_\phi[f,h] = d_\phi[f,g] + d_\phi[g,h] + \delta\phi[g; f-g] - \delta\phi[h; f-g] \tag{145}$$

where $\phi : L^p(v) \to \mathbb{R}$ is a strictly convex, twice-continuously Fréchet-differentiable functional. $\delta\phi[g;\cdot]$ is the Fréchet derivative of ϕ at g. KL divergence is a special form of functional Bregman divergence when $\phi = \int p(x)\log p(x)\,\mathrm{d}x$ whose Fréchet derivative at g is $\delta\phi[g;t] = \int(\log g(x)+1)t(x)\,\mathrm{d}x$. Plugging ϕ and $\delta\phi$ into Equation (145), we get

$$\begin{aligned} &KL(f||h) \\ =&KL(f||g) + KL(g||h) + \int(\log g(x)+1)(f(x)-g(x))\,\mathrm{d}x \\ &- \int(\log h(x)+1)(f(x)-g(x))\,\mathrm{d}x \\ =&KL(f||g) + \int f(x)\log\frac{g(x)}{h(x)}\,\mathrm{d}x \end{aligned} \tag{146}$$

All the bounds in the above inequalities are dependent on the parameters of the given distributions.

In our theorem, we allow all parameters are unknown or one Gaussian is fixed. Therefore, our theorems are suitable for contexts where the re can vary. This is common in deep learning where the parameters are learned by the model. Therefore, it is impossible to identify the parameters or the KL divergence before the model is trained. In some cases, we only know that some bound is guaranteed. In the next section, we discuss the applications of our theorems in deep learning.

5 Applications

5.1 Anomaly Detection with Flow-based Model

The research question in this paper comes from our research on deep anomaly detection using flow-based model [12], [13], [14], [22]. Flow-based model constructs diffeomorphism between data space to latent space. Compared with other generative models such as generative adversarial networks, flow-based model has the advantage of providing explicit likelihood $p_\theta(x)$ to input x, where θ refer to model parameters. Usually, flow-based model is trained by maximum likelihood estimate with Gaussian prior. Intuitively, it is natural to believe that samples from the training (in-distribution, ID in short) dataset should have higher likelihoods than out-of-distribution (OOD) data (*i.e.*, anomalies). However, Nalisnick *et al.* reveal that deep generative models including flow-based models may assign higher likelihoods to OOD data [23]. For example, Glow [14] assigns higher likelihoods for SVHN when trained on CIFAR-10. This observation is also verified by many other researchers including ourselves [24], [25], [26], [27], [28]. This brings obstacles to anomaly detection in flow-based model according to model likelihood [27], [28]. However, we can not sample these OOD data from the model although they may have higher likelihoods than training data. Nalisnick et al. explain this phenomenon by the discrepancy of typical set and high probability density regions of model distribution [27]. This can explain why we can not sample OOD data that have higher likelihoods than ID data. But their explanation fails when OOD data has coinciding likelihoods with ID data. Before our analysis, this counterintuitive phenomenon has not been satisfactorily explained.

In this context, we want to explain *why we can not sample OOD data from flow-based model with prior regardless of when OOD data have higher, lower, or coinciding likelihoods*. We investigate this problem from a statistical divergence perspective. Let $z = f(x)$ be the flow-based model which maps data x in data space to z in latent space. Assume that the prior distribution p_Z^r is the most commonly used Gaussian distribution. Suppose that $X_1 \sim p_X(x)$, $X_2 \sim q_X(x)$ represent distributions of ID and OOD datasets, respectively. We note $Z_1 = f(X_1) \sim p_Z(z)$, $Z_2 = f(X_2) \sim q_Z(z)$ to represent the distributions of representations of ID and OOD datasets, respectively. We also note the model induced distribution p_X^r such that $Z_r \sim p_Z^r$ and $X_r = f^{-1}(Z_r) \sim p_X^r$. Flow-based model is usually trained by maximum likelihood estimation. This is equal to minimizing forward KL divergence $KL(p_Z||p_Z^r)$ [11], [29]. In our experiments, we conduct generalized Shapiro-Wilk test for multivariate normality. Results demonstrate that p_Z is Gaussian-like for all datasets. Surprisingly, q_Z is also Gaussian-like for OOD datasets with higher or coinciding likelihoods except for just one case. These results allow us to approximate p_Z and q_Z with Gaussians. Note that, it seems that the normality of represetations of ID and OOD dataset is a characteristic of flow-based model. We did not get similar observations in variational autoencoders.

The theorems proved in this paper can help us to analyze the KL divergences between p_Z, p_Z^r, and q_Z. On one hand, according to Proposition 1, we can know $KL(p_X||p_X^r) = KL(p_Z||p_Z^r)$, so $KL(p_Z||p_Z^r)$ is trained to be small. By Theorem 1, we can know $KL(p_Z^r||p_Z)$ is small too. So we can assume $p_Z^r \approx p_Z$ when $KL(p_Z||p_Z^r)$ is sufficiently small. On the other hand, we can also assume that the distributions of ID and OOD data are far from each other. This implies that $KL(p_X||q_X) = KL(p_Z||q_Z)$ can be any large. By the relaxed triangle inequality, we can infer that $KL(p_Z^r||q_Z)$ must be large. This answers the question why we can not sample OOD data from flow-based model with prior. Furthermore, we decompose the large divergence $KL(q_Z||p_Z^r)$ into dimensional-wise KL divergence and total correlation (generalized mutual information) measuring the mutual dependence between dimensions. We demonstrate that the representations of OOD data are more correlated than that of ID data. From a geometric perspective, strong correlation indicates that the representations of OOD data locate in specific directions. In high dimensional space, it is hard to sample data residing in specific directions from prior. This gives the second explanation to the above question. Based on the theoretical analysis and further observation on the local pixel dependence in the representation of OOD dataset, we propose a KL divergence-based anomaly detection algorithm. Experimental results have shown the effectiveness of our method. More details of the application of our theorems in deep anomaly detection research can

be referred to in our manuscript [30], which is submitted independently.

Importantly, flow-based model constructs diffeomorphism between data space to latent space with thousands of dimensions. It is important that the bounds found in this paper are independent of the dimension. Furthermore, since both p_Z and q_Z are dependent on model parameters and q_Z is also dependent on the input OOD dataset, it is impossible to determine the parameters of p_Z and q_Z in advance. Our theorems do not dependent on the parameters of distributions and only requires some bound is restricted. This is why we need to prove the theorems in this paper rather than using existing theorems.

5.2 Safety Guarantee in Reinforcement Learning

The theorems proved in this paper can also be used as general conclusions in related fields. Since we post the last version of this manuscript on Arxiv [16], our manuscript has been cited by other researchers. For example, the relaxed triangle inequality (Theorem 4) has been used in the research of constrained variational policy optimization for safe reinforcement learning [15]. In their work, Liu *et al.* propose an Expectation-Maximization style approach for learning safe policy in reinforcement learning. After achieving one-step robustness guarantee, a natural question is extending to multiple steps policy updating robustness guarantee. This requires triangle inequality for consecutive updated policies. It is known that KL divergence does not has such property in general case. However, multivariate Gaussian is commonly used as policy in continuous action space tasks. In such context, our relaxed triangle inequality (Theorem 4) is used to extend one-step robustness guarantee to multiple steps. Particularly, Liu *et al.* use big-O to simplify the bound in Theorem 4 in case $\varepsilon_1 = \varepsilon_2$. Please see [15] for more details about the application.

6 Related Work

KL divergence is an important divergence and has a wide range of applications [2], [3], [4], [5], [31], [32], [33]. Researchers have investigated KL divergence between many distributions including Markov sources [34], GMM models [35], [36], multivariate generalized Gaussians [37], univariate mixtures [38], discrete normal distributions [39], *etc*. In [5], a bound of KL divergence between Gaussians is given. As we discussed in Remark 4, existing generalized Pythagorean inequality satisfied by KL divergence are all dependent on the parameters of distributions [4], [11]. As far as we know, there is no related work that focuses on the similar properties of KL divergence between Gaussians as this paper.

KL divergence is one member of more general divergences such as Bregman divergence [7], [9], [10], [40], f-divergence [5], [6], [41], Rényi divergence [5], [8], [11], and recently proposed (f,Γ)-divergence [42]. Bregman divergence defines a class of divergences [9] in vector space. KL divergence between multinomial distributions is a special form of Bregman divergence when the convex function for Bregman divergence is chosen as $\sum_{i=1}^{n} p_i \log p_i$, where $p_i \geq 0$ for $i = 1,\ldots,n$ define a multinomial distribution. Frigyik *et. al.* [10] extends vector Bregman divergence to functional Bregman divergence in L^p. Similarly, KL divergence is a special form of functional Bregman divergene. (functional) Bregman divergence also satisfies generalized Pythagoras theorem [9], [10]. We note that our relaxed triangle inequality has a different meaning in Remark 4.

f-divergence also defines a class of divergences based on convex functions [5], [6], [43], [44]. Many commonly used measures including the KL divergence, Jensen-Shannon divergence, and squared Hellinger distance are special cases of f-divergence. Many f-divergences are not proper distance metrics and do not satisfy the triangle inequality. KL divergence is the unique divergence belong to both f-divergence and Bregman divergence [45].

Rényi divergence defines another class of divergences [5], [8], [11], [46]. Rényi divergence with order of 1 becomes KL divergence. As discussed in Remark 4, Rényi divergence also satisfies a generalized Pythogras theorem [11].

KL divergence between general distributions does not have a closed form. In application, it is not easy to estimate KL divergence when only samples of distributions are available especially in high dimensional problems. A line of research is dedicated to the estimation of divergences [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58]. Unlike other distributions, the KL divergence between Gaussians has a closed form. The theorems presented in this paper can deepen our understanding of KL divergence between Gaussians.

The asymmetry of KL divergence has restricted the application of KL divergence in practical applications. Many other divergences have been investigated [9], [59], [60], [61], [62], [63], [64], [65], [66], [67]. Pardo gives a comprehensive survey on a wide range of statistical divergences in his book [5].

7 Conclusion

In this paper, we research the properties of KL divergences between Gaussians. First, we find the supremum of reverse KL divergence $KL(\mathcal{N}_2||\mathcal{N}_1)$ if the forward KL divergence $KL(\mathcal{N}_1||\mathcal{N}_2) \leq \varepsilon$ $(\varepsilon > 0)$. This conclusion quantifies the approximate symmetry of small KL divergence between Gaussians. We also find the infimum of $KL(\mathcal{N}_2||\mathcal{N}_1)$ if $KL(\mathcal{N}_1||\mathcal{N}_2) \geq M$ $(M > 0)$. We give the conditions when the supremum and infimum can be attained. Second, we find a bound for $KL(\mathcal{N}_1||\mathcal{N}_3)$ when $KL(\mathcal{N}_1||\mathcal{N}_2)$ and $KL(\mathcal{N}_2||\mathcal{N}_3)$ are bounded. This indicates that KL divergence between Gaussians follows a relaxed triangle inequality. Importantly, all the bounds in the theorems in this paper are independent of the dimension of distributions. The theorems presented in this paper is suitable especially for contexts where parameters may vary or can not be identified in advance (e.g., machine learning). Finally, we discuss the applications of our theorems in deep anomaly detection and safe reinforcement learning. We hope our research can shed light on more research in related field. In the future, we plan to explore the properties of KL divergence between more general distributions such as Gaussian mixture models and exponential family of distributions.

ACKNOWLEDGMENTS

This work is supported by NSFC Program (No. 62002107, 62172429, 61872371).

REFERENCES

[1] S. Kullback, *Information theory and statistics*. Courier Corporation, 1997.
[2] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.
[3] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.
[4] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
[5] L. Pardo, *Statistical inference based on divergence measures*. CRC press, 2018.
[6] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.
[7] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR computational mathematics and mathematical physics*, vol. 7, no. 3, pp. 200–217, 1967.
[8] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. University of California Press, 1961, pp. 547–561.
[9] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, no. 58, pp. 1705–1749, 2005. [Online]. Available: http://jmlr.org/papers/v6/banerjee05b.html
[10] B. A. Frigyik, S. Srivastava, and M. R. Gupta, "Functional Bregman divergence," in *2008 IEEE International Symposium on Information Theory*, 2008, pp. 1681–1685.
[11] T. van Erven and P. Harremos, "Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.
[12] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.
[13] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.
[14] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 215–10 224.
[15] Z. Liu, Z. Cen, V. Isenbaev, W. Liu, Z. S. Wu, B. Li, and D. Zhao, "Constrained variational policy optimization for safe reinforcement learning," in *The 5th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, 2022.
[16] Y. Zhang, W. Liu, Z. Chen, J. Wang, and K. Li, "On the properties of kullback-leibler divergence between gaussians," 2021. [Online]. Available: https://arxiv.org/abs/2102.05485
[17] J. H. Lambert, "Observationes variae in mathesin puram," *Acta Helveticae physico-mathematico-anatomico-botanico-medica, Band III*, pp. 128–168, 1758.
[18] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Advances in Computational mathematics*, vol. 5, no. 1, pp. 329–359, 1996.
[19] F. Nielsen, "An elementary introduction to information geometry," *Entropy*, vol. 22, no. 10, 2020. [Online]. Available: https://www.mdpi.com/1099-4300/22/10/1100
[20] L. Farmer, "The princeton companion to applied mathematics," *Reference Reviews*, vol. 30, no. 5, pp. 34–35, 2016.
[21] J. B. Lasserre, "A trace inequality for matrix product," *IEEE Transactions on Automatic Control*, vol. 40, no. 8, pp. 1500–1501, 1995.
[22] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3964–3979, 2021.
[23] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do deep generative models know what they don't know?" *International Conference on Learning Representations (ICLR)*, 2019.
[24] A. Shafaei, M. Schmidt, and J. J. Little, "Does your model know the digit 6 is not a cat? a less biased evaluation of" outlier" detectors," *arXiv preprint arXiv:1809.04729*, 2018.
[25] H. Choi and E. Jang, "WAIC, but why?: Generative ensembles for robust anomaly detection," *arXiv preprint arXiv:1810.01392*, 2018.
[26] V. Škvára, T. Pevnỳ, and V. Šmídl, "Are generative deep models for novelty detection truly better?" *KDD Workshop on Outlier Detection De-Constructed (ODD v5.0)*, 2018.
[27] E. Nalisnick, A. Matsukawa, Y. W. Teh, and B. Lakshminarayanan, "Detecting out-of-distribution inputs to deep generative models using typicality," *4th workshop on Bayesian Deep Learning (NeurIPS 2019)*, 2019.
[28] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Why normalizing flows fail to detect out-of-distribution data," *ICML workshop on Invertible Neural Networks and Normalizing Flows, 2020 (NeurIPS 2020)*, 2020.
[29] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," 2019.
[30] Y. Zhang, W. Liu, Z. Chen, J. Wang, Z. Liu, K. Li, and H. Wei, "Towards out-of-distribution detection with divergence guarantee in deep generative models," 2021. [Online]. Available: https://arxiv.org/abs/2002.03328v4
[31] S. Filippi, O. Cappé, and A. Garivier, "Optimism in reinforcement learning and kullback-leibler divergence," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2010, pp. 115–122.
[32] P. Guan, M. Raginsky, and R. M. Willett, "Online Markov decision processes with Kullback–Leibler control cost," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1423–1438, 2014.
[33] R. Agrawal, Y.-H. Chen, T. Horel, and S. Vadhan, "Unifying Computational Entropies via Kullback–Leibler Divergence," in *Advances in Cryptology – CRYPTO 2019*, A. Boldyreva and D. Micciancio, Eds. Cham: Springer International Publishing, 2019, pp. 831–858.
[34] Z. Rached, F. Alajaji, and L. L. Campbell, "The Kullback-Leibler divergence rate between markov sources," *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 917–921, 2004.
[35] J. . Durrieu, J. . Thiran, and F. Kelly, "Lower and upper bounds for approximation of the Kullback-Leibler divergence between gaussian mixture models," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4833–4836.
[36] J. R. Hershey and P. A. Olsen, "Approximating the Kullback-Leibler divergence between gaussian mixture models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–317.
[37] N. Bouhlel and A. Dziri, "Kullback–Leibler divergence between multivariate generalized gaussian distributions," *IEEE Signal Processing Letters*, vol. 26, no. 7, pp. 1021–1025, 2019.
[38] F. Nielsen and K. Sun, "Guaranteed bounds on the Kullback-Leibler divergence of univariate mixtures," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1543–1546, 2016.
[39] F. Nielsen, "On the Kullback-Leibler divergence between discrete normal distributions," *arXiv preprint arXiv:2109.14920*, 2021.
[40] W. Stummer and I. Vajda, "On Bregman distances and divergences of probability measures," *IEEE Trans. Inf. Theor.*, vol. 58, no. 3, p. 1277–1288, mar 2012. [Online]. Available: https://doi.org/10.1109/TIT.2011.2178139
[41] R. Agrawal and T. Horel, "Optimal bounds between f-divergences and integral probability metrics," *Journal of Machine Learning Research*, vol. 22, no. 128, pp. 1–59, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-867.html
[42] J. Birrell, P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and L. Rey-Bellet, "(f , Γ)-divergences: Interpolating between f -divergences and integral probability metrics," *Journal of Machine Learning Research*, vol. 23, no. 39, pp. 1–70, 2022. [Online]. Available: http://jmlr.org/papers/v23/21-0100.html
[43] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, 2006.
[44] I. Sason and S. Verdú, "f -divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.
[45] S.-I. Amari, "α-divergence is unique, belonging to both f -divergence and Bregman divergence classes," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 4925–4931, 2009.

[46] S. G. Bobkov, G. Chistyakov, and F. Götze, "Rényi divergence and the central limit theorem," *The Annals of Probability*, vol. 47, no. 1, pp. 270–323, 2019.

[47] H. Cai, S. Kulkarni, and S. Verdu, "Universal divergence estimation for finite-alphabet sources," *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp. 3456–3475, 2006.

[48] Q. Wang, S. R. Kulkarni, and S. Verdu, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2392–2405, 2009.

[49] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Inf. Theor.*, vol. 56, no. 11, p. 5847–5861, Nov. 2010.

[50] T. Kanamori, T. Suzuki, and M. Sugiyama, "f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 708–720, 2012.

[51] M. Gil, F. Alajaji, and T. Linder, "Rényi divergence measures for commonly used univariate continuous distributions," *Information Sciences*, vol. 249, pp. 124–131, 2013.

[52] K. R. Moon and A. O. Hero, "Ensemble estimation of multivariate f-divergence," in *2014 IEEE International Symposium on Information Theory*, 2014, pp. 356–360.

[53] K. Moon and A. Hero, "Multivariate f-divergence estimation with confidence," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/f4573fc71c731d5c362f0d7860945b88-Paper.pdf

[54] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "Estimation of KL divergence: Optimal minimax rate," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2648–2674, 2018.

[55] P. K. Rubenstein, O. Bousquet, J. Djolonga, C. Riquelme, and I. O. Tolstikhin, "Practical and consistent estimation of f-divergences." *Annual Conference on Neural Information Processing Systems*, vol. abs/1905.11112, pp. 4072–4082, 2019.

[56] P. Zhao and L. Lai, "Minimax optimal estimation of KL divergence for continuous distributions," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7787–7811, 2020.

[57] F. Nielsen, "Fast Approximations of the Jeffreys Divergence between Univariate Gaussian Mixtures via Mixture Conversions to Exponential-Polynomial Distributions," *Entropy*, vol. 23, no. 11, 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/11/1417

[58] S. Sreekumar and Z. Goldfeld, "Neural estimation of statistical divergences," *Journal of Machine Learning Research*, vol. 23, no. 126, pp. 1–75, 2022. [Online]. Available: http://jmlr.org/papers/v23/21-1212.html

[59] Z. Rached, F. Alajaji, and L. L. Campbell, "Rényi's divergence and entropy rates for finite alphabet markov sources," *IEEE Transactions on Information theory*, vol. 47, no. 4, pp. 1553–1561, 2001.

[60] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 209–216.

[61] K. T. Abou-Moustafa and F. P. Ferrie, "A note on metric properties for some divergence measures: The gaussian case," in *Asian Conference on Machine Learning*. PMLR, 2012, pp. 1–15.

[62] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/cedebb6e872f539bef8c3f919874e9d7-Paper.pdf

[63] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5769–5779.

[64] P. Donnat, G. Marti, and P. Very, "Toward a generic representation of random variables for machine learning," *Pattern Recogn. Lett.*, vol. 70, no. C, p. 24–31, Jan. 2016. [Online]. Available: https://doi.org/10.1016/j.patrec.2015.11.004

[65] A. Ghosh and A. Basu, "A new family of divergences originating from model adequacy tests and application to robust statistical inference," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5581–5591, 2018.

[66] P. Yang and B. Chen, "Robust Kullback-Leibler divergence and universal hypothesis testing for continuous distributions," *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 2360–2373, 2019.

[67] R. F. Vigelis, L. H. F. D. Andrade, and C. C. Cavalcante, "Properties of a generalized divergence related to Tsallis generalized divergence," *IEEE Transactions on Information Theory*, vol. 66, no. 5, pp. 2891–2897, 2020.

APPENDIX A
PROOF OF LEMMA 1

Proof 13.

(a) This is because $f'(x) = 1 - \frac{1}{x}$, $f''(x) = \frac{1}{x^2} > 0$

(b) We note $\Delta(x) = f(\frac{1}{x}) - f(x) = \frac{1}{x} - x + 2\log x$. Then $\Delta'(x) = -(\frac{1}{x} - 1)^2 \leq 0$ and $\Delta(1) = 0$ So it is easy to know Lemma 1b holds.

(c) We can verify this by definition.

$$\begin{aligned} y - \log y = x \Longleftrightarrow & e^{y-x} = y \Longleftrightarrow (-y)e^{-y} = -e^{-x} \\ \Longleftrightarrow & y = -W(-e^{-x}) \end{aligned} \tag{147}$$

(d) We can get Lemma 1d from 1c immediately.

(e) According to Equation (4), we can have

$$\begin{aligned} \frac{\mathrm{d}w_1(t)}{\mathrm{d}t} &= -\frac{\mathrm{d}(W_0(-e^{-(1+t)}))}{\mathrm{d}t} = \frac{-W_0(-e^{-(1+t)})}{-e^{-(1+t)}(1+W_0(-e^{-(1+t)}))} \\ &\times \frac{\mathrm{d}(-e^{-(1+t)})}{\mathrm{d}t} = \frac{W_0(-e^{-(1+t)})}{W_0(-e^{-(1+t)})+1} = \frac{-w_1(t)}{1-w_1(t)} \end{aligned}$$

The derivative of $w_2(t)$ can be computed in a similar way.

(f) From Lemma 1b, we can know Lemma 1f.

(g) This is because

$$f(x) \leq 1 + t \Longrightarrow w_1(t) < x < w_2(t) \Longrightarrow \quad \frac{1}{w_2(t)} < \frac{1}{x} < \frac{1}{w_1(t)}$$

Combining Lemma 1b, we have

$$f(\frac{1}{w_2(t)}) < f(w_2(t)) = 1 + t = f(w_1(t)) < f(\frac{1}{w_1(t)})$$

Thus Equation (7) holds. It is also easy to know that $S(t) = f(\frac{1}{w_1(t)})$ is continuous and strictly increasing with t.

(h) We have

$$\begin{aligned} f(x) \geq 1 + t \ &\Longrightarrow x \leq w_1(t) \vee x \geq w_2(t) \\ &\Longrightarrow \frac{1}{x} \leq \frac{1}{w_2(t)} \vee \frac{1}{x} \geq \frac{1}{w_1(t)} \end{aligned}$$

Combining Lemma 1b, we have $f(\frac{1}{w_2(t)}) < f(\frac{1}{w_1(t)})$, so we have Lemma 1h.

(i) Since $f'(x) = 1 - \frac{1}{x}$ and $w_2(t) \geq 1$ for $t \geq 0$, we have

$$\begin{aligned} f'(w_2(t)) =& 1 - \frac{1}{w_2(t)} = \frac{w_2(t)-1}{w_2(t)} \\ \leq& w_2(t) - 1 = -\left(1 - \frac{1}{\frac{1}{w_2(t)}}\right) = -f'(\frac{1}{w_2(t)}) \end{aligned} \tag{148}$$

(j)

$$\begin{aligned} & f(w_1(t_1)w_1(t_2)) \\ =& w_1(t_1)w_1(t_2) - \log w_1(t_1)w_1(t_2) \\ =& w_1(t_1)w_1(t_2) + (w_1(t_1) - \log w_1(t_1)) \\ & + (w_1(t_2) - \log w_1(t_2)) - w_1(t_1) - w_1(t_2) \\ =& w_1(t_1)w_1(t_2) + 1 + t_1 + 1 + t_2 - w_1(t_1) - w_1(t_2) \\ =& t_1 + t_2 + 2 + w_1(t_1)w_1(t_2) - w_1(t_1) - w_1(t_2) \end{aligned} \tag{149}$$

where the third equation follows from Lemma 1d. Equation (10) can be proved in a similar way.

□

APPENDIX B
PROOF OF LEMMA 3B

Proof 14.

The condition $KL(\mathcal{N}(0,I)||\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})) \leq \varepsilon$ is equal to the following conditions

$$\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) \leq n + \varepsilon_1 \tag{150}$$

$$\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \leq 2\varepsilon - \varepsilon_1 \tag{151}$$

$$0 \leq \varepsilon_1 \leq 2\varepsilon \tag{152}$$

We can apply Lemma 2 on Equation (150) and get

$$\begin{aligned} & -\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}) \\ \leq& \frac{1}{-W_0(-e^{-(1+\varepsilon_1)})} - \log\frac{1}{-W_0(-e^{-(1+\varepsilon_1)})} + n - 1 \end{aligned}$$

Applying Lemma 1g on Equation (150), we get

$$w_1(\varepsilon_1) < \lambda' < w_2(\varepsilon_1) \tag{153}$$

From Equation (151) we know $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \leq 2\varepsilon - \varepsilon_1$. Since $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \geq \lambda'_*\boldsymbol{\mu}^\top\boldsymbol{\mu}$ where λ'_* is the minimum eigenvalue of $\boldsymbol{\Sigma}^{-1}$, combining Equation (153), we can know

$$\boldsymbol{\mu}^\top\boldsymbol{\mu} \leq \frac{2\varepsilon - \varepsilon_1}{\lambda'_*} \leq \frac{2\varepsilon - \varepsilon_1}{w_1(\varepsilon_1)} \tag{154}$$

Adding the two sides of Inequalities (153), and (154), we get the same result as Equation (40). Therefore, we can get the same supremum as follows.

$$\begin{aligned} & KL(\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})||\mathcal{N}(0,I)) \\ \leq& \frac{1}{2}\left(\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - \log\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - 1\right) \end{aligned}$$

Inequality (155) is tight only when there exists one $\lambda'_j = -W_0(-e^{-(1+2\varepsilon)})$ and all other $\lambda'_i = 1$ for $i \neq j$, and $|\boldsymbol{\mu}| = 0$.

□

APPENDIX C
PROOF OF THEOREM 1

Proof 15.

For $X \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, there exists an invertible matrix B such that $X' = B^{-1}(X-\boldsymbol{\mu}) \sim \mathcal{N}(0,I)$ [2]. Here $B = PD^{1/2}$, P is an orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{\Sigma}$, $D = diag(\lambda_1,\ldots,\lambda_n)$ whose diagonal elements are the corresponding eigenvalues. For $X_1 \sim \mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $X_2 \sim \mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$, we define the following linear transformations T_1, T_2

$$X_1^1 = T_1(X_1) = B_1^{-1}(X_1 - \boldsymbol{\mu}_1) \text{ such that } X_1^1 \sim \mathcal{N}(0,I) \tag{155}$$

$$X_2^2 = T_2(X_2) = B_2^{-1}(X_2 - \boldsymbol{\mu}_2) \text{ such that } X_2^2 \sim \mathcal{N}(0,I) \tag{156}$$

and the reverse transformations T_1^{-1}, T_2^{-1} such that $X_1 = T_1^{-1}(X_1^1) = B_1X_1^1 + \boldsymbol{\mu}_1$ and $X_2 = T_2^{-1}(X_2^2) = B_2X_1^2 + \boldsymbol{\mu}_2$, where $p_{X_1^1} = p_{X_2^2} = \mathcal{N}(0,I)$. Besides, it is easy to know $X_1^2 = T_2(X_1) = B_2^{-1}(X_1 - \boldsymbol{\mu}_2)$ and $X_2^1 = T_1(X_2) = B_1^{-1}(X_2 - \boldsymbol{\mu}_1)$ are both Gaussian variables. We also have

$$X_1^2 \sim \mathcal{N}(B_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), B_2^{-1}\boldsymbol{\Sigma}_1(B_2^{-1})^\top) \tag{157}$$

$$X_2^1 \sim \mathcal{N}(B_1^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), B_1^{-1}\boldsymbol{\Sigma}_2(B_1^{-1})^\top) \tag{158}$$

With the help of invertible linear transformations, we can convert the KL divergence between two arbitrary Gaussians into that between one Gaussian and standard Gaussian. According to Proposition 1, diffeomorphisms preserve KL divergence. If we apply T_2 simultaneously on X_1, X_2, we can have

$$\begin{aligned} KL(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)) =& KL(p_{X_1^2}||p_{X_2^2}) \\ =& KL(p_{X_1^2}||\mathcal{N}(0,I)) \end{aligned} \tag{159}$$

Then we can apply T_2^{-1} on X_1^2, X_2^2 and also have

$$KL(\mathcal{N}(0,I)||p_{X_1^2}) = KL(p_{X_2^2}||p_{X_1^2}) \tag{160}$$

$$= KL(\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)||\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)) \tag{161}$$

According to precondition, it is easy to know $KL(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)) = KL(p_{X_1^2}||\mathcal{N}(0,I))$. Applying Theorem 3a on $KL(p_{X_1^2}||\mathcal{N}(0,I))$, we can prove

$$\begin{aligned} &KL(\mathcal{N}(0,I)||p_{X_1^2}) = KL(\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)||\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)) \\ \leq& \frac{1}{2}\left(\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - \log\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - 1\right) \end{aligned}$$

Similarly, if we use T_1 simultaneously on X_1 and X_2, we can get the same result.

Inequality (162) is tight when there exists only one eigenvalue λ_j of $B_2^{-1}\boldsymbol{\Sigma}_1(B_2^{-1})^\top$ or $B_1^{-1}\boldsymbol{\Sigma}_2(B_1^{-1})^\top$ is equal to $-W_0(-e^{-(1+2\varepsilon)})$ and all other eigenvalues λ_i $(i \neq j)$ are equal to 1, and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

□

APPENDIX D
PROOF OF THEOREM 2

Proof 16.

When ε is small, we can use the series expanding W_0 (see Section III.17 in [20]) to simplify the bound in Theorem 1.

Notice that when ε is small, $-W_0(-e^{-(1+2\varepsilon)})$ is close to 1. According to the series expanding W_0 (see Section III.17 in [20]), we have

$$W_0(-e^{-(1+2\varepsilon)}) = -1 + 2\sqrt{\varepsilon} - \frac{4}{3}\varepsilon + \frac{2}{9}\varepsilon^{1.5} + O(\varepsilon^2) \tag{162}$$

Now expand the log term around $-W_0(-e^{-(1+2\varepsilon)}) = 1$ using Taylor series for small ε.

$$\begin{aligned} &\log(-W_0(-e^{-(1+2\varepsilon)})) \\ =& \log(1 - W_0(-e^{-(1+2\varepsilon)}) - 1) \\ =& -W_0(-e^{-(1+2\varepsilon)}) - 1 - \frac{1}{2}\left(-W_0(-e^{-(1+2\varepsilon)}) - 1\right)^2 \\ &+ \frac{1}{3}\left(-W_0(-e^{-(1+2\varepsilon)}) - 1\right)^3 + O\left((-W_0(-e^{-(1+2\varepsilon)}) - 1)^4\right) \\ =& -2\sqrt{\varepsilon} + \frac{4}{3}\varepsilon - \frac{2}{9}\varepsilon^{1.5} + O(\varepsilon^2) \\ &- \frac{1}{2}\left(-2\sqrt{\varepsilon} + \frac{4}{3}\varepsilon - \frac{2}{9}\varepsilon^{1.5} + O(\varepsilon^2)\right)^2 \\ &+ \frac{1}{3}\left(-2\sqrt{\varepsilon} + \frac{4}{3}\varepsilon - \frac{2}{9}\varepsilon^{1.5} + O(\varepsilon^2)\right)^3 + O(\varepsilon^2) \end{aligned} \tag{163}$$

$$= -2\sqrt{\varepsilon} - \frac{2}{3}\varepsilon - \frac{2}{9}\varepsilon^{1.5} + O(\varepsilon^2) \tag{164}$$

Plugging Equation (162) and (164) into the bound in Theorem 1, we can have

$$\begin{aligned} &KL(\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)||\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)) \\ \leq& \frac{1}{2}\left(\frac{1}{1 - 2\sqrt{\varepsilon} + \frac{4}{3}\varepsilon - \frac{2}{9}\varepsilon^{1.5} + O(\varepsilon^2)}\right. \\ &\left. + \left(-2\sqrt{\varepsilon} - \frac{2}{3}\varepsilon - \frac{2}{9}\varepsilon^{1.5} + O(\varepsilon^2)\right) - 1\right) \\ =& \frac{1}{2}\frac{2\varepsilon - \frac{4}{3}\varepsilon^{1.5} + O(\varepsilon^2)}{1 - 2\sqrt{\varepsilon} + \frac{4}{3}\varepsilon - \frac{2}{9}\varepsilon^{1.5} + O(\varepsilon^2)} \\ =& \varepsilon + \frac{2\varepsilon^{1.5} + O(\varepsilon^2)}{1 - 2\sqrt{\varepsilon} + \frac{4}{3}\varepsilon - \frac{2}{9}\varepsilon^{1.5} + O(\varepsilon^2)} \\ =& \varepsilon + 2\varepsilon^{1.5} + O(\varepsilon^2) \end{aligned} \tag{165}$$

□

APPENDIX E
THE FIRST PROOF OF THEOREM 3

Theorem 3 can be proved using the similar method as that of Theorem 1, except that the proof uses W_{-1}. We put the key steps of the proof of Theorem 3 in Lemma 10 and Lemma 11.

Lemma 10. Given n-ary function $\bar{\boldsymbol{f}}(\boldsymbol{x}) = \bar{\boldsymbol{f}}(x_1,\ldots,x_n) = \sum_{i=1}^n x_i - \log x_i$ $(x_i \in \mathbb{R}^{++})$, if $\bar{\boldsymbol{f}}(x_1,\ldots,x_n) \geq n + M (M > 0)$, then

$$\begin{aligned} &\inf \bar{\boldsymbol{f}}(\frac{1}{x_1},\ldots,\frac{1}{x_n}) \\ =& \frac{1}{-W_{-1}(-e^{-(1+M)})} - \log\frac{1}{-W_{-1}(-e^{-(1+M)})} + n - 1 \end{aligned}$$

Proof 17.

The structure of proof of Lemma 10 is similar to that of Lemma 2. The constraint in the following optimization problem

$$\text{minimize } \bar{\boldsymbol{f}}(\frac{1}{x_1},\ldots,\frac{1}{x_n}) \tag{166}$$

$$s.t. \sum_{i=1}^n x_i - \log x_i \geq n + M \tag{167}$$

can be replaced by the following constraints

$$\bigwedge_{i=1}^n f(x_i) = x_i - \log x_i \geq 1 + M_i \wedge \bigwedge_{i=1}^n M_i \geq 0 \wedge \sum_{i=1}^n M_i \geq M \tag{168}$$

Given fixed $M_1,\ldots,M_n$ such that $\bigwedge_{i=1}^n M_i \geq 0 \wedge \sum_{i=1}^n M_i \geq M$, we define

$$\begin{aligned} \bar{\boldsymbol{I}}(M_1,\ldots,M_n) =& \inf_{\bigwedge_{i=1}^n f(x_i) \geq 1 + M_i} \bar{\boldsymbol{f}}(\frac{1}{x_1},\ldots,\frac{1}{x_n}) \\ =& \sum_{i=1}^n \inf_{f(x_i) \geq 1 + M_i} f(\frac{1}{x_i}) = \sum_{i=1}^n I(M_i) \end{aligned} \tag{169}$$

So we have

$$\inf \bar{\boldsymbol{f}}\big(\frac{1}{x_1},\ldots,\frac{1}{x_n}\big) = \inf_{\substack{\bigwedge_{i=1}^n M_i\geq 0\\ \sum_{i=1}^n M_i\geq M}} \bar{\boldsymbol{I}}(M_1,\ldots,M_n) \tag{170}$$

It is easy to know that $\bar{\boldsymbol{I}}(M_1,\ldots,M_n)$ is continuous and strictly increasing with $M_1,\ldots,M_n$. So the condition $\sum_{i=1}^n M_i \geq M$ in Equation (170) can be changed to $\sum_{i=1}^n M_i = M$.

The remaining proof consists of two steps. We find $\bar{\boldsymbol{I}}(M_1,\ldots,M_n)$ for fixed $M_1,\ldots,M_n$ in the first step, and then find $\inf \bar{\boldsymbol{I}}(M_1,\ldots,M_n)$ for any $M_1,\ldots,M_n$ satisfying $\bigwedge_{i=1}^n M_i \geq 0 \wedge \sum_{i=1}^n M_i = M$ in the second step.

Step 1: According to Lemma 1h, for fixed M_i, we get

$$I(M_i) = \inf_{f(x)\geq 1+M_i} f\big(\frac{1}{x}\big) = f\big(\frac{1}{w_2(M_i)}\big) \tag{171}$$

Combining Equation (169), we know

$$\bar{\boldsymbol{I}}(M_1,\ldots,M_n) = \sum_{i=1}^n f\big(\frac{1}{w_2(M_i)}\big) \tag{172}$$

Step 2: We define function

$$\begin{aligned}\Delta(M) =& f(w_2(M)) - f\big(\frac{1}{w_2(M)}\big)\\ =& w_2(M) - \frac{1}{w_2(M)} - 2\log w_2(M)\end{aligned} \tag{173}$$

Similarly, we can prove $\Delta(tM) \leq t\Delta(M)$ $(0 \leq t < 1)$ by showing $\Delta(0) = 0$ (apparently) and $\Delta(M)$ is strictly increasing and strictly convex. Combining Lemma 1e, we get the derivative of $\Delta(M)$ as

$$\frac{\mathrm{d}\Delta(M)}{\mathrm{d}M} = \left(1 + \frac{1}{w_2(M)^2} - \frac{2}{w_2(M)}\right) \times \frac{\mathrm{d}w_2(M)}{\mathrm{d}M} = 1 - \frac{1}{w_2(M)} \tag{174}$$

The second order derivative of $\Delta(M)$ is

$$\frac{\mathrm{d}^2\Delta(M)}{\mathrm{d}M^2} = \frac{1}{w_2(M)^2} \times \frac{w_2(M)}{w_2(M)-1} = \frac{1}{w_2(M)(w_2(M)-1)} \tag{175}$$

Since $w_2(M) \in (1,+\infty)$ for $M > 0$, so it is easy to know $\frac{\mathrm{d}\Delta(M)}{\mathrm{d}M} > 0$, $\frac{\mathrm{d}^2\Delta(M)}{\mathrm{d}M^2} > 0$ for $M > 0$. This implies that $\Delta(M)$ is strictly increasing and strictly convex. We can use the similar deduction as Lemma 2 to prove $\Delta(tM) \leq t\Delta(M)$. Thus, we have

$$\begin{aligned}\bar{\boldsymbol{\Delta}}(M_1,\ldots,M_n) &= \sum_{i=1}^n f(w_2(M_i)) - f\big(\frac{1}{w_2(M_i)}\big) = \sum_{i=1}^n \Delta(M_i)\\ &= \sum_{i=1}^n \Delta\big(\frac{M_i}{M}M\big) \leq \sum_{i=1}^n \frac{M_i}{M}\Delta(M) = \Delta(M)\end{aligned} \tag{176}$$

Inequality (176) is tight when there exists only one $M_j = M$ and all other $M_i = 0$ for $i \neq j$. Therefore, from Inequality (176), we can obtain

$$\begin{aligned}&\bar{\boldsymbol{I}}(M_1,\ldots,M_n)\\ &= \textstyle\sum_{i=1}^n f\big(\frac{1}{w_2(M_i)}\big) && \big\downarrow (172)\\ &\geq \textstyle\sum_{i=1}^n f(w_2(M_i)) - \Delta(M) && \big\downarrow (176)\\ &= \textstyle\sum_{i=1}^n (1+M_i) - \big(f(w_2(M)) - f\big(\frac{1}{w_2(M)}\big)\big) && \big\downarrow \text{Lemma 1d (173)}\\ &= n + M - (1+M) + f\big(\frac{1}{w_2(M)}\big)\\ &= \frac{1}{-W_{-1}(-e^{-(1+M)})} - \log\frac{1}{-W_{-1}(-e^{-(1+M)})}\\ &\quad + n - 1\end{aligned} \tag{177}$$

Finally, we can conclude that

$$\begin{aligned}&\inf \bar{\boldsymbol{f}}\big(\frac{1}{x_1},\ldots,\frac{1}{x_n}\big) = \inf_{\substack{\bigwedge_{i=1}^n M_i\geq 0\\ \sum_{i=1}^n M_i = M}} \bar{\boldsymbol{I}}(M_1,\ldots,M_n)\\ =& \frac{1}{-W_{-1}(-e^{-(1+M)})} - \log\frac{1}{-W_{-1}(-e^{-(1+M)})} + n - 1\end{aligned} \tag{178}$$

Similarly, $\bar{\boldsymbol{f}}(1/x_1,\ldots,1/x_n)$ reaches its infimum when there exists only one j such that $x_j = -W_{-1}(-e^{-(1+M)})$ and $f(x_i) = 1$ for $i \neq j$. □

The following Lemma 11 gives the infimum of KL divergence when one Gaussian is standard.

Lemma 11. For any n-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$,

(a) If $KL(\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})||\mathcal{N}(0,I)) \geq M (M > 0)$, then

$$\begin{aligned}&KL(\mathcal{N}(0,I)||\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}))\\ \geq& \frac{1}{2}\left(\frac{1}{-W_{-1}(-e^{-(1+2M)})} - \log\frac{1}{-W_{-1}(-e^{-(1+2M)})} - 1\right)\end{aligned}$$

(b) If $KL(\mathcal{N}(0,I)||\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})) \geq M$, then

$$\begin{aligned}&KL(\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})||\mathcal{N}(0,I))\\ \geq& \frac{1}{2}\left(\frac{1}{-W_{-1}(-e^{-(1+2M)})} - \log\frac{1}{-W_{-1}(-e^{-(1+2M)})} - 1\right)\end{aligned}$$

Proof 18. (a) We first consider the case when $KL(\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})||\mathcal{N}(0,I)) = M$. At the end of the proof, we deal with the case when $KL(\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})||\mathcal{N}(0,I)) \geq M$.

The condition $-\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top\boldsymbol{\mu} - n = 2M$ is equal to

$$-\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^n \lambda_i - \log\lambda_i = n + M_1 \tag{179}$$

$$\boldsymbol{\mu}^\top\boldsymbol{\mu} = 2M - M_1 \tag{180}$$

where $0 \leq M_1 \leq 2M$.

Applying Lemma 10 on Equation (179), we can get

$$\begin{aligned}&\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) = \sum_{i=1}^n \frac{1}{\lambda_i} - \log\frac{1}{\lambda_i}\\ \geq& \frac{1}{-W_{-1}(-e^{-(1+M_1)})} - \log\frac{1}{-W_{-1}(-e^{-(1+M_1)})} + n - 1\end{aligned}$$

Inequality (181) is tight when all eigenvalues λ_i of $\boldsymbol{\Sigma}$ are equal to 1 except for one $\lambda_j = -W_{-1}(-e^{-(1+M_1)})$.

From Equation (180), we know $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \geq \lambda'_*\boldsymbol{\mu}^\top\boldsymbol{\mu} = \lambda'_*(2M - M_1)$ where λ'_* is the smallest eigenvalue of

$\boldsymbol{\Sigma}^{-1}$. Here $\lambda^* = 1/\lambda'_*$ is the largest eigenvalue of $\boldsymbol{\Sigma}$. From Equation (179), Lemma 1a and 1g, we know $\lambda^* \leq -W_{-1}(-e^{-(1+M_1)})$. So we obtain

$$\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \geq \frac{2M - M_1}{-W_{-1}(-e^{-(1+M_1)})} \tag{181}$$

Note that, inequalities (181) and (181) become tight simultanously when the same condition holds. Now combining Equation (181) and (181), we obtain

$$\begin{aligned}
&\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - n \\
\geq &\frac{1}{-W_{-1}(-e^{-(1+M_1)})} - \log \frac{1}{-W_{-1}(-e^{-(1+M_1)})} \\
&+ \frac{2M - M_1}{-W_{-1}(-e^{-(1+M_1)})} - 1 \\
= &\frac{1 + 2M - M_1}{w_2(M_1)} - \log \frac{1}{w_2(M_1)} - 1 = L(M_1) \ (0 \leq M_1 \leq 2M)
\end{aligned} \tag{182}$$

It is easy to know that $L'(M_1) = \frac{M_1 - 2M}{w_2(M_1)(w_2(M_1)-1)}$. Since $M_1 \leq 2M$ and $w_2(M_1) > 1$ for $M_1 > 0$, so $L'(M_1) < 0$ $(M_1 > 0)$. This indicates that $L(M_1) > L(2M)$ for $0 < M_1 < 2M$. Thus, we can conclude

$$\begin{aligned}
&KL(\mathcal{N}(0, I) || \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \\
\geq &\frac{1}{2} L(2M) \\
= &\frac{1}{2} \left(\frac{1}{-W_{-1}(-e^{-(1+2M)})} - \log \frac{1}{-W_{-1}(-e^{-(1+2M)})} - 1 \right)
\end{aligned}$$

Inequality (183) is tight when there exist only one eigenvalue λ_j of $\boldsymbol{\Sigma}$ equal to $-W_{-1}(-e^{-(1+2M)})$ and all other eigenvalues $\lambda_i (i \neq j)$ are equal to 1, and $\boldsymbol{\mu} = 0$.

Finally, we can consider the case when $KL(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) || \mathcal{N}(0, I)) \geq M$. The bound in Equation (183) is strictly increasing with M. Therefore, the precondition $KL(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) || \mathcal{N}(0, I)) = M$ can be changed to $KL(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) || \mathcal{N}(0, I)) \geq M$.

(b) The proof of Lemma 3b is the similar to that of Lemma 3a. We list it here for clarity.

The condition $KL(\mathcal{N}(0, I) || \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = M$ is equal to

$$\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) = n + M_1 \tag{183}$$

$$\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} = 2M - M_1 \tag{184}$$

where $0 \leq M_1 \leq 2M$. Applying Lemma 10 on Equation (183), we can obtain

$$\begin{aligned}
&-\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}) \\
\geq &\frac{1}{-W_{-1}(-e^{-(1+M_1)})} - \log \frac{1}{-W_{-1}(-e^{-(1+M_1)})} + n - 1
\end{aligned}$$

From Equation (183) and Lemma 1a and 1g, we have $\lambda' \leq -W_{-1}(-e^{-(1+M_1)})$ where λ' is the eigenvalue of $\boldsymbol{\Sigma}^{-1}$. Now let λ'^* be the largest eigenvalues of $\boldsymbol{\Sigma}^{-1}$. It is easy to know

$$\begin{aligned}
&\lambda'^* \boldsymbol{\mu}^\top \boldsymbol{\mu} \geq \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} = 2M - M_1 \\
\Longrightarrow &\boldsymbol{\mu}^\top \boldsymbol{\mu} \geq \frac{2M - M_1}{-W_{-1}(-e^{-(1+M_1)})}
\end{aligned} \tag{185}$$

Inequalities (185) and (185) are tight simutanously when there exist only one eigenvalue $\lambda'_j = -W_{-1}(-e^{-(1+M_1)})$ and all other eigenvalues are equal to 1, and $|\boldsymbol{\mu}| = 0$. Therefore, combining Equation (185) and (185), we obtain

$$\begin{aligned}
&-\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \boldsymbol{\mu} - n \\
\geq &\frac{1}{-W_{-1}(-e^{-(1+M_1)})} - \log \frac{1}{-W_{-1}(-e^{-(1+M_1)})} \\
&+ \frac{2M - M_1}{-W_{-1}(-e^{-(1+M_1)})} - 1
\end{aligned} \tag{186}$$

Finally, using the similar analysis as Equation (183), we can conclude that

$$\begin{aligned}
&KL(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) || \mathcal{N}(0, I)) \\
\geq &\frac{1}{2} \left(\frac{1}{-W_{-1}(-e^{-(1+2M)})} - \log \frac{1}{-W_{-1}(-e^{-(1+2M)})} - 1 \right)
\end{aligned}$$

Similarly, the precondition $KL(\mathcal{N}(0, I) || \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = M$ can be changed to $KL(\mathcal{N}(0, I) || \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \geq M$ because the bound in Equation (187) is strictly increasing with M.

□

Notes. It needs strict conditions to reach the infimum in Lemma 11.

Now we can also obtain Theorem 3 on two general Gaussians. We can use linear transformation on Gaussians and apply Lemma 11 on them as what we do in the main proof of Theorem 1. The key steps have been proven in Lemma 10 and 11. More details are ommited.

APPENDIX F
PROOF OF LEMMA 5

Proof 19.

With the helper functions f_l, f_r (Equation (61)), we define function

$$\Delta_w(\varepsilon) = g_r(\varepsilon) - g_l(\varepsilon) = (w_2(\varepsilon) - 1) - (1 - w_1(\varepsilon)) \tag{187}$$

It is straightforward to know

$$\Delta_w(0) = w_2(0) - 1 - (1 - w_1(0)) = 0 \tag{188}$$

In the following, we prove $\Delta_w{}'(\varepsilon) > 0$ for $\varepsilon > 0$. Plugging Equation (62), we have

$$\begin{aligned}
g'_r(\varepsilon) = &f_r^{-1\prime}(\varepsilon) = \frac{1}{f'_r(f_r^{-1}(\varepsilon))} = \frac{1}{f'_r(w_2(\varepsilon) - 1)} \\
= &\frac{1}{1 - \frac{1}{w_2(\varepsilon)}} = \frac{1}{f'(w_2(\varepsilon))}
\end{aligned} \tag{189}$$

$$\begin{aligned}
g'_l(\varepsilon) = &f_l^{-1\prime}(\varepsilon) = \frac{1}{f'_l(f_l^{-1}(\varepsilon))} = \frac{1}{f'_l(1 - w_1(\varepsilon))} \\
= &\frac{1}{\frac{1}{w_1(\varepsilon)} - 1} = \frac{1}{-f'(w_1(\varepsilon))}
\end{aligned} \tag{190}$$

According to Lemma 1 and Lemma 1f, $f(x)$ is strictly decreasing in $(0, 1)$ and $f(w_1(\varepsilon)) > f(\frac{1}{w_2(\varepsilon)})$. So we can know $w_1(\varepsilon) < \frac{1}{w_2(\varepsilon)}$. Since $f(x)$ is convex and $f'(x) < 0$ in $(0, 1)$, we can know $f'(w_1(\varepsilon)) < f'(\frac{1}{w_2(\varepsilon)})$. Now combining Lemma 1i, we can obtain $0 < f'(w_2(\varepsilon)) \leq -f'(\frac{1}{w_2(\varepsilon)}) < -f'(w_1(\varepsilon))$ $(\varepsilon > 0)$. This leads to

$$g'_r(\varepsilon) = \frac{1}{f'(w_2(\varepsilon))} > \frac{1}{-f'(w_1(\varepsilon))} = g'_l(\varepsilon) \tag{191}$$

for $\varepsilon > 0$, which means $\Delta_w{}'(\varepsilon) = g_r'(\varepsilon) - g_l'(\varepsilon) > 0$ $(\varepsilon > 0)$. Now combining Equation (188), we can conclude

$$\Delta_w(\varepsilon) = g_r(\varepsilon) - g_l(\varepsilon) = (w_2(\varepsilon) - 1) - (1 - w_1(\varepsilon)) \geq 0 \tag{192}$$

□

APPENDIX G
PROOF OF LEMMA 6

Proof 20.

From Lemma 1g, we know $w_1(\varepsilon_x) \leq x \leq w_2(\varepsilon_x)$ and $w_1(\varepsilon_y) \leq y \leq w_2(\varepsilon_y)$. So we have $w_1(\varepsilon_x)w_1(\varepsilon_x) \leq xy \leq w_2(\varepsilon_x)w_2(\varepsilon_y)$. According to Lemma 1a, it suffices to show $f(w_1(\varepsilon_x)w_1(\varepsilon_y)) \leq f(w_2(\varepsilon_x)w_2(\varepsilon_y))$. By the definition of $f(x)$, we have

$$\begin{aligned}
&f(w_2(\varepsilon_x)w_2(\varepsilon_y)) - f(w_1(\varepsilon_x)w_1(\varepsilon_y)) \\
=&w_2(\varepsilon_x)w_2(\varepsilon_y) - \log(w_2(\varepsilon_x)w_2(\varepsilon_y)) \\
&- (w_1(\varepsilon_x)w_1(\varepsilon_y) - \log(w_1(\varepsilon_x)w_1(\varepsilon_y))) \\
=&w_2(\varepsilon_x)w_2(\varepsilon_y) - \log w_2(\varepsilon_x) - \log w_2(\varepsilon_y) \\
&- (w_1(\varepsilon_x)w_1(\varepsilon_y) - \log w_1(\varepsilon_x) - \log w_1(\varepsilon_y)) \\
=&w_2(\varepsilon_x)w_2(\varepsilon_y) - w_2(\varepsilon_x) + w_2(\varepsilon_x) - \log w_2(\varepsilon_x) \\
&- w_2(\varepsilon_y) + w_2(\varepsilon_y) - \log w_2(\varepsilon_y) \\
&- (w_1(\varepsilon_x)w_1(\varepsilon_y) - w_1(\varepsilon_x) + w_1(\varepsilon_x) - \log w_1(\varepsilon_x) \\
&- w_1(\varepsilon_y) + w_1(\varepsilon_y) - \log w_1(\varepsilon_y)) \\
=&w_2(\varepsilon_x)w_2(\varepsilon_y) - w_2(\varepsilon_x) + \varepsilon_x - w_2(\varepsilon_y) + \varepsilon_y \\
&- (w_1(\varepsilon_x)w_1(\varepsilon_y) - w_1(\varepsilon_x) + \varepsilon_x - w_1(\varepsilon_y) + \varepsilon_y) \\
=&w_2(\varepsilon_x)w_2(\varepsilon_y) - w_2(\varepsilon_x) - w_2(\varepsilon_y) \\
&- (w_1(\varepsilon_x)w_1(\varepsilon_y) - w_1(\varepsilon_x) - w_1(\varepsilon_y)) \\
=&w_2(\varepsilon_x)w_2(\varepsilon_y) - w_2(\varepsilon_x) - w_2(\varepsilon_y) + 1 \\
&- (w_1(\varepsilon_x)w_1(\varepsilon_y) - w_1(\varepsilon_x) - w_1(\varepsilon_y) + 1) \\
=&(w_2(\varepsilon_x) - 1)(w_2(\varepsilon_y) - 1) - (w_1(\varepsilon_x) - 1)(w_1(\varepsilon_y) - 1)
\end{aligned} \tag{193}$$

From Lemma 5, it is easy to know $w_2(\varepsilon_x) - 1 \geq 1 - w_1(\varepsilon_x)$ and $w_2(\varepsilon_y) - 1 \geq 1 - w_1(\varepsilon_y)$. Thus we can conclude

$$f(w_2(\varepsilon_x)w_2(\varepsilon_y)) - f(w_1(\varepsilon_x)w_1(\varepsilon_y)) \geq 0 \tag{194}$$

□

APPENDIX H
PROOF OF THEOREM 4

Proof 21.

For $X_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, there exists an invertible matrix B_2 such that $X_2' = B_2^{-1}(X_2 - \boldsymbol{\mu}_2) \sim \mathcal{N}(0, I)$ [2]. Here $B_2 = P_2 D_2^{1/2}$, P_2 is an orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{\Sigma}_2$, $D_2 = diag(\lambda_{2,1}, \ldots, \lambda_{2,n})$ whose diagonal elements are the corresponding eigenvalues. We define the following two invertible linear transformations T, T^{-1} on random vectors.

$$X' = T(X) = B_2^{-1}(X - \boldsymbol{\mu}_2),\ X = T^{-1}(X') = B_2 X' + \boldsymbol{\mu}_2 \tag{195}$$

Applying transformation T on X_1, X_2, X_3, we can get three Gaussians.

$$\begin{aligned}
X_1' = T(X_1) &\sim \mathcal{N}(\boldsymbol{\mu}_1', \boldsymbol{\Sigma}_1') \\
X_2' = T(X_2) &\sim \mathcal{N}(0, I) \\
X_3' = T(X_3) &\sim \mathcal{N}(\boldsymbol{\mu}_3', \boldsymbol{\Sigma}_3')
\end{aligned}$$

According to Proposition 1, T and T^{-1} preserve KL divergence. Thus, we have

$$KL(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) = KL(\mathcal{N}(\boldsymbol{\mu}_1', \boldsymbol{\Sigma}_1')||\mathcal{N}(0, I)) \tag{196}$$

$$KL(\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)||\mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)) = KL(\mathcal{N}(0, I)||\mathcal{N}(\boldsymbol{\mu}_3', \boldsymbol{\Sigma}_3')) \tag{197}$$

$$KL(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)) = KL(\mathcal{N}(\boldsymbol{\mu}_1', \boldsymbol{\Sigma}_1')||\mathcal{N}(\boldsymbol{\mu}_3', \boldsymbol{\Sigma}_3')) \tag{198}$$

Combining the preconditions and Equations (196), (197), we can know

$$KL(\mathcal{N}(\boldsymbol{\mu}_1', \boldsymbol{\Sigma}_1')||\mathcal{N}(0, I)) \leq \varepsilon_1,\ KL(\mathcal{N}(0, I)||\mathcal{N}(\boldsymbol{\mu}_2', \boldsymbol{\Sigma}_2')) \leq \varepsilon_2 \tag{199}$$

Now we can apply Lemma 9 on $\mathcal{N}(\boldsymbol{\mu}_1', \boldsymbol{\Sigma}_1')$, $\mathcal{N}(0, I)$ and $\mathcal{N}(\boldsymbol{\mu}_3', \boldsymbol{\Sigma}_3')$ and get the bound of $KL(\mathcal{N}(\boldsymbol{\mu}_1', \boldsymbol{\Sigma}_1')||\mathcal{N}(\boldsymbol{\mu}_3', \boldsymbol{\Sigma}_3'))$. Finally, combining Equation (198), we can prove Theorem 4.

□

APPENDIX I
PROOF OF THEOREM 5

Proof 22.

Suppose that $\varepsilon_1, \varepsilon_2$ are sufficiently small. According to the series expanding W_0 and W_1 (Section III.17 in [20]), we have

$$W_0(-e^{-(1+2\varepsilon)}) = -1 + 2\sqrt{\varepsilon} + O(\varepsilon) \tag{200}$$

$$W_{-1}(-e^{-(1+2\varepsilon)}) = -1 - 2\sqrt{\varepsilon} + O(\varepsilon) \tag{201}$$

So we can obtain

$$\begin{aligned}
&W_{-1}(-e^{-(1+2\varepsilon_1)})W_{-1}(-e^{-(1+2\varepsilon_2)}) + W_{-1}(-e^{-(1+2\varepsilon_1)}) \\
&+ W_{-1}(-e^{-(1+2\varepsilon_2)}) + 1 \\
=&(W_{-1}(-e^{-(1+2\varepsilon_1)}) + 1)(W_{-1}(-e^{-(1+2\varepsilon_2)}) + 1) \\
=&(2\sqrt{\varepsilon_1} + O(\varepsilon_1))(2\sqrt{\varepsilon_2} + O(\varepsilon_2)) \\
=&4\sqrt{\varepsilon_1\varepsilon_2} + o(\varepsilon_1) + o(\varepsilon_2)
\end{aligned} \tag{202}$$

and

$$\begin{aligned}
&- W_{-1}(-e^{-(1+2\varepsilon_2)})\left(\sqrt{2\varepsilon_1} + \sqrt{\frac{2\varepsilon_2}{-W_0(-e^{-(1+2\varepsilon_2)})}}\right)^2 \\
=&(1 + 2\sqrt{\varepsilon_2} + O(\varepsilon_2))\left(\sqrt{2\varepsilon_1} + \sqrt{\frac{2\varepsilon_2}{1 - 2\sqrt{\varepsilon_2} + O(\varepsilon_2)}}\right)^2 \\
\leq&(1 + 2\sqrt{\varepsilon_2} + O(\varepsilon_2))\left(4\varepsilon_1 + \frac{4\varepsilon_2}{1 - 2\sqrt{\varepsilon_2} + O(\varepsilon_2)}\right) \\
=&4\varepsilon_1 + o(\varepsilon_1) + o(\varepsilon_2) + \frac{4\varepsilon_2(1 + 2\sqrt{\varepsilon_2} + O(\varepsilon_2))}{1 - 2\sqrt{\varepsilon_2} + O(\varepsilon_2)} \\
=&4\varepsilon_1 + o(\varepsilon_1) + o(\varepsilon_2) + 4\varepsilon_2 + \frac{4\varepsilon_2(4\sqrt{\varepsilon_2} + O(\varepsilon_2))}{1 - 2\sqrt{\varepsilon_2} + O(\varepsilon_2)} \\
=&4\varepsilon_1 + 4\varepsilon_2 + o(\varepsilon_1) + o(\varepsilon_2) + O(\varepsilon_2^{1.5})
\end{aligned} \tag{203}$$

Using Equations (202) and (203), we can rewrite the bound in Theorem 4 as

$$\begin{aligned} &KL((\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\boldsymbol{\Sigma}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)) \\ &<3\varepsilon_1 + 3\varepsilon_2 + 2\sqrt{\varepsilon_1\varepsilon_2} + o(\varepsilon_1) + o(\varepsilon_2) \end{aligned} \tag{204}$$

□