Density Ratio Estimation via Sampling along Generalized Geodesics on Statistical Manifolds

Masanari Kimura **Howard Bondell**
The University of Melbourne

Abstract

The density ratio of two probability distributions is one of the fundamental tools in mathematical and computational statistics and machine learning, and it has a variety of known applications. Therefore, density ratio estimation from finite samples is a very important task, but it is known to be unstable when the distributions are distant from each other. One approach to address this problem is density ratio estimation using incremental mixtures of the two distributions. We geometrically reinterpret existing methods for density ratio estimation based on incremental mixtures. We show that these methods can be regarded as iterating on the Riemannian manifold along a particular curve between the two probability distributions. Making use of the geometry of the manifold, we propose to consider incremental density ratio estimation along generalized geodesics on this manifold. To achieve such a method requires Monte Carlo sampling along geodesics via transformations of the two distributions. We show how to implement an iterative algorithm to sample along these geodesics and show how changing the distances along the geodesic affect the variance and accuracy of the estimation of the density ratio.

1 Introduction

The density ratio of two probability distributions is one of the fundamental tools in mathematical and computational statistics. For example, its applications include estimation under the covariate shift assumption (Shimodaira, 2000; Sugiyama et al., 2007b), outlier detection (Azmandian et al., 2012a; Hido et al., 2011), variable selection (Azmandian et al., 2012b; Heck, 2019; Oh et al., 2016), change point detection (Hushchyn and Ustyuzhanin, 2021; Kawahara and Sugiyama, 2009; Liu et al., 2013) and causal inference (Matsushita et al., 2023; Reichenheim and Coutinho, 2010). For practical purposes, since the true density ratio is not accessible, density ratio estimation from finite size samples is required. Therefore, density ratio estimation (DRE) is a very important task that has attracted significant interest (Kanamori et al., 2010; Sugiyama et al., 2012a; Yamada et al., 2013).

A key challenge in DRE is the difficulty of the estimation when the source and target distributions are far apart. One approach to address this problem is DRE using incremental mixtures of two distributions. Rhodes et al. (2020) have shown the effectiveness of a method that creates $T > 0$ bridging distributions connecting the source and target distributions and then shrinks the gap between the two distributions by utilizing a divide-and-conquer framework. The method is based on the simple fact that a chain of density ratios using arbitrary bridge distributions restores the original density ratio. They employ a linear mixture of source and target distributions as the bridge distributions. In addition, Choi et al. (2022) extended this method to use infinitely continuum bridge distributions. We refer to the estimator that these methods produce as the Incremental Mixture Density Ratio Estimator (IMDRE).

We investigate the family of IMDRE within the framework of information geometry (Amari, 2016; Amari and Nagaoka, 2000), which allows for discussion on manifolds created by a set of probability distributions. From the viewpoint of information geometry, it is known that the set of probability distributions constitutes a Riemannian manifold, which is a non-Euclidean space, and that the parameter space of the probability distributions plays the role of a coordinate system. With this geometric tool, we can reinterpret IMDRE with linear mixtures as a sequential DRE along a specific curve called the m-geodesic. A geodesic here is de-

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

fined as a straight path on a manifold when a certain coordinate system is used. This means that different coordinate systems are equipped with different geodesics. We propose IMDRE with generalized geodesics called α-geodesics. We call this extension as Generalized IMDRE (GIMDRE). To realize GIMDRE, sampling along the α-geodesics connecting the source and target distributions is required. However, since the classes of the two probability distributions and their parameters are unknown, direct sampling is a non-trivial problem. To address this issue, by utilizing the Monte Carlo method with importance sampling (Elvira and Martino, 2014; Tokdar and Kass, 2010), we show that sampling according to α-geodesics can be achieved by giving the density ratio of the source and target distributions. The next problem here is that the density ratio estimation by GIMDRE relies on sampling along the α-geodesic, while the sampling depends on the density ratio. To resolve this interdependent deadlock, we develop an algorithm that alternately estimates the density ratio and updates the importance weights for sampling. We design numerical experiments to demonstrate the usefulness of the proposed method and its behavior. To summarize, our contributions are as follows.

- We geometrically reinterpret IMDRE through the lens of information geometry, and we show that IMDRE with linear mixtures can be viewed as a sequential DRE along a special curve called the m-geodesic. Furthermore, asymptotic analysis shows that certain geodesics appear in the evaluation of the DRE (Section 3, Theorems 1, 2 and Corollary 1).
- We consider extending IMDRE with generalized geodesics called α-geodesics. We refer to this extension as GIMDRE, and we show that the algorithm to obtain GIMDRE can be achieved by alternating algorithm of density ratio estimation and updating the importance weighting for sampling based on the Monte Carlo procedure (Section 4 and Figure 1).
- We design numerical experiments, including permutation test, an important task in statistics, to investigate the behavior of our algorithm and demonstrate its effectiveness (Section 5, Figures 2, 3, 4, and Tables 1, 2, 3).

Our study provides insights into the connections between statistical procedures and differential geometry. We show that the use of α-geodesics for the estimation of density ratios improves upon both the traditional IMDRE and the direct estimation. See the appendix for technical details such as proofs of theorems and experimental setups, as well as additional results including discussion on effective sample size, analysis of variance, and additional numerical experiments.

2 Background and Preliminary

In this section, we first provide the background and preliminary knowledge that is required throughout this study. See Appendix D for related literature.

2.1 Problem Setting

Let $\mathcal{X} \subset \mathbb{R}^d$ be the d-dimensional data domain with $d \in \mathbb{N}$. Suppose that we are given independent and identically distributed (i.i.d.) samples $\{\boldsymbol{x}_i^s\}_{i=1}^{n_s}$ and $\{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$ of sizes $n_s, n_t \in \mathbb{N}$, from $p_s(\boldsymbol{x})$ and $p_t(\boldsymbol{x})$, the source and target distributions, respectively. The goal is to estimate the density ratio

$$r(\boldsymbol{x}) \coloneqq p_s(\boldsymbol{x})/p_t(\boldsymbol{x}), \tag{1}$$

where we assume that $p_t(\boldsymbol{x})$ is strictly positive over the domain $\mathcal{X}$. Let $\hat{r}(\boldsymbol{x})$ be the estimator of $r(\boldsymbol{x})$, and what we want to achieve is to get $\hat{r}(\boldsymbol{x})$ and $r(\boldsymbol{x})$ as close as possible.

2.2 Incremental Mixture DRE

The IMDRE family includes Telescoping DRE (Rhodes et al., 2020) and ∞-DRE (Choi et al., 2022), and its formulation is based on a density ratio chain with arbitrary bridge distributions as follows.

$$r(\boldsymbol{x}) = \frac{p_s(\boldsymbol{x})}{p_t(\boldsymbol{x})} = \frac{p_s(\boldsymbol{x})}{p_1(\boldsymbol{x})}\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} \cdots \frac{p_{m-1}(\boldsymbol{x})}{p_m(\boldsymbol{x})}\frac{p_m(\boldsymbol{x})}{p_t(\boldsymbol{x})}, \tag{2}$$

where $p_1(\boldsymbol{x}), \ldots, p_m(\boldsymbol{x})$ are m bridge distributions. They consider the linear mixtures to gradually transport sample $\{\boldsymbol{x}_i^s\}_{i=1}^{n_s}$ from the source distribution $p_s(\boldsymbol{x})$ to sample $\{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$ from the target distribution $p_t(\boldsymbol{x})$ as $\boldsymbol{x}_i^k = (1-\lambda)\boldsymbol{x}_i^s + \lambda \boldsymbol{x}_i^t$ with $\lambda \in (0,1)$. Note that Rhodes et al. (2020) use $\lambda = \sqrt{1-a_k^2}$ where a_k form an increasing sequence from 0 to 1. For each successive density ratio, once the samples are taken, standard approaches can be used to estimate the density ratio, such as kernel logistic regression. Their experimental results report that such a framework yields good estimates.

3 Reinterpretation of IMDRE

To begin with, we consider the statistical manifolds constructed by a set of probability distributions in order to investigate the geometric behavior of IMDRE. Let $\mathcal{M} = \{p(\cdot;\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta\}$ be a statistical model parametrized by $\boldsymbol{\theta} \in \Theta$, where $\Theta \subset \mathbb{R}^b$ is a parameter space. Let $\boldsymbol{G} = (g_{ij})$ be the Fisher information matrix defined as $g_{ij} = \mathbb{E}[\partial_i \ell(\boldsymbol{\theta}) \partial_j \ell(\boldsymbol{\theta})]$, where

$\ell(\boldsymbol{\theta}) = \ln p(\boldsymbol{x};\boldsymbol{\theta})$ and $\partial_i = \partial/\partial\theta_i$. Here, it is known that $\mathcal{M}$ is a Riemannian manifold with Θ as the coordinate system and $\boldsymbol{G}$ as coefficients of the metric. A coordinate system is intuitively like location information for determining a point on a manifold, which can be understood from the fact that determining the parameters determines the corresponding probability distribution. The metric enables us to measure distances and angles on manifolds and to determine connections, and its requirements are to be symmetric, positive definite, and non-degenerate. Under appropriate regularity conditions, the Fisher information matrix satisfies these requirements. We avoid using geometry notations more than necessary and require only knowledge of basic covariant tensors. However, additional knowledge of differential geometry is included in Appendix B.2 to make it self-contained. For more details, refer to textbooks on differential geometry (Guggenheimer, 2012; Kreyszig, 2013) and Riemannian geometry (Gallot et al., 1990; Klingenberg, 1995; Lee, 2018; Sakai, 1996). Also see Appendix B.1 for proofs of the theoretical results in this section.

Consider a smooth curve $\boldsymbol{\theta}(\lambda)\colon [0,1] \to \Theta$ in the parameter space Θ with $\lambda \in [0,1]$, and a curve $\gamma(\lambda)\colon [0,1] \to \mathcal{M}$ on the statistical manifold $\mathcal{M}$ is defined as

$$\gamma(\lambda) \coloneqq p(\boldsymbol{x};\boldsymbol{\theta}(\lambda)), \quad \forall \lambda \in [0,1]. \tag{3}$$

The velocity of the curve $\gamma(\lambda)$ is given by $\dot{\gamma}(\lambda) = (d/d\lambda)p(\boldsymbol{x};\boldsymbol{\theta}(\lambda))$. A curve γ is called $\nabla^{(\alpha)}$-autoparallel if $\dot{\gamma}(\lambda)$ is parallel transported along $\gamma(\lambda)$, that is, the acceleration with respect to the $\nabla^{(\alpha)}$-connection vanishes

$$\nabla^{(\alpha)}_{\dot{\gamma}(\lambda)}\dot{\gamma}(\lambda) = 0, \quad \forall \lambda \in [0,1]. \tag{4}$$

Here, let $\mathcal{X}(\mathcal{M})$ be a set of vector fields on $\mathcal{M}$, and $\nabla^{(\alpha)}$-connection is defined as

$$\nabla^{(\alpha)}_X \coloneqq \frac{1+\alpha}{2}\nabla^{(1)}_X + \frac{1-\alpha}{2}\nabla^{(-1)}_X, \quad \forall X \in \mathcal{X}(\mathcal{M}), \tag{5}$$

where $\nabla^{(1)}$ and $\nabla^{(-1)}$ are operators satisfying

$$g(\nabla^{(1)}_{\partial_i}\partial_j, \partial_k) = \mathbb{E}\left[(\partial_i\partial_j\ell)(\partial_k\ell)\right], \tag{6}$$

$$g(\nabla^{(-1)}_{\partial_i}\partial_j, \partial_k) = \mathbb{E}\left[(\partial_i\partial_j\ell + \partial_i\ell\partial_j\ell)(\partial_k\ell)\right], \tag{7}$$

and 2-covariant tensor $g(X,Y)$ is a Riemannian metric on $\mathcal{M}$ as

$$g(X,Y) = \sum_{i,j=1}^{b} g_{ij}v_iw_j, \quad X = \sum_{i=1}^{b} v_i\partial_i,\ Y = \sum_{i=1}^{n} w_i\partial_i. \tag{8}$$

The operation $\nabla^{(\cdot)}\colon \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M})$ is called a linear connection, and see Appendix B.2 for the details. If the curve in Eq. (3) is $\nabla^{(\alpha)}$-autoparallel at some $\alpha \in \mathbb{R}$, it is called α-geodesic. The explicit form of α-geodesics connecting two probability distributions is given as follows.

Definition 1 (α-geodesics (Amari, 2016))**.** *Let $p, q \in \mathcal{M}$ be two probability distributions. The α-geodesic $\gamma^{(\alpha)}(\lambda)\colon [0,1] \to \mathcal{M}$ connecting $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ is defined as*

$$\gamma^{(\alpha)}(\lambda) = \begin{cases} \left\{(1-\lambda)p(\boldsymbol{x})^{\frac{1-\alpha}{2}} + \lambda q(\boldsymbol{x})^{\frac{1-\alpha}{2}}\right\}^{\frac{2}{1-\alpha}}, & (\alpha \neq 1), \\ \exp\left\{(1-\lambda)\ln p(\boldsymbol{x}) + \lambda \ln q(\boldsymbol{x})\right\}, & (\alpha = 1). \end{cases} \tag{9}$$

ignoring the normalization factor.

Note that the α-geodesic is a positive measure on the space, and the dependence on $\boldsymbol{x}$ is not explicitly written. For any λ, it is a positive measure, but it is in general not a probability measure, as it is not normalized as written. However, as stated later, we utilize the self-normalized importance sampling in our algorithm and we can ignore this normalization factor. Given an $\alpha \in \mathbb{R}$, the α-geodesics are known to be the minimizer of the following α-divergence $D_\alpha[p\|q]$, that is, for a fixed $\lambda \in [0,1]$, $\gamma^{(\alpha)}(\lambda) = \arg\min_{\pi\in\mathcal{M}}((1-\lambda)D_\alpha[p\|\pi] + \lambda D_\alpha[q\|\pi])$.

Definition 2 (α-divergence (Amari, 2009))**.** *Let $\alpha \in \mathbb{R}$. For two probability distributions p and q, the α-divergence $D_\alpha\colon \mathcal{M} \times \mathcal{M} \to [0, +\infty)$ is defined as*

$$D_\alpha[p\|q] = \frac{1}{\alpha(\alpha-1)}\left(1 - \int p(\boldsymbol{x})^\alpha q(\boldsymbol{x})^{1-\alpha}d\boldsymbol{x}\right). \tag{10}$$

Here, consider the following IMDRE with linear mixtures as bridge distributions.

$$\begin{aligned} r(\boldsymbol{x}) &= \frac{p_s(\boldsymbol{x})}{p_t(\boldsymbol{x})} \\ &= \frac{p_s(\boldsymbol{x})}{(1-\lambda_1)p_s(\boldsymbol{x}) + \lambda_1 p_t(\boldsymbol{x})} \\ &\quad \cdots \frac{(1-\lambda_m)p_s(\boldsymbol{x}) + \lambda_m p_t(\boldsymbol{x})}{p_t(\boldsymbol{x})}, \end{aligned} \tag{11}$$

where $\lambda_1, \ldots, \lambda_m \in (0,1)$ is a non-decreasing sequence. For such an IMDRE, a simple calculation yields the following.

Theorem 1. *The IMDRE given by Eq. (11) can be regarded as a DRE along a geodesic with $\alpha = -1$. Additionally, the curve to which the bridge distributions of this IMDRE belong is the minimizer of the KL-divergence between p_s and p_t.*

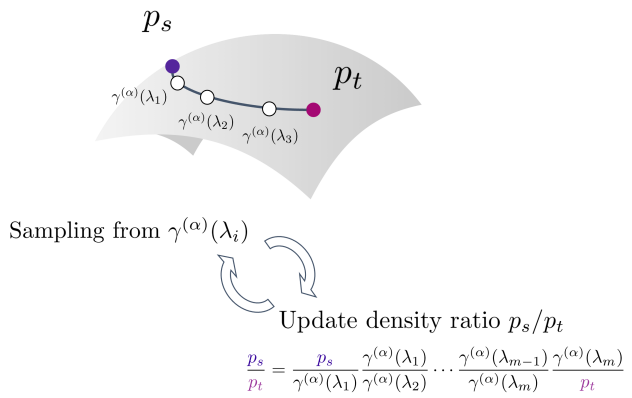

Figure 1: Overview of alternating algorithms for GIMDRE along α-geodesics. $\gamma^{(\alpha)}(\lambda_1), \ldots, \gamma^{(\alpha)}(\lambda_m)$ are m bridge distributions on the α-geodesic parameterized by $\lambda \in [0, 1]$, respectively.

Recall that π^* is a minimizer of KL-divergence $D_{\mathrm{KL}} \colon \mathcal{M} \times \mathcal{M} \to [0, +\infty)$ with respect to λ, p_s and p_t if it satisfies $\pi^* = \arg\min_{\pi \in \mathcal{M}} (1-\lambda) D_{\mathrm{KL}}[p_s \| \pi] + \lambda D_{\mathrm{KL}}[p_t \| \pi]$. Here, the geodesics of $\alpha = -1$ are specially called m-geodesics (m stands for mixture). Similarly, the geodesics of $\alpha = +1$ are specially called e-geodesics (e stands for exponential).

Now consider the evaluation of the density ratios obtained. Given that the Eq. (1) can be transformed as in $p_s(\boldsymbol{x}) = r(\boldsymbol{x}) p_t(\boldsymbol{x})$, a well-estimated density ratio should minimize

$$D_{\mathrm{KL}}[p_s \| \hat{r} \cdot p_t] \coloneqq \int_{\mathcal{X}} p_s(\boldsymbol{x}) \log \frac{p_s(\boldsymbol{x})}{\hat{r}(\boldsymbol{x}) p_t(\boldsymbol{x})} d\boldsymbol{x} - 1 + \int_{\mathcal{X}} p_t(\boldsymbol{x}) \hat{r}(\boldsymbol{x}) d\boldsymbol{x}, \quad (12)$$

where $D_{\mathrm{KL}}[p \| q]$ is the unnormalized version of KL-divergence between p and q. We can then show the following theorem.

Theorem 2. *The evaluation along the m-geodesics of any density ratio estimator $\hat{r}(\boldsymbol{x})$ is asymptotically evaluated along the α-geodesics with $\alpha = 2$.*

Furthermore, if the parameter α that determines the geodesic is well determined, the divergence that appears can be bounded above by the Lipschitz continuity. In fact, KL-divergence is not Lipschitz continuous since $x \to 0$ and $\ln x \to +\infty$, but the α-divergence in the case of $\alpha = 0$ satisfies this. Utilizing this, we obtain the following results from Sugiyama et al. (2008).

Corollary 1. *Under the appropriate assumptions, the convergence bound of a density ratio estimator $\hat{r}$ evaluated along α-geodesic with $\alpha = 0$ is $O_p(n^{-1/(2+\zeta_n)} + \sqrt{C})$, where ζ_n is a variable depending on the sample size n. and C is some constant.*

These results suggest the following:

- Choosing how to construct the bridge distributions in IMDRE is equivalent to choosing what curve on the statistical manifold to follow for density ratio estimation.
- The Behavior of the density ratio estimator depends on which curve to construct the bridge distributions along. In fact, Theorem 2 and Corollary 1 show that in asymptotic situations geodesics with $\alpha = 2$ and $\alpha = 0$ appear.

These suggestions naturally motivate the development of IMDRE along general α-geodesics.

4 GIMDRE along Geodesics

We now construct IMDRE along general α-geodesics, Generalized IMDRE (GIMDRE). To do so along the geodesics, we propose to utilize the following importance sampling technique (Tokdar and Kass, 2010): consider the expectation

$$I \coloneqq \mathbb{E}_{\gamma^{(\alpha)}(\lambda)(\boldsymbol{x})} [g(\boldsymbol{x})] = \int_{\mathcal{X}} g(\boldsymbol{x}) \gamma^{(\alpha)}(\lambda)(\boldsymbol{x}) d\boldsymbol{x}, \quad (13)$$

of any function $g(\boldsymbol{x})$ of a random variable $\boldsymbol{x} \sim \gamma^{(\alpha)}(\lambda)(\boldsymbol{x})$. Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ be the independent instances of sample size T, and the standard Monte Carlo integration is

$$\hat{I} \coloneqq \frac{1}{T} \sum_{i=1}^{N} g(\boldsymbol{x}_i). \quad (14)$$

From the law of large numbers, $\hat{I}$ converges to I as $T \to \infty$. Now, since it is difficult to sample directly from $\gamma^{(\alpha)}(\lambda)(\boldsymbol{x})$, considering another distribution $\pi(\boldsymbol{x})$, which allows for easy sampling, and we have

$$\begin{aligned} I &= \int_{\mathcal{X}} g(\boldsymbol{x}) \gamma^{(\alpha)}(\lambda)(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_{\mathcal{X}} \frac{\gamma^{(\alpha)}(\lambda)(\boldsymbol{x})}{\pi(\boldsymbol{x})} g(\boldsymbol{x}) \pi(\boldsymbol{x}) d\boldsymbol{x} \\ &= \mathbb{E}_{\pi(\boldsymbol{x})} \left[\frac{\gamma^{(\alpha)}(\lambda)(\boldsymbol{x})}{\pi(\boldsymbol{x})} g(\boldsymbol{x}) \right]. \end{aligned} \quad (15)$$

Then, let $w(\boldsymbol{x}) = \gamma^{(\alpha)}(\lambda)(\boldsymbol{x}) / \pi(\boldsymbol{x})$ and we have

$$\hat{I} = \frac{1}{T} \sum_{i=1}^{T} w(\boldsymbol{x}_i^*) g(\boldsymbol{x}_i^*) = \frac{1}{T} \sum_{i=1}^{T} \frac{\gamma^{(\alpha)}(\lambda)(\boldsymbol{x}_i^*)}{\pi(\boldsymbol{x}_i^*)} g(\boldsymbol{x}_i^*).$$

where $\boldsymbol{x}_i^* \sim \pi(\boldsymbol{x})$. Thus, the original problem falls into the problem of approximating the importance weighting $w(\boldsymbol{x})$. Note that the normalization constant of the

α-geodesic is unknown, and we can consider the self-normalized importance sampling technique. Then, let $\gamma^{(\alpha)}(\lambda)(\boldsymbol{x}) = q(\boldsymbol{x})/(\int_{\mathcal{X}} q(\boldsymbol{x})d\boldsymbol{x}) \propto q(\boldsymbol{x})$ for $q \in \mathcal{M}$, and rewrite Eq. (15) as $I = \int_{\mathcal{X}} g(\boldsymbol{x})\frac{q(\boldsymbol{x})}{\int_{\mathcal{X}} q(x)d\boldsymbol{x}} = \frac{\int_{\mathcal{X}} g(\boldsymbol{x})q(\boldsymbol{x})}{\int_{\mathcal{X}} q(\boldsymbol{x})d\boldsymbol{x}}$. Furthermore, if we assume that we are sampling from $p_s(\boldsymbol{x})$, we have

$$I = \frac{\int_{\mathcal{X}} g(\boldsymbol{x})\frac{q(\boldsymbol{x})}{p_s(\boldsymbol{x})}p_s(\boldsymbol{x})d\boldsymbol{x}}{\int_{\mathcal{X}} \frac{q(\boldsymbol{x})}{p_s(\boldsymbol{x})}p_s(\boldsymbol{x})d\boldsymbol{x}} = \frac{\mathbb{E}_{p_s(\boldsymbol{x})}\left[g(\boldsymbol{x})\frac{q(\boldsymbol{x})}{p_s(\boldsymbol{x})}\right]}{\mathbb{E}_{p_s(\boldsymbol{x})}\left[\frac{q(\boldsymbol{x})}{p_s(\boldsymbol{x})}\right]}. \quad (16)$$

Applying importance sampling to the denominator and numerator of Eq. (16) yields the following estimator.

$$\begin{aligned}\hat{I} &= \frac{\frac{1}{T}\sum_{i=1}^{T} g(\boldsymbol{x}_i^*)w(\boldsymbol{x}_i^*)}{\frac{1}{T}\sum_{i=1}^{T} w(\boldsymbol{x}_i^*)} \\ &= \sum_{i=1}^{T} \frac{w(\boldsymbol{x}_i^*)}{\sum_{j=1}^{T} w(\boldsymbol{x}_j^*)} g(\boldsymbol{x}_i^*) \\ &= \sum_{i=1}^{T} \frac{\frac{\gamma^{(\alpha)}(\lambda)(\boldsymbol{x}_i^*)}{\pi(\boldsymbol{x}_i^*)}}{\sum_{j=1}^{T} \frac{\gamma^{(\alpha)}(\lambda)(\boldsymbol{x}_j^*)}{\pi(\boldsymbol{x}_j^*)}} g(\boldsymbol{x}_i^*). \end{aligned} \quad (17)$$

In particular, suppose that the proxy distribution $\pi(\boldsymbol{x}) = p_s(\boldsymbol{x})$ for importance sampling, sampling from the α-geodesic connecting $p_s(\boldsymbol{x})$ and $p_t(\boldsymbol{x})$ can be accomplished by the following $w(\boldsymbol{x}^*)$.

$$\begin{aligned} w_\lambda(\boldsymbol{x}_i^*) = \frac{\gamma^{(\alpha)}(\lambda)(\boldsymbol{x}_i^*)}{p_s(\boldsymbol{x}_i^*)} &= \left\{1 - \lambda + \lambda\left(\frac{p_t(\boldsymbol{x}_i^*)}{p_s(\boldsymbol{x}_i^*)}\right)^{\frac{1-\alpha}{2}}\right\}^{\frac{2}{1-\alpha}} \\ &= \left\{1 - \lambda + \lambda\left(\frac{1}{r(\boldsymbol{x}_i^*)}\right)^{\frac{1-\alpha}{2}}\right\}^{\frac{2}{1-\alpha}} \end{aligned}$$

The key trick to obtain the density ratio $\gamma^{(\alpha)}(\lambda_i)(\boldsymbol{x})/\gamma^{(\alpha)}(\lambda_j)(\boldsymbol{x})$ is the following cancellation out of the proxy distribution $\pi(\boldsymbol{x}) = p_s(\boldsymbol{x})$ using the ratio of the weighting functions $w_{\lambda_i}(\boldsymbol{x})/w_{\lambda_j}(\boldsymbol{x})$.

$$\frac{w_{\lambda_i}(\boldsymbol{x})}{w_{\lambda_j}(\boldsymbol{x})} = \frac{\gamma^{(\alpha)}(\lambda_i)(\boldsymbol{x})/p_s(\boldsymbol{x})}{\gamma^{(\alpha)}(\lambda_j)(\boldsymbol{x})/p_s(\boldsymbol{x})} = \frac{\gamma^{(\alpha)}(\lambda_i)(\boldsymbol{x})}{\gamma^{(\alpha)}(\lambda_j)(\boldsymbol{x})}. \quad (18)$$

As a result, the ratio of the two distributions on the geodesic can be rewritten in a form that depends only on the density ratio $r(\boldsymbol{x})$. Thus, if a weighting function is obtained for each λ, the density ratios between distributions on the geodesic can be recovered. Note that the case $\pi(\boldsymbol{x}) = p_t(\boldsymbol{x})$ can be computed almost similarly. For example, if the function $g(\boldsymbol{x})$ is a loss function for a probabilistic classifier $f(\boldsymbol{x})$, as in logistic regression, with $y = +1$ for $\boldsymbol{x} \sim p_s$ and $y = -1$ for $\boldsymbol{x} \sim p_t$, the output learned by such a Monte Carlo integral approximation can be used for DRE.

However, for the importance sampling, it is necessary to obtain an importance weighting function $w_\lambda(\boldsymbol{x})$ that depends on the estimate of the density ratio $r(\boldsymbol{x})$. In other words, the density ratio estimator and the weighting function for sampling are interdependent and deadlocked.

To address this problem, we develop the following alternating algorithm.

i) The density ratio estimator $\hat{r}(\boldsymbol{x})$ is estimated from the samples $\{\boldsymbol{x}_i\}_{i=1}^{n_s}$ and $\{\boldsymbol{x}_i\}_{i=1}^{n_t}$.

ii) Obtained $\hat{r}(\boldsymbol{x})$ is used to update the weighting $w_\lambda(\boldsymbol{x})$ for the importance sampling.

iii) Using the updated weighting $w_\lambda(\boldsymbol{x})$, $\hat{r}(\boldsymbol{x})$ is estimated by GIMDRE along the α-geodesic.

iv) Iterate steps ii) and iii) over a predetermined number of times.

Figure 1 shows an overview of the above algorithm. The plug-in estimator that underlies our algorithm can be expected to have different statistical behavior depending on two parameters: α, which determines the shape of the curve, and λ, which determines the position on the curve.

Consistency of GIMDRE The initial step is a known and consistent estimator of the density ratio in the statistical sense of convergence in probability. Now, each iteration is then a transformation of the data that depends on the density ratio. Then this consistent density ratio estimator is applied to this transformed data. Since the transformation is smooth, and the initial estimator is consistent, the resulting estimator of the density ratio after that step remains consistent via traditional Taylor series expansion of the smooth transformation. Hence, we have another consistent density ratio estimator after that first iteration. Each iteration proceeds in the same way, starting with a consistent estimator and performing a smooth transformation. After each iteration, we have a consistent estimator of the density ratio after any iteration, and can stop at any point. This is akin to the idea of one-step M-estimators, which start from a consistent estimator and in that case take one Newton-Raphson step toward solving an estimating equation, which as in our case is a smooth transformation.

4.1 Choice of α

We provide some insight into the question about the choice of α through the following analysis.

Proposition 1. *For the estimator $\hat{I}$, we can see that the variance* $\mathrm{Var}(\hat{I})$ *decreases as $\alpha \to +\infty$.*

Proposition 2. *The efficiency in terms of the effective sample size of GIMDRE improves with $\alpha \to +\infty$.*

Table 1: Evaluation results at different step sizes m with $\alpha = 3$. The evaluation metric is the mean absolute error (MAE) between $\hat{r}$ and r. The means and standard deviations of 10 trials are reported.

$m = 10$	$m = 20$	$m = 30$	$m = 40$	$m = 50$	$m = 70$	$m = 100$
35.41(±3.55)	34.80(±3.09)	34.61(±2.60)	34.32(±2.54)	34.26(±2.47)	34.18(±2.39)	34.13(±2.34)

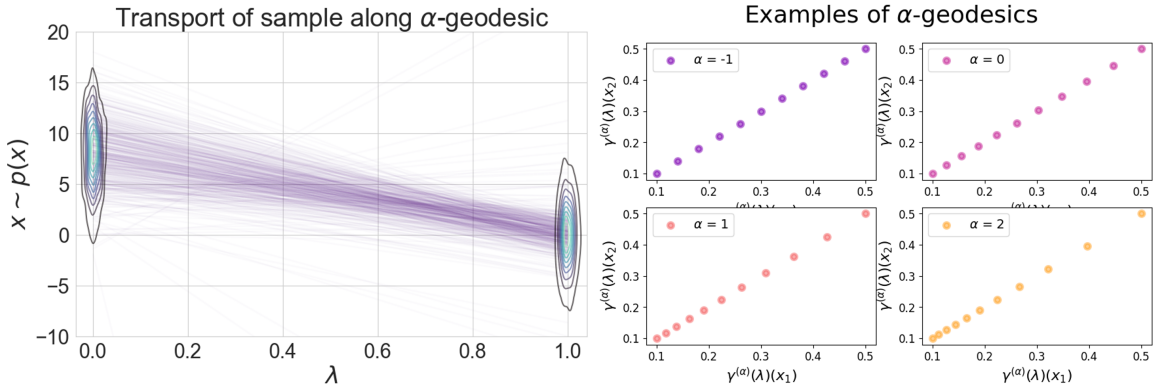

Figure 2: Illustrative examples of GIMDRE convergence. Left panel: an example of sample transport from p_s to p_t along α-geodesic. Right panel: examples of α-geodesics connecting $p_s(\boldsymbol{x})$ and $p_t(\boldsymbol{x})$ in two dimensional case. Here, we assume that $p_s(\boldsymbol{x}) = (0.1, 0.1)$ and $p_t(\boldsymbol{x}) = (0.9, 0.9)$.

These propositions show that using a large α improves the variance and effective sample size of the obtained estimator. See Appendix B.4 for proofs of these propositions.

4.2 Equally Spaced Transitions on Generalized Geodesic

The right panel of Figure 2 visualizes the trajectory approaching from the source point to the target point along several α-geodesics (see Section 5 for more details). In this figure, we have equally spaced transitions of λ from 0 to 1 in increments of 0.1. As another way of selecting λ, we can consider selecting equally spaced transitions on the geodesic. That is, we can choose $\lambda_1, \ldots, \lambda_m$ such that the arc lengths $\|\gamma^{(\alpha)}(\lambda_i)\|, \ldots, \|\gamma^{(\alpha)}(\lambda_m)\|$ of the α-geodesic at time $\lambda_1, \ldots, \lambda_m$ are all equal. Let $\gamma^{(\alpha)}(t)$ be the α-geodesic. The length of this curve connecting $\gamma^{(\alpha)}(0) = p_s$ and $\gamma^{(\alpha)}(1) = p_t$ is defined as

$$\|\gamma^{(\alpha)}\| := \int_0^1 \sqrt{\sum_{i,j} g_{ij} \frac{d\gamma^{(\alpha)i}(t)}{dt} \frac{d\gamma^{(\alpha)j}(t)}{dt}}, \tag{19}$$

where $\gamma^{(\alpha)i}$ is the i-th component of $\gamma^{(\alpha)}$. Then, we can see that equally spaced transitions $\lambda_1, \ldots, \lambda_m$ on the α-geodesic satisfy

$$\int_{\lambda_{k-1}}^{\lambda_k} \sqrt{\sum_{i,j} g_{ij} \frac{d\gamma^{(\alpha)i}(t)}{dt} \frac{d\gamma^{(\alpha)j}(t)}{dt}} = \frac{\|\gamma^{(\alpha)}\|}{m}, \tag{20}$$

for $1 \leq k \leq m$. As can be seen from the right panel of Figure 2, there is a large jump in the approaching distribution for large α. Therefore, adjusting the transition speed of λ may be useful in practical applications.

5 Numerical Experiments

In this section, we design numerical experiments to investigate the behavior of GIMDRE and report the results. Our series of numerical experiments consists of two parts: i) illustrative examples to clarify GIMDRE behavior, and ii) experiments on the two-sample test, a major application of density ratio in statistics. See Appendix C for more details on the experiments.

5.1 Illustrative Examples

First, we consider illustrative examples to clarify the behavior of GIMDRE. The purpose of the experiments in this section is to verify the following: does the alternating algorithm for GIMDRE converge and give good estimates, and what geodesics are better to estimate along when varying the sample size, the number of dimensions, and the distance between the two distributions. We generate samples $\{\boldsymbol{x}_i^s\}_{i=1}^{n_s}, \{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$ of sample sizes n_s and n_t, respectively, from multivariate Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s), \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ with $\boldsymbol{\mu}_s, \boldsymbol{\mu}_t \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_s, \boldsymbol{\Sigma}_t \in \mathbb{R}^{d \times d}$.

First, for illustrative purposes, we consider the case where $\alpha = 3$ and the samples follow univariate Gaussian

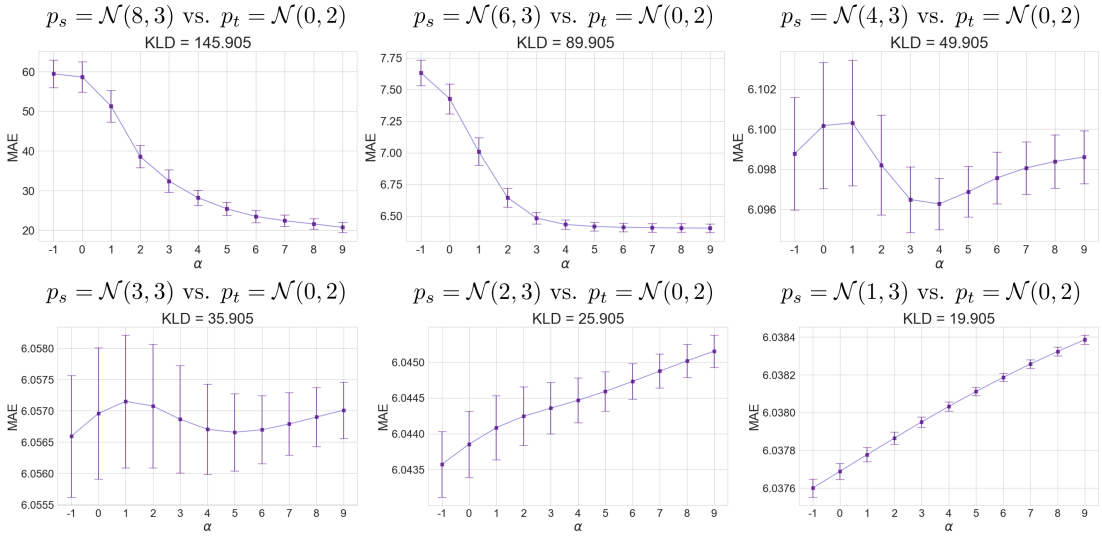

Figure 3: Behavior of GIMDRE according to α when two distributions are brought closer together. KLD is the analytically calculated KL-divergence value

distributions as $p_s = \mathcal{N}(8,3)$ and $p_t = \mathcal{N}(0,2)$. The sample sizes are $n_s = n_t = 500$. The base method for DRE is based on a probabilistic classifier using logistic regression. This setup is a typical example where classical DRE fails because the two distributions are far apart. That is, it is known that classical DRE is difficult unless the means of the two distributions are somewhat close and the variance of the distribution on the denominator side is sufficiently large compared to that on the numerator side. In fact, the mean absolute error (MAE) obtained by ordinary DRE with the base method is 264.07($\pm$116). Note that hereafter we will report the mean and standard deviation of the 10 trials unless otherwise specified. Table 1 shows the results of GIMDRE evaluations when different step sizes are used. These results show that even with a small number of steps, there is a significant improvement compared to the base method, and that increasing the number of steps further improves the mean and standard deviation of the evaluation results. The left panel of Figure 2 also shows an illustration of sample transport using the estimated density ratio $\hat{r}$. This figure shows that $\hat{r}$ is well able to transport from $p_s = \mathcal{N}(8,3)$ to $\mathcal{N}(0,2)$. In addition, the right panel of Figure 2 shows examples of α-geodesics in two dimensional case. In this figure, $\lambda \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ for all α. It can be seen that for $\alpha = -1$, one approaches the target distribution at a uniform rate as the λ increases.

Next, Figure 3 shows an illustration of the behavior of GIMDRE according to α when two distributions are brought together. The parameter $\mu_s \in \{8, 6, 4, 3, 2, 1\}$, and other parameters are fixed as $m = 100$, $n = n_s = n_t = 500$, $p_s = \mathcal{N}(\mu_s, 3)$ and $p_t = \mathcal{N}(0,2)$. To quantify the discrepancy between the two distributions, we analytically compute the KL-divergence. This result shows that the use of a large α consistently suppresses the variance of the estimator. In terms of bias, a large α is effective when the KL-divergence between the two distributions is large, while the effect is reversed as the distributions get closer together. Tables 2 and 3 show the evaluation results for different sample sizes and sample dimensions. It can be seen that the degradation of the estimator with decreasing sample size and increasing dimensionality is reduced by larger α.

From the above evaluation results, we can make the following suggestions: i) if the two distributions are close enough, a good estimator can be obtained with a small α, and as the two distributions move away from each other, the usefulness of using a large α increases. In particular, a large α consistently leads to a variance reduction in the estimator.

5.2 Two-Sample Test based on GIMDRE

In statistics, given two sets of samples, testing whether the probability distributions behind the samples are equivalent is a fundamental task. This problem is referred to as the two-sample test or homogeneity test (Stein, 1945; Conover et al., 1981; Gretton et al., 2006). In our experiments, we utilize the two-sample test based on the permutation test (Efron and Tibshirani, 1993; Sugiyama et al., 2011). Let X^s and X^t be two samples from p_s and p_t. We first estimate the

Table 2: Evaluation results with different α and sample sizes $n = n_s = n_t$.

	$n = 100$	$n = 200$	$n = 300$	$n = 400$	$n = 500$
$\alpha = -1$	180.06($\pm$48.14)	126.48($\pm$25.52)	80.67($\pm$16.16)	64.30($\pm$7.21)	59.88($\pm$4.50)
$\alpha = 3$	76.39($\pm$20.77)	54.99($\pm$14.23)	43.60($\pm$7.15)	36.36($\pm$3.59)	34.13($\pm$2.34)
$\alpha = 7$	69.14($\pm$18.51)	41.26($\pm$12.75)	35.33($\pm$5.62)	26.11($\pm$2.506)	23.27($\pm$1.86)

Table 3: Evaluation results with different α and sample dimensions d.

	$d = 2$	$d = 3$	$d = 4$	$d = 5$
$\alpha = -1$	260.55($\pm$58.23)	589.62($\pm$103.06)	6030.19($\pm$4845.87)	16337.70($\pm$9379.11)
$\alpha = 3$	123.83($\pm$18.23)	320.65($\pm$28.31)	504.50($\pm$47.42)	3664.92($\pm$287.02)
$\alpha = 7$	122.74($\pm$16.17)	331.91($\pm$24.26)	692.35($\pm$44.36)	3792.96($\pm$283.45)

Pearson divergence using X^s and X^t as

$$\hat{D}_{\mathrm{PE}}[X^s \| X^t] \coloneqq \frac{1}{2n_s} \sum_{i=1}^{n_s} \hat{r}(\boldsymbol{x}_i^s) - \frac{1}{n_t} \sum_{i=1}^{n_t} \hat{r}(\boldsymbol{x}_i^t) + \frac{1}{2}. \tag{21}$$

Next, we randomly permute the $|X^s \cup X^t|$ samples, and assign the first n_s samples to a set $\tilde{X}$ and the remaining n_t samples to another set $\tilde{X}'$. Then, we estimate the Pearson divergence again using the randomly shuffled samples $\tilde{X}$ and $\tilde{X}'$, and we obtain an estimate $\hat{D}_{\mathrm{PE}}[\tilde{X} \| \tilde{X}']$. This random shuffling procedure is repeated K times, and the distribution of $\hat{D}_{\mathrm{PE}}[\tilde{X} \| \tilde{X}']$ under the null hypothesis (that is, the two distributions are the same) is constructed. In this experiment, we set $K = 100$. Finally, the p-value is approximated by evaluating the relative ranking of $\hat{D}_{\mathrm{PE}}[X^s \| X^t]$ in the distribution of $\hat{D}_{\mathrm{PE}}[\tilde{X} \| \tilde{X}']$. Suppose that $n_s = n_t$, and let F be the distribution function of $\hat{D}_{\mathrm{PE}}[\tilde{X} \| \tilde{X}']$. Let $\beta \coloneqq \sup\{t \in \mathbb{R} \mid F(t) \leq 1 - \alpha\}$ be the upper 100α-percentile point of F. If $p_s = p_t$, it was shown that $\mathbb{P}(\hat{D}_{\mathrm{PE}}[X^s \| X^t] > \beta) \leq \alpha$. That is, for a given significance level α, the probability that $\hat{D}[X^s \| X^t]$ exceeds β is at most α when $p_s = p_t$. Thus, when the null hypothesis is correct, it will be properly accepted with a specified probability. Let $n = n_s = n_t = 500$, and $p_s = \mathcal{N}(0, 1)$. We consider the setups as $\mu_t \in \{0, 0.5, 1\}$ and $\sigma_t \in \{1, 1.5\}$. The histogram plots in Figure 4 show the distribution of estimated $\hat{D}_{\mathrm{PE}}[\tilde{X} \| \tilde{X}']$, with red crosses indicating $\hat{D}_{PE}[X^s \| X^t]$. Also, the line plots show the transition of the p-values with increasing sample size in the different settings. For all the alphas, when $p_s \neq p_t$, the p-value gets smaller with each increase in sample size, while if the two distributions are equal, it always produces a large p-value. The figure also shows that the GIMDRE with large α behaves conservatively.

6 Conclusion and Discussion

In this paper, we first geometrically reinterpreted the framework of incremental density ratio estimation using a mixture between two distributions. The effectiveness of such a framework for density ratio estimation has been reported in recent years, and it is very useful to study this behavior in detail. Our analysis demonstrated that considering what kind of bridge distributions to create is equivalent to choosing what kind of curve on a statistical manifold. Such a geometrically intuitive interpretation not only reveals properties of the algorithm, but also induces natural generalizations. Next, we considered density ratio estimation along arbitrary generalized geodesics on the manifold. This procedure requires sampling from curves on the manifold, called α-geodesics, which is non-trivial. This is because the two probability distributions are generally not given explicitly. Here we have shown that by using the importance sampling framework, this sampling can be written in a form that depends on the density ratio that we originally wanted to obtain. This is just the state of interdependence, and we have demonstrated that this deadlock can be resolved with a simple alternating procedure. We have shown that this strategy works through illustrative examples and experiments on hypothesis testing, an important task in statistics. Limitlations and broader impacts of this study are discussed in Appendix A. We believe that this study provides useful insights from a geometric perspective into a fundamental task in statistics.

References

Amari, S.-I. (2009). α-divergence is unique, belonging to both f-divergence and bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931.

Amari, S.-i. (2016). *Information geometry and its applications*, volume 194. Springer.

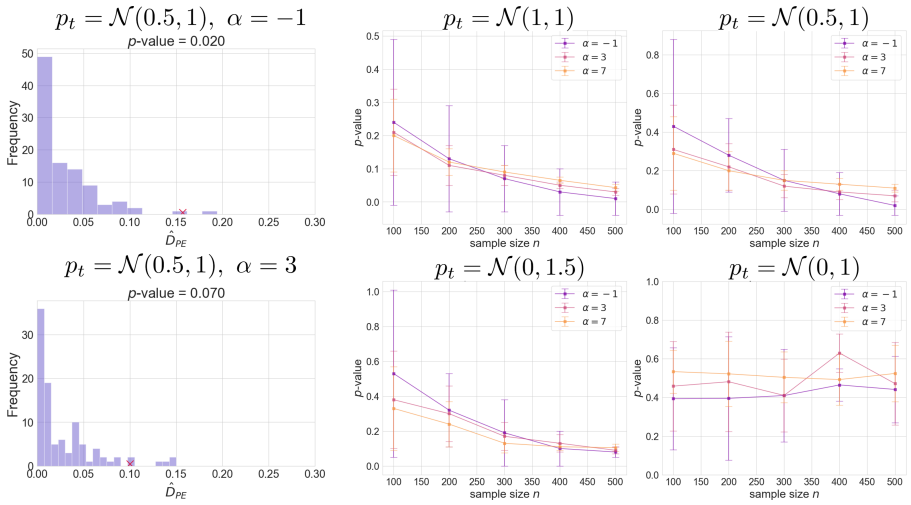

Figure 4: Experimental results of two-sample test. The histogram plots show the distribution of estimated $\hat{D}_{\mathrm{PE}}[\tilde{X}\|\tilde{X}']$, with red crosses indicating $\hat{D}_{PE}[X^s\|X^t]$. The line plots show the transition of the p-values with increasing sample size in the different settings.

Amari, S.-i. and Nagaoka, H. (2000). *Methods of information geometry*, volume 191. American Mathematical Soc.

Azmandian, F., Dy, J. G., Aslam, J. A., and Kaeli, D. R. (2012a). Local kernel density ratio-based feature selection for outlier detection. In *Asian Conference on Machine Learning*, pages 49–64. PMLR.

Azmandian, F., Yilmazer, A., Dy, J. G., Aslam, J. A., and Kaeli, D. R. (2012b). Gpu-accelerated feature selection for outlier detection using the local kernel density ratio. In *2012 IEEE 12th International Conference on Data Mining*, pages 51–60. IEEE.

Braga, I. (2014). A constructive density-ratio approach to mutual information estimation: experiments in feature selection. *Journal of Information and Data Management*, 5(1):134–134.

Byrd, J. and Lipton, Z. (2019). What is the effect of importance weighting in deep learning? In *International conference on machine learning*, pages 872–881. PMLR.

Chen, X., Monfort, M., Liu, A., and Ziebart, B. D. (2016). Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pages 1270–1279. PMLR.

Chen, Y.-T., Fang, W.-H., Lee, C.-Y., and Cheng, K.-W. (2015). Abnormal detection in crowded scenes via kernel based direct density ratio estimation. In *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, pages 230–234. IEEE.

Cheng, K. F. and Chu, C.-K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4):583–604.

Choi, K., Liao, M., and Ermon, S. (2021). Featurized density ratio estimation. In *Uncertainty in Artificial Intelligence*, pages 172–182. PMLR.

Choi, K., Meng, C., Song, Y., and Ermon, S. (2022). Density ratio estimation via infinitesimal classification. In *International Conference on Artificial Intelligence and Statistics*, pages 2552–2573. PMLR.

Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23(4):351–361.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103.

Efron, B. and Tibshirani, R. J. (1993). Permutation tests. *An introduction to the bootstrap*, pages 202–219.

Elvira, V. and Martino, L. (2014). Advances in importance sampling. *Wiley StatsRef: Statistics Reference Online*, pages 1–14.

Gallot, S., Hulin, D., Lafontaine, J., et al. (1990). *Riemannian geometry*, volume 2. Springer.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2006). A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19.

Guggenheimer, H. W. (2012). *Differential geometry*. Courier Corporation.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array

programming with numpy. *Nature*, 585(7825):357–362.

Heck, D. W. (2019). A caveat on the savage–dickey density ratio: The case of computing bayes factors for regression parameters. *British Journal of Mathematical and Statistical Psychology*, 72(2):316–333.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., and Kanamori, T. (2011). Statistical outlier detection using direct density ratio estimation. *Knowledge and information systems*, 26:309–336.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. (2006). Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19.

Hushchyn, M. and Ustyuzhanin, A. (2021). Generalization of change-point detection in time series data based on direct density ratio estimation. *Journal of Computational Science*, 53:101385.

Kanamori, T., Hido, S., and Sugiyama, M. (2008). Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection. *Advances in neural information processing systems*, 21.

Kanamori, T., Suzuki, T., and Sugiyama, M. (2010). Theoretical analysis of density ratio estimation. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 93(4):787–798.

Kanamori, T., Suzuki, T., and Sugiyama, M. (2011). f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2):708–720.

Kawahara, Y. and Sugiyama, M. (2009). Change-point detection in time-series data by direct density-ratio estimation. In *Proceedings of the 2009 SIAM international conference on data mining*, pages 389–400. SIAM.

Keziou, A. and Leoni-Aubin, S. (2005). Test of homogeneity in semiparametric two-sample density ratio models. *Comptes Rendus Mathématique*, 340(12):905–910.

Keziou, A. and Leoni-Aubin, S. (2008). On empirical likelihood for semiparametric two-sample density ratio models. *Journal of Statistical Planning and Inference*, 138(4):915–928.

Kimura, M. and Hino, H. (2022). Information geometrically generalized covariate shift adaptation. *Neural Computation*, 34(9):1944–1977.

Kimura, M. and Hino, H. (2024). A short survey on importance weighting for machine learning. *Transactions on Machine Learning Research*. Survey Certification.

Klingenberg, W. (1995). *Riemannian geometry*, volume 1. Walter de Gruyter.

Kreyszig, E. (2013). *Differential geometry*. Courier Corporation.

LaFond, R. and Watts, R. L. (2008). The information role of conservatism. *The accounting review*, 83(2):447–478.

Lee, J. M. (2018). *Introduction to Riemannian manifolds*, volume 2. Springer.

Lee, S. and Lee, D. K. (2018). What is the proper way to apply the multiple comparison test? *Korean journal of anesthesiology*, 71(5):353.

Li, K., Gao, X., Fu, S., Diao, X., Ye, P., Xue, B., Yu, J., and Huang, Z. (2022). Robust outlier detection based on the changing rate of directed density ratio. *Expert Systems with Applications*, 207:117988.

Liu, A. and Ziebart, B. (2014). Robust classification under sample selection bias. *Advances in neural information processing systems*, 27.

Liu, S., Yamada, M., Collier, N., and Sugiyama, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83.

Lu, Y.-h., Liu, A.-y., Jiang, M.-j., and Jiang, T. (2020). A new two-part test based on density ratio model for zero-inflated continuous distributions. *Applied Mathematics-A Journal of Chinese Universities*, 35(2):203–219.

Martin, J. I. S., Mazuelas, S., and Liu, A. (2023). Double-weighting for covariate shift adaptation. In *International Conference on Machine Learning*, pages 30439–30457. PMLR.

Matsushita, Y., Otsu, T., and Takahata, K. (2023). Estimating density ratio of marginals to joint: Applications to causal inference. *Journal of Business & Economic Statistics*, 41(2):467–481.

Nguyen, X., Wainwright, M. J., and Jordan, M. (2007). Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. *Advances in neural information processing systems*, 20.

Oh, M.-S., Choi, J., and Park, E. S. (2016). Bayesian variable selection in quantile regression using the savage–dickey density ratio. *Journal of the Korean Statistical Society*, 45(3):466–476.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Reichenheim, M. E. and Coutinho, E. S. (2010). Measures and models for causal inference in cross-sectional studies: arguments for the appropriateness

of the prevalence odds ratio and related logistic regression. *BMC medical research methodology*, 10:1–12.

Rhodes, B., Xu, K., and Gutmann, M. U. (2020). Telescoping density-ratio estimation. *Advances in neural information processing systems*, 33:4905–4916.

Sakai, T. (1996). *Riemannian geometry*, volume 149. American Mathematical Soc.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics*, 16(3):243–258.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1):187–205.

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007a). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5).

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., and Kawanabe, M. (2007b). Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20.

Sugiyama, M., Suzuki, T., Itoh, Y., Kanamori, T., and Kimura, M. (2011). Least-squares two-sample test. *Neural networks*, 24(7):735–751.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012a). *Density ratio estimation in machine learning*. Cambridge University Press.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012b). Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64:1009–1044.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., Von Bünau, P., and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746.

Suzuki, T. and Sugiyama, M. (2010). Sufficient dimension reduction via squared-loss mutual information estimation. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 804–811. JMLR Workshop and Conference Proceedings.

Suzuki, T., Sugiyama, M., Sese, J., and Kanamori, T. (2008). Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, pages 5–20. PMLR.

Tokdar, S. T. and Kass, R. E. (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., and Sugiyama, M. (2013). Relative density-ratio estimation for robust distribution comparison. *Neural computation*, 25(5):1324–1370.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

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

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

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

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

A Limitations and Broader Impact

In this section, we discuss about both the limitations and the broader impacts of our study.

A.1 Limitations

One limitation is that our algorithm does not include a neural network. Of course, on the other hand, this is a promising extension direction. Modification of our algorithm to take advantage of the expressive power of deep learning could lead to further advances in this research. Deep learning models have demonstrated remarkable capabilities in capturing complex patterns and relationships in data, offering a prospect for enhancing the performance and scope of our algorithm. On the other hand, an existing study reported that the effect of importance-weighted learning by deep learning decays with the number of steps (Byrd and Lipton, 2019), so it is not clear whether our method based on importance sampling is straightforwardly extendable or not. For the application of neural networks to our research, we need to take into account such negative results reported by existing studies, and it seems that we need to devote a large space to this.

A.2 Broader Impact

As highlighted in the Introduction section, the density ratio stands out as a fundamental tool in statistics, playing a pivotal role in various fields. Therefore, while our algorithm itself does not directly produce social impacts, its downstream applications may, depending on how it is used. One example is the observation in numerical experiments, where the choice of parameters leads to conservative hypothesis tests. How sensitive a hypothesis test should be depends on whether it is used in a life-threatening setting, such as the medical field, or a place where challenging decisions are allowed, such as marketing (Dudoit et al., 2003; LaFond and Watts, 2008; Lee and Lee, 2018; Storey et al., 2004). Our numerical experiments suggest that the choice of parameters can control how sensitive a hypothetical test is constructed, which could have a sufficient impact on society.

B Technical Details

In this section, we provide supplementary technical details, including proofs of the theorems and additional discussion of the ESS.

B.1 Proofs of Theoretical Results

Proof of Theorem 1. Recall that α-geodesics between p and q are given as

$$\gamma^{(\alpha)}(\lambda) = \left\{(1-\lambda)p(\boldsymbol{x})^{\frac{1-\alpha}{2}} + \lambda q(\boldsymbol{x})^{\frac{1-\alpha}{2}}\right\}^{\frac{2}{1-\alpha}}. \tag{22}$$

Substituting $\alpha = -1$, we see that this is an ordinary weighted average. Also, α-divergence becomes KL-divergence with $\alpha = -1$, thus obtaining the proof. □

Proof of Theorem 2. Considering the following Taylor expansion.

$$\begin{aligned}
\log \frac{p_s(x)}{\hat{r}(x)p_t(x)} &= \log \frac{r(x)}{\hat{r}(x)} \\
&= -\log \frac{\hat{r}(x)}{r(x)} \\
&= -\left(\frac{\hat{r}(x)}{r(x)} - 1\right) + \frac{1}{2}\left(\frac{\hat{r}(x)}{r(x)} - 1\right)^2 + O_p\left(\left|\frac{\hat{r}(x)}{r(x)} - 1\right|^3\right).
\end{aligned} \tag{23}$$

Let

$$J_1 = \int_{\mathcal{X}} p_s(x) \log \frac{p_s(x)}{\hat{r}(x)p_t(x)} dx - 1, \tag{24}$$

$$J_2 = \int_{\mathcal{X}} p_t(x)\hat{r}(x)dx. \tag{25}$$

Then,

$$
\begin{aligned}
J_1 &= \int_{\mathcal{X}} p(x) \log \frac{p_s(x)}{\hat{r}(x) p_t(x)} dx - 1 \\
&= \int_{\mathcal{X}} \left\{ -\left(\frac{\hat{r}(x)}{r(x)} - 1 \right) + \frac{1}{2} \left(\frac{\hat{r}(x)}{r(x)} - 1 \right)^2 \right\} p_s(x) dx - 1 + O(\|\hat{r}/r - 1\|^3) \\
&= \int_{\mathcal{X}} \left\{ -\frac{\hat{r}(x)}{r(x)} + 1 + \frac{1}{2} \left(\left(\frac{\hat{r}(x)}{r(x)} \right)^2 - 2 \frac{\hat{r}(x)}{r(x)} + 1 \right) \right\} p_s(x) dx - \int_{\mathcal{X}} p_s(x) dx + O(\|\hat{r}/r - 1\|^3) \\
&= \int_{\mathcal{X}} \left\{ -\frac{\hat{r}(x)}{r(x)} + \frac{1}{2} \left(\left(\frac{\hat{r}(x)}{r(x)} \right)^2 - 2 \frac{\hat{r}(x)}{r(x)} + 1 \right) \right\} p_s(x) dx + O(\|\hat{r}/r - 1\|^3) \\
&= \int_{\mathcal{X}} \left\{ -\frac{\hat{r}(x)}{r(x)} + \frac{1}{2} \frac{\hat{r}(x)^2}{r(x)^2} - \frac{\hat{r}(x)}{r(x)} + \frac{1}{2} \right\} p_s(x) dx + O(\|\hat{r}/r - 1\|^3) \\
&= \int_{\mathcal{X}} \left\{ -2 \frac{\hat{r}(x)}{r(x)} + \frac{1}{2} \frac{\hat{r}(x)^2}{r(x)^2} + \frac{1}{2} \right\} p_s(x) dx + O(\|\hat{r}/r - 1\|^3) \\
&= \int_{\mathcal{X}} \left\{ -2\hat{r}(x) + \frac{1}{2} \frac{\hat{r}(x)^2}{r(x)} + \frac{1}{2} r(x) \right\} \frac{1}{r(x)} p_s(x) dx + O(\|\hat{r}/r - 1\|^3) \\
&= \int_{\mathcal{X}} \left\{ -2\hat{r}(x) + \frac{1}{2} \frac{\hat{r}(x)^2 p_t(x)}{p_s(x)} + \frac{1}{2} \frac{p_s(x)}{p_t(x)} \right\} p_t(x) dx + O(\|\hat{r}/r - 1\|^3).
\end{aligned} \tag{26}
$$

Here,

$$
\|\hat{r}/r - 1\| \coloneqq \left(\int p_s(x) |\hat{r}(x)/r(x) - 1|^2 dx \right)^{1/2}. \tag{27}
$$

We have

$$
\begin{aligned}
D_{\mathrm{KL}}[p_s \| \hat{r} \cdot p_t] &= J_1 + J_2 \\
&= \int_{\mathcal{X}} \left\{ -2\hat{r}(x) + \frac{1}{2} \frac{\hat{r}(x)^2 p_t(x)}{p_s(x)} + \frac{1}{2} \frac{p_s(x)}{p_t(x)} \right\} p_t(x) dx + \int_{\mathcal{X}} p_t(x) \hat{r}(x) dx \\
&= \int_{\mathcal{X}} \left\{ -\hat{r}(x) + \frac{1}{2} \frac{\hat{r}(x)^2 p_t(x)}{p_s(x)} + \frac{1}{2} \frac{p_s(x)}{p_t(x)} \right\} p_t(x) dx \\
&= \frac{1}{2} \int_{\mathcal{X}} \left\{ -2\hat{r}(x) + \frac{\hat{r}(x)^2 p_t(x)}{p_s(x)} + \frac{p_s(x)}{p_t(x)} \right\} p_t(x) dx \\
&= \frac{1}{2} \int_{\mathcal{X}} \frac{1}{p_s(x)} \left\{ -2 p_s(x) \cdot \hat{r}(x) p_t(x) + \hat{r}(x)^2 p_t(x)^2 + p_s(x)^2 \right\} dx \\
&= \frac{1}{2} \int_{\mathcal{X}} \frac{(p_s(x) - \hat{r}(x) p_t(x))^2}{p_s(x)} dx + O(\|\hat{r}/r - 1\|^3) \\
&= D_{\mathrm{PE}}[p_s \| \hat{r}(x) p_t(x)] + O(\|\hat{r}/r - 1\|^3).
\end{aligned} \tag{28}
$$

Here, $D_{\mathrm{PE}}[\cdot\|\cdot]$ is the Pearson divergence. Hence, the asymptotic expansion of the unnormalized KL-divergence between $p_s(x)$ and $\hat{r}(x) p_t(x)$ is given as

$$
D_{\mathrm{KL}}[p_s \| \hat{r} \cdot p_t] = D_{\mathrm{PE}}[p_s \| \hat{r} \cdot p_t] + O(n^{-3/2}). \tag{29}
$$

We consider the following α-divergence.

$$
D_\alpha[p_s \| p_t] = \frac{1}{\alpha(\alpha - 1)} \left(1 - \int p_s(x)^\alpha p_t(x)^{1-\alpha} dx \right). \tag{30}
$$

From the simple calculation, we can confirm that Pearson divergence is a special case of α-divergene with $\alpha = 2$. Since $D_{\mathrm{KL}}[p_s \| p_t]$ and $D_\alpha[p_s \| p_t]$ are minimizers of α-geodesics with $\alpha = -1$ and $\alpha = 2$ between p_s and p_t, we have the proof. □

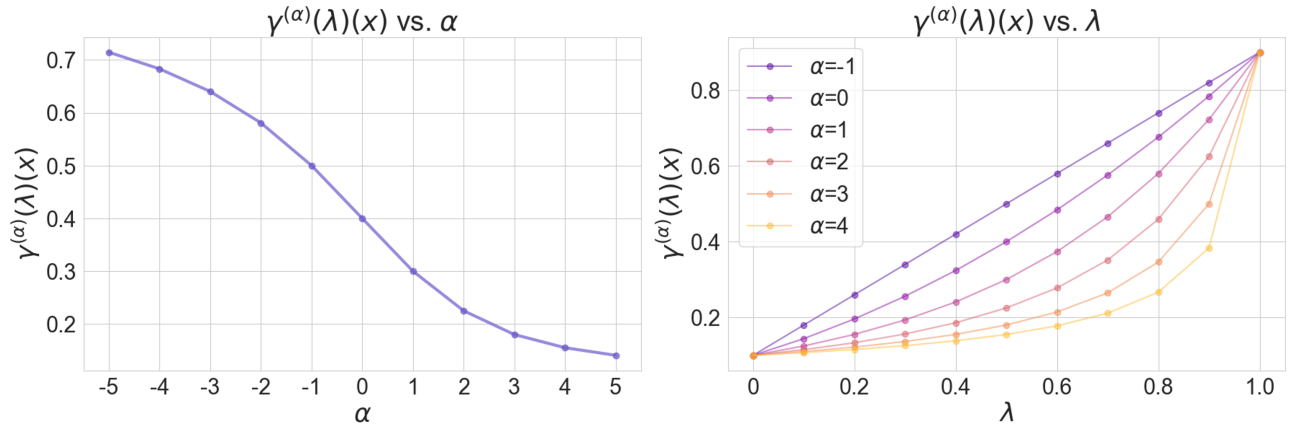

Figure 5: Monotonicity of α-geodesics. The left panel shows that the value of the α-geodesic decreases monotonically as α is increased. Here, $p_s(\boldsymbol{x}) = 0.1$, $p_t(\boldsymbol{x}) = 0.1$, and λ is set to a constant of 0.5. The right panel shows that the α-geodesic approaches p_s at $\lambda \to 0$ and p_t at $\lambda \to 1$ for all α.

Proof of Corollary 1. We assume that two distributions p_s and p_t are mutually absolutely continuous, and satisfy $0 < c_0 \leq dp_s/dp_t \leq c_1$ on the support of p_s and p_t. We also assume that for any function $f\colon \to \mathbb{R}$, there exist ϵ and δ such that $\mathbb{E}_{p_t}[f(x)] \geq \epsilon,\ \|f\|_\infty \leq \delta$. Finally, we assume that for some constants $0 < C < 2$ and M, $\sup_{q \in \mathcal{M}} \ln N(\epsilon, F^M, L_2(q))$, where F is a set of finite linear combinations of integrable functions f with positive coefficients bounded above by M, and $N(\epsilon, F^M, L_2(q))$ is the ϵ-covering number with L_2-distance (Nguyen et al., 2007). From the definition of ζ_n,

$$-\frac{1}{n_s}\sum_{i=1}^{n_2} \ln \hat{r}(\boldsymbol{x}_i^s) \leq -\frac{1}{n_s}\sum_{i=1}^{n_s} \ln \frac{r(\boldsymbol{x}_i^s)}{\frac{1}{n_i^t}\sum_{i=1}^{n_t} r(\boldsymbol{x}_i^t)} + \zeta_n. \tag{31}$$

By convexity of $-\ln x$,

$$-\frac{1}{n_s}\sum_{i=1}^{n_s} \ln \left(\frac{\hat{r}(\boldsymbol{x}_i^s) + r(\boldsymbol{x}_i^s)\frac{1}{n_i^t}\sum_{i=1}^{n_t} r(\boldsymbol{x}_i^t)}{2r(\boldsymbol{x}_i^s)\frac{1}{n_i^t}\sum_{i=1}^{n_t} r(\boldsymbol{x}_i^t)} \right) \leq \frac{\zeta_n}{2}. \tag{32}$$

Let $\xi_1 = \hat{r}(\boldsymbol{x}_i^s)$ and $\xi_2 = 2r(\boldsymbol{x}_i^s)\frac{1}{n_i^t}\sum_{i=1}^{n_t} r(\boldsymbol{x}_i^t)$, and we have

$$\begin{aligned}
(\hat{p}_t - p_t)(\xi_2 - \xi_1) - (\hat{p}_s - p_s)\ln\frac{\xi_2}{\xi_1} - \frac{\zeta_n}{2} &\leq -p_t(\xi_2 - \xi_1) - p_s\left(\ln\frac{\xi_2}{\xi_1}\right) \\
&\leq 2p_s\left(\sqrt{\frac{\xi_2}{\xi_1}} - 1\right) - p_t(\xi_2 - \xi_1) \\
&= p_t\left(2\sqrt{\xi_2\xi_1} - \xi_2 - \xi_1\right).
\end{aligned} \tag{33}$$

Here, the last line is the generalized version of Hellinger distance, and it corresponds to α-divergence with $\alpha = 0$. The rest of the proof follows immediately from Sugiyama et al. (2008). □

B.2 Generalized Geodesics and Information Geometry

In our manuscript, we have tried to deliver to the reader the usefulness of the geometric interpretation of statistical procedures, while eliminating as much as possible the notations of differential geometry. However, in order to make our manuscript more self-contained, in this section we introduce a few selected geometric concepts that may help in understanding our work.

Let g_{ij} be a Riemannian metric, particularly the Fisher information matrix for the statistical manifold. Then the most simple connection on the manifold is defined by the following Christoffel symbols of the first kind.

$$\Gamma_{ij,k} \coloneqq \frac{1}{2}\left(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}\right). \tag{34}$$

The Levi-Civita connection $\nabla^{(0)}$ is defined as

$$g(\nabla^{(0)}_{\partial_i}\partial_j, \partial_k) = \Gamma_{ij,k}. \tag{35}$$

Here, the operation $\nabla\colon \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M})$ on a differentiable manifold $\mathcal{M}$ is called a linear connection if it satisfies

i) $\nabla_X Y$ is $\mathcal{F}(\mathcal{M})$-linear in X,

ii) $\nabla_X Y$ is $\mathbb{R}$-linear in Y, and

iii) ∇ satisfies the Leibniz rule, that is

$$\nabla_X(fY) = (Xf)Y + f\nabla_X Y, \quad \forall f \in \mathcal{F}(\mathcal{M}), \tag{36}$$

where $\mathcal{X}(\mathcal{M})$ is the set of vector fields on $\mathcal{M}$ and $\mathcal{F}(\mathcal{M})$ is the set of all differentiable functions on $\mathcal{M}$. It is worth nothing that the superscript of $\nabla^{(0)}$ denotes a parameter of the connection, and the generalized connection induced by the α-divergence is denoted as $\nabla^{(\alpha)}$. In this case, $\alpha = 0$ corresponds to the Levi-Civita connection. Two other important special cases of the α-connection are the $\nabla^{(1)}$- and $\nabla^{(-1)}$-connections, given as follows.

$$g(\nabla^{(1)}_{\partial_i}\partial_j, \partial_k) = \Gamma^{(1)}_{ij,k} \coloneqq \mathbb{E}\left[(\partial_i\partial_j\ell)(\partial_k\ell)\right], \tag{37}$$

$$g(\nabla^{(-1)}_{\partial_i}\partial_j, \partial_k) = \Gamma^{(-1)}_{ij,k} \coloneqq \mathbb{E}\left[(\partial_i\partial_j\ell + \partial_i\ell\partial_j\ell)(\partial_k\ell)\right]. \tag{38}$$

Note that α-divergence with $\alpha = \pm 1$ is the KL-divergence and its dual. Here, the relationship between α-divergence and α-connection is that under α-connection, the α-geodesic becomes a straight path, and α-divergence becomes its minimizer. The α-connection is given as

$$g(\nabla^{(\alpha)}_{\partial_i}\partial_j, \partial_k) = \Gamma^{(\alpha)}_{ij,k} \coloneqq \mathbb{E}\left[\left(\partial_i\partial_j\ell + \frac{1-\alpha}{2}\partial_i\ell\partial_j\ell\right)\partial_k\ell\right], \tag{39}$$

and

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)}. \tag{40}$$

For example, let us consider the following Gaussian distribution.

$$p(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \tag{41}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \sigma)$. The Fisher information matrix is given by

$$g_{ij} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \tag{42}$$

Then, from the simple calculation, the Christoffel symbols of first and second kind are given as follows.

$$\Gamma_{11,2} = \frac{1}{\sigma^3}, \tag{43}$$

$$\Gamma_{12,1} = -\frac{1}{\sigma^3}, \tag{44}$$

$$\Gamma_{22,2} = -\frac{2}{\sigma^3}, \tag{45}$$

$$\Gamma^1_{ij} = \begin{pmatrix} 0 & -\frac{1}{\sigma} \\ -\frac{1}{\sigma} & 0 \end{pmatrix}, \tag{46}$$

$$\Gamma^2_{ij} = \begin{pmatrix} \frac{1}{2\sigma} & 0 \\ 0 & -\frac{1}{\sigma} \end{pmatrix}, \tag{47}$$

and we can see that the geodesics are solutions of the following ordinary differential equations.

$$\ddot{\mu} - \frac{2}{\sigma}\dot{\mu}\dot{\sigma} = 0, \tag{48}$$

$$\ddot{\sigma} + \frac{1}{2\sigma}(\dot{\mu})^2 - \frac{1}{\sigma}(\dot{\sigma})^2 = 0. \tag{49}$$

We can then have

$$\frac{\ddot{\mu}}{\dot{\mu}} = \frac{2\dot{\sigma}}{\sigma} \Leftrightarrow \dot{\mu} = c\sigma^2, \tag{50}$$

with some constant c. In the case of $c \neq 0$ we have

$$\sigma(s)^2 + (\mu(s) - K)^2 \frac{J}{\sigma^4} = c^2, \tag{51}$$

where $K > 0$ and $J > 0$ are positive constants, and the geodesics are half-ellipses, with $\sigma > 0$. In the case of $c = 0$,

$$\sigma(s) = \sigma(0)e^{\sqrt{J}s}, \tag{52}$$

and the geodesics are vertical half-lines. For the manifold of Gaussian distributions, the Christoffel coefficients of first kind are written as

$$\Gamma^{(\alpha)}_{11,1} = \Gamma^{(\alpha)}_{21,2} = \Gamma^{(\alpha)}_{12,2} = \Gamma^{(\alpha)}_{22,1} = 0, \tag{53}$$

$$\Gamma^{(\alpha)}_{11,2} = \frac{1-\alpha}{\sigma^3}, \tag{54}$$

$$\Gamma^{(\alpha)}_{12,1} = \Gamma^{(\alpha)}_{21,1} = \frac{1+\alpha}{\sigma^3}, \tag{55}$$

$$\Gamma^{(\alpha)}_{22,2} = \frac{2(1+2\alpha)}{\sigma^3}, \tag{56}$$

and the Christoffel symbols of second kind are as follows.

$$\begin{aligned}\Gamma^{1(\alpha)}_{ij} &= g_{11}\Gamma^{(\alpha)}_{ij,1} + g_{12}\Gamma^{(\alpha)}_{ij,2} = \sigma^2\Gamma^{(\alpha)}_{ij,1} \\ &= \sigma^2 \begin{pmatrix} 0 & -\frac{1+\alpha}{\sigma^3} \\ -\frac{1+\alpha}{\sigma^3} & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1+\alpha}{\sigma} \\ -\frac{1+\alpha}{\sigma} & 0 \end{pmatrix}.\end{aligned} \tag{57}$$

The ordinary differential equations for the α-autoparallel curves are given as

$$\ddot{\mu} - \frac{2(1+\alpha)}{\sigma}\dot{\sigma}\dot{\mu} = 0, \tag{58}$$

$$\ddot{\sigma} + \frac{1-\alpha}{2\sigma}\dot{\mu}^2 - \frac{1+2\alpha}{\sigma}\dot{\sigma}^2 = 0. \tag{59}$$

The first equation can be transformed as

$$\begin{aligned}\frac{\ddot{\mu}}{\mu} = 2(1+\alpha)\frac{\dot{\sigma}}{\sigma} &\Leftrightarrow \frac{d}{ds}\ln\dot{\mu} = 2(1+\alpha)\frac{d}{ds}\ln\sigma \\ &\Leftrightarrow \ln\dot{\mu} = 2(1+\alpha)\ln\sigma + c_0 \\ &\Leftrightarrow \dot{\mu} = c\sigma^{2(1+\alpha)},\end{aligned} \tag{60}$$

with constant c. Then, we have

$$\begin{aligned}\ddot{\sigma} + \frac{1-\alpha}{2\sigma}c^2\sigma^{4(1+\alpha)} - \frac{1+2\alpha}{\sigma}\dot{\sigma}^2 = 0 \\ \sigma^{k+1}du + \left(\frac{1-\alpha}{2}c^2\sigma^{4(\alpha+1)+k} - (1+2\alpha)\sigma^k\mu^2\right)d\sigma = 0,\end{aligned} \tag{61}$$

where $u = \dot{\sigma}$. Here, the following must be satisfied.

$$\frac{\partial}{\partial \sigma}\sigma^{k+1}u = \frac{\partial}{\partial \mu}\left(\frac{1-\alpha}{2}c^2\sigma^{4(\alpha+1)+k} - (1+2\alpha)\sigma^k\mu^2\right). \tag{62}$$

By considering this condition, we can determine $k+1 = -(4\alpha+2)$, and we have

$$u\sigma^{-(4\alpha+2)}du + \left(\frac{1-\alpha}{2}c^2\sigma - (1+2\alpha)u^2\sigma^{-(4\alpha+3)}\right)d\sigma = 0. \tag{63}$$

Hence, for a constant E, we can solve for σ as follows.

$$\begin{aligned}
\frac{\mu^2}{\sigma^2(2\alpha+1)} + \frac{1-\alpha}{2}c^2\sigma^2 = E &\Leftrightarrow \left(\frac{\dot{\sigma}}{\sigma^{2\alpha+1}}\right)^2 + \frac{1-\alpha}{2}c^2\sigma^2 = E \\
&\Leftrightarrow \left(\frac{\dot{\sigma}}{\sigma^{2\alpha+1}}\right)^2 = E - \frac{1-\alpha}{2}c^2\sigma^2 \\
&\Leftrightarrow \int \frac{d\sigma}{\sigma^{2\alpha+1}\sqrt{E - \frac{1-\alpha}{2}c^2\sigma^2}} = \pm s + s_0 \\
&\Leftrightarrow \int \frac{d\sigma}{\sigma^{2\alpha+1}\sqrt{C^2 - \sigma^2}} = (\pm s + s_0)\sqrt{\frac{1-\alpha}{2}}c,
\end{aligned} \tag{64}$$

where

$$C = C_\alpha = \frac{2E}{c}\frac{1}{1-\alpha}. \tag{65}$$

Let $t = \sigma^2$ and $v = \sqrt{C^2 - t}$, we have

$$\begin{aligned}
\int \frac{d\sigma}{\sigma^{2\alpha+1}\sqrt{C^2-\sigma^2}} &= \int \frac{dt}{2\sigma^{2(\alpha+1)}\sqrt{C^2-\sigma^2}} \\
&= \int \frac{dt}{2t^{(\alpha+1)}\sqrt{C^2-1}} \\
&= \int \frac{-2vdv}{2t^{\alpha+1}v} \\
&= -\int \frac{dv}{(C^2-v^2)^{\alpha+1}},
\end{aligned} \tag{66}$$

and then,

$$-\int \frac{dv}{(C^2-v^2)^{\alpha+1}} = (\pm s + s_0)\sqrt{\frac{1-\alpha}{2}}c. \tag{67}$$

The μ is given by

$$\mu = c\int \sigma^{2(1+\alpha)}(s)ds. \tag{68}$$

The Case of $\alpha = -1$ In the case of $\alpha = -1$, we have

$$-v - K = (\pm s + s_0)\sqrt{\frac{1-\alpha}{2}}c, \tag{69}$$

with solution

$$\sigma^2(s) = C^2 - \left((\pm s + s_0)\sqrt{\frac{1-\alpha}{2}}c + K\right)^2, \tag{70}$$

for a constant K. Then, we have

$$\mu(s) = cs + \mu(0). \tag{71}$$

The Case of $\alpha = 1/2$ In the case of $\alpha = 1/2$, since

$$\int \frac{dv}{(C^2 - v^2)^{3/2}} = \frac{v}{C^2\sqrt{C^2 - v^2}}, \tag{72}$$

and we have

$$-\frac{v}{C^2\sqrt{C^2 - v^2}} = (\pm s + s_0)\frac{c}{2} + K, \tag{73}$$

with solution

$$\sigma(s) = \frac{C}{\sqrt{1 + C^4\left(\frac{c}{2}(\pm s + S_0) + K\right)^2}}, \tag{74}$$

for a constant K. The, we have

$$\mu(s) = c\int \sigma^3(s)ds. \tag{75}$$

B.2.1 Important properties of α-connection

Uniqueness of α-connection Let ∇ and ∇^* are the linear connection and its dual w.r.t. the metric g. For $\alpha \in \mathbb{R}$, the α-connection is determined uniquely as

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^* + \frac{1-\alpha}{2}\nabla. \tag{76}$$

Relationship between α-connection and certain other connections The α-connection can be written in one of the following equivalent forms:

$$\nabla^{(\alpha)} = (1-\alpha)\nabla^{(0)} + \alpha\nabla^{(1)}, \tag{77}$$

$$\nabla^{(\alpha)} = (1+\alpha)\nabla^{(0)} - \alpha\nabla^{(-1)}, \tag{78}$$

or

$$\nabla^{(\alpha)} = \nabla^{(0)} + \frac{1}{2}\alpha(\nabla^{(1)} - \nabla^{(-1)}). \tag{79}$$

Duality of α-connection $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are dual connections with respect to the metric g.

Skewness of alpha-connection Let $C = \nabla g$ be the $(0,3)$-skewness tensor. Then, we have $\nabla^{(\alpha)}g = -\alpha C$. In addition, we have the following relation.

$$g(\nabla^{(\alpha)}_X Y, Z) = g(\nabla^{(0)}_X Y, Z) + \frac{\alpha}{2}C(X,Y,Z). \tag{80}$$

Curvature of α-connection and symmetry Let R and R^* be curvature tensors with respect to ∇ and ∇^*. Similarly, let $R^{(\alpha)}$ and $R^{(-\alpha)}$ be curvature tensors with respect to α-connection and its dual. Then if $R = R^* = 0$, we have $R^{(\alpha)} = R^{(-\alpha)}$, and the curvature tensor of the α-connection does not necessarily vanish. That is, α-connection is conjugate symmetric for all $\alpha \in \mathbb{R}$.

Moreover, the following relations hold.

$$R^{(\alpha)}(X,Y,Z) = \frac{1+\alpha}{2}R^*(X,Y,Z) + \frac{1-\alpha}{2}R(X,Y,Z), \tag{81}$$

and

$$R^{(\alpha)}(X,Y,X) - R^{(-\alpha)}(X,Y,X) = \alpha(R^*(X,Y,X) - R(X,Y,X)). \tag{82}$$

This means that the α-connection can be derived by using the skewness tensor.

Dually compatibility of α-connection For $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$, we have the following compatibility in dual sense.

$$g(\nabla_Z^{(\alpha)}X, Y) + g(\nabla_Z^{(-\alpha)}Y, X) = Zg(X, Y). \tag{83}$$

Difference tensor and difference between α-connection and its dual Let $K(X, Y) = \nabla_X^* Y - \nabla_X Y$ be the difference $(1, 2)$-tensor. Then, we have

$$\nabla^{(-\alpha)} + \nabla^{(\alpha)} = \alpha(\nabla^* - \nabla) = \alpha K. \tag{84}$$

The expression of curvature tensor being given by

$$R^{(\alpha)}(X, Y, Z) = \frac{1-\alpha^2}{4}(K(Y, K(X, Z)) - K(X, K(Y, Z))). \tag{85}$$

It follows that a necessary condition for all α-connections to have zero curvature tensors is that

$$K(Y(K(X, Z))) = K(X, K(Y, Z)). \tag{86}$$

This can be written in terms of the skewness tensor as

$$C(K(X, Z), Y, W) = C(X, K(Y, Z), W), \quad \forall X, Y, Z, W \in \mathcal{X}(\mathcal{M}). \tag{87}$$

B.3 Approximation of Arc Lengths of Generalized Geodesics for Equally Spaced Transitions

Let $\gamma^{(\alpha)}(t)$ be the α-geodesic, and we consider its Taylor expansion as

$$\gamma^{(\alpha)i}(t) = \gamma^{(\alpha)i}(0) + t\frac{d\gamma^{(\alpha)i}}{dt}\Big|_{t=0} + \frac{t^2}{2}\frac{d^2\gamma^{(\alpha)i}}{dt^2}\Big|_{t=0}. \tag{88}$$

Here, α-geodesics satisfy the following Euler-Lagrange equation.

$$\frac{d^2\gamma^{(\alpha)i}}{dt^2} + \Gamma_{jk}^{(\alpha)i}\frac{d\gamma^{(\alpha)j}}{dt}\frac{d\gamma^{(\alpha)k}}{dt} = 0, \tag{89}$$

where

$$\Gamma_{ij,k}^{(\alpha)} = \mathbb{E}\left[\left(\partial_i\partial_j\ell + \frac{1-\alpha}{2}\partial_i\ell\partial_j\ell\right)\partial_k\ell\right], \tag{90}$$

and $\ell = \ln p(x)$. Then, we can rewrite as

$$\gamma^{(\alpha)i}(t) = \gamma^{(\alpha)i}(0) + t\frac{d\gamma^{(\alpha)i}(t)}{dt}\Big|_{t=0} - \frac{t^2}{2}\left(\Gamma_{jk}^{(\alpha)i}\frac{d\gamma^{(\alpha)j}}{dt}\frac{d\gamma^{(\alpha)k}}{dt}\right)\Big|_{t=0}. \tag{91}$$

The length of this curve connecting $\gamma^{(\alpha)}(0) = p_s$ and $\gamma^{(\alpha)}(1) = p_t$ is defined as

$$\|\gamma^{(\alpha)}\| \coloneqq \int_0^1 \sqrt{\sum_{i,j} g_{ij}\frac{d\gamma^{(\alpha)i}(t)}{dt}\frac{d\gamma^{(\alpha)j}(t)}{dt}}. \tag{92}$$

Finally, we can see that equally spaced transitions $\lambda_1, \ldots, \lambda_m$ on the α-geodesic satisfies

$$\int_{\lambda_{i-1}}^{\lambda_i} \sqrt{\sum_{i,j} g_{ij}\frac{d\gamma^{(\alpha)i}(t)}{dt}\frac{d\gamma^{(\alpha)j}(t)}{dt}} = \int_{\lambda_{j-1}}^{\lambda_j} \sqrt{\sum_{i,j} g_{ij}\frac{d\gamma^{(\alpha)i}(t)}{dt}\frac{d\gamma^{(\alpha)j}(t)}{dt}}, \quad 1 \le i, j \le m. \tag{93}$$

B.4 Additional Discussion on Variance and Effective Sample Size

We provide some insight into the question about the choice of α through the following analysis.

Proposition 3. *For the estimator $\hat{I}$, we can see that the variance $\mathrm{Var}(\hat{I})$ decreases as $\alpha \to +\infty$.*

Proof. The variance of $\hat{I}$ with p_α^λ is calculated as follows.

$$\begin{aligned}
\mathrm{Var}_{\alpha,\lambda}(\hat{I}) &= \int \left(\frac{\gamma^{(\alpha)}(\lambda)(x)}{p_s(x)} g(x) - I \right)^2 p_s(x) dx \\
&= \int \left(\frac{\gamma^{(\alpha)}(\lambda)(x)^2}{p_s(x)^2} g(x)^2 - 2 \frac{\gamma(\alpha)(\lambda)(x)}{p_s(x)} g(x) I + I^2 \right) p_s(x) dx \\
&= \int \frac{\gamma^{(\alpha)}(\lambda)(x)^2}{p_s(x)} g(x)^2 dx - I^2 \\
&= \int \frac{\gamma^{(\alpha)}(\lambda)(x)}{p_s(x)} g(x)^2 \gamma^{(\alpha)}(\lambda)(x) dx - I^2 \\
&= \mathbb{E}_{\gamma^{(\alpha)}(\lambda)} \left[\frac{\gamma^{(\alpha)}(\lambda)(x)}{p_s(x)} g(x)^2 \right] - I^2
\end{aligned} \tag{94}$$

Thus, the variance of the estimator by sampling from the alpha-mixture depends on the expectation of $g(x)$ scaled by the weight $w(x) = \gamma^{(\alpha)}(\lambda)(x)/p_s(x)$. Here, for all α, we have $\frac{\partial}{\partial \alpha} \frac{\gamma^{(\alpha)}(\lambda)(x)}{p_s(x)} \leq 0$, and then we have the proof. □

The above proof is based on the monotonicity of α-geodesics with respect to α. Therefore, we give a few more observations on this important key monotonicity. Figure 5 shows the monotonicity of the α-geodesic. We use $p_s = 0.1$ and $p_t = 0.9$. The left panel shows that the value of the α-geodesic decreases monotonically as α is increased. Here, λ is set to a constant of 0.5. Recall that $\alpha = -1$ is an ordinary weighted average, and we can see that the geodesic value is $(p_s(\boldsymbol{x}) + p_t(\boldsymbol{x}))/2$ at $\alpha = -1$ in the figure. The right panel shows that the α-geodesic approaches p_s at $\lambda \to 0$ and p_t at $\lambda \to 1$ for all α. Further, for larger values of α, the speed at which the α-geodesic approaches p_t is found to be non-uniform. This behavior, in which the geodesic gradually approaches the target distribution from the source distribution, can be expected to affect the stability of the estimator. In addition, Figure 6 shows the examples of α-geodesics connecting $p_s(\boldsymbol{x})$ and $p_t(\boldsymbol{x})$ in two dimensional case. Here, we assume that $p_s(\boldsymbol{x}) = (0.1, 0.1)$ and $p_t(\boldsymbol{x}) = (0.9, 0.9)$. In this figure, $\lambda \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ for all α. It can be seen that for $\alpha = -1$, one approaches the target distribution at a uniform rate as the λ increases, while for larger values of α, there is an acceleration in approaching the target distribution. The above results follow from the monotonicity of α-geodesics with respect to α. Note that, for arbitrary $\lambda \in [0, 1]$, $\lim_{\alpha \to +\infty} w(\boldsymbol{x}) = \min\{1, p_t(\boldsymbol{x})/p_s(\boldsymbol{x})\}$. This can be seen as a truncation of the density ratio by 1 in the case of $\alpha \to +\infty$.

Finally, the choice of the geodesic parameter α is discussed in more detail. The appropriate parameter α for sampling from α-geodesics depends on $g(\boldsymbol{x})$, the target of the Monte Carlo integration. Let $\bar{Z}_\alpha \coloneqq \sum_{i=1}^T w(\boldsymbol{x}^*) \boldsymbol{x}_i / \sum_{j=1}^T w(\boldsymbol{x}_j)$, this variance is given by

$$\mathrm{Var}(\bar{Z}_\alpha) = \mathrm{Var}(\boldsymbol{x}) \frac{\sum_{i=1}^T w(\boldsymbol{x})^2}{\left(\sum_{i=1}^T w(\boldsymbol{x}) \right)^2} = \mathrm{Var}(\boldsymbol{x}) \frac{\sum_{i=1}^T \left\{ 1 - \lambda + \lambda \left(1/\hat{r}(\boldsymbol{x}_i^s) \right)^{\frac{1-\alpha}{2}} \right\}^{\frac{4}{1-\alpha}}}{\left(\sum_{i=1}^T \left\{ 1 - \lambda + \lambda \left(1/\hat{r}(\boldsymbol{x}_i^s) \right)^{\frac{1-\alpha}{2}} \right\}^{\frac{2}{1-\alpha}} \right)^2}. \tag{95}$$

On the other hand, the variance of $\bar{Z} = \frac{1}{T_e} \sum_{i=1}^{T_e} \boldsymbol{x}_i$ is $\mathrm{Var}(\bar{Z}) = \mathrm{Var}(\boldsymbol{x})/T_e$ for some $T_e > 0$. Then, let $\mathrm{Var}(\bar{Z}_\alpha) = \mathrm{Var}(\bar{Z})$ and solve for T_e to obtain the following.

$$\hat{T}_e = \frac{\left(\sum_{i=1}^T w(\boldsymbol{x}) \right)^2}{\sum_{i=1}^T w(\boldsymbol{x})^2} = \frac{\left(\sum_{i=1}^T \left\{ 1 - \lambda + \lambda \left(1/\hat{r}(\boldsymbol{x}_i^s) \right)^{\frac{1-\alpha}{2}} \right\}^{\frac{2}{1-\alpha}} \right)^2}{\sum_{i=1}^T \left\{ 1 - \lambda + \lambda \left(1/\hat{r}(\boldsymbol{x}_i^s) \right)^{\frac{1-\alpha}{2}} \right\}^{\frac{4}{1-\alpha}}}. \tag{96}$$

$\hat{T}_e$ represents the number of data required to obtain the same accuracy as estimation by $\bar{Z}_\alpha$ when estimating the expected value of a sample by $\bar{Z}$. It is called the effective sample size (ESS). ESS holds for $\hat{T}_e \leq T$, and the equality holds when w is a constant. When the value of $\hat{T}_e$ is small, estimation by $\bar{Z}_\alpha$ requires more samples than $\bar{Z}$, meaning that estimation is inefficient. From the monotonicity of the α-geodesic with respect to α, we obtain the following.

Proposition 4. *The estimation efficiency in terms of ESS of GIMDRE improves with $\alpha \to +\infty$.*

Recall that ESS $\hat{T}_e$ is given as

$$\hat{T}_e = \frac{\left(\sum_{i=1}^T w(\boldsymbol{x})\right)^2}{\sum_{i=1}^T w(\boldsymbol{x})^2} = \frac{\left(\sum_{i=1}^T \left\{1-\lambda+\lambda\left(1/\hat{r}(\boldsymbol{x}_i^s)\right)^{\frac{1-\alpha}{2}}\right\}^{\frac{2}{1-\alpha}}\right)^2}{\sum_{i=1}^T \left\{1-\lambda+\lambda\left(1/\hat{r}(\boldsymbol{x}_i^s)\right)^{\frac{1-\alpha}{2}}\right\}^{\frac{4}{1-\alpha}}}. \tag{97}$$

Here, we can rewrite this as follows.

$$\begin{aligned}\hat{T}_e &= \frac{T}{1+\frac{\frac{1}{T}\sum_{i=1}^T\left\{1-\lambda+\lambda\left(1/\hat{r}(\boldsymbol{x}_i^s)\right)^{\frac{1-\alpha}{2}}\right\}^{\frac{4}{1-\alpha}} - \left(\frac{1}{T}\sum_{i=1}^T\left\{1-\lambda+\lambda\left(1/\hat{r}(\boldsymbol{x}_i^s)\right)^{\frac{1-\alpha}{2}}\right\}^{\frac{2}{1-\alpha}}\right)^2}{\left(\frac{1}{T}\sum_{i=1}^T\left\{1-\lambda+\lambda\left(1/\hat{r}(\boldsymbol{x}_i^s)\right)^{\frac{1-\alpha}{2}}\right\}^{\frac{2}{1-\alpha}}\right)^2}} \\ &= \frac{T}{1+V^2},\end{aligned} \tag{98}$$

where

$$V \coloneqq \frac{\frac{1}{T}\sum_{i=1}^T \left\{1-\lambda+\lambda\left(1/\hat{r}(\boldsymbol{x}_i^s)\right)^{\frac{1-\alpha}{2}}\right\}^{\frac{4}{1-\alpha}} - \left(\frac{1}{T}\sum_{i=1}^T \left\{1-\lambda+\lambda\left(1/\hat{r}(\boldsymbol{x}_i^s)\right)^{\frac{1-\alpha}{2}}\right\}^{\frac{2}{1-\alpha}}\right)^2}{\left(\frac{1}{T}\sum_{i=1}^T \left\{1-\lambda+\lambda\left(1/\hat{r}(\boldsymbol{x}_i^s)\right)^{\frac{1-\alpha}{2}}\right\}^{\frac{2}{1-\alpha}}\right)^2}. \tag{99}$$

The value V is called the coefficient of variation, and from the fact that $V^2 \geq 0$, we can see that $\hat{T}_e \leq T$. Figure 7 shows the relationship between ESS and the parameters α and λ. In this figure, $p_s = \mathcal{N}(8,3)$ and $p_t = \mathcal{N}(0,2)$. The left panel of Figure 7 shows the relationship between α and ESS with $\lambda = 0.5$ as a constant. We see that the estimation efficiency in terms of the ESS is improved by using a large α. Furthermore, the right panel of Figure 7 shows that the estimation efficiency deteriorates as one moves away from the source distribution with larger λ. However, it can be observed that this degradation can be reduced by using a large α. Figures 8 and 9 show the results for the two skewed distributions, Log-normal and Power-law, and reveal similar observations. Here, the density functions of these two distributions $p_{\mathrm{LogNormal}}$ and p_{PowerLaw} are defined as follows.

$$\begin{aligned} p_{\mathrm{LogNormal}}(x;\mu,\sigma) &= \frac{1}{x\sigma\sqrt{2\pi}}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}, \quad x\in(0,+\infty),\ \mu\in(-\infty,+\infty),\ \sigma>0, \\ p_{\mathrm{PowerLaw}}(x;a) &= ax^{a-1}, \quad 0\leq 1,\ a>0. \end{aligned}$$

For Log-normal distributions $p_s = \mathrm{LogNormal}(\mu_s, \sigma_s)$ and $\mathrm{LogNormal}(p_t, \sigma_t)$, we use $\mu_s \in \{3,4\}, \sigma_s = 0.5$, $\mu_t = 0$ and $\sigma_t \in \{1.5, 2, 2.5\}$. Also, for Power-law distributions $p_s = \mathrm{PowerLaw}(a_s)$ and $p_t = \mathrm{PowerLaw}(a_t)$, we use $a_s = 3$ and $a_t \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$.

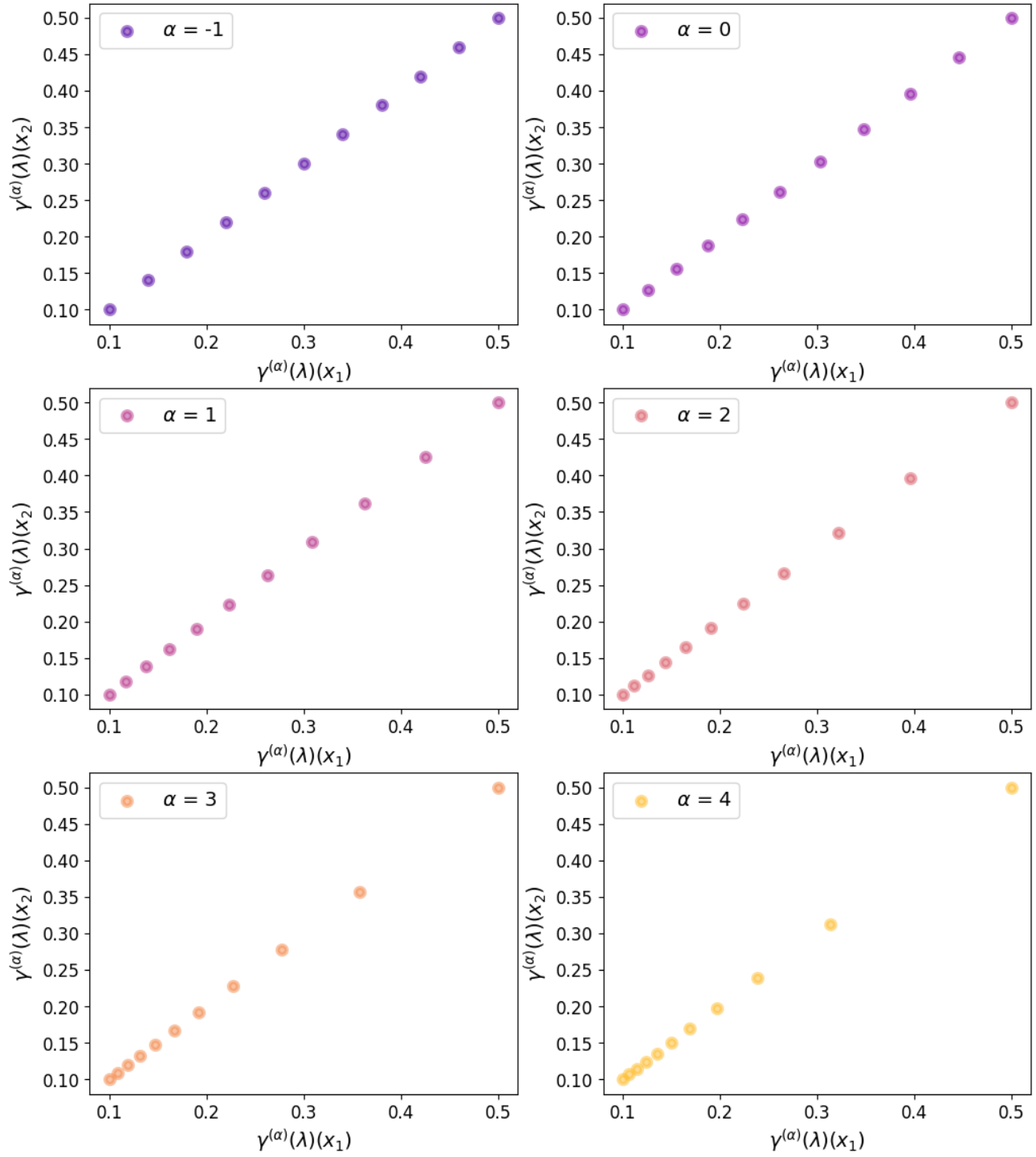

Figure 6: Examples of α-geodesics connecting $p_s(\boldsymbol{x})$ and $p_t(\boldsymbol{x})$ in two dimensional case. Here, we assume that $p_s(\boldsymbol{x}) = (0.1, 0.1)$ and $p_t(\boldsymbol{x}) = (0.9, 0.9)$.

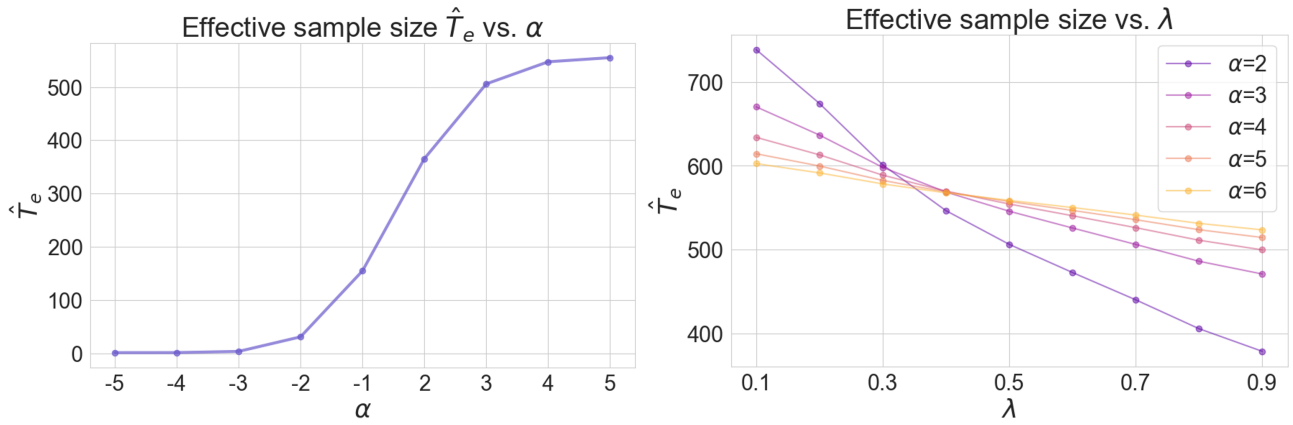

Figure 7: Relationship between ESS and the parameters α and λ.

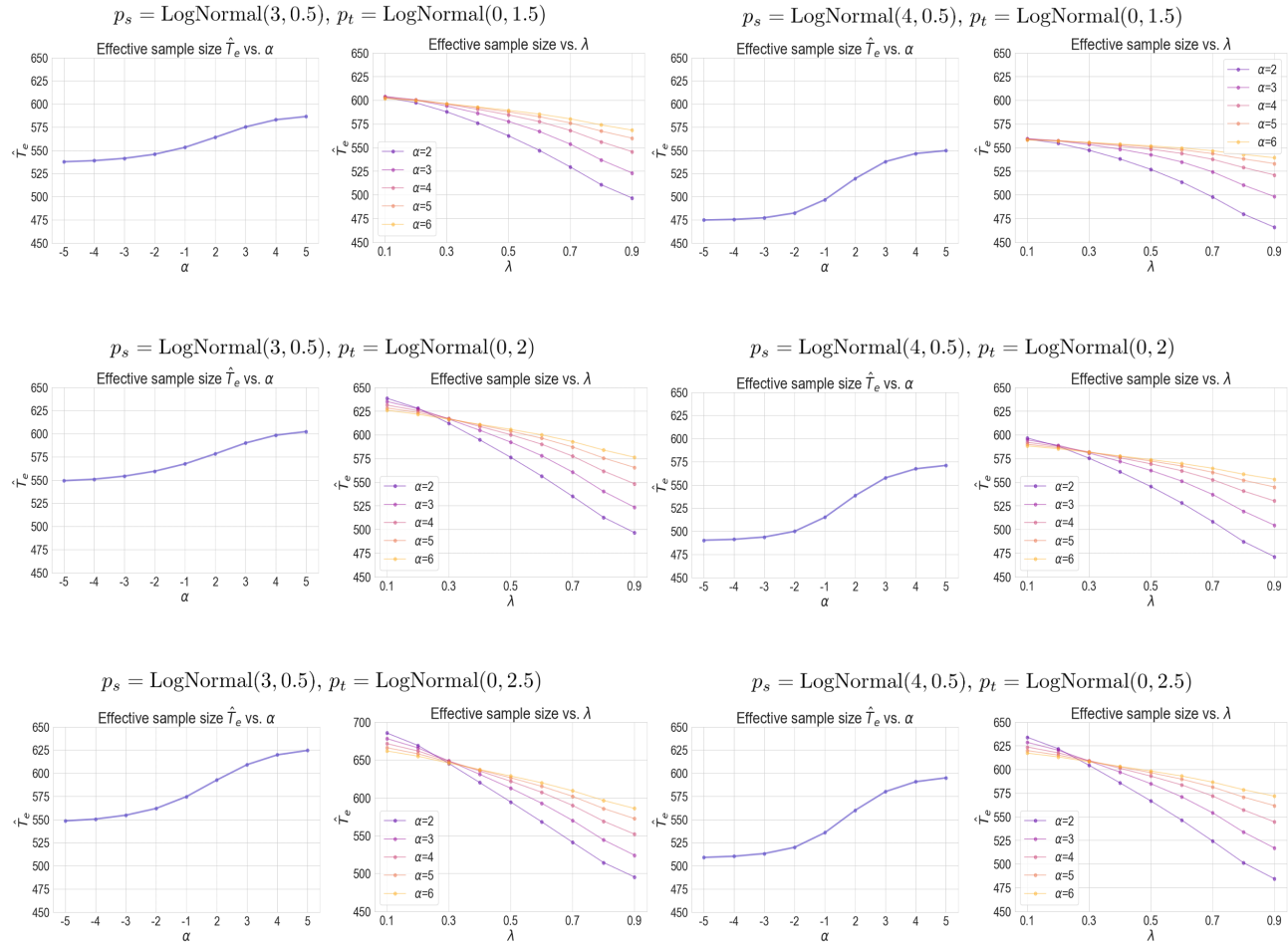

Figure 8: Relationship between ESS and the parameters α and λ for Log-normal distributions.

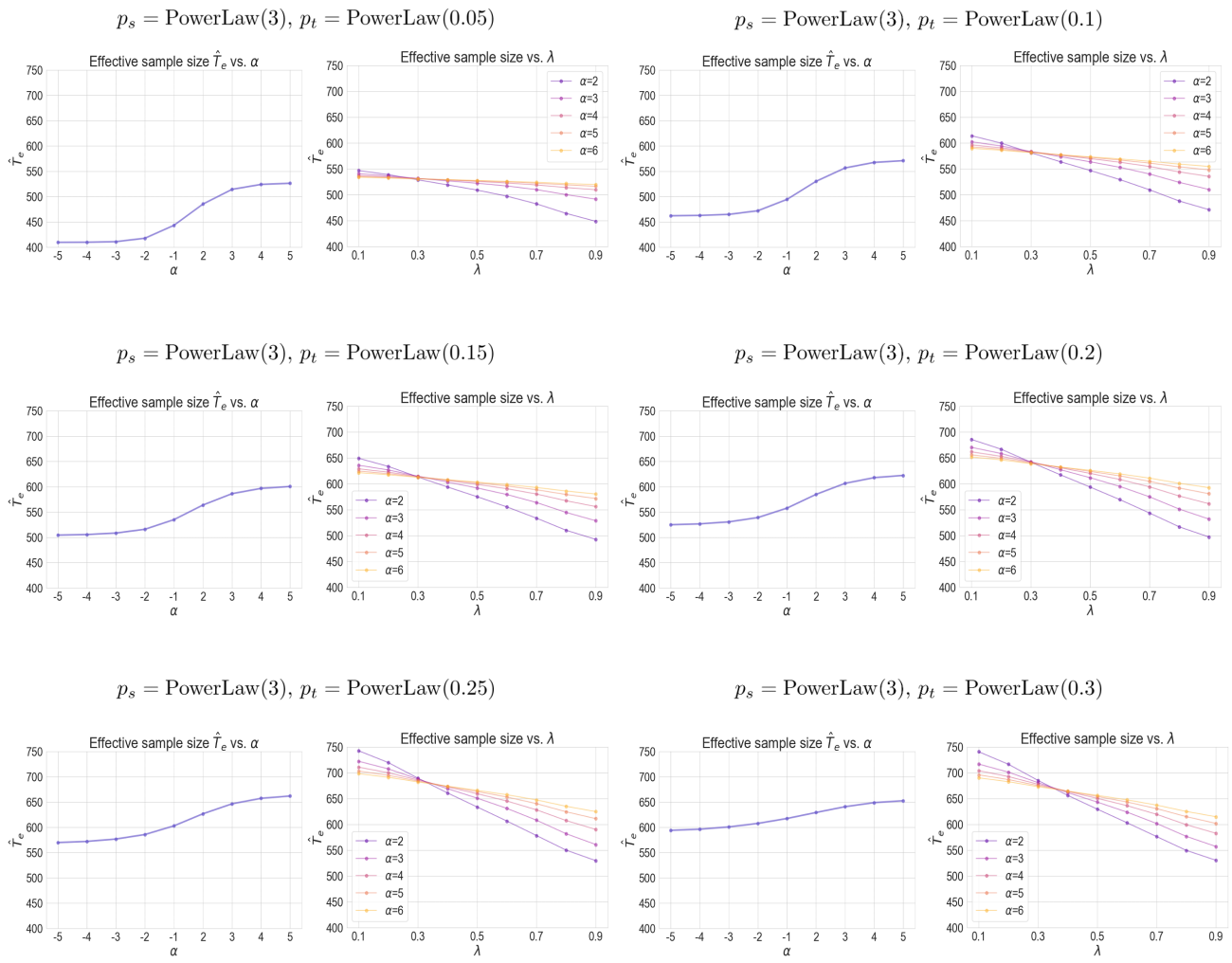

Figure 9: Relationship between ESS and the parameters α and λ for Power-law distributions.

C Details of Numerical Experiments and Additional Experimental Results

This appendix provides details of the numerical experiments and additional experimental results.

C.1 Implementation Details

C.1.1 Software

The results of all numerical experiments are obtained by our Python 3.11.0 [1] implementation. We use numpy (Harris et al., 2020) for random variable generation and matrix calculations, scipy (Virtanen et al., 2020) for computing the density function of probability distributions, matplotlib [2] and seaborn [3] for visualization of the results. In addition, scikit-learn (Pedregosa et al., 2011) is used to implement logistic regression, and default parameters are used unless otherwise noted.

C.1.2 Computing Resources

All our numerical experiments are performed on a machine with 16 GiB of system memory and 4 vCPUs

C.1.3 Base Model used for Density Ratio Estimation

In our experiments, we use a kernel logistic regression classifier as the density ratio estimator. A kernel logistic regression classifier employs a parametric model of the following form for expressing the class-posterior probability $p(y \mid \boldsymbol{x})$.

$$p(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{1 + \exp(-y\psi(\boldsymbol{x})^\top \boldsymbol{\theta})}, \tag{100}$$

where $\psi(\boldsymbol{x})\colon \mathbb{R}^d \to \mathbb{R}^c$ is a basis function vector and $\boldsymbol{\theta} \in \mathbb{R}^c$ is a parameter vector. In our experiments, we use the following basis functions corresponding to the polynomial kernel and B-spline kernel in addition to the linear kernel. The penalized log-likelihood maximization problem reduces the following minimization.

$$\hat{\boldsymbol{\theta}} \coloneqq \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^c} \left\{ \sum_{i=1}^{n} \ln\left(1 + \exp\left(-y_i \psi(\boldsymbol{x}_i)^\top \boldsymbol{\theta}\right)\right) + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta} \right\}, \tag{101}$$

where $\lambda \boldsymbol{\theta}^\top \boldsymbol{\theta}$ is a penalty term included for regularization process. Thus, the loss function $L(y_i, \boldsymbol{\theta})$ is

$$L(y_i, \boldsymbol{\theta}) \coloneqq \ln\left(1 + \exp\left(-y_i \psi(\boldsymbol{x}_i)^\top \boldsymbol{\theta}\right)\right). \tag{102}$$

From Bayes' theorem, the density ratio can be expressed as

$$\begin{aligned} r(\boldsymbol{x}) &= \frac{p_s(\boldsymbol{x})}{p_t(\boldsymbol{x})} \\ &= \left(\frac{p(y=+1 \mid \boldsymbol{x}) p(\boldsymbol{x})}{p(y=+1)}\right) \left(\frac{p(y=-1 \mid \boldsymbol{x}) p(\boldsymbol{x})}{p(y=-1)}\right)^{-1} \\ &= \frac{p(y=-1)}{p(y=+1)} \frac{p(y=+1 \mid \boldsymbol{x})}{p(y=-1 \mid \boldsymbol{x})}. \end{aligned} \tag{103}$$

The ratio $p(y=-1)/p(y=+1)$ can be approximated by

$$\frac{p(y=-1)}{p(y=+1)} \approx \frac{n_t/(n_t+n_s)}{n_s/(n_t+n_s)} = \frac{n_t}{n_s}. \tag{104}$$

[1] https://www.python.org/downloads/release/python-3110/
[2] https://matplotlib.org/
[3] https://seaborn.pydata.org/

A density ratio estimator $\hat{r}$ is then given by

$$\begin{aligned}\hat{r}(\boldsymbol{x}) &= \frac{n_t}{n_s}\frac{1+\exp\left(\psi(\boldsymbol{x})^\top\hat{\boldsymbol{\theta}}\right)}{1+\exp\left(-\psi(\boldsymbol{x})^\top\hat{\boldsymbol{\theta}}\right)} \\ &= \frac{n_t}{n_s}\frac{\exp\left(\psi(\boldsymbol{x})^\top\hat{\boldsymbol{\theta}}\right)\left\{\exp\left(-\psi(\boldsymbol{x})^\top\hat{\boldsymbol{\theta}}\right)+1\right\}}{1+\exp\left(-\psi(\boldsymbol{x})^\top\hat{\boldsymbol{\theta}}\right)} \\ &= \frac{n_t}{n_s}\exp\left(\psi(\boldsymbol{x})^\top\hat{\boldsymbol{\theta}}\right). \end{aligned} \tag{105}$$

C.1.4 Estimator of Pearson Divergence

The Pearson divergence $D_{\mathrm{PE}}[p\|q]$ is defined as follows.

$$D_{\mathrm{PE}}[p\|q] \coloneqq \frac{1}{2}\int_{\mathcal{X}}\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}-1\right)^2 q(\boldsymbol{x})d\boldsymbol{x}. \tag{106}$$

Here, we have

$$\begin{aligned} D_{\mathrm{PE}}[p\|q] &= \frac{1}{2}\int_{\mathcal{X}}\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}-1\right)^2 q(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{1}{2}\int_{\mathcal{X}}\left\{\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^2 - 2\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}+1\right\}q(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{1}{2}\int_{\mathcal{X}}\frac{p(\boldsymbol{x})^2}{q(\boldsymbol{x})} - 2p(\boldsymbol{x}) + q(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{1}{2}\int_{\mathcal{X}}\frac{p(x)^2}{q(\boldsymbol{x})}d\boldsymbol{x} - \int_{\mathcal{X}}p(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{2}\int_{\mathcal{X}}q(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{1}{2}\int_{\mathcal{X}}\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}p(\boldsymbol{x})d\boldsymbol{x} - \int_{\mathcal{X}}\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{2} \\ &= \frac{1}{2}\int_{\mathcal{X}}r(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} - \int_{\mathcal{X}}r(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{2} = \frac{1}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[r(\boldsymbol{x})\right] - \mathbb{E}_{q(\boldsymbol{x})}\left[r(\boldsymbol{x})\right] + \frac{1}{2}. \end{aligned}$$

The estimator of the Pearson divergence $\hat{D}_{\mathrm{PE}}[X^s\|X^t]$ is

$$\hat{D}_{\mathrm{PE}}\left[X^s\|X^t\right] \coloneqq \frac{1}{2n_s}\sum_{i=1}^{n_s}\hat{r}(\boldsymbol{x}_i^s) - \frac{1}{n_t}\sum_{i=1}^{n_t}\hat{r}(\boldsymbol{x}_i^t) + \frac{1}{2}. \tag{107}$$

Combining with the estimator by Eq. (105),

$$\begin{aligned}\hat{D}_{\mathrm{PE}}\left[X^s\|X^t\right] &= \frac{1}{2n_s}\sum_{i=1}^{n_s}\frac{n_t}{n_s}\frac{1+\exp\left(\psi(\boldsymbol{x}_i^s)^\top\hat{\boldsymbol{\theta}}\right)}{1+\exp\left(\psi(\boldsymbol{x}_i^s)^\top\hat{\boldsymbol{\theta}}\right)} - \frac{1}{n_t}\sum_{i=1}^{n_t}\frac{n_t}{n_s}\frac{1+\exp\left(\psi(\boldsymbol{x}_i^t)^\top\hat{\boldsymbol{\theta}}\right)}{1+\exp\left(\psi(\boldsymbol{x}_i^t)^\top\hat{\boldsymbol{\theta}}\right)} + \frac{1}{2} \\ &= \frac{n_t}{2n_s^2}\sum_{i=1}^{n_s}\frac{1+\exp\left(\psi(\boldsymbol{x}_i^s)^\top\hat{\boldsymbol{\theta}}\right)}{1+\exp\left(\psi(\boldsymbol{x}_i^s)^\top\hat{\boldsymbol{\theta}}\right)} - \frac{1}{n_s}\sum_{i=1}^{n_t}\frac{1+\exp\left(\psi(\boldsymbol{x}_i^t)^\top\hat{\boldsymbol{\theta}}\right)}{1+\exp\left(\psi(\boldsymbol{x}_i^t)^\top\hat{\boldsymbol{\theta}}\right)} + \frac{1}{2}. \end{aligned} \tag{108}$$

Table 4: Evaluation results with different α and sample dimensions d with polynomial kernel for Gaussian distributions. DRE refers the density ratio estimation without incremental mixtures.

	$d=2$	$d=3$	$d=4$	$d=5$
DRE	35.98($\pm$27.33)	1335.85($\pm$1447.51)	5038.17($\pm$2666.30)	48586.57($\pm$19828.46)
$\alpha=-1$	30.61($\pm$31.03)	715.52($\pm$143)	723.75($\pm$320.86)	1194.64($\pm$400.47)
$\alpha=3$	8.27($\pm$9.55)	63.14($\pm$80.29)	296.16($\pm$339.05)	480.41($\pm$89.07)
$\alpha=7$	6.09($\pm$0.02)	19.81($\pm$29.53)	42.52($\pm$126.12)	412.99($\pm$71)

Table 5: Evaluation results with different α and sample dimensions d with spline kernel for Gaussian distributions. DRE refers the density ratio estimation without incremental mixtures.

	$d=2$	$d=3$	$d=4$	$d=5$
DRE	38.13($\pm$4.04)	485.40($\pm$71.28)	5167.16($\pm$1710.77)	48656.67($\pm$14837.67)
$\alpha=-1$	6.09($\pm$0.01)	123.38($\pm$122)	489.74($\pm$409.68)	2767.53($\pm$3596.53)
$\alpha=3$	6.10($\pm$0.00)	7.69($\pm$0.01)	22.59($\pm$0.01)	119.83($\pm$0.02)
$\alpha=7$	6.10($\pm$0.00)	7.70($\pm$0.01)	22.60($\pm$0.01)	119.85($\pm$0.01)

C.1.5 Analytic Calculation of Divergence

In our numerical experiments, we compute KL-divergence between two Gaussian distributions $p(x)=\mathcal{N}(\mu_1,\sigma_1)$ and $q(x)=\mathcal{N}(\mu_2,\sigma_2)$ analytically. The calculation is performed as follows.

$$
\begin{aligned}
D_{\mathrm{KL}}[p\|q] &= \int_{\mathcal{X}} \{\ln p(x) - \ln q(x)\}\, p(x)dx \\
&= \int_{\mathcal{X}} \left\{ -\ln\sigma_1 - \frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \ln\sigma_2 + \frac{1}{2}\left(\frac{x-\mu_s}{\sigma^2}\right)^2 \right\} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} \\
&= \int_{\mathcal{X}} \left\{ \ln\frac{\sigma_2}{\sigma_1} + \frac{1}{2}\left(\left(\frac{x-\mu_2}{\sigma_2}\right)^2 - \left(\frac{x-\mu_1}{\sigma_1}\right)^2\right) \right\} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} dx \\
&= \mathbb{E}_{p(x)}\left[\ln\frac{\sigma_2}{\sigma_1} + \frac{1}{2}\left(\left(\frac{x-\mu_2}{\sigma_2}\right)^2 - \left(\frac{x-\mu_1}{\sigma_1}\right)^2\right)\right] \\
&= \ln\frac{\sigma_2}{\sigma_1} + \frac{1}{2\sigma_2^2}\mathbb{E}_{p(x)}\left[(x-\mu_2)^2\right] - \frac{1}{2\sigma_1^2}\mathbb{E}_{p(x)}\left[(x-\mu_1)^2\right] \\
&= \ln\frac{\sigma_2}{\sigma_1} + \frac{1}{2\sigma_2^2}\mathbb{E}_{p(x)}\left[(x-\mu_2)^2\right] - \frac{1}{2} \\
&= \ln\frac{\sigma_2}{\sigma_1} + \frac{1}{2\sigma_2^2}\left\{\mathbb{E}_{p(x)}\left[(x-\mu_1)^2\right] + 2(\mu_1-\mu_2)\mathbb{E}\left[(x-\mu_1)\right] + (\mu_1-\mu_2)^2\right\} - \frac{1}{2} \\
&= \ln\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.
\end{aligned} \tag{109}
$$

C.2 Additional Experimental Results

In additional experiments, we report results based on density ratio estimation by nonlinear kernel logistic regression. We use the following two-dimensional polynomial and cubic spline kernels.

$$
\begin{aligned}
K_{\mathrm{polynomial}}(\boldsymbol{u},\boldsymbol{v}) &\coloneqq (\boldsymbol{u}^\top\boldsymbol{v}+c)^2, \\
K_{\mathrm{spline}}(\boldsymbol{u},\boldsymbol{v}) &\coloneqq \|\boldsymbol{u}-\boldsymbol{v}\|^3.
\end{aligned} \tag{110}
$$

	$d = 100$	$d = 200$	$d = 300$	$d = 400$	$d = 500$
$\alpha = 3$	864.3($\pm$14.6)	942.7($\pm$18.3)	980.6($\pm$25.2)	1002.4($\pm$33.4)	1058.4($\pm$42.5)
$\alpha = 4$	858.6($\pm$12.8)	925.5($\pm$15.9)	964.2($\pm$23.8)	994.4($\pm$31.8)	1050.3($\pm$40.6)
$\alpha = 5$	841.6($\pm$10.7)	909.9($\pm$15.2)	955.5($\pm$22.3)	978.3($\pm$29.1)	1035.9($\pm$38.5)
$\alpha = 6$	823.6($\pm$9.4)	896.5($\pm$14.1)	949.7($\pm$20.1)	971.6($\pm$27.4)	1034.1($\pm$36.8)
$\alpha = 7$	814.4($\pm$9.1)	887.9($\pm$13.5)	940.7($\pm$18.2)	968.0($\pm$26.0)	1028.2($\pm$34.6)

Table 6: Evaluation results with different α and sample dimensions d in the case of higher dimensions.

Tables 4, 5, 6 and Figures 10, 11, 12 show the results of density ratio estimation based on nonlinear kernel logistic regression. The settings in these experiments are the same as in the main manuscript. These results show that density ratio estimation along geodesics with large α reduces variances, even when nonlinear kernels are used.

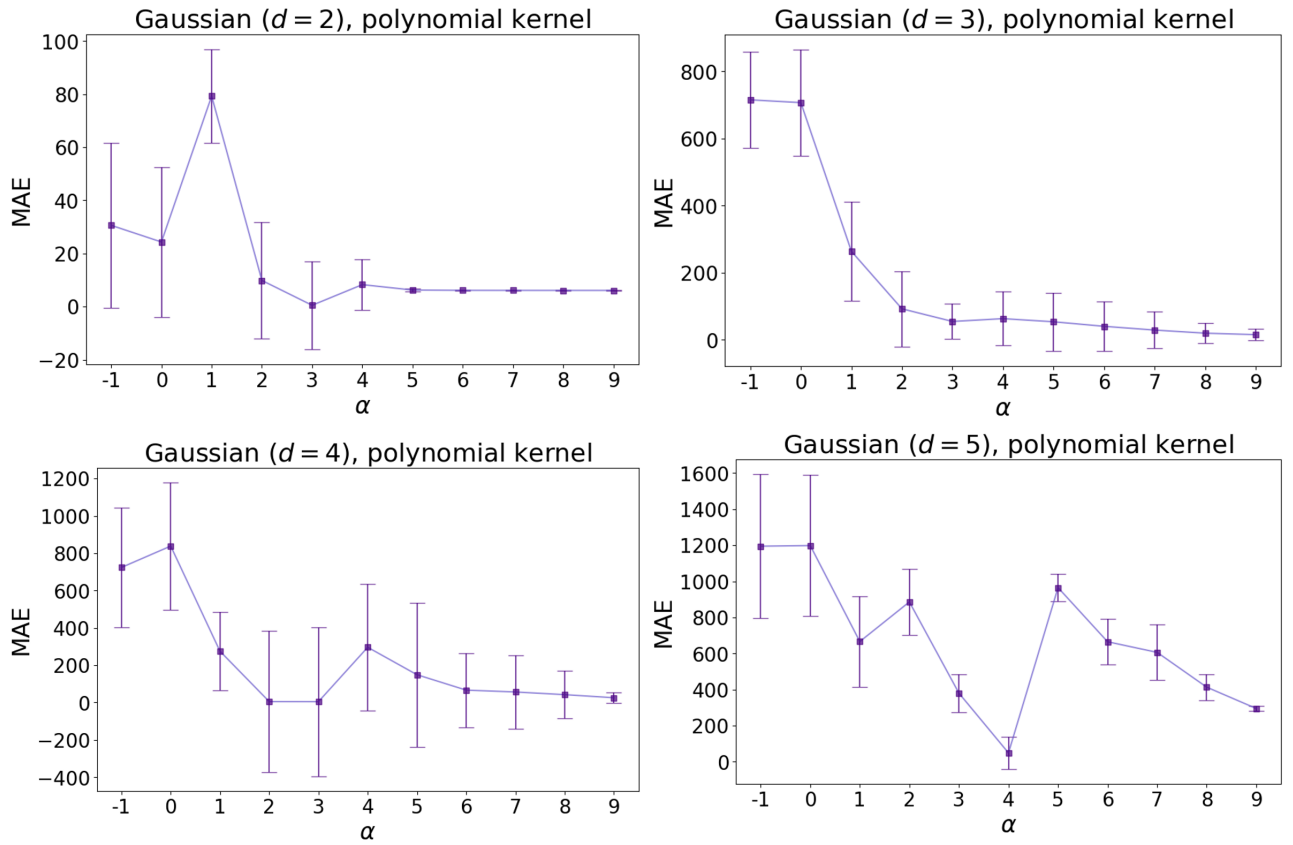

Figure 10: Evaluation results with different α and sample dimensions d with polynomial kernel for Gaussian distributions.

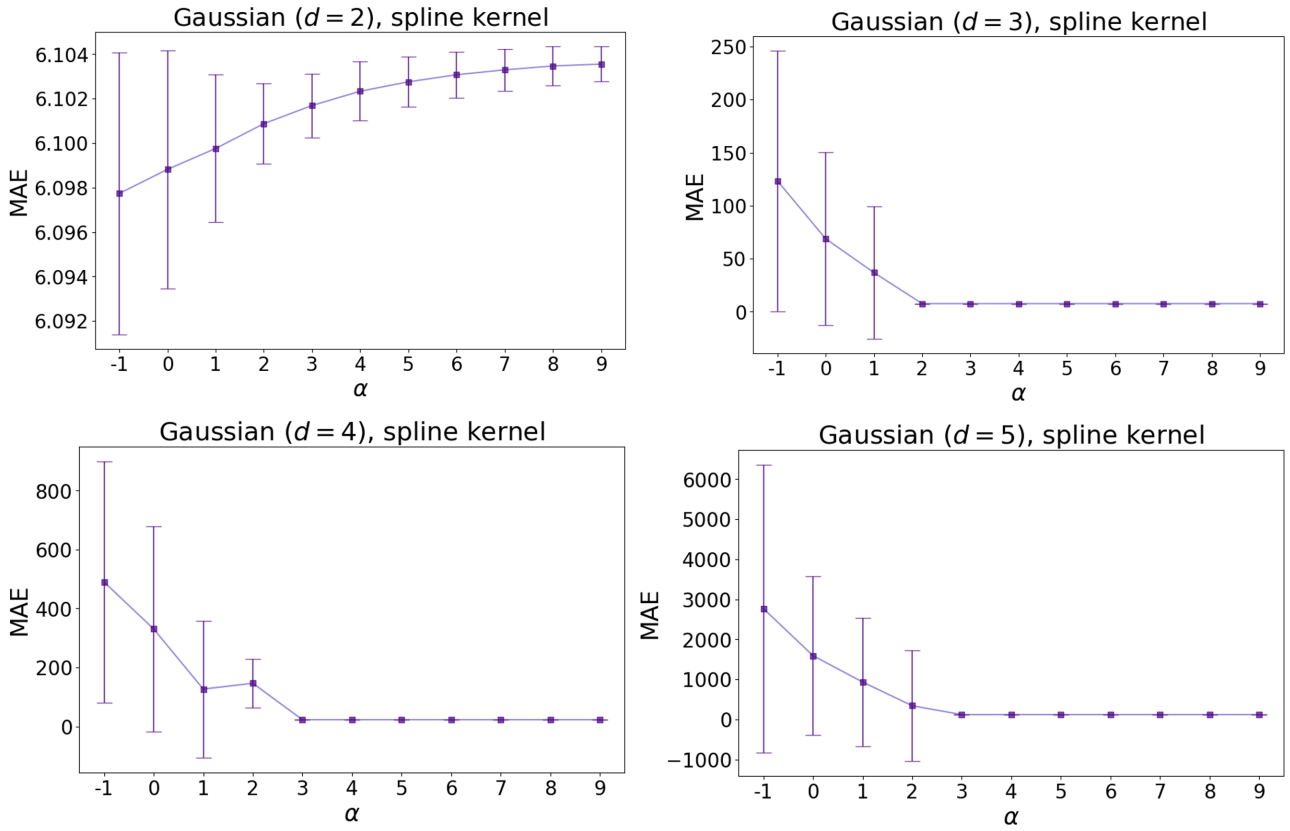

Figure 11: Evaluation results with different α and sample dimensions d with spline kernel for Gaussian distributions.

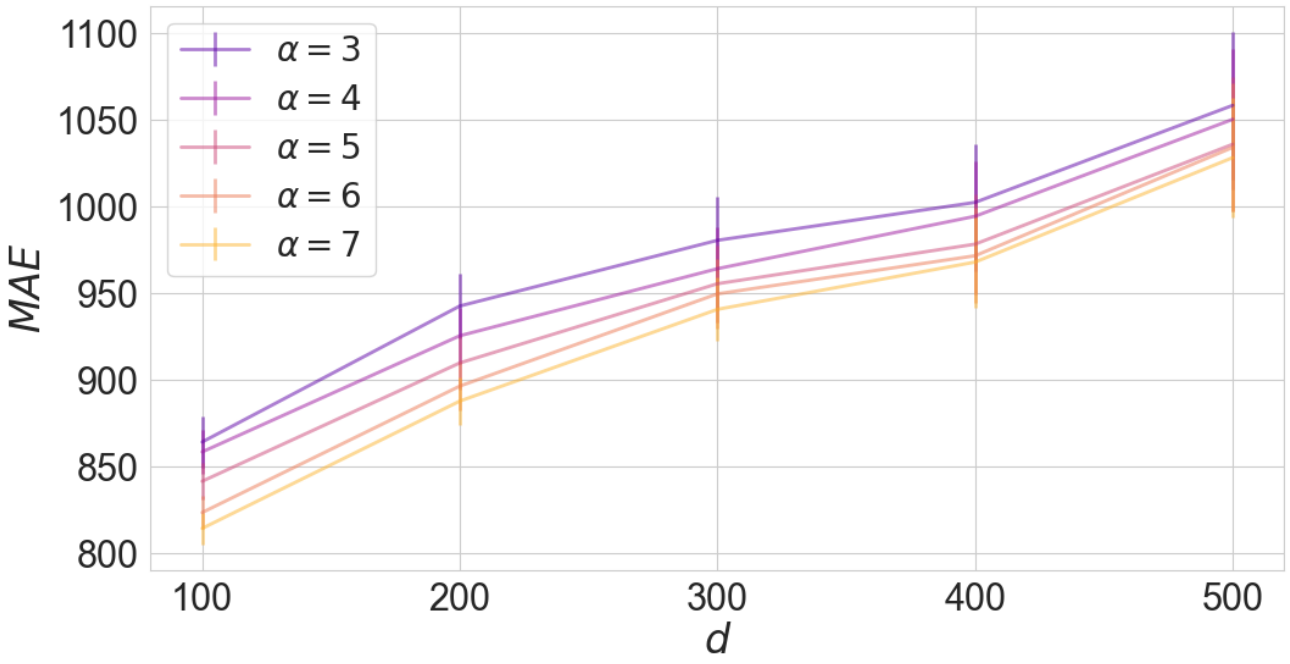

Figure 12: Evaluation results with different α and sample dimensions d in the case of higher dimensions.

D Related Work

D.1 Applications of Density Ratio

The density ratio of the two probability distributions has many known applications. Shimodaira (2000) shows that in supervised learning under the covariate shift assumption, where the marginal distribution of the input data differs between training and testing, learning by importance weighting using density ratios restores consistency. This framework is called Importance Weighted Empirical Risk Minimization (IWERM), and many variants have been investigated (Chen et al., 2016; Kimura and Hino, 2022, 2024; Liu and Ziebart, 2014; Martin et al., 2023; Sugiyama et al., 2007a; Yamada et al., 2013). Density ratios are also known to be effective for outlier detection, and many studies have used estimated density ratios from samples as outlier scores (Chen et al., 2015; Hido et al., 2011; Kanamori et al., 2008; Li et al., 2022). Another useful application of the density ratio is the two-sample test, which considers whether, given two sample pairs, they are generated from the same distribution (Cheng and Chu, 2004; Kanamori et al., 2011; Keziou and Leoni-Aubin, 2005, 2008; Lu et al., 2020). Further applications of density ratios include variable selection (Oh et al., 2016), dimensionality reduction (Suzuki and Sugiyama, 2010), causal inference (Matsushita et al., 2023), and estimation of mutual information (Braga, 2014; Suzuki et al., 2008). Many of them are also discussed theoretically in the use of density ratios.

D.2 Methods of Density Ratio Estimation

Because of this wide range of applications, density ratio estimation has become one of the most important tasks. The simplest idea is to estimate the density ratio based on a separate density estimation, but these two-step approaches are known to be computationally unstable (Sugiyama et al., 2012a). A more promising idea is a framework for direct density ratio estimation. One idea is the moment matching approach which tries to match the moments of $p_s(\boldsymbol{x})$ and $p_t(\boldsymbol{x})$ by considering $\hat{r}(\boldsymbol{x})$ as a transformation function (Huang et al., 2006). Another well-known approach to density ratio estimation is to optimize the divergence between one distribution $p_s(\boldsymbol{x})$ and the transformed other $\hat{r}(\boldsymbol{x})p_t(\boldsymbol{x})$ (Sugiyama et al., 2012b, 2007b). Recently, algorithms for density ratio estimation based on generative models have also been proposed (Choi et al., 2021).